BERT Feature Based Model for Predicting the Helpfulness Scores of Online Customers Reviews

Shuzhe Xu, Salvador E. Barbosa, and Don Hong
Computational Science

ABSTRACT

Online product reviews help consumers make purchase decisions when shopping online. As such, many computational models have been constructed to automatically evaluate the helpfulness of customer product reviews. However, many existing models are based on simple explanatory variables, including those extracted from low quality reviews that can be misleading and lead to confusion. Quality feature selection is essential for predicting the helpfulness of online customer reviews. The Bidirectional Encoder Representations from Transformers (BERT) is a very recently developed language representation model which can attain state-of-the-art results on many natural language processing tasks. In this study, a predictive model for determining helpfulness scores of customer reviews based on incorporation of BERT features with deep learning techniques is proposed. The application analyzes the Amazon product reviews dataset, and uses a BERT features based algorithm expected to be useful in help consumers to make a better purchase decisions.

DATA

In this study, 2 datasets are generated for the experiment. The first dataset is for testing the model. It is collected by Amazon.com, which is downloadable at
https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt.

It contains 2 product categories (cameras and video games) with different product types (search goods or experience goods) that were used in prior research [1, 2, 3]. The original data has 15 columns, and this model focused on 4 factors, which are listed below:

Star rating: The 1 to 5 star rating of the review.

Helpful votes: Number of helpful votes the review received.

Total votes: Number of total votes the review received. This is equal to the number of helpful votes plus the number of not helpful votes.

Review body: The review text.

Fig .1. Helpful votes on Amazon.com

10000 reviews were selected randomly (from those where the number of total votes was greater than 30) from 2 different categories (Camera and Video Game).

After filtering the data, the extracted data is divided into 3 subsets: 80% of the data as the training set, 10% as the development set and another 10% as the testing set.
Therefore, for each category, there are 4000 reviews to train the regression model, 500 reviews to optimize the parameters of the model (development set), and 500 reviews for testing.

NEURAL NETWORK BASED MODEL

In this study, the proposed NN based model is formed by incorporating BERT pretrained model with one additional output layer to predict the helpfulness score. The regression model contains 3 main terms: BERT features, star rating, and product type. The BERT features are obtained from the pre-trained model as vectors [4]. Though the vectors are difficult to explain and understand directly, they are data driven outcomes and hence, the BERT features are reliable. The star rating can be extracted from the data directly and the product type is determined by the previous research [1, 2, 3]. The model is designed as:

$$s = f(w^T x + b)$$

Here s is the helpfulness score, f is the activation function ReLU, x is the vector of input and w is a vector as the weight of x, where the length of x and w is determined by the length of the input. b is bias and it is a single value since there is only one output which is the predicted helpfulness score. The equation above can be extended in detail:

$$s = f(w_1^T \, FeaturesBERT + w_2^T \, StarRating + w_3^T \, ProductType + b)$$

where w_1, w_2 and w_3 are vectors of weight corresponding to specific terms.

The loss function of this NN based model is measured by mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |S_i - \hat{s}_i|$$

where S_i represents the real helpfulness score from the original dataset, $\hat{s}_i$ is the corresponding predicted helpfulness score, and n is the number of reviews in dataset. The input of the NN are elements in x. For each element in x, it multiples the corresponding weight in the hidden layer. The bias b is added to this. The output of the NN is the summation of outputs that come from the activation function as showed below.

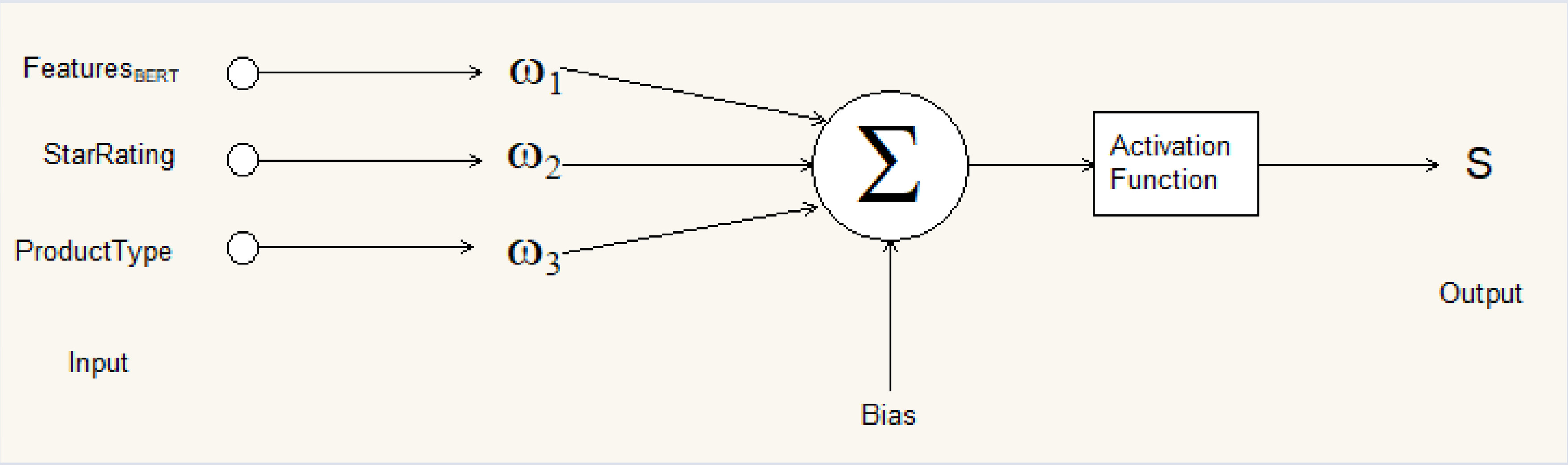

Fig. 2. Regression Process in Neural Network

ALGORITHM PROCEDURE

The program is based on the original BERT release. The source code is downloadable at
https://github.com/google-research/bert#fine-tuning-with-bert

Data pre-processing The files are generated for testing and the comparison. The process is listed below:

Step 1: Select data to match the condition, for example, the number of total votes is greater than 30.

Step 2: Convert the data file to .tsv file.

Step 3: Randomly divide the data into training, development and test sets.

Step 4: Repeat step 3 until enough data is collected.

Neural Network incorporating BERT The general steps to run the code on Tensor-Flow are:

Step 1: Extract the BERT features from each dataset.

Step 2: Combine the BERT features with star rating and product type as the input to the NN.

Step 3: Train the model in NN using training data and optimize the parameters with the development set.

Step 4: Test the model and calculate the error.

Step 5: Repeat Steps 1 to 4 on 10 random data subsets, and calculate the average error and standard deviation.

RESULTS

In this section, the exact same data in [5] is used for comparison, and the best results (best average MAE and smallest stand deviation) in that paper are compared. The data is collected from
http://jmcauley.ucsd.edu/data/amazon/
and reviews with more than 10 total votes are selected, to match what that paper did (this data is different from the testing data used in section 3). Ten different data sets are generated randomly by using these selected data. For each data set, 80% is chosen as the training set, 10% as the development set, and another 10% as the test set. There are 520 reviews for each test of Cellphone products and 836 reviews for each test of Beauty products, as was done in [5], and the proposed model does not contain the product type (experience or search goods) for this comparison.

Data set	MAE(NN)
1	10.7091
2	11.3106
3	11.0028
4	9.9965
5	10.0641
6	10.6486
7	10.4135
8	10.6583
9	11.3135
10	10.7398
Average	10.6857
SD	0.4505

Table: MAE of different testing data set(Cellphone)

Data set	MAE(SVR)	MAE(M5P)
1	11.1354	12.1070
2	11.5486	12.5098
3	10.4736	11.7598
4	10.3240	10.6372
5	11.8700	12.2581
6	11.5203	11.8930
7	11.9216	12.4486
8	12.1916	11.8303
9	13.3505	12.2984
10	12.9721	12.6732
Average	11.7308	12.0415
SD	0.9178	0.5495

Table: Test results in Park's paper(Cellphone)

Fig. 3. Comparison with explanatory variables based regression model

In this study, a BERT features approach was investigated for linguistic and psychological information to incorporate into a NN model for predicting scores for helpfulness of online reviews. The comparison shows that the predictive results of the proposed model are not only better in terms of MAE, but also yield a smaller standard deviation. This means that the BERT based NN model is more stable than the models with limited explanatory variables across different product categories. These results show the model is reliable in predicting the helpfulness scores of customer reviews, while avoiding limitations such as the time consuming process associated with selecting and extracting explanatory variables.

REFERENCES

[1] Nelson P (1970) Information and consumer behavior. Journal of political economy 78(2):311–329

[2] Bragge J, Storgards J (2007) Utilizing text-mining tools to enrich traditional literature reviews. case: digital games. In: Proceedings of the 30th Information Systems Research Seminar in Scandinavia IRIS, pp 1–24

[3] Weathers D, Sharma S, Wood SL (2007) Effects of online communication practices on consumer perceptions of performance uncertainty for search and experience goods. Journal of retailing 83(4):393–401

[4] Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

[5] Park YJ (2018) Predicting the helpfulness of online customer reviews across different product types. Sustainability 10(6):1735

CONTACT

Shuzhe Xu
Email: sx2b@mtmail.mtsu.edu