Sentiment Analysis of Amazon Reviews

Connor Galvin

Opportunity
- Can I build a model that takes in the text of a product review and predicts the sentiment? (positive or negative)
- Machine learning model to be implemented with Python
- Possible Difficulties: Sentiment is subjective and even humans will disagree with each other
- Language is complex: certain words can convey positive and negative sentiment and it is difficult to analyze irony and sarcasm

Impact
- Insights from these reviews can be used to manage product catalogs and alter marketing strategy to maximize sales and revenues
- Sentiment analysis can be used on other categories of reviews (movies, TV, Employers)
- Can be used to gauge large scale public reactions to media events

Approach
- Randomly sample the data and pick 30,000 reviews to analyze for computational simplicity
- Run term frequency, inverse document frequency, and cosine similarity analysis on individual reviews
- Remove filler words that do not convey sentiment (and, the, but, to)
- Assign sentiment value to individual words
- Assign sentiment value to multiword phrases and negations
- Build a Naive Bayes model to classify a review as positive or negative
- The goal is to obtain around 80% accuracy, as this generally the human agreement on sentiment

Data/Results
- Dataset of reviews written in English for a large variety of products on Amazon.com
- The data was obtained from Kaggle, including a training dataset of 3.6 million reviews and testing data of 400,000 reviews
- Each review is either labeled as positive (4 or 5 star review) or negative (1 or 2 star review)

Acknowledgment: EECE 2300 Professor Joseph Robinson