Carnegie Mellon University

Various Approaches to Sentiment Classification for Yelp Reviews

Guannan Tang*, Neil Ashim Mehta*, Joohyung Shin*

Introduction

Motivation

- **To explore novel techniques that utilize the language of a Yelp review to predict ratings.**

Dataset

- **Yelp Review Dataset**
- **~36,000 Reviews in Dataset**
- **Features: Review Text,** Restaurant Name, Sentiment Score, Usefulness Score, Number of Characters, Number of Words, Restaurant Category

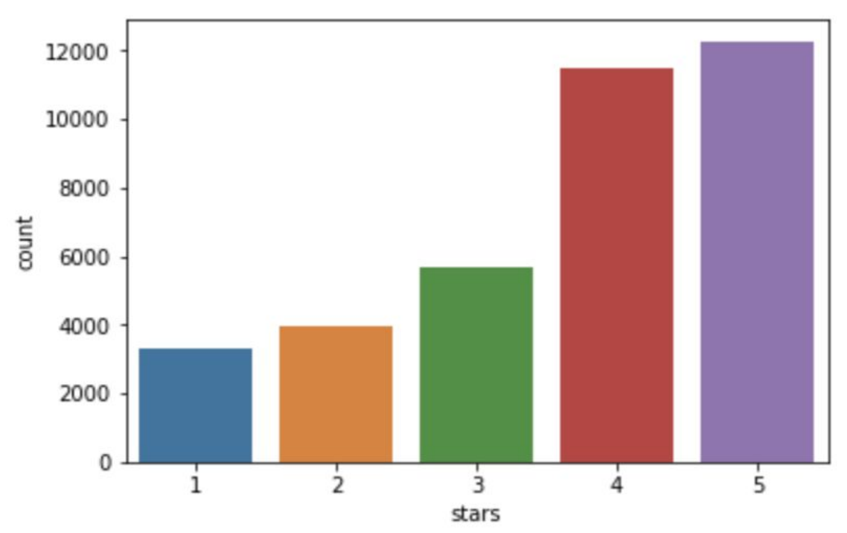


Fig 1. Distribution of Stars

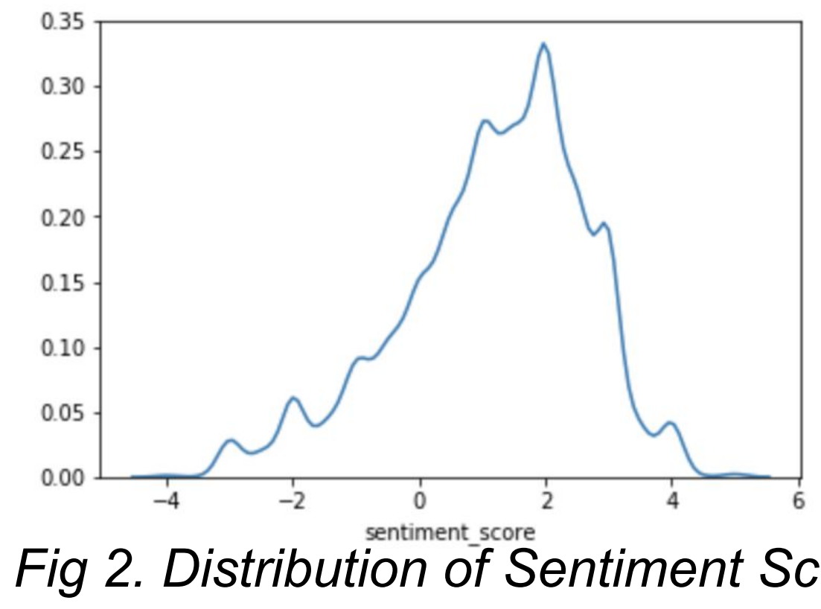


Fig 2. Distribution of Sentiment Score

Feature Engineering

- **Text (User Reviews)**

  stemming -> removing stop words and punctuations -> lower case
- **One-hot encoding**
- **Filtering NaN values**

Experimental Methodology

- **Sentiment analysis task with variable sequence length of review text and other relevant features.**
- **Test different encoding styles on text data:**
- **one-hot encoding based on occurrences (binary)**
- **one-hot encoidng based on word count**
- **tf-idf**
- **implicit word embeddings**
- **top frequent words**
- **Metric: Classification Accuracy**
- **Training epochs: 100**

References

- Bessou, Sadik, and Rania Aberkane. "Subjective Sentiment Analysis for Arabic Newswire Comments." *arXiv preprint arXiv:1911.03776* (2019).
- Carbon, Kyle, Kacyn Fujii, and Prasanth Veerina. "Applications of machine learning to predict Yelp ratings." (2014).
- Yu, April and Chang, Daryl. “Multiclass Sentiment Prediction using Yelp Business Reviews” (2015)

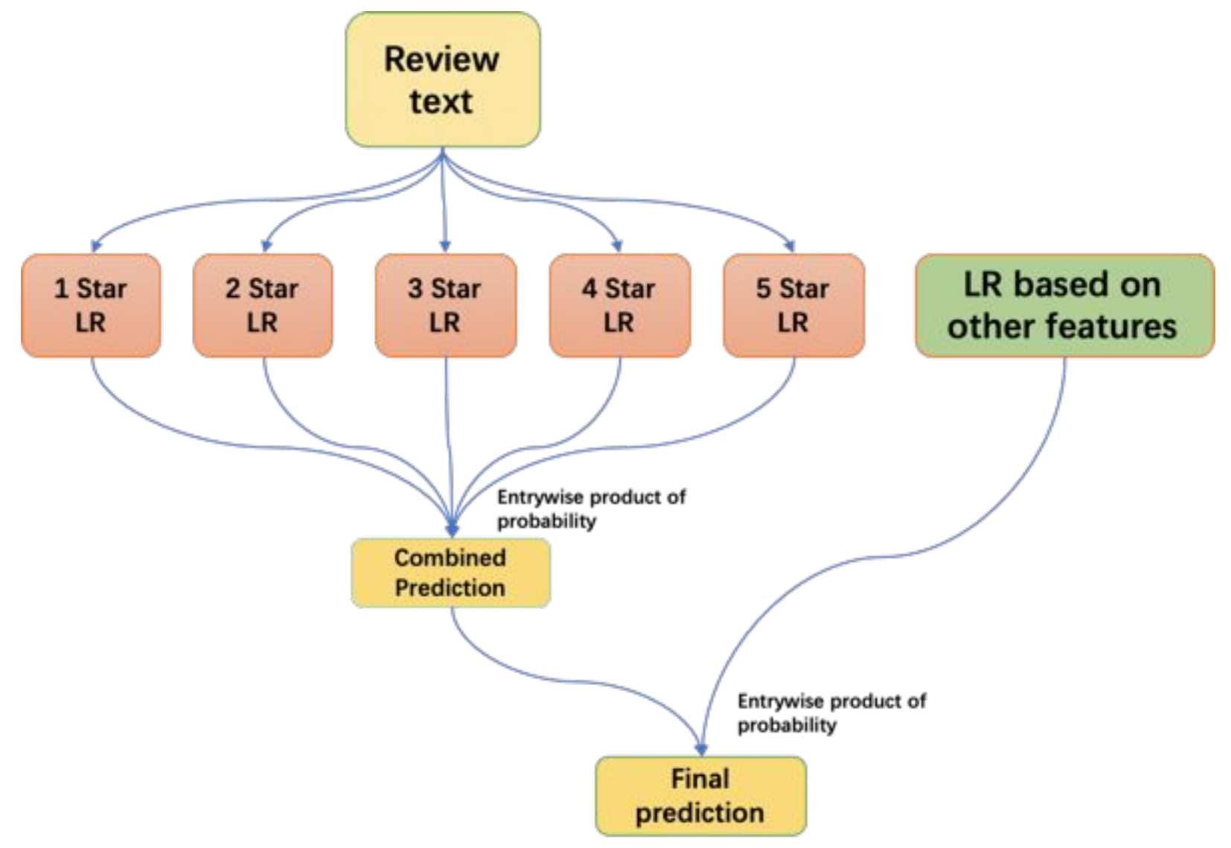
Models

1. Multinomial Naive Bayes

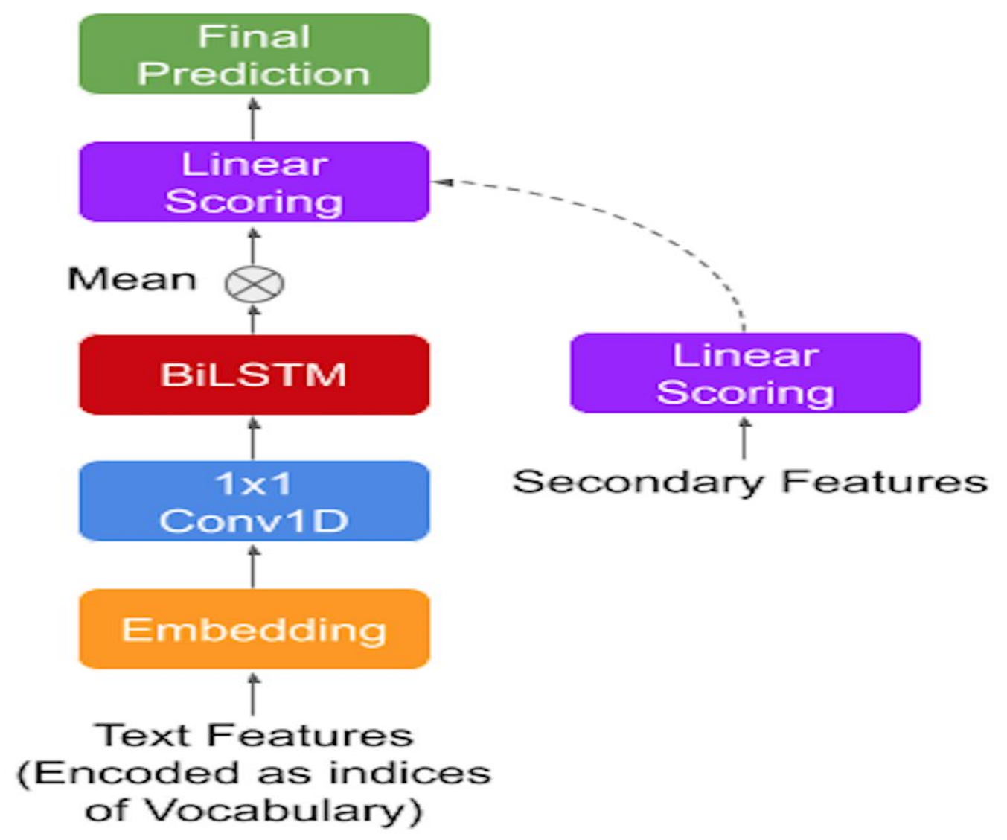
$$\frac{P(C=c^1|x_1,\ldots,x_n)}{P(C=c^2|x_1,\ldots,x_n)} = \frac{P(C=c^1)}{P(C=c^2)}\prod_{i=1}^{n}\frac{P(x_i|C=c^1)}{P(x_i|C=c^2)}$$

Expression of Naive Bayes eqaution - Probabilistic Graphical model (Stanford)

2. Ensemble Logistic Regression



3. Recurrent Neural Network



Result of Our Models

Comparison of Classification Accuracies for Models

# words	10000	15000	20000	25000
MNB(tf-idf)	0.5368	0.5263	0.5183	0.5068
MNB(count)	0.5704	0.5725	0.5783	0.5789
MNB(bin)	0.5739	0.5795	**0.5819**	0.5762

Table 1. Results of Classification Accuracies on Multinomial Naive Bayes

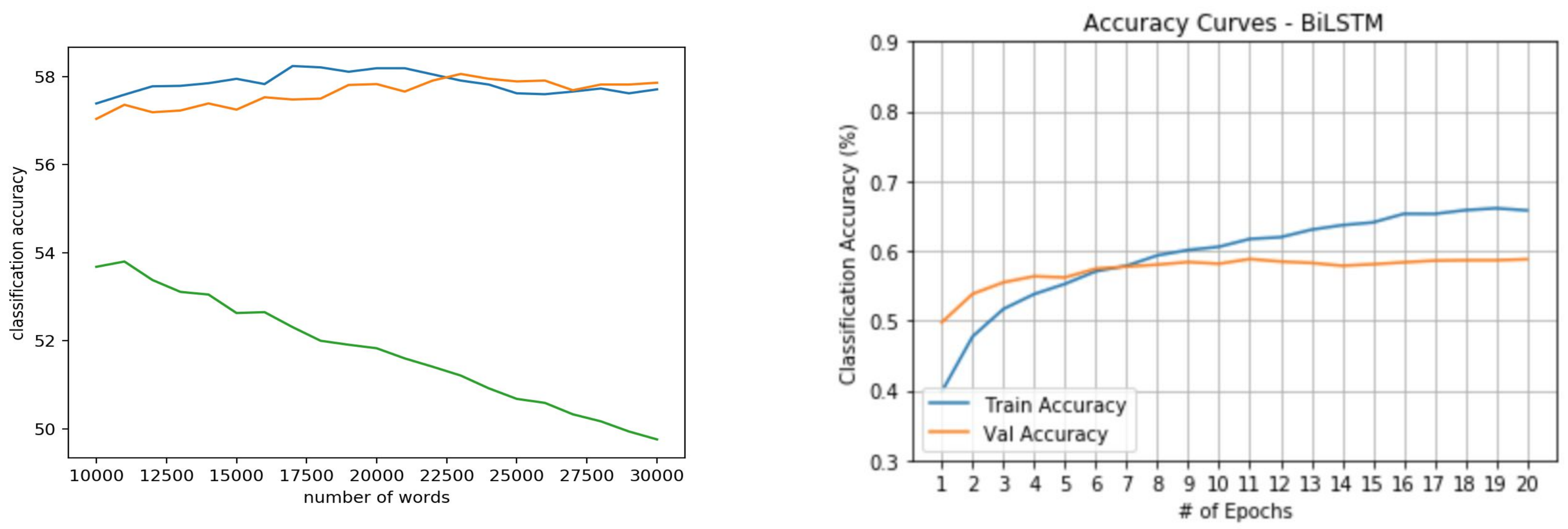


Fig 3. Comparison of Accuracies between different methods (MNB) (Left), BiLSTM with Dropout and Mean Reduction (Right)

# words	1000	3000	4000	10000
ELR(count)	0.54268	0.5594	0.5518	0.5418
# top words	500	1000	1500	2000
ELR(top frequent)	0.5464	**0.566**	0.5631	0.5604

Table 2. Results of Classification Accuracies on Logistic Regression

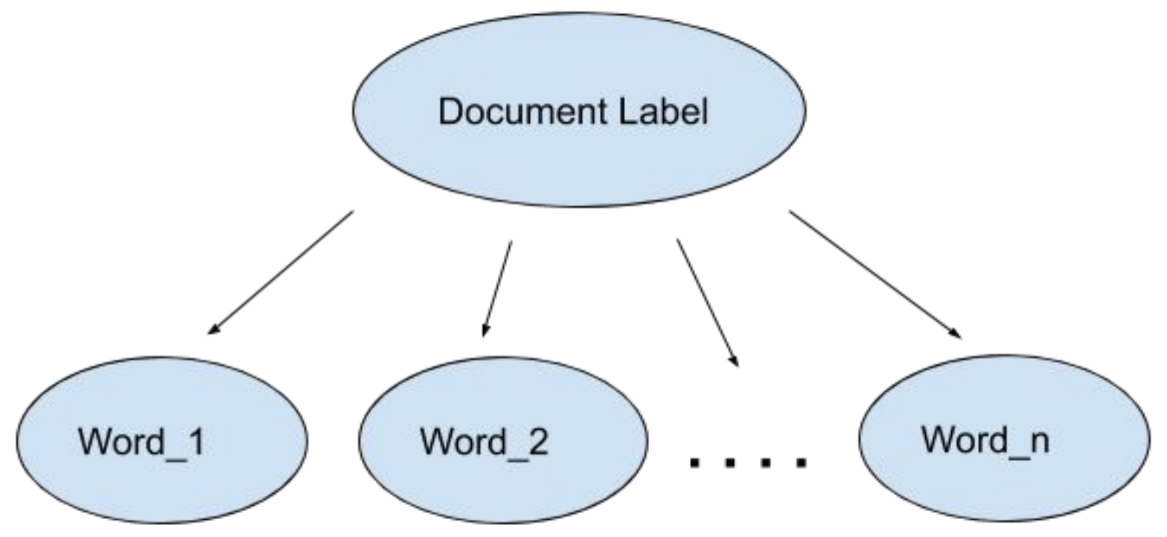
	Baseline (Yu *et al.*)	BiLSTM (Dropout, Mean reduction)	Ensemble BiLSTM	BiLSTM Regression Model
Best Accuracy	0.51	**0.592**	0.561	0.424

Table 3. Results of Classification Accuracies on BiLSTM. Results were achieved with batch size of 32, 64 hidden neurons, and word embeddings of size 100. Results shown are best out of 20 epochs. Best results were achieved without removing stop words or punctuation.

Discussion

Model 1: Multinomial Naive Bayes

- Baseline model: Multinomial Niave Byaes For subjective News Comment analysis
- Each of word occurrences are expressed as conditional probability distribution
- On the test, tried different expressions - binary count, word count, tf-idf
- Strong independence assumptions reduce performance many features are strongly correlated

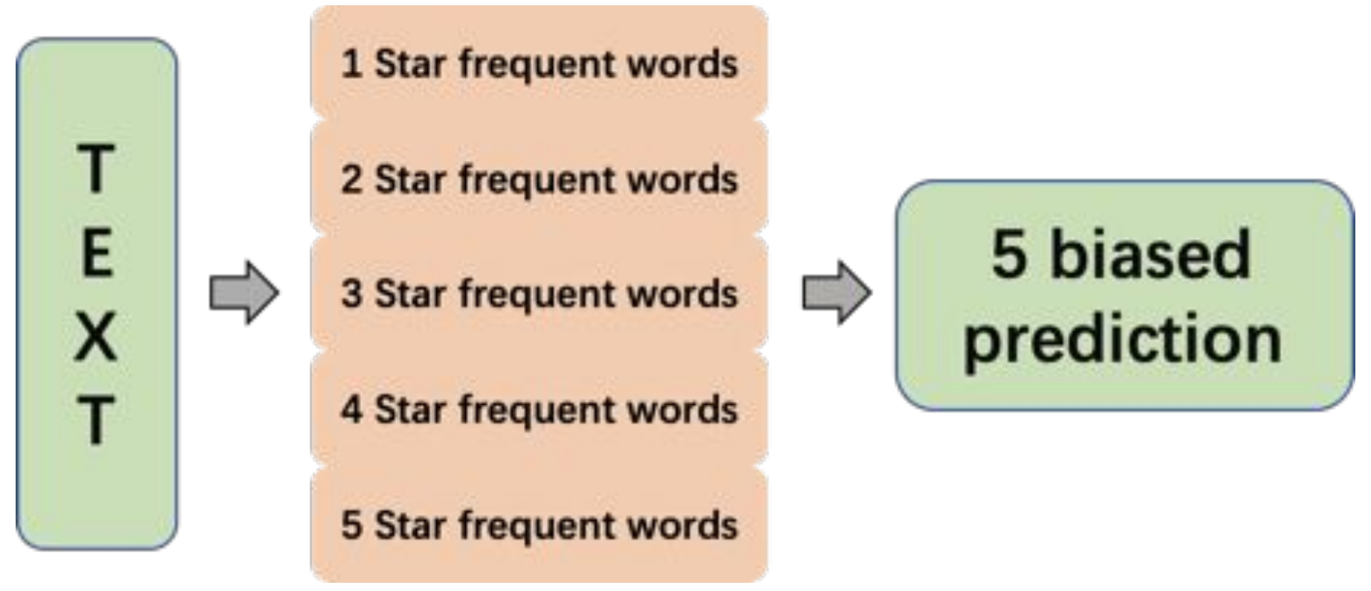


Improvement:

- Experiment on limited number of words by reducing the dimnesionalities
- Try probability smoothing
- Consider making word bag manually which can be regarded as more important

Model 2: Ensemble Logistic Regression

- Baseline model: Multinomial Logsitic regression based on top frequent word-bag
- Five word- bags built upon top frequent words corresponding to each categrory of stars



- Product of prediction probability from the same label was used at the end rather than summation or averge (This allows us to capture the variance in each prediction probability matrix)
- LR based on one-hot encoding was performed as a comparison

Improvement:

- Build LR model with more sophisticated hierachy (e.g. Add another regression model or classifer at the end to combine the result)
- Some words appear to be more important than others in terms of predicting the sentiment, meaning our models can be improved by feature engineering(e.g. Training only with adjective)

Model 3: RNN Models

- Variable Sequence Length RNN with Embedding Layer, 1x1 Conv Layer, 1 BiLSTM Layer, Linear Scoring Layer
- Vocabulary size is reduced and inputs are encoded in terms of indices of the new vocabulary (empirically found best results with 5000 words)
- Experimented with different methods to reduce sequence length dimension (last output only, mean reduction, label application at every timestep)
- Ideas:
  - Variable Sequence Length BiLSTM with mean reduction across sequence length dimension. Yielded best results. BiLSTM makes each timestep’s output more informative which makes use of mean reduction reasonable.
  - Ensemble method with various features injected in to scoring network post-LSTM (sentiment score, usefulness score, number of characters, number of words).
  - RNN Regression on last output using MSELoss with and without Tanh layer to scale outputs.

Improvement:

- Experiment with use of other secondary features or more informative use of secondary features
- Fixed length model, removal of mean reduction may help in making stronger weight updates
- Implement attention using Pyramidal BiLSTMs to capture context level sentiment