Summer 6-27-2016

PUBLIC OPINION ANALYSIS BASED ON PROBABILISTIC TOPIC MODELING AND DEEP LEARNING

Baojun Ma
Beijing University of Posts and Telecommunications, mabaojun@bupt.edu.cn

Hua Yuan
University of Electronic Science and Technology of China, yuanhua@uestc.edu.cn

Yan Wan
Beijing University of Posts and Telecommunications, wanyan@bupt.edu.cn

Yu Qian
University of Electronic Science and Technology of China, qiany@uestc.edu.cn

Nan Zhang
Tsinghua University, nanzhang@tsinghua.edu.cn

See next page for additional authors

Authors
Baojun Ma, Hua Yuan, Yan Wan, Yu Qian, Nan Zhang, and Qiongwei Ye

PUBLIC OPINION ANALYSIS BASED ON PROBABILISTIC TOPIC MODELING AND DEEP LEARNING

Baojun Ma, School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China, mabaojun@bupt.edu.cn

Hua Yuan, School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China, yuanhua@uestc.edu.cn

Yan Wan, School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China, wanyan@bupt.edu.cn

Yu Qian, School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China, qiany@uestc.edu.cn

Nan Zhang, School of Public Policy and Management, Tsinghua University, Beijing, China, nanzhang@tsinghua.edu.cn

Qiongwei Ye, School of Business, Yunnan University of Finance and Economics, Kunming, China, yqw@ynufe.edu.cn

Abstract

With the rapid development of Internet, especially the social media technologies, the public have gradually published their perception of social events online through social media. In Web2.0 era, with the concept of extensive participation of public in social-event-related information sharing, the effective content analysis and better results presentation for these media published online thus possesses significant importance for public opinion analysis and monitoring. In view of this, this paper proposes a novel method for public opinion analysis on social media website. First, the probabilistic topic model of Latent Dirichlet Allocation (LDA) is adopted to extract the public ideas about the distinct topics of certain event, and then the deep learning model named word2vec is used to calculate the emotional intensity for each text. Next, the underlying themes in the whole as well as the events of emotional intensity are investigated, and the variation trend of public's emotion intensities is tracked based on time series analysis. Finally, the rationality and effectiveness of the method are verified with the analysis of a real case.

Keywords: Social media, public opinion analysis, probabilistic topic modeling, deep learning, emotion evolution analysis.

1 INTRODUCTION

With the development and popularization of Internet, the participation of the public to social events is gradually increasing. Now, people are not only getting information from the Internet, but also participating actively in information diffusing and opinion expressing. Recently, with the extensively usages of various social platforms, more and more information about certain hot social event is revealed through online social media and then attracts more people to concern and discuss. Hence, social platforms begin to act as an important source of Internet public opinion. In this context, mining public opinions and viewpoints regarding to certain hot issue from social platforms, and forecasting the trend of social events according to the changes of public's emotional intensities, are of great importance for public opinion monitoring and early warning (Ma, et al., 2016).

However, two challenges will prevent people from efficient mining and analyzing social media contents. Firstly, since there are normally great amount of data for an event on social platforms, people are often unable to understand such various viewpoints within a short time (Ma and Wei, 2012). The second issue is about the evolution of public opinions (Zhang and Ma, 2015). It is known to all that the public would generate new understanding and viewpoints for that event as time goes on. In addition, the emotional intensity of the public may change continuously along with the event development. Therefore, the critical point lies on not only how to calculate public's emotions for an event, but also the way to investigate the evolution of these emotions with temporal change.

Since the information about a hot event is normally very complex, for example, the public are often holding different opinions for that event. In literature, the most researches are more focused on how to discover hot topics from social media contents while less attention has been paid on the issue of emotional intensity tracking (Luo, et al., 2013). By taking the perspectives of probabilistic topic modeling and deep learning with time series analysis, this paper proposes a novel social event public opinion analysis method based on emotion intensity to realize the idea of event evolution tracking on a social platform. To that end, the Latent Dirichlet Allocation (LDA) model is used firstly to explore the public's different viewpoints and opinions about a hot event on social platforms. Then, the deep learning model named word2vec is introduced to calculate the vector for each term in the corpus, and then obtain emotional words and their corresponding emotional intensity by calculating their cosine distance; next, by using time series analysis, the change of public's emotional intensity along with the evolution of the event is tracked. Finally, a real case study is conducted to show the effectiveness of the proposed method.

2 RELATED WORK

The essence of this research is to use the LDA model, word2vec model, and time series analysis to analyse the public emotion intensity for a given social event. Thus the related researches are around three parts of probabilistic topic modeling, neural network language modeling and emotion analysing.

Probabilistic topic model is one of the most common topic modeling methods and its essence is using statistical methods and theories to find and extract thematic information from the huge amount of text, among which the Latent Dirichlet Allocation (LDA) model uses a three layers Bayesian framework and assumes that the topics and words in documents are all polynomially distributed (Blei, et al., 2003). LDA can be used to identify the hidden topics in large document collection or corpus. For these reasons, this work will use the LDA model for the task of topic extracting and analyzing.

The goal of statistical language model is to learn a joint probability function for the word sequences in a document collection. However, the biggest difficulty of these models lies in the dimension disaster. Fortunately, the newly emergence of neural network language model can be used to address the problem. In 2013, Google Inc. has opened a deep learning tool of word2vec (Tomas Mikolov, et al., 2014), in which the former proposed models and algorithms, i.e., NNLM, CBOW, Skip-gramm models, etc. (Bengio, et al., 2003; Hinton, et al., 1986; Tomáš Mikolov, et al., 2009), are integrated. It has been proved that word2vec is very efficient in capturing the latent semantic similarity between the words. In this work, the analysis of emotion intensity of text is based on the calculation word2vec.

Moreover, the emotional dictionary based sentiment analysis has a higher accuracy in semantic polarity classification, and it enables people to conduct the fine-grained emotion research in text, in which the semantic-similarity-based methods (Zhu, et al., 2006), weighting priority method (Li, et al., 2008) and the improved SO-SD algorithm (Xue, et al., 2014) are typical ones. However, due to the limitations of natural language processing technology as well as the related data extraction method, it is difficult to find and get the hidden information from data, thus there is a very big space for improving the emotional-dictionary-based method in subsequent research.

3 RESEARCH METHODOLOGY

3.1 The Research Framework

An overview of our research framework is presented in Figure 1.

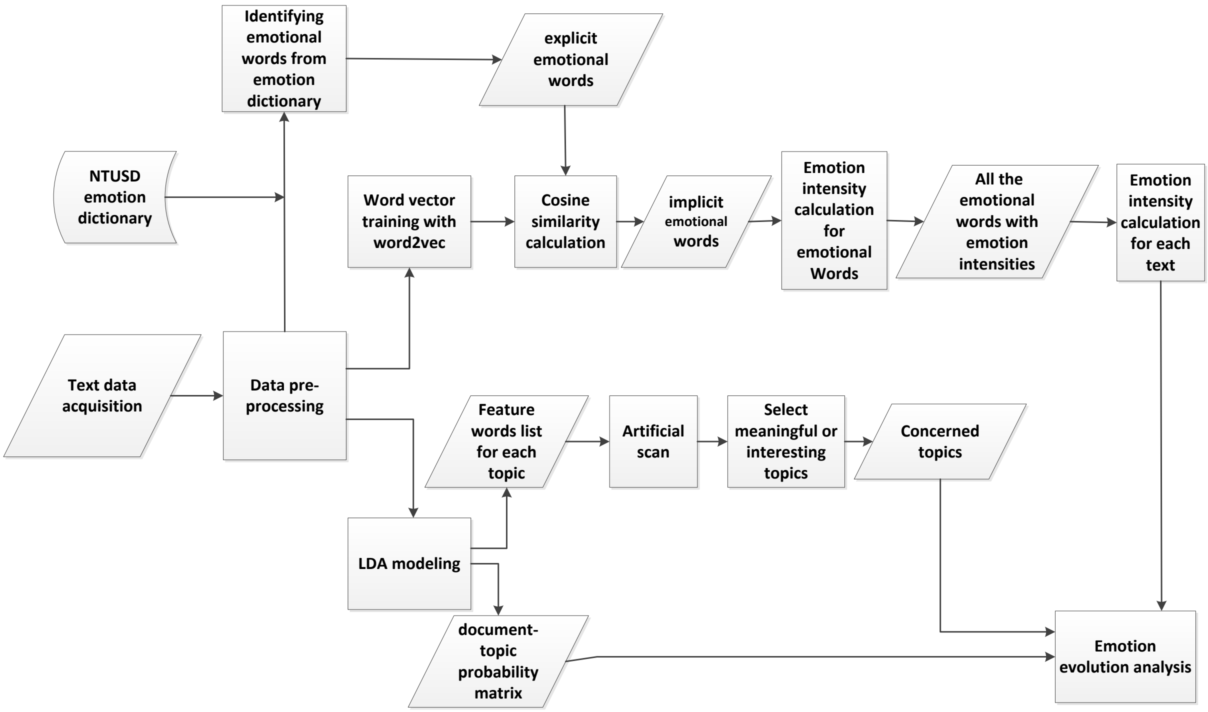

Figure 1. *The research framework in this study.*

In general, the analysis framework consists four parts:

- **Data processing**: crawling related text data from the social platform and carrying on the data preprocessing;
- **Topic modeling**: using the LDA method to model the topics distributed in the crawled text, and retrieving meaningful or interesting topics through artificial analysis of manual annotation;
- **Emotion intensity calculation**: Based on emotion dictionary, using word2vec to obtain implicit emotional words in corpus and calculate the emotion intensity of each piece of the text;
- **Emotion evolution analyzing**: For a selected topic or topics, using time series analysis to observe the change trends of emotional intensity.

3.2 Data Processing

Let t_i denotes a piece of message published online, and then all the data crawled from the selected social media platforms is

$$T = \{t_1, ..., t_i, ...\} \tag{1}$$

In order to implement efficient semantics related analyzing, necessary data pre-processing is needed. Hence, the general processing in NLP, such as word segmentation, spam (noise phrases) filtering and stop words removing are involved in this phase. The direct result of data pre-processing is to transform text t_i into a set of words:

$$t_i = \{w_{ii}, ..., w_{ij}, ...\} \tag{2}$$

3.3 Topic Modeling

In this work, LDA modeling is followed after data pre-processing, through which, two groups of useful information can be obtained: (1) feature words list for each topic, i.e., words with top relevance to the hidden topic; and (2) document-topic probability matrix, in which, each line corresponds to a document and each column corresponds to an underlying topic, thus each single numerical value in matrix represents the probability that a document belongs to a corresponding topic. Based on these two groups of information, artificial scan is required to make a judgment of the feature words for each topic to filter out some meaningful or interesting topics.

3.4 Emotion intensity calculation

Word vector training is performed on aggregated global word-word co-occurrence statistics from a corpus and the resulting representations showcase interesting linear substructures of the word vector space, especially, how frequently words co-occur with one another in a given corpus. Formally speaking, given a sequence of training words $w_1, ..., w_T$, the objective of the word vector model is to maximize the average log probability:

$$\frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{j \in nb(t)} \log(p(w_j \mid w_t)) \tag{3}$$

where $nb(t)$ is the set of neighboring words of word w_t. Based on this genius work, word2vec, a neural network implementation proposed by Google in 2014 to learn distributed representations for words, learns quickly relative to other models. In addition, word2vec does not need labels in order to create meaningful representations. This is useful, since most data in the real world is unlabeled. In the proposed research framework in this paper, word2vec will train each of the word in T into a K-dimensional numerical vector:

$$Word2vec : w_{ij} \to \omega_{ij} \tag{4}$$

Here, $\omega_{ij} = \{\omega_{ij}^1, ..., \omega_{ij}^K\}$ and K is a predefined parameter in word2vec. Obviously, the results generated by word2vec can be used to obtain the potential emotional words. Moreover, we can define the emotional intensity of a word as the cosine distance between itself and the words in an emotion dictionary. As we mentioned above, there may be two types of explicit (emotion word stated) and implicit (emotion only implied) emotional words lie in text t_i, both of which have important contributions for the emotion of t_i. In general, we can take all the words in a known "emotion words dictionary", i.e. D, as the explicit emotional words. To facilitate the calculation, we partition D as:

$$D = D^+ + D^- \tag{5}$$

where D^+ means the set of all the words with positive sentiment while D^- means all the words with negative sentiment. All the explicit emotional words in t_i can be acquired as $D \cap t_i$. Now, the challenge is how to identify the implicit emotional words from $\bigcup_{i=1}^{|T|} t_i$. Obviously, they exist in data set $\bigcup_{i=1}^{|T|} t_i \setminus D$. Thus, we implement the *k*-nearest-neighbor algorithm on both D^+ and D^- to gather 2*k* nearest neighbors for word w_{ij}, namely $D_{w_{ij}}^+$ and $D_{w_{ij}}^-$, where $w_{ij} \in \bigcup_{i=1}^{|T|} t_i \setminus D$. The emotion intensity w_{ij} can be calculated as following (Xue, et al., 2014):

$$SO - SD(w_{ij}) = \Sigma_{w_s \in D_{w_{ij}}^+} SD(w_{ij}, w_s) - \Sigma_{w_t \in D_{w_{ij}}^-} SD(w_{ij}, w_t) \tag{6}$$

where $SD(w_1, w_2)$ denotes the similarity distance between *word*$_1$ and *word*$_2$, its value can be obtained commonly by the cosine similarity as following:

$$SD(w_1, w_2) = \frac{\Sigma_{m=1}^{K} \omega_1^m \omega_2^m}{\sqrt{\Sigma_{m=1}^{K} (\omega_1^m)^2} \sqrt{\Sigma_{m=1}^{K} (\omega_2^m)^2}} \tag{7}$$

According to the value of $SO - SD(w_{ij})$, we can make an evaluation for the sentiment polarity as follows:

- w_{ij} is labeled as an (implicit) positive emotional word if $SO - SD(w_{ij}) > p$;
- w_{ij} is labeled as an (implicit) neutral emotional word if $SO - SD(w_{ij}) \in [q, p]$;
- w_{ij} is labeled as an (implicit) negative emotional word if $SO - SD(w_{ij}) < q$.

Here, *q* and *p* (*q* < *p*) are two thresholds. Without losing generality, we define the emotion intensity of w_{ij} as following:

$$EI(w_{ij}) = \begin{cases} 1, & \text{If } w_{ij} \in D^+; \\ -1, & \text{If } w_{ij} \in D^-; \\ SO - SD(w_{ij}), & \text{If } w_{ij} \in \bigcup_{i=1}^{|T|} t_i \setminus D. \end{cases} \tag{8}$$

Finally, according to the acquired emotional words as well as their emotional intensity, the emotional intensity of text t_i can be calculated:

$$EI(t_i) = \Sigma_j EI(w_{ij}) \tag{9}$$

3.5 Emotion evolution analysis

In this part, we cast the emotion evolution analysis task as a kind of time series analysis for emotion intensity. Thus, all the text in $\bigcup_{i=1}^{|T|} t_i$ should be sorted firstly according to their publishing time stamp. Two types of emotion intensity analysis are involved: *analysis for the whole event* and *the analysis for some topics generated by the event*.

For the former, emotion intensity for each text is calculated firstly. Then, evolution of the public's emotion for the whole event can be seen by cumulating all the emotional intensity values within the same time slot. For the latter, the emotion evolution analysis is topic-oriented. At the first step, both the emotion intensity for text t_i and the probability that t_i belongs to a particular topic should be taken into consideration, such that to obtain the relative emotion intensity for text t_i with respect to the specified topic. In general, their product will be calculated. Then, all the emotional intensities (regarding to the topic) within the same time slot are all cumulated.

4 EXPERIMENTAL RESULTS ON A REAL CASE

4.1 The data

The training data were crawled from an online social website of weibo.com. In order to demonstrate the effectiveness of the proposed research method, we selected the recent hot social event of “A women driver in Chengdu is beaten” as a case for analysis, which has attracted widespread attentions and been discussed extensively online.

In weibo.com, the data for the targeted event has an identical hashtag of “#A women driver in Chengdu is beaten for changing lanes#”. So, we can use the AJAX based web crawler for the HTML parsing to grab the corresponding microblog data. These texts are mostly published before 2015-05-17 which is seen as the ending time point for such a hot event. Altogether mixed types of 11,899 valid messages in form of short text (each text is limited within 140 Chinese characters) are crawled from the target website. At last, duplicate documents and document with invalid content are removed. In total, there are 6,989 pieces of valid blogs for the experiments.

In the initial data pre-processing stage, we use the ICTCLAS toolkit for word segmentation. Moreover, a stop-words-glossary presented by Harbin Institute of Technology is adopted to filter out the stop words. In addition, the word vector dimension, i.e., K, is set to 50 in word2vec training, and emotional dictionary of NTUSD (published by Taiwan University) is used for the emotional intensity calculation.

4.2 Hidden topic discovery

In the crawled dataset, we conduct the LDA model with 5-15 as the topic number for probabilistic topic modeling. Based on additionally artificial evaluating, 11 may be the best number for topic modeling with LDA which holds the best semantic information of the whole dataset. The most representative of 5 topics as well as 10 of their associated feature words are shown in Table 1. With the results of LDA modeling, the overview of all kinds of opinions for the targeted event published by the public can be constructed, through which the follow-up readers may have a preliminary learning and understanding about the event.

Topic 0 (Criticizing the male driver from the legal point of view)	**Topic 1** (Arguing by the woman driver’s mother)	**Topic 2** (The woman driver was searched by human flesh)	**Topic 6** (The woman driver’s condition got worsened)	**Topic 8** (The male driver beat the woman driver too hard)
人 (person)	女司机 (woman driver)	人肉 (human flesh)	宣称 (declare)	女司机 (woman driver)
法律 (law)	视频 (video)	转发 (repost)	女司机 (woman driver)	下手 (beat)
社会 (society)	慈善 (charity)	中国 (China)	舆论 (public opinions)	脸 (face)
暴力 (violence)	母亲 (mother)	开车 (drive)	女儿 (daughter)	凶狠 (fierce)
危险 (danger)	机构 (institution)	删除 (delete)	加重 (get worsened)	变道 (change lanes)
交通 (transportation)	女儿 (daughter)	挑衅 (provoke)	病情 (state of an illness)	男司机 (male driver)
文明 (civilization)	搞 (make)	道德 (morality)	母亲 (mother)	显示 (display)
强行 (force)	别车 (hit other cars)	四川 (Sichuan)	别车 (hit other cars)	视频 (video)
马路 (road)	辩解 (argue)	女人 (woman)	呕吐 (throw up)	脑震荡(cerebral concussion)
网络 (internet)	采访 (interview)	开房 (open room)	发高烧 (have a high fever)	骨折 (fracture)
……	……	……	……	……

Table 1. *The most representative of 5 topics each with 10 feature words.*

4.3 Implicit emotion words identification

In total, 612 explicit positive emotional words and 1,315 explicit negative emotional words are identified directly from the weibo data with the help of NTUSD emotion dictionary. Further, with the method proposed in this work, i.e., applying the *SO-SD* method by measuring the similarity distance between any two word vectors (generated by word2vec with dimension K =50). In the calculation, we take the k =20 to identify the nearest neighbours for each word, and the thresholds in *SO-SD* are set as p = 0.05 and q = -0.05.

Finally, we identify 6,876 implicit emotional words from the corpus, in which 5,533 are corresponding to positive emotion and the remainder 1,343 are negative emotional words. Due to the limitation of space, in Table 2, we only illustrate 10 representative emotional words respectively. Combining the explicit emotion words with the identified implicit emotion words, we thus build a dedicated emotional dictionary for studying the emotion evolution of the hot event.

Implicit positive emotion words	*SO-SD value* (positive)	Implicit negative emotion words	*SO-SD value* (negative)
合情理 (reasonable)	0.328784	心如蛇蝎 (have the heart of a devil)	-1.553315
耐心 (patience)	1.362254	违法乱纪 (violate the law and discipline)	-0.933532
坚守 (stand fast)	1.215394	错误百出 (full of mistakes)	-2.803829
正名 (rectification of name)	1.177136	不规范 (irregularity)	-2.271391
理所当然 (of course)	1.147314	坎坷 (bumpy)	-2.188523
霸气 (domineering)	0.595432	轻狂 (frivolous)	-2.182535
温良恭俭 (respectful)	0.662376	酒驾 (drunk driving)	-1.943230
有理有据 (well-founded)	0.309414	大打出手 (strike violently)	-1.874091
以理服人(convince people by reasoning)	0.621868	病入膏肓 (beyond recovery)	-1.872049
规范化 (normalization)	0.588630	吃闭门羹 (kiss the post)	-1.226578

Table 2. *Representative implicit emotional words identified from the weibo data.*

4.4 Emotion evolution with time

Using the dedicated emotion dictionary and the emotional intensity of each emotional word, at this step, we can calculate the emotional intensity for each piece of message. Meantime, these messages can be categorized into corresponding time slots according to their publishing time.

Based on the five representative topics, we take continuous 12 hours as time slot, and draw out the changes of emotional intensity of each topic over various time slots from 12:00, May 3rd to 24:00, May 17th (see Figure 2 (left)). The evolutionary process of emotion intensity for the whole event is shown in Figure 2 (right). As we can see from these two figures, most of the public's emotions for the event are absolutely negative, thus we should pay more attention on the changes of emotional value along with time in understanding the event rather than the polarity of emotion. The key points for the event evolution in reality are collected in Table 3 for the comparative purpose.

Key time point	**Important incident happened**
2015-05-03 18:22	The event has been exposed online
2015-05-04 18:46	The male driver apologized for his beat.
2015-05-04 20:22	The monitoring video clips have been released, which made the public opinions reversed.
2015-05-05 13:01	Since the woman driver was searched by human flesh, she claimed to investigate and affix netizens' legal liabilities.
2015-05-05 15:41	The woman driver's father was not willing to accept the apology.
2015-05-06 17:38	The woman driver's mother argued that the reason why they changed lanes is to do charity.
2015-05-06 22:20	The woman driver's mother claimed that her daughter's condition got worsened.
2015-05-07 11:49	The hospital denied that the woman driver's condition got worsened.
2015-05-11 10:42	The woman driver apologized for this event.

Table 3. *The key time points of the event evolution in this case.*

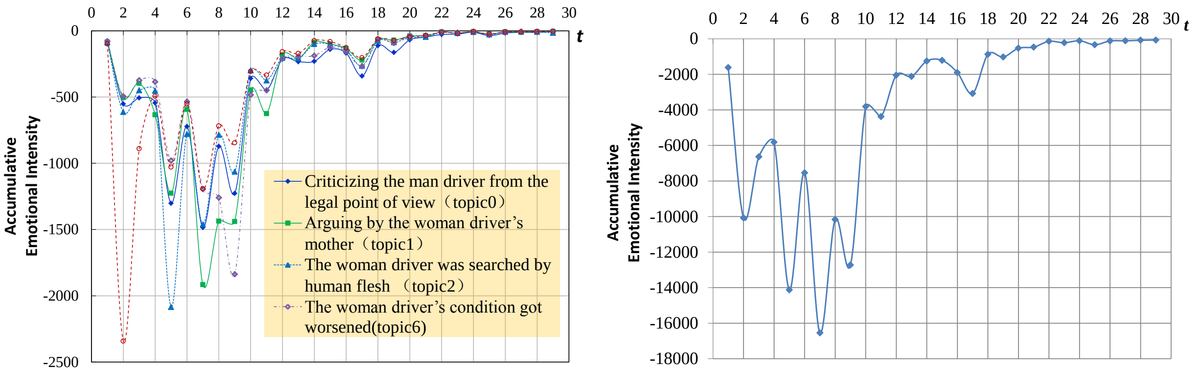

Figure 2. *The evolutionary process of emotional intensity of each topic (left) and the whole event (right) over various time slots.*

In combination with the contents in Figure 2, it is easy to verify that this information has strong relation to the subsequent evolution of the observed event in reality:

- The topic for condemning the male drivers (i.e., topic 8), has the very high emotional intensity from the very beginning of event since it was exposed by the media. Along with the apology made by the male driver, and the monitoring video clips released by a third-part, the emotion intensity for topic 8 dropped rapidly.
- Especially, the video recorder released that the woman had a relative poor driving habits on the road, critical messages thus reduced dramatically for the male driver and then aimed in succession at the woman side. Accordingly, the emotion intensity of the relevant topic (i.e., topic 2) increased immediately at time point 5.
- Another key point for the event was the justification making by the mother of the woman driver, however, the emotion intensity of related topic (i.e., topic 1) increased immediately after that, which shows that the excuse made by the mother did not accepted by online readers. Meanwhile, with the woman driver's condition claimed from the hospital side, the corresponding topic (i.e., topic 6) also got an immediately rise in emotion intensity.

In comparing with the event evolutionary process, we can draw a conclusion that the evolutions of the emotion intensity calculated by our method for the topics are conformed to that happened in real. Therefore, the research method can be thought as reasonable and effective for emotion evolution analysis and could be helpful for scholars and public administrators to conduct public opinion analysis.

5 CONCLUSION AND FUTURE WORK

This study is aimed at the viewpoint and opinion analysis for hot events diffused on the social platforms. To that end, a novel research framework for emotion evolution is proposed based on the probabilistic topic modeling and deep learning methods. The experimental results with the real event data crawled from online social media show both the rationality and validity of the proposed methods in this paper. The limitation and direction of future work are about involving more reasonable emotion intensity calculation algorithms to improve the precision of the proposed method, and considering with more online comments data as well as rigorous evaluation method to make the results comprehensive, completed and robust.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No. 71402007/ 71473143/71471019) and Yunnan Provincial E-Business Innovation and Entrepreneurship Key Laboratory of Colleges and Universities (YES 2014 [16]).

References

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research, 3* (Feb), 1137-1155.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*993-1022.

Hinton, G. E., Mcclelland, J. L. and Rumelhart, D. E. (1986). *Distributed representations*: MIT Press, Cambridge, MA.

Li, D., Cao, F., Cao, Y. and Wan, Y. (2008). Text Sentiment Classification Based on Phrase Patterns. *Computer Science, 35* (4), 132-134.

Luo, X., Zhang, J. and Duan, W. (2013). Social media and firm equity value. *Information Systems Research, 24* (1), 146-163.

Ma, B. and Wei, Q. (2012). Measuring the coverage and redundancy of information search services on e-commerce platforms. *Electronic Commerce Research and Applications, 11* (6), 560-569.

Ma, B., Zhang, N., Liu, G., Li, L. and Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management, 52* (3), 430-445.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2014). word2vec. In: accessed 2014-04-15. https://code.google.com/p/word2vec.

Mikolov, T., Kopecký, J., Burget, L., Glembek, O. and Černocký, J. H. (2009). Neural network based language models for highly inflective languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)* (pp. 4725-4728). Taipei: IEEE.

Xue, B., Fu, C. and Shaobin, Z. (2014). A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec. In *2014 IEEE International Congress on Big Data (BigData Congress)* (pp. 358-363): IEEE.

Zhang, N. and Ma, B. (2015). Constructing a Methodology Toward Policy Analysts for Understanding Online Public Opinions: A Probabilistic Topic Modeling Approach. In E. Tambouris, H. J. Scholl, M. Janssen, M. A. Wimmer, K. Tarabanis, M. Gascó, B. Klievink, I. Lindgren, M. Milano, P. Panagiotopoulos, T. A. Pardo, P. Parycek & Ø. Sæbø (Eds.), *Electronic Government and Electronic Participation* (Vol. 22, pp. 72-79): IOS Press.

Zhu, Y., Min, J., Zhou, Y., Huang, X. and Wu, L. (2006). Semantic Orientation Computing Based on Hownet. *Journal of Chinese Information Processing, 20* (1), 14-20.