Analysis of the Passion Criminal Tendencies Based on the Topic-Emotion Model

Jiaqian Sun¹, Menghui Wang¹, Jiasheng Huang¹, Anqi Wang¹ and Linying Jiang¹,²
¹Department of Software Engineering, Northeastern University, Shenyang, Liaoning, 110169, China
²Corresponding author
*Corresponding author’s e-mail: Jiangly@swc.neu.edu.cn

Abstract. With the penetration of the Internet into daily life, more and more crime clues could be discovered on the Internet. Therefore, this paper uses a new method to predict whether the instant message on the internet has the tendency of emotional crime. It is because that the single theme-based analysis of criminal tendency, which ignores the impact of emotions, may confound the declarative sentences and normal expression with statements that contain criminal tendencies. The method used in this study introduces the analysis of emotional fluctuations while analysing themes. In order to solve the problems above, the study adopts seq2seq model to analyse emotional fluctuation. Themes and the values of emotional fluctuation are put into linear SVM as input vector to judge whether the instant message has an emotional crime tendency. The average predication accuracy of the model reached 86%, which basically accords with the assumption made before the experiment. This model can not only be used to analyse emotional crime, but also to screen the radical opinions on the internet.

1. Introduction
Text Data Mining is a process to understand the content and meanings by analyzing texts with rich semantics. It has become an increasingly popular and important research field in the data mining domain.

Instant messages on the Internet are one of the fastest growing social ways. The Internet has replaced texts and become the main way to spread the news. In such a complex network environment, various potential problems are also derived. For example, the frequency of cyberbullying and cybercrime are increasing year by year. The number of users on the Internet is huge: as far as QQ is concerned, according to the official website, the average number of online users during prime time (18:00-19:00) exceeds 300 million [1]. Such a large amount of user traffic is inevitably accompanied with a large number of instant information records. However, there's no perfect solution of obtaining useful information by analyzing instant message so far.

2. Issue Raised
The author before [2] classifies the conversation reply into two types, the short response waiting for time class-SRWTC and long response waiting time class-LRWTC. He claims that he can predict response time by analyzing conversation content. In another passage [3], the author establishes classification system that creates a concept-based profile, which represents a summary of the topics,
discussed in a chat room or by an individual participant and determines the subject of the conversation texts by analyzing interactive texts. However, there is no further way to analyze emotions in the passages.

In the previous study, the researchers have made an effective analysis of the relationship between conversation topics and criminal tendency [4]. However, the accuracy is low. For example, a conversation may involve the topic of guns and ammunition. If the analysis of this conversation just considers about the topic, it is more likely that the conversation is involved in the field of crime. However, if we analyze the content, it is also possible that the conversation simply discusses the relevant knowledge in a plain tone. In the definition of safe cybercrime [5], it can belong to criminal offense only if the cybercriminals realize that their behavior will cause the result of jeopardizing national security and hope the danger happens. And if an individual is in a state of negative passion and stress, it is easier to cause criminal behavior [4].

In this article, in addition to analyzing the topic of the texts [6] [7], it also introduces the analysis of conversational sentiment [8] [9]. This article mainly analyzes local conversation which consists of instant message and finds out the relationship between topics and sentiment value (use the sentiment dictionary [10] and NN model) according to the result of an analysis [11]. Then we can analyze whether the content of the conversation has a criminal tendency. After the emotional extremes are introduced in the local conversation analysis which consists of instant messages between two people, the results of the text analysis can be expressed as the difference of sentiment value between A and B on the X topic. For example, if the emotional extreme value is high when A and B talk about guns, we can infer that the conversation is involved in the crime field.

3. The construction of the sentiment model and theme model

3.1. Terminology and Definitions
There are several features of conversations generated from instant messages: each message is labelled with sender and timestamp, so chat record could be represented as a set of messages. For example: in chat record \( C = \{ M_{a1}, M_{b1}, M_{a2}, M_{a3}, M_{a4}, M_{a5}, M_{b2}, M_{a6}, \ldots\} \), \( M_{ai} \) means the \( i_{th} \) message sent by sender A. Similarly, \( M_{bj} \) means the \( j_{th} \) message sent by sender B. We use \( T_i \) to represent the timestamp of the \( i_{th} \) message, so we could calculate interval time \( \Delta t \) between adjacent messages. By analyzing 187 chat records of 5 people, we find that distributions of most persons’ interval times approximately match skewed distribution (Recorded as \( f(x)_{\text{interval distribution}} \)). As a result, we choose interval times that lie at 95 percentiles to represent \( t_{\text{max}} \), as in equation (1), which is regarded as the condition of segmenting chat record into local conversations. Besides, the mean of all interval times in a chat record excluding interval times lie at 95 percentiles and above, is regarded as the condition of single conversation, recorded as \( \Delta \bar{t} \), as in equation (2).

\[
\int_0^{t_{\text{max}}} f(x)_{\text{interval distribution}} \, dx = 0.95
\]

\[
\Delta \bar{t} = E\left\{ \sum_{0}^{t_{\text{max}}} x \right\}
\]

3.1.1. Local Conversation: The set of messages that sent by both sides. Generation rule: First, get all the \( \Delta t \) of the chat record and check each \( \Delta t \) from the beginning of the record. If \( \Delta t \geq t_{\text{max}} \), regard the message before the interval as the end of the last local conversation, and regard the message after the interval as the beginning of the next new local conversation. \( LC_i \) means the \( i_{th} \) local conversation in the chat record.
3.1.2. Single Conversation: A complete sentence with subject and predicate generated by one side of chatters. A turn could be constructed by one or several messages. \(\text{Turn}_{An}\) is the \(n\)th turn produced by A. The generation method is written in part 4.1.

3.1.3. Turn: A set of instance message generated by both sides of chatters in short term, (the single conversation is the subset of the local conversation; local conversation is composed of several single conversations). Generation rule: First, get all the \(\Delta t\) of the local conversation, and check each \(\Delta t\) from the beginning of the record. If \(\Delta t \geq \Delta T\), regard the message before interval as the end of last single conversation, and regard the message after the interval as the beginning of the next new single conversation. \(SC_i\) means the \(i\)th single conversation in the local conversation.

3.1.4. Turn A: The set of Turn \(A_i\).

3.1.5. Word: The basic unit to analysis plain text, produced by analysing plain text by using word segmentation tools.

3.1.6. The theme of single conversation: The theme of a complete instance message produced by two persons, recorded as \(s\).

3.1.7. The themes of local conversation: The set composed by the themes of single conversation, recorded as \(S\).

3.1.8. The sentiment of Turn \(A_{mn}\): The sentiment value of \(m\)th turn produced by person A by using sentiment dictionary.

3.1.9. Sentiment \(n'\): The sentiment value produced by A in local conversation by using seq2seq model.

3.2. Constraint of Research
The same point between turn and short text is that their numbers of words are small. However, there are also some differences between the above two. Firstly, as more and more turns cumulate, the sentiments of the speaker may change a lot, they may change their emotion from happy to sad. Comparing with this, the short text is usually with a single emotion. Secondly, the short text is always with a personal idea or one theme, but the dialogue made up by instant message (local conversation) can represent two persons’ or even several people’s communicational relationship, so a local conversation can either be related to one theme or several themes. The two uncertainties above bring a lot of difficulties for the analysis of local conversations. What’s more, it is inaccurate to get the specific themes by only analyzing one person’s dialogue. It is better to consider the time sequential and the themes of one turn to get an accurate theme or several accurate themes of one local conversation. This paper limits the number of persons in the local conversation to 2.

3.3. The process of the generation of theme model
The original input of the model is a local conversation, which includes pure texts (excluding pictures, emoticons and etc.) sent by two participants, timestamps and usernames along with each message. As figure 1 shows, to get the themes of the input, the model starts with breaking up the local conversation into a number of single conversations according to the definition (using the interval time calculated by timestamps of adjacent messages). Then, using the word segmentation tool to get a set of words of every single conversation. After this, KNN algorithm is applied. Use the occurrence frequency of every word in the single conversation to create a vector which represents the single conversation. And regard the vector as the input of KNN algorithm to get the theme of each single conversation. Last, the theme set of local conversation is collected by gathering all the themes together.
3.4. The process of generation of sentiment model
Figure 2 shows the construct of the sentiment-text tree. Since this paper will analyse the sentiment value of A and B respectively, we divide the local conversation into the sets of turns of A and B respectively. As 3.3 shows, the words have research value only when they are put into a single turn. So firstly, we divide the sets of turns of A and B to several turns. And secondly, we split the turns into words (We analysis all of the words in units of turns). Thirdly, we generate the sentiment value of the single turn by using the model shown in the right model in figure 3. Finally, we generate the final sentiment value of Turn A by calculating the mean of sentiment values of Turn $A_i$.

4. The analysis of Criminal Tendency

4.1. Data pre-processing

4.1.1. The selection of data sets: The original corpus comes from actors’ lines of movies and teleplays, since it’s difficult to collect conversations involving crime topic. Another benefit of subtitles is that the movies or teleplays are labelled with the theme, which cuts the cost of labelling data. In this paper, we prepared 30 subtitle files as training set and 4 subtitle files as the validation set.
4.1.2. **Label theme for lines:** To obtain the original corpus, we crawled subtitles of movies which belong to 6 themes: criminal, shopping, daily, food, hospital, travel. Of each theme, there are 5 subtitle files of movies or teleplays. And labelled those subtle files with 6 themes and take them as the training set.

4.1.3. **Generation of sentiment dictionary:** This paper gets the complete sentiment dictionary by crawling data from comments of Amazon and T-mall, the steps are as followed:

- Get the score of comments and cut the comments into words.
- Make the scores as the input values, make the comments sentences as tags and put scores into CNN model, quantify the weight of words of different sentiment and generate a preliminary sentiment dictionary
- Integrate the preliminary sentiment dictionary into the existing sentiment dictionary, calculate the mean of the weight of words which both exists in the preliminary sentiment dictionary and existing sentiment dictionary; add the non-existing words into sentiment dictionary.

4.1.4. **Sentence construction and word segmentation:** Since the conversation may be informal, it is possible that a message sent by chatter would be an incomplete sentence. So, we have to group the scattered message into a turn. First, put together all the messages sent from one side. Then use the dependency analysis method from HANLP [12] to analyse the sentence elements grammatically. Last, according to the position of subjects, we separate turns from the message aggregate.

4.2. **Generation of model**

4.2.1. **Generation of KNN model for theme analysis:**

- Segment sentence in the training set into words, then generate a word frequency library by calculating the relative word frequency.
- Rank words that appear in the training set by relative word frequency and choose the top 100 words to constructing 30 vectors of 100 Dimensions to represent training set. Each dimension is valued by the word’s frequency. At last, use those vectors to generate the KNN model.

4.2.2. **Generation sentiment vector for sentiment analysis:**

- Generation of local conversation: Since the number of characters in movie or TV series is a lot, which is the same as group chat, to meet the limit of our research, this paper chooses two persons with the most interaction as the two sides of conversation and makes the lines as local conversation, then makes the two persons’ dialogue as two turns.
- Analysis the sentiment value of two person’s turns

\[ x_i = \sum_{i=1}^{l} \left[ (-1)^j \times m_{ij} \right] \]  

4.3. **Evaluation of model**

4.3.1. **Theme analysis:** Input the validation set to analysis the theme of them and the results show as followed:
Figure 3. shows the first dialogue is related to travel

Figure 4. shows the second dialogue is related to criminal and daily

Figure 5. shows the third dialogue is related to food

Figure 6. shows the fourth dialogue is related to shopping

4.3.2. Sentiment value analysis: We use the comparison method to determine the specific model for generating mode vector and final sentiment value of local conversation by using the validation set.

(1) seq2seq: Taking one side of local conversation “A” as an example, firstly, we use formula 1 to calculate $x_t$, before this we have already calculated the sentiment value of Turn $A_{t-1}$ as $h_{t-1}$ and put it as $x_{t-1}$; secondly we put $x_t$, $x_{t-1}$ into the formula $x_t' = f(x_{t-1}', x_t, c)$ ($c$ is the correction factor to prevent model from divergence). Analysising B is the same as analysising A.

(2) CNN&NN: Taking one side of local conversation “A” as an example, firstly, we use formula 1 to calculate $x_t$, then we input $x_t$ to the two formula: $x_t' = f(x_t)_{con}$; $x_t' = f(x_t)_nn$. When $x_t = 0$, the two models have no output value.

We use seq2seq, CNN and NN model to generate three polyline charts, compare them with the standard polyline chart and the results are as followed:
Figure 7 to 10 show the differences between standard value and the output of the three models (seq2seq, CNN, NN). Figure 8 is the enlarged version of figure 7.

As figure 8 and 9 show, the output of NN model and CNN model diverge; as all of the figure above show, the output of seq2seq is the most stable and accurate one.

Table 1 shows the standard deviation and the recall rate of the three models.

| Model    | Standard Deviation | Recall Rate |
|----------|-------------------|-------------|
| seq2seq  | 0.68              | 0.89        |
| CNN      | 0.29              | 0.78        |
| NN       | 14.9              | 0.43        |

From the results above, we use seq2seq to analyze sentiment value.

4.4. Analysis of criminal tendency

Firstly, we put the training data on the label “whether it is a crime-themed conversation”; secondly, we use seq2seq model to calculate the sentiment value, KNN model to calculate the themes of local conversation. We bring the sentiment value and the number that the local conversation’s crime subject relevance is greater than 3 as two dimensions into the SVM classifier. From the SVM classifier, we know the “sentiment value” and “the number of themes associated with criminal in a chat” in a chat with “high criminal tendency”. Finally, we put the validation data on the label “whether it is a crime-themed conversation” according to the classifier (1 for YES and 0 for NO) and calculate the accuracy of the model. The result is as followed:

| Model          | Accuracy |
|----------------|----------|
| seq2seq&KNN    | 86%      |
5. Conclusion
In this paper, the accuracy of the final criminal tendency analysis reached 86%. Because the sample data of the training set is not enough, some topics of the new subtitles cannot be correctly classified. In the future study, we will increase the amount of sample data, which requires us to further improve the efficiency of the algorithm and apply distributed computing on this algorithm, etc.

References
[1] Tencent (2018) Tencent Online Statistics. http://www.txzxtj.com/index.html
[2] Bengel, J., Gauch, S., Mittur, E., Vijayaraghavan, R. (June 2004) Chattrack: Chat room topic detection using classification. In: 2nd Symposium on Intelligence and Security Informatics (ISI-2004). Tucson, Arizona. pp. 266-277.
[3] Ikoro, G.O., Mondragon, R.J., White, G. (2017) Predicting response waiting time in a chat room. In: Computing Conference. London, UK. pp. 18-20
[4] Guo, Z.X. (2000) Discussion on Emotional Crime. Public Security Research, 03: 76–79.
[5] Peng, J.Y. (2015) Research on Some Problems of Internet Crime. Hunan University of Technology, Zhuzhou
[6] Li, S.Y., He, W. (2007) A Study on the Method of Feature Selection in Chat Text. Computer Science., 34:202-204
[7] Peng, J., Shi, Y.G., Gao, B.S. (2016) Session Topic Mining for Interactive Text Based on Conversational Content. Nanchang University, Nanchang
[8] Hu, Y.P. (2017) Research on the Theory of Crime Degree in the Big Data Environment. Jilin University, Changchun.
[9] Pang, B., Lee L. (2008) Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, 2: 1-135.
[10] Zhou, J., Lin, C. and Li, B.C. (2010) Research of Sentiment Classification for Netnews Comments by Machine Learning. Journal of Computer Application., 30(4):1011-1014
[11] Lin, L. (2005) A Research of the Conversation Mechanism of the Talk Show. Central China Normal University, Wuhan
[12] HanLP v1.2.8 (2014) Shanghai Linyuan Info Tech Ltd. http://hanlp.linrunsoft.com