Sentiment Analysis in E-Commerce: Uncovering Insights from Customer Reviews Using Machine Learning

Zachary Yancey — *Given selection, I would want to participate in this group*

DESCRIPTION OF THE PROJECT

We live in a unique era of online shopping, where every product on the metaphorical digital shelf comes attached with a myriad of text reviews. These reviews contain a collective sentiment of all the consumers that purchased the product. This project would aim to harness the power of machine learning to unlock the insights hidden within this trove of feedback. The primary goal of the project is to create a robust sentiment analysis pipeline that is capable of automatically gauging customer sentiment from e-commerce product reviews.

Sentiment Analysis provides a lot of benefit to companies ranging from tech companies to digital vendors. By parsing and summarizing the data in this way these businesses can gain rapid access to consumer sentiment and identify issues quickly.

At the most basic level I would expect this model to be able to classify reviews as either positive, neutral, or negative based on the language used within. As an extra mile it would be really beneficial if it could determine which nouns were most commonly attached to words which affected the sentiment score. This could be beneficial in that it would provide a short list of two or three features for each product that people liked and disliked respectively, providing context to a consumer trying to decide between two products or a seller seeing how their product stacked up to similar products.

Some challenges that I expect may come up, Natural Language Variability and Model Generalization. The English language is quite complex and there is a lot of nuance to word usage, we need to be careful in how the model addresses those nuances to accurate capture sentiment. We also should ensure that the model generalizes will to different product categories and domains.

WHAT FEATURES THE DATA SET MIGHT INCLUDE

I think that for the sake of fairness and generalization we should only concern ourselves with the text in the review and disregard any rating that was given. Initially I think the input data will look something like this:

Product ID	Product Category	Text	Rating (Hidden from Training Dataset)
1	1	This device came broken and was super slow to arrive…	0.2
2	1	Fantastic product, I love it!	0.9
…	…	…	…

Obviously though this wouldn't provide the features needed to actually train a model in its current state. We would likely need to process the text by stripping unneeded words, converting the words to vectors and storing the data in a more model-friendly state:

Product ID	Product Category	Word 1	Word 2	Word 3	Word 4	…	Word N	Rating (Hidden)
1	1	Vector	Vector	Vector	Vector	…	Vector	0.2
2	1	Vector	Vector	Vector	Vector	…	Vector	0.9
…	…	…	…	…	…	…	…	…

HOW AND FROM WHERE WOULD THE DATA SET BE GATHERED AND LABELED

I did a brief search and found a few datasets on Kaggle of scraped e-commerce reviews. I think that taking that data and merging it with another dataset as well as adding manually scraped reviews from a few choice products would provide a lot of variety with regards to product categories and give us a good amount of data to work with.