COSC274

Deliverable 2: Features Plan

Charles Carver, Mingi Jeong, Sam Lensgraf

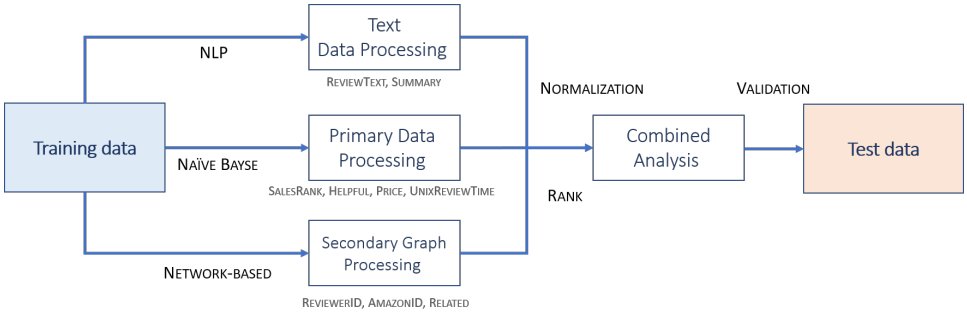

Fig 1. Flow chart of the proposed model for the final project.

1. Text Data Processing

INPUT: REVIEWTEXT, SUMMARY

- Remove stop words from textual fields (REVIEWTEXT and SUMMARY) using NLP.
- Run sentiment analysis on REVIEWTEXT and SUMMARY fields. Give each review a numerical rating (0-2) depending on if it's negative (0), neutral (1), or positive (2).
    - To do so, we can utilize a method like Naïve Bayes to predict whether the review went into different brackets of star ratings: [0-2) being negative,2-3 being neutral, and 4-5 being positive
    - Features which would be useful for this step would be the set of unique non-stop words and the set of non-stop word containing 2 and 3-grams
- This should hopefully convert the qualitative data to quantitative data.

2. Primary Data Processing

INPUT: SALESRANK, HELPFUL, PRICE, UNIXREVIEWTIME

- SALESRANK, HELPFUL and PRICE are already numerical values which we can directly feed to the classifier.
- We could order the reviews based on UNIXREVIEWTIME in case the tone of early reviews influences the tone of later reviews.
- We can consider mapping additional fields, such as ROOT-GENRE or CATEGORIES, to numerical values. After processing, we can feed all numerical data through a Naïve Bayes classifier.

3. Secondary Graph Processing

INPUT: REVIEWERID, AMAZONID, RELATED

- We're wondering if we can apply graph processing to the REVIEWERID, AMAZONID, and RELATED FIELDS to find connections between commonly purchased products.
- We might need to normalize/rank the features in importance given that the range of each variable is different.

4. Combined Analysis

INPUT: OUTPUT OF ABOVE

- Given heterogeneous features and algorithms (e.g., NLP for text, Naïve Bayes for quantitative data, secondary graph processing), we believe there should be some final step to return the star rating. It's possible we could feed it through a second Naïve Bayes classifier, or another algorithm that we'll learn in the future.