Sentiment Analysis on Shenzhen Fintech News

Chang Liyan
Decision Analytics

Summer Research Fellowship 2020 for science Students

Poster No.:B1
Name: Chang Liyan
UID: 3035534880
Major: Decision Analytics

Introduction

FinTech is one of the main keywords in the development of the Greater Bay Area in the past three years. It applies novel technologies, such as mobile payment, blockchain and virtual currency, to improve financial activities. A sentiment analysis, based on 1-dimentional convolutional neuron networks (**1DCNN**), was performed on Shenzhen FinTech news to identify mass media's perspective on FinTech, whether the outlook for FinTech is promising or worrying. The results of this analysis can serve as references for future development in the Greater Bay Area.

Methodology

In this research, 1DCNN was applied to conduct sentiment analysis. Unlike normal CNN, the filters only contain one dimension. The hyperparameters mainly contains the number of filters and the filter sizes. Kim (2014) proved that for sentence-level classification tasks, a two-layer CNN model was adequate to perform well.

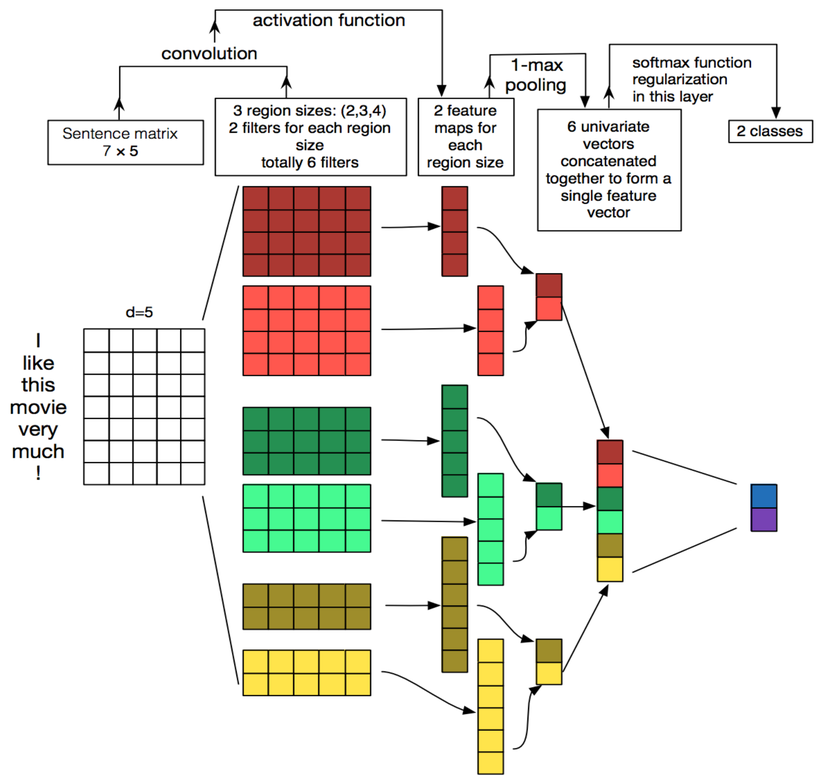

Fig.1 1DCNN Model architecture for an example sentence.

Data Preprocessing

We collected news scripts by crawling Factiva. After filtering a list of keywords, we obtained 12000 pieces of Shenzhen Fintech news from 2017 to mid-2020. Figure 2 gives a sketch of the preprocess pipeline.

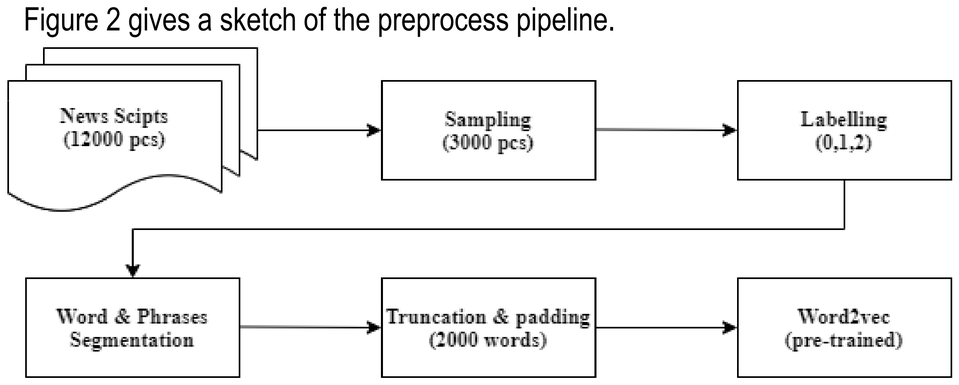

Fig.2 Data preprocessing pipeline

1. **Sampling & Labelling:** We first manually labelled the random sample of 3000 pieces. The news fell into three categories.
   (0-Non Shenzhen Fintech related; 1-Positive; 2-Negative).
2. **Words Segmentation:** The second step was to cut the news scripts into words and phrases, through 'jieba' python library.

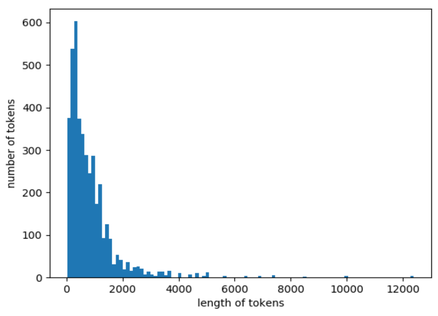

Fig.3 The distribution of script length

3. **Truncating & Padding:** The third step was to truncate or pad the script, according to a specific length.
4. **Word-to-vector Transformation**: Either random initialization or pre-trained transformation. Here we utilized a pre-trained model. These vectors were trained by Li et al. (2018) on 2.7 million words of Chinese financial news [1].

Model Design and Experiments

After preprocessing the scripts, we took 600 pieces of news from 2020 as the test set, and the remaining 2400 pieces before 2020 as the training set. From figure 4, we see that label 0 and label 1 samples are comparable. But label 2 sample only accounts for 16% of the data. Since the size of sample is small, we oversampled label 2 to a comparable level to label 1 and 2.

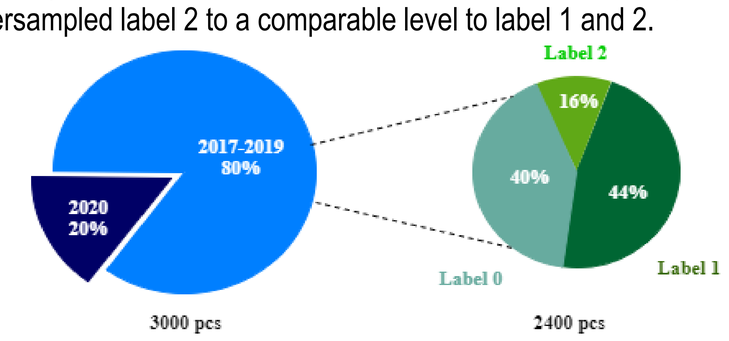

Fig.4 Pie chart of training and test samples.

Now, we consider the architecture of the 1DCNN model. This is a multi-label and paragraph-level classification task. Here, we used cross validation to select the best combination of two parameters, i.e. number of filters and filter sizes. It turned out that the accuracy can achieve **96.18%**, given filters with size 3, 4, 5 and 10, 64 each. This high accuracy can be attributed to the local and global interpretability of this model.

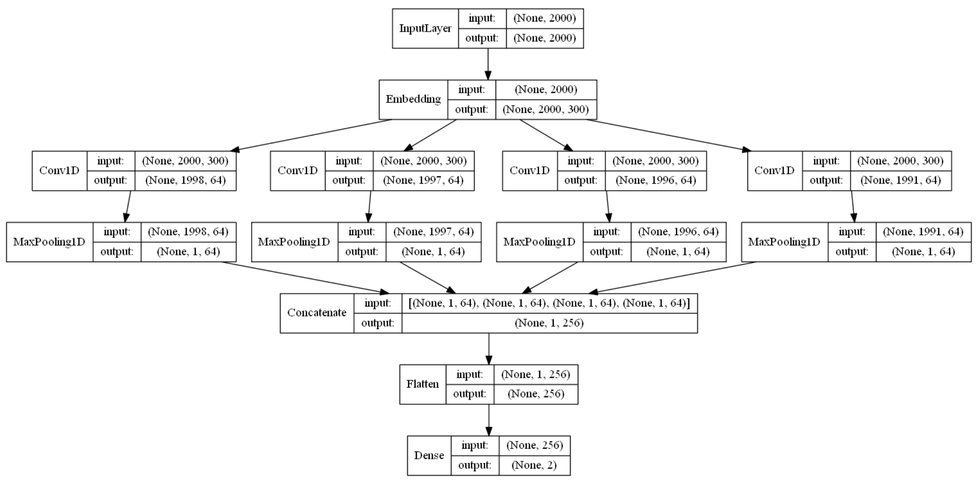

Fig.5 The architecture of 1DCNN model.

Furthermore, notice that there exists a hierarchy in the labels. We may classify whether the news is about Shenzhen Fintech or not. Then, we feed the samples that are predicted label 0 in another CNN model to classify the sentiment. This becomes an ensembled 1DCNN model. Mathematically speaking, the accuracy can be boosted if two models are moderately accurate. The accuracy of the ensembled model is raised up to **98%**. Although the accuracy is improved, from the resulting confusion matrices of two models in Table 1, the performances are not significantly different.

Table 1. Comparison of confusion matrices of two models

True\Predict	1DCNN			Ensembled		
	0	1	2	0	1	2
0	368	6	1	375	3	0
1	11	152	0	13	152	0
2	1	0	59	3	0	61

Conclusion

The proposed 1DCNN model can greatly classify the sentiment of Shenzhen Fintech news. An ensembled model may have a more accurate result. This research can serve as references for future Fintech development in the Greater Bay Area. For instance, we can examine the weights in the CNN model and check which signal determines the negative sentiment. In this way, we can further obtain the keywords for the downside of Fintech and these are the future development focus.

Reference

1. Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical Reasoning on Chinese Morphological and Semantic Relations. *arXiv preprint arXiv:1805.06504.*
2. Kim, Y.(2014). Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1408.5882*.