ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 567-570
doi: 10.18178/wcse.2016.06.096

LINE Chat Record Emotion Classification Using SVM

Hsin-I Huang, Jeanne Chen, Tung-Shou Chen[+] and Yong-Ru Jhang

Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, No.129, Sec. 3, Sanmin Rd., North Dist., Taichung City 404, Taiwan

Abstract. There is abundance of precious information that can be tapped from the Social network sites. Much research has been done using data mining to tap information that can be used to forecast user emotion. It can also be used in user behavior forecasting like pop forecasting, stock analyses and more. However, most information is easily collected from the public domains. Most research is focused on public information analysis while very little research is focused on the private less accessible information. In this paper, the public post records from twitter are used to analyze private chat record from LINE. Training of the data is performed using support vector machine (SVM). The network public data are training data and the personal private data are testing data for emotion classification. Results showed that the public data used to analyze private data is feasible and resulted in significant level of accuracy.

Keywords: data mining, text mining, emotion classification.

1. Introduction

In recent years, much focus has been on text mining. The mined information contains precious information that is important to businesses. The amount of information from social network sites like LINE, Facebook and Twitter are incredible. Much research is focused on mining information from the social network sites. The information aids businesses to understand user behaviours and can also be used in many fields, such as pop forecasting and stock analyses.

Today, much research has been focussed on mining emotional information for the social network sites. Go et al. [1] proposed an approach to automatically classify sentiment in Twitter. The test data emphasized largely on emoticons and were significantly accurate predicting the sentiments. Xu et al. [2] classifies and analyses emotion on hierarchy for the Chinese blog posts which significant results. Gunarathne et al. [7] developed the Intellemo system for the instant messager used in smart phones. Other success included using the support vector machine (SVM) in emotion classifications [3].

There is abundance emotional information to be mined from the social websites like Facebook and Twitter and instant messagers like LINE, Whatsapp and Weibo. However, the information from Facebook and Twitter are easily available public data while LINE is difficult to access private data. The characteristic of data from LINE is short format and free style. Collecting this kind of data for data mining is difficult. Also this type of data is private and users do not have to consider public opinions when chatting and can, therefore, personal opinions can be expressed freely. This private data are different from the public data, and its analysis could provide a rich amount of information for future research on text mining.

This research proposed an emotion classification framework based on machine learning. The public social data is used to analyse private personal data for proving that the public Twitter data can supplement the difficulty of mining the private LINE data for training.

2. Method

[+] Corresponding author. Tel.: +886937275091
E-mail address: MashiroKinji@gmail.com

The proposed method is focused on Chinese data analysis. Training data is accessed from Twitter, and the test data is chat records provided by volunteers using LINE. Chinese word segmentation problem (CWS) is resolved using the CKIP[4] system, developed by the Academia Sinica, Taiwan. Finally, LibSVM [5] is used for classification. Fig. 1 shows the flow of the research process. The process included four phases, namely, training data collection, test data collection, feature selection, and classification phases. Detail on each phase is as shown in Fig. 1.

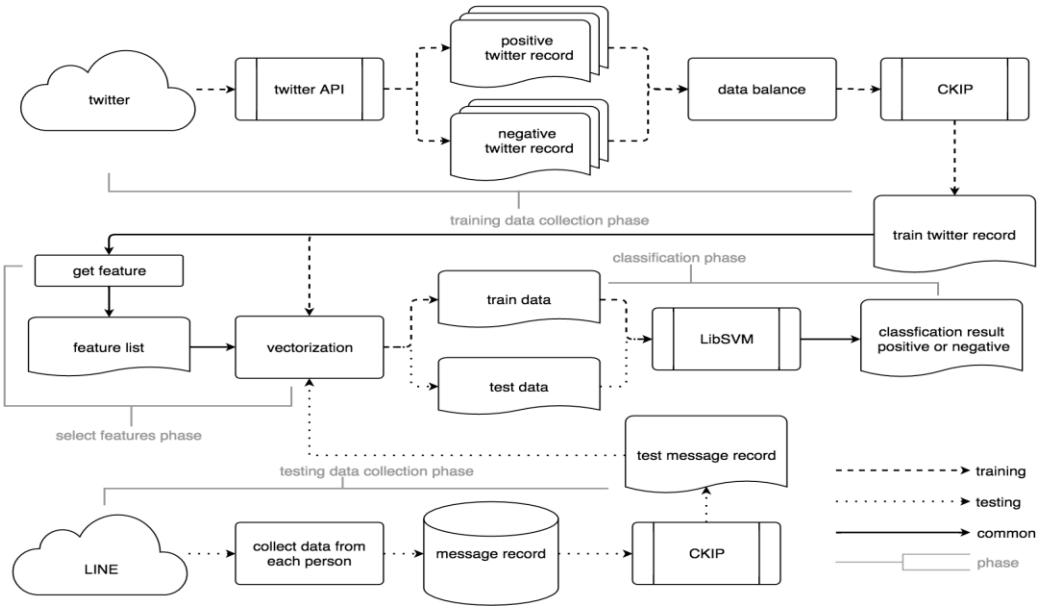

Fig. 1: Chat record emotion classification process.

The Chinese word segmentation problem (CWS) is due to the Chinese character display characteristic. Characteristically, each Chinese character is part of a word. The characters are stringed to make up a word or sentence and do not require any space separation. Other groups having the same problem include the Japanese and Korean language. Before doing text mining, text context has to be divided in words of units. CWS proposed three solutions for the text problems, namely, dictionary-based approach, statistics-based approach and hybrid approach. Many tools have been developed for solving this question. NLPIR and CKIP are popularly used segmentation systems for three basic reasons. Firstly, is because the systems have a default dictionary which is continuously updated. Secondly, is due to the systems' ability to detect new words. Third and the last reason, is the systems provide API for external users to access. In this research, we will use the CKIP system to resolve the segmentation problem.

Twitter posts will be used for training data. Positives and negatives are collected by using a Twitter API. The Twitter API emotion classification is based on emoticons like ":)", ":D" and ":(". Twitter cannot be collected for nonemotion data. Therefore, there are only two classes of training data, negative and positive. In actuality, our collection resulted in more positives than negatives. This is disadvantageous for support vector machine (SVM). In this research, collection of the positive/negative data will pick up as much in volume as the number of negative data based on the positive data for positive/negative data balance. These data has to pass a filter that is designed for Twitter proposed by Go et al. [1]. The selection is based on:

- Delete unnecessary emoticons
- Delete the posts included positive and negative
- Delete hyperlink, @username and any repost
- Delete the posts include ":P"
- Delete repeat posts

Test data is collected, randomly, from the LINE chat records provided by volunteers. The LINE special tags like "Sticker", "Video", "File" and "Photo" were deleted. The habit of most people using instant message chatting are also considered. Most speak a long statement followed by many short statements. Therefore, if a person does not speak over one hour, the system will merge the statements into one long statement. The data are classified manually and compared with the SVM result. Since 60% of the sentences

are nonemotion, so the data are classified into three classes, positive, negative and neutral.

Every word is regarded a feature in this research, and some rules are set for the Chinese data. First, Delete data include any non-Chinese word like Japanese or Korean words. These could cause mistakes when segmenting Chinese. Second, Punctuations are deleted to avoid inclusion in the feature list. The algorithm is as shown in Fig. 2. The data will be quantized after getting the feature list. Each word is compared with the feature list. If the word is included in feature list, the system will be updated one (see Fig. 3).

```
input:
  train_messages ← input training message
output:
  features ← output using features
process:
  for each train_message in train_messages do
    for each word in train_message do
      if word is punctuation then continue endif
      if word is not Chinese word then continue endif
      if word is in features then continue endif
      add word into features
    end for
  end for
```

Fig. 2: Feature list selection.

SVM is used in the classification phase. In recent years, many related tools have been used such as LibSVM [5] and SVMpref [6]. The SVM results showed high performance in the emotion classification. In this research, LibSVM is used. Due to the Twitter API, training data only have two classes, positive and negative. However, 60% of the sentences are neutral. Therefore, this research also made used of the reference algorithm proposed by Xu [2], where the use regression value is used to get value by LibSVM. The best value of threshold is obtained by using iteration to locate sentences of neutral class. Fig. 4 shows the algorithm.

```
input:
  features← input training message
  messages ← input using features
output:
  msgfeatures← output of each message and feature
process:
  for m to messages size do
    for each word in messages do
      for f to features size do
        if features"f" == word then
          msgfeatures"m""f"++
          break
        end if
      end for
    end for
  end for
```

Fig. 3: Feature list quantization algorithm.

```
input:
  train_messages ← training message that is included value of features
  test_messages ← test message that is included value of features
output:
  test_messages ← test message and classification information
process:
  model = svm_train(train_messages)
  for each test_message in test_messages do
    result = svm_predict(test_message)
    if result < negative_threshold then
      test_message.classification = NEGATIVE_TAG
    else if reulst > positive_threshold then
      test_message.classification = POSITIVE_TAG
    else
      test_message.classification = NEUTRAL_TAG
    end if
  end for
```

Fig. 4: Classification algorithm.

3. Experiment Results

The experimental data is from the Chinese data collected from October 2015 to December 2015. The data are divided into three groups, {S, ST0, STB}. Class S uses only SVM. Class ST0 uses SVM and classification algorithm in Fig 4. The value of threshold is not specified. This meant that the value of the threshold is zero. Class STB is different from class ST0 in that the threshold is its best value. Experimental results will show accuracy, threshold and training size.

In Table 1, the accuracy of S is very low. The reason is that the training data is from the Twitter API which cannot locate the neutral class. However, ST0 and STB can locate neutral class which resulted in higher accuracy than S. The best value of positive threshold and negative threshold is found by using iteration. The range of positive threshold is from 0 to 1, and the range of negative threshold is -1 to 0. In Fig 5, the value of positive threshold is 0.00004, and value of negative threshold is -0.00085.

Table 1: Result accuracy

Class	S	ST0	STB
Accuracy	0.23	0.56	0.65

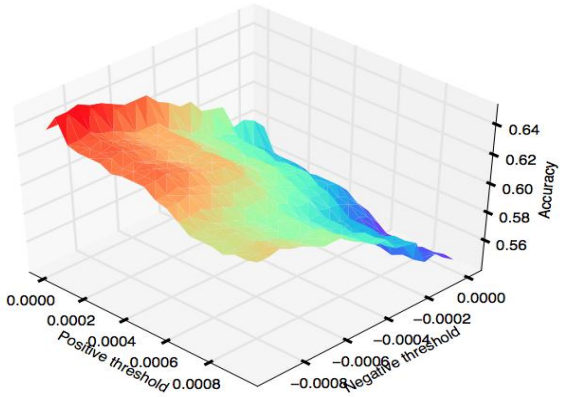

Fig. 5: Threshold iteration.

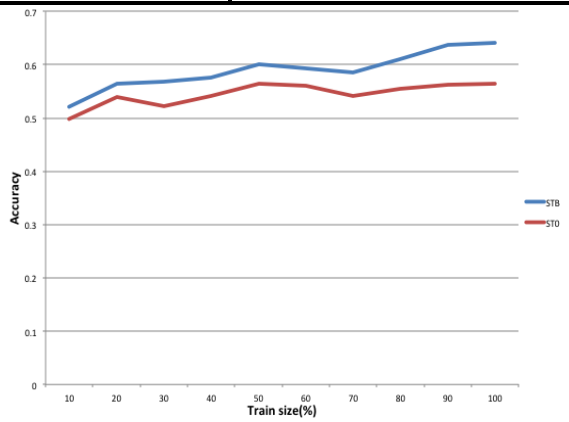

Fig. 6: Training size with accuracy.

In training the size of experiment, the maximum training size is 5MB. The lowest accuracy is 0.43 in the experiment. Fig 6 shows that accuracy has positive growth when training data size is higher. 5MB training data were passed thru train data collection phase, selecting feature phase and support vector machine phase. It has done within an hour, so this framework can refresh training model frequently. This reduces the effect of new words on emotion classification system which is often time consuming.

4. Conclusion and Future Work

In the accuracy experiment, this research shows the possibility of using public data to analyse private data. The highest accuracy is 0.65 which confirmed that the emotion classification framework has basic accuracy. Although it is not very high in comparison with other emotion classification, considering the free style and shortness of the instant message chat records the accuracy value is acceptable. Also, the training data size of experiment requirement is small for achieving significant accuracy. The maximum size used in the experiment is 5MB. The small size requirement allows for frequent updating of the training model.

In the near future, we plan to focus on the unique features of the chat record like speaking, specific speaker and speaking space. These features will be included into the framework of this research for accuracy improvement. When accuracy is high, text mining of private data is possible. The result could be used by marketing or human behavioral analysis.

5. References

[1] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009: 1-12.

[2] H. Xu, W. Yang, and J. Wang. Hierarchical emotion classification and emotion component analysis on chinese micro-blog posts. Expert Systems with Applications, 2015. 42: 8745-9752.

[3] D. Zhang, H. Xu, Z. Su, and Y. Xu. Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 2015.

[4] K.J. Chen and S.H. Liu. Word identification for mandarin chinese sentences. Proceedings of COLING 1992, 1992: 101-107.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2011. 2.

[6] T. Joachims. Training linear SVMs in linear time. Proceedings of the ACM conference on knowledge discovery and data mining, 2006: 217-226.

[7] S. R. Gunarathne, J. De Silva, E. M. C. P. Ekanayake, I. Samaradiwakara, P. S. Haddela, and P. A. Fernando. Intellemo: A mobile instant messaging application with intelligent emotion identification. Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on, 2013: 627-632.