Data Wrangling.

The studies utilized reviews submitted for five Echo products on Amazon by the end of February, 2019. The selected Echo products include three devices with a screen, including Echo Show 1st Generation, Echo Show 2nd Generation, and Echo Spot 1st Generation, and two speaker devices without a screen, including Echo Plus 1st Generation and Echo Plus 2nd Generation. The reviews were downloaded using the web scrapper implemented in Chrome. The detailed procedures are outlined as follow:
https://www.scrapehero.com/amazon-review-scraper/

The scrapper extracts review information, including author, title, date, review content, and star rating, from each product page, and the reports, with up to 4000 reviews per product, are saved as csv files.

The reviews are relatively clean. This study will focus on the review titles, review contents, star ratings and review submission dates. For text data, including titles and contents, punctuations were removed, and lower case was applied. Both reviews and titles were tokenized using the NLTK package, with stop words removed. Bigram and trigram tokens were also generated by linking unigram tokens with underscore. Character count and word count were also computed for both title and content to measure text length. Star ratings, captured as a string in the format of "X.0 out of 5 stars", was converted to integers. A handful of invalid reviews have a star rating of 0, which were removed from the dataset. No missing data was otherwise found in these key columns.