NITMZ-JU at IJCNLP-2017 Task 4: Customer Feedback Analysis

Somnath Banerjee
Jadavpur University, India

Partha Pakray
NIT Mizoram, India

Riyanka Manna and **Dipankar Das**
Jadavpur University, India

Alexander Gelbukh
Instituto Politécnico Nacional, Mexico

Abstract

In this paper, we describe a deep learning framework for analyzing the customer feedback as part of our participation in the shared task on Customer Feedback Analysis at the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017). A Convolutional Neural Network (CNN) based deep neural network model was employed for the customer feedback task. The proposed system was evaluated on two languages, namely, English and French.

1 Introduction

Most of the companies provide their clients or customers provision to feedback in terms of register comments, complaints or suggestions in order to enhance service quality and the reputation of the company. Companies even provide toll free numbers for interactive communication where customers can speak with a service representative. Drastic change has been observed in customer feedback scenario after the advent of the computer. The use of computer has simplified the acquisition of information that is provided by customers. The use of Internet now allows companies to receive customer comments via electronic mail (email) and web page feedback techniques. Recently, a popular system is online dialogue interface where the service representatives interact with the customers live. Additionally, companies nowadays collect the customers feedback from various social media sites.

Understanding customer feedback is the most fundamental task to provide good customer services. The customer feedback is so valuable today that customer feedback analysis nowadays has become an industry on its own. A number of internet companies (also referred to as app companies) provide support to process and analyse the customer feedback for other companies. The business model for these app companies is to acquire customer feedback data from their clients and after analyzing the data using their internal tools these companies provide the reports to their clients periodically (Freshdesk, Nebula).

It is quite understandable that the reports which are generated by the app companies for their client are confidential materials. Also, the app companies keep the categorization of customer feedback as business secrets. However, three categorizations of customer feedback are openly available. The most commonly used categorization adopted by many website (SurveyMonkey[1]) is the five-class categorization, namely, *Excellent, Good, Average, Fair, Poor* (Yin et al., 2016). Another categorization, adopted by an app company Freshdesk, is also a five class classification which is a combined categorization of sentiment and responsiveness. The categorization includes the classes- *Positive, Neutral, Negative, Answered, Unanswered*. However, another app company called Sift adopted a seven-class classification which includes the classes- *Refund, Complaint, Pricing, Tech Support, Store Locator, Feedback, Warranty Info*. Although other categorizations for customer feedback analysis are present, most of them are not available publicly (Clarabridge, Inmoment[2], Equiniti[3]).

A common approach to text classification is to use Bag of Words (Harris, 1954) , N-gram (Cavnar et al., 1994), and their term frequency-inverse document frequency (TF-IDF) (Sparck Jones, 1972) as features, and traditional models such as SVM

[1] https://www.surveymonkey.com

[2] http://www.inmoment.com/products/

[3] https://www.equiniticharter.com/services/complaints-management

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 180–183,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

(Joachims, 1998), Naive Bayes (McCallum et al., 1998) as classifiers. However, recently, many researchers (Collobert et al., 2011; Conneau et al., 2016; Kim, 2014; Zhang et al., 2015), using deep learning model, particularly the Convolutional Neural Networks. Our model is highly motivated by the CNN architecture described in (Collobert et al., 2011).

2 Task Description

In general perspective, this task is a classification problem. In a global multilingual environment, the two main challenges for international companies (such as Microsoft) to automatically detect the meanings of customer feedback are: i) no widely acknowledged classes for understanding the meanings for customer feedback, ii) the classification may not be applicable in multiple languages. The participants were provided real world samples of customer feedback from Microsoft Office customers in four languages (namely English, French, Spanish and Japanese) and a five-plus-one-classes categorization (namely *comment, request, bug, complaint, meaningless and undetermined*) for customer feedback meaning classification. A participant has to develop a system that can predict the class(es) for customer feedback sentences across four defined languages.

3 Dataset and Resources

The organizer of the customer feedback analysis provided the participants customer feedback sentences which were collected from Microsoft Office customers as part of the joint ADAPT-Microsoft research project. The sentences were annotated with six classes: comment, request, bug, complaint, meaningless, and undetermined. The dataset was provided in four languages, namely English, French, Spanish and Japanese. Each sentence has at least one tag assigned to it and may be annotated with multiple tags. We did not use any external resources as additional data, i.e., we used only the dataset which was provided for this task.

4 Proposed Architecture

The proposed model is inspired by the CNN architecture described in (Collobert et al., 2011). The proposed model for the customer feedback categorization is shown in Figure 1.

Pre-processing: We pre-processed the dataset provided for building the system. Initially, we removed the index from the data samples. In the provided dataset, a number of data samples were tagged with multiple classes. If a data sample was tagged with multiple tags, the data sample was modified. The following example shows the processing of a data sample tagged with two tags for training as a single tagged.
Original:

Renovations underway...dated. ⇒ comment, complaint

After pre-processing:

Renovations underway...is dated. ⇒ comment

Renovations underway...is dated. ⇒ complaint

Embedding layer: Instead of using any pre-trained word embedding scheme, we have built a vocabulary table which is learnt from training data. The embedding layer works as a lookup table which maps vocabulary word indices into low-dimensional vector representations. The proposed architecture works for all the languages except Japanese because the Japanese sentences were not segmented.

Convolutional layer: This layer is the heart of the architecture. This layer made up of three 1D-Convolution layers. Each 1D-Convolution layer has kernel size equal to 3 and feature map equal to 128. During convolution operation filters are applied to extract the features. Then, max-pooling operation is applied to the feature map s to obtain the maximum value $s' = \max\{s\}$ for a particular filter. The objective of the max pooling is to capture the most important feature with the highest value for each feature map. Thus, one feature is extracted from one filter. However, the proposed architecture uses multiple filters with varying window sizes to obtain multiple features. We used the ReLU (Nair and Hinton, 2010) as our nonlinear activation function. .

Fully Connected layer: The max-pooling operation selects the best features from each convolutional kernel. Thus, all the resulting features which are selected from the max-pooling are combining in the fully-connected layer. The output of fully connected layer is passed to the output layer.

Output layer: The final layer (i.e., the output layer) made of 6 neurons as the customer feedback has 6 target classes. The output layer uses softmax as the nonlinear activation function.

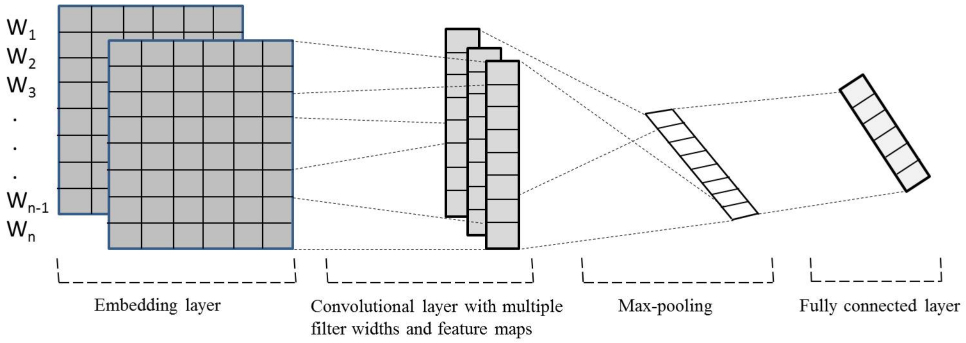
W1
W2
W3
.
.
.
Wn-1
Wn
Embedding layer
Convolutional layer with multiple filter widths and feature maps
Max-pooling
Fully connected layer

Figure 1: Model architecture

5 Results and Discussion

We have submitted the runs for English and French customer feedback. For English feedback, three runs were submitted. However, four runs were submitted for the customer feedback in French. In this shared task, for overall performance accuracy, macro-averaging and micro-averaging was used. The class specific performance was measured using precision, recall and F1-measure. For English, the best run achieved 0.388 in terms of accuracy. The evaluation is shown in Table 5. It is observed from Table 5 that the proposed model identified the 'comment' class more effectively than other classes. For English, the model did not identified the 'bug' and 'request' classes.

Overall performance			
Accuracy	0.388		
Macro-Avg	0.245895	0.204983	0.223583
Micro-Avg	0.427746	0.427746	0.427746
Class specific performance			
Tag	**Precision**	**Recall**	**F1-score**
comment	0.5549	0.6211	0.5861
complaint	0.2901	0.2621	0.2754
bug	0.0000	0.0000	NA
meaningless	0.1304	0.0968	0.1111
request	0.0000	0.0000	NA
undetermined	0.5000	0.2500	0.3333

Table 1: English Customer Feedback Evaluation

For French, we submitted 4 runs and the best run achieved 0.6675 in terms of accuracy. The macro-average and micro-average values were also satisfactory. The evaluation is shown in Table 2. It is observed from Table 2 that the proposed model performed the best for the 'complaint' class (F1-score: 0.8526). However, the model performed well on 'comment' (F1-score: 0.5455) and 'bug' (F1-score: 0.4717) classes. The model did not identified the 'meaningless', 'request' and 'undetermined' classes.

Overall performance			
Accuracy	0.6675		
Macro-Avg	0.303275	0.330574	0.316337
Micro-Avg	0.709443	0.697619	0.703481
Class specific performance			
Tag	**Precision**	**Recall**	**F1-score**
comment	0.5745	0.5192	0.5455
complaint	0.8664	0.8392	0.8526
bug	0.3788	0.6250	0.4717
meaningless	NA	NA	NA
request	0.0000	0.0000	NA
undetermined	0.0000	0.0000	NA

Table 2: French Customer Feedback Evaluation

During data pre-processing, we made the multiple tag data samples into single tagged. Therefore, for the same feedback text the system trained with two different tags. This created the ambiguity for the system. We believed that this ambiguity is the main reason for the decrease in performance of the proposed system.

6 Conclusions

We present this paper as part of our participation in the Customer Feedback Analysis shared task at IJCNLP. We proposed a CNN based deep learning framework for English and French. However, the proposed model performed well on French data than English data. Our embedding scheme did not work on Japanese as the sentences were not segmented. In future, we will employ Recurrent Neural Network to tackle the customer feedback analysis.

Acknowledgments

We would like to thank the organizers for organizing a wonderful research problem and giving opportunities for researchers to participate in it.

References

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al. 2016. Ranking relevance in yahoo search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 323–332. ACM.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.