Identifying the Relative Importance of Customer Issues on Product Ratings through Machine Learning

Himanshu Tiwari
HP R&D Center
Bangalore
himanshu.tiwari@hp.com

Shameed Sait
HP R&D Center
Bangalore
shameed.sait@hp.com

Md Imbesat Hassan Rizvi
HP R&D Center
Bangalore
md.imbesat.hassan.rizvi@hp.com

Niranjan Damera-Venkata
HP Labs
niranjan.damera.venkata@hp.com

ABSTRACT

Millions of customer reviews for products are available online across hundreds of different websites. These reviews have a tremendous influence on the purchase decision of new customers and in creating a positive brand image. Understanding which of the product issues are critical in determining the product ratings is crucial for marketing teams. We have developed a solution which can derive deep insights from customer reviews which goes significantly beyond keyword based analysis. Our solution can identify key customer issues voiced in the reviews and the impact of each of these on the final rating that a customer gives the product. This insight is very actionable as it helps identify which customer concerns are responsible for bad ratings of products.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**;
• Computing methodologies → Neural networks;

KEYWORDS

Customer Review Analysis, Attention Models

1 INTRODUCTION

Figure 1 shows a typical review of a product. A review often discusses both the pros and cons of the product and associated customer support/service experience. The final rating in a product review is often based on only a few critical sentences that explain the rating.

Understanding the key attributes and how much they impact customer perception of the product is very important for R&D and marketing teams to maintain brand image and reduce support costs.

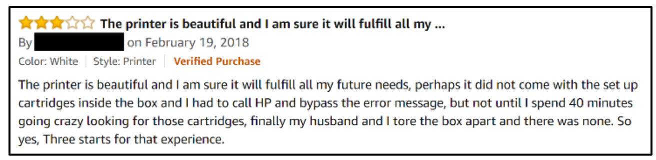


Figure 1: Sample review of a product

Employing people to read reviews is not a scalable solution although it is still used to gain anecdotal insight into a product. Basic text analytics can be employed to extract significant keywords (ex: tag clouds) in good and bad reviews and visualize them using dashboards. However, such analysis struggles with the variability in how two different reviewers express the same intent.

Recently the field of NLP has been revolutionized by deep learning [1]. Neural networks are able to create excellent representations of words and sentences [2, 3]. Denil et. al [4] apply CNN based prediction of sentiment based on deep representation words, and sentences in a paragraph. They are able to infer the importance of a particular sentence to the final sentiment by finding the derivative of the output prediction with respect to an input sentence using backpropagation. Lin et. al. [5] introduces the idea of attention into deep sequence analysis using recurrent neural networks (RNNs). They are also able to discover which phrases of a paragraph are most important in predicting ratings. Barzilay et. al. [6] develop the concept of "hard" attention to identify sentences that best explain a particular rating.

The methods described above are focused on explaining individual ratings. What we are after in this paper is a higher-level understanding of what issues (which may be expressed in various ways in different reviews) are most important in determining overall ratings. We are also interested in quantifying their impact.

Our solution derives key factors and their importance from the reviews. The output of our solution is a visualization of the relative importance of different customer issues on product ratings. The insights produced by our solution will help to identify the key customer issues. Our solution also helps prioritize the issues that need to be remedied soon to boost product ratings.

DocEng '18, August 28–31, 2018, Halifax, NS, Canada

ACM ISBN 978-1-4503-5769-2/18/08…$15.00
https://doi.org/10.1145/3209280.3229113

2 DETAILS OF OUR SOLUTION

2.1 Attention Model to Identify Key Sentences

To identify the key sentences within a customer review, we leverage a deep learning based mechanism known as attention models. Attention models are based on the visual attention mechanism found in humans. Humans can understand an image by focusing on certain parts of the image with focus (attention) while paying less attention to the rest of the image. Attention mechanism have also been successful in improving machine translation tasks by allowing the decoder to decide on the parts of the source sentence to pay attention to [7]. In general, attention mechanism uses deep learning techniques to identify parts of the input which are critical in determining the output. In our case, it boils down to identifying key phrases within a review which had more influence in determining the product rating given by the customer. An example of such key phrase identification is presented in Figure 2. An overview of our solution pipeline is presented in Figure 3.

I am giving it 2 stars because it makes a noise like a jet engine when it starts to print. I had no idea it would be this noisy-- during initialization I actually thought something was wrong and shut it down to verify that i had removed all packaging. But no, this noise seems par for the course. Update 2 weeks later: I am downgrading to 1 star. Print quality is fine for logos, but printing text in color gives very blurry results. I make product labels and it was not acceptable; I changed the text to black so it's better but not great; I use a Mac, and the driver does not have a "fine quality" option; apparently only the Windows driver has it. It looks like you are printing text in draft or economy mode. Also, the connection drops unexpectedly. I have to fix by shutting down and restarting the printer. If you want to print professional-looking product labels this is not the printer to buy. Am looking at alternatives.

I am giving it 2 stars because it makes a noise like a jet engine when it starts to print. I had no idea it would be this noisy-- during initialization I actually thought something was wrong and shut it down to verify that i had removed all packaging. But no, this noise seems par for the course. Update 2 weeks later: I am downgrading to 1 star. Print quality is fine for logos, but printing text in color gives very blurry results. I make product labels and it was not acceptable; I changed the text to black so it's better but not great; I use a Mac, and the driver does not have a "fine quality" option; apparently only the Windows driver has it. It looks like you are printing text in draft or economy mode. Also, the connection drops unexpectedly. I have to fix by shutting down and restarting the printer. If you want to print professional-looking product labels this is not the printer to buy. Am looking at alternatives.

Figure 2: Visual representation of Attention Model output. Red text indicates the sentences which got attention

The attention model pipeline comprises of Sentence Embedding Model and Classification Model. The Sentence Embedding model utilizes bidirectional LSTM layers to obtain hidden state representations at each sentence tokens and then applies a self-attention mechanism in the form a set of summation weight vectors for these state representations. An embedding for the sentence is obtained by element-wise dot operating the summation weight vectors with the LSTM hidden states. This model is applied to classification model, combined with a fully connected layer and a SoftMax layer. The model as shown in Figure 4, takes the LSTM hidden states **H** as input, and outputs a matrix **A,** whose rows correspond to attention weight vectors:

$$A = softmax(W_{s2} \tanh(W_{s1} H^T)) \qquad (1)$$

where $\{W_{s2}, W_{s1}\}$ are the weight parameters. The embedding vector M is evaluated by multiplying the annotation matrix A and LSTM hidden states H, resulting in sentence embedding matrix:

$$M = AH \qquad (2)$$

After this computation, the values inside the matrix A will give the attention weights assigned to each individual word. For a sentence level attention value, we average the attention value weights of all the words in the sentence. After sorting these values, we determine which of the sentences within a review got the most attention.

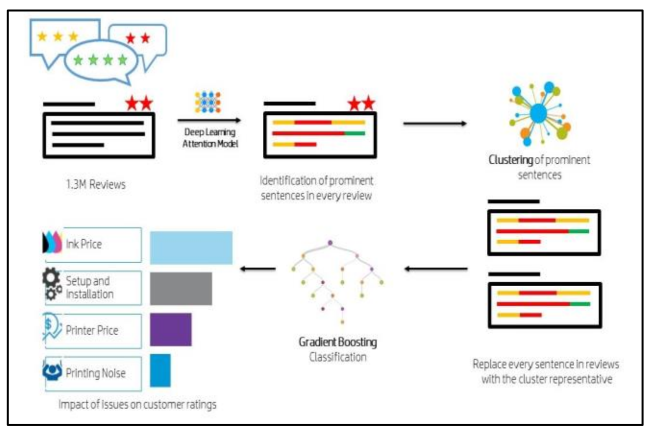


Figure 3: Our Solution Pipeline

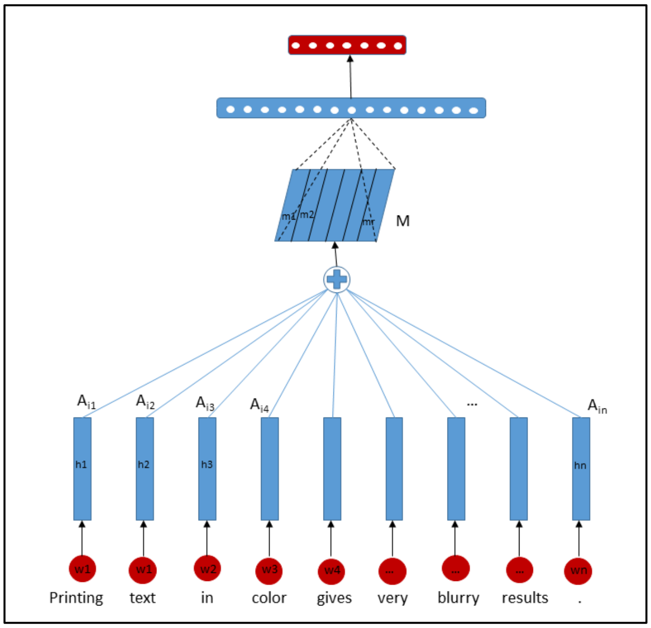


Figure 4: Multiple weighted sums of hidden states from a bidirectional LSTM as sentence embedding

2.2 Clustering of Sentences after Attention

After the attention model provides the important sentences in each review, we perform a clustering operation on all the important sentences. The aim of this module is to group together the sentences which might be talking about the same issue. For sentences like "wi-fi connection with the printer was lost frequently" and "wireless keeps disconnecting" will be clustered together by this approach. For clustering of sentences, we create a vector representation for each sentence by averaging the word embedding vectors of all the words in a sentence. We then use K-means clustering algorithm to cluster the similar sentences together. The optimal number of clusters (K) is found using an approach called the elbow method.

In the elbow method, the clustering is performed over a range of values of K (say, K from 2 to 10), and then for each value of K, compute the sum of squared errors (SSE) of the distances of the observations from the centroid of the cluster they are assigned to. When this graph is plotted it will look like the example shown in Figure 5.

As shown in Figure 5, the graph looks like an arm and then the elbow of the arm represents the optimal value of K. The intuition behind this approach is that the elbow denotes the value of K after which there is only a small reduction in SSE with an increase of K.

After we find the optimal value of K, we perform the clustering which assigns a cluster number to each sentence (values 1, 2,…, K).

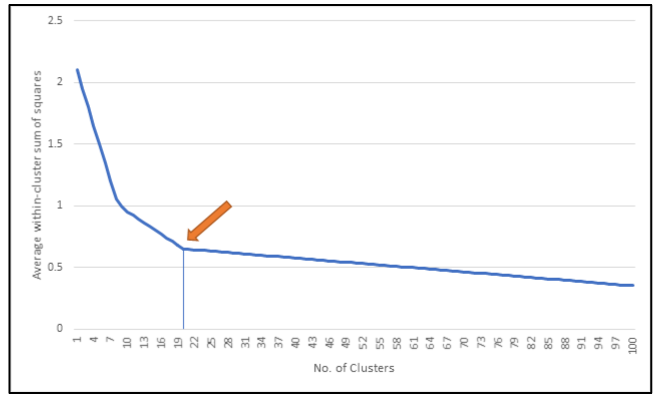
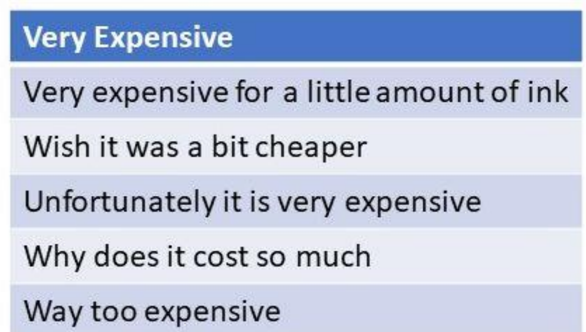


Figure 5: Sum of Squared Errors (SSE) vs the Number of Clusters (K) plot



Very Expensive
Very expensive for a little amount of ink
Wish it was a bit cheaper
Unfortunately it is very expensive
Why does it cost so much
Way too expensive

Figure 6: An example cluster and its member sentences

2.3 Computing Representative Sentence for each Cluster

For each of the cluster of sentences, we compute a representative sentence for the cluster by using a scoring method. For each word in the cluster, we attach a score based on totally how many times it appears in the cluster. Then for each sentence in the cluster, we compute a sentence score by adding up the word scores for all the words in the sentence and dividing it by the number of words in the sentence. After this, we select the sentence with the maximum score. This method has worked well for choosing a representative sentence for the cluster. The sentence thus chosen always reflected most of the sentences within a cluster. The representative sentence is required to finally visualize the impact of each cluster in the product ratings.

2.4 Identifying the relative importance of Customer Issues

After the clustering sentences which got attention, each review is converted to a feature vector of K dimensions (equal to the number of clusters). For every sentence which got attention in the review, based on the cluster it belongs to, we assign the value 1 in the corresponding column of the vector. At the end of this operation, each review will be a K dimensional vector containing 0's and 1's. The modified reviews are passed to a gradient boosting classifier. For our classifier, the input is the feature vectors and the output label is the rating (the number of stars the customer gave).

Gradient Boosting [8] is a machine learning technique for classification which produces an ensemble of weak prediction models which maximizes the performance of the classifier. The gradient boosting classifier creates a model which can predict the rating of a new given review and also it gives the relative feature importance of the features it is trained on. In our case, the features are the issues (clusters) the customers write about in the reviews and the feature importance values identify the effect each issue has on the customer rating.

We use the feature importance aspect [8] of Boosting algorithm to compute a score which indicates how valuable each feature was in the construction of the model. This is followed by the creation of visualization for the feature importance. This output will demonstrate which customer issues are most responsible for the lower ratings of products.

3 RESULTS AND DISCUSSION

3.1 Visualization of Feature Importance

We collected customer reviews for printer domain from more than 100 websites. The dataset consists of more than 1.3 million customer reviews. The main components of the dataset are the verbatim customer reviews and the ratings the customer gave for the product (which usually ranges from 1 star to 5 stars). As a first step, we perform rating normalization so that we have a uniform rating range across the data from the different websites. We also perform various text cleaning techniques like stop-word removal and stemming so that we can remove the noisy components in the text data. After the data pre-processing, we pass the data through our solution pipeline.

The gradient boosting classifier which we trained on our full dataset can predict the product ratings from the reviews. We trained our classifier with 80% of the data and 20% was kept aside for testing. Our classifier was able to achieve an average accuracy of 82%. As discussed in the earlier section, the output of our solution is the visualization of feature importance, produced by the gradient boosting classifier.

Figure 7: Output of our solution for one product

Sample output of our solution for one of the product models is shown in Figure 7. As shown in the figure, our output makes it very easy for other teams to understand We have generated similar outputs for more than 500 printer models. These results are shared in the format of a dashboard which can be used by different teams within the companies to identify the key issues with their products and also prioritize their tasks.

4 CONCLUSIONS

In summary, we have been able to identify the key customer issues from millions of reviews worldwide. Our solution also provides which of the customer issues are critical in determining the product rating given by the customer. Our solution can process millions of reviews and generate an output which the experts feel is very relevant. As part of the future work, we are looking at using the attention model itself to compute the feature importance.

REFERENCES

[1] Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa. 2011 . *Natural Language Processing (almost) from Scratch*
[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013 . *Distributed Representations of Words and Phrases and their Compositionality*
[3] Jeffrey Pennington, Richard Socher, Christopher D. Manning. 2014. *GloVe: Global Vectors for Word Representation*
[4] Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, Nando de Freitas . 2014 . *Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network*
[5] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu,Bing Xiang, Bowen Zhou& Yoshua Bengio . 2017. *A Structured Self-attentive Sentence Embedding*
[6] Tao Lei, Regina Barzilay and Tommi Jaakkola. 2016 . *Rationalizing Neural Predictions*
[7] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio . 2014 . *Neural Machine Translation by Jointly Learning to Align and Translate*
[8] Jerome H. Friedman. 1999. *Greedy Function Approximation: A Gradient Boosting Machine*