SHORT TEXT TOPIC MODELING WITH EMPIRICAL LEARNING

Supriya A. Kinariwala
Department of Computer Science and Engineering
Marathwada Institute of Technology, Aurangabad, Maharashtra, INDIA
sakinariwala@gmail.com

Sachin N. Deshmukh
Department of Computer Science and IT
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, INDIA
sndeshmukh@hotmail.com

Abstract - In the present modern digital era, use of social media has been increasing exponentially. People have started using short text for expressing their thoughts. Social media websites like Twitter, Facebook are generating vast amount of short text at every second that reveals good knowledge of real time information. Extensive research is going on to discover knowledge from it. Short text is very sparse and ambiguous; hence there is a big challenge to find latent topics from it. This can be resolved by using unsupervised machine learning approach referred as topic modeling. This paper covers various topic modeling methods like Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Semantics-assisted Non-negative Matrix Factorization (SeaNMF) and their comparative analysis. These three methods have been tested on ABCNews headline dataset, results have been analyzed using average Normalized Google Distance (NGD) score; which is 67.88%, 58.60%, 59.32% for SeaNMF, NMF and LDA respectively. The quantitative result shows that more meaningful and semantically similar words are clustered under each topic by SeaNMF model.

Keywords:Topic Modeling, Short text, Latent Dirichlet Allocation, Non-negative Matrix factorization, Semantic assisted NMF.

1. Introduction

Today's world has witnessed the global rise in the use of technology like Social Media tools .As a result of this lot of unstructured data is generated over the World Wide Web. This vast amount of electronic data can be made useful using text mining which is a process of extracting high-quality content from collections of documents. In order to acquire required information from this data, it is needed to identify the relevant documents. To mine such a pattern of words from a given set of documents, a statistical model called topic model is used [Likhitha *et al.* , (2019)]. Topic modeling is an unsupervised machine learning technique which helps in:

- Discovering hidden topical patterns that are present across the collection
- Annotating documents according to these topics
- These annotations are used in organizing, searching and summarizing text

To cluster words from a set of documents, topic modeling relies on the bag-of-words, their frequency and not on order of words[Alghamdi and Alfalqi, (2015)]. Various techniques used to obtain topic models are Latent Dirichlet Allocation, Non-negative Matrix Factorization, and Semantics-assisted Non-negative Matrix Factorization.

The rest of this paper is organized as follows; Section 2 presents various topic modeling methods. The experiments and results of topic model methods on ABCNews headline dataset are given in Section 3. Finally Section 4 concludes the paper.

2. Methods in Topic Modeling

2.1 Latent Dirichlet Allocation (LDA)

LDA is probabilistic generative topic model used to identify topics in a given document. It works on the assumption that similar words are used to represent similar topics and each document as a mixture oftopics that represent whole corpus. The model considers that each word is mapped to at least one of the topic of document [Blei *et al.* ,(2003)].

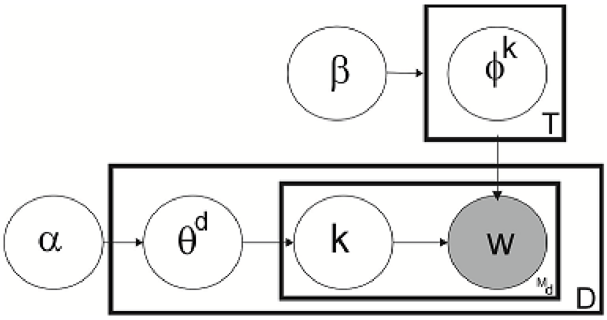

Fig. 1: Graphical Representation of LDA

In LDA documents are considered as probability distribution over topics and further each topic in a document as probability distribution over words. Having T topics, the probability distribution over ithword(w) in a document is given as in Eq.(1),

$$P(w_i) = \sum_{j=1}^{T} P(w_i \,|k_i = j) P(k_i = j) (1)$$

wherek_i is the topic $j \in T$, which is a latent variable from which i^{th} word is taken. $P(w_i \mid k_i = j)$ is the probability of i^{th} word; under topic j. $P(k_i{=}j)$ is the probability of choosing topic j under given document, which can be different for various documents[Blei *et al.* ,(2003)];[Chhatbar,(2010)].

$P(w|k)$ can be represented as a set of T multinomial distributions ϕ over W unique words, such that $P(w|k = j) = \phi_w^j$and $P(k)$ as D multinomial distribution θ over T available topics, such that for each word in document $d \in D$, $P(k = j|d) = \theta_j^d$. Multinomial distributions word-topic (ϕ) and topic-document (θ) are computed as [Chhatbar,(2010)]:

$$\theta_j^{(d)} = \frac{M_{dj}^{DT} + \alpha}{\sum_{t=1}^{T} M_{dt}^{DT} + T\alpha} \qquad (2)$$

$$\phi_i^{(j)} = \frac{M_{ij}^{WT} + \beta}{\sum_{w=1}^{W} M_{wj}^{WT} + W\beta} \qquad (3)$$

Using Gibbs sampling, along with Dirichlet Prior on ϕ and θ , probabilities can be estimated by following equation

$$P(k_i = j \,|\, k_{-i,} W_{-i,} \ldots) = \frac{M_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} M_{d_i t}^{DT} + T\alpha} \frac{M_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} M_{wj}^{WT} + W\beta} \qquad (4)$$

Here, considering probability of topic $k_i = j$ given all other word and document assignments. α is a prior weight of topic in a document, is initialized as $\alpha<1$ to prefer few topics per document. β is a prior weight of word w in topic, is initialized as very less than 1 to prefer fewer words per topic. W_{-i} stands for all other word instances than the current one. M^{WT} and M^{DT} are matrices of counts for word-topic and document-topic assignments respectively. These steps are repeated till it reached a steady state where assignments are good. These assignments are then used to determine word-topic matrix which gives words from each document and document-topic matrix which yields belongingness of each document to particular topic [Chhatbar,(2010)].For a given collection of documents posterior distribution of the hidden variables determines topic-wise distribution of documents. These hidden variables are used in information retrieval and document browsing. LDA has achieved great results in modelling collection of normal length text like news article, research papers and blogs [Zuo *et al.*, (2016)].

2.2Non-negative Matrix Factorization

Non-negative matrix factorization is a linear algebraic model used for dimensionality reduction. This method is suitable where underlying factors are non-negative. As NMF yields good clustering results for high dimensional data, it is used for topic modeling [Yan *et al.*,(2013)];[Choo *et al.,(2013)*]. Given N text documents, are represented as the term-document matrix in which each column represents a document and each element of matrix is the weight calculated through tf-idf [Kuang *et al.,(2015)*]. NMF decomposes given original matrix *D* (term-document matrix) into two matrices *W* (word-topic matrix) and Z (document-topic matrix) such that,

$$D_{M\times N} = W_{M\times K} * Z_{K\times N} \tag{5}$$

Consider a corpus with *N* documents and *M* distinct words in vocabulary. Term-document matrix is defined as $D \in R_+^{M\times N}$, where R_+ denotes positive real numbers. Representation of bag of words of document j in terms of *M* keywords is given by column vector $D_{(:,j)} \in R_+^{M\times 1}$. The term-document matrix is $D \approx W Z^T$, where $W \in R_+^{M\times K}$ and $Z \in R_+^{N\times K}$, $K << min(M, N)$ is the number of latent factors (i.e., topics). Usually, this approximation can be devised as in Eq. (6) [Choo et al.,(2015)]:

$$\min_{W,Z\geq 0} ||D - WZ^T||_F^2 \tag{6}$$

Elements in column vector of term-topic matrix $W_{(:,k)} \in R_+^{M\times 1}$ are weights of M^{th} keyword under K^{th} topic , and row vector $W_{(i,:)} \in R_+^{1\times K}$ is semantic representation of word *i*. Row vector of document-topic matrix $Z_{(j,:)} \in R_+^{1\times K}$ represents weights for j^{th} document corresponding to K^{th} topic. As short text is very sparse, the factor matrices *W* and *Z* are updated using Block Coordinate Descent (BCD) algorithm. It is a divide-and-conquer strategy which divides data in blocks and updates data block by block [Kim *et al.,(2014)*].

Update W:

$$W_{(:,k)} = \frac{W_{(:,k)} + (DZ)_{(:,k)} - (WZ^T Z)_{(:,k)}}{(Z^T Z)_{(k,k)}} \tag{7}$$

Update Z:

$$Z_{(:,k)} = \frac{Z_{(:,k)} + (D^T W)_{(:,k)} - (ZW^T W)_{(:,k)}}{(W^T W)_{(k,k)}} \tag{8}$$

The association between different documents strongly depends on keywords and vice-versa. In short text each document has few keywords and NMF does not consider semantic relationship between keywords, hence clustering performance is poor [Shi *et al.*,(2018)].

2.3 SeaNMF Model

The Semantic-assisted NMF (SeaNMF) is a method which reveals semantic relationship between keywords and their context to discover topics from short text. During training, word embedding is used to find semantic association between keyword and their contexts. Semantic correlation matrix(S) between word-context is obtained from vocabulary of words (*V*) using skip gram model [Levy and Glodberg,(2014)];[Milolov *et al.*,(2013)].

$$S_{ij} = \left[\log\left(\frac{\#(w_i, c_j)}{\#(w_i).p(c_j)}\right)\right] \tag{9}$$

where $w_i \in V$, $c_j \in V$, , p(c_j) is a distribution of words based on its frequency in corpus.

In SeaNMF algorithm bag-of-word representation is used to construct term-document matrix (*D).* Then, semantic correlation matrix *S* is computed by Eq.(9). The latent factor matrices of words (*W),* Context *(W_c)* and documents (*Z)* are randomly initialized with non-negative numbers. In each iteration, their weights are updated by using Block Coordinate Decent (BCD) algorithm [Kim *et al.,(2014)*]. Updated weights $W_{(:,k)}$ and $W_{c(:,k)}$ are normalized by F -norm. Process is repeated until the algorithm converges [Shi *et al.*,(2018)]. The *W,* W_c ,*Z* are updated for k topics as in Eq. (10) and (11) :

Update W

$$W_{(:,k)} = \frac{W_{(:,k)} + (DZ)_{(:,k)} + \alpha(SW_c)_{(:,k)} - (WZ^T Z)_{(:,k)} - \alpha(WW_c^T W_c)_{(:,k)}}{(Z^T Z)_{(k,k)} + \alpha(W_c^T W_c)_{(k,k)}} \tag{10}$$

Update Wc

$$W_{c(:,k)} = \left[W_{c(:,k)} + \frac{(SW)_{(:,k)} - (W_c W^T W)_{(:,k)}}{(W^T W)_{(k,k)}}\right] \tag{11}$$

where $\alpha \in R_+$ is ascale parameter.

Consideration of word-context semantic relationship in the overall updating procedure yields highly correlated top keywords under each topic.

3. Results and Discussion

3.1 Dataset Used

Experimentation is carried out on the ABC Millions Headlines dataset published on Kaggle, sourced from Australian Broadcasting Corporation (ABC). This contains Million news headlines with published date and headline text. It includes the entire corpus of articles published by the ABC website in the given time range 2003-2019 with focus on international news. Headlines from various domains like financial crisis, Iraq war, various elections, ecological disasters, terrorism, famous people, local crimes etc. are covered [Kulkarni (2017)].

3.2 Results

This section shows performance of various topic modeling algorithms on ABCNews dataset.

For experimentation 2,57,848 headlines are considered. Dataset is preprocessed which includes tokenization, removal of stop words, lemmatization [Vijayarani *et al.*,(2016)]. This creates dictionary of words. It's tf-idf score is calculated and given as input to LDA, NMF and SeaNMF model. Topics discovered by these topic modeling methods are listed in Table 1, Table 2 and Table 3.

Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
Cup	South	search	High	killed	claims	set	govt	Water	police
world	Talks	continues	Rise	crash	inquiry	opposition	council	Probe	man
decision	Howard	missing	market	two	rejects	gold	plan	Group	court
Win	Qld	family	Anti	day	New	return	urged	Ban	face
takes	Iraq	open	Year	dead	Aid	coast	health	Safety	charged
Toll	North	Body	Price	three	Govt	attacks	boost	New	murder
Test	Aust	accused	workers	one	Chief	clash	funds	Call	drug
Rate	West	sought	Strike	attack	drought	abuse	new	Council	death
Park	East	Lead	Prices	injured	Fire	iraqi	indigenous	Warning	charges
australia	Blaze	First	expected	hopes	farmers	debate	public	Prompts	Case

Table 1: Top 10 keywords for 10 topics discovered by LDA

Table 2: Top 10 keywords for 10 topics discovered by NMF

Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
iraq	war	police	Govt	man	new	says	council	Iraqi	world
troops	anti	death	Nsw	court	plan	qld	water	Baghdad	cup
baghdad	plan	anti	Qld	death	water	troops	plan	Troops	win
war	world	nsw	Plan	qld	court	baghdad	claims	Claims	claims
claims	Nsw	claims	claims	water	qld	win	qld	Plan	rain
win	Win	water	water	win	nsw	rain	rain	Rain	death
world	claims	win	death	rain	baghdad	death	court	Water	water
water	iraqi	qld	Rain	nsw	death	court	win	Death	anti
death	govt	rain	Anti	world	win	water	world	War	nsw
man	baghdad	troops	court	govt	claims	cup	iraq	World	police

Table 3: Top 10 keywords for 10 topics discovered by SeaNMF

Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
new	Govt	police	Council	downer	Man	win	Weather	market	killed
us	urged	probe	Plan	un	Court	edge	Damage	Prices	kills
council	Plan	investigate	Govts	blair	charged	draw	Cyclone	Stocks	least
laws	Nsw	man	Proposal	annan	murder	finals	Ses	Profit	dead
plan	Vic	missing	Proposed	peace	Jailed	blues	firefighters	markets	blast
zealand	Says	crash	Mp	pm	charges	tigers	Winds	Asx	kill
centre	Wa	search	Consultation	powell	Guilty	final	Flooding	higher	baghdad
chief	Qld	car	Councilor	eu	Bail	match	Storms	profits	bomb
iraq	funds	seek	Unhappy	bush	alleged	clash	Flood	Price	kashmir
may	Fire	death	Education	iran	Teen	play	Crews	Sales	wounded

3.3 Topic Model Evaluation

Evaluation of topic model is done by topic coherence [Ramirez and Brena(2011)];[Cilibrasi and Vitanyi (2001)]. There are two measures in topic coherence Intrinsic and Extrinsic. Intrinsic measure needs ordered word set as it compares word with its preceding and succeeding words. In extrinsic measure every word is paired with every other word under given set [Alhawarat and Hegazi (2018)];[Stevens and Kegelmeyer(2012)];[Roder and Both(2015)].

The Normalized Google Distance (NGD) is an extrinsic semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords[Cohen and Vitanyi(2013)]. Keywords with similar or same meaning in natural language sense tend to be “close” in units of NGD, while words with dissimilar meaning tend to be farther apart. The NGD score is calculated by the formula:

$$NGD(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log(x,y)}{\log M - min\{\log f(\mathrm{x}), \log f(y)\}} \tag{13}$$

where $f(x)$ and $f(y)$ are frequency of term x and term y, M is overall number of web pages indexed by Google. NGD(x,y) is a nonnegative score [Alguliev *et al. ,(2011)*].

The set of Keywords under each topic obtained after execution of LDA, NMF and SeaNMF are compared using NGD. Following results are obtained:

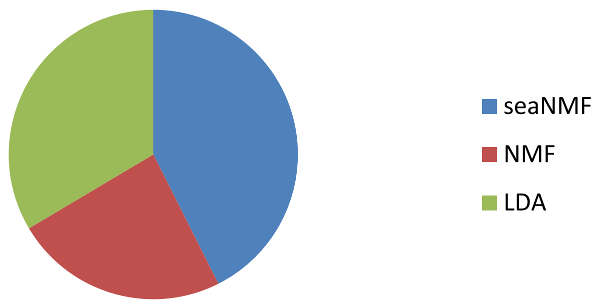

Fig 2: Number of correlated terms

The terms having NGD greater than 0 and less than 1 are said to be correlated terms.[Cilibrasi and Vitanyi (2001)] Fig. 2 shows SeaNMF is giving more correlated terms as compare to NMF and LDA.

Degree of correlativity of keywords is calculated by computing averaged NGD score per topic for all mentioned methods, results are listed in Table 4:

Table 4: Average NGD of 10 Topics

Topic	SeaNMF	NMF	LDA
1	0.8905	0.6897	0.5529
2	0.8332	0.6397	0.3504
3	0.6167	0.4242	0.5014
4	0.5226	0.7585	0.5163
5	0.8643	0.5032	0.6776
6	0.6194	0.5933	0.8502
7	0.7359	0.4223	0.7262
8	0.6577	0.6186	0.7947
9	0.5372	0.7205	0.3189
10	0.5105	0.4663	0.6434
Average	**0.6788**	**0.58363**	**0.5932**

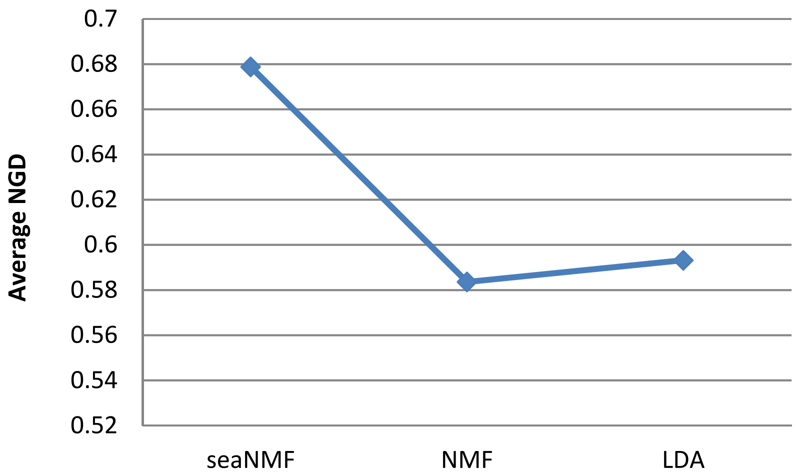

Fig 3: Average NGD for 10 Topics

The degree of correlativity of keywords computed is 67.88%, 58.6% ,59.32% for SeaNMF, NMF and LDA respectively. Fig. 3 shows that words clustered under each topic by SeaNMF are highly correlated.

4. Conclusions

Learning meaningful topics from short text is considered to be a challenge due to limited contextual information in it. This paper includes empirical study of three state-of- the-art methods of topic modeling. LDA is good for normal length text but not so for short text as it does not consider the relationships among keywords during topic discovery. NMF is a dimension reduction technique which yields clustering results based on the words in same region using term-document matrix whereas SeaNMF gives grouping of words using word-context semantic correlation matrix and skip-gram view of corpus that reveals word semantic association.SeaNMF outperforms NMF and LDA as it discovers more relevant topics from short text.

References

[1] Alghamdi Rubayyi, Khalid Alfalqi, (2015), "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, pp 147-153
[2] Alguliev Rasim, Aliguliyev Ramiz, Makrufa S Hajirahimova, Chingiz A Mehdiyev,(2011) "MCMR: Maximum Coverage and Minimum redundant text summarization model", Article in Expert Systems with Applications, Elsevier.
[3] Alhawarat M., Hegazi M.(2018), "Revisiting K-means and topic modeling, a comparison study to Cluster Arabic Documents", IEEE Access.
[4] Blei D, A. Ng, M. Jordan,(2003),"Latent Dirichlet Allocation", Journal of Machine Learning Research, 3: 993-1022
[5] Chhatbar Cirag Dilip, (2010), "Improving Statistical Topic Models by Using Ontological Concepts", COMP8740 Project Report, Dept., Computer Science, Australian National University.
[6] Choo Jaegul, Changhyun Lee, Chandan K. Reddy, Haesun Park,(2013), "Utopian:User-driven Topic Modeling based on Interactive Non-negative Matrix Factorization", IEEE transactions on Visualization and Computer Graphics, vol.19
[7] Choo Jaegul, Changhyun Lee, Chandan K. Reddy, Haesun Park, (2015) "Weakly supervised nonnegative matrix factorization for user-driven clustering", Data Mining and Knowledge Discovery, 1598–1621.
[8] Cilibrasi Rudi, Vitanyi Paul,(2001) "Automatic Meaning Discovery Using Google", BSIK/BRICKS project, Netherland
[9] Cohen Andrew R. and Paul M. B. Vitanyi,(2013), "Normalized Google Distance of Multisets with Applications", CoRR, abs|1308.3177
[10] Kim Jingu, Yunlong He,and Haesun park,(2014), "Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework", Journal of Global Optimization 58, 2(2014), 285-319
[11] Kuang Da, Jaegul Choo, and Haesun Park,(2015) "Nonnegative matrix factorization for interactive topic modeling and document clustering", In Partitional Clustering Algorithms, Springer, 215 - B
[12] Kulkarni Rohit,(2017), A Million News Headlines [CSV Data file], doi:10.7910/DVN/SYBGZL,Retrieved from: https://www.kaggle.com/therohk/million-headlines
[13] Levy Omer and Goldberg Yoav,(2014), "Neural Word Embedding as Implicit Matrix Factorization". In Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2177–2185.
[14] Likhitha S. , B. S. Harish, H. M. Keerthi Kumar,(2019), "A Detailed Survey on Topic Modeling for Document and Short Text Data", International Journal of Computer Applications, vol. 178, no. 39
[15] Milolov Tomas, Chen Kai, Corrado Greg, Dean Jeffrey,(2013), "Efficient Estimation of word representation in vector space", arXiv preprint arXiv:1301.3781
[16] Ramirez Eduardo H., Brena Ramon,(2011), "Topic Model Validation", Elsevier.
[17] Roder Michael, Both Andreas, (2015),s "Exploring the space of topic coherence Measures", ACM
[18] Shi Tian, Kyeongpil Kang , Jaegul Choo, Chandan Reddy, (2018),"Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations", ACM ISBN 978-1-4503-5639-8.
[19] Stevens Keith, Kegelmeyer Philip, (2012), "Exploring Topic Coherence over many models and many topics", proceeding of 2012 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952-961
[20] Dr. Vijayarani S., J. Ilamathi, Ms. Nithya , (2016)," Preprocessing Techniques for Text Mining-An Overview", International Journal of Computer Science & Communication Networks, vol 5.
[21] Yan Xiaohui, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, Yanfeng Wang,(2013), "Learning Topics in short Text by Non-negative Matrix Factorization on Term Correlation Matrix" , In Proceedings of the SIAM international Conference on Data Mining.
[22] Zuo Yuan, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, Hui Xiong, (2016),"Topic Modeling for Short Text: A Pseudo-Document View", KDD'16, ACM, August 13-17