# Deception Detection

*Detecting deceptive Hotel and Electronic reviews using BLT dataset*

**Nikhil Sulegaon**

09.09.2019

## INTRODUCTION

People today rely a lot of online reviews before making any kind of purchases. In the recent times, there has been an increase in the number of fake online reviews - from spams to misleading opinions. Thus, there is a critical need for a system that can detect and filter deceptive reviews. In this project we come up with a way to detect deceptive/fake Hotel and Electronic reviews by making use of the Boulder Lies and Truth (BLT) dataset.

## DATASET

The BLT dataset is composed of online reviews of Hotels and Electronics that was written by Amazon Mechanical Turkers. It categorizes reviews as either Truth, False or Deceptive. Each review has a Sentiment polarity associated with it. There are about 1580 reviews in total. The table below shows the distribution of Truthful and Deceptive reviews by sentiment.

| **Sentiment** | **Truth** | **False** | **Deceptive** |
|---|---|---|---|
| + | 240 | 237 | 314 |
| - | 238 | 240 | 311 |

Reviews with a 'Truth' label are those reviews where the Turker submitting the review has used a product/service and is giving a genuine opinion about his/her experience.

Reviews with a ‘False’ label are those reviews where the Turker has made use of the product/service but are the opposite of what the Turker actually experienced - positive or negative. ‘Deceptive’ reviews are those reviews where the Turker has not used a product/service and is fabricating a fake opinion.

## MODEL

### FEATURES

Research has shown that certain words in fake reviews have much higher frequencies than in non-fake reviews. These high frequency words actually imply pretense and deception[2]. Thus, the reviews itself are cleaned, tokenized and used as features.

It is also very important to consider certain behavioral features of the person writing the review as writing a fake review you has a lot of common psychological traits associated with it. Thus the following aspects were used as behavioral features.

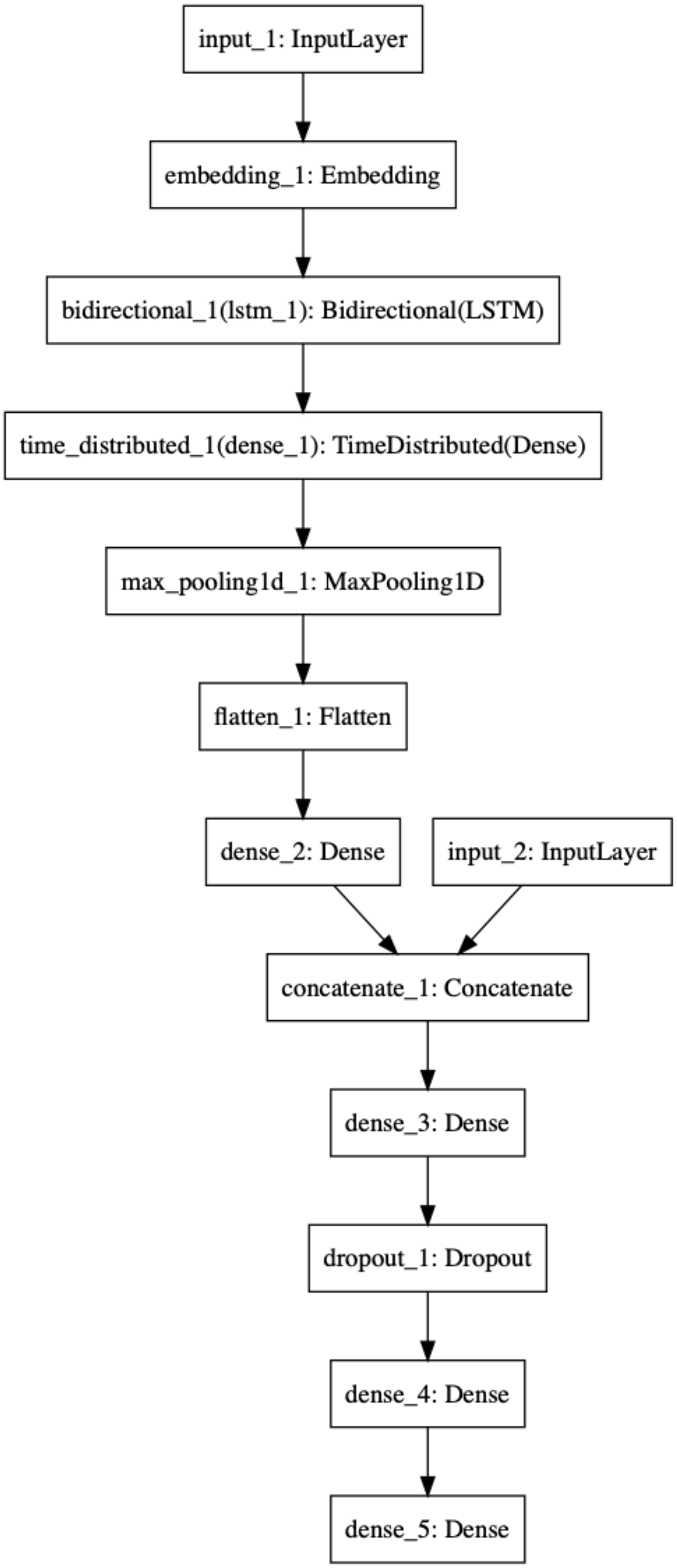
- *Review Count*: Number of reviews that a Turker has submitted.
- *Review Length:* The length of the review
- *Sentiment Polarity:* The actual sentiment of the fake review - whether positive or negative.
- *Similarity Scores:* The average Word Mover’s Distance between all the reviews written by a Turker per Domain(Hotel or Electronics). This score tells us how similar the fake reviews submitted by a Turker are. Thus, it will be easier for the model to predict the Truth Value of the review given the Worker ID, Domain and how similar the reviews by the Turker usually are.

### ARCHITECTURE

Since we are classifying sequential patterns(online reviews) and patterns lying ahead in the sequence influence the classification of the current pattern, using a Bidirectional LSTM make for a good choice to learn the linguistic features in online reviews.

However, for the model to learn behavioral features like the ones mentioned above, the behavioral features are concatenated with the output of the Bidirectional LSTM and is fed into a 2-layer Neural Network that classifies the review into one of the 3 categories - T, F, or D.

The figure below shows the entire architecture of the model with the Embedding layer, Dropout layer and a MaxPooling layer.


input_1: InputLayer
embedding_1: Embedding
bidirectional_1(lstm_1): Bidirectional(LSTM)
time_distributed_1(dense_1): TimeDistributed(Dense)
max_pooling1d_1: MaxPooling1D
flatten_1: Flatten
dense_2: Dense
input_2: InputLayer
concatenate_1: Concatenate
dense_3: Dense
dropout_1: Dropout
dense_4: Dense
dense_5: Dense

## TRAINING AND TESTING

Information in the BLT dataset is organised into the following columns - ['Review ID', 'Review Pair ID', 'Worker ID', 'Review', 'Domain', 'Sentiment Polarity', 'Truth Value', 'URL Origin', 'Length in Bytes', 'Avg. Quality', 'Accuracy in Detecting Truthfulness', 'Avg. Star Rating', 'Time to Write a Review Pair (sec.)', 'Quality Judgments', 'Truth vs. Deception Judgments', 'Star Rating Judgments', 'Known', 'Used/Stayed', 'URL', 'Plagiarized', 'Rejected'].

This dataset is fit cleaned by extract the reviews, tokenizing them and removing all the stop words from them. Then all the behavioral features are either computer (Average Word Mover's Distance) or are selected from the dataset(Sentiment Polarity, Domain, etc).

Word Embeddings of theses cleaned, tokenized reviews are first computed using Keras with Embedding space = 100. These embeddings are then passed into the Bidirectional LSTM. The output of which is concatenated with the Behavioral features and passed into a 2-layer Neural Network that finally classifies the reviews.

Each review is written by a Turker and belongs to a Domain. Thus, reviews are grouped by Worker ID and Domain, and the Mean and Standard Deviation of Word Mover's distance is computed for each Worked ID and Domain. The same is assigned to every review depending on the Worker ID and Domain. These Mean and Std. Dev Sim scores are used a Behavioral feature of the Review.

The Bidirectional LSTM and 2-layer Neural network are trained together. 10-fold Cross Validation is performed while training the composite model. The train-test split is set at 90-10. The validation set being 10% of the training set. The model is trained for 50 epochs with a batch size of 32.

The program also provides an option to train just the Bidirectional LSTM as well. It also gives the user an option to condense 'False' and 'Deceptive' labels into one category while cleaning the dataset. Refer to the Readme for more information.

## RESULTS

Training the composite model with behavioral features on Google Cloud Platform gave an average Test accuracy of **70.45%**. However, training just the Bidirectional LSTM to learn Linguistic features led to an accuracy of **66.78%**. The Training logs in the results directory of the project indicate the same. These results are a huge improvement to the results achieved by previous work solving the same problem using the BLT dataset!

## REFERENCES

1. A Tangled Web: The Faint Signals of Deception in Text - Boulder Lies and Truth Corpus (BLT-C)
2. V. Sandifer, Anna & Wilson, Casey & Olmsted, Aspen. (2017). Detection of fake online hotel reviews
3. A. Mukherjee, V. Venkataraman, B. Liu and N. Glance, "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews
4. Automatic detection of deceptive opinions using automatically identified specific details Nikolai Vogler
5. Sentence classification using Bi-LSTM
6. From Word Embeddings To Document Distances
7. Gensim
8. bidirectional LSTM + keras
9. Evaluate the Performance Of Deep Learning Models in Keras