Twitter US Airline Recommendation Prediction

Xiaotong Duan, Tianshu Ji, Wanyi Qian

Abstract

The goals of this project are 1) build a model for six major U.S. airlines that performs sentiment analysis on customer reviews so that the airlines can have fast and concise feedback, 2) make recommendations on the most important aspect of services they could improve given customers' complains. In this project, we performed multi-class classification using Naive Bayes, SVM and Neural Network on the Twitter US Airline data set from Kaggle. Significant accuracy has achieved, which shows that our models are reliable for future prediction.

1 Introduction

In recent years, Twitter has become the de facto online customer service platform. Thus, a companys image on Twitter is of central importance and this is especially true for airlines given that many tweets are travel-related in nature. In fact, research has shown that responding to tweets has revenue generating potential, drives higher satisfaction than other customer service channels, and perhaps most importantly, satisfied Twitter users spread the word. In this project, we use tweets gathered from Twitter to learn about people's flight experiences and give airline companies suggestions on how to make their trip more enjoyable.

The data set contains about 15,000 tweets, collected from February 2015 on various airline reviews. Every review is labeled as either positive, negative or neutral. First, we want to build a model to perform sentiment analysis on the data set. Second, more interestingly, we want to assign a reason to each negative response, such as late flight, lost luggage, etc. In our data set, about 80% of the negative reviews has a negative reason label, yet the rests are labeled as "can't tell". Our goal is to assign a label to this unspecified group. By knowing every review's negative reason, we can give specific suggestions to different airline companies on how to improve their service.

2 Background and Related Work

Nowadays, developing and testing different models for a natural language processing problem is an interesting and challenging task. However, due to the nature of the problem, the accuracy of sentiment analysis on single sentence like movie reviews never reaches above 80% for the past 7 years [1]. Looking at last years project on twitter [2], their accuracy was 59.32% to 63.71%, depending on different models. In our project, we achieved near 20% more than their result, which is a significant improvement.

Since tweets texts are usually short and verbal, the same problem presents in our data set as well. However, even though the tweets are short, there are strong indicative words. Specific words can be used as indicators for spam/ham emails and achieve good test accuracy. Therefore, we believe that tweets review, without many negating negatives, can be predicted well using the frequency vector representation. To prove this, we will use Recurrent Neural Network model and the GloVe word vector [3] to compare the result.

3 Approach

3.1 Dataset

The sentiment analysis labels are positive(20%), negative(60%), and neutral(20%). The negative reason labels are bad flight(7.45%), canceled flight(9.62%), customer services issues(39.77%), damaged luggage(0.84%), flight attendant complaints(6.05%), flight booking problems(6.19%), late flights(1.99%), long lines(19.97%), and lost luggage(8.23%).

3.1.1 Preprocess of Dataset

In the preprocessing step, non-English word, symbols and website links are eliminated. Then the whole data set is randomly separated into training set (10000 samples, 70%) and test set (4636 samples, 30%).

3.1.2 Dictionary

The dictionary is made based on the training data and all sentences are broken down into list of words: (1) Delete common words such as a, an, to, of, on etc. with high frequency but little semantic usage. (2) Stem words, such as "thanks" and "thank" as one word. (3) Delete low frequency words that appear once to reduce the size of dictionary for calculation efficiency.

3.1.3 Frequencies Matrix

A feature matrix is built to convert the textual information into numerical information. In the feature matrix, the number of rows indicates the number of samples, the number of columns is the length of the dictionary, and each element indicates whether the specific word has appeared in the current review, 1 for existence and 0 for absence.

To get a sense of correlation presented in our feature matrix, i.e. "bad" and "suck" may have a higher chance to present together, we perform PCA to capture the variance. The result shows that for the first component, variance explained is 2.3%, and for the next nine components, the variance explained is all around 1.0%. This shows that there isn't significant correlation between words and to achieve better accuracy, we include all the words in the dictionary. We propose that the lack of correlation comes from the nature of the text data. Most of them are very short sentences and extremely verbal.

3.2 Models

1. Naive Bayes with multinomial event model from sklearn is used. Input is the frequency vector and Laplace smoothing is used.

$$\hat{\theta_{yi}} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{1}$$

2. Support vector machines with linear kernel and RBF kernel are used in this project. SVM uses the same input and implementation package as Naive Bayes.

$$K(u, v) = u^T v \tag{2}$$

$$K(u, v) = exp(-\frac{\|u - v\|^2}{2\sigma^2}) \tag{3}$$

3. Neural Network
   Tensorflow is used in implementation. Input is the frequency vector that represents a review. The output is a vector with probabilities for different classes and the highest is selected as prediction. Label is a one-hot vector that represents the class. Loss function is cross entropy plus a regularization term. The vanilla Neural Network that we use:

$$h = Wx + b \tag{4}$$

$$\hat{y} = softmax(Wh + b) \tag{5}$$

4. Recurrent Neural Network
   A Bi-directional Gated Recurrent Unit Network (GRU) can capture the structure features of a sentence. Also, it solves the vanishing gradient problem which many recurrent neural network models have. Bi-directional GRU is commonly used in text analysis, which we want to compare with our models. Package scikit is used for implementation. In GRU, word vectors, instead of frequency vector, will be used and we choose glove.twitter.27B.zip.[3] These are pre-trained word vectors that are trained on twitter data set. The math for GRU is shown as follows:
   For right direction $\overrightarrow{h}_t^{(i)}$

$$\overrightarrow{z}_t^{(i)} = \sigma(\overrightarrow{W}_i^{(z)} x_t^{(i)} + \overrightarrow{U}_i^{(z)} h_{t-1}^{(i)}) \tag{6}$$

$$\overrightarrow{r}_t^{(i)} = \sigma(\overrightarrow{W}_i^{(r)} x_t^{(i)} + \overrightarrow{U}_i^{(r)} h_{t-1}^{(i)}) \tag{7}$$

$$\widetilde{h}_t^{(i)} = tanh(\overrightarrow{W}_i x_t + r_t \circ \overrightarrow{U}_i h_{t-1}) \tag{8}$$

$$\overrightarrow{h}_t^{(i)} = z_t^{(i)} \circ h_{t-1}^{(i)} + (1 - z_t^{(i)}) \circ \widetilde{h}_t^{(i)} \tag{9}$$

   Similarly for $\overleftarrow{h}_t^{(i)}$
   Output

$$y_t = softmax(U[\overrightarrow{h}_t^{(top)}; \overleftarrow{h}_t^{(top)}] + c) \tag{10}$$

4 Experiments & Results

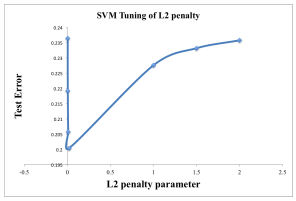
4.1 Sentiment Analysis

4.1.1 Naive Bayes Classification With Laplace Smoothing

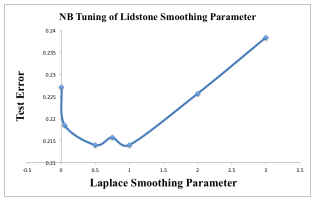
α (smoothing parameter) is tuned. The lowest test error (0.214) is achieved when α is 0.5 or 1. $\alpha = 1$ is used in further experiments.

4.1.2 Support Vector Machine With Linear Kernel

Before tuning the regularization, SVM with linear kernel, RBF kernel results in 0.23, 0.21 test error, respectively. Therefore, SVM with RBF kernel is excluded from future tuning due to the higher initial test error. L2 regularization is used to avoid overfitting. According to the graph shown below, the lowest test error (0.200) is achieved when L2 regularization is 0.02.



(a) SVM Tuning Result



(b) NB Tuning Result

Figure 1: Sentiment Analysis test error

4.1.3 One layer neural network

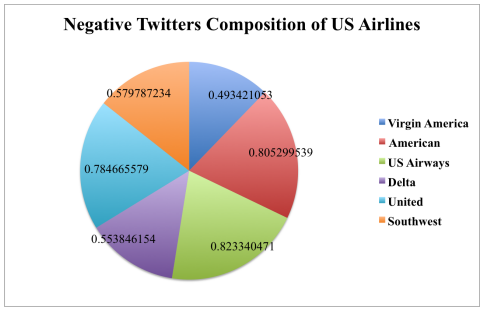
In each stochastic gradient descent step, only a batch of 100 samples is used in SGD to increase the training speed. Tuning parameters are learning step and regularization term, which are 0.01 and 0 respectively. The best test error for this one layer neural network is 26.3374%

4.1.4 Bi-directional Gated Recurrent Unit Network

Word vectors from GloVe with a dimension of 50 will be used in GRU. A 2-layer GRU has a test error of 26.5%. A 3-layer GRU has a test error of 25.6%. Learning step is 0.01; l2 is 0.

4.1.5 Sentiment Analysis Result

In sentiment analysis task, SVM with linear Kernel achieves the best test accuracy. Therefore, SVM is recommended in this section. According to the result from linear SVM, Virgin American performs the best according to its lowest negative review composition in its total reviews.



4.2 Negative Reason Prediction

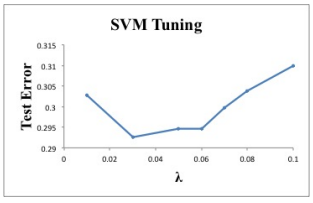
In this section, the goal is to determine the most negative reason on flight services. All the negative reviews have been collected for this task, and separated into labeled set and unlabeled set. We will make predictions on the unlabeled set.

4.2.1 Naive Bayes Classification

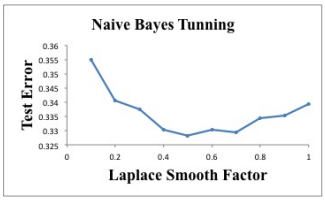
Using Naive Bayes Classification, the test error is average to 29.26% after ten-fold cross-validation, with a Laplace smooth factor of 0.5.

4.2.2 Support Vector Machine

L2 regularization is tuned. The best test error for SVM is 32.82%, when l2 regularization = 0.03.



(a) SVM Tuning Result



(b) NB Tuning Result

Figure 2: Negative reason test error

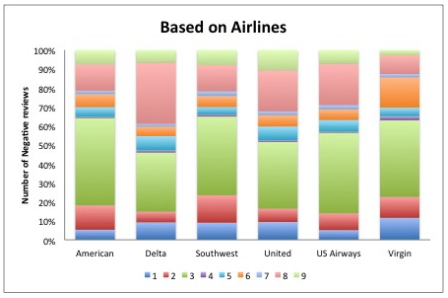
4.2.3 One layer neural network

Input is same as before. Label's dimension changes to 9. Learning step and regularization term are 0.01 and 0 respectively. The test error is 37.82%

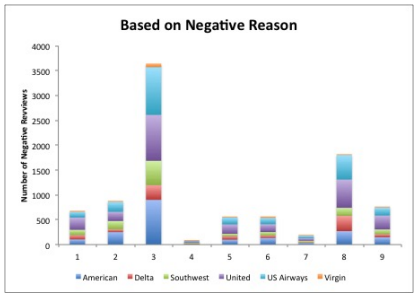
4.2.4 Negative reasons classification

Given by the lowest test error, Naive Bayes is used for the prediction of unclassified data. Result is shown below that most complaints are on customer service. One postulate might be due to the high

volume of contact. Since various reasons can lead to calling customer service. Thus correlations between classes may play a factor in determining this result.



(a) Major negative reasons for each airlines



(b) Airlines performance on service issues

Figure 3: Negative Reason Classification

Airline	1st Negative Reason	2nd Negative Reason
Virgin America	Customer Service Issue	Flight Booking Problems
United Airlines	Customer Service Issue	Late Flight
Southwest	Customer Service Issue	Late Flight
Delta	Late Flight	Customer Service Issue
US Airways	Customer Service Issue	Late Flight
American Airline	Customer Service Issue	Late Flight

5 Conclusion

- It is pleased that our vectors work. It is surprising that SVM and Naive Bayes perform better than deep learning methods. And the accuracy is very high, 80%. We think the reason behind this is that while movie reviews have a lot of sarcasm[1], which is very difficult for any model to grasp, twitter reviews are much more straight forward, and thus most of the sentiments are expressed directly at the word level. That is to say, with specific word appearance, sentiment is indicated clearly, which justifies our feature representation using frequency vector. It is possible to judge a twitter airline review's sentiment only by identifying positive words in a review. Therefore, given the nature of our data set, the task can be solved at bag-of-word level well.
- However, it is too early to say that neural network can not perform better than bag-of-word models. The frequency vector used in vanilla neural network is so large that takes enormous time to train, roughly 6 hours for 10,000 iterations now. Therefore, clever ways of reducing frequency vector size are needed. Meanwhile, better tuning parameters can be figured out once training time is significantly decreased.
- Another possible reason is that for recurrent neural network, GRU in our project, labeling every node is very important. While this model can achieve as high as above 80% accuracy using Stanford Sentiment Tree Bank dataset[4], Our results show that without sufficient labeling, this model is not able to achieve an accuracy above 80%, which means RNN family needs strong supervision. However, most of the online reviews and other documents only have limited labels. Better labeling algorithm on new data set should be thought about in future work.

6 Reference

1. Pang, Lee. 2005. CS224D Slides.
2. Yuan, Zhou. Twitter Sentiment Analysis with Recursive Neural Networks.
3. Pennington, Socher, Manning. GloVe: Global Vectors for Word Representation
4. Stanford Sentiment Tree Bank http://nlp.stanford.edu/sentiment/treebank