ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse105852019.pdf
https://doi.org/10.30534/ijatcse/2019/105852019

Efficient Topic Level Opinion Mining and Sentiment Analysis Algorithm using Latent Dirichlet Allocation Model

Vamshi Krishna.B[1], Dr. Ajeet Kumar Pandey[2], Dr. A. P Siva Kumar[3]
[1]Research Scholar, Dept. of CSE, JNTUA, Anantapuramu, India, vamshikrishna.bs@gmail.com
[2] Delivery Manager, Dept. of RAMS, L&T Technology Technology Services, India, ajeet.mnnit@gmail.com
[3]Asst. Professor, Dept. of CSE, JNTUA, Anantapuramu, India, sivakumar.cse@jntua.ac.in

ABSTRACT

This paper discusses an efficient algorithm for topic level opinion mining and sentiment analysis of online text reviews by using unsupervised topic model, latent dirichlet allocation (LDA) for topic extraction and sentiment analysis of text reviews. The model accuracy is validated on twitter data by evaluating parameters perplexity and loglikelihood and compared with earlier models.

Key words: Opinion mining, Sentiment Analysis, Topic models, Latent Dirichlet Allocation

1. INTRODUCTION

As there is tremendous increase in posting online text reviews in social networking and obtaining feedback have gained more popularity in decision making to purchase a product or seeking feedback on any user interested topic. This user generated text is more in unstructured and un grammatical in nature. Opinion mining and sentiment analysis in topic level is used to identify the sentiment hidden in the text reviews by forming cluster of words among topics based on the probability distributions.

Twitter is a microblogging platform which has large content of user generated text and actively mined for patterns. Twitter have become more popular for exchanging opinions and ideas. which consists of various emotions of different intensities. For handling these discrete emotions in trending topics, there is need of efficient topic level sentiment classification techniques.

Many supervised learning algorithms are used in sentiment classification of text data and perform only binary classification either positive or negative sentiment. As these are supervised in nature and are trained using labelled data performs well for a trained domain, and the classifier has to retrained when the domain is shifted. Hence unsupervised techniques are in more demand to resolve the data annotations and domain shifting problems and achieve cross domain sentiment classification. This model gives flexibility to extract topics level features even for cross domain sentiment classification.

2. RELATED WORKS

Topic models like latent dirichlet allocation (LDA) are used to extract topic level features in an unsupervised manner. Neural and topic models were combined for the latent topic extraction and used in sentiment analysis related tasks.

Luo and Yi proposed an unsupervised topic specific emotion mining model by considering relationship between opinion words and aspect words [1]

Xu et al. proposed a topic recognition method for network sensitive information based on the sensitive word weighted-latent dirichlet allocation model and semantic similarity between the words [2] .

Yu et al. Proposed a bi-term topic model by using the semantically similar words based on word embedding model in mining of short texts which has data sparsity [3] .

Esmaeili et al. proposed a neural topic model with the help of inference neural network to learn aspect based structural representations of reviews [4] .

Matsumoto et al. proposed a slang classification method by constructing topic models to analyze topic change on twitter data by extracting slang features [5] .

Huang et al. proposed a multitask neural model for domain adaptation with the help of demographic factors through adversarial training and by representing the text with a topic model [6] .

Yu et al. proposed a hierarchical topic model for twitter data to mine the dimension hierarchy of tweet topics from large amount of unstructured text data [7] .

Viegas et al. proposed a novel topic models by generating cluwords by exploring nearest words in pre-trained word embedding model based on syntactic and semantic relationships between the word embedding space [8] .

Zheng et al. proposed a hierarchical three-layered interest network to mine user interests of microblogs and combined

timeliness and interactivity of users to evaluate dynamic interest hierarchical orientation [9] .

Li et al. proposed a novel model to mine the dynamics of research topics from large corpora by combining LDA and wordnet [10] .

Ali et al. proposed latent dirichlet allocation based on ontology and word embeddings for sentiment classification of transportation content from social networks [11] .

Wang et al. proposed a neural topic model which can learn end to end by latent topic model and predict key phrases on the user generated content on social media [12] .

Lin et al. proposed combined neural and topic models which can identify latent topic sparsity from online social media by providing sparse posterior distribution of topics [13] .

Recalde and Baeza Yates proposed a method to extract the multidimensional preference of user on twitter data set by using Expectation maximization algorithm [14] .

Rodrigues et al. proposed supervised two topic models for classification and regression, which can learn from multiple annotators and crowds with high dimensional text data sets [15] .

Peng and Vylomova proposed an algorithm for automatic classification of idiomatic and literal expressions using unsupervised clustering L.D.A by treating the idioms as sematic outliers [16] .

Movahedi et al. proposed a deep neural network based on attention mechanism to identify different aspect categories of a given review sentence based on different topics [17] .

Vamshi et al. proposed topic model-based opinion mining and sentiment analysis algorithm using LDA and SVM models [18] .

3. RESEARCH BACKGROUND

3.1 LDA topic model

Topic models are generative probabilistic models and unsupervised like clustering methods. Represent topics as multinomial distribution of words. Few of the topic models are latent dirichlet allocation (LDA), latent sematic analysis (LSA) and Non-negative matrix factorization (NMF). LDA works on probabilistic graphical modelling and NMF works on linear algebra. LSA uses singular singular value decomposition (SVD) method to reduce the document term matrix which is input to topic models.

Topic model frame work is shown in Figure 1. Topic model take text reviews as input and latent parameters including number of topics are set. Topics are detected automatically using probability distribution between words and topics. Distribution of topics are produced as output and topics are evaluated by using parameters like perplexity, log likelihood and topic coherence.

Topic vectors are derived from topic distributions which are the features to sentiment classification and prediction. Perplexity is the evaluation criteria of estimating the LDA model. Smaller the perplexity value, model will generate topics with high performance.

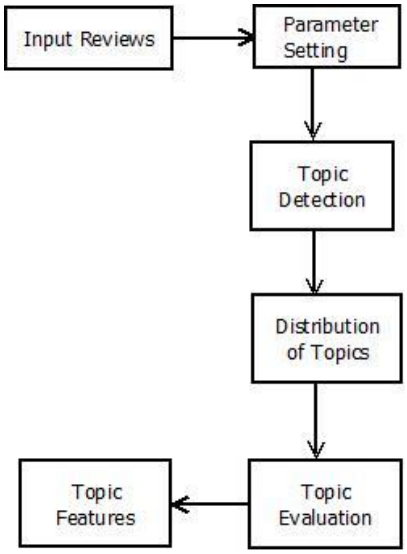

Figure 1: Framework for the Topic model

4. PROPOSED MODEL

The model architecture is shown in the figure 2, which uses LDA topic model for extracting the topic features from text reviews of twitter data sets in an unsupervised fashion. These topic features are fed to classifier which is used for sentiment classification purpose and implemented by using Python libraries sklearn, numpy, kerns and tensorflow.

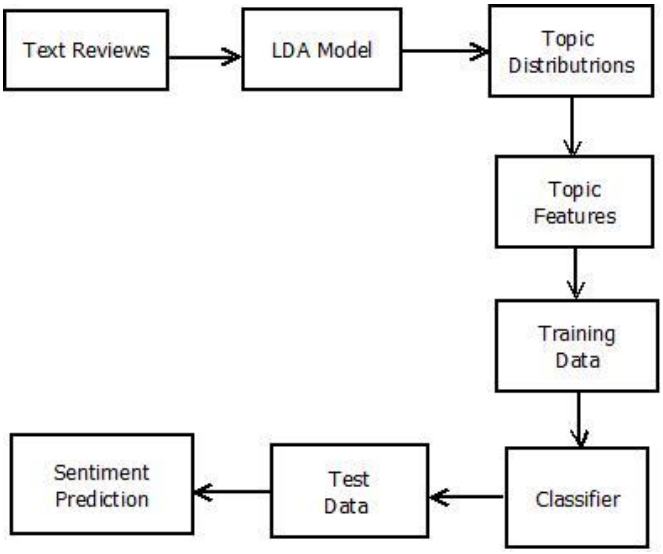

Figure 2: Proposed model

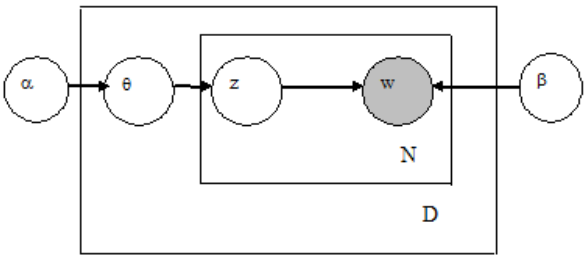

Figure 3: LDA model

4.1 Data-set used

Data set of twitter reviews of size 5000 text reviews are used as input for this model after doing necessary preprocessing.

4.2 Proposed Algorithm

Below is the proposed algorithm:
Input: Corpus of text reviews
Output: Distribution of topics

1) Build topic model on text reviews with number of topics and setting of latent parameters K, α, β as per figure 3
   // K - No. of topics, α- Dirichlet of document -topic and
   // β – Dirichlet of topic- word
2) Extract topic distributions for every review
3) Evaluate perplexity and log likelihood of topics
4) Use topic distributions as topic feature vectors to classification model
5) Apply sentiment classifier
6) Use LDA model to get topic distributions from new data set
7) Predict sentiment of text reviews by running classifier on new data set.

5. RESULTS AND DISCUSSION

LDA model is used to extract the topics and features respectively and are fed to sentiment classifier. Topic features generated are combined with SVM classifier and few variants of neural models are experimented.

5.1 Parameter setting

Hyper parameters set for the LDA model are shown in the table 1.

Table 1: Parameter settings

Parameter	Value
No. of Topics (K)	10
No. of words for each topic	15
Alpha (Dirichlet -Document Topic)	0.1
Beta (Dirichlet- Topic Word)	0.1
Iterations	100
Learning Decay	0.7
Batch size	128

5.2 Detection of topics

Topic are detcted by LDA model as shown in table 2.

Table 2: Topics generated in LDA model

Topic # no	Topics in LDA model
Topic #0	love happy lol great went mad days ill come let hope start amazing kpop friends
Topic #1	best going man miss yes yeah big ass thanks wanna make mo sure dany omg
Topic #2	like day really trash look yall feel mothers long episode good gameofthrones got men night
Topic #3	amp way right watch oh better god help th things sorry believe live cause family
Topic #4	dont know time good say mean got said doesnt getting wouldnt actually work baby sad
Topic #5	thank today fuck gonna week heart loved lost

	hit nice free sir online wow talking
Topic #6	shit hes fucking queen music saw says daenerys story sleep shot thinking home hand coming
Topic #7	need youre thats game literally thought couple thrones season didnt bad post read wrong character
Topic #8	im think want life people stop real beautiful cute apparently ok trying politics morning makes
Topic #9	got new wait girl little end year thing saying follow away probably wish learn left

5.3 Model evaluation

Data sparsity is 0.2765. Perplexity and log likelihood are used to evaluate performance of LDA model and are shown in table 3. Figure 4 show the variation of these parameters with respect to topics.

Table 3: Model evaluation

No. of Topics (K)	Perplexity	log likelihood
10	1056.0874	99164.4137
20	1208.9709	-101090.0458
30	1359.0465	-102756.6742
40	1405.8716	-103239.14366
50	1556.8376	-104691.9162
60	1621.2189	-105269.0674
70	1820.9578	-106923.8804
80	2010.6444	-108335.2572
90	2191.0260	-109558.9395

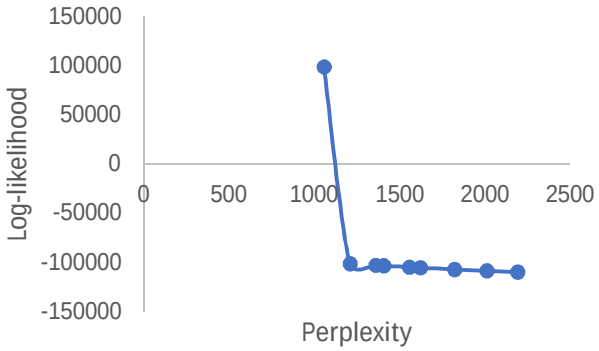

Figure 4: Variation of Perplexity and Log-likelihood

5.4 Classifier accuracy

LDA model is combined with SVM classifier and neural models, convolutinary neural network (CNN) and recurrent neural network (RNN) with variants RNN-LSTM and RNN-GRU. Training and testing accuracies of the respective models are shown in table 4. LDA combined with CNN has higher accuracy when compared to all other models.

Table 4: Accuracy of the models

Model	Training Accuracy	Testing Accuracy
LDA + SVM	0.6220	0.6220
LDA + CNN	0.6500	0.6364
LDA + RNN-LSTM	0.6415	0.6318
LDA + RNN-GRU	0.6415	0.6318

6. CONCLUSION

A topic level model for opinion mining and sentiment analysis for the twitter data set has been presented by integrating LDA model with SVM and neural models respectively. Some useful directions for future work include detecting abusive language present in textual reviews.

REFERENCES

[1] Luo, X. and Yi, Y. **Topic-Specific Emotion Mining Model for Online Comments**, Future *Internet*, Vol. 11(3), pp. 79, March 2019. https://doi.org/10.3390/fi11030079

[2] Xu, G., Wu, X., Yao, H., Li, F. and Yu, Z. **Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model**, *IEEE Access*, vol. 7, pp. 21527-21538, 2019.

[3] Yu, K., Zhang, Y. and Wang, X. **Topic Model over Short Texts Incorporating Word Embedding**, in *Proc. 2nd International Conf. of AEECS*, 2018, Vol. 155. https://doi.org/10.2991/aeecs-18.2018.34

[4] Esmaeili, B., Huang, H., Wallace, B. and van de Meent, J.W. **Structured Neural Topic Models for Reviews**, in *Proc. 22nd International Conference on Artificial Intelligence and Statistics*, April 2019, Vol. 89, pp. 3429-3439.

[5] Matsumoto, K., Ren, F., Matsuoka, M., Yoshida, M. and Kita, K. **Slang feature extraction by analysing topic change on social media**, *CAAI Transactions on Intelligence Technology*, vol. 4(1), pp. 64-71, 2019.

[6] Huang, X. and Paul, M. **Neural User Factor Adaptation for Text Classification: Learning to Generalize Across Author Demographics**, in *Proc. of the Eighth Joint Conference on Lexical and Computational Semantics*, June 2019, 136-146. https://doi.org/10.18653/v1/S19-1015

[7] Yu, D., Xu, D., Wang, D. and Ni, Z. **Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing**, *IEEE Access*, vol. 7, pp. 12373-12385, 2019.

[8] Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L. and Gonçalves, M.A.

CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling, in *Proc. of the Twelfth ACM International Conference on Web Search and Data Mining*, Feb 2019, 753-761.

[9] Zheng, W., Ge, B. and Wang, C. **Building a TIN-LDA Model for Mining Microblog Users' Interest**, *IEEE Access*, vol. 7, pp. 21795-21806.

[10] Li, C., Feng, S., Zeng, Q., Ni, W., Zhao, H. and Duan, H. **Mining dynamics of research topics based on the combined LDA and WordNet**, 2019, *IEEE Access*, vol. 7, pp. pp.6386-6399. https://doi.org/10.1109/ACCESS.2018.2887314

[11] Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H. and Kwak, K.S. **Transportation sentiment analysis using word embedding and ontology-based topic modeling**, 2019, *Knowledge-Based Systems*, vol. 174, pp. pp. 27-42.

[12] Wang, Y., Li, J., Chan, H.P., King, I., Lyu, M.R. and Shi, S. 2019, **Topic-Aware Neural Keyphrase Generation for Social Media Language**, arXiv preprint arXiv:1906.03889.

[13] Lin, T., Hu, Z. and Guo, X. **CluWords: Sparsemax and relaxed Wasserstein for topic sparsity**, in *Proc. of the Twelfth ACM International Conference on Web Search and Data Mining*, Feb 2019, pp. 141-149. https://doi.org/10.1145/3289600.3290957

[14] Recalde, L. and Baeza-Yates, R. 2018, **What kind of content are you prone to tweet? Multi-topic Preference Model for Tweeters**, arXiv preprint arXiv:1807.07162.

[15] Rodrigues, F., Lourenço, M., Ribeiro, B. and Pereira, F.C. **Learning supervised topic models for classification and regression from crowds**, *IEEE transactions on pattern analysis and machine intelligence*, vol. 39-12, pp. 2409-2422, 2018.

[16] Peng, J., Feldman, A. and Vylomova, E., 2018. **Classifying idiomatic and literal expressions using topic models and intensity of emotions**. arXiv preprint arXiv:1802.09961.

[17] Movahedi, S., Ghadery, E., Faili, H. and Shakery, A., 2019. **Aspect Category Detection via Topic-Attention Network**. arXiv preprint arXiv:1901.01183.

[18] Vamshi, K.B., Pandey, A.K. and Siva, K.A. **Topic Model Based Opinion Mining and Sentiment Analysis**, in *Proc. 2018 International Conference on Computer Communication and Informatics*, IEEE, Jan 2018, pp. 1-4. https://doi.org/10.1109/ICCCI.2018.8441220