Fake Review Detection on Yelp Restaurant Data

Ganapathy Sankararaman*(ganasank)*, Khaled Aounallah*(kaounall)*, Ferhat Turker Celepcikay*(turker)*

CS229 FINAL PROJECT PRESENTATION, STANFORD UNIVERSITY

Motivation & Objectives

- Online product reviews are a fundamental part of the decision-making process for customers in e-commerce. However, since people can freely write their own contents, their opinions might not always be accurate and unbiased.
- Implications of such fake reviews can affect both the customers and the servicers and can pose a significant challenge in the e-commerce industry.
- Our goal in this project is to predict if a given review is fake or real, based on the content from Yelp Restaurants review data for New York City.

Data Set

- Yelp Restaurants Review Dataset was provided by the authors of the "Collective Opinion Spam Detection: Bridging Review Networks and Metadata".
- 11124 training reviews from year 2011, 2000 validation reviews from 2012 and 10000 testing reviews from 2013. All the three sets have 50% fake reviews.
- Each review has the review content, user id, product id, date of the review, and the rating given by the user to the product.

Features and Feature Engineering

- **5 Raw features**: user id, product id, rating, date of review, review comment
- **16 Derived features:**

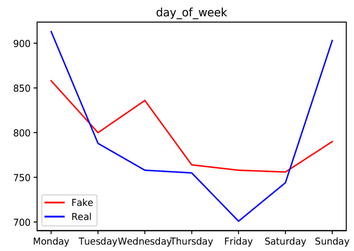
i. *4 Reviewer Centric:* No. of reviews given by user, average rating given by user, Average no. of words in reviews by user, Max reviews given by user in each day

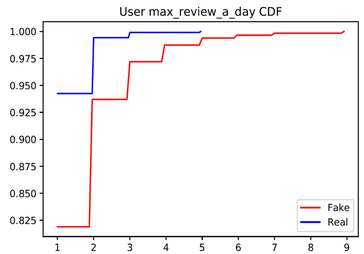
ii. *4 Product Centric:* no of reviews for restaurant, average rating for restaurant, average no. of words for restaurant review, max reviews received by restaurant in a day

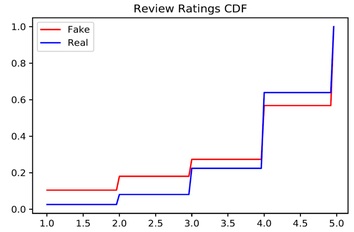
iii. *8 Review Centric:* Character count, word count, word density, punctuation count, upper case word count, top 1000 unigrams and bigrams that occur more frequently in fake reviews than real reviews, top 1000 unigrams and bigrams that occur more frequently in real reviews than fake reviews

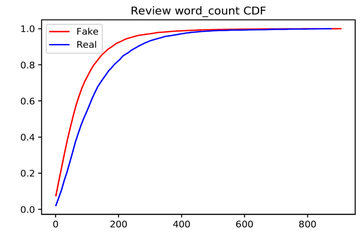
- The derived features give more information about the fake reviewer, and the restaurant that receives the fake reviews, and helps us to classify them better.
- Our primary goal is to achieve the classification with just the review and look at how the metadata performance.

Feature and Data Visualization



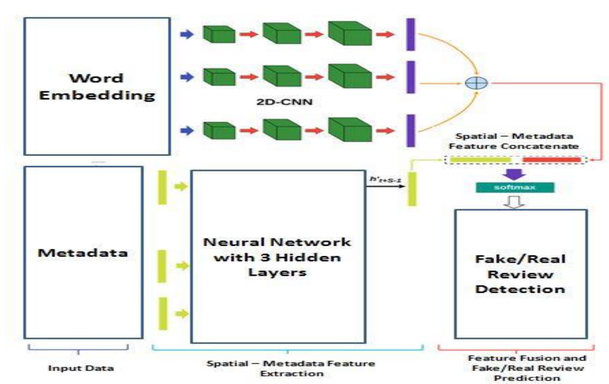






Models and Methods

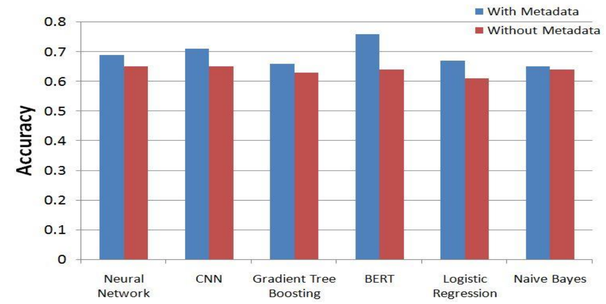
- **Naive Bayes:** A generative model. Generate TF-IDF of words and use it to classify a review as fake or real.
- **Logistic Regression:** also uses TF-IDF of words, minimize loss to achieve linear separation between fake and real reviews.
- **Neural Networks:** Two models: one model with 3 hidden layers and another model with 6 hidden layers - 3 for reviews and 3 for metadata. All hidden layers use ReLu activation and use softmax for classification. All hidden layers are L2 regularized with a 50% dropout rate at before the output layer.
- **Convolutional Neural Network:** two models both having 128 convolution 2D filters for 5 different filter sizes (1-5), and have ReLu activation, followed by max pooling. The convolution layer is L2 regularized and has ReLu activation and normal initialization for the kernels. The second model has an extra neural network with 3 hidden layers for the metadata, all of which have ReLu activation. The output layer uses softmax and there is a 50% dropout before the output.
- **Gradient Boosting Tree:** combines an ensemble of weak prediction learners into a single strong learner through iteration. An extensive parameter tuning was performed through varying 4 key parameters (learning rate, maximum features, number of estimators, max depth)
- **BERT (Bidirectional Encoder Representations from Transformers):** a recent paper published by researchers at Google AI Language. BERT's key technical innovation is applying the bidirectional training of Transformer.



Results

Model	Accuracy		AUROC		F1 Score	
	Train	Test	Train	Test	Train	Test
Neural Network	0.73	0.69	0.80	0.75	0.75	0.72
CNN	0.77	0.71	0.84	0.77	0.78	0.74
Gradient Boosting	0.74	0.66	0.83	0.71	0.74	0.68
BERT	0.98	0.76	0.98	0.76	0.98	0.76
Logistic Regression	0.69	0.67	0.76	0.73	0.71	0.69
Naive Bayes	0.71	0.65	0.77	0.71	0.72	0.68

Discussion and Error Analysis



- For Error Analysis, we individually (3 members) tried classifying 20 misclassified reviews by worse margin from BERT. We took the classification that was most chosen by the three of us and compared it with BERT's prediction. 50% of what BERT predicted was correct. This might be due to an accumulation of error due the misclassification of reviews by YELP's filtering algorithm.
- Additionally, with bootstrapping of test set with sample size of 2000, we found that the standard error is a bit more than 1%. So this could have contributed to the lower accuracy in testing dataset as well.

Future Work

- We will work on fine tuning the hyperparameters on CNN with grid search and try to successfully implement a bidirectional LSTM.
- Bidirectional LSTM is already implemented but would be interesting to tune the hyperparameters for the model.
- Gradient Boosting Tree has more potential for improvement, and with better computational resources, we should be able to increase the complexity of the model by expanding parameters such as the maximum depth of the tree.

References

1. Rayana, Shebuti, and Leman Akoglu. "Collective opinion spam detection: Bridging review networks and metadata." Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. ACM, 2015.
2. Ahmed, Hadeer, Issa Traore, and Sherif Saad. "Detecting opinion spams and fake news using text classification." Security and Privacy 1.1 (2018): e9.
3. Yoon Kim. "Convolutional Neural Networks for Sentence Classification."(2014)
4. Shervin Minaee, Elhaam Azimi, Amir Ali Abdolrashidi. "Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models." (2019)