UCSanDiegoX DSE200x
Final Project

Sentiment analysis to determine whether Amazon Kindle products are likely to have a high or low rating

Megan A Holborn

Abstract

Using customer review data from the Amazon Kindle Store collected between May 1996 and July 2014, this analysis aimed to determine whether sentiment analysis of textual product reviews could be used to predict whether Kindle products received high or low ratings. A Naïve Bayes classifier model was utilised and rating polarities were successfully predicted at a high accuracy. The results suggest that sentiment analysis of textual reviews can effectively be used to categorise the rating levels of Amazon Kindle products.

Motivation

- Online product reviews provide valuable insight into customer sentiment that can be utilised by businesses to make informed decisions about product purchasing and advertisement, and reputation management.

- The Amazon Kindle store provides a popular range of e-book and e-reader product offerings, attracting a large amount of customer reviews.

- Understanding the customer sentiments associated with positive and negative ratings in Kindle reviews, could assist e-book/e-reader buyers in making informed purchasing decisions. Additionally, it can provide feedback to Amazon that aids in the selection of recommended products and highlights key product issues.

- This analysis aimed to determine whether the sentiments expressed in Amazon Kindle product reviews could serve as indicators for a product’s overall rating.

Dataset

- Amazon Kindle Store customer reviews collected between May 1996 and July 2014 were used for this analysis.
- The data is available at:

  https://www.kaggle.com/datasets/bharadwaj6/kindle-reviews/versions/3?resource=download
- The dataset was retrieved in Json format and consists of 982 619 review entries and 9 features, prior to filtering.
- The dataset provides information on (i) customer reviews, including reviewer ID, review text, review summary, and the review time, (ii) the reviewed product ID, and (iii) the overall rating of a product.

Data Preparation and Cleaning

The dataset was prepared for sentiment analysis as follows:

1. Entries with null values and duplicate entries were located and removed if present.
2. Relevant dataset features, namely, the review text and overall rating of a product, were selected for the analysis.
3. Review entries with a corresponding rating greater than 3, equal to 3 and less than 3 were categorised as positive, neutral, and negative, respectively.
4. The dataset consisted of 982 483 review entries. Random samples of 20 000 positive and 20 000 negative reviews were selected for sentiment analysis.
5. In preparation for sentiment analysis, review text was converted to lowercase and tokenized using NLTK. Stopwords and punctuation were removed. Negation handling was also performed to ensure that words preceded by negations would not be misinterpreted.

Research Question

Can sentiment analysis on textual reviews be used to predict which Amazon Kindle products received high or low ratings?

Methods

1. Reviews with a corresponding rating greater than 3, equal to 3 and less than 3 were categorised as positive, neutral, and negative, respectively.

2. Using a bag-of-words model, words associated with each positive or negative review were collected.

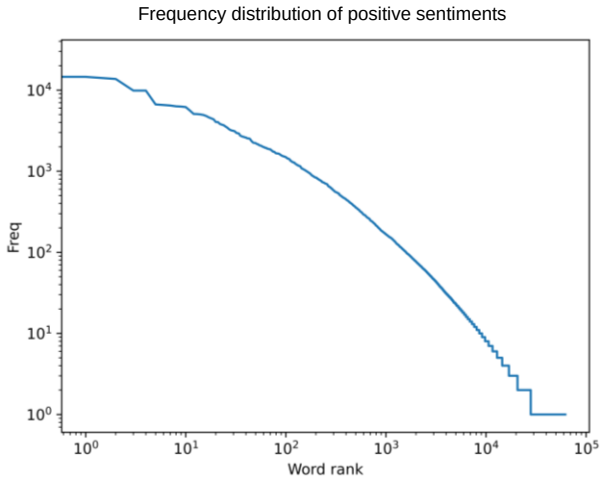
3. Log-log plots were used to visualize the frequency distribution of all positive and negative words. Additionally, word clouds were used to visually represent the most common positive and negative words found in these reviews.

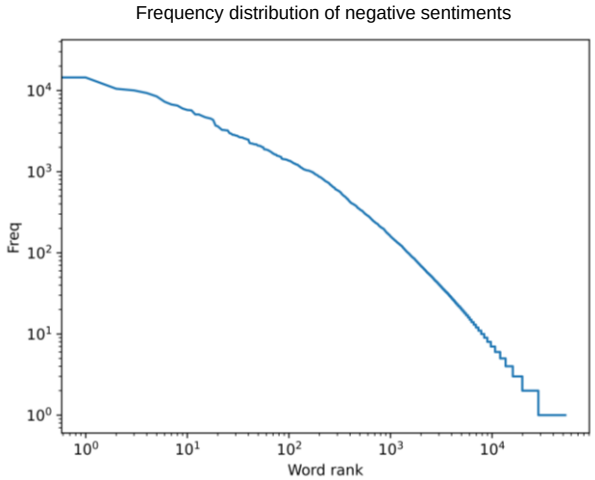
4. To predict which reviews were likely to have a positive (high) or negative (low) rating, a Naïve Bayes machine learning classifier was trained on the features associated with positive and negative reviews. The classifier was trained on 80% of the data. The remaining data was used for testing purposes.

Findings

Frequency distribution of positive and negative sentiments

- The gradual decline of the distribution curves indicated a broad vocabulary usage in both positive and negative reviews.
- Positive reviews contained more words than negative reviews (62137 vs 52560 words).



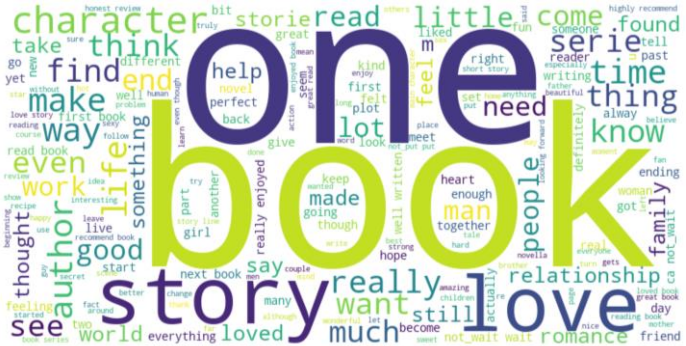


Findings

Common positive and negative sentiments

- Several of the most common words were shared between positive and negative reviews.
- In the positive reviews, the most frequent words, in descending order, were book (n=25096), story (n=14574) and one (n=9870). Similarly, in the negative reviews, the most frequent words, in descending order, were book (n=25452), story (n=14471) and one (n=9307).
- The shared common words are unlikely to be informative in distinguishing between positive and negative reviews.

Common positive sentiments



Common negative sentiments

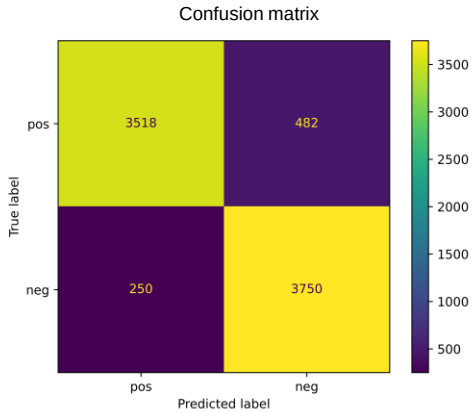
Findings

Naïve Bayes classification of textual reviews

- A Naïve Bayes classifier, trained on 80% of the review data, predicted which Amazon Kindle products received positive (high) or negative (low) ratings with 90% accuracy.
- According to the confusion matrix, of the 8000 reviews tested using the classifier, 7268 were correctly classified as positive or negative, while 732 were incorrectly classified.

Accuracy report

Train accuracy	Test accuracy
95.09%	90.02%



Findings

Informative features for classification

- The classification model was trained using the most informative features, which were words that were likely to occur in positive reviews but not in negative reviews, and vice versa. The top 5 features are listed in the table below.

Informative features

Feature	Occurrence ratio
not_disappoint	pos : neg = 104.3 : 1
4.5	pos : neg = 73.0 : 1.0
Deleted	neg : pos = 62.1 : 1.0
not_impressed	neg : pos = 52.3 : 1.0
not_finish	neg : pos = 44.7 : 1.0

Limitations

- For this analysis, a single machine learning algorithm was employed. No other algorithms were tested to determine which would be most suitable.

- Many words were shared between both negative and positive reviews. In future analyses, it may be useful to group words in phrases to aid in distinguishing between sentiments expressed in positive and negative reviews.

- It is also worth noting that the classifier did not predict the exact rating of a product, but rather whether a product received a high or low rating. It would be beneficial to incorporate more fine-scale predictions of product rating in future analyses.

Conclusions

Despite the presence of a wide range of vocabulary and shared common words between positive and negative reviews for Amazon Kindle products, a Naïve Bayes classifier model was successfully utilised to accurately predict which Amazon Kindle products received high or low ratings using textual review data. This was made possible due to the utilisation of informative features, which were words that were likely to occur in positive reviews but not in negative reviews, and vice versa.

Acknowledgements

The data used in this analysis was obtained from:
https://www.kaggle.com/datasets/bharadwaj6/kindle-reviews/versions/3?resource=download

This data was originally compiled by Julian McAuley:
http://jmcauley.ucsd.edu/data/amazon/

References

Taparia, Ankit and Bagla, Tanmay, Sentiment Analysis: Predicting Product Reviews' Ratings using Online Customer Reviews (May 15, 2020). Available at SSRN: https://ssrn.com/abstract=3655308 or http://dx.doi.org/10.2139/ssrn.3655308