Predicting Medication Review Usefulness from Text and Metadata

Abrar Ahmed
Georgia Institute of Technology
Atlanta, GA
abrar.ahmed@gatech.edu

Rachit Bhargava
Georgia Institute of Technology
Atlanta, GA
rachitb@gatech.edu

Joseph Miano
Georgia Institute of Technology
Atlanta, GA
jmiano@gatech.edu

Abstract

In this paper, we study relationships between medication review text, metadata, and usefulness. Using a dataset of medication reviews from Drugs.com, we perform exploratory analyses, train models to predict review usefulness, and visualize model attention in models predicting condition and usefulness. https://github.gatech.edu/jmiano3/NLP_Fin_Proj

1 Introduction

Product reviews are ubiquitous across a range of industries and help potential customers make purchasing decisions. Medication reviews, such as those curated on the website Drugs.com, are especially sensitive and important because they can impact a person's choice of medication and therefore their health. Understanding and predicting the usefulness of reviews, before they have been online long enough to be upvoted, can help businesses prioritize their reviews and help users improve the quality of their reviews.

In this work, we leverage a dataset of free-text medication reviews and their metadata from Drugs.com (Gräßer et al., 2018) to perform exploratory analyses of the review usefulness and then train baseline as well as attention-based models to make usefulness classification, regression, and ordinal regression predictions. We also investigate the relationships between review usefulness and condition classification. Finally, we visualize model attention as it relates to condition and usefulness classification to investigate the relationships between these tasks.

2 Related Work

Much existing literature related to review analysis has focused on the statistical impact estimation of reviews on product sales (Lin, 2014; Zhang et al., 2013), user-wise evaluation of posted reviews (Forman et al., 2008), and evaluation of reviews through estimation of user intent (Jarvenpaa and Leidner, 1998). Recent work has also suggested the use of some Natural Language Processing (NLP) techniques to predict review usefulness. Kim et al. (2006) suggest using an SVM regressor to predict review usefulness with metadata and word features, which assume equal word importance. However, their approach comparatively ranks reviews and does not give an independent score. Another study (Liu et al., 2008) uses the review metadata along with expert rules to check the writing quality to predict review usefulness. Some other relevant papers include an IEEE study (Ghose and Ipeirotis, 2010) that aims to use the review metadata and writing quality to predict review usefulness and a feature extraction study (Liu et al., 2013) that combines semantics and phrasing to find relationships with review usefulness. These studies, while able to find relationships, completely or partially ignore the textual data itself.

3 Methods

3.1 Dataset and Data Processing

We focus our project on a dataset consisting of medication reviews from 2008 to 2017 from Drugs.com (Gräßer et al., 2018). The dataset is split into a training set and a test set, with 113,494 medication reviews in the training dataset and 48,439 reviews in the test set. In addition to the reviews themselves, the dataset contains a rating (from 1 to 10), a condition, a drug name, a date, and a number of upvotes (useful count) associated with each review. To increase review freshness and make our modeling more computationally manageable, we consider the time-span of 2013 to 2017 and the top 10 conditions overall by review count. Furthermore, we cap the outliers (in terms of useful count) down to the 99th percentile value to help standardize the review usefulness across conditions. We then generate our **target variable**, called the usefulness score, which

is normalized to be between 0 and 1 by dividing the capped (99th percentile) number of upvotes for each review by the maximum (capped) number of upvotes in the date range. We also develop an age score column, which is a normalized age in days (compared to the most recent review).

3.2 Modeling Usefulness

We formulate modeling usefulness as regression, classification, and ordinal regression problems. We also study the impact of metadata on the usefulness prediction by training model variants that use text and metadata. For classification, we split our data into 2, 3, 4, or 5 equal-sized parts by using equidistant quantiles. Finally, we combine regression and classification to perform ordinal regression by bucketing the outputs of the regression model across equidistant quantiles to obtain class labels; the reasoning behind this approach is that if it performs well, it could enable a single regression model to make predictions across arbitrary class buckets.

We also investigate linear and neural meta-data baselines, bag-of-words baselines, and transformer-based models of text-only and text with metadata. We leverage a pretrained DistilBERT (Sanh et al., 2019) encoder from the HuggingFace transformers library (Wolf et al., 2019), to which we append a 2-layer Leaky ReLU neural network to output regression or classification predictions. For the text-only model, we use the DistilBERT embeddings as inputs to the 2-layer neural network, while for the text with metadata models, we concatenate the metadata features to the DistilBERT embeddings before passing them through the neural network. We use the Adam optimizer (Kingma and Ba, 2014) with a batch size of 8 and learning rate of 0.0001; we use Mean Squared Error (MSE) loss for regression and cross-entropy loss for classification.

3.3 Visualizing Attention

To better understand what makes reviews useful or not, we visualize the attention maps for trained text-based DistilBERT models using the ktrain library (Maiya, 2020), for which adapted this tutorial to our dataset. After loading and preprocessing our data, we train a DistilBERT text-only model to predict condition, and another to predict usefulness (binary classification). Then, we use the ktrain's built in "explain" function to generate attention maps for the predictions made on specific reviews. This enables us to compare and contrast attention for condition prediction and usefulness prediction.

4 Experimental Results

We began with exploratory analyses of the useful count and its relationships with the other metadata variables. In Figure 1 (A), we show violin plots of the distribution of useful count by condition, for the top 10 conditions by review count. For the uncapped useful count variable for data from 2013 to 2017, the Weight Loss condition had the single review with the highest useful count, at 556 upvotes. On the bottom of Figure 1 (A), we show the useful count distributions capped to the 99th percentile to help standardize review usefulness comparisons across conditions.

In Figure 1 (B), we show the relationship between the useful count and the age score. Figure 1 (B) shows a positive correlation between the age score and the useful count. Intuitively, the older a review, the more time it has had to be displayed on the website and collect upvotes. However, it appears that after a certain age score (about 0.2 in Figure 1 (B)), most of the density below the useful count of 100 flattens out. In Figure 1 (C), we demonstrate the relationship between the review ratings and the useful count. It appears that reviews on either extreme of the rating spectrum tend to be more useful, with a bowl-shape forming from reviews with ratings of 1 to 10; however, the positive reviews (7 and up) appear overwhelmingly more useful, especially the 9s and 10s.

In Figure 2, we show our results for training regression, classification, and ordinal regression models to predict the usefulness score. From (A), we observe that our DistilBERT text model outperformed the Linear BOW text model, obtaining a Mean Absolute Error (MAE) score of 0.1203 on the test dataset versus 0.1300 with a Linear BOW model (Shen et al., 2016). Furthermore, DistilBERT + Metadata outperformed a neural network which used only the metadata, with an MAE of 0.0774 versus 0.0805 with the baseline neural network (Liu et al., 2008). As shown in Figure 2 (B), the standard classification and ordinal regression models performed approximately equally well across 4 different data splits; however, the performance of both models suffered as the number of useful score classes increased. In Figure 2 (C), we show how giving more features to our text-based DistilBERT model improves its performance. Given only the review text data, the 2-class classifier obtained an F1 score of 0.76792. Adding the metadata (condition and rating) increased the

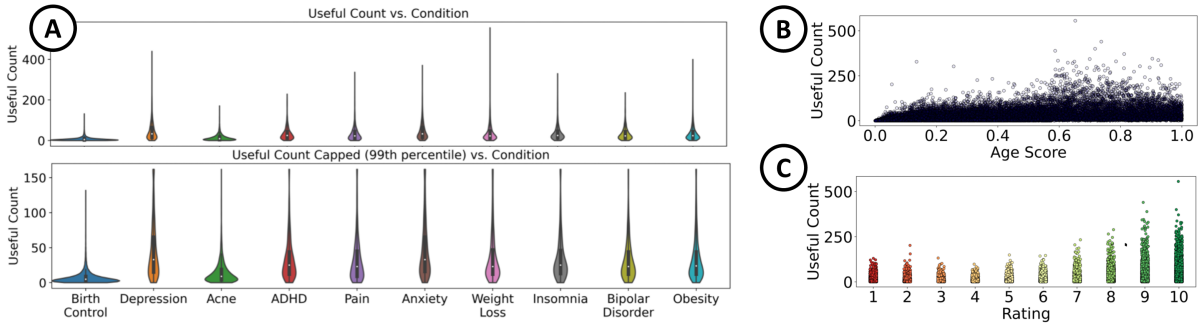

Figure 1: In A, we show violin plots of uncapped and capped (to the 99th percentile) useful counts by condition. In B, we show a plot of useful counts against the age score, which is a normalized measure of how old each review is in days. In C, we show strip plots of the useful count against the rating for the same top 10 conditions.

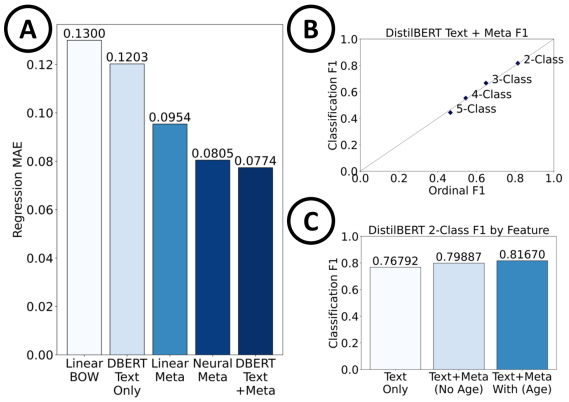

Figure 2: In A, we compare the MAE on the test dataset for 5 models, where DBERT refers to DistilBERT. In B, we compare the classification DistilBERT Text + Metadata model to the regression-based ordinal classifier and plot their F1 scores. In C, we show the DistilBERT performance across 3 training conditions.

F1 score to 0.79887, and adding the age feature further increased the F1 score to 0.81670.

In Figure 3, we show the attention maps on specific test-set reviews for a model we trained using the ktrain library (Maiya, 2020) to predict condition classification (for the top 10 conditions) and another model to predict 2-class usefulness (i.e., useful vs. not useful). Specifically, we fine-tuned DistilBERT models for each task. Figure 3 (A) and (B) show attention outputs for reviews predicted to be useful, and 3 (C), (D) and (E) show attention outputs for reviews predicted to be not useful. Further discussion and interpretation of these results are included in the Discussion section below.

5 Discussion

Our results demonstrate that transformer-based models are indeed able to predict review usefulness using the review text only, but that augmenting the text features with metadata features can further improve the performance. From our exploratory analyses, we observed that the distribution of upvotes varied by condition, age, and rating; thus, we can expect that enabling a model to use these features along with the text to make predictions would improve its performance. Furthermore, we observed that these combined models of text + metadata outperformed baselines that only used metadata, which show that our combined model is able to learn something about review usefulness beyond what is available in the metadata. We also showed that a model trained for usefulness regression can be applied towards ordinal classification by bucketing its outputs, which enables the training of a single model whose outputs can be grouped into any desired number of categories.

In our analyses of attention (Figure 3), we observe that for useful reviews, it appears that the model pays attention to the medication (e.g., "adipex" for the 2nd Useful review), aspects of the condition (e.g., "90lbs" for the same review), and to the method of administration (e.g., "tablet" for the first Useful review). On the other hand, when predicting *not* useful, the model attended to misspellings (e.g., "workedad" for the first Not Useful review), and mixed information, which may be difficult for it to parse (e.g., "the first week was not so bad... I do not recommend this drug"). Words that our condition model pays attention to when predicting condition are the medication (e.g., "wellbutrin" for Depression), the condition (e.g., "insomnia"), and relevant words and concepts (e.g., "focus", "school", and "grades" for ADHD).

6 Conclusion

In this paper, we presented exploratory analyses and predictive modeling of medication review use-

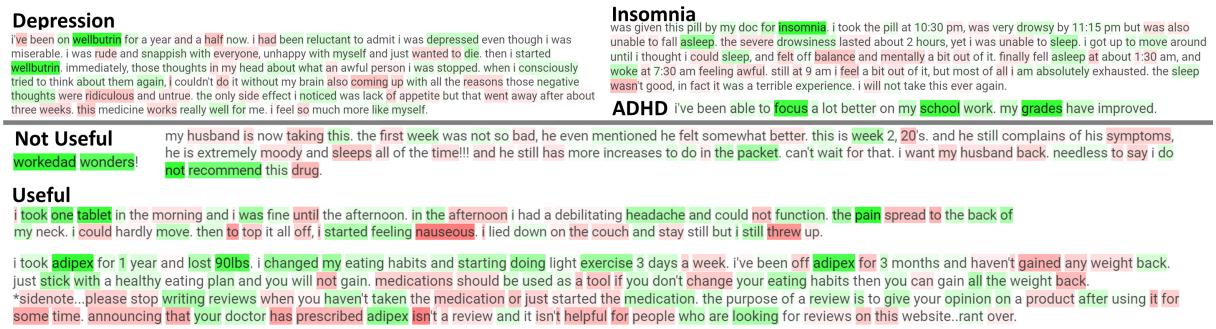

Figure 3: On top, we show attention maps for condition predictions, and at the bottom, ones for usefulness predictions. Green is for words that push models towards their prediction and red for ones that push them away.

fulness. We found that review usefulness varies by condition, age, and rating, and that models that combine both the review text and metadata perform best. Furthermore, we showed how a DistilBERT model trained on a regression task using review text and metadata can be applied to classification by bucketing its predictions. Finally, we analyzed the attention maps for specific reviews using fine-tuned DistilBERT models to predict condition and usefulness. Overall, we found that the attention maps for condition were often more interpretable than those for usefulness, since a given review may be *not* useful due to the absence of key information. Some potential future directions for our work include multi-task learning and applications to other review datasets or review prioritization systems.

References

Chris Forman, Anindya Ghose, and Batia Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information systems research*, 19(3):291–313.

Anindya Ghose and Panagiotis G Ipeirotis. 2010. 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.

Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health*, pages 121–125.

Sirkka L Jarvenpaa and Dorothy E Leidner. 1998. Communication and trust in global virtual teams. *Journal of computer-mediated communication*, 3(4):JCMC346.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhijie Lin. 2014. An empirical investigation of user and system recommendations in e-commerce. *Decision Support Systems*, 68:111–124.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *2008 Eighth IEEE international conference on data mining*, pages 443–452. IEEE.

Ying Liu, Jian Jin, Ping Ji, Jenny A Harding, and Richard YK Fung. 2013. Identifying helpful online reviews: a product designer's perspective. *Computer-Aided Design*, 45(2):180–194.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ruhui Shen, Jialiang Shen, Yuhong Li, and Haohan Wang. 2016. Predicting usefulness of yelp reviews with localized linear regression models. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 189–192. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ying Zhang, Khim Yong Goh, and Wang Qingliang. 2013. Unraveling the information role of online reviews: Distinguishing between the competing effect, local and global peer effects on consumer choice.