Sentimental Analysis with Amazon Review Data

Mingxiang Chen
Stanford University
450 Serra Mall, Stanford, CA 94305
ming1993@stanford.edu

Yi Sun
Stanford University
450 Serra Mall
ysun4@stanford.edu

1. Abstract

Analyzing and predicting consumers behavior has always been a blooming and promising area of study with great value of research. Given the existing methods in the field, sentimental analysis of texts could be feasible especially in regarding to E-commerce. By analyzing the polarity of the text, decision maker can effectively inspect the strengths and weaknesses of their products, or even anticipate the complaints and the sales amount. In this project, we used the Amazon review dataset and try to predict reviewers' rating based on their review texts. With the help of the LSTM RNN model, we reached a final accuracy of 46.66%.

2. Introduction

Sentimental analysis often refers to using a combination of techniques like natural language processing and text analysis to identify positive or negative opinion, emotions or evaluations accurately. It is a relatively hot topic and widely applied in areas relating to human interaction such as customer review or comment. There are significant research and business values hidden in the sentimental analysis as this technique allow us to quantize subjective information for further investigation. For example, by analyzing the polarity of the text, decision maker can effectively inspect the strengths and weaknesses of their products, or even anticipate the complaints and the sales amount. Due to its vast potential, there are lots of researchers applying or studying the topic of sentimental analysis.

Figure 1 shows that the number of studies working on the issue of sentimental analysis. We can see that the number is steadily increasing but showing sign of decreases. The trend indicates the sentimental analysis is becoming quite mature and has entered a practical stage. There has been plenty of traditional methods and models built in sentimental analysis such as SVM and KNN. With the development of deep learning, modern deep learning techniques have also been introduced in classifying positive and negative emotions.

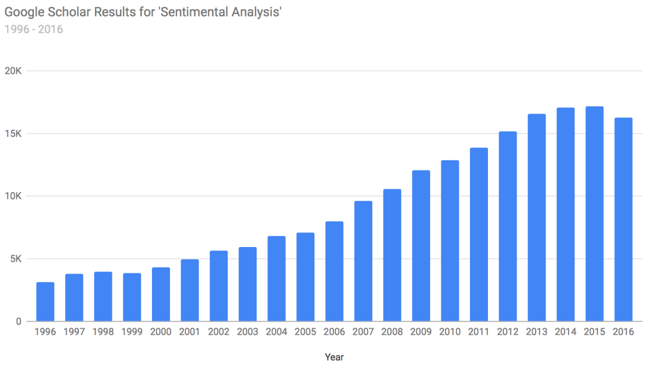

Figure 1. Research Papers relating to 'Sentimental Analysis'

3. Problem Statement

To fully implement our course material and fulfill our interest, our goal is to compare different classification method in sentimental analysis on Amazon dataset to see if it works better in some particular aspects. Though we suspect modern method like deep learning can achieve higher accuracy, traditional methods would be less data-hungry and might have similar performance in some settings.

The dataset[1] we are working on is Amazon product review data. It contains product reviews and meta-data from Amazon, including 142.8 million reviews spanning May 1996 July 2014. We would pick only 8.9 million reviews on the subset of books sold on Amazon not only because it is one of the representative services provided by Amazon but also the data size of it is reasonable for a single machine to train or evaluate with efficiency. The variable we trained on is the review by customers with labels overall comment from 1 to 5. The model can take in a new customer review and predict an overall score from 1 to 5.

4. Related Work

The sentimental analysis could be hugely instrumental for us to get an overview of a paragraph. For example, it is widely used in social media monitoring. Many researchers use this as a tool to investigate problems in the field of social science, such as predict stock price [2], and political preference [3]. Some even use it as a tool to predict president election [4].

By far, a very popular corpus to do the work is Twitter [5]. Since microblogging today has become a very popular communication tool among Internet users. The big data from microblogging reflects up-to-date states of bloggers. Amazon reviews, on the other hand, though not as timely as Twitter and Facebook, can also be a good resource of microblogging, which indicates consumers' attitudes toward each product [6].

Pang et al.[7] is relatively early attempt to classify the document with sentiment instead of topics. They utilized the simple bag-of-words method with machine learning techniques, which are Naive Bayesian and SVM. The accuracy achieves by sentimental based classification is worse than the one based on topic based classification. The difference in performance showed the difficulty in predicting sentiment based on the text. However, their work only implemented traditional methods where we can use deep learning technique which had achieved state-of-art accuracy. On the other hand, we try to predict in categories of five (overall view from 1 to 5) instead of two categories (positive or negative), which might hurt our accuracy.

5. Experiment

5.1. Binary Classification

A sentence can be either happy or not happy. Before figuring out how 'good' or how 'bad' the reviews are, we would firstly like to know "is it good?" and "is it bad?", which give rise to the binary classification model.

So far, there are already a bunch of natural language processing tool kits on the Internet. Here, we are comparing our Naive Bayes model with the one called Textblob. The Naive Bayes model is calculated from the probabilities of how likely a single word can lead to a good review (4-star or 5-star). Then multiply the probabilities for all the words in the sentence, and we can get the probability of happy for the sentence.

For example, we have a review written as "good product", and the corresponding rate is 5. The word 'good' can be appeared in a phrase like 'good book', 'very good', or it can be 'it is not very good.' Let's say 'good' appears in 1000 reviews, and 700 of them are good reviews (4 or 5 stars). And the word 'product' also appears in 1000 reviews, but only 400 of which are good. Then the probability of good review for this piece of text is

$$\frac{0.7 * 0.4}{0.7 * 0.4 + 0.3 * 0.6} = \frac{0.28}{0.46} = 60.87\%$$

Since it is above than 50%, it would be a good review. The test set is 100 thousand book reviews. Since the Naive Bayes is trained on a similar dataset, it achieves a higher accuracy at 78.78%, where the Textblob model achieves an accuracy at 63.34

5.2. Multi-category Classification

The very first thing to do when starting the multi-category classification is extracting a vector from a paragraph, where in our model, we call it a review vector. There are many ways to do this, and we start from the very straightforward one, where we use the mean vector of the word vectors for the corresponding words in a review. The word vectors in this section is the 100 dimension pre-trained word vectors (https://nlp.stanford.edu/projects/glove/).

However, the dataset itself is biased. The ratio of 1-star, 2-star, 3-star, 4-star, and 5-star reviews is around 1:1:2.5:5:10. Most of the reviewers give five stars to the product (note that this is only the dataset of books, which may not hold true in other categories). Hence, we deleted many of the data, to make sure the number of reviews in each class is the same. The final dataset contains 1,546,730 reviews, with 309,346 in each class.

Our first model is a set of K-nearest neighbor (KNN) models, with k value equals to 1, 5, and 10. The x here is the review vectors, and the y here is the number of stars for each review. The test set is a subset of the review data with 46730 reviews and stars. The ratio of correct classification is 25.52%, 27,05%, and 27.92% for k = 1, 5, and 10 respectively.

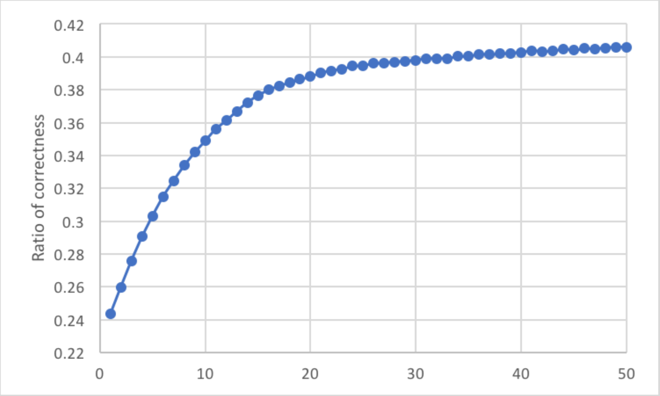

Figure 2. Training history for the simple deep learning model

The second model is a multi-class linear support vector machine, which implements the one-vs-the-rest multi-class strategy. So in this case, four models are generated. Much better than KNN, the linear SVM model achieved a ratio of the correctness of 37.94% on the test set, which is already close to the limit of using the mean vector as the sentence vector.

After learning deep learning in class, we realized that using deep learning may also be very helpful for classification.

So we implemented a model with 2 hidden layers with 200 units and the activation functions for each layer is Relu. The shape of the output layer is Batch size*5. Using the softmax function, we can get the probabilities. The loss function is the cross-entropy function. We used the ADAM optimizer to minimize the loss. Since the dataset is huge, we set a batch size of 100. And trained 5 epoches. In this example, the learning rate is 0.0002, and we set alpha = 0.99 as the decay factor for the learning rate. The final result slightly better than SVM, where the accuracy in the training set is 40.59% and 40.53% in the test set.

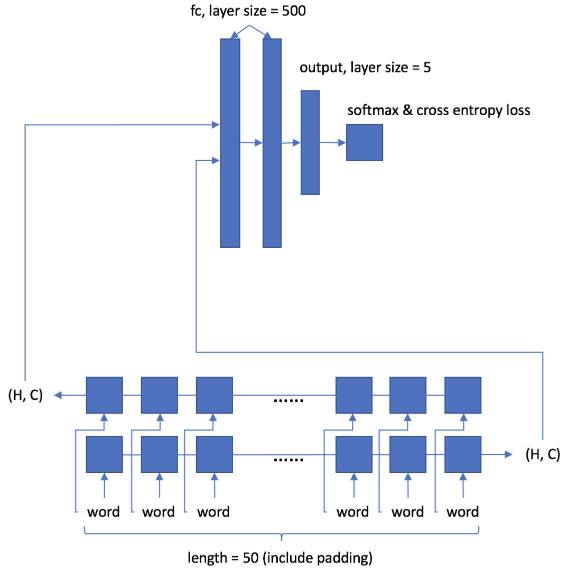

Figure 3. Model Framework

As mentioned above, 40% is almost the limit of using the mean vector as the sentence vector, so a more advanced method to extract sentence vector is using a recurrent neural network.

According to the fact that using an RNN models requires us to record every word vector (Actually not every. If the length of the sentence is lower than 50, we would add some paddings (all-zero vectors) to the end of the sentence; if above than 50, we would delete the rest.) in a sentence, the computation complexity and the requirement for memory is much more larger than a straightforward fully connected layer. Because of this, we only used a subset of 100,390 reviews with 20,078 reviews in each class (based on the number of stars).

The RNN model has two parts. The first part is a bidirectional LSTM layer. The hidden layer's dimension in the LSTM cell is 128. By combining the h state and the c state at the endpoints, we can get a review vector with dimension of 512. The second part is one fully connected layer which directly output the result. Similar to the deep learning model discussed above, we used softmax function for output, and an adam optimizer to minimize the cross-entropy loss.

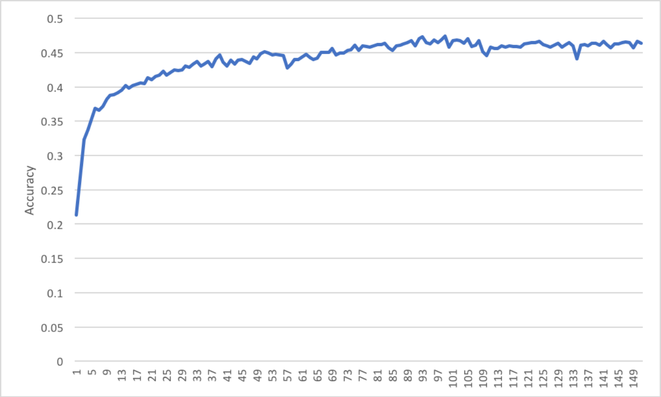

Figure 4. Training History for the RNN model

Since the model is much more complicated, it is natural that it took more time to train, and the result should be better. The whole training process (not include the data preprocessing) took 12 hours, with no GPU involved.

However, tunning the parameters in this model is much harder than straightforward fully connected layers. By far, the models shows a much better result than the simple neural network model and the SVM model, where the accuracy in the training set is 46.38% and 46.66% in the test set. The accuracy for the test set is even higher than the training set, which means the model is not overfitting the data.

6. Discussion

The following are some example of misclassification:

Have purchased many in the past like in the 1970's after I received one as a gift. I recommend to all ages. (Overall: 1 Prediction: 5)

In this example, I think the algorithm is giving us a good result. Maybe the reviewer want to express his sarcasm here, or maybe the reviewer do not know how to use the rating system.

A must in everyone's library for graceful poetry, inpirational reading and lessons to live by and to even to let go. A new printing would be a welcome as the üsedönes in good condition are quite expensive. (Overall: 4 Prediction: 2)

The word 'inspirational' is absolutely a good word, but the predictor didn't get it, because the reviewer spell it wrong. However, the algorithm do understand what is 'expensive' mean.

This work still holds up decades later. Every reflective, thoughtful person should have The Prophet in the home library. It can be read one chapter at a time. The chapter on love is as good as anything I've ever read. (Overall: 4

Prediction: 5)

For me, if someone says ’as good as anything I’ve ever done’, I would apparently regard this as a great compliment. The algorithm gives 5 which is reasonable. But speaking of accuracy, this won’t count.

7. Future Works

Sentimental analysis could be useful in many more aspects. It is a very good tool helping us to understand the trend in fashion, people’s opinions, what do people like and what do we hate without looking and searching for the information on the Internet by ourselves. However, a set of pre-trained word vectors has its own flaw. The meaning of the same word can be different under different situations. For instance, the word ”Trump” on wikipedia could be neutral, but on twitter, the meaning of ”Trump” is not only a name, but has its political meanings. So it could be interesting to see if we can reach a better result when training our own word vectors on the review dataset.

8. Contributions

Mingxiang Chen: Wrote the deep learning models (both the straight forward one and the LSTM models), the SVM, and KNN model. Searched related papers.

Yi Sun: Wrote the binary classification models. Helped to tune the parameters and train the LSTM model. Did the error analysis. Searched related papers.

We wrote the report together.

References

[1] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. pages 507–517, 2016.

[2] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. 1:492–499, 2010.

[3] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358, 2014.

[4] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[5] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. 10(2010), 2010.

[6] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. pages 79–86, 2002.