NLP Models Classifying Helpful Ratings in OpenTable Dataset

Sunnyoung Lee, *Hsiu-Ping Lin, Jihyun Park, Eugene Lim, *Jongwook Woo
Dongguk Business School, Dongguk University
Seoul, South Korea
*Department of Information Systems, California State University, Los Angeles
[e-mail: enhasu19@dongguk.edu, {zzhxxxn, euginoo}@gmail.com, {hlin54, jwoo5}@calstatela.edu]
*Corresponding author: Jongwook Woo

Abstract

This paper aims to build models to predict the helpfulness of reviews by analyzing review data from OpenTable, a restaurant reservation application spanning 14 years (2009 to 2023). Our goal is to accurately classify the helpfulness of the review column of the dataset. We adopted a DistilBERT Huggingface model, an example of LLM (Large Language Model), as a distilled version of the BERT base model for the Natural Language Process (NLP) to classify sentences. We built an NLP classification model by fine-tuning the pre-trained DistilBERT model of Huggingface with the OpenTable data. We also made traditional machine learning models using Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boost Tree (GBT) algorithms to compare the performance of the models. We observed that the fine-tuned DistilBERT model has 4 - 14 % better accurate Precision and AUC than the traditional models. The traditional models' computing time is 90 – 1,600 seconds, while the fine-tuned DistilBERT model's is 490 seconds. In summary, the fine-tuned DistilBERT model has the best performance among the models. This analysis allows us to identify consumers' (satisfied or dissatisfied) attitudes toward OpenTable's app service and pinpoint areas for improvement. Moreover, from a marketing perspective, featuring the predicted helpfulness review data at the top of the application can enhance the experience for existing customers and provide valuable insights to potential customers, which will gain more potential customers.

Keywords: LLP, Huggingface, Scalable Computing, OpenTable, BERT, NLP, Classification

1. Introduction

Online review systems have influenced the decision-making of potential customers [1]. In particular, classifying and predicting helpful reviews among many online reviews significantly impact customer acquisition [2]. This study aims to build models to predict reviews' helpfulness, comparing the performance of the traditional Machine Learning (ML) and Deep Learning (DL) Large Language (LLM) Models.

2. Related Work

Yu et al. built category classification models with The THUCNews dataset, combining neural network models such as word2vec, BiGRU, ELMo, BERT, CNN, and RNN. It shows that BERT-BiGRU model has the highest accuracy [4].

Garrido-Merchan et al. built sentiment prediction models with the traditional ML models and BERT with the datasets: IMDB, Tweets, News, Hotel Reviews. It concludes that BERT has the highest accuracy [5].

Our paper re-trains and builds the **DistilBERT** model with OpenTable data. The paper classifies helpfulness ratings in accuracy and computing time, comparing the traditional ML models and DistilBERT.

3. Background

3.1 Review of OpenTable Dataset

This paper collected online review data from **OpenTable**, a global restaurant reservation application, to build a model to predict the helpfulness of reviews. The 46,392 sample data was collected from the OpenTable's app from September 2009 to February 2023, about 14 years. The data includes the user's nickname, review content, star rating, helpfulness, and date and time.

3.2 Data Clustering Analysis

Topic modeling was performed to determine the frequency of co-occurrence of words. The consistency score (0.3952) and the inflection point were selected for 9 topics [3].

Table 1 presents how consumers rated the app with the 9 topics. Topics 1 and 6 provide an objective assessment of the quality of the app. It lists features that are important to them: [“booking”, “new places”, “recommendations”]. Topics 2 and 4 evaluate the UI and UX aspects of the app, particularly mentioning: [“adding application”, “convenience”], indicating that consumers value functional convenience in using apps. Topic 3 evaluates the restaurant experience, focusing on the experience of a good restaurant, which is important and often leads to evaluating the app based on the reviewers' detailed restaurant experience. Topics 5 and 8 evaluate the app experience in terms of hedonic aspects. Emotional words such as "love" and "wonderful" are predominant, suggesting that users are highly satisfied when using the app. Topics 7 and 9 evaluate the accuracy and reliability of the app's content and system.

Table 1. Top 10 terms of co-occurrence of words.

Topic	Words	%
1	table, reserve, place, new, friendly, travel, easier, available, recommend, want	13.6
2	make, make reservation, simple, easy make, need, options, cancel, problem, handy, make easy	12.8
3	time, way, experience, reliable, best, wish, dinner, reward, restaurant use, great experience	11.6
4	quick, use app, helpful, quick easy, service, try, app easy, nice, navigate, years	11.4
5	love, love app, book, super, issue, super easy, help, easy book, know, wonderful	11.3
6	great app, excellent, point, opentable, use opentable, use opentable, accurate, food, excellent app, website, add	10.7
7	useful, din, efficient, date, review, user, amaze, fast, look, time	10.3
8	good, awesome, opentable, thank, good app, email, let, love opentable, book, love	9.2
9	convenient, work, phone, easy convenient, work great, great restaurant, people, app work, convenient easy, pasword	9.2

4. NLP Rating Classifying Models

Spark is a distributed parallel computing platform that supports machine learning APIs. It also provides NLP APIs such as Tokenizer, StopWordsRemover, and HashTF.

LLM in AI DL have received highlights lately, which can process, understand, and generate natural language. ChatGPT and BERT are the most popular LLM models in NLP. ChatGPT is built by Open AI that can generate texts. Google introduced BERT in 2018 that is suitable for text classification and Q&A.

4.1 ML and LLP Models

We built classification models to predict Helpful labels in the Spark platform - **LR**, **DT**, **RF**, **GBT** - with the users’ review text from OpenTable data. We adopt Hashing Term Frequency for the models and set the parameters as {maxDepth:5, numTrees:3, maxIter: 20, regParam: 0.01}.

In addition, **BERT** can be used in text classification as well. It is a bidirectional transformer pre-trained model with raw texts without human labeling on the *BookCorpus* and *English Wikipedia* datasets, combining next-sentence prediction and mask language modeling objectives. **DistilBERT** model is the small and fast version of BERT based on the BERT model. It is pre-trained on the same datasets as BERT. The model can determine a mask in a sentence to fill in. We train and build a text classification model on **DistilBERT** with Auto Tokenizer to predict Helpful labels.

4.2 Experimental Result

We built and executed the traditional machine learning models in single node **PySpark** with Spark 3.5.0, and **DistilBert** model in **Google Colab** as *e2-standard-4* with CPU speed 2.25 GHz, 4 vCPU, and 16 GB RAM.

Table 2. Performance Comparison of the Models

	Precision	AUC	Computing Time: *log(sec)*
LR	0.682	0.644	201
DistilBERT	0.712	0.705	493
RF	0.667	0.524	634
DT	0.622	0.593	709
GBT	0.637	0.616	3,141

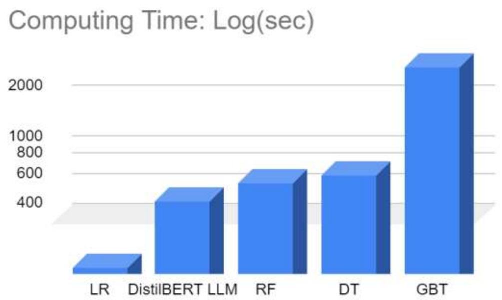

Fig. 1. Computing Time of the Models

We observed performance results of the models that predict the Helpful label ratings in **Table 2**. The precision and AUC of the DistlBERT LLM model are 71.2 % and 70.5 %, respectively. So, it has 4 - 14 % better Precision and AUC than the traditional models. LR has the shortest computation time to build the model. And, the next shortest time is the DistlBERT model, as shown in **Table 2** and **Fig. 1**.

5. Conclusions

This paper compares the performance of the traditional classification models and the DistilBERT LLM model to classify the review text of the OpenTable dataset. We observe that the DistilBERT LLM model shows the highest accuracy and the second fastest computing time compared to the traditional ML models to predict the *Helpfulness* label with the feature of the users' comments in the OpenTable data set as an NLP.

From a marketing perspective, helpful reviews significantly influence the decisions of potential customers. Therefore, developing a strategy to place helpful reviews at the top of your app using accurate and fast predictive models is critical.

References

[1] Guo, Y., Barnes, S. J., & Jia, Q. ,"Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation", Tourism Management, 59, 467–483, 2017.

[2] Zhan, Mengmeng, et al. "The role of explained actions and reactions in the helpfulness of online reviews." Electronic Commerce Research, 1-30, 2023.

[3] O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P., "An analysis of the coherence of descriptors in topic modeling", Expert Systems with Applications, 42(13), 5645–5657, 2015.

[4] Q. Yu, Z. Wang, K. Jiang, "Research on Text Classification Based on BERTBiGRU Model ," Journal of Physics: Conference Series, doi:10.1088/1742-6596/1746/1/012019, 2019.

[5] E. C. Garrido-Merchán, R. Gozalo-Brizuela, S. González-Carvajal, "Comparing BERT against traditional machine learning text classification ," Journal of Computational and Cognitive Engineering, https://doi.org/10.47852/bonviewJCCE3202838, 2023.

[6] DistilBERT base model (Oct 2023), Retrieved from https://huggingface.co/distilbert-base-uncased