Capstone NLP Project Proposal

1. The Problem Statement:

E-commerce, also known as electronic commerce or internet commerce, refers to the buying and selling of goods or services using the internet, and the transfer of money and data to execute these transactions. Ecommerce is often used to refer to the sale of physical products online. There are two companies in e-commerce; one is the online retailer, which wants to maximize e-commerce sales, and the other is the consumers, who want to buy the best product over alternatives.

Customer reviews come to play a huge role in this life cycle of e-commerce. The success of online marketing has quickly become dependent on consumer reviews. According to a study;

a. 70% of customers consult reviews or ratings before making a final purchase.

b. 63% of consumers are more likely to purchase from a site if it has product ratings and reviews.

c. 67% of consumers read 6 reviews or less before they feel they can trust a business enough to make a purchase.

d. 80% of consumers have changed their mind about purchases based on negative information they have found online.

As it might be seen above, user reviews are proven sales drivers. For that reason, online retailers should assess those reviews and ratings as a precious data pool and make smart decisions about their customers' perception and products. Based on the data they have, online retailers' mind will be clearer about which products are most consumed, what problem areas they have in their low consumed products, which consumer has tendency to buy their products, how can they improve their sales, etc.

2. Who is your client and why do they care about this problem?

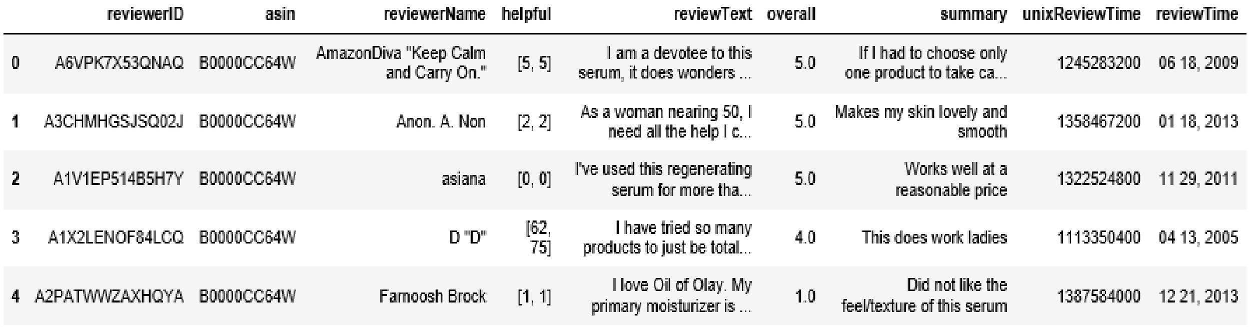
The largest internet retailer in the world as measured by revenue and market capitalization in Amazon.com, Inc. which is an American e-commerce and cloud computing company. It became the fourth most valuable public company in the world (behind only Apple, Alphabet, and Microsoft), the largest Internet company by revenue in the world, and after Walmart the second largest employer in the United States.

As mentioned above, Amazon is my client in this project, and I will focus on the beauty products' review. I will use sentiment analysis which is most general use in e-commerce activities. Websites allow their users to submit their experience about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. Customers can easily view opinions and recommendation information on whole product as well as specific product features. Graphical summary of the overall product and its features is presented to users. Popular merchant websites like amazon.com provides review from editors and customers with rating information. Sentiment analysis helps such websites by getting a current general idea of the customers about a specific product, using that idea in decision making process about the future of specific products, extracting the complaining points from negative reviews, and powerful points from positive reviews, and converting dissatisfied customers into promoters by analyzing this huge volume of opinions.

3. Description of the Data Set:

Beauty dataset revolving around the reviews written by customers. This is a real commercial data.

This data includes 28798 rows and 9 feature variables. Memory usage is 2.2+ MB.



	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A6VPK7X53QNAQ	B0000CC64W	AmazonDiva "Keep Calm and Carry On."	[5, 5]	I am a devotee to this serum, it does wonders ...	5.0	If I had to choose only one product to take ca...	1245283200	06 18, 2009
1	A3CHMHGSJSQ02J	B0000CC64W	Anon. A. Non	[2, 2]	As a woman nearing 50, I need all the help I c...	5.0	Makes my skin lovely and smooth	1358467200	01 18, 2013
2	A1V1EP514B5H7Y	B0000CC64W	asiana	[0, 0]	I've used this regenerating serum for more tha...	5.0	Works well at a reasonable price	1322524800	11 29, 2011
3	A1X2LENOF84LCQ	B0000CC64W	D "D"	[62, 75]	I have tried so many products to just be total...	4.0	This does work ladies	1113350400	04 13, 2005
4	A2PATWWZAXHQYA	B0000CC64W	Farnoosh Brock	[1, 1]	I love Oil of Olay. My primary moisturizer is ...	1.0	Did not like the feel/texture of this serum	1387584000	12 21, 2013

Each row corresponds to a customer review, and includes the variables:

reviewerID : ID of the reviewer, e.g. A2SUAM1J3GNN3B - type: object
asin : ID of the product , e.g. 0000013714 – type: object
reviewerName : name of the reviewer – type: object
helpful : helpfulness of the review, e.g. 2/3 – type: object
reviewText : text of the review – type: object
overall : Rating – type: float64
summary : summary of the review – type: object
unixReviewTime : time of the review (unix time) – type: int64
reviewTime : time of the review (raw) – type: object

I will download the beauty product review file via link below and open the file with coding in Jupyter notebook.

Data Source:
http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_10.json.gz

4. Approach to solving the problem

I will approach this NLP Sentiment Analysis project by following the steps below:

a. Understand the business problem
b. Create a repository
c. Gather the data from Amazon review link and load it into Jupyter notebook.
d. Analyze the data to determine the data quality
e. Preprocessing
   (1) Data Set Basic Formatting
   (2) Missing Values

(3) Cleaning the text feature
(4) Creating a new column consists of the classification of the ratings

f. Data Storytelling (tableau, word cloud, ….)
g. Hypothesis testing
h. Apply feature extraction and NLP techniques
i. Selecting Evaluation Metric
j. Modeling
k. Selecting Best Model
l. Prepare a report

5. Project Deliverables

My deliverables will be a milestone report, a PowerPoint presentation, and a Jupyter notebook associated with my project.