Sentiment Analysis from Amazon Reviews

Group 2: Carolyn Nina Fiore and Syed Hani Haider

MOTIVATION

Sentiment analysis is the process by which models process text strings to determine the emotion associated with the text. It uses natural language processing, or NLP, and models can range from the very simplistic (associating a single word with positive or negative emotion) to the very complex (for example, identifying sarcasm or contradictory statements). Applications for sentiment analysis include analysis of public opinion, publications (anything from news articles to tweets), to optimizing automated processes like recommender systems, customer service bots, and more.

Our dataset is from Stanford University[1] and consists of 28,978 rows with 8 feature variables, detailing various Amazon reviews. We have chosen this dataset because of the combination of text and user rating, which can be used as a proxy for sentiment. A high rating is a positive review, and therefore can be used as a “positive” label for that text string. This is a public dataset and therefore competing work exists; there are similar projects with different datasets and different machine learning techniques in Kaggle and in online article publications such as Toward Data Science. From our review, there is currently no completed work – published articles, tutorials, or reference book chapters - using this dataset.

METHODOLOGY

We have broken down the project goals into milestones for low, medium and high risk steps. Our immediate goal, which is low-risk, includes learning and executing all the tasks to prepare our dataset for embedding. We will process the data and conduct EDA, ensuring standardization so that our models won’t get confused by things like contractions, accented characters, etc. We will explore the data to ensure that the reviews are long enough for the model to use and to ensure a balanced train-test stratification. We will determine whether to impose a minimum review length for the selected measurements and we will use existing python tools to standardize the text strings for processing. Estimated completion for this phase is 31 March 2023.

The medium risk goal is to perform 3 different embedding techniques on 3 different models to conduct binary classification. We will compare the performance of Count Vectorizer, TF - IDF, and Word 2 Vec with Logistic Regression, Naive Bayes, and Random Forest. For the binary classification we will use ratings of 1-2.5 as the negative classification and 2.6-5 as the positive classification. We will examine the results based on accuracy and F1 score, which is determined using precision and recall. Estimated completion for this phase is 10 April, 2023.

Our high risk goal is to attempt to replicate the results of a 2022 paper by Vernikou, Lyras, and Kanavos[2] to our dataset. Their research resulted in higher performance metrics when using three-class sentiment analysis compared with two-class. We believe that this will hold true for our data as well. For our dataset, we will assign "negative" to ratings of 1-2, "neutral" to ratings of 3, and "positive" for ratings from 4-5. This phase will attempt to run all models with the multiclass rating, but we know that some models may not produce results with this technique. This phase will use the same evaluation metrics as the medium risk models. Estimated completion for this phase is 20 April, 2023.

IMPACT

Our approach will provide a comparison of multiple NLP embedding and sentiment analysis techniques to a practical dataset, informing the existing work for data scientists interested in utilizing sentiment analysis on a variety of projects. Hopefully, we will be able to provide rigorous and replicable results that support Vernikou, Lyras, and Kanavos' results for a new dataset and different methods. Should the results contradict the paper's results, that will also provide useful information to generate further research attempts.

[1] Amazon Reviews Dataset [April 26, 2016]. Stanford University. [Online] available: http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_10.json.gz [accessed February 26, 2023].

[2] Vernikou, S., Lyras, S. and A. Kanavos. "Multiclass sentiment analysis on COVID-19-related tweets using deep learning models" [August 6, 2022]. Neural Computational Application 34(22). [Online] available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9362523/ [accessed February 23, 2023].