Article Title

Frederik Vincent Primdahl Mikael Steenberg Pasovski
Andreas Peter Brodersen Andreas Wede Gustavsen

Copenhagen School of Design and Technology

May 30, 2022

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

I. Introduction

In the globalized economy of today, consumers are presented with a range of products to choose from when shopping. This can present some difficulty, as quality of the product is not always clear from its presentation. The customer review seeks to solve this by providing feedback on the purchase. Still, determining the value of a review can be difficult, as reviews can be misleading or provide little information.

Voting systems are often set in place to identify helpful reviews, and help consumers circumvent this issue, yet these reviews still get lost in the sheer number of unlabeled reviews. Our machine learning model aims to further filter reviews mainly by analyzing use of words, verbosity, among other aspects, to discover what makes a review helpful.

II. Problem Formulation

The main focus of this paper is to analyze Amazon review data across different categories and answer the question: "What makes an Amazon product review helpful?" through the use of classification models. In addition, answering "can you tell the helpfulness of a review based on name?", "how does year affect the sentiment of the review?"

III. Methods

Maecenas sed ultricies felis. Sed imperdiet dictum arcu a egestas.

- Donec dolor arcu, rutrum id molestie in, viverra sed diam
- Curabitur feugiat
- turpis sed auctor facilisis
- arcu eros accumsan lorem, at posuere mi diam sit amet tortor
- Fusce fermentum, mi sit amet euismod rutrum
- sem lorem molestie diam, iaculis aliquet sapien tortor non nisi
- Pellentesque bibendum pretium aliquet

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Text requiring further explanation[1].

IV. Analysis

i. Data Preprocessing

Initial data preparation was carried out on the whole dataset, dropping review time, image, style, ASIN, and reviewer id columns, as these were not needed. Unverified reviews were also filtered out, to avoid reviews from customers who did not purchase the product. Finally, null values were removed from the review text column.

As our input data mainly consists of unstructured text reviews, text normalization was done on the review text column to reduce randomness and reach a more uniform structure. This was done in accordance with common NLP data cleaning guidelines. Steps in this process involved removing non-word elements from the text – punctuation, symbols -, along with transforming each remaining word into lowercase. Because our model concerns itself with classification rather than sentiment, stop words like “not” did not provide any additional information to our model and were subsequently removed, as with other stop words contained in the NLTK “stopwords” library.

Each review was then tokenized into separate words using whitespace as a delimiter, as part of lemmatization to reduce morphological variation. Shortening some words to their root was thought important to improve model training efficiency. Lemmatization was chosen over stemming, as the accuracy of a lemmatizer in properly shortening words was more important than the speed of a stemmer.

[1]Example footnote

As the final step of our data cleanup a column “voteSuccess” was added to the dataframe, calculating helpfulness of a review by dividing number of votes from the “votes”-column by quarters elapsed since the review was submitted. This was done to create a fairer indicator of helpfulness, since newer reviews would have had less time to accumulate helpfulness-votes and therefore appear less helpful.

ii. Ethics and bias

When you consider ethics in NLP, you need to first look at the bias the can occur. A review might say “Great product, used it for a week!”. This is perceived as helpful by some customers, since the reviewer expresses their opinion on the product with sentiment information. But some other consumer might not find it helpful due to the lack of more details or an analysis of the product. This leads to reviews being biased towards the consumers uses, and needs.

Another thing that frequently occurs in reviews is herd behaviour. If we have a product, then depending on the first review given to it, other reviews might tend to follow. If the first review is very positive, other reviews might have a tendency to also grade the product higher. This leads to reviews being untrustworthy. [Yves Rychener, 2018].

V. Findings

i. Subsection One

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo.

Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VI. Conclusion

References

[Figueredo and Wolf, 2009] Figueredo, A. J. and Wolf, P. S. A. (2009). Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330.

[Yves Rychener, 2018] Yves Rychener, Nicolas Zimmermann, Loïs Bilat Detecting Bias in Amazon reviews `https://ada-lyn.github.io/`