UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

COS781 - Data Mining

Semester Project Proposal

Reuben Jooste
Student number: u21457060
Email: u21457060@tuks.co.za

Topic: Sentiment Analysis and Topic Modelling of Amazon Software Reviews

Problem Statement: Businesses such as Amazon try to improve their products and services daily and sometimes it's challenging to understand how they can improve them. By using data mining techniques such as *topic modelling* and *sentiment analysis* we can derive meaningful insights from customer's reviews on Amazon's software products. It is quite interesting how something as simple as understanding customer feedback can lead to an increase in customer experience. I believe that analysing the reviews can reveal sentiment trends, trending topics/products, and what customers like and dislike, allowing the business to then make changes to their services to better fit customer needs which will lead to an increase in customer satisfaction and potentially increased revenue for the company.

Data Description: We will use the publicly available *Software* reviews dataset created by Amazon. The dataset can be downloaded by clicking the 'review' link next to the 'Software' category. The dataset contains 2.6M records in JSON format and contains the following fields: *rating (1.0 to 5.0), title, text, asin (product ID), parent_asin (Parent product ID for variants), user_id, timestamp, verified_purchase, helpful_vote.*

Approach: **Sentiment Analysis:** I will first preprocess the data to remove any stop words, as well as apply tokenisation and lemmatisation techniques. Using a pre-trained model such as *TextBlob*, I will classify the processed subset of data as positive, negative, or neutral and label the data. Based on the labelled data I will then analyse the sentiment trends by product ratings, time, and purchase verification. **Topic Modelling:** I will apply similar preprocessing techniques as the sentiment analysis approach, with an emphasis on text cleaning. This will remove any special characters and irrelevant content. Using the *Latent Dirichlet Allocation (LDA)* I will extract the key topics from all the reviews. Based on the extracted topics, we can then examine the distribution across different product ratings as well as track how certain topics got more popular or less popular over time.

Expected Outcomes based on the Approach: **Sentiment Analysis Expectations:** I expect to observe a strong correlation between the sentiment of products and their ratings. Lower ratings will likely have negative sentiment, and, in contrast, higher ratings will have more positive sentiment. In terms of the trends, I expect that the sentiment may shift due to user experience and product upgrades. **Topic Modelling Expectations:** I expect to uncover popular topics among the reviews such as easy-of-use and customer service. I also anticipate observing a varying topic distribution for different ratings.

Evaluation: I will compare the sentiment polarity with the review rating to measure whether the sentiment predictions was accurate or not. I will use time-series plots to measure if a product's sentiment improved over time. For Topic Modelling, I will use coherence scores to evaluate the topics generated by the LDA algorithm. Distribution plots will also be used to verify whether meaningful patterns emerge e.g. if praising comments are concentrated in highly-rate reviews. **Baseline**: I will assume that reviews with low ratings (below 3) correspond to negative sentiment and vice versa.

Expected Outcomes: At the end of the semester, I expect to have a detailed analysis on the extracted insights from the sentiment analysis with visualisations showing sentiment trends by product ratings, time, and purchase verification. Furthermore, I hope to produce a list of key topics found in customer reviews with visualisation showing the topic frequency and how topics vary in terms of ratings and verified or unverified purchases. Finally, I hope to produce a comprehensive report explaining the insights gathered from our approached and how these insights could benefit stakeholders such as marketers and product managers.