CS6140 - Final Project Proposal

Team members:

1. Gmon Kuzhiyanikkal
2. Mahvash Maghrabi
3. Sruti Keerti Munukutla

Data Source: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

This dataset consists of reviews of products from Amazon marketplace from 1995 until 2015. It has 130M+ customer reviews written in multiple languages.

The dataset consists of reviews for multiple categories (Kitchen appliances, Books, Mobile Electronics, Furniture, Software, Tools, Sports etc). We decided to compile data from Books, Mobile Electronics and Furniture and perform sentiment analysis on their reviews. Dataset is compiled in a Tab Separated Value format (TSV). It has 15 features and a total of 104855 (Mobile Electronics reviews) + 100000 (Books) + 791702 (Furniture) = 996557 data points.

We are trying to predict the sentiment analysis of the common day products that we buy from online. We are using three different amazon dataset for working on this. The feature vector of the dataset are:

marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	review_date

A similar research has been conducted and published in the following research paper:

https://www.researchgate.net/publication/350973881_Sentiment_Analysis_of_Amazon_Product_Reviews_Using_Hybrid_Rule-based_Approach

We are planning to implement three ML methods. Two classic ML methods and one deep networking method.

1) Decision Trees

2) Logistic Regression

3) Convolution Neural Network

We are still planning to come up with novel ideas. Hoping to come up with something good by the time we start implementing our project.