Sentiment Analysis with Emotion Detection using Prediction Algorithm

Jayta Kelzare[1] Alsaba Qureshi[2] Neha Dongre[3] Khushboo Ansari[4] Ritesh Shrivastav[5]

Abstract— The volume of microblogging messages is increasing exponentially with the popularity of microblogging services. With a large number of messages appearing in user interfaces, it hinders user accessibility to useful information buried in disorganized, incomplete, and unstructured text messages. In order to enhance user accessibility, we propose to aggregate related microblogging messages into clusters and automatically assign them semantically meaningful labels. However, a distinctive feature of microblogging messages is that they are much shorter than conventional text documents. These messages provide inadequate term co-occurrence information for capturing semantic associations. To address this problem, we propose a novel framework for organizing un structured microblogging messages by transforming them to a semantically structured representation. The proposed first captures informative tree fragments by analyzing a parse tree of the message, and then exploits external knowledge bases to enhance their semantic information. Twitter dataset shows that our framework significantly outperforms existing state-of-the-art methods.

Key words: Microblogging, Accessibility, Clustering, Labeling

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 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 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 microblog. 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. 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 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].

II. EXISTING PROBLEM

An important issue that needs to be dealt with carefully when using term frequency vectors to represent texts is the "sparse data" problem. Apply deep networks to map the high-dimensional representations of microblog. Creating proper low-dimensional feature space for microblog texts and also investigate how to utilize the expansion approach to learn better features. Exploring potentially useful information from huge amount of textual data produced by micro blogging services has attracted much attention in recent years. An important preprocessing step of micro blog text mining is to convert natural language texts into proper numerical representations. Due to the short-length characteristics of micro blog texts, using term frequency vectors to represent micro blog texts will cause "sparse data" problem. Finding proper representations of micro blog texts is a challenging issue. Exploring potentially useful information from huge amount of textual data produced by microblogging services has attracted much attention in recent years. An important preprocessing step of microblog text mining is to convert natural language texts into proper numerical representations. Due to the short-length characteristics of microblog texts, using term frequency vectors to represent microblog texts will cause "sparse data" problem. The result of dimensionality reduction is constant.

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. Separate topics for positive words and a separate one for negative words links. Are spammers more likely to link to each other than regular people. Is content similarity more important than linguistic similarity when predicting links. 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.

A. Proposed work to over come

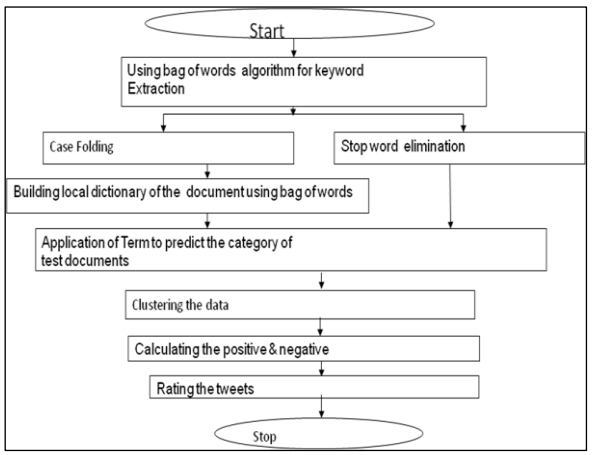

Fig. 1: Microblog dimensionality reduction along with rating

- Text preprocessing
- Semantic text analysis
- Features Generation
- Bag of words
- Features Selection
- Simple counting 1
- Text/Data Mining
- Classification- Supervised learning
- Analyzing results.

Fig. 2: Data Analysis

Fig 3: Analysis Result

B. Deep network-based dimensionality reduction

DR techniques are proposed as a data preprocessing step. This process identifies a suitable low dimensional representation of previous data. Dimensionality Reduction (DR) in the dataset improves the computational Efficiency and accuracy in the analysis of data [8]. Dimensionality reduction is the process of reducing the number of random variables under some consideration. A word matrix(documents *terms) is given as input to reduction techniques like Principal Component Analysis (PCA).

1) Of offline tweets:

To perform this task database containing tweets should be needed. So we have used a MySQL database containing 3000 tweets for offline or document classification. Whenever the tweets of particular link or profile are retrieved, they all are imported to the database in our framework.

2) Data preprocessing:

It is a very important stage in our framework where it involves following steps.

1) Tokenization: It is a process of splitting up of stream of text into tokens (words, phrases or other meaningful elements). Here each tweet is broken up into words i.e. tokens and passed as input to next stage.
2) Stop word removal: In text mining, most of the frequently used words in English are unwanted words for mining. So such words are called stop words. And the division of the natural language are the stop words. Stop words include a, an, the, and, are, as, at, be, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with etc. In this step the stop words should be eliminated to make the tweets look less heavier and more important for analytics. Dimensionality of the term space will be reduced by removing stop words from the tweets. Here we have used classic method for removing stop words.
3) Stemming: In this step root or stem of the words are identified. For example, the words define, defined, defining and definition all can be stemmed to the word "define". Here we have used stemming algorithm to reduce number of words, eliminate various suffixes, to have accurately matching stems, and to save time and memory space.

3) Extraction of Features:

Extraction of features is also an important step in text classification and clustering. It is process of selecting the subset of the words form the training dataset. This subset acts as features in the process of text classification. In this work method used to represent text is Bag-Of-Words model i.e. BOW model. BOW model is the most widely recognized strategy to depict text. The text is divided into words in this technique and every word depicts a feature. For example, for sports dataset, Sachin Tendulkar, cricket, football, India etc. Act as features. This step is useful for the classification and clustering in two ways: (1) by reducing the size of the vocabulary which can be trained used for applying the classier algorithm efficiently, and (2) by eliminating the noise features which in turn makes the clustering and feature extraction to increase the accuracy of classification.

Many forms of text databases contain a large amount of side-information or meta-information, which may be used in order to improve the clustering process. In order to design the clustering method, we combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. This general approach is used in order to design both clustering and classification algorithms.

IV. CONCLUSION

To map the high-dimensional representations of microblog texts to low-dimensional representations. To improve the result of dimensionality reduction, we take advantage of the semantic similarity derived from two types of microblog-specific information, namely the retweet relationship and hashtags. Two types of approaches, including modifying training data and modifying the training objective of deep networks, are proposed to make use of microblog-specific information. We investigated how to optimize the learning of representations towards specific microblog mining tasks.

REFERENCES

[1] Microblog Dimensionality Reduction—A Deep Learning Approach. Lei Xu, Chunxiao Jiang, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 7, JULY 2016

[2] Sentiment Analysis of Chinese Micro-blog Using Vector Space Model. Zhi-Qiang Xian 23-8 © 2014 .
[3] Social Networking text Classification in Big Data Environment Amit mittal . 2016 IJlEET.
[4] X. Yan and H. Zhao, "Chinese microblog topic detection based on the latent semantic analysis and structural property," J. Netw., vol. 8, no. 4, 2013, pp. 917–9233.
[5] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in Proc. 4th Int. Conf. Weblogs Social Media, 2010, pp. 130–137.
[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., 2011, pp. 775–784.
[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., 2012, pp. 536–544.
[8] M. A. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in Proc. 25th Int. Conf. Mach. Learning, 2008, pp. 792–799.
[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, 2008, pp. 792–799.
[12] S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semi-supervised sentiment classification," Neurocomputing, vol. 120, pp. 536–546, 2013.

