Sentiment Analysis on Business Reviews with LSTM Attention and Multi-Task Learning

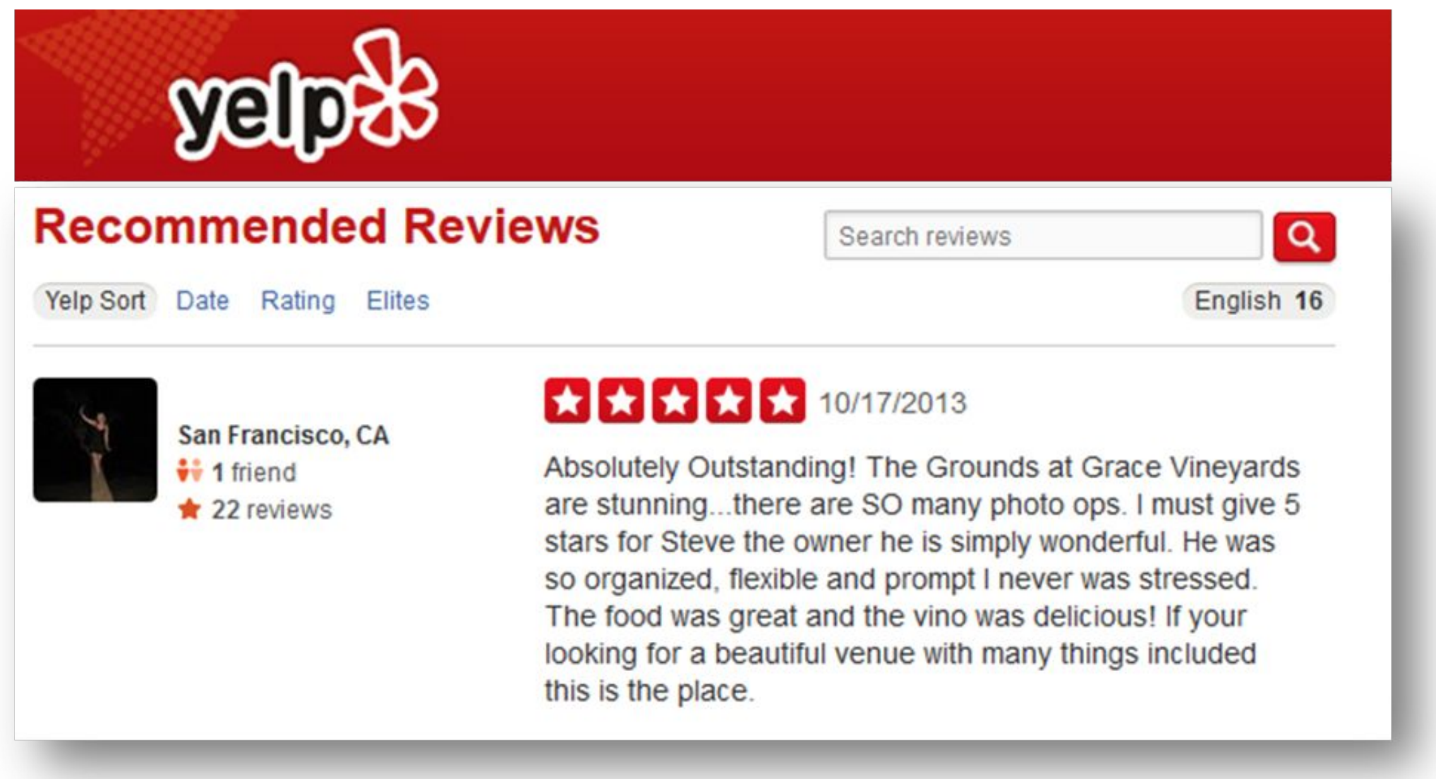
Mingyang Ling - *mingyanl@stanford.edu*

Stanford

Video URL: https://www.dropbox.com/s/kbj98187n8eosws/poster.mov?dl=0

Introduction

Problem: Platforms like **Yelp** maintains a review forum by crowd sourcing, aiming to help users find businesses through descriptions, textual **comments** and **ratings**.



Recommended Reviews

Search reviews

Yelp Sort Date Rating Elites English 16

San Francisco, CA
1 friend
22 reviews

★★★★★ 10/17/2013

Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place.

Since each review's rating is important in calculating the average rating of the business, it is important that contents in reviews and final ratings are consistent. By automatically analyzing the sentiment of textual reviews, ratings can be auto generated and recommended to reviewers, ensuring the ratings are coherent with the comments.

Project goal: we analyze reviews' sentiment or the satisfaction level of a customer towards a business. We formulate multiple models to learn and predict star rating with review comments, and improve accuracy with different methods such as LSTM sentence feature extraction, attention mechanism, and multitask learning.

Data and Preprocessing

We collected Yelp Dataset (https://www.yelp.com/dataset) including reviews, business and user information. As the main focus of our system, we used reviews as input, and star rating as output. Also, we incorporated other information such as business category for multitask learning.

- We associated 1-star and 2-star reviews with a negative sentiment, 4-star and 5-star reviews with a positive sentiment, and 3-star reviews with a neutral sentiment.
- We sampled 100k samples randomly for each of positive, negative and neutral rating, and splitted into train/eval/test.
- We tokenized each review, removed punctuations and the english stop words outlined in the Natural Language Toolkit.
- We obtained the root business category for each review by recursively traversing the category tree which narrows down the number of categories to predict from over 300 to 22.

LSTM Sentence Model with Attention

Word Feature Models

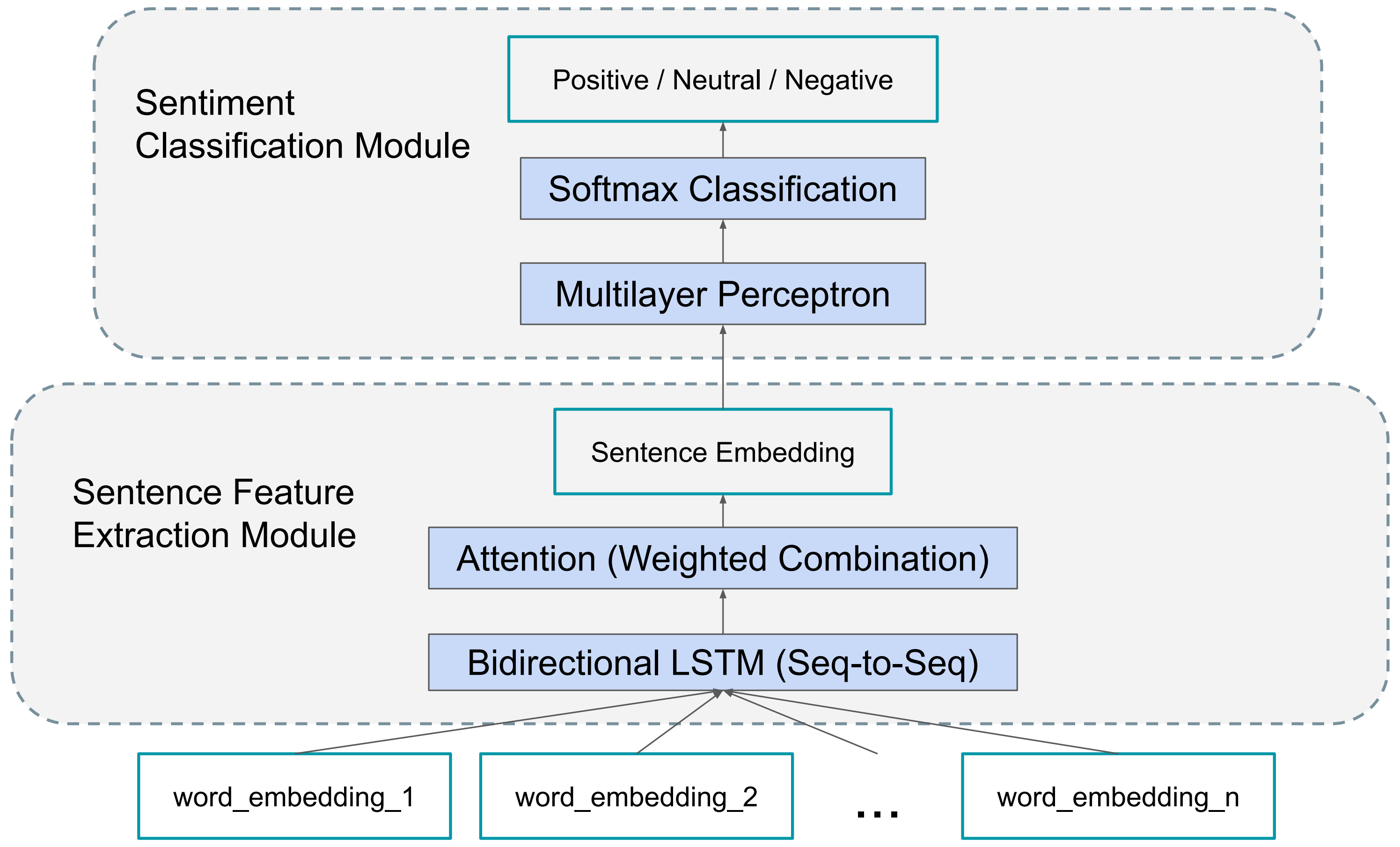
- We created **bag-of-words** feature vectors as our baseline.
- To captures semantic information and able to generalize to unseen words, we adopted **GoogleNews word2vec** model.

Sentence Feature Models

- We **averaged** the word embeddings as sentence feature baseline.
- To capture long term dependencies among the words in reviews, we introduced bidirectional **LSTM** sentence classifier.
- To highlight the importances of each word in sentiment analysis, we leveraged **Attention** mechanism to learn the distribution.

Sentiment Classification Model

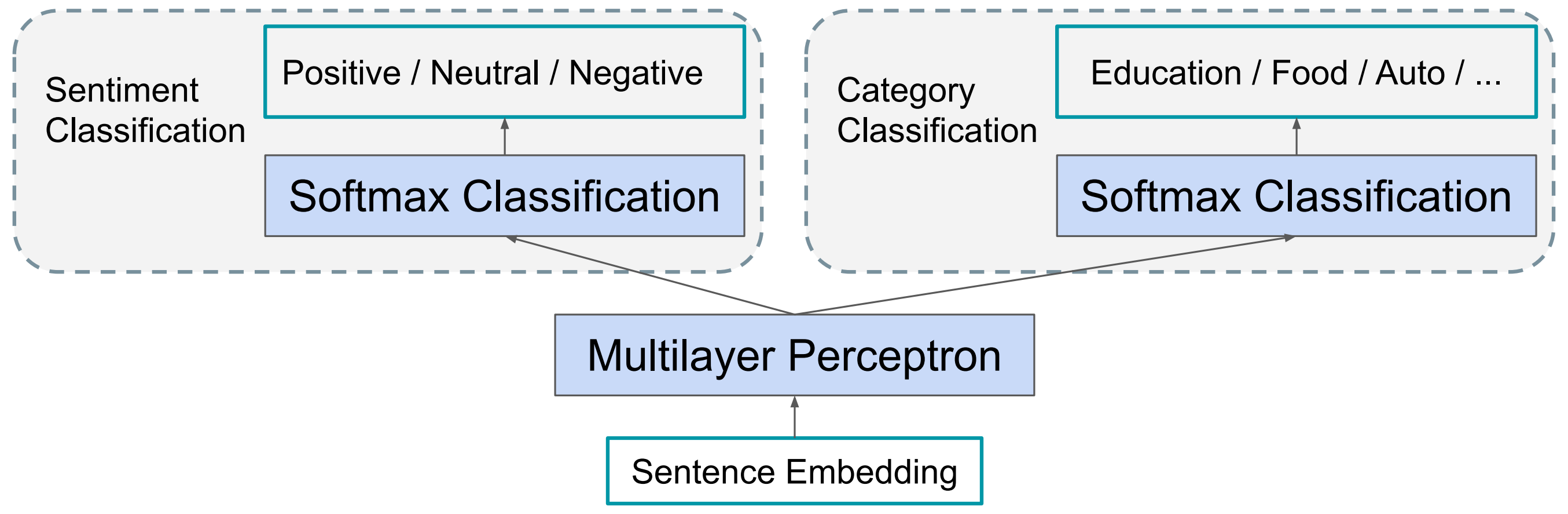
- We combined multilayer perceptron and softmax cross-entropy loss.



Sentiment Classification Module
Positive / Neutral / Negative
Softmax Classification
Multilayer Perceptron
Sentence Feature Extraction Module
Sentence Embedding
Attention (Weighted Combination)
Bidirectional LSTM (Seq-to-Seq)
word_embedding_1
word_embedding_2
...
word_embedding_n

Multitask Learning

We expect other information could improve sentiment classification. Especially for **business category**, it contains the context of review. Metaphorically, we analyze review condition on certain category so that it reduces the complexity of review recognition. Rather than appending one-hot category feature which doesn't work well, we leverage multitask learning to inject the category information. Also, it reduces the memory cost by sharing sentence model and MLP features.



Results and Discussions

	Precision	Recall	Accuracy	F1
Baseline	0.66	0.67	0.67	0.66
Avg w2v	0.73	0.72	0.72	0.73
LSTM	0.77	0.78	0.77	0.78
LSTM + Attention	0.79	0.79	0.79	0.79
LSTM + Attention + Multitask	0.8	0.8	0.8	0.8

Table 1: Comparison of models. We select label classes with highest score from predictions, and evaluate against ground truth.

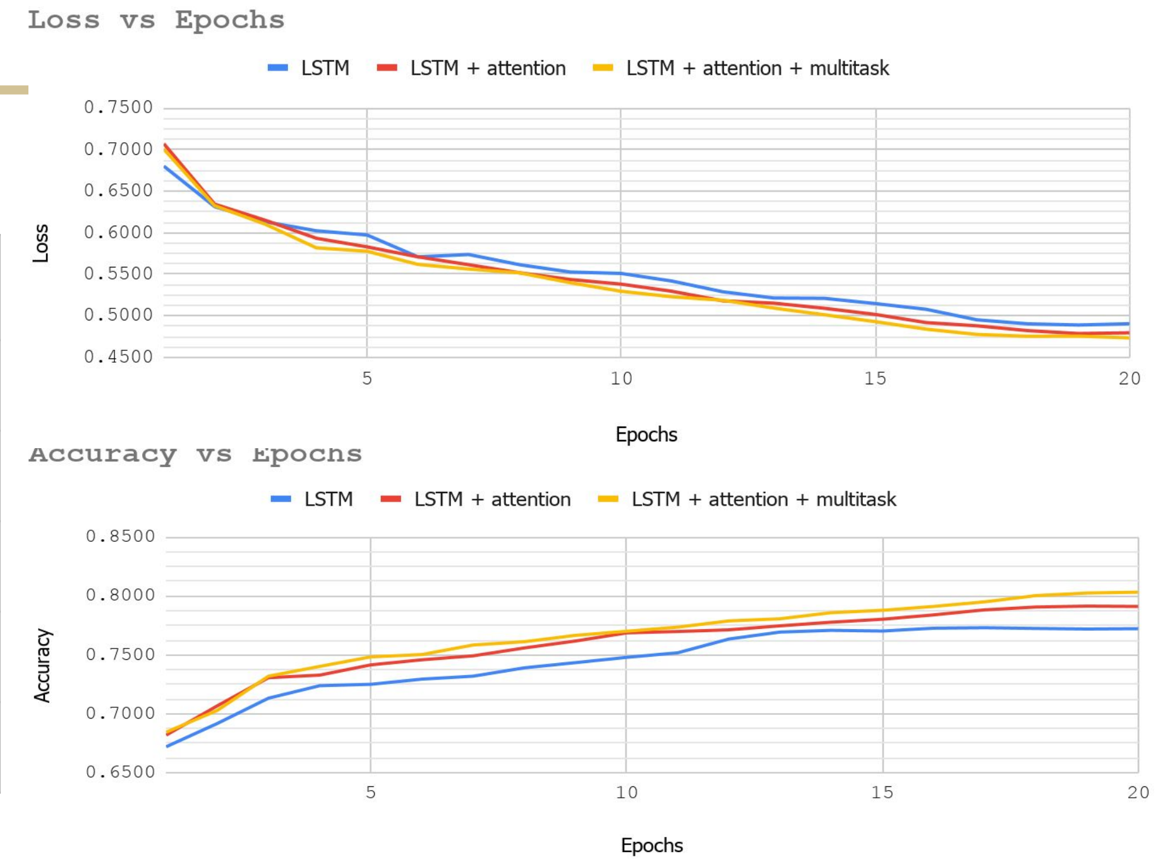


FIgure 1: Learning curves with mse loss and accuracy in terms of epochs

Based on the model comparison, we find there are multiple machine learning techniques able to gradually improve the accuracy of sentiment level classification. Specifically,

- Word2vec is better than bag-of-words given its generalization and semantic information.
- On top of GoogleNews word2vec, LSTM can boost the performance by encapsulating sentence level information.
- The attention mechanism introduces importance distribution on LSTM features to weight more on sentiment related words.
- In addition, multitask learning will incorporate category information so that the sentiment classification has prior knowledge on the business categories.

Conclusion and Future Work

To conclude our report of progress made in our sentiment analysis exploration, we have investigated various methods in tackling such problem. We collected and processed dataset from Yelp, and explored two methods of creating word embeddings for each review. Next, we established six workflows that utilize the neural network techniques to understand reviews. We also examine the performance of these workflows and achieved the best performance with w2v, LSTM, Attention, and Multitask.

We plan to explore creating even more high level embeddings at a paragraph level to handle reviews with long paragraph. As another potential next step, we could evaluate performance on certain categories to analysis the effect of multitask learning.

References

[1] Asghar, Nabiha. "Yelp dataset challenge: Review rating prediction." arXiv preprint arXiv:1605.05362 (2016).
[2] Kiritchenko, Svetlana, et al. "NRC-Canada-2014: Detecting aspects and sentiment in customer reviews." Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014.
[3] Palangi, Hamid, et al. "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24.4 (2016): 694-707.
[4] Tang, Duyu, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural network for sentiment classification." Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.
[5] Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp's Ratings Based on Text Reviews." (2014): 1-5.
[6] Yu, Boya, et al. "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews." arXiv preprint arXiv:1709.08698 (2017).