News-text Sentiment Classification Research Based on JST Model

Yan ZHAN, Hao CHEN, Hongyan MA, Guochun ZHANG
College of Mathematics and Information Science, Hebei University, Baoding, CHINA

Abstract: JST model is a hybrid model with both topic and sentiment, which adds emotional elements on the basis of the LDA model, and it can make text sentiment analysis and subject extraction in textual level. JST's sentiment classification accuracy of the model is a little low, therefore we put forward the appraise dictionary for JST model of prior knowledge, which is to change the method assigning emotional tags randomly into the method assigning emotional tags the dictionary after comparing.

Keywords: Sentiment Analysis; JST Model; Appraise Dictionary

1. Introduction

With the rapid development of network technology, the Internet has become an indispensable part of people's lives. Internet makes more and more people willing to express their views through the internet, and get the information they need by its features as follows: freedom; virtually and concealment, etc. These news websites release a large number of various categories of news every day, including both domestic news and international news, among which exists not only positive news but also negative news. News media uses network news as a carrier, supervising the community and government, spreading news events that have happened every day. Network's public opinions have already been able to exert influence on the relevant decision-making departments, which make analysis of the information about network's public opinions, thus making it an important basis on which to guide and monitor network's public opinions.

This article selects the network news as material, and makes analysis of emotional tendency to the news text. Traditional text representation models ignore the semantic level of the text, thus not able to make good emotional classifications for news. This article uses the Joint sentiment/topic model(JST) model developed based on topic LDA model, which not only can solve the semantic problem, express the text increasing from words to topic, but also solve high dimension problems at the same time. News is a description of the facts, usually exhibiting the feature of "low degree of subjective", most of which are long texts, and screen emotional sentence of news text topic. Extracting sentences that expressing topics and emotions for analysis is a research focus of this article .

Text sentiment analysis, also known as opinion mining, in simple terms, is a process of analyzing, processing, summarizing and reasoning the subjective text with emotional colors [1]. Currently, most of materials used in the sentiment analysis derive from the blog, professional review sites, news sites, and some e-commerce sites, among which user reviews of various products are the first choice of many researchers [2]. The emotional tendency of commenting text usually seems obvious, and most are short text, in which the contents of most of the comments are related to the topic. With respect to the comment text, news text is a description of an event with relatively weak subjectivity in which most are long text. Since there are obvious difference between news text and comment, it applies to the method of comment datasets instead of news datasets. The main task of sentiment analysis is to deal with the text related to the topic which active users on the network released, to identify the subjective sentence contained in the text, and to judge their emotional tendency [2]. This extends the focus of the sentiment analysis study: subjective and objective classification of text and emotional tendency of subjective text classification.

Subjective and objective classification task of the text is to divide the text into the subjective and objective text types or extracts subjective text from the text, which is to identify subjective text. Subjective text is emotional, whereas objective statement of facts has no emotion. As for subjective and objective classification of the text, researchers usually choose to study from the following three levels: words, sentences and chapters.

Wiebe et al. are early researchers of the task [1-9]. Wiebe et al. [3,4] made statistics of each document on vocabulary with subjective tendencies, constructed k-nearest neighbor classifier which achieved good results through the study on sentiment classification problem by chapter level. Wiebe et al. [5] presented a method that studying potential subjective collocation from the corpus. Experiment has verified that by using existing subjective clues (adjective, verb) and the mining of subjective match of the text may be used in conjunction to obtain a better subjective and objective classification. Wiebe et al.[6] added the concept of "subjective expression density" to the study to help determine whether the match mode is a subjective expression or not. Wiebe et al[7] explored the expression syntax by a large number of features includ-

ing syntax analysis ,and they made identification and intensity classification on subjective sentences according to deeply nested clues. Yu Hong et al[8] proposed a subjective and objective text classification which completed the chapter level by naive Bayes classifier characterized by a term. Each features' combination of the experiment was selected to construct the classifier, and the experiment compared classifier's performance whose features including word, bigrams, trigrams, part of speech, polarity, and the classification accuracy achieved up to 92%. Pang Bo et al.[9] studied the extraction of subjective sentences by using contextual information based on graph's minimum cut classification algorithm. The relationships between sentences are not independent, however, there exists emotional connection between them. Therefore, inserting this important feature into the classification between sentences, and using minimum graph cut method to find out the relationship between the context can improve classification accuracy.

Because of different expression habits, language structures and cultures between China and America, it makes subjective and objective classification in the context of Chinese more challenging.

FangYuan et al[10] constructed classifier which used speech and emotional dictionary as characteristics, thus completing the subjective sentence recognition of the text. At first, construct a weak classifier by a single feature, then, by using AdaBoost method and Bootstrapping Iterative process reconstruct a much stronger multi-feature classifier. The results showed that constructing a stronger classifier can improve accuracy rate up to 78.82%.

As for the classification of the emotional tendency of subjective text, at present, there are methods mainly based on the rules, semantics, and statistics. Sentiment analysis method based on rules, costs high and has a big workload. Meanwhile, due to the appearance of new words and the change of expressions, the extensibility of this method is poor. Therefore, the later two sentiment analysis methods are more adopted by scholars.

Lin ChengHua et al[11] proposed JST hybrid model based on the LDA topic model. What make JST model and LDA model different can be embodied as follows: LDA model only uses the topic tag for each word, however, JST model uses topic and emotion as two tags for each word; LDA gets distribution of subject text finally but JST eventually gets distribution of emotional text. Similarity is that they both can get all the vocabulary potentially under the topic.

2. JST Model

Joint sentiment/topic model(JST) is a probabilistic modeling framework based on the LDA model which is proposed by Lin ChengHua et al. [11] in the meeting CIKM in 2009. JST model is a four-layer plate model: document is associated with sentiment tags, topic is associated with sentiment tags, words are associated with sentiment tags and topic tags. JST model has a good effect on a review of data sets, but there is no one applying it to the analysis in the emotional news text. This article makes sentiment analysis for Chinese news text by JST model. Figure 1[12] is a model diagram of JST.

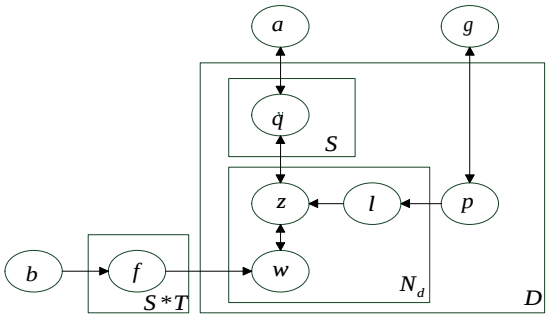

Figure 1. JST model diagram

The basic idea of JST Model: Suppose there are documents whose total number is D, denoted as $C=\{d_1,\ldots,d_D\}$; each document is represented by a sequence of words whose total number is N_d, denoted as $d=(w_1,w_2,\ldots,w_{N_d})$; after removing the repeated words in the corpus, the remaining words are moved into the dictionary, the dictionary size was denoted as V ,as a result each word in the document corresponds to an index entry; suppose the number of different topics is T, the number of different emotions is S. The process that the generation of words in the document is summarized as follows: First, select a sentiment tag l from the document distribution with emotion; and then, select a topic randomly in the topic distribution $q_{l,d}$ with sentiment tag l; finally, generate a word from word distribution f with topic and sentiment tag .

Corresponding to the hierarchical bayesian model shown in figure 1, words is generated as the following steps [12] :

1). For each document in the corpus, extract polynomial distribution p_d from dirichlet distribution with parameter g, namely sampling $p_d \sim Dir(g)$;

2). For each sentiment tag in document d, extract polynomial distribution $q_{d,l}$ from dirichlet distribution with parameter a ,namely sampling $q_{d,l} \sim Dir(a)$;

3). For each word in the document w_i:

(a) choose a sentiment tag l_i from p_d , namely sampling $l_i \sim p_d$;

(b) according to the sentiment tag l_i, select a topic tag randomly from q_d, l_i, namely sampling $z_i \sim q_{d,l_i}$;

(c) select one word from distribution with topic z_i, sentiment tag l_i.

3. JST Model Building

(1) Web crawler

The basis of the JST model implementation is to collect pages of data, the amount of which are typically huge, which makes it hardly can be accomplished by artificial means. Therefore, the web crawler technology comes into being. Web crawler is an automated crawling web content program. At first, it finds an initial links, under which it stores the web content, adds other pages' link in the web page to the link queue at the same time. Then it removes the following links to be accessed from the link queue, and conducts the first step of the operation. It will repeat the process until it satisfies the initial predetermined stop condition. Web crawler involves in web page filtering analysis, the efficiency of the crawler and other issues. The web page filtering analysis refers to the need to design algorithms, selecting topic related links to add to the link queue, and remove the duplicate in the link queue, and so on. The efficiency of the crawler is an indicator of measuring the crawler algorithm. Establishing an optimized index can improve the speed of the crawler, and save the time of users. At present, the crawler technology has already been very mature, so in this article we use the existing open source web crawler tools - Heritrix.

(2) HTML parsing

Crawling down data directly from the pages is often HTML format with html tags. Therefore, HTML needs to be parsed. HTML is hypertext markup language, composed by the "head" section, and the "body" section structure, in which the "head" section includes the page title, language of the web pages and other information; "body" section is equivalent to the specific content of the page in the body of the article. Specific content contains various types of tags, which use "<" and ">" in quotes.

This article uses the Java language to parse HTML, thus extracting data which meets the experimental needs in this article. The main method is to use regular expressions and Java's basic processing capabilities of the string.

(3) Word segmentation and part-of-speech tagging

English text, because of its unique writing specifications: each word is separated by space, is very clear that there are boundaries between words, therefore it does not require segmentation. However, Chinese writing is a string of characters sequence which requires segmentation. Chinese word segmentation refers to cutting the sequence of characters into separate words, which provides the basis for further analysis. Commonly used segmentation tools are ICTCLAS, SCWS, HTTPCWS, and it's like, in which ICTCLAS is an open source word segmentation platform developed by Chinese Academy of Sciences, and won first place in a number of domestic and international evaluation. That platform applies leading natural language understanding, web search and text mining technology, provides a set of basic tools for technical secondary development. Developing platform consists of several parts: the word segmentation and part of speech, neologisms discovery, and the like, in which each part can be independently be called by Java, C, C #, and other programming languages, and is compatible with a variety of different operating systems. Through comprehensive comparison of multiple open source experiment segmentation tool, we decided to adopt ICTCLAS segmentation system in this article. Directly batching news text word by ICTCLAS makes prepared for subsequent analysis.

(4) Delete stop words

After performing word segmentation, the presence of some words only play a role in the structure, and they may not reflect any actual significance, such as prepositions, adverbs and the like. There is a high frequency of some words appearing in the entire corpus whose frequency of each document appears much the same, take Sohu news as an example: the text of each news will be a "Sohu" which needs to be added to the disable dictionary. Further analysis and research on such words is pointless, so we need to remove those words completely. The two kind of words mentioned above are referred to delete stop words. The presence of stop words will increase vector's dimension, and deleting stop words plays a role in crude dimensionality reduction.

In this article we downloaded a Chinese commonly used stop list from the Internet, in which contains 1208 stop words. These words cover not only the auxiliary verb, but also some non-significant punctuation. Based on the actual corpus this article added some stop words: words without research value but appear in every news such as "Sohu", "News" and so on. After being adjusted the stoplist can meet the needs of this article experiments on the whole.

(5) JST model algorithm

After preprocessing word segmentation and stop words of the text corpus, we achieved JST models by the java language. Algorithm can be embodied as follows:

Input: pretreated news text

Output: Matrix f (words × topics × emotions, $V \times T \times S$), matrix q (emotions × topics × documents, $T \times S \times D$) and matrix p (emotions × documents, $S \times D$).

Basic steps:

a. Read words of the text, and make serialized representation of them, then load these words into dictionary V.

b. Initialize the matrix f (words ×topics × emotions, $V \times T \times S$), matrix q (emotions × topics × documents, $T \times S \times D$) and matrix p (emotions × documents, $S \times D$);

c. Perform Gibbs sampling iterations form $m = 1$ to M , an iteration process is as follows:

(1) reads a word from document, randomly assigned to which topic tag and emotional tag;

(2) According to the formula (1)

$$P(z_t = j, l_t = k \mid w, z_{-t}, l_{-t}, a, b, g) \propto \frac{\{N_{w,j,k}\}_{-t} + b}{\{N_{j,k}\}_{-t} + Vb} \cdot \frac{\{N_{j,k,d}\}_{-t} + a}{\{N_{k,d}\}_{-t} + a} \cdot \frac{\{N_{k,d}\}_{-t} + g}{\{N_{d}\}_{-t} + g} \quad (1)$$

Computing the probability of words with emotional tag k for and topic tag j ;

(3) According to the probability estimated in (2), reselect a topic tag j for words according to the Markov chain;

(4) In topic tag j, reselect an emotional tag k;

(5) According to the new sampling results, update the matrix f, the matrix q and the matrix p;

(6) Return and perform (1) until all the words processed. Thus the iteration is completed one time.

In this article we start recording results from the 100th iteration, record 100 times per iteration, and the total iteration reaches 2,000 times. In the 1000th iteration, the experimental results tend to be stable. Therefore we selected 10 groups' average iteration after 1000th iteration as the final classification.

4. JST Model Combined with Prior Knowledge

Using JST model directly makes classification accuracy of news text is not very high. In order to improve the classification performance of JST model, we introduced praise or blame righteousness dictionary as the prior knowledge of JST model. This article selects HowNet emotional dictionary, NTUSD simplified Chinese emotional polarity dictionary in Taiwan university, praise or blame righteousness dictionary (Tsinghua university), respectively to conduct experiments. Eventually we selected HowNet emotional word sets which has the best effect as the experiment's praise or blame righteousness dictionary, including Chinese positive emotional words, 836, and negative emotional words in Chinese, 1254.

After using praise or blame righteousness dictionary as prior knowledge, the algorithm in steps (1) of step c becomes:

Read a word from the document, and assign topic tags to the word randomly, then contrast the word with praise or blame righteousness dictionary. If the word is the same as the one word of praise or blame righteousness dictionary, then assign the word to the corresponding emotional tag; otherwise, it is still adopt the method of assigning randomly emotional tags to the word.

5. Conclusion

Through the sentiment analysis for news text based on statistical topic sentiment model -JST model, we verify the feasibility of JST model in the news text. The JST model is unsupervised and needn't to mark the training sample. Therefore it does not exist the problem of the field transfer. To join prior knowledge - appraise dictionary, JST model can further improve the classification accuracy.

6. Acknowledgments

This work is supported by the Soft Science Research Program of Hebei Province (Grant no. 14450318D) and NSF of Hebei Province (No. F2013201064, F2014201100).

References

[1] Wiebe J. Tracking point of view in narrative.Computational Linguistics, 20(2):223-287 (1994).

[2] Wiebe J,Bruce R,O' Hara T. Development and use of a gold standard dataset for subjectivity classifications//Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).Seattle,USA:246-253 (1999).

[3] Wiebe J. Learning subjective adjectives from corpora//Proc. of the 17th National Conf. on Artificial Intelligence(AAAI-2000).Texas,USA,(2000).

[4] Wiebe J,Wilson T,Bruce R,et al.Learning subjective language.Technical Report TR-02-100,Pennsylvania ,USA (2002).

[5] Wiebe J, Wilson T, Bell M. Identifying collocations for recognizing opinions. In: Webber BL, ed. Proc. of the ACL/EACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. Morristown: ACL, 24−31 (2001).

[6] Wiebe J, Wilson T. Learning to disambiguate potentially subjective expressions. In: Roth D, van den Bosch A, eds. Proc. of the Conf. on Natural Language Learning (CoNLL). Morristown: ACL: 112−118 (2002).

[7] Wilson T, Wiebe J, Hwa R. Recognizing strong and weak opinion clauses. Computational Intelligence, 22(2): 73−99 (2006).

[8] Yu Hong,Hatzivassiloglou V. Towards answering opinion questions:Separating facts from opinions and identifying the polarity of opinion sentences//Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics:129-136 (2003).

[9] Pang Bo,Lee L. A sentimental education:Sentiment analysis using subjectivity summarization based on minimum cuts//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics:271-278 (2004).

[10] Yuan Fang,Duansheng Chen,Yangyang Wu. Semi-supervised learning approach to recognize subjective sentences in microblogs. Application Research of Computers,2014,31(7):2035-2039 (2014).

[11] Lin ChengHua,He Yulan. Joint sentiment/topic model for sentiment analysis//Proceedings of the 18th ACM conference on Information and knowledge management. ACM:375-384 (2009).

[12] Salton G,Wong A,Yang C S.A vector space model for automatic indexing.Communications of the ACM,18(11):613-620 (1975).