Consumer Finance Complaint Product Classification

Hong Yang, Haifeng Lin
December 2019

Abstract

More and more consumers voice their opinions of products and services via online platforms and expect fast response from companies. This phenomenon creates both opportunities and challenges for many companies. By understanding consumers' complaints, businesses can get product feedback quickly, resolve complaints properly, improve customer satisfaction, and ultimately deliver higher quality of products and services than peers can. This research applied a variety of NLP techniques (LSTM, Attention, and RNN) to automatically identify products from consumer complaint narratives. Findings from this research have implications for businesses as well as for other areas, such as education institution.

Introduction

Our goal for this project was to build a language model that, given a consumer complaint narrative, automatically predict which product the complaint is about. Given the dataset we used in this project, this project focuses entirely on financial products and services. But such language model can be widely used in real world where businesses want to know more about complaint or review written by their customers. By automating the prediction and building capacity of processing large amount of data points, companies can see the trending view of which their products have been complained the most and monitor product feedback on real-time. With this insight, businesses can better understand the performance of their products and quickly react to the complaint or feedback by making more pointedly change to their products.

Text classification It has broad applications including topic labeling, sentiment classification, and spam detection (Yang et al., 2016). So if we can utilize NLP techniques to classify new complaints with high accuracy and route them to the right team to resolve the issue, that will be a win and time saving to any business. Benefiting from its recurrent structure, RNN, as an alternative type of neural networks, is very suitable to process the variable-length text (Zhou et al., 2016). In this research, we used a variety of NLP techniques such as LSTM, Attention, and RNN.

Background

With services-businesses going online and due to enormous scale, it has been a lot easier for consumers to vent out the anger online as well as publicizing the issue in expectancy of quick actions. A customer complaint highlights a problem, whether that's a problem with product, employees or internal processes. Research finds that customers' whose complaints are handled quickly can often turn into loyal customers and even brand advocates. A study by Harvard Business Review found that customers who have a complaint but received quick response likely spend more on future purchases than other customers (Huang et al., 2018).

Hence, it has become imperative for businesses, such as banks and other financial institutions, to be able to know what products or services are being mentioned in customer complaints. In the past, performing this kind of task was done manually by employees. Of course, the human approach is very costly to the business and often delays swift response to the complaints received.

There have been a number of articles that describe various approaches to this problem. A 2018 article by Ashwin K, who is a data scientist, tackled exactly same issue by using Logistic Regression mode, Random Forest, SVM, GBM, Neural Networks and utilize hyperparameter tuning (Ashwin, 2018). The author was able to get high accuracy, around 90%, for some products but low accuracy for other products. Another article by Susan Li, another data scientist, went after the same problem. She used LSTM and achieved 82% overall accuracy (Li, 2018).

Approach

Objective

The objective of this project is to classify consumer complaint text into product categories. We are not attempting to perform sentiment analysis of such complaints as most of the complaints would be highly negative anyways.

Data

We used the Consumer Complaint Database that was published by the Consumer Financial Protection Bureau. The Consumer Complaint Database is a collection of complaints about consumer financial products and services that the bureau sent to companies for response. Complaints are published after the company responds, confirming a commercial relationship with the consumer, or after 15 days, whichever comes first [6].

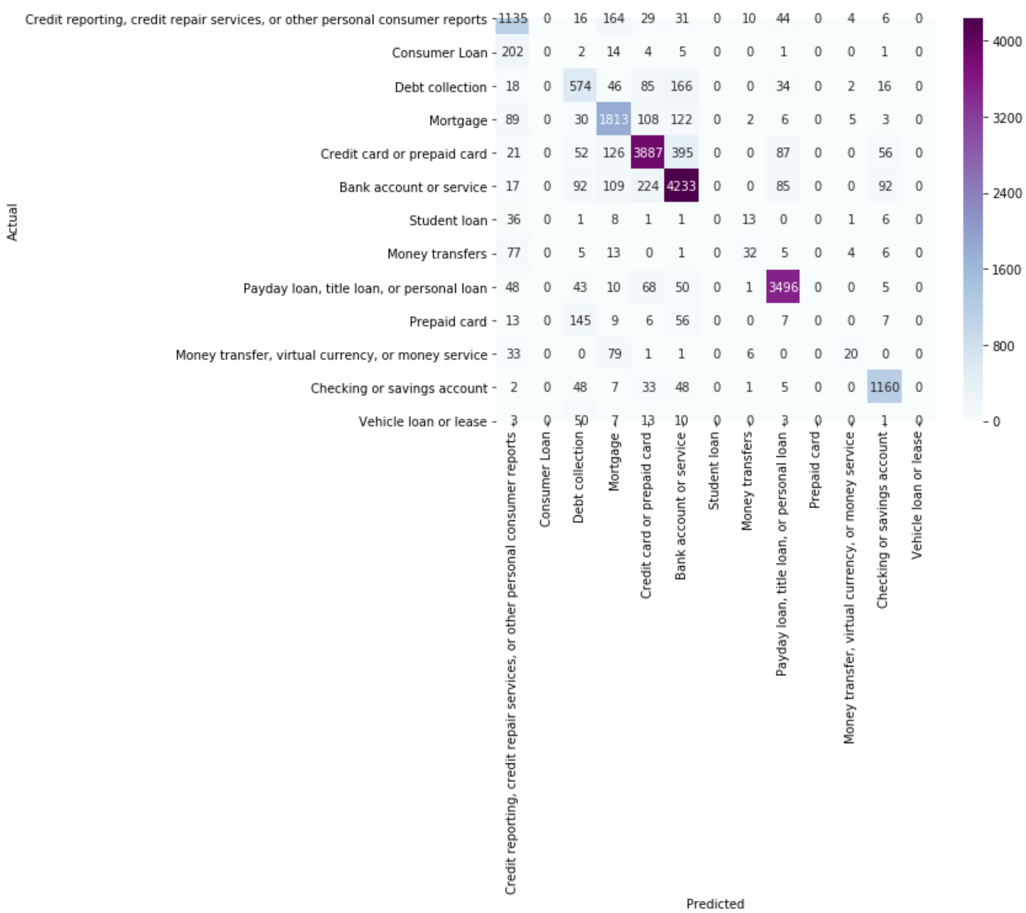
Baseline Model and Results

<u>LSTM model</u>

The baseline provided an example of how a Recurrent Neural Network (RNN) using the Long Short Term Memory (LSTM) architecture can be used for this task.

Despite the overall decent accuracy of 82%, this model has some shortcoming flaws. For example, product categories that have few data points, such as Student loan, Vehicle loan or lease, don’t produce any correct prediction. This model does well, in terms of prediction accuracy, on product categories that have massive amount of data points but terribly on those with relatively fewer data points. Hence, the Macro Average Precision from this model is only 50%.

Confusion Matrix for baseline model



Our Models and Results

We then decided to use Attention with the goals of improving the overall accuracy as well as the accuracy among products. The reason is that in any particular complaint, a lot of words are used to describe the background or the consumer’s emotions. So paying attention to some input words within complaint narratives could improve the LSTM model. We attempted two other models: Attention sequence based classification model and LSTM with Attention.

Model 1: Attention sequence based classification model

We leveraged the Position_Embedding and Attention classes from this open source: https://github.com/foamliu/Self-Attention-Keras/blob/master/attention.py

We used the 10 vectors embedding 13 categories dense layer for output as seen in LSTM model. Additionally, we replaced LSTM layer with attention as Multi Headed attention model that used the embedding directectly to generate keys, value and queries. The final prediction is based on the pooled attention results.

Although we only see modest improvement on the overall accuracy (83%) compared to the baseline model - LSTM model, the accuracy of product categories with few data points has been improved significantly. For example, this model picked up predictions for product like “Student Loan” or “Prepaid Card”, which have 0 prediction from the baseline model. Additionally, the Macro Average Precision is 74%. Both metrics shows that Model 1 did make good improvements over baseline model.

However, few product categories, such as “Vehicle loan or lease” or “Consumer loan”, still have very poor accuracy. So we decided to combine LSTM and Attention and see if we can make more improvement.

	precision	recall	f1-score	support
Credit reporting, credit repair services, or other personal consumer reports	0.69	0.80	0.74	1439
Consumer Loan	1.00	0.00	0.01	229
Debt collection	0.65	0.61	0.63	941
Mortgage	0.80	0.81	0.81	2178
Credit card or prepaid card	0.86	0.86	0.86	4624
Bank account or service	0.82	0.87	0.85	4852
Student loan	0.94	0.24	0.38	67
Money transfers	0.68	0.43	0.52	143
Payday loan, title loan, or personal loan	0.92	0.95	0.94	3721
Prepaid card	0.61	0.26	0.36	243
Money transfer, virtual currency, or money service	0.75	0.52	0.62	140
Checking or savings account	0.88	0.89	0.88	1304
Vehicle loan or lease	0.00	0.00	0.00	87
accuracy			0.83	19968
macro avg	0.74	0.56	0.58	19968
weighted avg	0.83	0.83	0.82	19968

Model 2: LSTM with Attention

We used the 10 vector embedding 13 category dense layer for output as seen in the Model 1 and LSTM on embedding input, then passed results to the multi headed attention layer. All keys, values and queries are based on LSTM.

Overall results from Model 2 are similar to these from the LSTM model except for Macro Average Precision, which has modest improvement. We think the reason why this Model 2 does not perform better than the other 2 models is that the additional complexity in Model 2 overfitted the model, or it needs more dataset to train. As far as the issue of producing no prediction for product categories such as “Student Loan” or “Prepaid Card”, the Model 2 does not show any improvement on that area.

	precision	recall	f1-score	support
Credit reporting, credit repair services, or other personal consumer reports	0.66	0.80	0.72	1439
Consumer Loan	0.00	0.00	0.00	229
Debt collection	0.63	0.57	0.60	941
Mortgage	0.76	0.82	0.79	2178
Credit card or prepaid card	0.78	0.92	0.84	4624
Bank account or service	0.89	0.78	0.83	4852
Student loan	0.00	0.00	0.00	67
Money transfers	0.50	0.38	0.44	143
Payday loan, title loan, or personal loan	0.92	0.94	0.93	3721
Prepaid card	0.57	0.16	0.26	243
Money transfer, virtual currency, or money service	0.71	0.34	0.46	140
Checking or savings account	0.90	0.86	0.88	1304
Vehicle loan or lease	0.00	0.00	0.00	87
accuracy			0.82	19968
macro avg	0.56	0.51	0.52	19968
weighted avg	0.80	0.82	0.80	19968

Both models achieved similar overall accuracy but different level of accuracy among products. Model 2 seems to produce high level of precision from highly populated classes but low level of precision from less populated classes. In contrast, the results from Model 1 are more balanced across the broader set of classes. Overall, we think that Model 1 is a more mature model than Model 2 because it is further away from the trivial approach of assigning everything to the most populated class.

Conclusions

Our struggle to improve prediction accuracy for product categories that have few data points despite applying complex methods reflects the fundamental challenge of how to train NLP, or even machine learning in general, models with insufficient data. Fortunately, data have been created in an unprecedented pace. Additionally, the pace will only go faster. Although we have theory behind why Model 2 does not produce better results, we think that we could spend more time to understand whether our theory is correct and to figure out how we change improve the model for better results.

There are several other NLP models we could also try in this task. One of them is to use multi-headed attention for more than a single layer. We have shown that by adding attention can improve accuracy. With multi-header attention, we think we may see some improvement on overall accuracy. We also think BERT is a good choice because BERT has achieved amazing results in many natural language understanding (NLU) Tasks (Sun et al., 2019). Kaushal Trivedi, who works in AI & Machine Learning field, used BERT for a multi-label text classification Kaggle project and produced one of the best scores (Trivedi, 2018).

Python Code:
https://github.com/hong6us/W266_Fall2019_FinalProject/blob/master/Customer_Complaints_Analysis.ipynb

References

1. **Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016.** Hierarchical attention networks for document classification**. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.** **https://www.aclweb.org/anthology/N16-1174.pdf**
2. **Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016.** Text classification improved by integrating bidirectional lstm with twodimensional max pooling**. In COLING.** **https://arxiv.org/pdf/1611.06639.pdf**
3. **Wayne Huang, John Mitchell, Carmel Dibner, Andrea Ruttenberg, and Audrey Tripp. 2018:** How Customer Service Can Turn Angry Customers into Loyal Ones https://hbr.org/2018/01/how-customer-service-can-turn-angry-customers-into-loyal-ones
4. **Ashwin K. 2018:** NLP - Consumer Complaints Classification using Machine learning and Deep Learning **https://ashwin-ks.github.io/2018-08-15-NLP-Consumer-Complaints-Classification-ML-DL/**
5. **Susan Li:** Multi-Class Text Classification with LSTM **https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17**
6. **Consumer Complaint Database** **https://www.consumerfinance.gov/data-research/consumer-complaints/**
7. **Kaushal Trivedi:** Multi-label Text Classification using BERT – The Mighty Transformer https://medium.com/huggingface/multi-label-text-classification-using-bert-the-mighty-transformer-69714fa3fb3d
8. **Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang. 2019**: How to Fine-Tune BERT for Text Classification? https://arxiv.org/pdf/1905.05583.pdf