Binning Reviews - Sentiment Analysis & TripAdvisor Reviews

Philippe Gagnon, Christopher Ratsimbazafy-Da Silva
August 2023

Abstract

Achieving high levels of customer satisfaction and expanding their clientele base are two of the foremost goals for participants (i.e. hotels, restaurants, etc.) within the global tourism industry. However, analyzing customer reviews to predict accurate sentiments has proven difficult to both participants and researchers alike due to the inherent skewed nature of user review data as well as the lack of the contextualization and aspect-based sentiment analysis of these reviews[1]. Our research applied a variety of modern techniques (CNNs, RNNs, BERT models) to both predict a user's eventual rating and subsequently classify a good hotel from a bad one. Our results suggest that a BERT classifier model can provide a marginal improvement in accuracy over the Deep Averaging Network (DAN) baseline and can address some of the issues related to contextualization faced by other researchers. Our findings have implications for service providers within the global tourism industry in their ability to (1) improve customer service, (2) target marketing campaigns more effectively, and (3) identify emerging trends.

Keywords: tourism industry; user reviews; aspect-based sentiment analysis

Introduction

Our goal for this project was to build a classifier model that could better leverage – in comparison to traditional sentiment analysis approaches – text elements from user reviews to predict a user's eventual rating and ultimately distinguish a good hotel from a bad one. TripAdvisor is a prominent platform for travel reviews which has achieved success namely through its effective use of user-generated content to connect customers, advertisers, hotels, restaurants and other participants in the global travel industry. Tourist reviews are a valuable source of information for potential travelers. For example, they can help travelers make informed decisions about where to stay, eat and visit in a given location. Furthermore, reviews provide necessary and valuable feedback to businesses who can thereby improve both their products and services[2].

Nevertheless, there are many challenges associated with leveraging tourist reviews as a proxy for widespread consumer sentiment towards an establishment. For example, reviews tend to be biased - i.e. individuals with a negative experience with a business are more likely to leave a review than an individual who had a positive experience, which leads to both a skewed perception of the business and class imbalance issue which the researchers address in this study. Secondly, online reviews can be difficult to evaluate both by human reviewers and machine learning methods. For example, if a customer were to leave a TripAdvisor review such

[1] Sun, Fengdong & Chu, Na & Du, Xu. (2020). Sentiment Analysis of Hotel Reviews Based on Deep Learning. 627-630. 10.1109/ICRIS52159.2020.00158.

[2] Cornell Hospitality Report • May 2016 • www.chr.cornell.edu • Vol. 16, No. 10

as “our stay at the hotel was *good*,” some potential travelers could interpret this as a positive review, while others as a neutral one.

Sentiment analysis – or the process of extracting feeling or opinion from text – is a technique that can be used to address the aforementioned challenges associated with tourist views and can be conducted through a combination of lexicon-based, machine learning or deep learning methods. For example, lexicon-based methods can be used to identify biased reviews, weigh these reviews accordingly, and ultimately improve the accuracy of said reviews. Similarly, aspect-based sentiment analysis can be leveraged to identify the *aspects* (i.e. quality of service, accommodations, etc.) of a business that are mentioned in reviews, which can aid an establishment to better understand and improve their offerings[3].

Nevertheless, some NLP techniques can struggle to accurately assess sentiment by failing to take into account context or discern between objective and subjective texts as well as differentiate between negation and sarcasm. It is the aim of this study to improve on such methods.

Background

According to the 2023 UNWTO World Tourism Barometer, international tourism for 1Q2023 indicates that the industry is “well on its way to returning to pre-pandemic levels, with twice as many people traveling during the first quarter of 2023 than in the same period of 2022.” As such, hotels and other participants of the global tourism industry have increasingly devoted more resources to (re)capturing tourist inflows and expanding their clientele base back to pre-pandemic levels[4]. Specifically, the Center for Hospitality Research’s report “Hotel Performance Impact of Socially Engaging with Consumers” found that “hotels that encouraged…responded to [and incorporated] guest reviews saw increases in their ratings, as compared to their competitive set.” Moreover, “when [hotel] management responded to reviews, their sales and revenue improved.” Transactions data from Travelocity highlights that “if a hotel increases its review scores by 1 point on a 5-point scale (e.g., from 3.3 to 4.3), the hotel can increase its price by 11.2 percent and still maintain the same occupancy or market share.[5]”

Such analysis illustrates that hotels are incentivized to both evaluate and incorporate their customer reviews; however, there are many associated challenges in doing so. In their seminal 2002 paper, Pang et al. leveraged “the presence of positive and negative words…and the sentiment of the words surrounding a target word” to predict the sentiment of the overall movie review. The authors achieved an accuracy of 82.1% on a test set of movie reviews and found that their model was able to generalize well to new reviews that it had not encountered before;

[3] Kim, D., Shin, S., & Park, J. (2014). Sentiment analysis of hotel reviews using deep learning. In Proceedings of the 23rd ACM international conference on conference on information & knowledge management (pp. 2661-2664). ACM.

[4] World Tourism Organization. Tourism on Track for Full Recovery as New Data Shows Strong Start to 2023. (2023, January 17).

[5] Cornell Hospitality Report • May 2016 • www.chr.cornell.edu • Vol. 16, No. 10

however, model accuracy could have improved had it incorporated domain-specific lexicons leveraged by subject matter experts to readily identify positive, negative or neutral sentiments[6].

Liu et al. (2012) contribute to Pang et al.'s findings through a comprehensive taxonomy of opinion words and phrases. The authors employed the opinion mining approach or "the task of automatically extracting subjective information from text, such as opinions, evaluations, and emotions" to domains outside of customer feedback and product reviews, such as political campaign analysis. Nevertheless, the authors found that while lexicon-based methods can be simple to implement, deep learning methods prove to be the most accurate methods for sentiment analysis but require a large amount of training data and can be computationally expensive[7].

In their 2014 paper, Kim et al. attempt to "address the subjectivity of language" by employing a CNN to extract features from TripAdvisor reviews, training a SVM classifier to categorize positive, negative and neutral sentiments, and ultimately attaining 82.3% accuracy on a test set of reviews. Their results mark a significant improvement against their baseline model (bag-of-words approach) in predicting user review sentiment. Our study aims to expand on the work conducted by the aforementioned authors and, specifically, Kim et al. – using the same baseline, bag-of-words model – with two additional contributions: (1) by leveraging newer classifier models (e.g. BERT) and (2) by incorporating aspect-based analysis to help us better predict user review ratings and classify good hotels from bad ones[8].

Methodology

Exploratory Data Analysis

Our data was a sample of approximately 148,000 TripAdvisor reviews of establishments in New Delhi, India in 2023[9]. The data had 2 columns: *rating_review* which was an integer between 1-5 and *review_full* which was the user-generated text review. For example, the first user reviewer gave a rating of 5 and left a review of "Totally in love with the Auro of the place, really beautiful and quite fancy at the same time." A distribution analysis of the score ratings; however, reflects a stark class imbalance with 93.9% of ratings falling between 3 and 5. When computing the length of each review (*review_length*), we found the expected result of a right-skewed distribution with there being a handful of reviews with more than 2000 characters. Similarly, we found the correlation between review length and rating score to be negligible (-0.096). The most compelling insight from our exploratory analysis; however, was that the distribution of polarity, or sentiment scores ranging from -100 (very negative) to 100 (very positive), was centered around 30-40, which indicates that the median user review tended to be mildly positive.

[6] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL 2002 workshop on empirical methods in natural language processing (pp. 79-86). Association for Computational Linguistics.

[7] Liu, B., Hu, M., & Cheng, J. (2012). Opinion mining and sentiment analysis. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[8] Kim, D., Shin, S., & Park, J. (2014). Sentiment analysis of hotel reviews using deep learning. In Proceedings of the 23rd ACM international conference on conference on information & knowledge management (pp. 2661-2664). ACM.

[9] https://www.kaggle.com/datasets/arnabchaki/tripadvisor-reviews-2023

Data Cleansing

Following our exploration of the data, we conducted a number of data preprocessing for categorical and text columns. For the text reviews, in a similar approach to Kim et al. (2014), we removed stop words, incorporated a dictionary of positive and negative words, and computed the frequency of both classes of words across each rating. Following a review of the literature of Hwang et al. (2022), we employed two different methods of handling the class imbalance within the data. The first approach was a random selection of a number of reviews for each score based on the category with the fewest number of reviews (i.e. 1)[10]. The second was to create a binary score where if the rating was 3 or above, it was classified as positive and negative if below. From there, our data was split into 80/20 train-test splits.

Code: EDA & Data Cleansing

Baseline

For our baseline, we employed a similar bag-of-words approach to Kim et al. (2014) with an Averaging Network (DAN) model on our selection of randomized reviews. Following 20 epochs, the DAN model achieved modest accuracy results between 49.2% and 56.1%[11].

Code: Baseline Model

Modeling

Outside of our AN/DAN baselines, we built models using CNN and RNNs. In keeping with the Kim et al. (2014) study, we leveraged CNNs and experimented with various hyperparameters (e.g., dropout rates, dense layer dimensions) to attempt to match their results of 82.3% accuracy on a similar dataset of hotel reviews. Despite the lack of sequential nature within our data, we experimented with RNNs to corroborate the findings in the Wei et al. (2019) paper that RNNs perform well on task classification tasks within smaller datasets[12]. As per our contributions to the existing literature on sentiment analysis within hotel reviews, we built two different BERT classifier models: the first was a binary classifier that was trained and tested on our negative and positive review classifications (0 and 1). The second was a multiclass classifier that made predictions on the same dataset with an equally-distributed, random sampling of all scores (1-5).

[10] Kim M, Hwang KB. An empirical evaluation of sampling methods for the classification of imbalanced data. PLoS One. 2022 Jul 28;17(7):e0271260. doi: 10.1371/journal.pone.0271260. PMID: 35901023; PMCID: PMC9333262.

[11] Kim, D., Shin, S., & Park, J. (2014). Sentiment analysis of hotel reviews using deep learning. In Proceedings of the 23rd ACM international conference on conference on information & knowledge management (pp. 2661-2664). ACM.

[12] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Code: CNNs & RNNs, BERT Model (Binary), BERT Model (Multi-Class)

Results

The chart below shows the accuracy of each model (in descending order).

Model	Test Accuracy	Test Loss
BERT (Binary Classifier)	84.09%	0.3979
Convolutional Neural Network	63.17%	0.2882
Recurrent Neural Network	62.16%	0.2931
Deep Averaging Network	60.19%	0.3080
BERT (Multiclass Classifier)	59.64%	0.9245
Averaging Network (Baseline)	55.58%	0.3401

We achieved our highest validation accuracy (84.09%) using the BERT Binary Classifier model and the lowest using our bag-of-words baseline model. Given that AN/DANs are "syntactically ignorant" models, it is therefore unsurprising that they performed worse on the text classification tasks required for sentiment analysis[13]. While the NNs underperformed the BERT Binary Classifier by a margin of ~21-22%, more hyperparameter tuning could have been conducted through adjustments to the filter sizes and the number of filters to potentially match the results found in Kim et al. (2014). Other modifications such as L2 Regularization did not seem necessary given that the NNs showed no indication of overfitting the training data.

BERT (Multiclass Classifier) - Classification Report

Ratings	Precision	Recall	F1-Score
1	0.71	0.70	0.71
2	0.50	0.48	0.49
3	0.54	0.60	0.57
4	0.59	0.37	0.46
5	0.62	0.83	0.71

[13] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Accuracy			0.60
Macro Avg.	0.59	0.60	0.59
Weighted Avg.	0.59	0.60	0.59

While the BERT Multiclass Classifier underperformed (59.64%) many of the other models, it is interesting to see that it maintained the highest precision and recall levels for ratings 1 and 5 which had the fewest and highest number of reviews respectively. More broadly, the poorer performance of the Multiclass Classifier seems to corroborate the findings of Sahare et al. (2012) that the class imbalance within the dataset can have an outsized influence on model accuracy of multiclass models[14]. Leveraging BERT for multiclass classification was expensive both in terms of memory and computation; however, future iterations of this study will experiment further with hyperparameter tuning (i.e. batch size, maximum length of reviews, number of hidden layers, learning rate, etc.) to optimize computational performance.

Conclusions

The aim of our study was to leverage newer sentiment analysis tools to predict user review ratings and to apply context-based methods for text classification to distinguish good hotels from bad ones. While our BERT binary classifier significantly outperformed our baseline bag-of-words method as expected and was able to distinguish both good reviews from bad and subsequently good hotels from bad, the generalizability of our results is limited by certain caveats and assumptions underlying our study. For example, our text data did not include any non-English reviews and our exclusion of stop words within the data pre-processing filtered out non-standard writing (i.e. slang, typos, emojis, etc.), elements which Liu et al. (2012) assert to be crucial for the improvement of NLP techniques[15]. As a result, we should expect our BERT classifier model to perform poorly against non-English and non-standard texts.

Separately, we made an assumption at the onset of the study that the range of human sentiment can be mapped onto a spectrum that falls neatly into a universal set of basic emotions that are expressed with similar words. However, further research is required as to how we can improve text classification for instances of negation and sarcasm as exemplified by a number of text reviews within our sample. Within the realm of sentiment analysis for tourist reviews, the researchers would like to employ more domain-specific lexicon-based methods that would enhance the positive and negative word buckets used to categorize the review texts. Lastly, the limited scope of the underlying dataset (in terms of both time and location) suggests that our analysis may not be externally valid in other contexts. As stated by Pang et al. (2002), “if the training dataset is not representative of the target domain, the NLP technique may not be able

[14] Sahare, Mahendra & Gupta, Hitesh. (2012). A Review of Multi-Class Classification for Imbalanced Data. International Journal of Advanced Computer Research. 2. 160-164.

[15] Liu, B., Hu, M., & Cheng, J. (2012). Opinion mining and sentiment analysis. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

to generalize well to new data.[16]" Nevertheless, our findings have implications for service providers within the global tourism industry in their ability to (1) improve customer service, (2) target marketing campaigns more effectively, and (3) identify emerging trends.

[16] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL 2002 workshop on empirical methods in natural language processing (pp. 79-86). Association for Computational Linguistics.

References

- Kim, D., Shin, S., & Park, J. (2014). Sentiment analysis of hotel reviews using deep learning. In Proceedings of the 23rd ACM international conference on conference on information & knowledge management (pp. 2661-2664). ACM.
- Sun, Fengdong & Chu, Na & Du, Xu. (2020). Sentiment Analysis of Hotel Reviews Based on Deep Learning. 627-630. 10.1109/ICRIS52159.2020.00158.
- Liu, B., Hu, M., & Cheng, J. (2012). Opinion mining and sentiment analysis. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL 2002 workshop on empirical methods in natural language processing (pp. 79-86). Association for Computational Linguistics.
- Zhu, S., Chen, L., & Wang, H. (2015). Hotel review sentiment analysis based on ensemble learning. Expert Systems with Applications, 42(22), 9042-9051.
- Chinnalagu A, Durairaj AK. Context-based sentiment analysis on customer reviews using machine learning linear models. PeerJ Comput Sci. 2021 Dec 17;7:e813. doi: 10.7717/peerj-cs.813. PMID: 35036535; PMCID: PMC8725657.
- Cornell Hospitality Report • May 2016 • www.chr.cornell.edu • Vol. 16, No. 10
- Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Lorena, A.C., de Carvalho, A.C.P.L.F. & Gama, J.M.P. A review on the combination of binary classifiers in multiclass problems. Artif Intell Rev 30, 19 (2008). https://doi.org/10.1007/s10462-009-9114-9
- Sahare, Mahendra & Gupta, Hitesh. (2012). A Review of Multi-Class Classification for Imbalanced Data. International Journal of Advanced Computer Research. 2. 160-164.
- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1681–1691, Beijing, China. Association for Computational Linguistics.
- World Tourism Organization. Tourism on Track for Full Recovery as New Data Shows Strong Start to 2023. (2023, January 17). https://www.unwto.org/news/tourism-on-track-for-full-recovery-as-new-data-shows-strong-start-to-2023#:~:text=New%20Data%20from%20UNWTO%3A%20What%20We've%20Learned&text=It%20shows%20that%3A,the%20same%20period%20of%202022.
- Kim M, Hwang KB. An empirical evaluation of sampling methods for the classification of imbalanced data. PLoS One. 2022 Jul 28;17(7):e0271260. doi: 10.1371/journal.pone.0271260. PMID: 35901023; PMCID: PMC9333262.