JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Microblog Dimensionality Reduction With Semantic Analysis

M.S. Masram[1*], T. Diwan[2]

[1*]Dept. of CSE, Shri Ramdeobaba College Of Engineering and Management, Nagpur,India
[2]Dept. of CSE, Shri Ramdeobaba College Of Engineering and Management, Nagpur,India

[*]*Corresponding Author: masramms@rknec.edu*

Available online at: www.ijcseonline.org

Received: 01/Jan/2018, Revised: 10/Jan/2018, Accepted: 25/Jan/2018, Published: 31/Jan/2018

Abstract— Much attention in recent years has been attracted by the process exploring useful information from a large amount of textual data produced by microblogging services such as Twitter. A very important preprocessing step is to convert natural language texts of microblog text mining into proper numerical representations. The short-length characteristics of microblog texts result in using the term frequency vectors to represent microblog texts and it will cause "sparse data" problem. Finding proper representations for microblog texts is a challenging issue.In the previous paper, they applied deep networks so that they can map the high-dimensional representations to low-dimensional representations.The retweet and hashtags have been used as the semantic similarity. They used two types of approaches which includes modifying the training data and modifying the training objective. They have also shown that deep models perform better than traditional methods such as latent Dirichlet allocation topic model and latent semantic analysis.

***Keywords*—**Microbloging, Accessibility, Sentiment Classfication, Latent Semantic Analysis

I. Introduction

In the last few years, microblogging services such as Twitter has gained great popularity among the Internet users. The high volume of textual data produced by the microblogging services is very attractive to the researchers in the text mining field. Transforming natural language texts into numerical vectors is an important preprocessing step for many text mining tasks, such as cluster analysis and sentiment classification. The most widely adopted model for text representation is vector space model [1], where each document in a corpus is represented by a vector with each dimension corresponding to a separate term and the elements denoting the frequencies of the terms. An important issue that needs to be dealt with carefully when using term frequency vectors to represent texts is the "sparse data" problem.

Exploring potentially useful information from the huge amount of textual data produced by microblogging has gained much attention. Transforming natural language texts into numerical vectors. We apply deep networks to map the high-dimensional representations of microbiology. Creating proper low-dimensional feature space for microblog texts and also investigate how to utilize the expansion approach to learn better features. For the classification of documents, given the document training data, the most well-known approaches start by assessing the words' co-occurrence matrix versus documents. Dimensionality reduction means converting high dimensional representation to low dimensional representation. Also, it is the process of reducing random variables by obtaining variables that are known as principal variables. Latent Semantic Analysis is related to analyzing the relationship between term and document by obtaining the set of related concepts.Researchers have proposed many approaches to enhance the representations of short text, such as expanding the original short document by adding semantically related terms [2], [3], or mapping a high-dimensional term frequency vector to a low-dimensional feature vector via latent semantic analysis (LSA) [4]. There are a few approaches specifically proposed for dimensionality reduction of tweets. Instead, low dimensional representations of tweets are usually obtained as the by-products of topic modeling [5], [6], [7].

For the classification of documents when the document training data is given the most well-known approach is to start assessing the words that are co-occurrence matrix versus documents. It is understood that in any case, the count matrices have a very high tendency to be noisy and sparse especially when training data is moderately small[13]. Usually, the documents are presented in a high dimensional feature space, this is long way best for algorithms that are used for classification. A known methodology to reduce dimensionality is the feature selection. In this one has to choose a subset of words using some pre-defined paradigm and as features, it is able to use only the words that are chosen for classification. In the information retrieval, other similar strategies are Latent

Semantic Indexing (LSI) and Probabilistic LSI, LDA. There is a methodology so that any one can minimize the feature dimensionality by collecting 'similar' words into a much smaller number of groups and then use these groups as features[14].

Rest of the paper is organized as follows, Section I contains the introduction of microblog dimensionality reduction and some basic techniques used in the procedure. Section II contain the related work of how microblog dimensionality reduction is done in different papers using different techniques and also the description of techniques are also mentioned which were used in them. Section III contain some limitiation from the techniques present so that to implement technique which can overcome this limitations. Section IV contain the results and discussion followed by future scope in section V.

II. Related Work

2.1 Mapping Messages

Tweets and retweets of a user's followers appear alongside the user's own tweets in reverse chronological order. People often have only enough patience to skim through the first 20 - 50 messages. When the messages become overwhelming, it is impractical for a user to quickly gauge the main subjects from their followers' posts. To make a large collection of messages which can be accessed by the users the web systems should provide not only accurate clusters for subtopics in messages but also meaningful labels for each cluster[2]. Enhancing the accessibility of microblogging messages entails two tasks: (a) cluster microblogging messages into manageable categories, and (b) assign readable and meaningful labels for each cluster of messages. Unlike standard text with many sentences or paragraphs, microblogging messages are noisy and short[4]. In addition, microbloggers, when composing a message, may use or coin new abbreviations or acronyms that are uncommon in conventional text documents.[5] Furthermore, these short messages do not provide enough information so that they can capture the semantic meanings.

2.2 Traditional text mining

Traditional or General text mining methods when applied to messages from microblog lead to unsatisfactory results. In this paper they presented a framework to enhance the accessibility of microblogging message. The proposed framework by them helps in improving the message representation by mapping the messages from an feature space which is unstructured to a semantically meaningful knowledge space[8]. First, in order to reduce the noise yet keep the key information as expressed in each message, we propose to use natural language processing (NLP) techniques to analyze the message and extract informative words and phrases.To overcome the probem that is happened because of sparsity of messages they map the terms to concepts that are structured and derived from external knowledge bases that are semantically rich[9]. Because of conducting feature selection to refine the feature space they were able to cluster all message accurately and generate humanly understandable labels efficiently from the related concept.

2.3 Data Sparseness

Most of the recent related work concentrates on the elimination of the problem of data sparseness. One solution to solve this is to increase the short text with additional information to make it like a large document of text. Then the algorithms for clustering or classification can be applied easily to it. As discussed in [11], their main focus is on combining short text messages with web search engines such as Bing, Google to extract more information about the short text. In [12] they said that if two hashtags co-occur in a tweet, then they are similar. With co-occurrence frequency as a distance measure, they created a clustered graph. The other works on clustering text -related entities typically focus on a bag –of - words(BOW). BOW model takes all the words of entity followed by reduction of dimensionality. It makes clustering computationally feasible as in [15,16].

One of the biggest data sources is Wikipedia. Short text messages can be strengthened with additional semantic knowledge by combining knowledge accessible within the Wikipedia For the classification of documents, given the document training data, the most well-known approaches start by assessing the words co-occurrence matrix versus documents. It is understood that in any case such count matrices have the tendency to be excessively noisy and sparse particularly when training data is moderately small. Usually the documents are depicted in a feature space that is high dimensional which is the long way from optimal for algorithms of classification.

2.4 Feature Dimensionality

A standard and basic methodology to reduce the dimensionality of features is feature selection. In this methodology one can choose a subset of words using some pre-defined paradigm as features, it uses only the chosen words for classification. In the task of information retrieval, other similar strategies for dimensionality reduction are Latent Semantic Indexing (LSI) and Probabilistic LSI. An alternative methodology is to minimize the feature dimensionality by gathering 'similar' words into a much smaller number of word groups and utilize these groups as features.

2.5 Probabilistic Topic Models

To capture semantic similarity among words from the message they derived a model of documents which learns word representations which is probabilistic. This component does not require labeled data and shares its foundation with probabilistic topic models such as LDA. The sentiment component of their model use sentiment annotations such that the words expressing a similar sentiment to have similar representations. It can efficiently learn parameters for the joint objective function using alternating maximization. The model which is presented does not capture information from the sentiments from the message.By applying the algorithm to documents will produce representations, where which are occurring together in documents, will be given similar representations. This approach which is unsupervised has no way of capturing which words are predictive of sentiment.[17][18] Much previous work in natural language processing achieves better representations by learning from multiple tasks. Following this theme, they introduced a second task to use labeled documents so that improve their model's word representation.

2.6 Hierarchical Dirichlet process (HDP)

Each and every group of data being modeled with a mixture with the number of components which are open-ended and they are inferred automatically by the model. Then the components can be shared across various groups and allow the dependencies across groups to be modeled effectively as well as conferring generalization to new groups. Such clustering problems occur in practice, e.g. in the problem of topic discovery in document corpora. They reported experimental results on three text corpora showing the effective and superior performance of the HDP over previous models. They considered the application which has hierarchical Bayesian ideas to a problem in "multi-task learning" where the "tasks" are clustering problems, and their goal is to share clusters among multiple clustering problems that are related. They were motivated by the task of discovering topics in document corpora [19]. A topic (i.e., a cluster) is a distribution across words while documents are viewed as distributions across topics.They wanted to discover various topics which are common across various multiple documents in the same corpus and also as across multiple corpora. Their work is based on a tool which is from nonparametric Bayesian analysis also called (DP) mixture model [20].

III. Limitation

To address the sentiment classification problem with a small number of labeled reviews. We studied the problem of finding bursty topics from the text streams on microblogs. Another limitation of the current method is that the number of topics is predetermined that is separate topics for positive words and a separate one for negative words links. And more types of meta-information contained in tweets, such as emoticons and the embedded hyperlinks, will be explored.

IV. Results and Conclusion

In this, they investigated how to apply deep networks to perform dimensionality reduction on microblog texts. A priori knowledge about semantic similarity which is derived from retweet relationships and hashtags was explored to train the deep networks. Two types of approaches are there namely modifying the training data and modifying the objective of fine-tuning, were proposed to utilize such prior knowledge. Experiment results validated that deep models can learn better representations than LSA and LDA, and the use of microblog-specific information can further improve the performance of deep models. The evaluation results also demonstrate that the proposed modifications of training data can help to learn better representations. Later, they give some detailed explanations to the evaluation results.

In this, they have proposed Real-Time Big data analytical Architecture for emotion extraction from social networks tweets. The proposed architecture by them was capable of efficiently processing and analyzing real-time and offline data and classification of tweets in positive and negative .They used the big data technologies to classify text in a distributed manner and can decrease the execution time. LMClassifier, OpenNLPClassifer andNaiveBayesclassifer these classifiers were used with MapReduce approach for reducing execution time for a large amount of text data. If they used NaiveBayesclassifer then they achieve the highest accuracy with less execution time because in normal Naivebayesapprocah it took more time to execute than OpenNLPClassifer andLMClassifer.They used MapReduce with all three classifiers to reduce the execution time. Text data can arrive with high velocity, veracity, Volume, variety so in these conditions, they used MapReduce approach because it handles all these scenarios in execution. Execution time is decreasing when size is increasing in MapReduce approach. In conventional use of classifiers execution time increases with an increment of the size of dataset. So applying classification techniques in distributed environment provides better execution time.

In the first Map-Reduce pass, the mapper takes the labeled tweets from the training data and outputs category and word as a key-value pair. The Reducer then sums up all instances of the words for each category and outputs category and word-count pair as a key value. The Map-Reduce then deals with the making of model for the classifier. The next Map-Reduce pass does the classification by calculating the conditional probability of each word (i.e. feature) and outputs category and the conditional probability of each word as a key-value pair. Then final reducer calculates the final probability of each category to which the tweet may belong to and outputs the predicted category and its probability value as key-value pair The final step after preprocessing of tweets is that they label the tweets based on categories namely politics, sports, and technology.

In this,they proposed a semi-supervised learning called Active Deep Networks algorithm (ADN) to solve the sentiment classification problem with a very few number of labeled data. ADN can choose the proper training data to be labeled manually and also it fully exploits the embedding information from a large amount of unlabeled data to improve the robustness of the classifier. They proposed a new architecture to guide the output vector of samples belong to different regions of new Euclidean space and use an exponential loss function to maximize the separability of labeled data in global refinement for better discriminability. The ADN can make a right decision about which training data should be labeled based on the data which is not labeled and labeled data. By using the both supervised and unsupervised learning , ADN is able to choose the training data to be labeled and train the architecture at the same time.

The deep architecture is then re-trained using the labeled data and all the data which is not labeled. They also conduct experiments to verify that ADN method is effective with the various number of labeled data, and also demonstrated that ADN is able to reach very competitive performance for classification just by using few labeled data. These shows that the ADN method which needs only a fewer manual labeled reviews so that it can reach a higher accuracy also can be used to train a high-performance sentiment classification system.

V. FUTURE SCOPE

This is one of the directions studies in the future. Another limitation of the current method is that the number of topics is predetermined. We planned to look into methods that allow the appearance and disappearance of topics along with the timeline. There are also other properties, such as mass data, real-time updates, and so on which makes the detection method can not be applied to microblog topic detection. But, on the other hand, the microblog structure gives us a new idea to improve the performance of the microblog topic detection. The best of these structural properties, and propose an effective microblog topic detection method which aims at the above three properties. In future studies, to optimize the learning of representations towards specific microblog mining tasks, such as sentiment classification. And more types of meta-information contained in tweets, such as emoticons and the embedded hyperlinks, will be explored.

VI. REFERENCES

[1] Lei Xu, Chunxiao Jiang,“Microblog Dimensionality Reduction—A Deep Learning Approach,” *Ieee Transactions On Knowledge And Data Engineering, Vol. 28, No. 7, July 2016.*

[2] Zhi-Qiang Xian , “Sentiment Analysis of Chinese Micro-blog Using Vector Space Model,” *APSIPA,2014.*

[3] Amit mittal , “Social Networking text Classification in Big Data Environment,” *IJlEET, 2016*

[4] X. Yan and H. Zhao, “Chinese microblog topic detection based on the latent semantic analysis and structural property*,” J. Netw., vol. 8, pp. 917–9233, no. 4, 2013*.

[5] D. Ramage, S. T. Dumais, and D. J. Liebling, “Characterizing microblogs with topic models*,” in Proc. 4th Int. Conf. Weblogs Social Media, pp. 130–137, 2010.*

[6] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, “Transferring topical knowledge from auxiliary long texts for short text clustering,” *in Proc. 20th ACM Int. Conf. Inf. Knowl. Manag., pp. 775–784, 2011*.

[7] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, “Finding bursty topics from microblogs,” *in Proc. 50th Annu. Meet. Assoc. Comput. Linguistics: Long Papers-Vol. 1. , pp. 536–544, 2012.*

[8] M. A. Ranzato and M. Szummer, “Semi-supervised learning of compact document representations with deep networks,” *in Proc. 25th Int. Conf. Mach. Learning, pp. 792–799, 2008.*

[9] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Sci., vol. 313, no. 5786, pp. 504–507, Jul. 2006.*

[10] R. Salakhutdinov and G. Hinton, “Semantic hashing,” *Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, Jul. 2009.*

[11] M. A. Ranzato and M. Szummer, “Semi-supervised learning of compact document representations with deep networks,” *in Proc. 25th Int. Conf. Mach. Learning, pp. 792–799, 2008.*

[12] S. Zhou, Q. Chen, and X. Wang, “Active deep learning method for semi-supervised sentiment classification,” *Neurocomputing, vol. 120, pp. 536–546, 2013.*

[13] M. R. Min, L. Maaten, Z. Yuan, A. J. Bonner, and Z. Zhang, “Deep supervised t-distributed embedding*,” in Proc. 27th Int. Conf. Mach. Learn. , pp. 791–798, 2010.*

[14] D. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learning Res., vol. 3, pp. 993–1022, 2001.*

[15] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis*,” Discourse Processes, vol. 25, pp. 259–284, 1998.*

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR, vol. abs/1301.3781, 2013.*

[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” *in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Human Language Technol.-Volume 1., pp. 142–150,2011.*

[18] J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu, “Enriching short text representation in microblog for clustering,” *Frontiers Comput. Sci., vol. 6, no. 1, pp. 88–101, 2012.*

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Sharing clusters among related groups: Hierarchical dirichlet processes*,” in Proc. Int. Conf. Neural Information Processing Syst, pp. 1385– 1392, 2004.*

[20] C. E. Grant, C. P. George, C. Jenneisch, and J. N. Wilson, "Online topic modeling for real-time twitter search," *in Proc. Text Retrieval Conf. , pp. 1–9, 2011.*

[21] X. Wang, F. Zhu, J. Jiang, and S. Li, "Real time event detection in twitter," *in Proc. 14th Int. Conf. Web-Age Inf. Manag., pp. 502–513, 2013.*

[22] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Process. Mag., vol. 28, no. 1, pp. 145–154, Jan. 2011.*

[23] Y. Bengio, A. C. Courville, and P. Vincent, "*Unsupervised feature learning and deep learning: A review and new perspectives,"CoRR, vol. abs/1206.5538, 2012.*

[24] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *in Proc. 25th Int. Conf. Mach. Learning, pp. 160–167,2008.*

[25] J. P. Turian, L.-A. Ratinov, and Y. Bengio, "Word representations:A simple and general method for semi-supervised learning," *in Proc. 48th Annu. Meet. Assoc. Comput. Linguistics, pp. 384–394, 2010*.

Authors Profile

MissMegha Masram has done Bachelor of engineering in computer Science and engineering from Rashtrasant Tukdoji Maharaj Nagpur University,India in 2016 and currently pursuing P.G in computer science and engineering from Ramdeobaba college of engineering and management,Nagpur,India.And research work focus on Data Mining,Information Retrieval and machine learning.

Dr.Tausif Diwan received M.Tech. And Ph.D. in Computer Science and Engineering from VNIT college, Nagpur, India in 2011 and 2017 respectively. Since June 2012, he has been with the Department of Computer Science and Engineering, RCOEM Nagpur, India as an Assistant Professor. His research area includes parallel computing and algorithms on multicore and manycore architectures.