LN - MP2 Report

João Cardoso — ist199251, Sara Monteiro — ist1108402, Group 75

1 Models

In our initial approach, we applied a basic Naive Bayes classifier to the dataset preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) method, based on this notebook. By doing so, we were able to achieve an accuracy score of 84%. Despite its good performance, Naïve Bayes works on the assumption that input words are independent, and, as such, could fail to represent the complexity of the classification classes. After this baseline, we then tested a few BERT-based models. Contrary to the Naïve Bayes classifier, BERT explicitly models the contextual relationships between words, whether it be through word embeddings or through its attention mechanism that assigns different weights to each word based on their relationship with the other words in the sequence. After some experimenting with BERT-like models (including distilBERT, BERT, RoBERTa, and ALBERT), we ended up using a model we found on Hugging Face, which had been fine-tuned on movie reviews (IMDB dataset). Despite the differences in the problem domain, we found this transfer-learning approach to have a slight edge over the others, with faster convergence towards optimal scores.

2 Experimental Setup and Results

We used a manual 85-15 test split (we singled out 200 reviews into another file, although we use sci-kit learn in the provided code), and trained the model using 4-fold cross-validation on the training set, keeping an eye on the relationship between the score on the test and validation sets, and tuning the hyperparameters accordingly. We ended up using Adam as an optimizer, with a learning rate of 3×10^{-5}, a dynamic number of epochs according to validation set performance, batch size of 16, and a cross-entropy loss function. For further implementation details, please refer to the code.

Due to the small dataset, we wanted to use as much of the training data as possible. As such, we carried over these parameters onto a training run of the model on the entirety of the provided dataset. Despite not being able to know how our model will perform on the unlabeled set, we are confident it will perform significantly better, since the training set is so small. This means, however, that the results on the test set we present here may be considerably lower than on the unlabeled reviews.

The evaluation results for the BERT model fine-tuned with the IMDB dataset are shown in figures (1) and (2). The overall accuracy was 89%, which is on par with the current state of the art for this task.

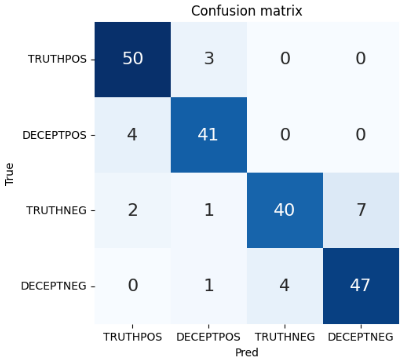

Figure 1: Confusion matrix

Label	Precision	Recall	F1-Score	Support
TRUTHPOS	0.89	0.94	0.92	53
DECEPTPOS	0.89	0.91	0.90	45
TRUTHNEG	0.91	0.80	0.85	50
DECEPTNEG	0.87	0.90	0.89	52

Figure 2: Evaluation Metrics

3 Discussion

From the get-go, we surmised that the task would be challenging, even for a human classifier. Since all reviews were written by humans, there were no clear instances of nonsensical grammar or phrasing. Furthermore, there was significant overlap between the topics covered across all reviews.

However, looking at the failed predictions, we did notice that deceptive reviews exhibited awkward phrasing, with collocations that would seem atypical for a native English speaker. There also seemed to be a repetition of topics, such as inventing a family member the reviewer had stayed with. By looking at the TF-IDF scores for the different labels, we noticed further that truthful reviews mentioned some practical details (such as booking through Priceline), which nonetheless are quite subtle.

The performance of our model seems to be in line with what we identified. The difficulty of the domain was evident - no matter the models we tried, or the pre-processing we did (testing included cased vs. uncased model, separating labels into 2 categories of 2 labels each, passing n-grams to the model, adding POS labels to each token, using different pre-trained models, tuning hyperparameters), we never seemed to be able to get scores well above 90%. A cursory look at the confusion matrix reveals that mislabeling took place mostly across truthfulness/deceptiveness lines, especially for negative reviews, with largely correct sentiment attributions. Some examples are now in order:

```
truth DECEPTIVENEGATIVE; prediction TRUTHFULNEGATIVE
I arrived at the hotel 15 minutes prior to check in time as they recommend. When I get
there I noticed (...)

truth DECEPTIVEPOSITIVE; prediction TRUTHFULPOSITIVE
On our first trip to Chicago, my fiance and I stayed 2 nights on our anniversary and
were pleasantly surprised by the Hotel (..)
```

In the first instance, the model was not able to pick up on the unnatural and incorrect verb tense (”get”). In the second instance, the mention of a ”fiancé” does not tip the model off, something a human might have picked up on. Having said this, what the model struggles with overall is the awkwardness or unnaturalness of the review. There is a ”robotic” quality to the deceptive reviews which is quite hard to quantify or pin down. Lastly, the model failed to identify some sentiments:

```
truth DECEPTIVENEGATIVE; prediction DECEPTIVEPOSITIVE
My recent stay (...) couldn’t have gone worse. (..) There are many attractions in the
area and indeed leaving the hotel is the best part of the experience.
```

In this case, the model seems to be misled by the ironic language at the end, and misses the semantics of the sentence, although it correctly identifies the review as deceptive.

Lastly, some reviews appear to be mislabeled in the dataset:

```
truth TRUTHFULNEGATIVE; prediction TRUTHFULPOSITIVE
The location is ideal. They have very high ceilings so the rooms appear much larger
than they are. The views are great. There were fresh flowers in the bathroom. They had
all white bedding. You feel like a princess there.
```

4 Future work

If given more time, we would fine-tune a model to be able to classify sentences as written by a native speaker or not. From the original paper describing the creation of the dataset, deceptive reviews were sourced using Mechanical Turk, whose workers are spread worldwide, and are often non-native. In real life scenarios, the provenance of deceptive reviews would be similar. We believe a model trained on picking up subtle differences in phrasing between native and non-native text might avoid some of the aforementioned issues. Given the availability of models and datasets for Native Language Identification(NLI) tasks, we have reason to believe fine-tuning a model for this task would be feasible, as it is a relaxed version of a standard NLI task.