Truthful vs. Deceptive Hotel Reviews (Group 18)

1st Ricky Xu
Computer Science and Engineering
University of Lisbon
Lisbon, Portugal
ricky.xu@tecnico.ulisboa.pt

2nd João Vasco Reis
Computer Science and Engineering
University of Lisbon
Lisbon, Portugal
joao.vasco.sobral@tecnico.ulisboa.pt

I. Models

In this project, we used a Linear Support Vector Classifier, a Gradient Boosting Classifier, a K-Nearest Neighbors Classifier, and a Multinomial Naive Bayes Classifier for our classification tasks.

Linear SVM, a type of Support Vector Machine, classifies data by identifying the hyperplane that best divides a dataset into classes. It is frequently used in text classification and can handle large datasets.

Gradient Boosting Classifiers combines predictions from multiple weak learners, normally decision trees, in order to create a better model. It builds trees sequentially, with each one of them fixing the errors from the previous ones, obtaining the final result from all trees. It can capture complex patterns and is resistant to overfitting.

KNN classifies new data points based on the classification of their neighbors. It is simple and intuitive and can adapt to new data easily.

Multinomial Naive Bayes is a probabilistic classifier based on Bayes' theorem with an assumption of multinomial distribution for the features. It is simple and fast and is suitable for multi-class classification.

The Linear Support Vector Classifier (Linear SVC) emerged as the top-performing model in our experiments, delivering the most promising results. Consequently, we will focus our discussion on this particular model in our paper.

In the preprocessing stage, our approach began with converting all words to lowercase. This initial step ensured uniformity in the input data, preventing unnecessary word duplications. Following this, the reviews were tokenized to facilitate lemmatization. During lemmatization, each word was checked against a stopword list from NLTK, and if it was not identified as a stopword (with some stop words removed based on their importance for classification), it was lemmatized. This process helped in reducing words to their base forms and improving the consistency of the data. Lastly, the tokens were rejoined to reconstruct the fully preprocessed review. Punctual signs were not removed due to their importance for the classification.

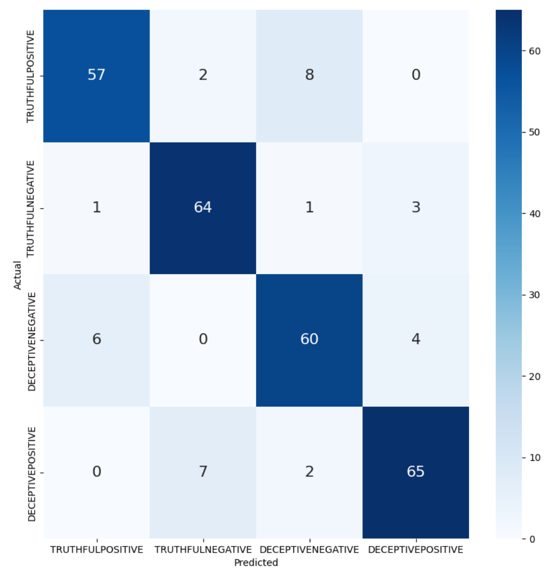

Fig. 1. Confusion Matrix of Best Model

Model	**Accuracy (%)**
Linear SVC	87.86
Gradient Boosting	69.93
KNN	64.29
MultinomialNB	79.29
Linear SVC(TP)	87.84
Linear SVC(TN)	85.71
Linear SVC(DN)	85.07
Linear SVC(DP)	92.75

TABLE I
ACCURACY OF DIFFERENT MODELS. NOTE: TP = TRUTHFULPOSITIVE, TN = TRUTHFULNEGATIVE, DN = DECEPTIVENEGATIVE, DP = DECEPTIVEPOSITIVE.

II. Experimental Setup and Results

Our dataset comprised of 1400 reviews, distributed as follows: 350 TRUTHFULPOSITIVE, 352 TRUTHFULNEGATIVE, 346 DECEPTIVEPOSITIVE, and 352 DECEPTIVENEGATIVE reviews. This distribution demonstrates a near-equal representation for all four categories, signifying a balanced dataset. In the context of review classification, accuracy served as our pivotal evaluation metric.

The results were gleaned by employing the train_test_split function on the train.txt dataset, setting the test_size parameter to 0.2 and fixing the random_state at 17. The reviews were processed using TfidfVectorizer, facilitating the quantification of each word's relevance within the corpus. Our top-performing model was trained using a 'linear' kernel.

In the realm of high-dimensional text data, KNN's efficiency tends to dwindle due to the "curse of dimensionality." This phenomenon results in reduced discriminatory power for distances. In contrast, Linear SVC, which adeptly identifies hyperplanes in such spaces, often trumps KNN in text classification tasks, such as sentiment or deception detection. This discrepancy was evident in our study, with KNN registering as the least accurate model we evaluated. Other models like Gradient Boosting and MultinomialNB also didn't achieve the anticipated results, as showcased in **table I**.

III. Discussion

In our recent **NLP analysis**, we encountered discrepancies between our model's classifications and the expected outputs. Below, we highlight some of these mismatches, provide abridged versions of the statements, and offer insights into potential reasons for the misclassification. Being TN,DN,TP,DP the abbreviations for (True Negative, Deceptive Negative, True Positive and Deceptive Negative):

Statement 1 : The Sheraton was a wonderful hotel! (...) The beds are absolutely to die for (...) They had a really nice restaurant inside with excellent food.

Model Classification: **DP**

Expected Classification: **TP**

Insight: Our model might have been influenced by terms like "I wanted to take it home with me", "to die for", which in this context is positive, which our model got right, but it might be considered as an exaggeration, and our model classified it as a Deceptive result.

Statement 2 : It is always interesting going to a high quality Hotel. (...) But, the room was not well cleaned (...) they stole a watch that I left in the room. (...) Stay elsewhere.

Model Classification: **DN**

Expected Classification: **TN**

Insight: While the hotel description suggests quality, the theft of a watch is an intense negative experience and it might be considered as an exaggeration. We concur with our model's classification, emphasizing the unexpected negative twist.

Statement 3 : Great hotel with beautiful scenery! (...) They even had a docking station for my iPod!

Model Classification: **TP**

Expected Classification: **DP**

Insight: This statement is ambiguous. The details presented can easily be associated with genuine experiences, leading to the model's potential confusion.

Statement 4 : We stayed for a one-night getaway with family on a Thursday (...) Only flaw was breakfast was pricey (...) A gem in Chicago.

Model Classification: **TN**

Expected Classification: **TP**

Insight: The use of words like "steal" in relation to the AAA rate, combined with some unfamiliar terms, might have skewed the model's perspective, leading to a perception of negativity.

Statement 5 : When I was asked to stay at the Swissotel in Chicago, (...) The room was neatly fixed (...) If I were ever to be in that area again, the Swissotel is where I'm staying.

Model Classification: **DN**

Expected Classification: **DP**

Insight: This seems to be an anomaly, as there are no glaring indicators for such classification. It might be an oversight or a statistical aberration from the model.

Statement 6 : Upon first entering the hotel, we were greeted (...) It was a very enjoyable stay.

Model Classification: **TP**

Expected Classification: **TN**

Insight: We concur with our model's classification as the narrative is overwhelmingly positive. The training data might have mislabeled this instance.

Summary: The discussed misclassifications highlight the intricate challenges in NLP tasks. Factors like contextual interpretations, inherent ambiguities, and possible dataset mislabeling can influence model predictions. Our findings underscore the importance of iterative model refinement and comprehensive dataset scrutiny to achieve higher accuracy.

IV. Future work

In our pursuit of an advanced and accurate hotel review classification system, we acknowledge certain avenues that would further enhance its capabilities. Given more time, here's what we would prioritize:

- **Data Preprocessing**: Refine preprocessing to address colloquialisms and unconventional terms, e.g., transforming "yr" to "year" and filtering terms like "IMO".
- **Dataset Alignment**: Some of the labels from the training where disagreed by us. In case of more time we would remove some of the reviews that are labelled with something we disagree. For example on Statement2 we disagree with the training set label.
- **Deep Learning**: Explore advanced architectures like RNNs, LSTMs, and Transformer models for enhanced text understanding.
- **Data Augmentation**: Enhance dataset diversity using techniques such as back translation and synonym replacement.

Conclusively, as we advance, focusing on these dimensions will be critical. Our end goal is to have a hotel review classifier that isn't just technically robust, but also sensitive to the myriad nuances of human expression and guest experiences in the hospitality sector.

V. Bibliography

https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/

https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/

https://www.geeksforgeeks.org/removing-stop-words-nltk-python/