Project Proposal

Topic: Amazon Product Reviews Multiclass Classification

Text mining is a vital field that extracts insights from massive textual data, offering crucial tools in the information-rich digital age. Our research aims to harness text mining's capabilities to address a significant research question of identifying the specific category to which the product belongs based on its reviews. By utilising different algorithms and techniques, we intend to uncover hidden patterns and insights within text data.

The dataset 'amazon_reviews_multi' has 7 features and contains 1.2M rows in the training set, 30k rows in the test set, and 30k rows in the validation set across six languages: English, German, Spanish, Japanese, French and Chinese. The corpus is balanced across stars, so each star rating constitutes 20% of the reviews in each language. The product_category column has the multi-class target labels, which are categories under which the review falls. The dataset can be found here: https://huggingface.co/datasets/amazon_reviews_multi

Example of classification of review: "*Inferior quality, the jacket is bursting at the seams, poor communication, unfriendly service. The only positive thing is fast delivery. Wouldn't recommend the jacket or the seller. That's why only 1 star.*" The review is classified under "apparel" and has a rating of 1 star.

We wish to answer the below questions with respect to the dataset using text mining algorithms:

- There are two columns containing text about the product reviews, namely, review_title and review_body. The idea is to identify which column can be a good candidate for identifying the product category.
- We want to identify the common words or phrases that are used for the different product categories.
- Also, we intend to find out which product categories have had the highest ratings and vice versa.
- Potentially, we plan on working with reviews belonging to 2-3 different languages and comparing the performance of the same algorithm across the different languages.

We need to perform various preprocessing steps to get the data ready for analysis. These include dropping irrelevant columns such as review_id and reviewer_id. Selecting and filtering 2-3 languages for further experimentation. Other basic NLP preprocessing steps include converting the reviews to lowercase, removing punctuation, stopword removal, stemming, and lemmatization.

To answer the questions described above, we plan on using a variety of algorithms that are popular in the Natural Language Processing and Text Mining domain. These include Support Vector Machines (SVMs), a couple of Recurrent Neural Networks (RNNs) like LSTM, and BERT.

To evaluate the results, we will be using standard classifier metrics like confusion matrix, accuracy, recall, precision, and F1 score. Besides this, we will be using any algorithm-specific metrics that can help us examine our results.