Stony Brook University

Amazon Product Reviews Analysis based on NLP

AMS 561 Project Presentation

Zhilin Yang, MS in AMS (STAT)

FAR BEYOND

Overall idea and significance

- The objective of this project is to help me learn how to use NLP techniques to deal with text data and doing sentiment analysis.

- The significance of this project lies in its application of NLP techniques to extract valuable insights from unstructured textual data. By using various feature extraction techniques and machine learning algorithms, we were able to perform sentiment analysis on a large dataset of reviews and accurately classify them as reviews' score 1 to 5.

Expertise level and what I learned during the project

- Finished this project on my own.
- AMS graduate student. 2 years data analysis experience. Never dealt with text data before.
- What I learned
  - Text cleaning and preprocessing
  - Feature extraction for text data
  - Sentiment analysis

Techniques and tools used

- Text Preprocessing
  - Expand contractions
  - Replace abbreviations
  - Remove HTML tags, URLs and special characters
- Feature Extraction
  - Tokenization
  - Stop word removal
  - Stemming
  - Bag-of- words representation
  - TF-IDF representation
- Sentiment Analysis
  - Logistic Regression
  - Support vector machines
  - Naive Bayes
  - LSTM

Stony Brook University

Results and Conclusions

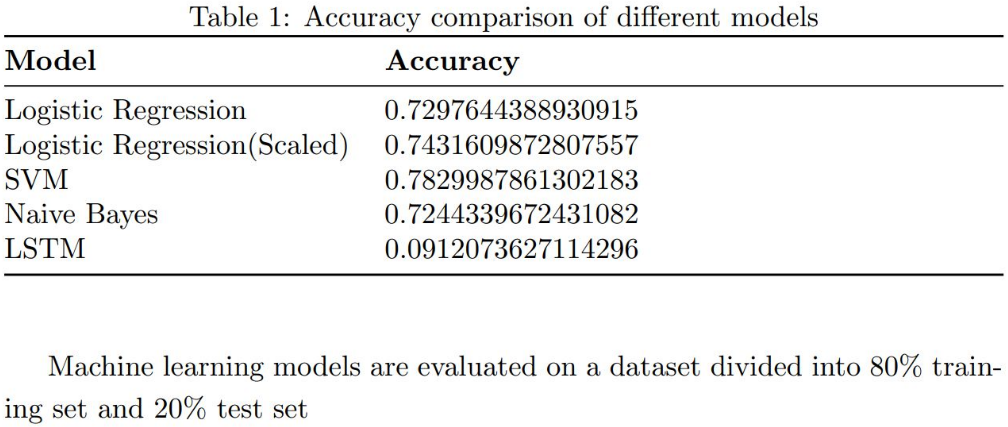


Table 1: Accuracy comparison of different models

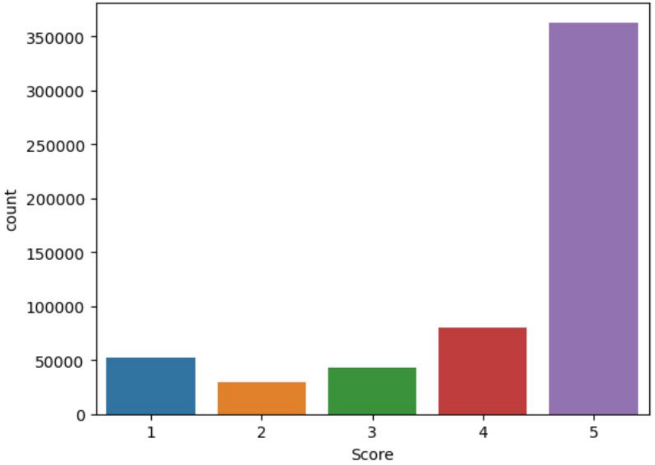
Model	**Accuracy**
Logistic Regression	0.7297644388930915
Logistic Regression(Scaled)	0.7431609872807557
SVM	0.7829987861302183
Naive Bayes	0.7244339672431082
LSTM	0.0912073627114296

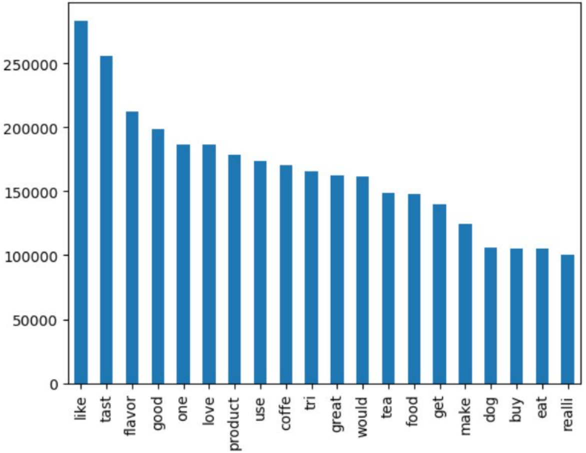
Machine learning models are evaluated on a dataset divided into 80% training set and 20% test set

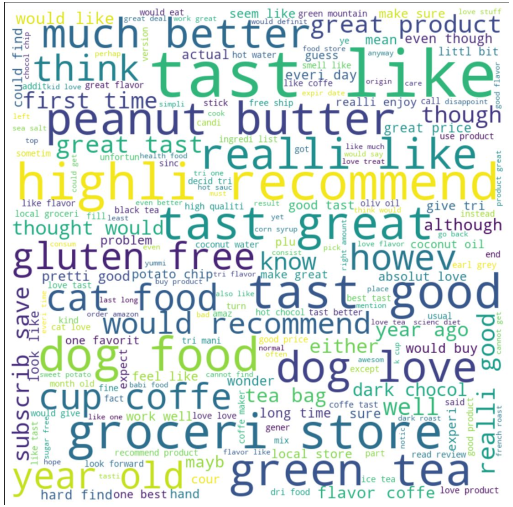
- SVM performed the highest accuracy
- The LSTM model did not perform as well as the other models, which may indicate that it is not the most suitable choice for sentiment analysis of short textual data.
- Scaled data performed better for logistic regression
- NLP techniques can help extract valuable insights from unstructured textual data

FAR
BEYOND

Pics and graphics







Challenges and Deficiencies

- Only focused on sentiment analysis and did not consider other aspects of NLP, such as named entity recognition, part-of-speech tagging, and text summarization.
- Did not explore the use of unsupervised learning techniques, such as clustering and topic modeling, which could provide additional insights into the underlying structure of the reviews
- Did not evaluate the robustness of the models to adversarial attacks, which could be important in real-world applications.
- Although demonstrated the potential of deep learning models for sentiment analysis, the LSTM model did not perform as well as expected, indicating that there may be limitations to the use of such models for short textual data.

Stony Brook University

Thank You

FAR
BEYOND