KOREAN MOVIE REVIEW SENTIMENT PREDICTION

Sooeun Oh | ANLY 580 - NLP | Georgetown University M.S. Analytics

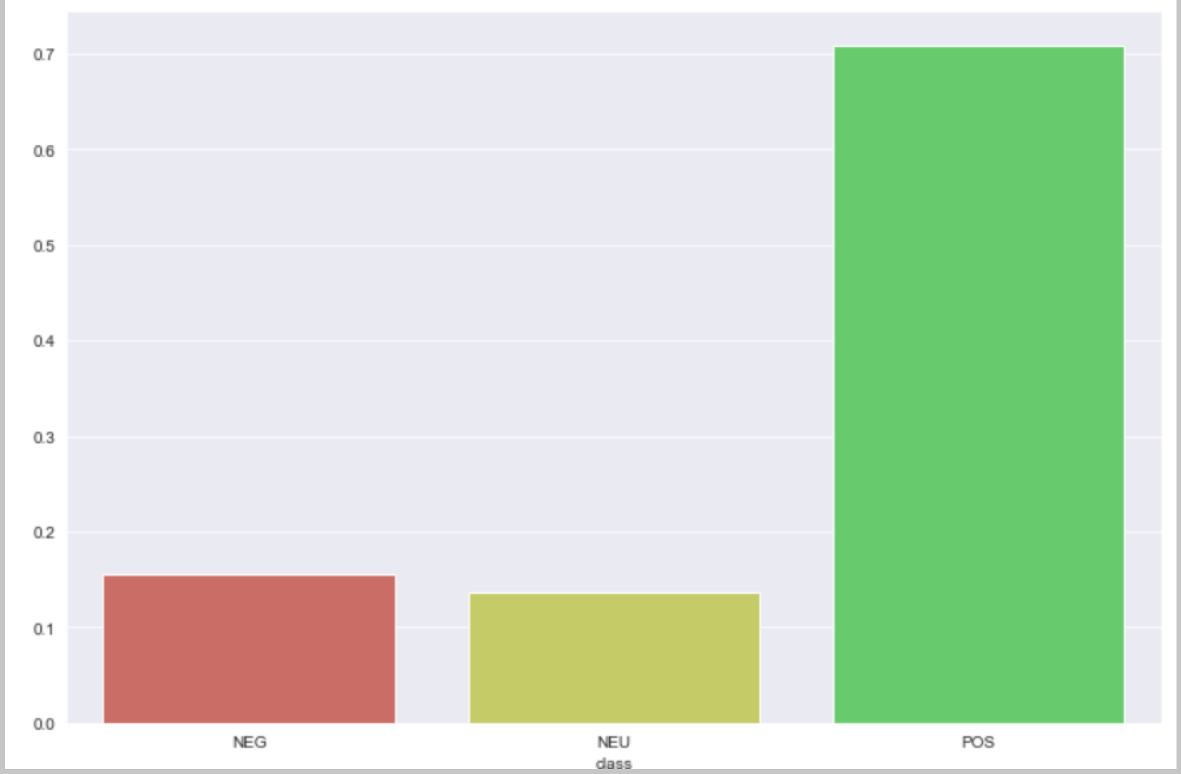
Project Overview

While there have been many successful attempts at sentiment analysis and prediction of English text-based content, fewer attempts have been made to classify text in Korean.

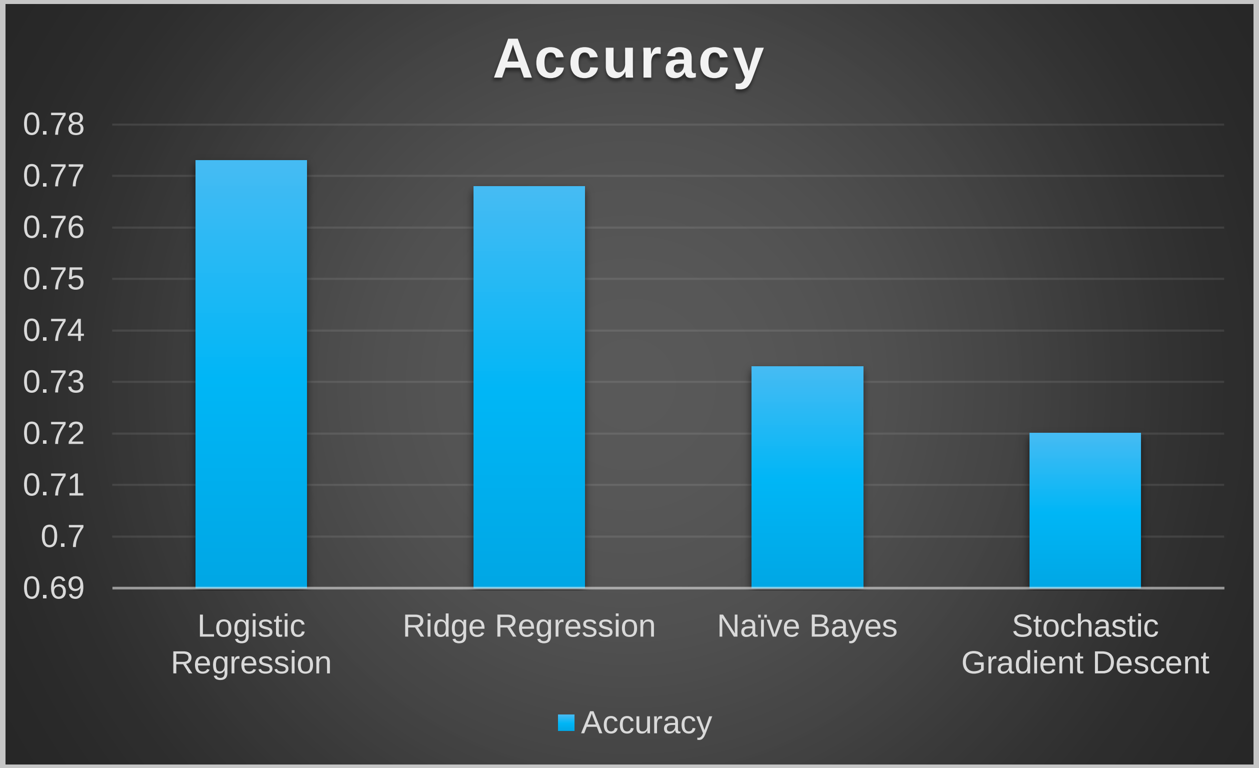
The research project aims to <u>develop a model using machine learning techniques to predict the rating of movie review.</u>

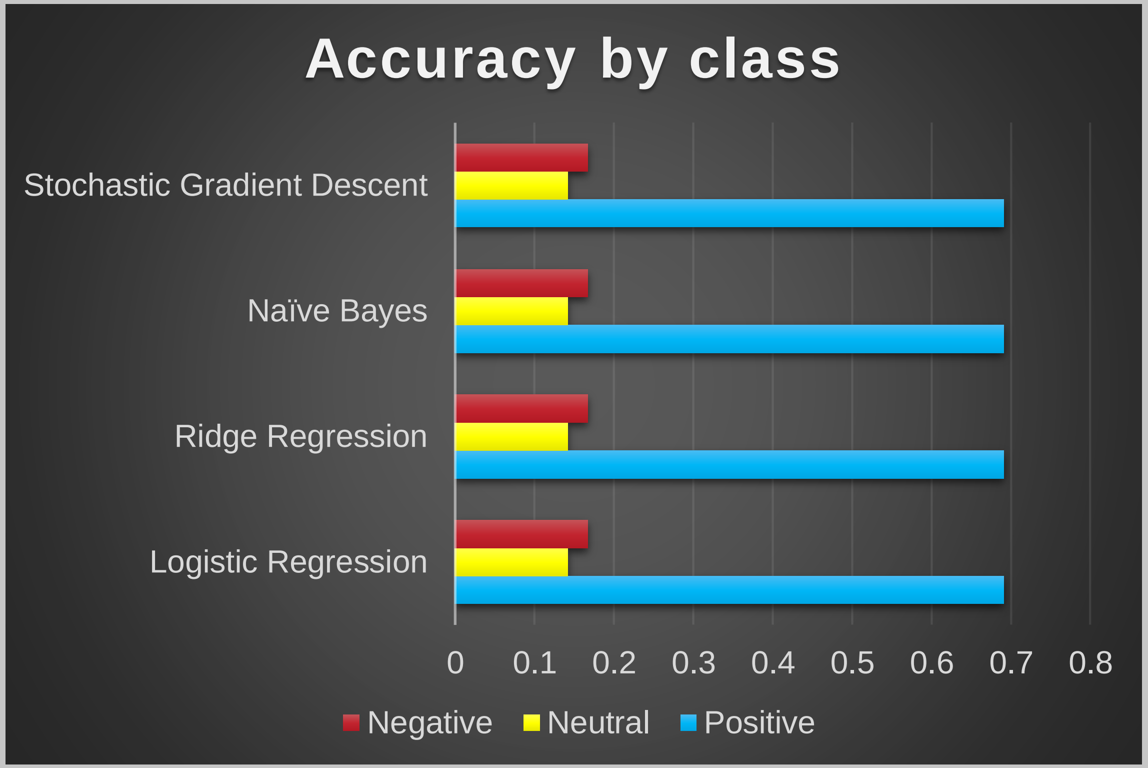
Method

- Collected 9,220 Korean movie reviews by scrapping via Beautiful Soup
- Pre-processed the data
  - **Data-cleaning**: only kept review, rating
  - **New feature**: 3 classes (NEG, NEU, POS)
  - **Tokenization**: KoNLPy
  - Defined **stop words**
  - **Vectorization**: TfidfVectorizer
  - Classifiers
    - Logistic Regression
    - Ridge Regression
    - Naïve Bayes
    - Stochastic Gradient Descent



Results


Accuracy
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
0.69
Logistic Regression
Ridge Regression
Naïve Bayes
Stochastic Gradient Descent
Accuracy


Accuracy by class
Stochastic Gradient Descent
Naïve Bayes
Ridge Regression
Logistic Regression
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Negative
Neutral
Positive

Discussion

- **Findings**
  - Regardless of classifiers, positive reviews have better prediction due to an enough training dataset
  - Though the accuracy may be heavily relying on the training dataset, the research is meaningful as one of a few researches in Korean Natural Language Processing.
- **Key issues**
  - Lack of data – web scrapping did not allow more than 9,200 sample data
  - Skewed distribution of data – relatively more positive ratings
  - Need a better, refined stop word