# Classifying Useful Amazon Reviews Using a Pre-Trained BERT Model

David Miller

## Abstract

There is an overwhelming amount of information generated in reviews for products on Amazon. Because there is no limit or filter for who creates a review, most of that information is not helpful. With rapid advances in language modeling techniques, I aim to use a BERT model pre-trained for sequence classification to predict the usefulness of Amazon reviews. To test the hypothesis that this state-of-the-art model can be leveraged to improve this prediction task, I will compare performance across a series of baseline models.

## Introduction

As more and more shopping shifts to ecommerce, research for making purchasing decisions can be self-contained within a given product's page on a website in the form of user reviews and ratings. No online retailer presents this wealth of valuable information better than Amazon. Aggregated metrics, keyword search, and filtering options allow potential consumers to conduct rapid research and comparison before purchasing a product. Identifying useful reviews and extracting the information contained has many potential applications. Consumers are informed with key product information and retailers can track what people like or dislike about their products.

However, an inherent shortcoming is that review data relies on the expertise and thoroughness of the users generating it. Amazon attempts to enrich the data to overcome this shortcoming by marking reviewers as verified and allowing other users to vote on whether they found a review to be useful.

## Background

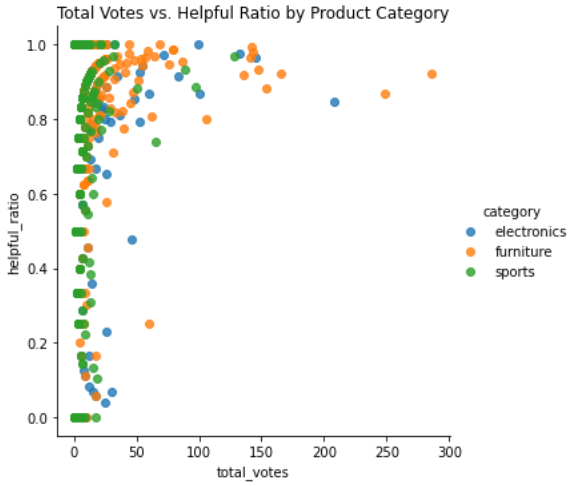
Various forms review usefulness prediction research has been performed in the past to further solve for the inherent shortcoming of review data. One of the earliest attempts (Kim et al., 2006) takes a regression-based approach to ranking Amazon reviews with a combination of text features and review and product metadata. A similar but more recent approach adds additional features such as the reviewer's reputation (Baek et al., 2012). Another work is one of the earliest to consider both regression and classification (Ghose and Ipeirotis, 2011) by first quantifying usefulness as a continuous value and then producing a binary prediction of whether or not the review is useful. These were very helpful in defining the problem space and scope of available data. Two more recent works made an important distinction between this problem and generic sentiment analysis by hypothesizing that useful reviews contain arguments. The first (Liu et al., 2017) considers a set of 110 hotel reviews and manually labels arguments contained within the review before training using a set of traditional machine learning classification algorithms. The second (Passon et al., 2018) applies a more automated approach by using off-the-shelf argumentation mining and undergoing a transfer learning task to apply these embeddings to review data.

**Methods**

My goal is to predict whether a review is useful by using BERT; one of the latest and most advanced models for language understanding. Specifically, I will use an out-of-the-box implementation of the model: BERT for sequence classification; and fine tune the model structure and parameters based on feedback from training.

Data will come from publicly available Amazon reviews. The review text body, helpful votes, and total votes are of particular importance for the task. The review text will form the basis for the predictor variable. The binary target variable is calculated as the ratio of helpful votes to total votes for the review, split at 0.8 (i.e. – a review with 8 helpful votes out of 10 total votes would be classified as being useful).

After significant exploratory data analysis, it is important to highlight the underlying distribution of the predictor variables.




| Total Votes: | Skew | Kurtosis |
|---|---|---|
| Electronics | 11.29 | 163.80 |
| Furniture | 10.85 | 153.58 |
| Sports | 12.39 | 201.45 |
| Total | 12.36 | 205.83 |

I have chosen to remove any reviews with 0 total votes, but otherwise include all reviews as eligible. Although skew statistics for votes are similar across several product categories, I will sample from 3 distinct categories for modeling (electronics, furniture, and sports) to examine the intuitive hypothesis that review content may differ across distinct categories.

Standard text cleaning procedures were used to prepare the data for use in all three models. These steps include removal of html, special characters, and extra spaces. Because BERT models have their own preprocessing layer, certain further preprocessing steps were only applied for use in the two baseline models discussed below. These steps include lemmatization to reduce words to their root form, converting words to lowercase, removing punctuation, and removing stopwords.

To compare against the eventual BERT model, I have developed two distinct baseline models. The first is a very simple Naïve Bayes classifier, trained on tf-idf vectorizations of the processed

text. The second is a more robust, ground-up implementation of a 1-dimensional convolutional neural network. The CNN will have 2 Dense layers (with a relu and sigmoid activation, respectively), and be optimized with the default parameterization of the Adam optimizer. It will be trained on tokenized representations of the processed text. The BERT model is taken from TensorflowHub, specifically the small-uncased model with a classifier output layer. All models will use SMOTE to resolve class imbalance in the training data and prevent overfitting.

**Results and discussion**

See below for a summary of performance metrics across the 3 models and 3 product categories.

| | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| **Electronics** | | | | |
| Naïve Bayes | .5927 | .5814 | .5927 | .5748 |
| CNN | .5887 | .5770 | .5887 | .5313 |
| BERT | **.6008** | **.5937** | **.6008** | **.5937** |
| **Furniture** | | | | |
| Naïve Bayes | .6199 | .6011 | .6199 | .6091 |
| CNN | .5337 | .5858 | .5337 | .5512 |
| BERT | **.6550** | **.6154** | **.6550** | **.6263** |
| **Sports** | | | | |
| Naïve Bayes | .4799 | .4603 | .4799 | .4660 |
| CNN | .5861 | .5687 | .5861 | .5608 |
| BERT | **.5897** | **.5802** | **.5897** | **.5811** |

For each product category, the BERT model outperformed both baselines. The magnitude of outperformance was anywhere from 0-12% depending on model and metric. The Naïve Bayes model struggled to generalize to test data even when applying oversampling methods, while the CNN and BERT models take significantly more compute power. While significant improvement would be needed to consider making business (or even personal) decisions based on outputs from the model, there is some evidence that BERT's out-of-the-box features can save significant development time while not sacrificing performance.

**Conclusion**

The overall performance metrics across each model, as well as the performance of the BERT model in comparison to the baselines, indicate that even current state-of-the-art models require a significant amount of fine-tuning and thoughtful approaches to data enrichment and processing. Model performance was worse in comparison to the referenced literature of prior works, indicating that additional features and more robust text processing may significantly improve performance.

There are several key priorities for potential future work to refine and improve the utility of the models developed in the scope of this work. The first is to conduct additional exploratory data analysis and text processing. Analysis conducted in the scope of this work indicates that there is significant skew to the distribution of votes on reviews, at the product category and individual

product level. Additionally, the quality and reliability of text data in the review body is low as it is entirely user generated. Adjusting the definition of whether a review is useful, changing the filter process for which reviews to include in the modeling exercise, and changing the text processing steps could all have tangible impact on model performance. The second is data enrichment. BERT has ample flexibility to ingest additional metadata as features pre-pended to the review body text. Based on learnings from other works in this space, potentially valuable features include reviewer reputation, reviewer history, and product metadata. The third is to conduct further model fine-tuning and hyperparameter optimization.

An additional scope of future work includes adding justification to model predictions. Model interpretability is extremely important in enterprise-grade machine learning models and algorithms. Interpretability for classification based on text data can take the form of extracting pieces of the review body or a summarization of the review. I reviewed the work performed by the Google research team in creating T5 (Narang et al., 2020) as a viable use case for transfer learning applied to this problem space.

**References**


Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining. *Proceedings of the 5th Workshop on Argument Mining*, pages 35–39. Association for Computational Linguistics.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. 2012. Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2):99–126.

A. Ghose and P. G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using argument-based features to predict and analyse review helpfulness. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1363. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, Karishma Malkan. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv preprint arXiv:2004.14546*, 2020.