Predicting Helpfulness In Amazon Fine Food Reviews

Victor Zhou, Elizabeth Tian, Devina Singh

ABSTRACT

Amazon is increasingly establishing itself as the largest and most extensive e-commerce company in the world. Its seemingly unlimited product offerings and millions of users result in hundreds or even thousands of reviews on popular products. The helpfulness rating of a review can add more dimensionality to the review and help Amazon better identify which sort of reviews that customers are more interested in reading. However, because this rating is completely manual, there is a significant lack of helpfulness data tackled on all of the reviews. Thus, being able to identify which reviews might be helpful to consumers is an problem that could be through semantic analysis, natural language processing, or other machine learning methods, with applications to providing better and more concise information for customers looking to purchase an item.

BACKGROUND AND APPROACH

Data

We used 500,000 Amazon reviews of fine foods published over 19 years [1]. Each review sample consists of a user ID, product ID, product score, time stamp, review summary, and text review. In order to determine the review labels, we used the "Helpfulness Denominator" (HD), the number of users who indicated whether they found the review helpful or not, and the "Helpfulness Numerator" (HN), the number of users who found the product helpful. We used the equation below to determine a helpfulness ratin

$$R = \frac{HN}{HD}$$

Related Work

There exists a large body of previous work pertaining to the helpfulness of amazon reviews - Kim, Pantel et al. applied an SVM regression on MP3 player and digital camera reviews to predict helpfulness [2]. They found that the most useful predictive features included length of the review, product review and review unigrams, further suggesting that comparative words such as "more" or "better than" played a large role in these predictions. Furthermore, Korfiatis et al. [3] relied on using a Random Forests model along with the readability features of a review i.e. reviews without grammatical or spelling errors. They found a high correlation between readability and review helpfulness.

Methods

Classification Models

1. Recurrent Neural Networks (RNN):
   (a) Long Short-Term Memory (LSTM)
   (b) Bi-directional Long Short-Term Memory (BiLSTM)
2. Random Forests (RF)
3. Support Vector Machine with RBF Kernel (SVM)

Features

In addition, we decided to run these models on different feature sets provided and selected features below:
- Full text review
- Review summary
- Text length

Results

Random Forests

	Accuracy	Precision			Recall			F1		
Class Labels		**Not Helpful**	**Possibly Helpful**	**Helpful**	**Not Helpful**	**Possibly Helpful**	**Helpful**	**Not Helpful**	**Possible Helpful**	**Helpful**
Full Text	0.83	0.67	0.71	0.77	0.28	0.27	0.97	0.39	0.39	0.86
Summary	0.76	0.95	0.97	0.81	0.42	0.44	0.99	0.59	0.58	0.89
Text and Summary	0.83	0.96	0.97	0.81	0.42	0.41	0.99	0.58	0.57	0.89

Table 1: Results from the random forests classifier on text review data with different feature sets (summary, text, summary and text)

Class	Most Important Review Words
Overall	great, best, product, excellent, delicious, horrible
Unhelpful	yuck, awful, worst, nasty, terrible, disgusting
Neutral	horrible, weak, hated, bland, okay, disappointment
Helpful	great, best, product, excellent, delicious, good

Tabel 2: Most important words from the RF model on different classes and overall using the summary feature set

Class	Most Important Review Words
Overall	great, best, br, love, like, good
Unhelpful	worst, awful, return, threw, disgusting, msg, charged
Neutral	stream, horrid, swears
Helpful	great, best, br, love, like, good

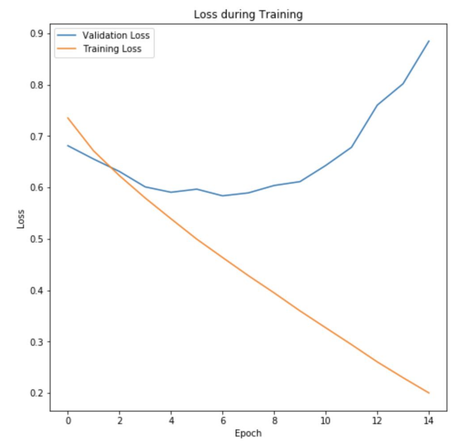
Tabel 3: Most important words from the RF model on different classes and overall using the text feature set

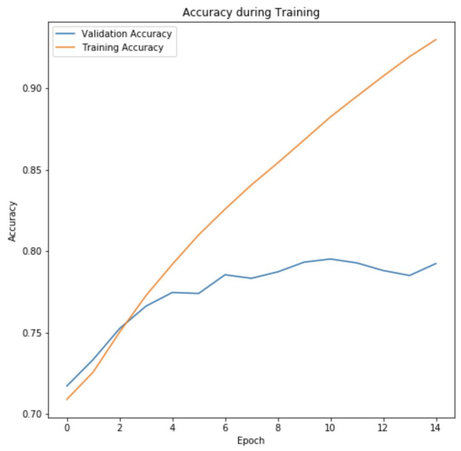
SVM

	Accuracy	Precision			Recall			F1		
Class Labels		Not Helpful	Possibly Helpful	Helpful	Not Helpful	Possibly Helpful	Helpful	Not Helpful	Possible Helpful	Helpful
	0.71	1	0.89	0.71	0.01	0.02	0.99	0.03	0.04	0.8

Table 4: Results from the SVM classifier on text review data

LSTM





Discussion

Our more complex LSTM-based model actually did not manage to achieve the same accuracy in predicting unhelpfulness that our Random Forests models did, despite the fact that LSTM had more parameters, and took much longer to train. This is not altogether surprising - it's possible that review helpfulness is much more correlated to the actual specific words used in the reviews and not the *sentiments* behind those words. That's the key difference between an LSTM-based model and a BOW-based model (which we used for the Random Forests) - the former represents words via GloVe vectors that encapsulates their meanings, while the latter represents words for what they themselves are.

Future Work

- Building models with hand-picked features. The Random Forests models helped us determine many of the most useful Bag of Words features, and we could use a select group to build a model that could potentially improve the final accuracy.
- Building more advanced recurrent models. The initial LSTM model we used only consistend of a single LSTM layer with one fully connected layer afterwards. We could try increasing the number of stacked LSTM layers to 2 or 3, or we could try using different recurrent layers like GRUs.
- Performing unsupervised analysis.

REFERENCES

1] "Amazon fine foods dataset," Stanford Network Analysis Project (Published on Kaggle) Jan 2016.
2) Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006, July). Automatically assessing review helpfulness. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 423-430). Association for Computational Linguistics.
3) Korfiatis, N., García-Bariocanal, E., & SánchezAlonso, 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.