Yelp Review Sentiment Analysis

Introduction

Why sentiment Analysis

92% of marketing professionals think that social media has profound impact on their business. Sentiment analysis can help business win more customers.

It will help business adjust marketing strategy, develop product quality, improve customer service. Social media is a place where your customers chit chat about your business and sentiment analysis gives you insight into how your brand/service is perceived by your customers. Social media has network effect where positive reviews bring in more customers and they in turn experience good service will likely to leave a good review and the positive cycle begins. Therefore it's crucial for a business to get to know what their customers' opinions are about their product/service and evaluate if their initiatives to improve quality and service actually align with customers' tastes.

Where is the data coming from

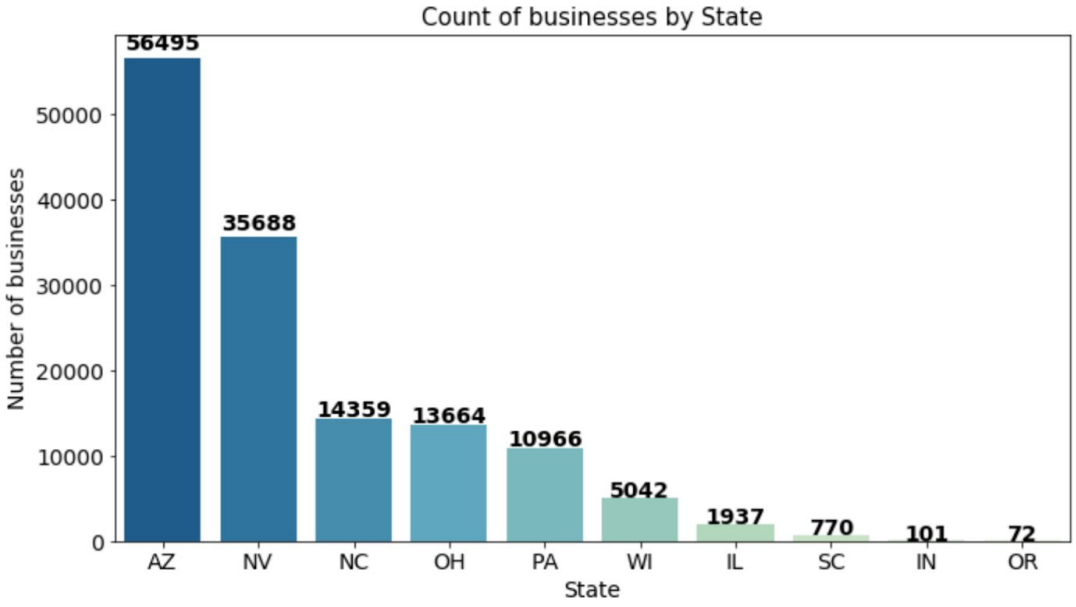
The Yelp dataset is from Yelp challenge [https://www.yelp.com/dataset/challenge]. The data format is in json file.

Business.json: Contains business data including location data, attributes, and categories.
Review.json: Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

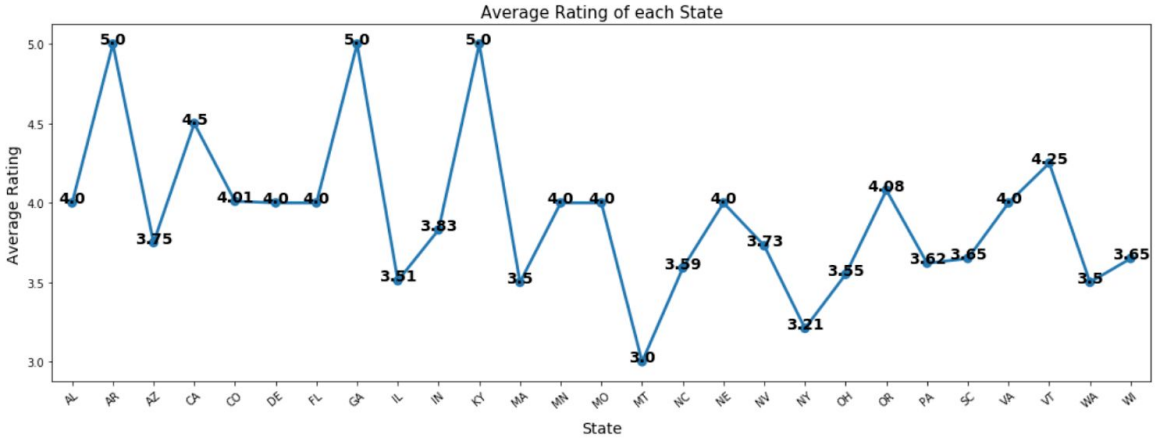
Exploratory Data Analysis

Understand Business

There are 188593 unique businesses in the Yelp data set. The businesses are spread all over the country. The 10 states that have the most businesses in this data set are:

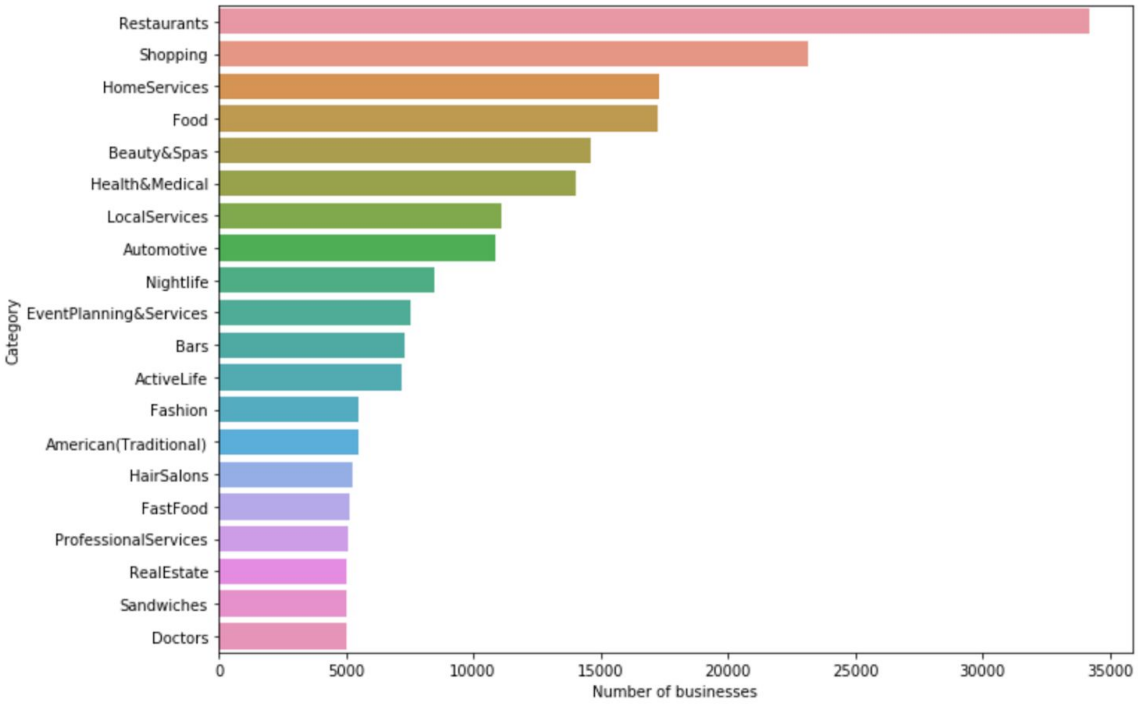


The average rating of all business in a state:



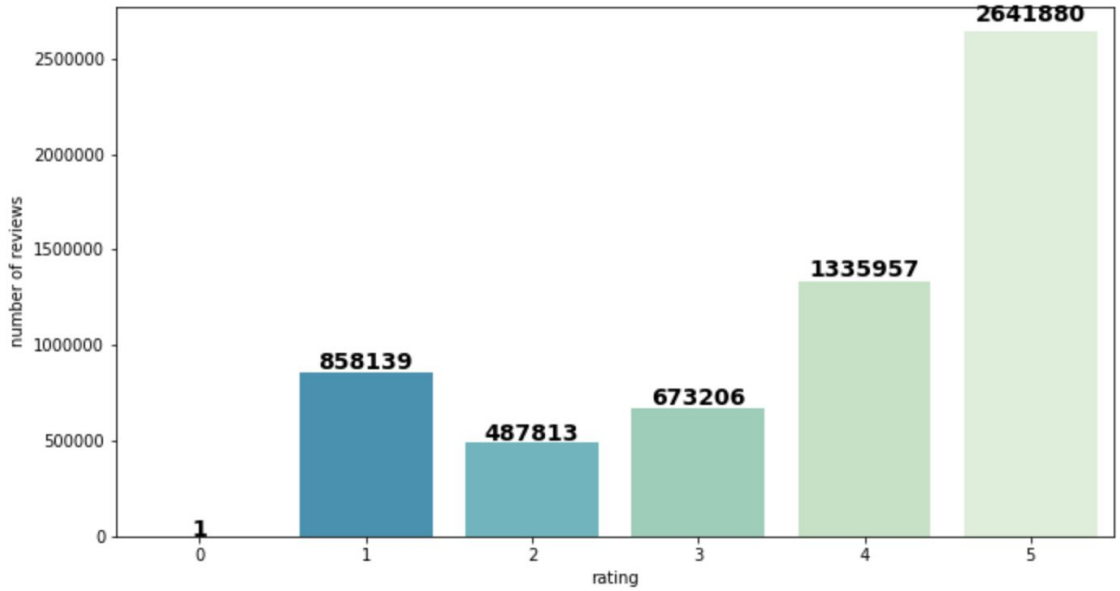
It's reasonable to say that customers get generally good experience if they give >3 star to a business. From above graph, we can see almost all states customers are generally feel good.

The distribution of business categories in the data set is show below. The majority of business that uses Yelp are restaurants or food related businesses.



Understand Reviews

There are ~6M reviews by ~1.5M unique users about ~190K businesses.



There are more positive ratings than negative ratings. In this project, I will discard neutral ratings (ratings of 3) and label positive sentiment if a user gives 4 or above stars for a business, negative sentiment if a user gives 2 or less stars for a business.

Modeling

LSTM

In NLP applications, the sequence of the word can dramatically change the meaning of a sentence. For example:

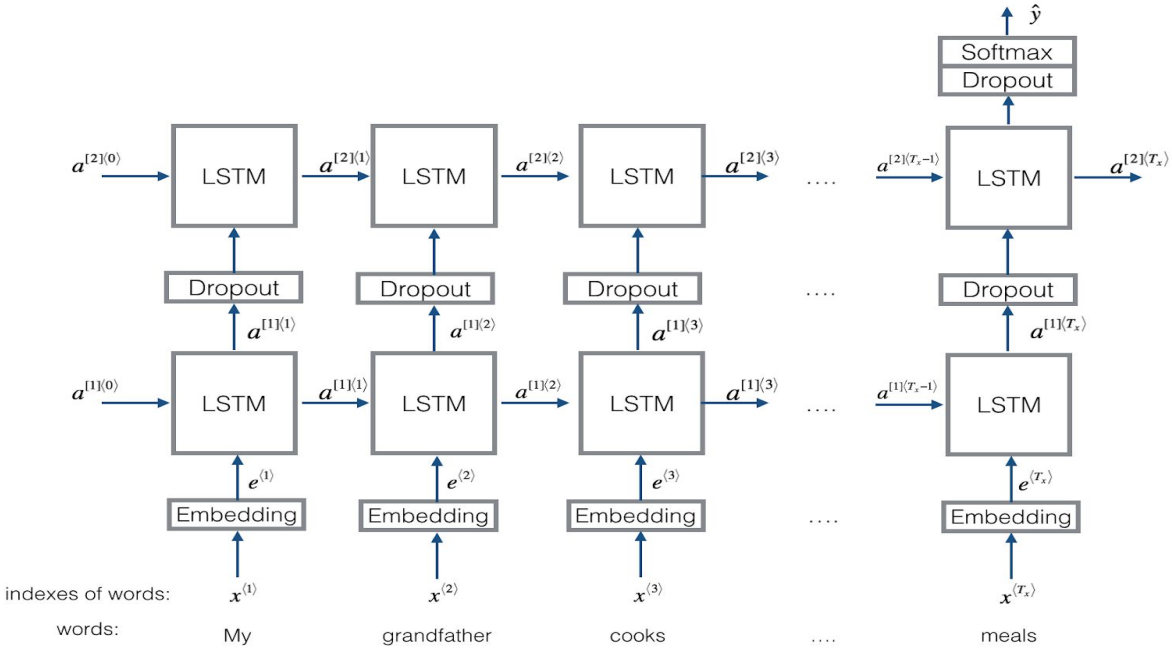
The food tasted good. However the waitress was rude when we asked for our bill. We got stomach pain the next morning, which makes us wondering the hygiene of the restaurant.

The first few words seem like a positive sentiment. However the sentiment changes dramatically after first sentence. An LSTM model is perfect for this task as it has memories over the sequence of words. It will pick up memory that is important for the task and forget unimportant memories. In this example, the first sentence is irrelevant for the overall sentiment of the review.

Word Embedding

Word embedding are a type of word representation that allows similar words to have similar representation. In this project I used the Glove 50 presentation where each word is represented using 50 dimension vector. The benefits of using word embedding are: it's a dense representation, allows for more efficient computation than bag of words (which is sparse). The word embedding is trained on much larger corpus, which allows us to use transfer learning. The representation captures the meaning of a word, i.e. the context of a word.

The overall architecture of the model is shown below. The dropout layer is added to prevent overfitting. In the hyper-parameter tuning section we will see the effect of using different layer of LSTM, the number of LSTM cells on the loss and accuracy of the training and test set.



Data Preparation:

1. Obtain labels

   I label stars >3 as positive review and stars <3 as negative review. Leave out the neutral review of stars=3. I fetched 140K reviews in total, half of them are positive and half are negative.

2. Clean data

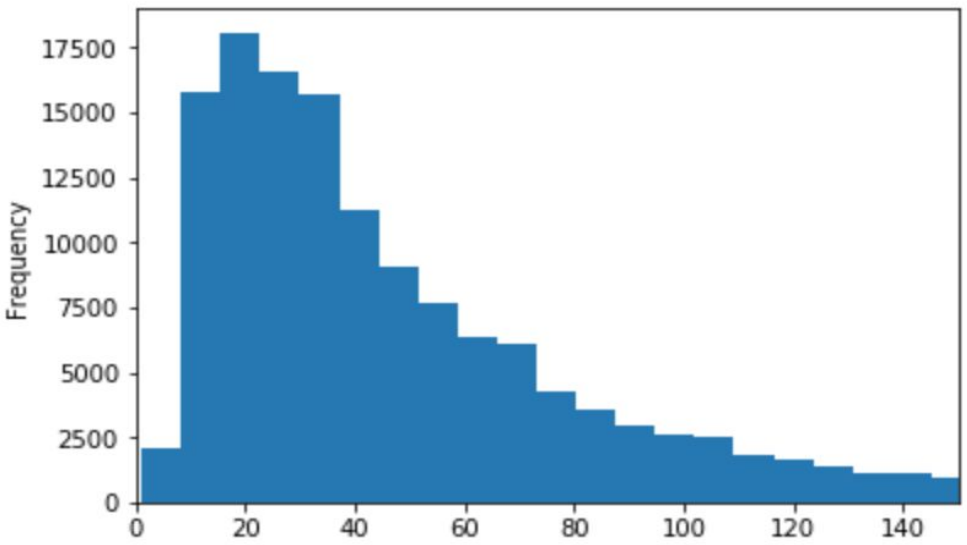
   Several steps of data cleaning are performed to work with Glove model. They are: expand contractions, remove non-alphabet words (such as punctuations, special characters such as @,#, links etc), lower-case everything, and remove stop words. Tokenize and represent each review with the indexes of words from Glove 50. In the model, I discard the words that are not in Glove 50 vocabulary. (glove 50 vocabulary is 400K words).

3. Choose the length

   We need to choose a cap for sentence length we will process. If we pick a number that is too large, the model will train much slowly and not necessarily leads to a

good model. If max_len is too small, then most of reviews will be truncated and may not leads to a good model either.

Plotting the length of review histogram, we can see that most reviews are 20-60 words long, therefore max_len=60 seems a reasonable choice.



Model Tuning:

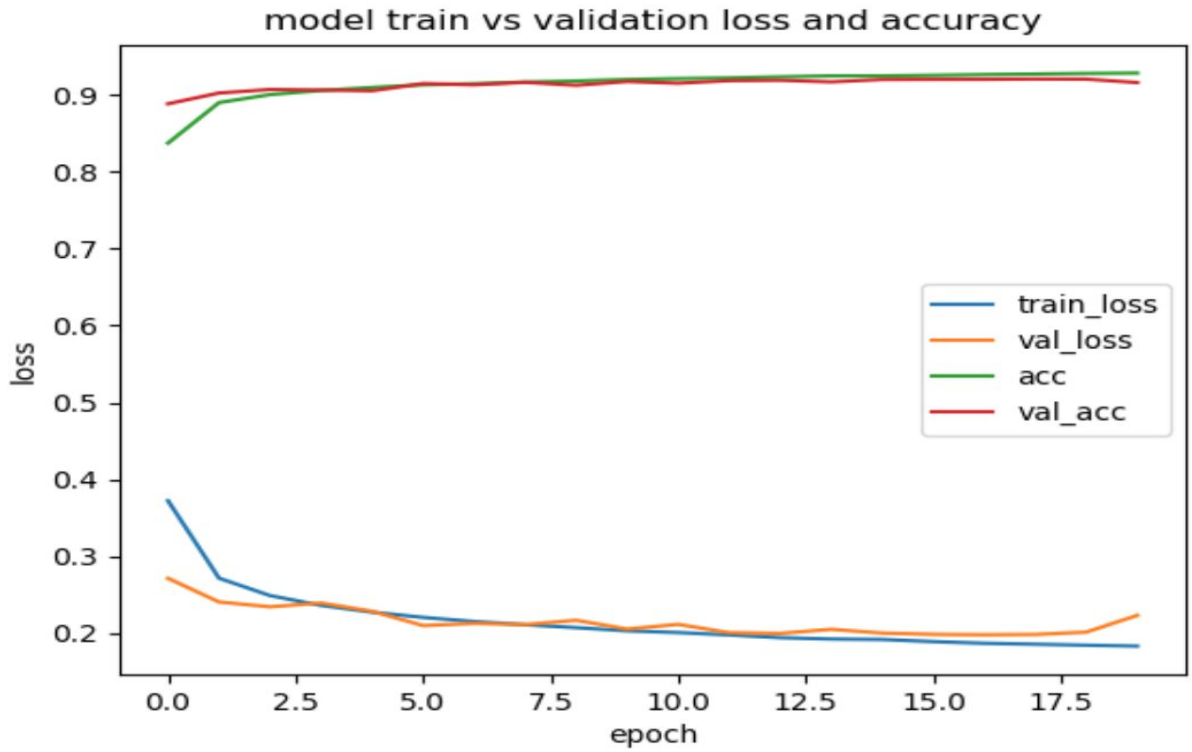
(layers, LSTM cells per layer)	Train loss	Validation loss	Train acc	Validation acc	Test acc
(1,6)	0.2754	0.2398	0.8913	0.9021	0.9026
(2,6)	0.2580	0.2264	0.8974	0.9080	0.9069
(1,12)	0.2288	0.2215	0.9106	0.9109	0.9110
(1,24)	0.1832	0.2231	0.9282	0.9158	0.9157
(1,24) *	0.1698	0.1864	0.9341	0.9278	0.9287
(1,24) * backward	0.1509	0.1587	0.9432	0.9387	0.9383

From Row 1-4 from above table, we can get some intuitions:

Generally speaking, having multiple LSTM layers will help to capture structure in the text. It will help complex tasks such as language translation. Since we only predict sentiment (positive or negative) the structure of the text might not matter that much. Therefore we've seen from 1 layer to 2 layer LSTM, the performance doesn't change much.

Having more LSTMs in a layer will help capture different types of reviews, for examples if it's a negative reviews, there might be several categories of negative reviews (high price, bad service, long waiting time etc). It gives more granularity in the model. We've seen that accuracy improves for 1% when we increase LSTMs from 6 to 24.

The above table shows results of training for 20 epochs. I also plotted the training loss/validation loss vs epochs and accuracy vs epochs. When we have 24 LSTMs, training for 20 epochs will result in overfitting. We can stop at around 15 epochs.



Can We Do Better?

Sentiment analysis is even hard for human being due to subjectivity, tone, lack of context, irony, and sarcasm so a typical 80% accuracy score is expected according to domain experts. The accuracy on yelp data set is higher due to two reasons: 1) yelp reviews are individually labeled by each user so there is less ambiguity to user sentiment. 2) yelp reviews are more straightforward compared to other sources such as twitter. Here we will print out some misclassification examples to see if there are room for improvement.

Misclassification examples:

Example 1: Original review:

Came in for my chronic knee problem. Got evaluated. Dr. Todd Winton took enough time to patiently listen to me & see my condition. But I didn't really get confidence in his understanding of the condition. A couple of the pathological/evaluation terms he used were not right (I verified that on the internet afterwards). Can't comment on the treatment as i didn't undertake it.

After cleaning:

came chronic knee problem got evaluated dr todd winton took enough time patiently listen see condition but really get confidence understanding condition a couple pathological evaluation terms used right verified internet afterwards cannot comment treatment undertake

The model mis classifies it as a positive review. We've seen that stopwords removal got rid of all "not" and "no", which are crucial for sentiment. A "not" or "no" could totally change the sentiment of a sentence. Therefore we could try modify the stopwords list in NLTK to exclude 'not' and 'no'. In the summary table, row with * refers to the model with modified stopwords.

Example 2:

Original review:

Hi, It's a Long one! I brought my 25 yr. old car that I cherish to Danny's car wash to get the works. Oil change, and full Detail done. The first salesman I came upon was nice enough, as he was gouging me for every dollar he could get from me. He bulldozed right over me! I told him I really wanted it to be perfect as I Love my car. He quoted me $750! I balked at that and said I

wouldn't pay over $400. (Mind you, I have a 2 dr. Little car!) Then another specialist-oil this time-came over and Told Me that with the age and rollover of the miles on my car, that I'd need synthetic oil change. Which is $50 more, and you can Never go back to basic oil again. I told him, that those are Original miles, Hasn't rolled over. He said, No, probably rolled over a couple times. THEY ARE ORIGINAL MILES. BUT, he didn't listen... they told me it would be done by 4:00pm-5 hrs later. (I had to pick my son up from school, so it Had to be done.) I came back at 4:00, not done til 5:00. Price was still $400, but vents, & windows not cleaned, upholstery not vacuumed or shampooed in back, my chrome rims were black, not even touched. And, overall sucky! I came home and cried, and cleaned my whole car myself. I was worried my partner was going to be mad that I spent so much money, and nothing was done, but a basic car wash, and Wrong oil change!...She was... So, the reason for the 4 star rating, is because she wrote a letter to the owner.(would've given them 5 star if not for the 1st visit!) Apparently, under New ownership now-Not Danny's Family, Now "Jackson's." The manager called us and asked us to please bring it back, and give them another chance. We did. We left my car with them all day, but it was worth it. One of the managers talked us through Everything, he and 2 other managers had been versed on my car. Top to bottom.

The model mis classifies this one as positive. Since we cap the sentence at 60 words. We saw in this example, the first half of the review was negative. Then things got better b/c his partner wrote a letter to the owner and the second time they were there, the experience was positive. When we cap the sentence at 60 words, we only see the first half of the review. One thing we can do is to check if a high percentage of misclassified examples have longer length than 60 words. If this is true, we can try increase max_len and train again.

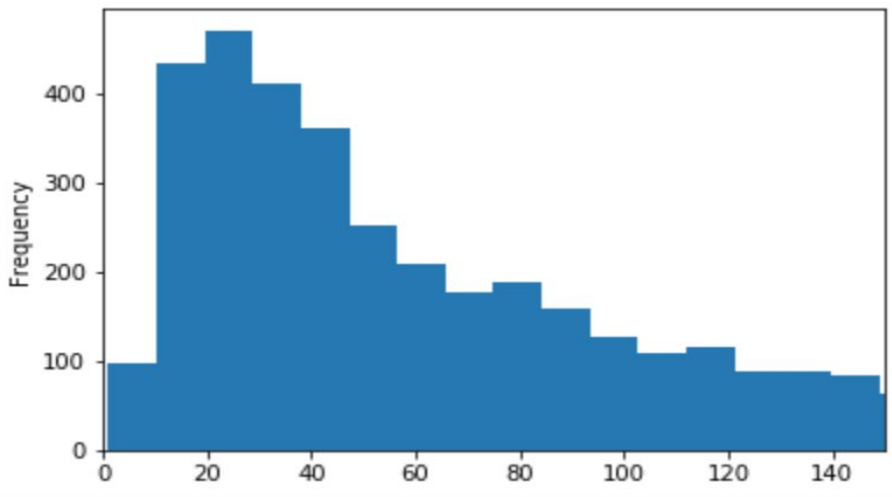
We will see results from above two improvement suggestions in next section--Experiments.

Experiments

1. Modified stopwords list:

   Exclude “not” and “no” from NLTK stopwords list. With this approach, we see a 1% improvement in the accuracy. The validation and test set accuracy is 92.8% (compared to 91.5% before)

2. Length distribution among misclassified examples



The length distribution is pretty similar to the whole data set we've seen in previous sections. There is no evidence showing misclassified examples tend to be longer reviews. Therefore I don't expect performance gain if we increase max_len parameter.

However, we do notice that for long reviews, second half of the review is more important than first half. A lot often the twist of the story happens in the middle as we've seen in one of the misclassification examples. In addition people tend to say nice things first if they want to critique something. It's worth trying to keep max_len the same but trim sentence starting from the end of the review than from the beginning of the review. In addition, it might also be the case that future time step helps better to understand the context and meaning of the words. For example, "He said, Teddy bears are on sale", "He said, Teddy Roosevelt was a great President". The previous two words "He said" doesn't help understand what Teddy is, we need to look ahead in the future to tell whether Teddy refers to a president or Teddy bear. Last model I tried combined these two observations. It's a backward in time model where the x(t), t(t-1)... are feeding into model sequentially. (x(t) represents last word in a review). This model achieves 1% improvement over the last model. (93.8% accuracy on test data set)

Mis-classification examples:

Be sure to schedule your appointments an hour ahead of when you need to see the doctor. Have been here multiple times and always end up waiting 30min -an hour! Otherwise the doctor and staff are good.

This is a hard one since the review is mixed with both positive and negative comments. The model predicts positive sentiment, the user gives negative review.

En général quand j'entends parler de marché je me dis cool je vais voir des fruits, des légumes et pleins de produis locaux En plus decela quand j'ai vu la façade du bâtiment je me suis dis ce Marché est vraiment magnifique (refait à neuf on dirait) et lorsque l'on rentre c'est la déception. En effet il n'y a à l'intérieur aucun marché mais plus des boutiques type souvenir ou autres. C'était je pense un marché dans le temps mais c'est maintenant plus un lieu touristique qui selon moi n'a d'intérêt que l'extérieur.

We also see some misclassification examples are written foreign language. For future improvement, we can filter out these comments and only include reviews in English in the data set (since Glove 50 only includes English vocabulary).

Conclusion

In this report, we built a LSTM model with transfer learning using Glove to perform sentiment analysis on Yelp reviews, achieving 93.8% accuracy on test data. We've seen that performance doesn't improve much with increasing number of LSTM layers, but increasing the number of LSTM cells in a layer will help capture more categories of different types of positive/negative reviews. In addition, we've found that removing the "not" and "no" from stopwords list will improve the model performance since themselves are important indicators for sentiment. Finally we've found that backward in time model (future dependency of words) further improves the model by another 1%.