Diego Valdes

IST 736

Home Work 2

10/19/2019

**Introduction**

The truth has been one of the most discussed topics among the public in recent times. More than ever, what is the truth entered American’s daily lives, from reading shopping reviews before clicking the buy button, the candidates running for office, and reading the news. While one could argue that truth was never a part of politics (a topic for a different analysis), the news has always had a presumption from consumers of truth. With the increase in social media and the importance consumers place on online reviews, it is not surprising that this medium has also been pierced with falsehoods as companies look to use every avenue to remove dollars from consumers.

In person, detecting lies and sentiment may be a simple matter for some due to social queues. Those queues are lost in text. It is extremely difficult to read a text and make a determination on truth and sentiment. Sarcasm is a great example of something very easy to detect in person due to tone of voice and delivery but can be taken literally (which is the incorrect meaning and interpretation) if read in text.

It is also a data question. Although it is advantageous to have data in numbers and categories, most data is unstructured and text. Any party interested in data (which should be all of them) needs a way to make sense of text data. It is the only way to derive value from that data.

**Analysis**

**About the Data**

The data is a collection of restaurant reviews. Each row is one review and it is labeled with a sentiment (positive/negative) and lie (false/true). The dataset is presented in both comma and tab delimited files. The csv file was chosen to clean, due to when both files were initially brought in, the csv file was less disastrous.

The dataset is intended to have 3 columns (lie, sentiment, review), but due to the reviews having standard punctuation, when attempting to process, more columns are generated; each review is broken up where a comma appears. Each review was stripped of all punctuation to prevent this from occurring. No letters were removed during this cleaning, instead, the step of removing words based on length or category would be made while prepping for analysis. The final clean dataset appears as intended with 3 columns and each review stripped of punctuation.

For analysis, multiple datasets were generated by controlling stop words, word length, and word frequency. No dataset with less than 100 columns was used for modeling, with the biggest having over 1500 columns.

**Models**

Multinomial Naïve Bayes was used to model all the generated datasets. Implementation required separating the sentiment/lie from the reviews. Testing and training sets were generated using a 70/30 split.

**Results**

For lie detection, predictions varied depending on the dataset, but the results were 100% in one label, either they were all true or false. Multiple attempts and strategies were tried to produce variations in the results, from varying the size of the training data, changing the length of words to be removed, and the inclusion/exclusion of stop words. Unfortunately, in every test, the results were always true. At best, results ranged ~50%, but this was more based on the test set than the model’s accuracy.

For sentiment, predictions varied between 100% positive and 100% negative, depending on the parameters used to generate the dataset. Datasets with more columns tended to produce positive sentiment while smaller sets, generally ~100 columns tended to produce negative sentiment. Word length variance ranged from 2-5, but this didn’t effect the outcome more than the size of the data. As with lie detection, this resulted in results no greater than 50% accuracy.

A model was generated using a new generated column where sentiment and lie was combined into one label, ie: pt, pf, nt, nf. Results from this did produce more variation than previous models, but the results were either pt(positive true) or nf(negative false) and the accuracy was in line with the frequency of possible results: 25%.

**Conclusion**

Neither model produced an accuracy greater than simple guessing could have produced, even though much of the academic readings and presentations dictate that Naïve Bayes is one of the best models for text classification, the results here do not show that. These results can be interpreted in one of two ways.

The first is that Naïve Bayes isn’t as good for sentiment analysis and lie detection as many of the academic readings and presentations have led to believe. This may be due to the assumption of independence in the algorithm and language word choice does not have that trait. One word following another is not independent of the word that came before.

The second is that the analysis performed here is flawed. While multiple datasets and variations were attempted to produce increased results, none did better than a guess. Perhaps the solution is in the preparation of the data, the tokenization of the words, and the selection/inclusion of stop words needs to be handled in in differing manner.

With those choices laid out, the answer is that there is a flaw in the analysis. Due to the guesses being 100% in one direction for two models, and a coin flip in the other, and various observations in results due to alterations in stop words, and word length/frequency. The models created in this analysis are not suitable for any reliable analysis, and the best course of action is to reexamine the approach in data preparation and vectorization.