Description of the dataset

The data was obtained from Kaggle.com dataset titled “Consumer Reviews of Amazon Products”. The dataset contain over 34,000 reviews of Amazon products such as the Alexa or Kindle. It has columns containing the rating (1-5 stars), review text, username of the review poster, etc. For this project’s purposes, we are only interested in the “reviews_text” column which contains the user’s written review of the product.

Preprocessing steps

To clean the data, first we extract only the “reviews_text” column through indexing by its column name. Then, we create a specified function to perform text pre-processing. The function converts the text into lowercase, removes any digits/numbers as well as any symbols/punctuations and whitespaces. It also removes any stop-words predefined in the SpaCy library.

Results

In general, the model appears to do a decent job at performing sentiment analysis on a random sample of 100 reviews from the dataset. For example, it gave the review “Purchased for my son for Christmas and he loves it. Great price” a sentiment polarity score of 0.8, which is very positive. It also gave the review “I will not recommend this one, do slow and I ended up returning this product” a polarity score of -0.3, which is generally negative. However, its performance is quite inconsistent as sometimes very positive sentiment reviews will be scored as negative or neutral. For example, the review “This product works very well and represents an extremely useful piece of artificial intelligence.” was given a polarity score of -0.15.

Model strength & limitations

One of the main strengths of the model is its ease of use and simplicity. It only takes a few seconds to process 100 data entries which may not be the case for more advanced NLP models. And while simple, it is able to predict the sentiment of our data sample relatively accurately with a small number of inconsistencies. However, because of its simplicity, it is not able to achieve highly accurate sentiment analyses as more advanced models may be capable of. In addition, our model makes use of the TextBlob package where its sentiment analysis model is based on predefined sentiment scores for words. This limits the ability for model fine-tuning/customisation which may also contribute to our model’s occasional incorrect sentiment scoring.