Sentiment analysis and truthfulness assesment of hotel reviews

Group 74: Jascha Jestel, Piotr Migdałek

October 26, 2023

Abstract

In this project, we developed a sentiment analysis system for hotel review classification, a crucial aspect of understanding customer satisfaction in the hospitality industry. We applied various preprocessing and text embedding techniques, alongside fine-tuning machine learning and deep learning algorithms, to categorize reviews as positive or negative. Additionally, our system distinguishes whether reviews are truthful or deceptive. We assessed the models' generalization through cross-validation and identified the optimal epochs for pre-trained solutions. The outcomes of this project hold wide-ranging applications, enabling real-time customer satisfaction monitoring for hotel management and providing travelers with decision-making insights. Our model represents a valuable tool for the hospitality sector, boasting high accuracy, especially in distinguishing positive from negative sentiment.

1 Data

The provided corpus consists of 1399 hotel reviews with human-assigned labels for sentiment and truthfulness (TRUTHFULPOSITIVE, TRUTHFULNEGATIVE, DECEPTIVEPOSITIVE, and DECEPTIVENEGATIVE). Each class represents approximately a quarter of the data, making the dataset balanced. In the course of our examination of this linguistic corpus, we unearthed intriguing characteristics of the natural language conveyed in this written form.

An illustrative discovery from Figure 1 portrays a noteworthy trend: as the prevalence of words in all capital letters in a review escalates, so does the proportion of TRUTHFULNEGATIVE class. To capitalize on this insight, we thoughtfully incorporated it into our preprocessing framework, preserving the original case of words typed in all caps, thereby leveraging their potential significance.

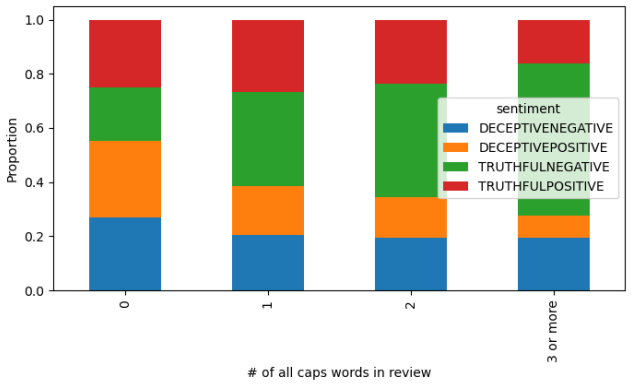

Figure 1: Proportion of all caps words per class in all reviews.

From Figure 2, we can infer that positive sentences tend to be shorter. However, the distribution shapes related to truthfulness heavily overlap, which may highlight a potential issue with masking classes, a matter we delve deeper into in the subsequent discussion.

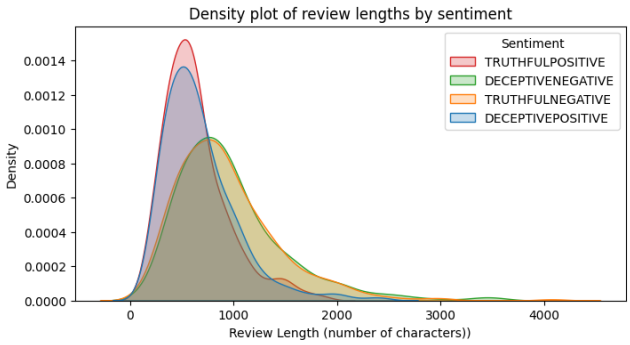

Figure 2: Kernel density estimators of the review length distribution per sentiment.

2 Models

We initially used a simple LSTM model with tokenized sequences, padding, and training the embedding layer as the first step of the network. We also experimented with traditional machine learning models, as outlined in Table 1, and a stacking classifier that combined outputed probabilities of LR, SVM, and Naive Bayes classifiers. We enhanced preprocessing with part-of-speech tagging, lemmatization, lowercasing and stop word plus punctuation removal, and transformed reviews into a 7288-dimensional TF-IDF space.

However, we later improved our approach by adopting a fine-tuned DistilBERT [1] model with lowercasing, double spaces removal and tokenizing truncated reviews (it didn't affect model's accuracy) as preprocessing is concerned, which achieved the best accuracy in our experiments. DistilBERT is a compact and computationally efficient version of the BERT (Bidirectional Encoder Representations from Transformers) model, developed by Hugging Face. It retains much of the language understanding capabilities of BERT but with fewer parameters, making it quicker and less resource-intensive for natural language understanding tasks. DistilBERT achieves this efficiency through a process called "knowledge distillation," where it is trained to mimic the behavior of a larger pre-trained model.

We also tried DistilBERT embedding with machine learning models, but that combinations performed worse than ones with TF-IDF.

3 Experimental Setup and Results

The modeling phase commenced with a 80%-20% train-test data split. Neural network models utilized a validation set from the training data for monitoring training and determining optimal epochs, while traditional machine learning models underwent 5-fold cross-validation for hyperparameter tuning. Model performance was assessed using accuracy measure.

The LSTM results (see Table 1) were unsatisfactory, primarily due to insufficient data for training the embedding layer from scratch. Notably, the TF-IDF feature space seemed to favor linear models like LR and SVM with linear kernel over more complex, non-linear tree-based models like Random Forest and LightGBM. Stacking the best-performing models further improved accuracy on the hold-out dataset.

Model	Val acc	Test acc
DistilBERT	**87.05%**	**86.79%**
Stacking Ensemble	–	84.64%
Logistic Regression	82.04%	83.93%
SVM	82.13%	83.21%
Naive Bayes	80.97%	81.43%
Random Forest	78.02%	73.57%
LightGBM	73.38%	72.5%
LSTM	70.54%	66.79%

Table 1: Accuracy for evaluated models.

In terms of the highest validation accuracy, the DistilBERT classifier outperformed all others, achieving approximately 87% accuracy on both validation and test sets. To fine-tune the model, we utilized the Adam optimizer with a learning rate of 10^{-5} and employed sparse categorical cross-entropy as the loss function, ultimately yielding the best validation accuracy after 16 training epochs.

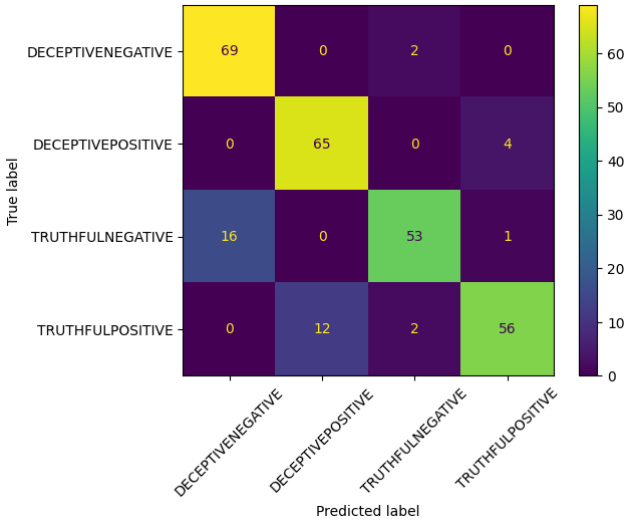

Figure 3: Confusion matrix for the DistilBERT model predictions on the test set.

4 Discussion

We selected DistilBERT as our final model primarily because it exhibited the most coherent error patterns in comparison to the predictions of other models, as illustrated in Figure 3. Notably, the final model displayed no instances of double errors, successfully distinguishing either sentiment or truthfulness. It made only three sentiment recognition errors in both deceptive and truthful review groups, showcasing a notable ability to discern changes in the tone of these reviews. Analyzing these three reviews reveals a distinct shift in their tone. In one instance, a customer initially voiced dissatisfaction over stolen Gucci glasses but later praised the hotel. Similarly, another reviewer began with contentment about their choice but transitioned into complaints about additional charges. This trend is mirrored in the case of a family who concluded a positive review with grievances about noise levels and cleaning hours.

Significant challenges were predominantly encountered when classifying suspicious reviews. Notably, six deceptive reviews were mistakenly classified as truthful, and an even greater number, 28 truthful reviews, were inaccurately classified as deceptive. These errors may be attributed to distinct characteristics observed within most of the truthful reviews, which often manifest as more vivid, personal, and comprised of longer sentences. It's plausible that these traits led to confusion for the model, resulting in misclassifications.

Conversely, the model displayed a discernible bias in its sensitivity towards suspicious reviews. This bias appears to be linked to the presence of shorter, more factual, and machine-like sentences in the text. This pattern was observed to be prevalent in reviews where sentences exhibited a matter-of-fact tone, such as "We were disappointed with this hotel. The staff seems hurried and unhelpful. Our room was not made up until after 5 PM despite several requests..." or "I recently stayed here for the Chicago triathlon. This was my third stay at this hotel. I have not had any issues until this visit...".

5 Future work

To elevate the performance of our existing system, we might consider delving into more computationally demanding sentiment analysis solutions such as XLNet or RoBERTa. This exploration would involve not only adopting these advanced models but also fine-tuning them to harmonize with optimal preprocessing techniques tailored to our specific dataset.

Furthermore, we could explore a strategic approach by segregating the problem into two distinct facets: sentiment classification and truthfulness assessment. This separation would allow us to develop specialized models for each aspect, fine-tuning them independently to excel in their respective domains. Finally, we would unify the strengths of these specialized solutions, potentially through ensemble techniques, to create a comprehensive and robust sentiment and truthfulness analysis system. This multifaceted strategy aligns with our goal of improving the accuracy and depth of our analytical capabilities.

References

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.