Sentiment Classification on Amazon Food Reviews & Beyond

W266 Final Project

David Larance & Matthew Prout

August, 2018

Project Introduction

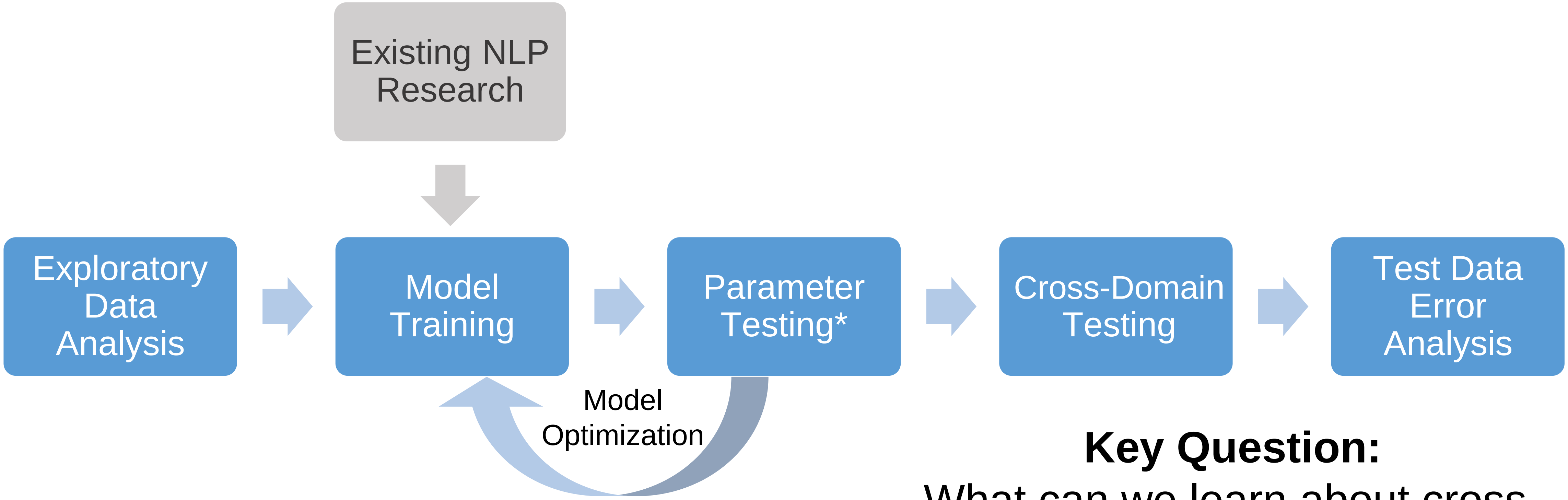
Objective: Train and optimize a sentiment classification model and then measure sentiment accuracy across other domain data sets

1. **Primary model built off Amazon Food Reviews dataset**
   a) **568k reviews**
   b) **Text reviews paired with 1-5 rating scale**
2. **Bulk of project was focused on optimizing model, not just using first working model**
3. **Test data sets selection (restaurants, beer, and movies) based on diversity of domain, review style, and dataset size**

Key Question:
What can we learn about cross-domain sentiment analysis?

Research Process

The team spent significant amount of time on model training, testing, optimization, and error analysis.



Key Question:
What can we learn about cross-domain sentiment analysis?

*Includes Base Model Error Analysis

Baseline Models

Naïve Bayes

- Split the data 2/3 training and 1/3 test
- Used the Scikit-learn BernoulliNB classifier
- 87.6% test accuracy on the Amazon reviews

Neural Bag-of-Words

- Split the data 60% training, 10% validation, and 30% test
- Used the model from our assignment 2 (Bengio 2003)
- 93.7% test accuracy on the Amazon reviews

LSTM Model

Data Preparation

- Ratings need to be converted to a 0 or 1
- Ratings are converted to tokens, which are converted to word IDs.

Word Embeddings

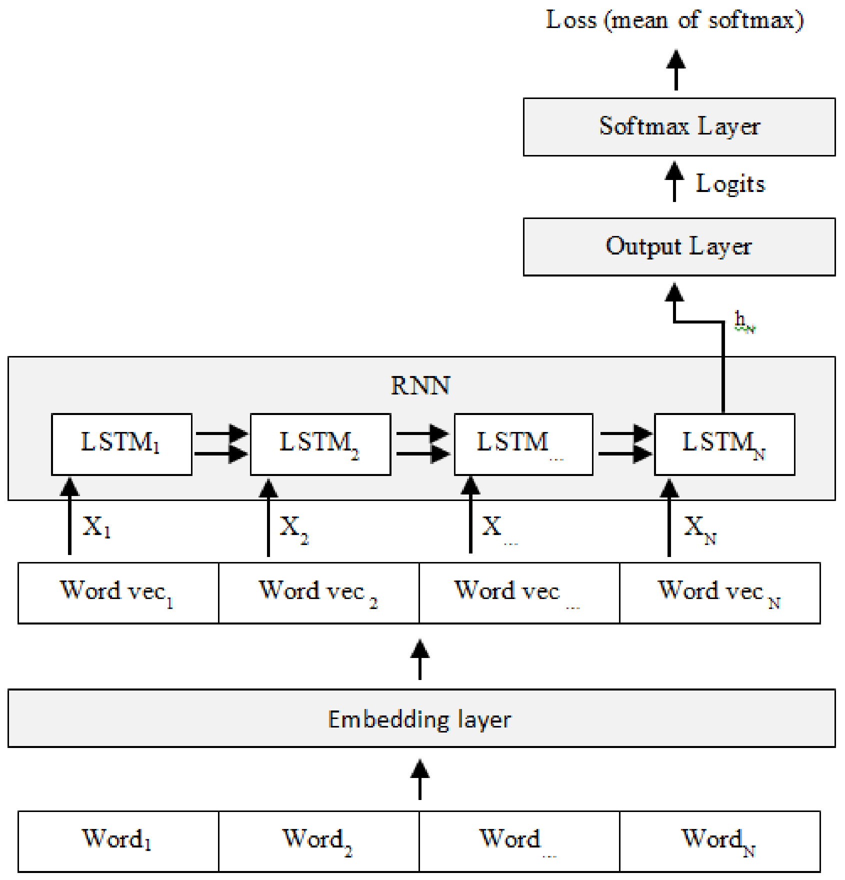
- We use pre-trained GloVe vectors (dimension 50)

TensorFlow Graph

- Consists of an embedding, RNN, output, and softmax layer.

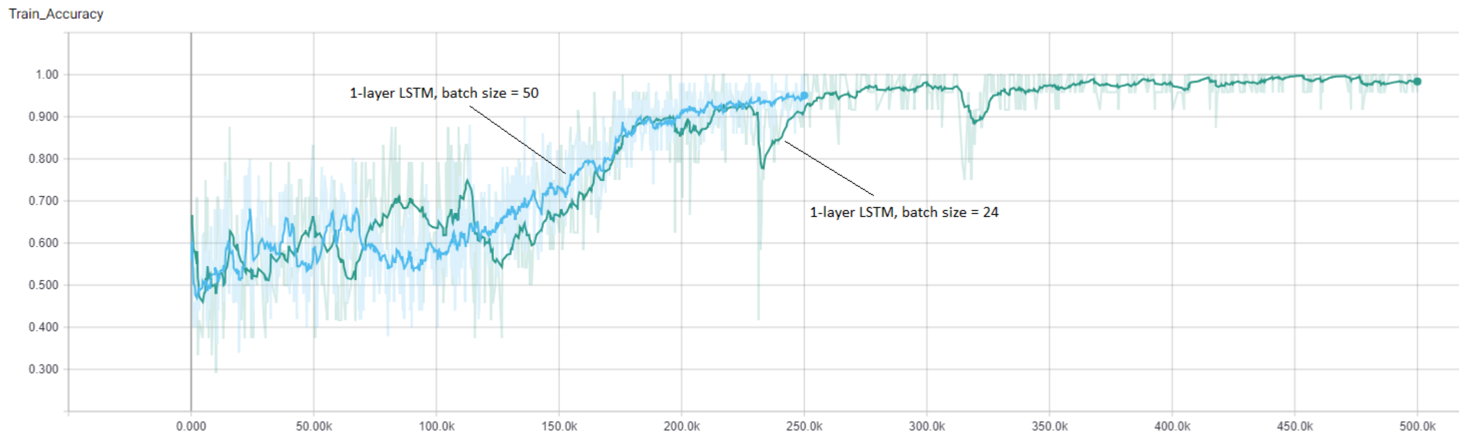
Training

- A loop feeds training data to the TF graph.
- The graph is then saved as a checkpoint that can be loaded by the testing code.

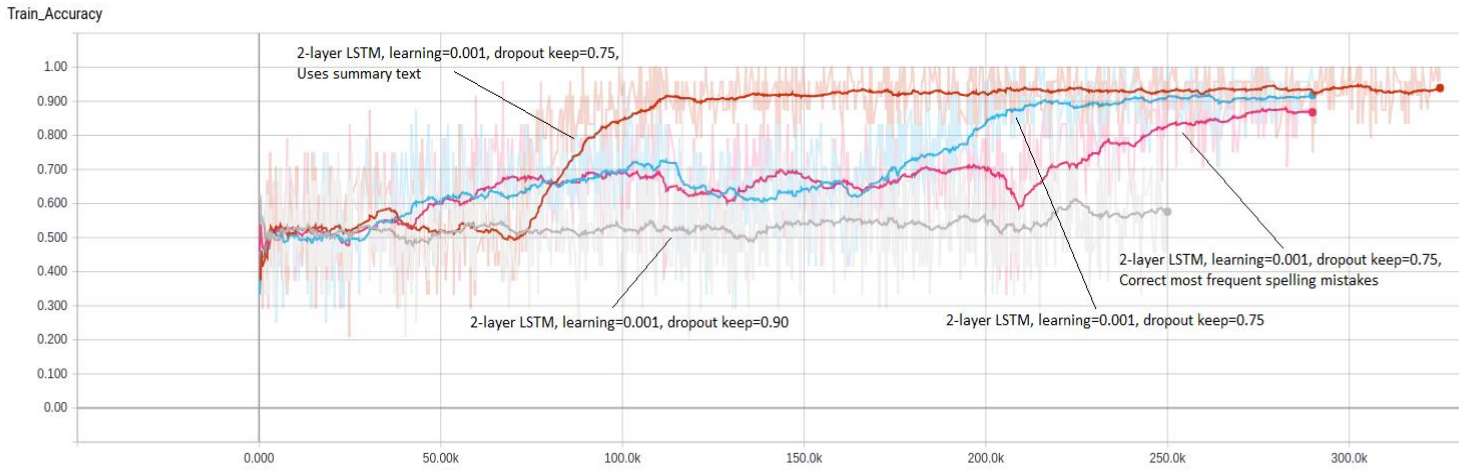


LSTM Model Optimizations

Test	LSTM Layers	Learning Rate	Dropout keep probability	Batch size	Test Accuracy
1-layer LSTM model	1	0.001	0.75	24	0.87
1-layer LSTM model with larger batch size	1	0.001	0.75	**50**	0.81



Test	LSTM Layers	Learning Rate	Dropout keep probability	Batch size	Test Accuracy
2-layer LSTM model	**2**	0.001	0.75	24	0.91
2-layer LSTM model with less dropout	2	0.001	**0.9**	24	0.63
2-layer LSTM model using summary text	2	0.001	0.75	24	0.91
2-layer LSTM model correcting common misspellings	2	0.001	0.75	24	0.86



Note: We did test changing the learning rate (to 0.01), but model was not learning, so the training was aborted

Base Model Error Analysis

Mis-predicted reviews were analyzed to understand where the model failed

Failure Categories

1. **Fact Heavy:** Review is description based (price, neutral phrasing, little sentiment) instead of opinion based
2. **Category Specific:** Text contains taste specific descriptions unique to a food type (i.e. coffee, dog food)
3. **Incorrect Rating:** Binary based model forces fringe ratings (2-4) to fall on one side of the fence, sometimes incorrectly
4. **Review Commingling:** Text mixes focus between products for comparison purposes

Failure Corrections

1. **Summary Text (R):** Concatenate summary text to main review text
2. **Misspellings (RNI):** Build incorrect spelling dictionary to correct misspellings
3. **Category Specific (DNR):** Train category specific models to differentiate between category specific language
4. **Fact/Opinion Tagging (DNR):** Tag phrases as fact or opinion and de-value the role that pure facts play in sentiment weighting

R = Revise; RNI = Revised, No Improvement; DNR = Did Not Revise

Cross Domain Testing

	Amazon: Fine Foods	Yelp: Restuarants	RateBeer: Beer	IMDB: Movies
Size	568k reviews	1.5k reviews	71k reviews	2k reviews
Qualitative Review		- Short reviews - Mix of restaurant description (factual) and reviewer experience (opinion) - High use of taste, smell, visual descriptions	- Long reviews - Fact based descriptions (color, viscosity, etc.) with little expression - Opinion and sentiment usually at end of review - Review distribution skewed higher	- Extremely long reviews - Descriptive in nature - No food, taste, smell descriptors used - Opinion often clouded by complex sarcasm or phrasing
Accuracy	Accuracy: 0.91	Accuracy: 0.72	Accuracy: 0.66	Accuracy: 0.44
Loss	Loss: .27	Loss: 0.59	Loss: 0.72	Loss: 1.38
Error Analysis Comments		**Forecast:** Good performance due to food related overlap in vocabulary and strong opinions. **Post Testing:** Errors due to restaurant comparison, situational descriptions, and implicit price/quality descriptions	**Forecast:** Good performance due to visual, taste, and smell related descriptions. **Post Testing:** Almost all errors have domain specific descriptions (head, color, beer specific taste)	**Forecast:** Moderate performance due to no food domain overlap. **Post Testing:** Length of reviews and complex sarcasm / phrasing led to terrible performance

Conclusion

Achieving 72% accuracy on Yelp restaurant reviews shows it is possible to share classifier models across review data sets. The most impactful considerations are review length and dataset domain overlap.

Lessons Learned:

- Don’t underestimate workload to train, optimize, and re-test LSTM models on very large datasets
- Binary sentiment is one thing, but multi-class sentiment could be more beneficial
- Many variables present in test data sets can impact accuracy
- Error analysis processes could be more robust to handle advanced problems (word weighting, word specific sentiment, etc.)