Diego Valdes

IST 736

Home Work 6

11/10/2019

**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 two datasets were generated, one with word frequency and the other a binary representation of word existence in the review.

**Models**

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

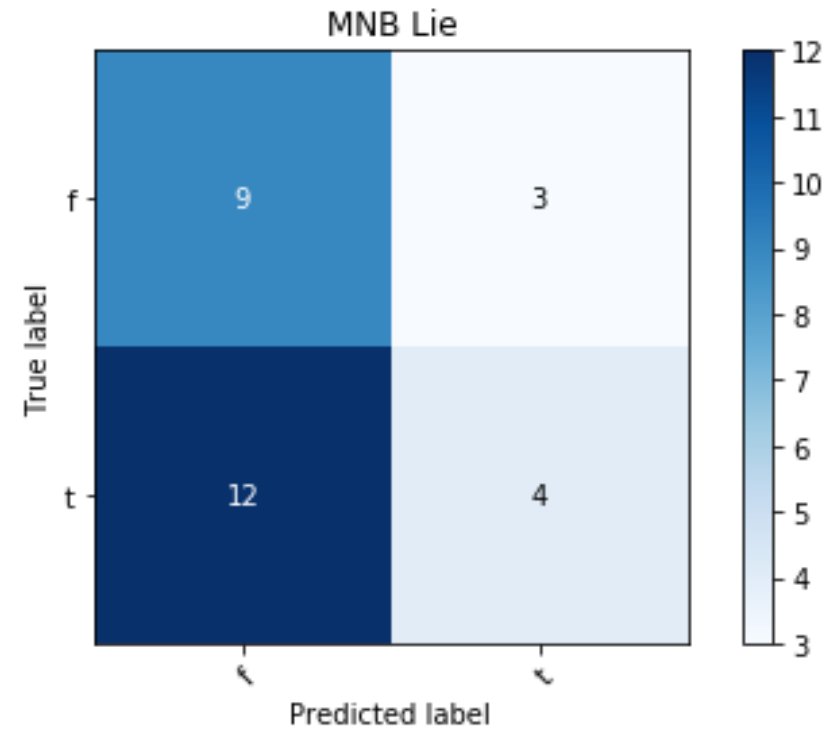
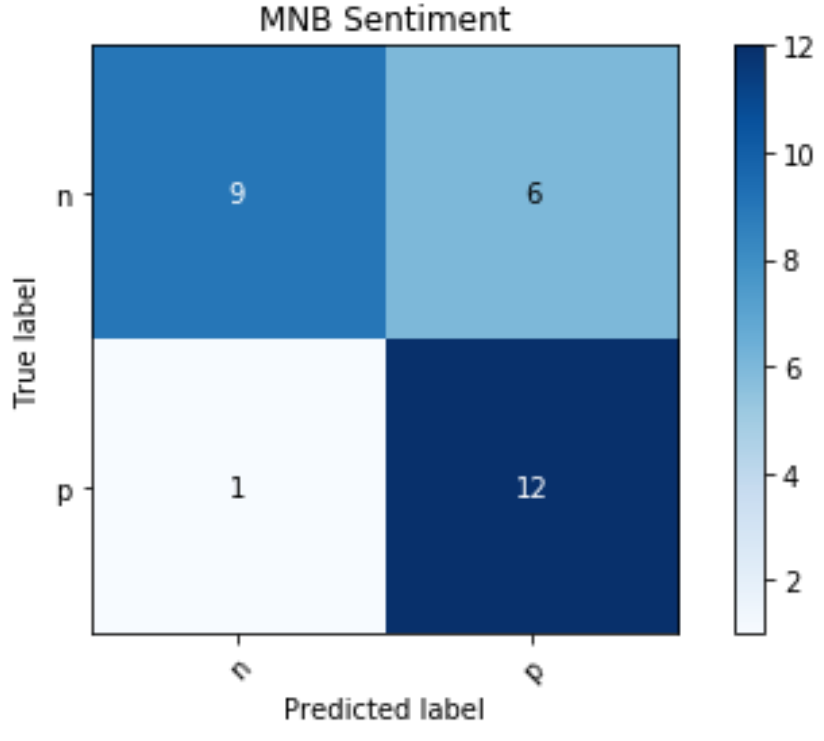
Based on results of a previous analysis, the frequency dataset was not converted to a percent. This was due to the poor performance of the model and a desire to try a new approach.

**Results**

**Multinomial Naïve Bayes**

Regardless of alterations made to word choice (inclusion of stop words) and word length, results for lie detection rarely exceeded 50%. When compared to previous models generated in past analysis, this model’s predictions didn’t lean 100% in one direction, but the results remained roughly the same.

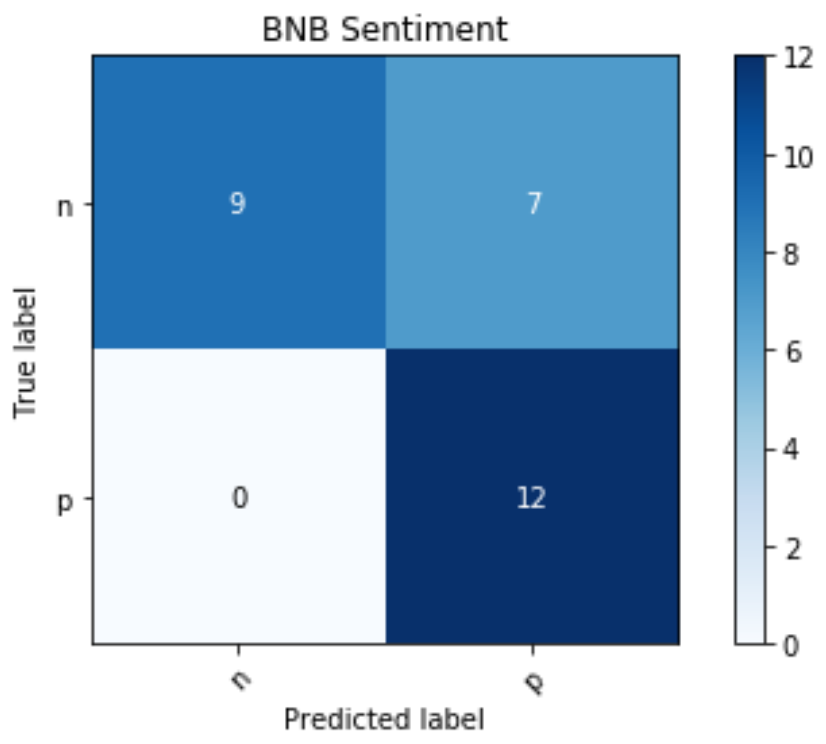
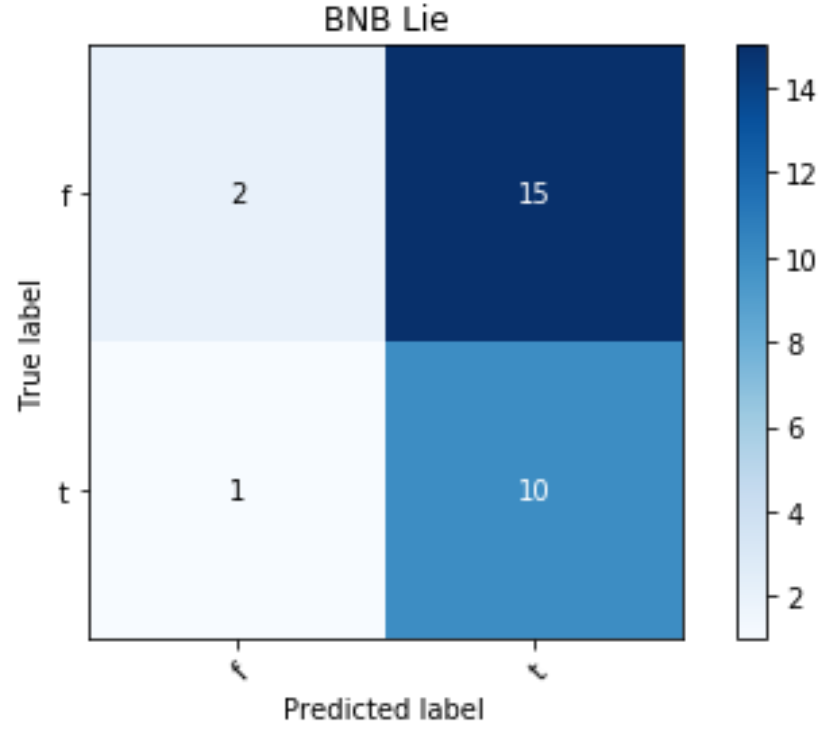
When examining sentiment, the model performed much better and didn’t require much coercion of the dataset. The model performed at 75% accuracy, with false positive being where the model had most error. The model performed well predicting negative sentiment, only erroring on one example.

**Bernoulli Naïve Bayes**

The same strategies used to attempt to derive better results in the Multinomial Naïve Bayes model were applied to the Bernoulli model. Results also remained the same, in none of the runs the accuracy of the model didn’t achieve an accuracy greater than 50%.

Sentiment, like in the Multinomial model, was easier to achieve good results. An accuracy of 75% was regularly achieved in multiple runs. These results match what was achieved with Multinomial and similarly, the most error was caused in false positive predictions.

**Conclusion**

Attempting to predict honesty with either Naïve Bayes model didn’t produce results better than random guessing. There could be a multitude of reasons for this, but at it’s core, it may be difficult to ask a computer to do something that humans can’t do on a regular basis. Unless the lie defies science or known facts, it is a matter of intuition to determine if a person is lying. It is more difficult to asses honesty in the written word, as much of the social queues that may ‘give away’ a lie are absent.

Predicting sentiment, however, is completely possible. Both models performed better than random guessing. If that is the measure, then both models are more than adequate. If a more reliable accuracy than 75% is needed, more work is required to get these models ready for deployment. Considering that the results were easily achieved, it may be possible to better prepare the data in order to raise accuracy. Another approach may be to provide more examples of positive sentiment, as both models largest errors occurred in the false positive predictions.

Overall, no real winner could be chosen between the Naïve Bayes models for predicting sentiment. If dealing with a larger dataset, then the options may be different, but for the movie reviews examined in this analysis, there was no clear choice.