Santa's Little Helper Recommendation

Data Source

I scouted the web for datasets and found two which were suitable for my model. I got these two datasets from Kaggle website. The first dataset is the Twitter sentiment dataset, 'Sentiment140' dataset which can be found via this link, https://www.kaggle.com/code/paoloripamonti/twitter-sentiment-analysis?select=training.1600000.processed.noemoticon.csv. The second dataset is the Amazon Review Data 2023, a very large dataset, so I had to download it as JSON files then combine them into one dataset. It can be found via this link, https://www.kaggle.com/datasets/wajahat1064/amazon-reviews-data-2023/data?select=README.md.

Using python, I further selected the necessary features from the respective datasets to use, cleaned and preprocess these features for my model.

Algorithm Choice

I used two classification algorithms for this project. To process user inputs, I implemented Naïve Bayes model utilizing the sentiment dataset. To recommend a gift, I used K-Nearest Neighbour (KNN).

Naïve Bayes:

With every user input, the constructor splits each line of the training data into sentiments and text and store the two columns from the sentiment dataset, sentiments and text into dictionaries, counting and storing the frequency of each sentiment after filtering stop words and removing non-alphabetic characters. I further update the word counts for positive and negative sentiments and computes the overall probability of positive and negative sentiments.

I use the "SumWordCount" method to calculate the total word count from a frequency dictionary.

The "GetWordProbability" method is used to compute the probability of a word occurring in positive or negative sentiment data using Laplace Smoothing.

Finally, the "Predict" method checks whether the given input text has positive or negative sentiment. It filters the user's input and remove stop words, then get the probability of negative and positive sentiments. A comparison between the resulting scores is done and returns true for positive sentiment and false for negative sentiments.

K-Nearest Neighbour (K-NN):

The returned Boolean from the Naïve Bayes model is used together with the user input to recommend a gift. This constructor in the KNN class reads all lines of the amazon dataset, removes comas and split by a coma. I remove the special characters from each column of this dataset, main category and title and then store them into a List from a class with same properties as the column names.

I further tokenizes the product's title and the product’s main category into individual words and adds these tokens to a set “allTitlesSet” and “allMainCategoriesSet”. I then create a dictionary where each unique word is assigned an index, enumerating and storing the results in a dictionary with the word as the key and index as value.

The “EncodeInput” method breaks the words into tokens and removes stop words. It uses the vectorize method to convert tokens into numerical vectors.

The “EncodeProduct” breaks down the product's main category and title into tokens. It then uses the vector to map tokens to indices in a dictionary called, ”vocabulary”.

The “Recommend” method compare input with products and converts each product into a numerical vector using “EncodeProduct”. the cosine similarity between the input vector and the product vector using “CalculateCosineSimilarity” method and stores it in a list of candidates. I remove candidates with a similarity score of 0.1 or less and outputs “No suitable products found” if no candidates remains. I then sort candidates by similarity in descending order for positive sentiment and ascending order for negative sentiment. Finally, I pick the top match and output the product’s main category and title.

Ethical Reflection

Data Privacy: Assessing users personal data can raise concerns for their privacy regarding how this information is being processed after recommending a gift. Without the consent of the user, it would be illegal to store their information without prior consent according to GDPR.

Filter Bubble: Since the model solely relies on user preference, that is their input, the user risked being stock in a Filter Bubble where there would only get recommendations based on what they like/know. This limits the users exposure to other diverse products.

Bias Data: If the sentiment dataset contains societal biases, like gender, race or culture, the model could make biased predictions. Similarly, if the product dataset is biased towards certain demographics or categories, recommendations may unfairly favour certain groups while excluding others.

Overgeneralization of Sentiments: Sentiments are complex and context-dependent. A Naïve Bayes model might oversimplify sentiments, categorizing nuanced feedback as simply "positive" or "negative," which could result in generalization or inappropriate recommendations.

Handling Negative Inputs: If users express dissatisfaction or provide negative input (e.g., “I hate this”), the recommendation of the least similar product may not address their dissatisfaction effectively. This could lead to dissatisfaction.

Exclusion of Diverse User Preferences: Users with unique or non-mainstream preferences might not receive accurate recommendations since the data reflects popular preferences and the model caters primarily to majority preferences.

Explainability: Users may not understand how their inputs influence recommendations since the recommendation system doesn't explain why a particular a particular product was recommended or how sentiments influenced their result.

Manipulative Recommendations: A recommendation system like this one could be exploited for commercial purposes, prioritizing products from specific sellers or categories that provide monetary benefits to the platform, rather than genuinely matching user needs.

Some of the concerns raised can be address by respecting data and privacy laws like the GDPR, ensuring ethical and transparent business practices as well as utilizing neutral and unbiased data for the model training. Another solution could be to collect data that reflects diverse groups and cultures. The issue with this is the extra computational power needed to handle the processing of large datasets.