Is Review Content the Most Important Feature for Jointly Predicting Review Helpfulness and Sentiment?

Introduction

In the past decade, online reviews have received a significant amount of attention from the research and data science communities. One research area that has been heavily explored is predicting the rating of online reviews. Online reviews are typically rated on a 5-star scale, and many researchers have built models to predict the rating of online reviews using various features (Martin, 2016; Chen et al., 2017). Another research area that has received significant attention is related to predicting helpful online reviews. Several platforms provide "helpful" or "unhelpful" buttons for readers to rate the quality of reviews. Researchers studying review helpfulness aim to predict which reviews will acquire either the most helpful reviews, or the highest ratio of helpful to unhelpful reviews (Korfiatis et al., 2012).

Researchers have shown that the rating of Amazon reviews play an important role in predicting helpfulness (Quaschning et al., 2014). Therefore, several studies have demonstrated that models that predict review helpfulness may benefit from jointly learning review helpfulness and valence, as these models learn to better incorporate features related to valence in their predictions of helpfulness when trained on both tasks (Fan et al., 2018; Du et al., 2020; Amazon. Du et al., 2020).

Most studies that jointly predict helpfulness and valence only utilize the content of a review for the prediction task. Online reviews typically contain more information than solely the review content. This information can include a "review summary", a short, high-level overview about the reviewer's thoughts about a product, as well as the date the review was posted. It is unclear from these studies if the review content itself is the most important feature for jointly predicting helpfulness and valence, or if non-content features, such as the review summary and date would better perform this task. Moreover, it is unclear to what extent performance would increase by incorporating non-content-related information into models that jointly predict valence and helpfulness. These issues lend themselves nicely to the following research questions: Is the review content itself the most important feature for jointly predicting helpfulness and valence? Also, does incorporating non-content-related features improve performance for this prediction task, and to what extent? Additionally, since models that predict both helpfulness and valence are predicting two tasks simultaneously, it seems plausible that performance gains from incorporating non-content features may be more pronounced for one task than another. Consequently, as a third research question, I will investigate if incorporating non-content-related information provides a more pronounced improvement in model performance towards predicting helpfulness or towards predicting valence.

To answer these research questions, I will train machine learning models to jointly predict the helpfulness and valence of Amazon reviews by formulating the problem as a six-label classification task. The features used for this classification task will include the review content,

the review summary, and the year the review was posted. I will train models using several combinations of these features to determine which features are most useful for predicting helpfulness, as well as if including non-content-related features in addition to content features improves model performance.

Literature review:

There is a plethora of studies that used machine learning models to predict the helpfulness or the rating of Amazon reviews. Most studies only predict one of these tasks (Alsmad et al., 2020; Chen et al., 2017). Some studies have examined the role of the review summaries for predicting helpfulness or valence, respectively. Malik (2020) determined that classifiers that utilized both the review summary and content performed better than classifiers that used review content alone. Zhou et al. (2020) discovered that reviews received more helpful votes when the content of the review was similar to the summary. Mudambi et al. (2014) predicted review valence using the review content and summary and determined that there was occasionally a mismatch between the sentiment of the review and the star rating making review rating classification more challenging.

Various studies have also examined the role of timing on review helpfulness. Wan (2015) determined that the earliest posted reviews for a given product tended to acquire the most helpful votes and maintained their status as the "most helpful review" for the product's entire lifecycle on Amazon due to the Matthew effect. Lu et al. (2018) determined that helpfulness is a dynamic concept. As a result, they claimed the factors that influence helpfulness change over time (Lu et al., 2018).

The task of predicting review helpfulness and rating appear to be linked, since the characteristics of helpful reviews depend on the rating they receive (Quaschning et al., 2014). Fan et al. (2018) trained various models that jointly predict helpfulness and valence of reviews in the electronics category of Amazon. Du et al. (2020) determined that a CNN architecture could be effectively used to predict both star rating and helpfulness. Qu et al. (2020) trained BERT models to jointly predict review helpfulness and valence using a multitask loss function and achieved high classification accuracy on various categories of Amazon reviews. They also showed that their models could achieve decent accuracy predicting helpfulness on other platforms. All three of these studies used content alone to jointly predict the review's helpfulness and rating. From my review of literature, I could not find studies that use non-content features to jointly predict review helpfulness and valence.

Methods

Data Cleaning

I used the Amazon 2014 dataset of He & McAuley (2016) to answer my research question. The reason for selecting the 2014 dataset instead of the most recent one from 2018 is because the 2018 dataset has no information about the number of "unhelpful" votes a review received, unlike

the 2014 dataset. Most studies, including recent ones, appear to use the 2014 dataset to calculate the ratio of helpful to unhelpful votes and calculate helpfulness this way.

I originally started with three categories from the 5-core subset of reviews: Movies and TV, books, and electronics. My data cleaning steps were similar to those of many existing studies in literature examining review helpfulness (Alsmad et al., 2020; Passon et al., 2018). I began by removing all reviews that had a sum of helpful and unhelpful votes less than ten. I subsequently removed all non alphabetical symbols, lowercased all letters, and tokenized the text from the review content and summary. I also extracted the year from the review timestamp and normalized the value to be between 0 and 1. Next, I filtered all reviews whose content was less than 25 words and more than 500 words. I then removed all duplicate reviews by removing reviews that had identical content. After this step approximately 920,000 reviews remained.

Of the remaining data, helpful reviews were defined as those that received more than or equal to 70% positive votes out of their total number of votes. All reviews that did not reach this threshold were considered unhelpful. This was a similar approach to the study of Alsmad et al. (2020). Next, positive valence was defined as reviews with four or five stars, neutral valence as three stars, and negative valence as one or two stars. The reviews were then naturally categorized into 6 classes, based on the combinations of helpfulness and valence they received. Originally the electrical dataset was to be included in my analysis. However, due to small amounts of data for the unhelpful positive and unhelpful neutral classes, I removed this category to ensure I could have similarly large datasets for categories included in my analysis. I subsequently downsampled all classes to the category and class with the least amount of data and rounded down to the nearest multiple of 1000 for convenience. After downsampling, 9,000 examples remained per class for each category, creating a final dataset of 54,000 reviews per category. Next, the data was split into a training, development and test set with balanced classes. The test and development sets contained 1000 examples per class for each category, and the training set the remaining 7000.

Summary Statistics and Exploration

I performed various steps for my exploratory analysis. I originally computed the Jaccard similarity using all words in the vocabulary. However, the Jaccard similarities between all classes were nearly identical, likely because there were so many rare words in the vocabulary of each class. Therefore, I computed the Jaccard similarity between classes while filtering out all words that appeared in less than 5 documents. The Jaccard similarity matrix for the Movies and TV category is shown in Figure 1. The matrix of Jaccard similarities for the book section looks very similar and is provided in the appendix. From this matrix it appears that the most separable classes are unhelpful negative sentiment review, and those that are helpful with positive sentiment. The most similar classes are helpful, neutral reviews and helpful positive sentiment. Therefore, I expect my classifier to misclassify these two classes the most, and this should be reflected in my confusion matrices. This type of error appears to be related to the ability to distinguish sentiment, rather than helpfulness, so it seems that my classifiers may struggle to

accurately determine the sentiment of reviews, rather than the helpfulness when using review content alone.

I also computed the Jaccard similarity using the review summaries to determine which classes summary-based features would struggle distinguishing. From Figure 2 it appears that for a given valence, the Jaccard similarity between classes is highest between classes with the opposite helpfulness label. For instance, negative helpful reviews have the highest Jaccard similarity with negative unhelpful reviews. Consequently, it appears a classifier that utilized summary-based features would struggle to classify helpfulness but may perform well classifying review sentiment.

I also generated rank frequency plots to examine the distribution of word frequencies of the review content. From Figure 3, we can see that the rank frequency plots of both categories of reviews were similar and had long tails. This was also the case for the rank frequency plots of the reviews associated with each class (found in the appendix). Therefore, it is clear there were many words in both categories that occurred only a few times. I have no reason to believe that accurate counts for rare words are necessary for my classification task, so I do not anticipate that this will affect my model's performance.

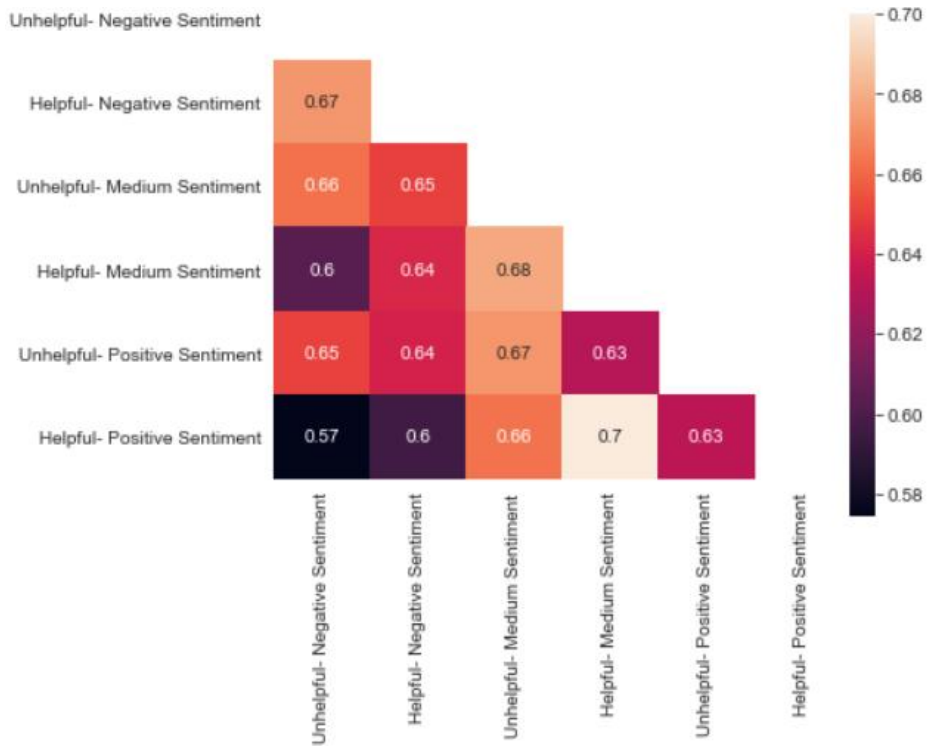

Figure 1: Heat map of Jaccard similarities for the review content for movies and TV category after filtering words with a document frequency less than 5. A similar heatmap for the books section is similar and is provided in the Appendix

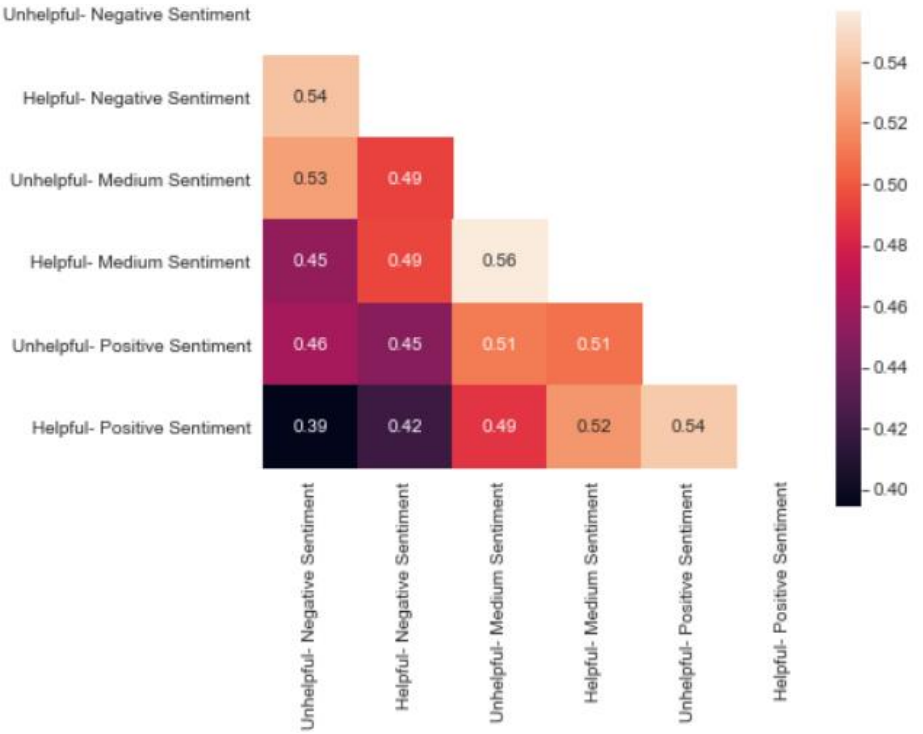

Figure 2: Heat map of Jaccard similarities for the review summaries for movies and TV category after filtering words with a document frequency less than 5. A similar heatmap for the books section is similar and is provided in the Appendix

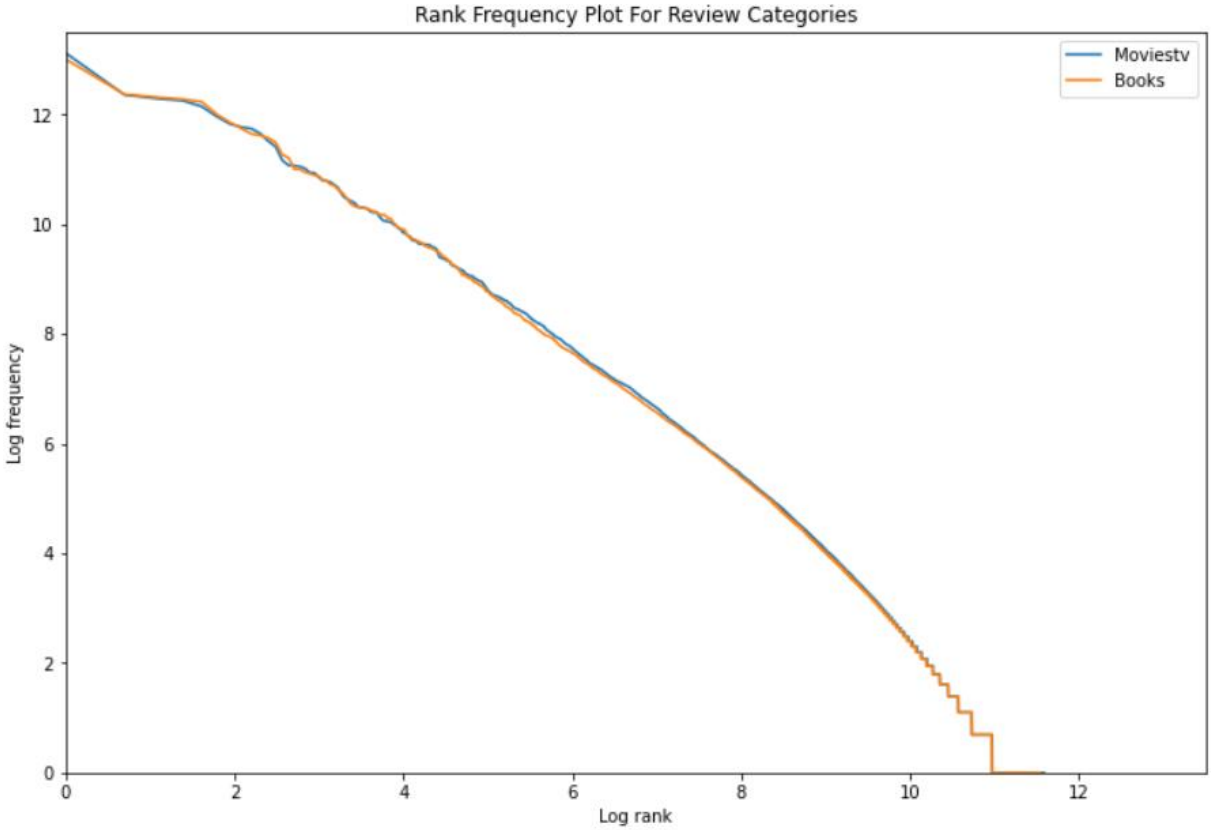

Figure 3: Rank frequency plot of words in review content by category

Next, I examined the distribution of years reviews were posted in my dataset for each category. Figure 3 shows that the number of reviews for the Movies and tv category peaks in 2005, and gradually declines thereafter. The number of reviews for the book category increases until 2012. There is a steep drop off for both categories in 2014 since the data was collected in the middle of that year.

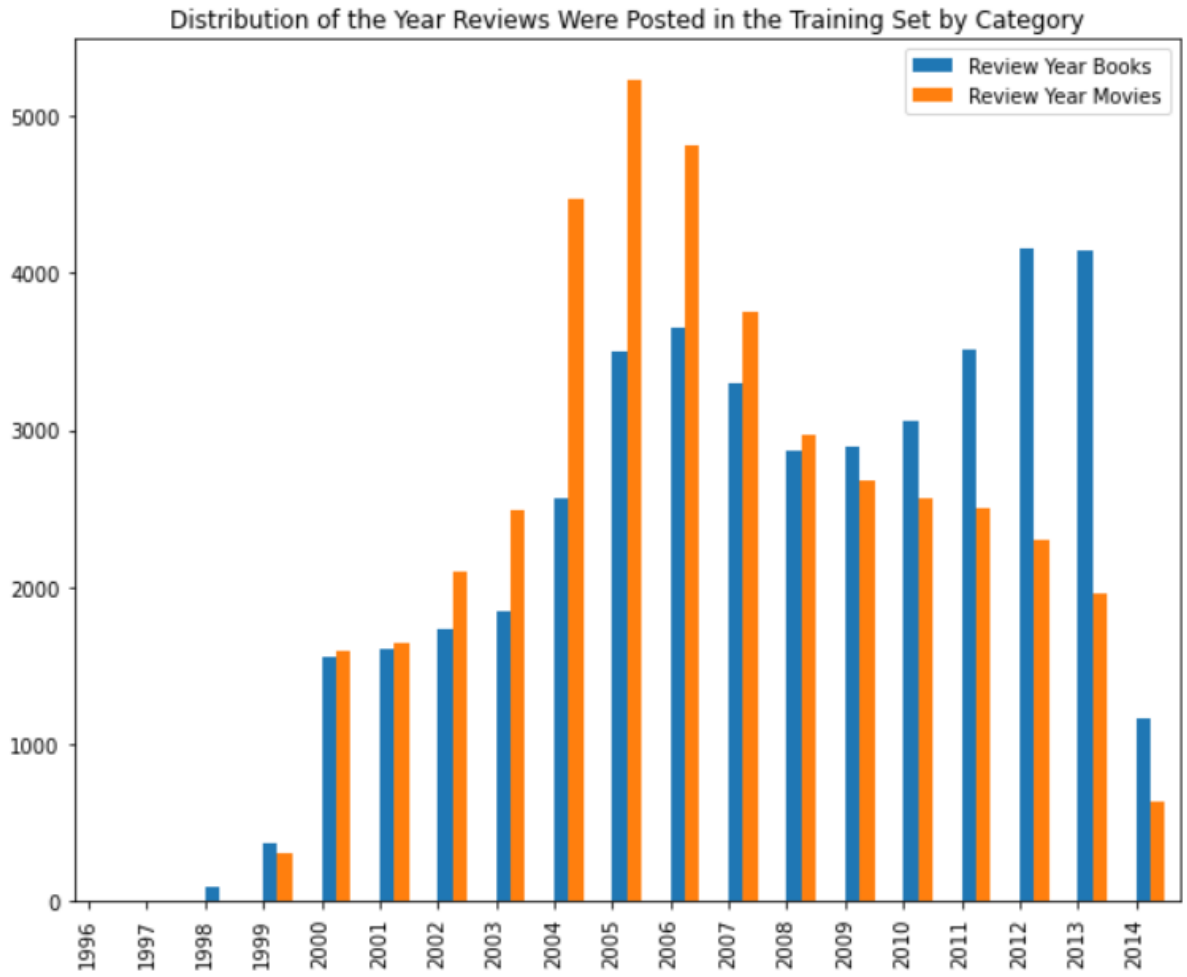

Figure 3: Distribution of the year reviews were posted for each category.

As part of my exploratory analysis, I also examined several metrics to compare the review content and summary to determine how review summaries may provide additional information for predicting review sentiment and helpfulness. First, I measured the degree which the sentiment of review content and the review summaries differ. To this end, I utilized Valence Aware Dictionary and sEntiment Reasoner (VADER) for sentiment analysis (Hutto & Gilbert, 2014). This tool was designed for social media but has been shown to work relatively well for Amazon reviews (Nguyen & Veluchamy, 2018). For each review, I calculated the absolute difference between the VADER sentiment between the review content and summary. This VADER sentiment score can be between -1 and 1, therefore I considered a difference of more than 1 to be a different sentiment. Table 1 shows that a noticeable percentage of reviews have

summaries and contents with different sentiments. Therefore, it seems plausible the review summaries provide additional context related to review sentiment that the content alone is lacking.

	Percent of Reviews Where the Sentiment Between the Summary and Content is Different	Percent of Reviews Where the Readability Between the Summary and Content is Different
Movies and TV	16.36	55.29
Books	15.46	55.23

Table 1: Measuring the differences between the review content and summary in terms of sentiment and readability. The sentiment column is the percentage of reviews where the content and summary have a different VADER sentiment score of more than 1. The readability column is the percentage of reviews where the content and summary have a different Flesch–Kincaid readability score of more than 6.

Next, to measure how much the review summary might contribute additional information to predicting helpfulness, I compared the reading ease between the review content and summaries. It has been shown in past research that readable reviews receive better helpfulness scores on Amazon (Korfiatis et al., 2012). Therefore, I calculated the absolute difference between the Flesch–Kincaid readability score (Jones, 1952) of the review content and summary for each review. The score corresponds to the average US grade level that could easily read the text. Therefore, I decided that the readability between the content and summary were different if the absolute difference of these was less than a threshold. The threshold was chosen somewhat arbitrarily as 6, as I believed a difference of more than 6 US grades was sufficient to determine if the difference in readability was significant. From Table 1 it appears that the readability of the review content and summary differ substantially, where the majority of reviews have content that is at least 6 US grades easier (or more challenging) to read. However, this could be due to several reasons. For instance, many review summaries are simply one word, which would receive a very low Flesch–Kincaid grade. Overall, it is unclear to what extent the summary may improve helpfulness or rating predictions without training and comparing models, so I proceeded to create models to jointly predict helpfulness and valence.

Models

Five models were trained for the 6-label classification task. These models included naïve bayes, logistic regression, a 1-hidden layer feed forward neural network, a Long Short Term Memory (LSTM) classifier, as well as a bidirectional LSTM. Each of these models have been used in existing literature on predicting review helpfulness or review valence (Alsmad et al., 2020; Chen et al., 2017; Du et al., 2020). There does not appear to be any consensus in literature on the best baseline model for these tasks, so I simply considered naïve bayes as my baseline. Although state of the art accuracy was not the objective of my study, I will also compare my accuracy to an existing study that jointly predicted helpfulness and review valence on the same review categories as mine (Qu et al., 2020).

Naïve bayes was trained with smoothing, and the optimal smoothing hyperparameter was obtained using 5-fold cross validation. The logistic regression and 1-hidden layer feedforward

neural network models were trained using TFIDF vectors as the input features. These models also included the length of the review as a feature, as it was found that these models performed significantly better when I included this feature. The length of the review is related to the content of the review itself, therefore it was considered as a content feature. I acknowledge that the comparison to the naïve bayes model is not entirely fair due to this additional feature, but the accuracy of the naïve bayes model was still competitive even without this feature. The models were trained using cross entropy loss, and the Adam optimizer.

Two hyperparameters related to the TFIDF vectors needed to be tuned. These were the maximum number of features and the maximum document frequency hyperparameters. The maximum features hyperparameter is self explanatory and is related to the maximum number of features included in the TFIDF vector. The "maximum document frequency" hyperparameter is used to filter words that occur too often in documents and likely provide minimal information for a classification task. Due to long run times on my local machine, the optimal hyperparameters were obtained from the set of hyperparameters that maximized the accuracy of the development set rather than cross validation. Not all hyperparameters could be tuned due to long run times. Therefore, the optimal hyperparameters for the learning rate and number of hidden layers for the 1-hidden layer neural network were determined heuristically. The parameters used for these models can be found in the appendix. Once the optimal hyperparameters were obtained, I trained the models on the test set with early stopping. Early stopping was used to help prevent overfitting, by stopping the training process once the validation loss stopped decreasing for more than five epochs. I did not use dropout for either model, because I found they were not overfitting, especially with early stopping.

The unidirectional and bidirectional LSTM models were trained using the pretrained wiki-news-300d FastText embeddings from the work of Mikolov et al. (2018). I originally wanted to fine tune these word embeddings, as Alsmad et al. (2020) achieved promising results for predicting review helpfulness by fine-tuning FastText word embeddings on a large amount of Amazon reviews. However, due to limited computational resources, I decided to use the embeddings out of the box for my prediction task. Therefore, I could expect better results by fine tuning the embeddings.

The LSTM models contained one hidden layer (with dropout) and were trained using cross entropy loss, and the Adam optimizer. While training the models using both review content and review summary, the original objective was to have two separate LSTM layers and concatenate the outputs of both LSTM layers into a single hidden layer. However, it was determined while training the model in this manner that the optimizer would typically get stuck at a local minimum, and the development set would experience sudden and substantial drops in accuracy, with seemingly no warning. A similar issue was encountered while attempting to concatenate the date information to the hidden layer. Therefore, instead of combining features into a single model, I built an ensemble of LSTM models, where each LSTM was trained using one type of review feature, the review content and the review summary. I decided to exclude date information, since it did not make sense to build an LSTM model that utilized date information alone. Therefore, for LSTM models, only 3 combinations of features were used, the content

alone, the summary alone, and a combination of both. The ensemble was created by summing the output probability of both models (after a softmax layer) and taking the class with the highest summed probability. The hyperparameters used for the LSTMs can be found in the appendix.

Overall, for each category of reviews, the naïve bayes, logistic regression and neural network models were trained using five combinations of features. The unidirectional and bidirectional LSTMs were each trained using three combinations of features. Therefore, a total of 21 models were trained per category. After training these models I extracted accuracies on the review sentiment alone, as well as the review helpfulness. This was done to determine if non-content features provided more information towards resolving review helpfulness, or review valence.

Results:

The naïve bayes models used as my baselines achieved competitive accuracies compared to other models across all features. Logistic regression and the neural network only slightly outperformed Naïve bayes and used the review length as an additional feature. The LSTMs appear to be the best performing models, where both the LSTM models trained on content alone achieve higher accuracy than all other models, even with all features included. Results for each category are shown in Tables 2 and 3. I have included confusion matrices in the appendix.

	Content Alone	Non-Content Alone	Content and Summary	Content and Year	All Features
Naïve Bayes	41.45	35.17	43.08	41.40	43.42
Logistic Regression	42.42	32.63	43.88	42.68	44.00
Neural Network	41.88	32.63	43.88	42.22	44.40
LSTM	44.93	34.87			46.77
Bidirectional LSTM	46.93	36.05			48.72

Table 2: Accuracies for each model and combination of features for movies and tv reviews

	Content Alone	Non-Content Alone	Content and Summary	Content and Year	All Features
Naïve Bayes	37.75	35.38	40.43	37.80	40.48
Logistic Regression	40.65	31.68	42.43	41.05	42.93
Neural Network	40.67	30.58	42.10	40.95	41.88

LSTM	45.72	37.35			47.72
Bidirectional LSTM	44.57	36.13			46.97

Table 3: Accuracies for each model and combination of features for book reviews.

I report another set of results related to the accuracy of my models for predicting helpfulness and review valence individually. For instance, in the case of helpfulness, if my classifier successfully predicted the helpfulness of a review, irrespective of the review's valence, it would be considered as an accurate prediction in Tables 4. This table only includes models that were trained using content alone, as well as those that used review summaries, to determine the relative gain of including summaries towards either helpfulness or valence prediction. This table also only includes reviews from the movies and tv category. The same table for the book category can be found in the appendix.

	Helpfulness Accuracy			**Sentiment Accuracy**		
	Content Only	All Features	Difference in Accuracy	Content Only	All Features	Difference in Accuracy
Naïve Bayes	67.97	68.35	0.38	58.27	61.27	2.77
Logistic Regression	68.19	68.12	-0.07	59.8	62.23	2.43
Neural Network	67.40	67.98	0.58	59.90	62.52	2.61
LSTM	67.73	67.63	-0.22	66.07	68.78	2.72
Bidirectional LSTM	68.77	69.10	0.33	66.88	69.67	2.78

Table 4: Relative difference in accuracies after including non-content features for the movies and TV category. The same table for the book category can be found in the appendix.

In terms of comparing my models to the results of Qu et al. (2020), I only compare accuracy in terms of helpfulness, as this study only reports their results on helpfulness. This is because the authors wanted to compare their models to existing benchmarks that predicted helpfulness alone, even though they jointly predicted valence and helpfulness. A comparison of my results to their study is shown in Table 5.

	Naïve Bayes	Logistic Regression	Neural Network	LSTM	Bidirectional LSTM	(Qu et al., 2020) BERT

Movies and TV	68.35	68.12	67.98	67.63	69.10	**75.30**
Books	63.98	64.80	64.45	65.37	65.42	**71.25**

Table 5: Comparing my helpfulness prediction accuracy using all features to the work of Qu et al. (2020). They achieved higher helpfulness accuracy than my models but used a BERT model.

Discussion

It can be seen from Tables 2 and 3 that Models that utilize the content alone achieve much higher accuracy than those that use solely non-content features. Therefore, this provides an answer to my first research question, where the review content itself appears to be the most important feature for jointly predicting helpfulness and valence. In terms of my second research question, it appears that adding non-content-related features in addition to content features improves model performance. The review summary appears to provide noticeable increases in model accuracy. Although the date was only used for three models, it appears this feature provided minimal gain in terms of accuracy compared to the review summary. For instance, the best performing non-LSTM models utilize all three types of features, but they only perform marginally better than just using the review content and the summary.

After examining the confusion matrices (found in the appendix), it is clear that all my models struggle with neutral reviews. This makes sense, as the definition of a neutral review is narrower than positive or negative reviews, since class membership for neutral reviews is defined as exactly three stars, whereas negative and positive reviews are defined as one to two and four to five stars, respectively. Unlike my expectations from the Jaccard similarity matrices, it appears that models trained on content alone do not struggle most with differentiating between helpful neutral and helpful positive reviews. Instead, all models appear to struggle with unhelpful neutral reviews, and confuse this class with helpful or unhelpful negative reviews.

I examined the errors made by my naïve bayes classifier more closely to better understand the kinds of classification errors that are occurring related to sentiment and valence, respectively. Errors related to the review valence appear to occur in cases where the models have not properly learned if the reviewer is referring to the quality of the book or movie itself, or an event that took place in the movie or book. For instance, there are several cases where the classifier incorrectly predicted a negative review because the reviewer detailed negative events that happened within a book, while they still enjoyed reading the book overall. In terms of errors related to helpfulness, it appears the models struggle to classify reviews that contain controversial or sensitive topics, such as religion or politics. Examples of the reviews that were misclassified on the basis of sentiment as well as helpfulness can be found in the appendix.

Table 4 highlights that the performance gain from incorporating non-content features into my models are primarily in terms of increased accuracy towards sentiment prediction, rather than helpfulness. This was expected from my exploratory analysis using the Jaccard similarities of the

review summaries, since the Jaccard similarity was high between classes with different levels of helpfulness and relatively low between classes with different levels of valence. Therefore, it appears the answer to my third question is that non-content information provides the most additional information towards predicting review sentiment, rather than helpfulness.

I wanted to better understand what kinds of corrections were made towards my models' valence predictions after including information related to the summaries. To this end, I examined reviews where my naïve bayes model trained on content features alone failed to accurately predict the correct rating, but the model trained on both the content and summary succeeded. From these examples, it appears that the summary occasionally provides useful information related to the sentiment of the review in cases where the sentiment in the content of the review may be ambiguous. For instance, one of the reviews had the following content: "*This book was the first finance book I ever read back in the 90s now. I loved it at the time. I recently bought a copy and it just does not really grab me. I would suggest something by Swedroe or Bogle*". In my opinion, it is unclear whether this review is neutral or negative (my Naïve bayes classifier predicted neutral). However, the summary makes the negative sentiment significantly clearer, as the reviewer wrote "*Eh kinda sucks now*". Other examples are provided in the appendix.

It can be seen in Table 5 that my models do not perform as well in terms of predicting helpfulness as Qu et al. (2020). This could be due to various reasons, but one evident reason is this study used a BERT model, which has been shown to achieve state-of-the-art performance for a variety of NLP tasks. Another possible reason could be that I did not use a multitask loss function to properly weigh the different error types, as they did in their study. Multitask loss has been shown to often increase performance for a variety of applications, so it is reasonable to believe that incorporating this loss function would have improved my model's performance.

Limitations

My study has various limitations that limits the generalizability of the findings. For instance, I only examined online reviews on Amazon, and reviews on this platform are not representative of reviews on other platforms. Moreover, many platforms do not even have an option to vote reviews as unhelpful, so classes could not be categorized in the same manner as I performed in my study. Additionally, I only included two categories of Amazon reviews, so it is unclear if these results would hold for other categories. There is clearly a noticeable difference in classification accuracy for my model on both categories of products I analyzed, so it is possible results would be different on other categories of goods. I also only included reviews with lengths between 25 and 500, as well as reviews with at least ten total votes. Therefore, my models would likely perform poorly on very short or long Amazon reviews, as well for low-traffic items where reviews for these items tend not to acquire many helpful or unhelpful votes (Wan, 2015). Another limitation to my study is that I would need to change the definition of helpfulness if I wanted to run my analysis on more recent Amazon reviews, since Amazon discontinued its "unhelpful" button in 2018.

An additional limitation is related to the review date features I used in my study. I only incorporated the year the review was posted in my analysis, so I could have provided more granularity related to the timing of the review. Furthermore, the date information on its own likely does not provide much information without context on when the product being reviewed was first posted on Amazon. The work of Wan (2015) determined that reviews posted early for a *given product* received more helpfulness votes. Therefore, it appears the timing of when the review was posted on its own may provide much less information compared to the relative timing between the review posting date and the product posting date. These issues were likely part of the reason the date provided little information gain in my analysis, and future work should address these issues to better understand the effects of including the date for jointly predicting review helpfulness and valence.

Another limitation to my study is the fact that I did not fine tune the FastText embeddings I used for my LSTM models. It has been shown in previous work that fine tuning embeddings could significantly improve accuracy for predicting review helpfulness (Alsmad et al., 2020). Therefore, the performance of my models likely could have improved further by fine-tuning the FastText embeddings on a large number of Amazon reviews.

Conclusion

In this study, I trained various models to jointly predict review helpfulness and valence on the movies and tv as well as book category of Amazon reviews. Five models were trained using various combinations of content and non-content related features to determine if content was the most important feature for this joint prediction task. I determined that content was the most important feature, but the review summary provided additional information that was not always available in the review content alone, resulting in improved performance when these two types of features were combined. This increase in performance was towards better predicting the sentiment of the review itself, rather than predicting the review helpfulness. Despite the limitations of my study, I believe it still provides useful findings related to additional features that can be used to improve the accuracy of existing models that jointly predict review valence and helpfulness.

References

Martin, M. (2016). Predicting ratings of Amazon reviews - Techniques for imbalanced datasets.

Chen, M., & Sun, Y. (2017). Sentimental Analysis with Amazon Review Data. *European Conference on Information Retrieval*, 127–138.

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205–217. https://doi.org/10.1016/j.elerap.2011.10.003

Mudambi, S. M., Schuff, D., & Zhewei Zhang. (2014). Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings. *2014 47th Hawaii International Conference on System Sciences*. https://doi.org/10.1109/hicss.2014.389

Malik, M. S. (2020). Predicting users' review helpfulness: the role of significant review and reviewer characteristics. *Soft Computing*, *24*(18), 13913–13928. https://doi.org/10.1007/s00500-020-04767-1

Zhou, Y., Yang, S., li, yixiao, chen, Y., Yao, J., & Qazi, A. (2020). Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Information Processing & Management*, *57*(2), 102179. https://doi.org/10.1016/j.ipm.2019.102179

Quaschning, S., Pandelaere, M., & Vermeir, I. (2014). When Consistency Matters: The Effect of Valence Consistency on Review Helpfulness. *Journal of Computer-Mediated Communication*, *20*(2), 136–152. https://doi.org/10.1111/jcc4.12106

Wan, Y. (2015). The Matthew Effect in social commerce. Electronic Markets, 25(4), 313–324. https://doi.org/10.1007/s12525-015-0186-x

Fan, M., Feng, Y., Sun, M., Li, P., Wang, H., & Wang, J. (2018). Multi-Task Neural Learning Architecture for End-to-End Identification of Helpful Reviews. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. https://doi.org/10.1109/asonam.2018.8508623

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, *48*, 17–27. https://doi.org/10.1016/j.chb.2015.01.010

He, R., & McAuley, J. (2016). Ups and Downs. *Proceedings of the 25th International Conference on World Wide Web*. https://doi.org/10.1145/2872427.2883037

Jones, R. L. (1952). How to test readability. *Journal of Applied Psychology*, *36*(2), 144–144. https://doi.org/10.1037/h0050326

Lu, S., Wu, J., & Tseng, S.-L. (A. (2018). How Online Reviews Become Helpful: A Dynamic Perspective. *Journal of Interactive Marketing*, *44*, 17–28. https://doi.org/10.1016/j.intmar.2018.05.005

Passon, M., Lippi, M., Serra, G., & Tasso, C. (2018). Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining. *Proceedings of the 5th Workshop on Argument Mining*. https://doi.org/10.18653/v1/w18-5205

Qu, X., Li, X., Farkas, C., & Rose, J. (2020). An Attention Model of Customer Expectation to Improve Review Helpfulness Prediction. *Lecture Notes in Computer Science*, 836–851. https://doi.org/10.1007/978-3-030-45439-5_55

Du, J., Zheng, L., He, J., Rong, J., Wang, H., & Zhang, Y. (2020). An Interactive Network for End-to-End Review Helpfulness Modeling. *Data Science and Engineering*, *5*(3), 261–279. https://doi.org/10.1007/s41019-020-00133-1

Alsmad, A., AlZu'bi, S., & Al-Ayyoub, M. (2020). Predicting Helpfulness of Online Reviews.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media*, *8*.

Nguyen, H., & Veluchamy, A. (2018). Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches. *SMU Data Science Review*, *1*(4).

Mikolov, T., Grave, E., & Bojanowski, P. (2018). Advances in Pre-Training Distributed Word Representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Appendix:

Jaccard Similarities for Books category

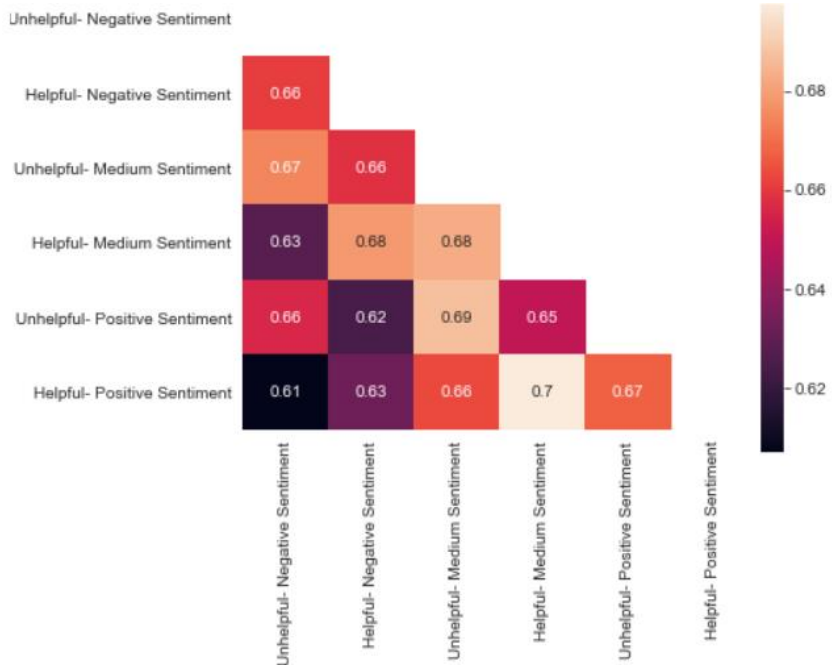

Figure A1: Jaccard similarities for review content of books section

Figure A2: Jaccard similarities for review summaries of books section

Rank frequency plots per class

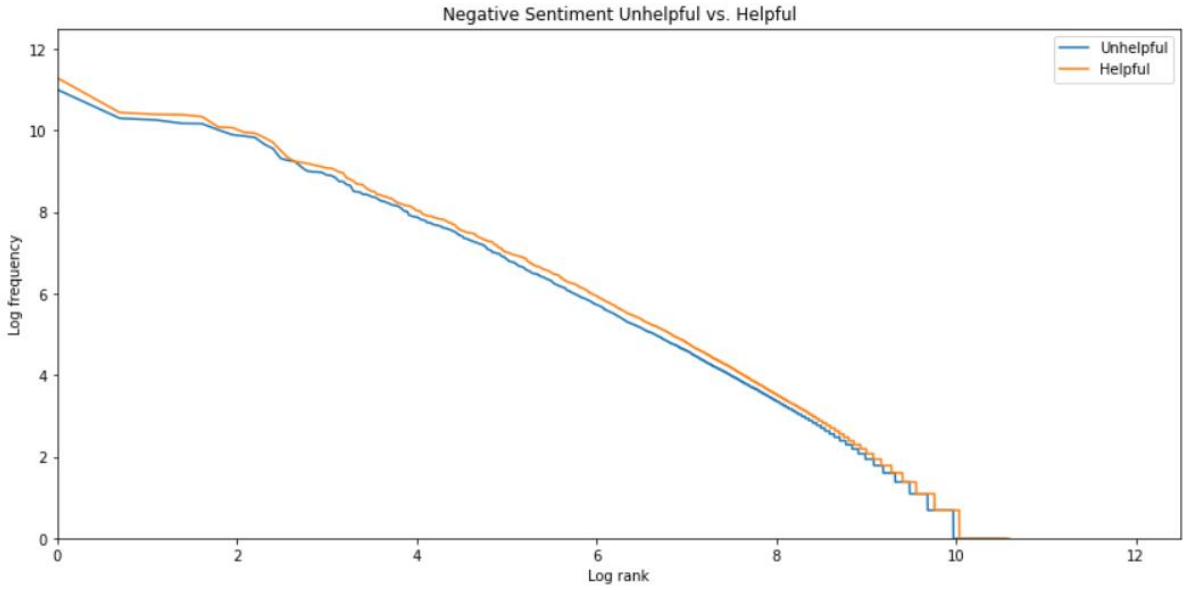

Figure A3: Rank frequency plot of negative sentiment reviews

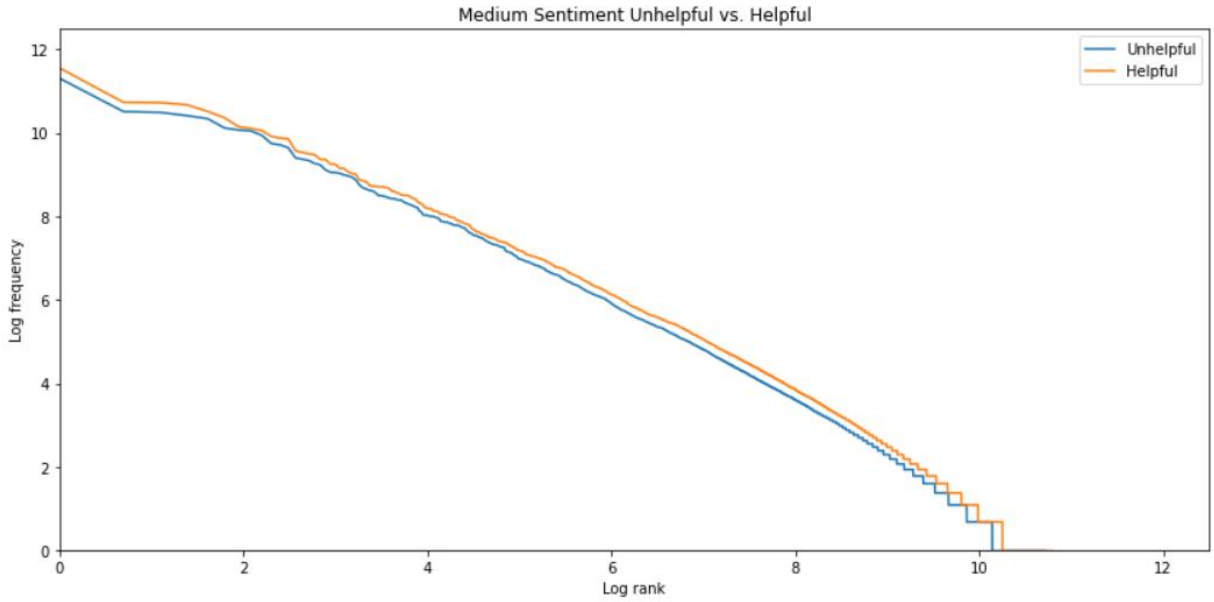

Figure A4: Rank frequency plot of neutral sentiment reviews

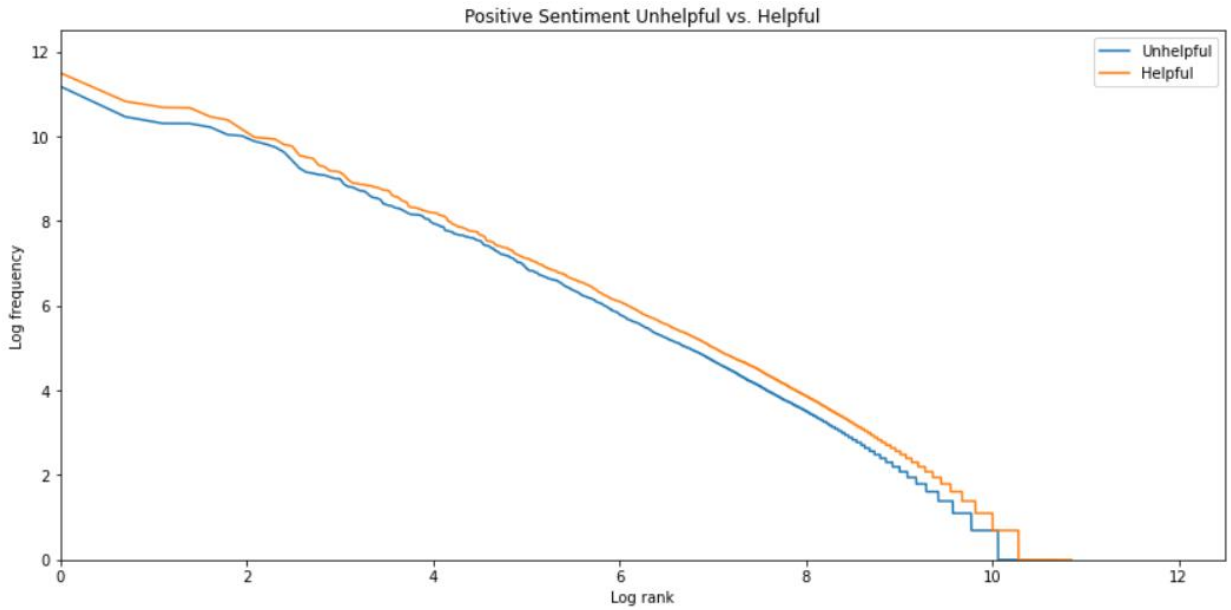

Figure A5: Rank frequency plot of positive sentiment reviews

Model Hyperparameters

Max features	Maximum document frequency	Learning rate
2000	0.7	0.02

Table A1: Logistic regression hyperparameters

Max features	Maximum document frequency	Learning rate	Hidden Neurons
2000	0.7	0.02	500

Table A2: Neural Network hyperparameters

Dropout	Learning rate	Hidden Neurons
0.25	0.002	200

Table A3: unidirectional LSTM hyperparameters

Dropout	Learning rate	Hidden Neurons
0.25	0.002	400

Table A4: bidirectional LSTM hyperparameters

Confusion Matrices

Due to the large number of models trained, I have only included a subset of 2 confusion matrices in this section. These confusion matrices are from the Naïve Bayes and bidirectional LSTM models that were trained on content alone in the movies and tv section.

Predicted class / True class	Unhelpful and negative rating	Helpful and negative rating	Unhelpful and neutral rating	Helpful and neutral rating	Unhelpful and positive rating	Helpful and positive rating
Unhelpful and negative rating	55.5	17.2	13.8	4.2	5.8	3.5
Helpful and negative rating	20.3	51.8	9.0	12.2	2.9	3.8
Unhelpful and neutral rating	24.4	13.6	23.9	15.3	10.9	11.9
Helpful and neutral rating	8.1	22.7	16.2	28.8	4.3	19.9
Unhelpful and positive rating	11.8	16.2	11.9	6.7	35.4	18.0
Helpful and positive rating	4.0	9.9	8.9	13.3	15.8	48.8

Table A5: Confusion matrix on Naïve Bayes using content only and the movies and t category. It is clear that the model struggles most with predicting unhelpful, neutral sentiment reviews.

Predicted class / True class	Unhelpful and negative rating	Helpful and negative rating	Unhelpful and neutral rating	Helpful and neutral rating	Unhelpful and positive rating	Helpful and positive rating
Unhelpful and negative rating	34.5	27.8	15.9	8.2	12.1	1.5
Helpful and negative rating	18.7	45.1	12.4	19.2	3.1	1.5
Unhelpful and neutral rating	9.3	11.6	25.9	25.5	17.7	10.0
Helpful and neutral rating	4.1	10.5	16.1	48.6	5.8	14.9
Unhelpful and positive rating	5.1	1.1	7.9	5.9	58.5	21.5
Helpful and positive rating	1.3	0.9	5.0	8.9	29.1	54.8

Table A6: Confusion matrix on Bidirectional LSTM using content only and the movies and t category. This model also struggles most with predicting unhelpful, neutral sentiment reviews, similarly to Table A5

Accuracy Breakdown by Sentiment and Helpfulness for Books Category

This is the same plot as Table 4 in the main body of the report, except for the books category.

	Helpfulness Accuracy			**Sentiment Accuracy**			
	Content Only	All Features	Difference in Accuracy (DH)	Content Only	All Features	Difference in Accuracy (DS)	Relative Difference (DS-DH)
Naïve Bayes	63.30	63.98	0.68	57.22	61.48	4.27	3.59
Logistic Regression	65.83	65.20	-0.63	61.23	64.15	2.92	3.55
Neural Network	65.02	65.27	0.25	60.33	63.07	2.73	2.48
LSTM	65.00	65.42	0.42	69.20	72.43	3.23	2.81
Bidirectional LSTM	65.22	65.37	0.15	67.68	71.52	3.83	3.68

Table A7: Relative difference in accuracies after including non-content features for the books category.

Example classifier mistakes for sentiment

Example 1

Summary: “There's a Fine Line Between Madness and Genius”

Content “Right out of the gate this book creeped me out. Being a brunette myself, the title alone was enough to scare me! This was a seriously good book.Words like unhinged, grotesque, shocking, intriguing, and macabre were spinning through my head as I read the stunning and suspenseful pages. In this book, the lines between profiler and killer severely blur constantly. Cristyn West keeps you guessing, "Just WHO is the killer?" And just when you think you've figured it out, she takes you through to another labyrinthine twist.I found myself wondering, "Just who is watching you?" This was a creepy, gritty, horrifying and shocking ride. Right until the last page I found myself exclaiming, "No way!" This book will make you jumpy, especially if you're a brunette. You'll find yourself checking your doors and windows, and looking over your shoulder. You'll catch yourself watching everyone around you, wondering.....The suspense is drawn out to the last drop, right up until you discover the shocking truth! To catch a killer, just how far would you go? Read Plain Jane: Brunettes Beware and you will discover that there's a fine line between madness and genius!”

Predicted class Negative

Actual class: Positve

Example 2

Summary: “Dated, draggy, and just monumentally useless, even in japan”

Content: “For one thing, do not expect to learn anything about "Japanese Business", which seemed to have played out well for this title commercially in 1975, when Japan was this mysterious Godzilla across the Pacific.Nothing, that is, aside from some pithy insights such as: "Actually japanese companies do not really have a strategic planning capability, they usually have one person, or a few persons, who has/have an intuitive pulse of the market."Intrigued yet? There's more. You'll learn that strategy is the art of thinking on three major vectors: company based, customer based and competitor based. You can enjoy a truckload of charts and jargon. You can savor dated explanations of how American companies organize themselves and the anachronisms about Soviet-style central planning (I can recognize a relic when I see one.)Guess I bought an expensive paperweight. Do yourself a favor and ignore the drooling reviews this book has garnered as recently as last month. Look instead for names like Porter, Drucker and Mintzberg.”

Predicted class Positive

Actual class: Negative

Example 3

Summary: “Great collection of essays”

Content: “Piece by piece, these essays stack up to a devastating portrait of a man of fathomless selfishness. Sobran is the best columnist of latter 20th century, a man of profound, concise insights and a great writer. He's witty, too.This is a book you'll read again in 20 years when the revisionists have repainted Clinton as a flawed genius (or some other sort of twaddle). Hey, they did it with FDR...”

Predicted class Negative

Actual class: Positve

Examples classifier mistakes for helpfulness

Example 1

Summary: Shocking Truth

Content: This is a really good book. I flipped through the pictures and data mostly, but you couldn't help reading the text as well for the information is shocking. There are sharp comparisons between the past and now, in stunning photographic form: Alaska, glaciers around the world, coral bleaching, hurricanes, ocean currents, coastlines, lives affected by the damages and humans' continued invasion. Al Gore's conviction and unique access to information gave this book immense value

Predicted class Helpful

Actual class: Unhelpful

Example 2

Summary: Another Excellent Pagan Reference Book

Content: in the short time I've followed this path, I've had the opportunity to read a handful of wonderfully written and informative books about Wicca/Paganism. This book is no exception. While the dictionary of Celtic Gods and Goddesses, Hereos and Heroines which takes up most of the book can be a bit daunting to a newcomer like me, I have to also admit it's probably the most extensive one I've seen. Another reviewer pointed out that while McCoy points out the care that should be used in invoking certain deities, but does not go into greater detail, I too agree that the practioner should use care and caution when working with them (but isn't that true of doing any task?).I like too that this book explains in detail various rituals of evocation (house protection spell, money spell, etc.) and invocation (healing spell, eco-magick, for example), the numerous Sabbats (I had yet to learn before reading this book for example that Samhain is considered the beginning of the "New Year," not Yule), and Pagan Life Cycle events.This is a wonderful book for anyone like me who wishes to explore Celtic Paganism.

Predicted class Helpful

Actual class: Unhelpful

Examples Where the Sentiment was Corrected by Including the Review Summary:

Example 1

Review Summary: "Eh kinda sucks now"

Review Content: "This book was the first finance book I ever read back in the 90s now. I loved it at the time. I recently bought a copy and it just does not really grab me. I would suggest something by Swedroe or Bogle"

Predicted sentiment from content alone: neutral

Actual sentiment and predicted using both content and the title: negative

Example 2

Review Summary: "Rather Disappointing!"

Review summary: “Dr. Lee's book begins with the well-known Scott Peterson and Elizabeth Smart cases. He details all the evidence collected (and sometimes not collected), and takes readers from the crime to the trial. Unfortunately, forensic evidence did not play a role in either case. In Scott Peterson's case, Dr. Lee concluded that it wasn't the evidence that did Peterson in - rather his post-disappearance actions and court attitude. On the other hand, it was interesting to not that the original jury foreman was removed (reason unknown) - since he was both an attorney and an M.D. the jury might have been led to a greater focus on the inconclusive evidence and the verdict turned out differently.As for Elizabeth Smart's nine-month disappearance, the case seemed to have been solved in spite of the Salt Lake City Police. Elizabeth's younger sister (in the same room when the kidnapping took place) was convinced that the man police suspected was not the one, identified the correct individual, and helped in the drawing of his portrait - thus, leading to Elizabeth's safe return. (The kidnapper and his wife were judged mentally incompetent for trial; nonetheless, they had brainwashed Elizabeth so much that she did not try to escape, and originally denied that she was the one everyone was looking for.)The third case involved an individual whose wife was found dead at the bottom of the stairs. While Dr. Lee was called as a witness for the defense, it was not enough to overcome the eerie fact that the defendant's first wife had similarly died, and that her injuries seemed to great for having simply fallen part-way down the stairs.The fourth case was quite straight-forward - an arson, followed by the murder of a key witness, and the fifth, while rather salacious, was also not that challenging.”

Predicted sentiment from content alone: positive

Actual sentiment and predicted using both content and the title: neutral

Example 3

Review summary: I feel cheated.

Review content: “The Talisman is one of my most favorite Stephen King novels, and when I found out there was going to be a sequel I was thrilled. However, I hated Black House. If you are not familiar with King's Dark Tower series, you probably won't appreciate this book. I'm very impressed with King's ability to interweave so many stories, however, I didn't want to have to read a bunch of other things, i.e. The Gunslinger, The Drawing of the Three, The Wastelands, Wizard & Glass, "Low Men in Yellow Coats", and Insomnia, just so this one would make sense. I prefer The Talisman as a stand alone work.”

Predicted sentiment from content alone: neutral

Actual sentiment and predicted using both content and the title: negative