# **HW Assignment 3 (A3): Sentiment Analysis**

# CS6120: Natural Language Processing Spring 2024

Due Date: Tuesday, \_\_\_\_\_\_\_

**(10 points)**

# Amazon Ratings vs. Amazon Reviews: What's the Difference? | 1440.io

# Scenario

Sentiment analysis or opinion mining focuses on identifying and categorizing opinions expressed in text toward a particular topic, product, or service is positive, negative, or neutral. This technology is widely used in monitoring and analyzing customer feedback, market research, and social media monitoring to understand consumer sentiment.

Amazon reviews are a rich data source for sentiment analysis because they consist of a textual review and a star rating, typically on a scale from 1 to 5 stars, which clearly indicates the customer's sentiment towards the product. These reviews are written by customers who have purchased and used the products, offering their opinions, experiences, and satisfaction levels. Sentiment analysis on Amazon reviews involves processing and analyzing these texts to extract insights about general sentiment, specific features of the product that customers liked or disliked, and overall customer satisfaction. Such analysis helps businesses improve products, address customer concerns, and make strategic decisions.

Use Amazon reviews for one or more categories of products found here:

<https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html>

Clean and preprocess the data including removing irrelevant information, stop words, lower casing and standardizing the text format for analysis. dataset. Apply sentiment analysis techniques to the preprocessed review texts as discussed in the class and starter program. Finally, analyze the results to identify patterns and insights.

# Tasks

**1. Data Preprocessing:**

* Load the dataset and perform initial exploration to understand its structure.
* Clean the text data, including removing special characters, stopwords, applying lowercasing, and other tasks as you deem necessary.

**2. Word2Vec, fasttext embeddings**

* Create 100D, 200D or 300D vectors using both Word2Vec (CBOW and SkipGram separately), and fasttext algorithms
* Average the vectors to create new average vector columns in the df
* Perform EDA to analyze associations between vectors from the three methods above.

**3. Sentiment Analysis:**

* Use Overall rating (1-5 scale) column and convert it to a binary column (1&2 = Negative, 4&5 = Positive, remove 3 category data)
* Perform dimension reduction using techniques of your choice (e.g. PCA, LSA, UMAP, LLE, t\_SNE etc.)
* Use the reduced dimensions to create classification model (sentiment analysis)
* Use proper ML pipeline techniques by splitting the data into test and train and tuning hyperparameters.
* Perform SA using the above setting (Method 1) + Method 2 (CNN, RNN)
* Lastly, apply VADER and TextBlob on your choice of reviews.

**4. Comparison:**

* Compare various methods in 3 and provide reasoning for the most optimal model.
* Analyze SA on the test data and describe the results in detail.
* Look for patterns in the instances where the model went wrong. For example, filter out reviews where the model made False Positive (or False negative) errors and look for patterns.

# **Expected Output**

Please submit a fully executed jupyter notebook identifying question number and steps. Make sure to add comments to your solution.