STATS 507: Project Proposal **Heleyna Tucker**

Overview

The internet has become a useful resource for many people, allowing us to stay connected and spread information about important topics. Through platforms like twitter and Instagram and online shopping, like Amazon, we are more connected than ever. For this project, I aim to expand my knowledge of sentiment analysis and how it can be applied to real world situations. As of now, I have looked at Amazon review data to explore what emotions are linked to bad reviews/good reviewed products, on average. Of course, there are "good" and "bad" sentiment analysis models. However, I wanted to explore a model that included various emotions like anger, disgust, fear, joy, sadness, and surprise to get a more in depth understanding of a 'good' (above 2.5 stars) or 'bad' (below 2.5 stars) review. Some insights I hope to gain from this model and analysis are how to effectively analyze company data, use sentiment analysis, and give visualizations and summaries to represent something I would show a higher up. For example, if there is a product that should be improved in some way, what leading emotion could customers be feeling? Identifying this emotion can help companies better understand their product and what aspects they may aim to change.

Prior Work

To look into possible models to use, I came across a model in hugging face by Jochen Hartmann called "Emotion English DistilRoBERTa-base". This model utilized text analysis to get percentages of anger, disgust, fear, joy, neutral, sadness, and surprise in a phrase. In this site, there are multiple papers outlined that use this model. The paper I looked at was "Computers in Human Behavior" that looks at rumor and non-rumour tweets and their reactions, finding the emotions behind each. In this analysis, they found "Fear and Sadness are the two most instigated

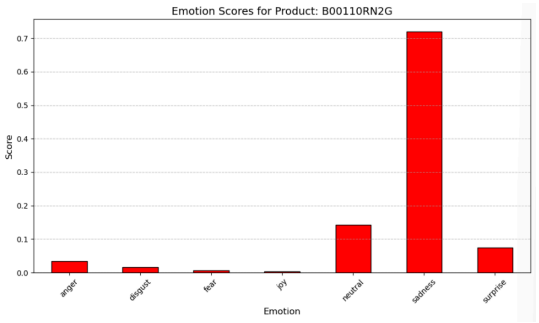
emotions in the rumor tweets" and resulted in responses that and "anger and surprise" (Butt, S., Sharma, S. Sharma, R., Sidorov, G., & Gelbukh, A.). A similar strategy could be used to understand each product individually. For example, maybe a product that has a bad review could have disgust as its leading emotion and another bad review could have more sadness. Laying out the percentages of each emotion for each product based on reviews could help companies better understand their customers and how they are receiving their products.

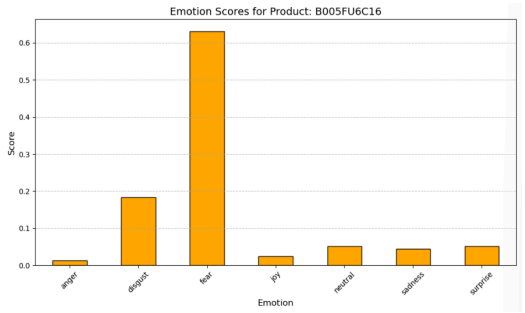
Additionally, this paper looked into certain groups of events when looking into rumored/non-rumored tweets like the Ottawa shooting and the Germanwings plane crash. Grouping Amazon review data by product-type could be beneficial because Amazon is such a big packaging company. Grouping data could help centralize what products are doing well and not so well. I am not entirely sure on what methods could be used to do this with Amazon data but would be worth looking into.

Preliminary Results

The specific dataset I'm using is an Amazon review dataset that contains around 200,000 observations from Kaggle. Notable variables being the 'asin' of the product (product code), 'reviewText' (review of the product), 'overall' rating, and 'summary' of the review. Initially, I attempted to run some tests on the reviews for each product and found that the model has a limit of 512 units. Therefore, not all reviews could be counted. Additionally, the number of observations is quite large to run through the model efficiently. Therefore, I used pandas to adjust the data I'm looking at to look at each specific product using 'asin' and concatenate the 'summary' text for each review. Doing this, I still had the word limit issue but the number of observations were shortened immensely to around 10,500. Because of this unit limitation on the model, I had to truncate some summaries to meet the requirements. Doing this I could get a very

rough idea about different lower rated products, for example. Two poorly rated product emotional analysis results are shown below:





Above, we can see that these products with lower ratings have differing sentiment analysis. Some initial limitations are that I have not pinpointed on how to get the specific product each 'asin' is for. However, the hope would be that Amazon would have this data somewhere. Finding ways to group the data more with existing variables like 'helpful' indicating whether or not the review was helpful could result in new findings. Additionally, looking more into how these 'bad' and 'good' average ratings vary on emotional text analysis would be interesting. These further analysis goals can be looked into using the existing pandas tools and maybe even some regular expressions on the text data to help sift through certain key phrases in the reviews. Graphical displays can be done using Matplotlib or seaborn.

Project Deliverables

A successful project will produce usable visuals and statistics to be used for Amazon to better understand its products and how they are being received/used. My sub goals are mostly listed above. Finding ways to group the data and visualize these findings. Additionally, looking into the 'helpful' variable and utilizing that information to better understand the quality of these 'good' and 'bad' reviews.

Timeline

Given the December 9th deadline, I suspect to have a more concrete goal (what is achievable) with some further data analysis after this first week. Then the second week, I will work on creating my report and helpful visuals that would be used to present for higher-up personnel.

References:

Butt, S., Sharma, S., Sharma, R., Sidorov, G., & Gelbukh, A. (2022). What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses. Computers in Human Behavior, 107345.

Jochen Hartmann, "Emotion English DistilRoBERTa-base". https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

Kaggle dataset: https://www.kaggle.com/datasets/abdallahwagih/amazon-reviews