Sentiment Analysis of Amazon Reviews for the Tapo C210 Camera

A Comparison Between Lexical and Contextual Methods

Alfarano Giovanni
Durante Federico
Perico Dario

Content

1 Introduction

In this study, the sentiment of Amazon UK reviews related to the product "Tapo C210 2K 3MP," a home surveillance camera, is analyzed. The goal is to identify strengths, weaknesses, and the overall user experience by processing a large corpus of reviews. Data were collected by implementing a web scraping algorithm in R, with filters for ratings and review sorting. Sentiment analysis was conducted using two approaches: the Bing method, which is lexicon-based, and an enhanced version leveraging UDPipe, which additionally accounts for the presence of negators, amplifiers, and deamplifiers to provide a more contextual sentiment assessment. The results of the two methods were then compared to evaluate differences in the distributions of sentiment scores and their correlation with the star ratings given by users.

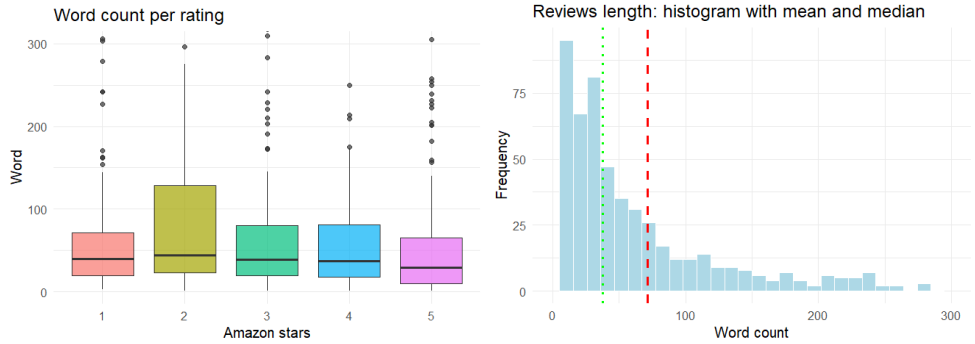
2 Data

Amazon implements various strategies to limit web scraping activities. In our case, the main restriction is that the platform displays a maximum of 500 reviews per product, divided into batches of 100 reviews for each rating level (from one to five stars). By changing the review sorting option from "top reviews" to "most recent," it is possible to access a different batch for each star rating, although with significant overlap between the two sorting orders.
By combining sorting options and filters for each rating level, we managed to collect a total of 1,000 reviews. Among these, 420 were duplicates (overlap), reducing the actual number of unique reviews to 580. Eight reviews were discarded due to unresolvable errors or incomplete data. Of the remaining 572 reviews, some were in languages other than English; to preserve as much information as possible, the language of each review was detected and, if it was not English, the text was translated into English using Google Translate. The final dataset consists of 572 unique reviews, each containing the following information: title, text, and star rating. According to Amazon's official data, the global average rating for the product is 4.6 out of 5.

Due to the collection method, based on stratified sampling by rating level for operational reasons, the arithmetic mean of the ratings in our dataset is 2.98. This average, significantly lower than the actual mean reported by Amazon,

reflects the distortion introduced by stratified sampling and does not represent the true distribution of user satisfaction. However, this distortion does not pose an issue for our objectives, as the analysis aims to identify recurring patterns and themes in both positive and negative reviews rather than to estimate the actual distribution of ratings.

The reviews have an average length of 71 words, with a median of 38 words and a standard deviation of 96 words. The shortest review consists of a single word (e.g., "ok," "great"), while the longest extends to 921 words.



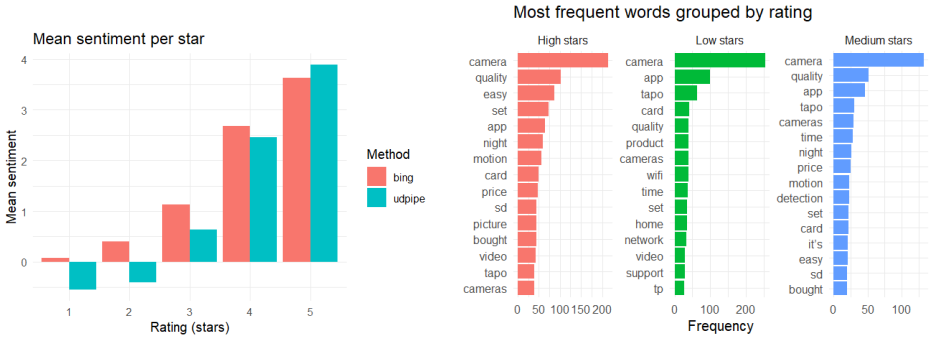
Analyzing the average review length based on the assigned rating reveals that reviews with 2, 3, and 4 stars tend to be longer than those with 1 or 5 stars. This is consistent with expected behavior: in extremely negative or positive reviews (1 and 5 stars), users often express their dissatisfaction or satisfaction concisely, whereas in intermediate ratings they tend to provide more detailed descriptions of their experience.

3 Results and discussion

For the sentiment analysis, we implemented the Bing dictionary, which is based on the binary polarity of words (-1 for negative terms, +1 for positive terms). The corpus was tokenized at the word level, and for each document, the polarities of the tokens were summed to determine an overall sentiment score. The UDPipe method follows a similar approach but with important differences: in addition to considering the word itself, the two preceding tokens are analyzed. If these belong to lists of negators, amplifiers, or deamplifiers, their effect modifies

the polarity of the subsequent token. The resulting scores are then aggregated for each document, as in the Bing method, to obtain the final sentiment score.

The correlation between bing sentiment scores and star ratings is 0.39; the correlation between udpipe sentiment scores and star ratings is 0.46. The analysis of sentiment scores by rating reveals a trend consistent with expectations:



The distribution of sentiment scores according to the two methods shows significant differences: the average sentiment calculated using the Bing method is 1.56 in our balanced dataset, while the arithmetic mean of the ratings is 2.98. In a balanced sampling context, one would expect an average sentiment closer to zero. The more complex UDPipe approach, which takes into account context and linguistic modifiers, instead yields a lower average sentiment that aligns more closely with the expected distribution.

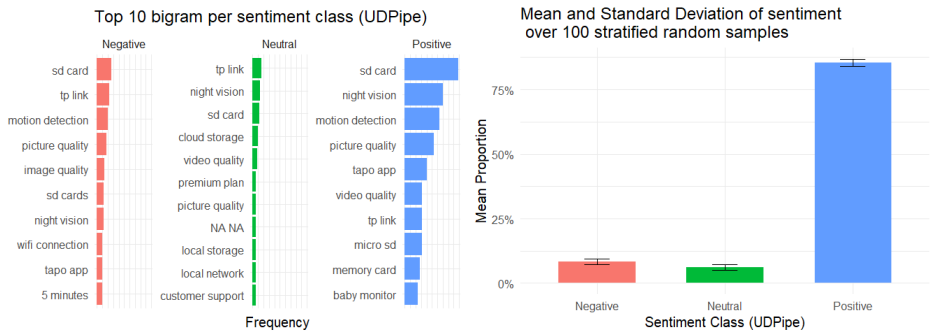
Despite these differences, the scores from the two methods show a strong mutual correlation of 0.85, confirming that both capture the general direction of sentiment.

Analyzing the average sentiment by assigned rating shows that the Bing method tends to underperform in negative reviews (1–3 stars), likely due to its inability to capture negations, amplifications, and other linguistic nuances. This limitation reduces Bing's sensitivity in detecting sarcasm or detailed complaints, aspects that UDPipe manages more effectively. Dividing the reviews by rating groups reveals that positive terms such as good, easy, quality, and great appear more frequently in high-rating reviews (4–5 stars), while words like no, problem, issue, and disconnect occur predominantly in low-rating reviews (1–2 stars). This distribution aligns with the expected polarity.

The presence of the term "camera" does not provide useful information for distinguishing sentiment but instead reflects an inevitably frequent domain term. This phenomenon is typical of domain-specific words.

The word tapo also appears often in negative reviews, suggesting that the brand itself becomes a target of criticism in certain cases.

The analysis of bigrams highlights recurring topics of interest or concern for users. Frequent bigrams include sd card, memory card, micro sd, which indicate widespread discussion of issues or features related to memory management. Night vision emerges as a recurring theme, both positively and negatively, while motion detection confirms users' interest in automatic detection features. Comments about picture quality, video quality, and image quality reflect widespread attention to visual performance, and bigrams like tp link and tapo app emphasize the relevance of the brand and its application to the overall user experience.



Grouping reviews based on sentiment estimated with UDPipe shows that some bigrams appear across both positive and negative polarities, such as motion detection and picture quality, while others are almost exclusively associated with positive reviews, including night vision and tapo app. Conversely, the bigram tp link is more frequent in negative reviews, suggesting direct criticism of the brand. Criticism related to picture quality is understandable given that video quality is not the primary expected feature of a security camera; however, some users purchase the product for alternative purposes, such as pet monitoring, where image quality becomes a more significant factor.

Our reviews are evenly distributed across star ratings (about 20% per star), which doesn't match the real Amazon star distribution. We perform stratified sampling 100 times to mirror the actual distribution and compare aggregated sentiment.

Since reviews use stars but our sentiment scores are numerical, we classify UDPipe scores below -1 as negative, above 1 as positive, and the rest as neutral, corresponding to Amazon star groups (1–2 negative, 3 neutral, 4–5 positive). Sampling 100 times, we find the average sentiment distribution is close to the real one: 85% negative (vs. 92% real), 6% neutral (vs. 4%), and 8% positive (vs. 4%).

4 Conclusions

The results obtained highlight the intrinsic complexity of sentiment analysis applied to real product reviews. The maximum correlation achieved by the UDPipe method, which reached 0.46 with the user-assigned ratings, is not particularly high and would be insufficient for building a robust predictive model, especially in a multiclass context. The difficulties encountered can be attributed to several factors, including the use of static dictionaries for sentiment classification, the limited ability of lexicon-based methods to capture linguistic context, and the heterogeneity of user behavior. In some cases, users assign ratings that are inconsistent with the textual content of the review, such as positive texts paired with one-star ratings.

Supporting this hypothesis, the experimental implementation of a classification model based on BERT showed an accuracy of 31% on the test set, confirming that correctly predicting the rating from free-form text is a challenging task, even when using state-of-the-art models.

Despite these limitations, the analysis enabled the clear identification of the product's strengths and weaknesses. Specifically, ease of use and the quality of night vision emerged as appreciated features, while the main issues reported by users concerned connection problems, compatibility with memory cards, and, in some cases, image quality. These findings demonstrate the validity and effectiveness of sentiment analysis based on relatively simple text mining techniques, especially when considering the low computational cost compared to more complex methods such as Large Language Models.

In conclusion, while recognizing the predictive limitations of the methods used, the analysis confirmed their utility in extracting relevant insights from textual reviews, providing a detailed overview of user experience and the key drivers of satisfaction or dissatisfaction related to the Tapo C210 product.