Yu Heng (Henry) WU   **Supervisor: Dr. Farid Shirazi**

Toronto Metropolitan University — Data Science & Analytics, Graduate Studies

Objective: Sentiment analysis: Systematically categorizing people's feelings/opinions from textual data, most often distinguishing between positive, negative and neutral statements. This research seeks to compare traditional machine learning algorithms with the BERT model in sentiment analysis on the Yelp dataset in categorizing positive/negative opinions and predicting specific 5 star ratings of a review.

Background: Online reviews have become a central component in driving consumer choices one online review could influence the decision of countless other consumers. Through the advent of computers (and machine learning) it became possible to analyze the massive amounts of reviews online.

Machine learning models like Naive Bayes and SVMs dominated early sentiment analysis. But Neural networks began to reshape sentiment analysis, then Transformers architectures were introduced, leading to development of BERT, GPT and variants. With the recent gold standard being pre-trained transformer models, fine-tuned on specific sentiment analysis tasks.

Methodology:
Undersampling 5-star reviews in the data to remove bias - trained on 500'000 reviews of each star

Number of reviews by star
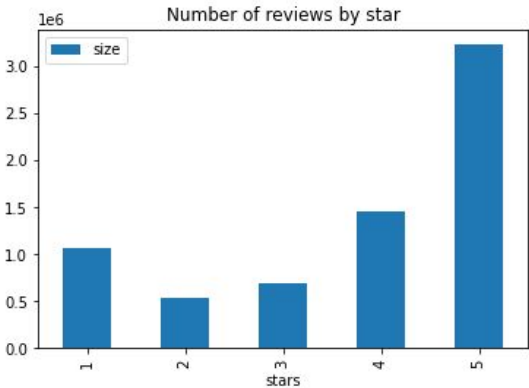
Machine learning models (Naive bayes, Logistic Regression, Random Forest, Support Vector Machines) trained on TF-IDF Vectorized text reviews

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) \qquad idf(t) = \log \frac{n}{1 + df(t)}$$

For BERT - the entire review is input as training

2 sets of classifiers - one for positive/negative (negative was 3 or less stars, positive was 4 or 5 star rating), the 2nd for 5-star classification

Both sets of classifiers were evaluated on testing data made up of 500'000 randomly selected reviews

Results:

Accuracy for test data for 5-star rating

Naive Bayes	Logistic Regression	Random Forests	SVMs	BERT
0.556568	0.650252	0.615466	0.572264	0.658952

Accuracy for test data on positive/negative sentiment

Naive Bayes	Logistic Regression	Random Forests	SVMs	BERT
0.82453	0.908618	0.810738	0.813022	1.0

The models performed much better on positive/negative sentiment classification compared to 5-star rating classification, with BERT having a notable 100% accuracy in its predictions.

Conclusions:
This study applied both traditional machine learning and deep learning model BERT, to conduct sentiment analysis on Yelp reviews. Traditional models performed well for extreme sentiments (1-star and 5-star reviews), but struggled with nuanced sentiments (2-4 star reviews), whereas BERT exhibited exceptional accuracy in both sentiment classification and nuanced rating prediction. For future research, exploring other datasets like travel blogs or e-commerce reviews, as well as trying different data handling strategies and newer deep learning architectures like RoBERTa, XLNet, or DistilBERT.