Attention-Based Sentiment Analysis on Movie Reviews

Li-Hen Chen, Yun He, Qifan Li, Fan Yang, Yining Zhou

Texas A&M University

Abstract

In the web era, sentiment classification has been becoming a significant problem nowadays, due to the tons of online available textual information. To better facilitate human in doing online textual analysis for data mining, machine learning and deep learning techniques have been widely utilized in real-world applications. However, most of those applied learning systems typically lacks interpretability, which largely hinders developers to know whether their deployed models are reasonable or not.

In this project, We classify the reviews from movie websites into positive and negative comments using different machine learning models as well as deep learning models, including logistic regression, linear/non-linear SVM, Random Forest, CNN , bidirectional GRU, etc. We get the dataset from SST-2 and we use pre-trained word-embedding tools "Glove" for obtaining embedding vector representations for words. Besides, we also provide interpretations to the sentiment classification task with attention-based method. The results turn out that Bi-RNN gives the best accuracy performance.

Attention Methodology

In this project, we employ the attention-based bidirectional recurrent neural network (RNN) to conduct the sentiment classification task, where the attention weights of words are used to deliver the explanations of classification results. The specific architecture of the system is shown in Figure 1.

The designed attention layer is built on the top of the recurrent layer and used to weigh the contributions of each hidden states of RNN. The particular update equation can be expressed as follows:

$$M = \tanh(H)$$

$$\alpha = softmax(w^T M)$$

$$r = H\alpha^T$$

where ***H*** is the RNN output, ***w*** is the trained weights, ***α*** is the attention vector.

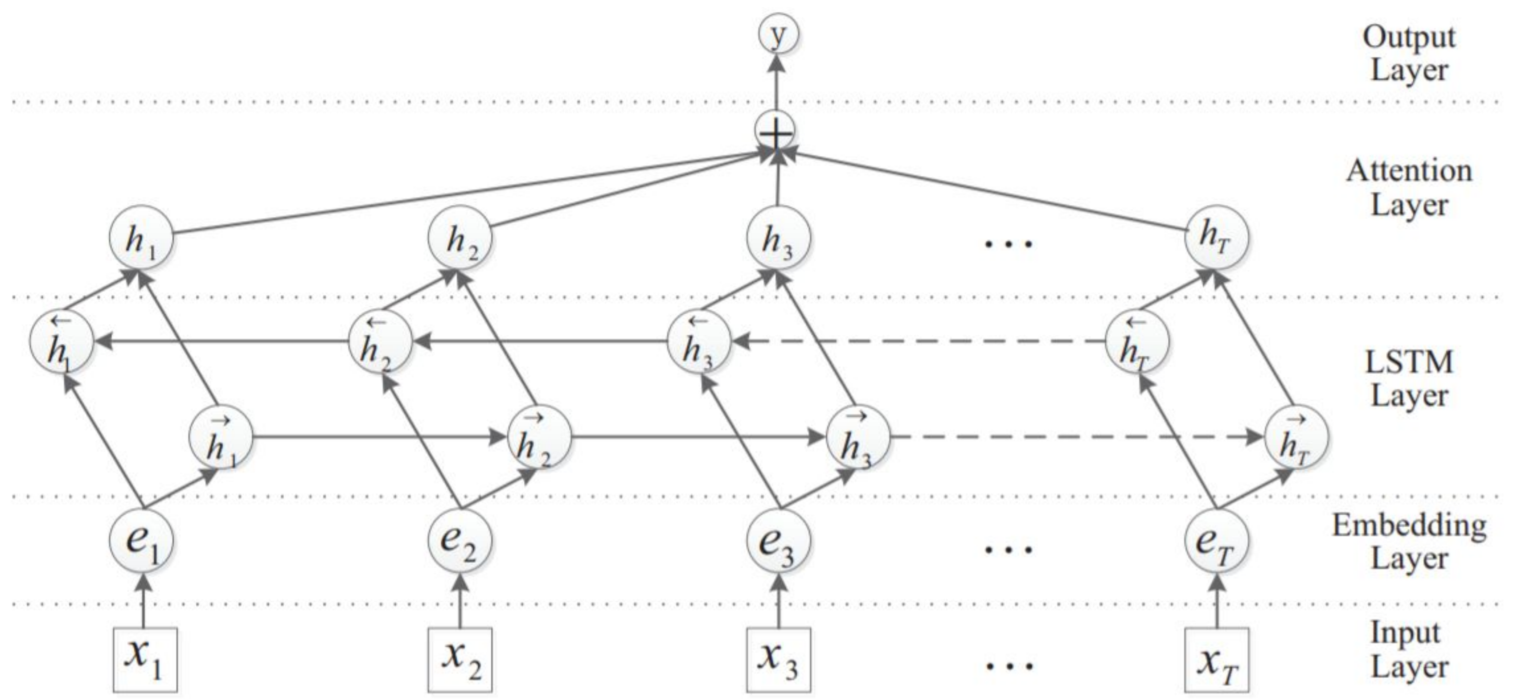

Figure 1. Attention-based bidirectional recurrent neural network architecture.

Results

In the figure 2 that most of the words are below 10,000 on both X-axis and Y-axis, and we cannot see meaningful relations between negative and positive frequency. However, if we combine harmonic mean of rate CDF and frequency CDF to interpret sentiment data, it has created an interesting pattern on the plot. If a data point is near to the upper left corner, it is more positive, and if it is closer to the bottom right corner, it is more negative. With the figure, we can see what token each data point represents by hovering over the points. Not every point has the correlation as we expect, we believe that's the reason why our accuracy is around 85% eventually.

The pre-trained model we used is Glove300d, which is an unsupervised learning algorithm for obtaining vector representations for words. 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. The results are shown in the below table. The models we train include some deep learning models such as bidirectional GRU + attention, bidirectional RNN, CNN, and NN and some baseline models such as SVM, Random Forest and Logistic Regression.

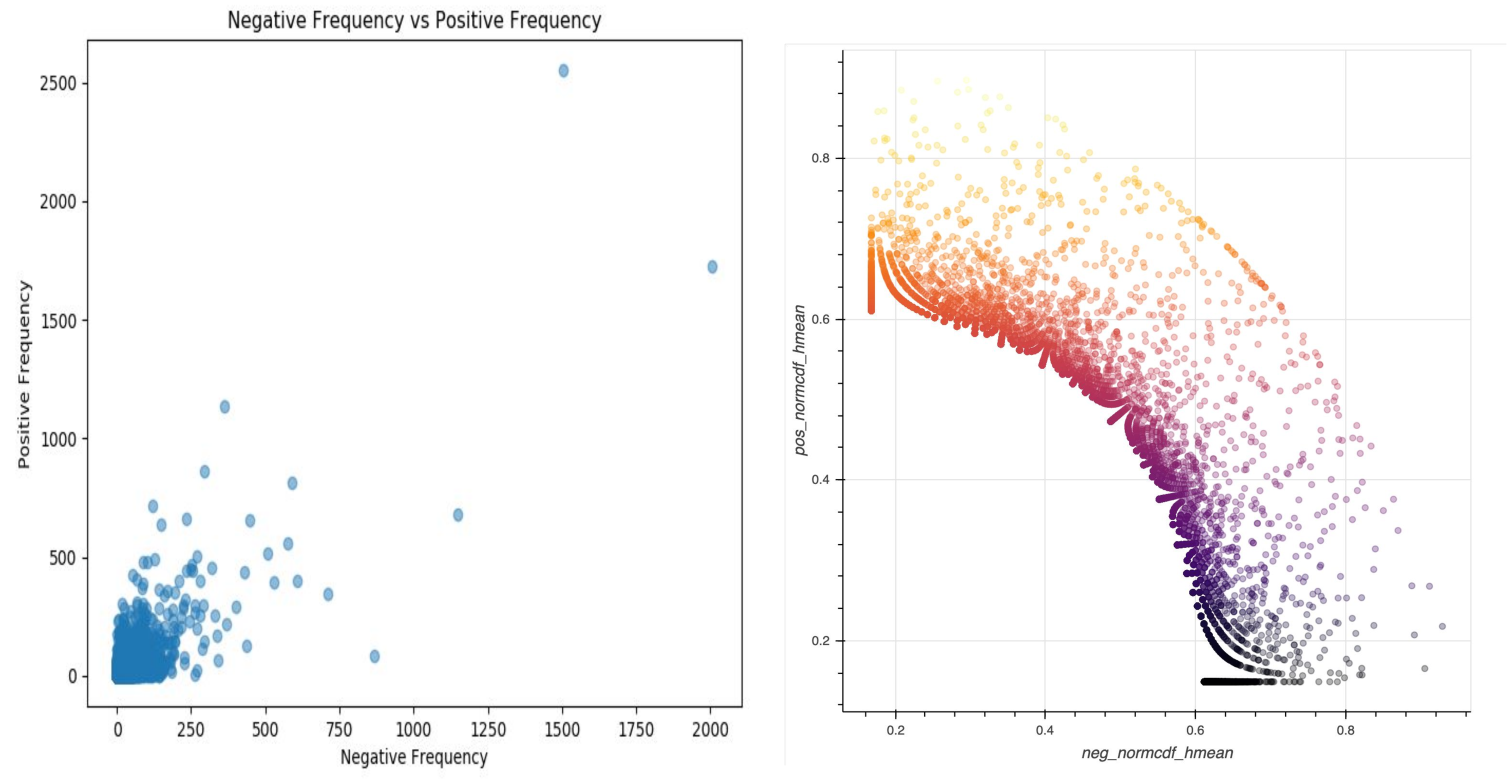

Figure 2. Correlation of words frequency (left) Frequency CDF Harmonic Mean(right)

Table 1. Training Results

Model	bidirectional GRU + attention	bidirectional RNN	CNN	NN
Accuracy	86.35%	86.04%	84.65%	80.53%
Model	**Linear SVM**	**Non-linear SVM**	**Random Forest**	**Logistic Regression**
Accuracy	67.43%	69.15%	66.83%	68.23%

Case Study

In this section, we provide a case study of our model, where the word tokens and their responding attention scores are shown below. The darker the color is, the higher the attention score and more important of the token is.

Positive reviews:

0.02013708	0.01178347	0.01283708	0.03860016	0.029	0.02718525	0.04630307	0.04459512
it	s	a	charming	and	often	affecting	journey

0.00976174	0.02263067	0.00873657	0.0125226	0.01936398	0.012414727	0.01000419	0.01270897	0.02105907	0.01785653
it	provides	the	grand	intelligent	entertainment	of	a	superior	cast
0.02827017	0.02795154	0.02159456	0.03524882	0.04450357	0.049330834	0.02783003			
playing	smart	people	amid	a	compelling	plot			

Negative reviews:

0.04870348	0.07349443	0.01424661	0.00288323	0.00236665	0.002737909	0.00292143	0.00482289	0.01049423	0.00379167
it	does	nt	believe	in	itself	it	has	no	sense
0.00348484	0.00428188	0.00600788	0.00850687	0.01453435	0.023409057	0.06658567			
of	humor	it	s	just	plain	bored			

0.03314108	0.03910998	0.03396684	0.0896245	0.06989624	0.032917414	0.0192147	0.04619217
a	string	of	rehashed	sight	gags	based	in
0.06085098	0.03581252						
insipid	vulgarity						

Based on the cases, we observe that our model is able to attend key words for sentiment classification problem. For example, "charming", "compelling" are strong signals for positive sentiment and obtain higher attention scores from our model. Thus, our model can unexplainably classify the sentiment polarity of movie reviews.

Conclusions

In this project, we implement an interpretable deep model for conventional sentiment classification task, using the attention-based bidirectional RNN architecture. With the aid of the attention weights corresponding to each word, end-users could know how the trained model makes the classification beyond the prediction results. By conducting a series of experiments, we observe that our interpretable attention-based model could achieve a competitive performance among all other baselines, including other non-interpretable deep models. Meanwhile, our model could generate explanations for each new coming instance (movie review), so as to help users understand of contribution of each word for sentiment classification. Besides, with some case studies, we found that our generated explanations is sensible, which effectively captures the important words for sentiment analysis.

Contact

Li-Hen Chen (UIN:928003907), Yun He (UIN:326005850), Qifan Li (UIN:127004551), Fan Yang (UIN:525004621), Yining Zhou (UIN:927009507)

Texas A&M University

Email: leo0215667@gmail.com, yunhe@tamu.edu, excaliburea6@tamu.edu, nacoyang@tamu.edu, zynzyn135@tamu.edu

References

1. Kim, Y. (2014). Convolutional neural networks for sentence classification.
2. Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification.
3. Zhou, Peng, et al. "Attention-based bidirectional long short-term memory networks for relation classification." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2016.
4.
5.