COL774 Assignment 4.1

2022CS11621 - Jakharia Rishit

October 14, 2024

1 Competitive Part

In the competitive part we try to use all the columns of the LAIR dataset to get the best prediction of the credibility of the news:

1. Preprocessing
   - Normalization of Numerical Rows
   - Stemming, Stopwords Removal, and Tokenization
2. Trying Different Models
   - Bernoulli Naive Bayes with unigrams
   - Bernoulli Naive Bayes with bigrams
   - Multinomial Naive Bayes with unigrams
   - Multinomial Naive Bayes with bigrams
   - Multinomial Naive Bayes with trigrams
   - Logistic Regression
3. Varying Hyperparameters

1.1 Pre-Processing

Stemming, stopword removal, and tokenization simplify text for better analysis. Stemming reduces words to their root forms, stopword removal eliminates irrelevant words, and tokenization breaks text into smaller units. These steps help models focus on meaningful content and improve performance.

Normalization is important in logistic regression because it scales features to a common range, preventing large values from dominating the model's learning process. This ensures that all features contribute equally to the prediction, improving model performance and convergence speed.

1.1.1 Results

Though the order of 1e-3 change was observed it was not enough to conclude that tokenization proved to be too much helpful. However, in Logistic regression, a high benefit was observed after normalizing the numerical columns.

1.2 Data Exploration

1.2.1 Text Data

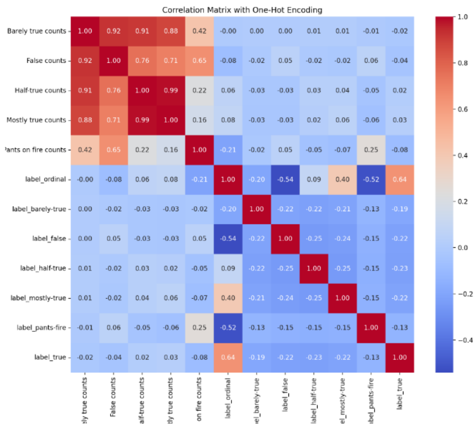
The following table shows the most common words vs the label from the text of the lair dataset.

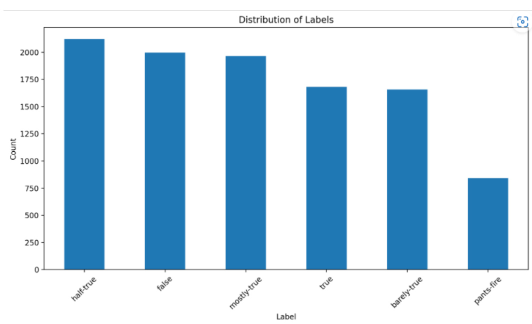
Table 1: Top 9 Features for Each Class

Rank	Barely-true	False	Half-true	Mostly-true	Pants-fire	True
1	republican	republican	republican	democrat	republican	republican
2	democrat	democrat	democrat	republican	none	democrat
3	us	state	state	state	democrat	state
4	state	new	us	us	say	us
5	senat	us	senat	senat	new	senat
6	texa	none	new	new	texa	texa
7	none	senat	texa	none	state	new
8	new	interview	none	florida	obama	none
9	say	texa	florida	interview	us	ohio

1.2.2 Numerical Data

Below are the correlation plots of the numerical columns and the one_hot, as well as the ordinal encoding of the labels and the distribution of a number of examples for each label.





From here, we can observe that, at least linearly, there does not exist a good correlation between the labels and the numerical data.

1.3 Varying Models

Several models were applied to model this data. The table below shows the results of the initial testing. Experimented with methods like **tf-idf** however it did not yield good output.

Table 2: Model Performance Summary

Model	Train Accuracy	Test Accuracy
Bernoulli Unigrams	0.6637	0.2399
Bernoulli Bigrams	0.7080	0.2157
Multinomial Unigrams	0.7392	0.2461
Multinomial Bigrams	0.9608	0.2539
Multinomial Trigrams	0.9802	0.2531
Multinomial Trigrams (α = 1.73)	0.9539	0.2562
Multinomial Trigrams (α = e - 1)	0.9540	0.2562
Multinomial Quadgrams (α = e - 1)	0.9791	0.2523

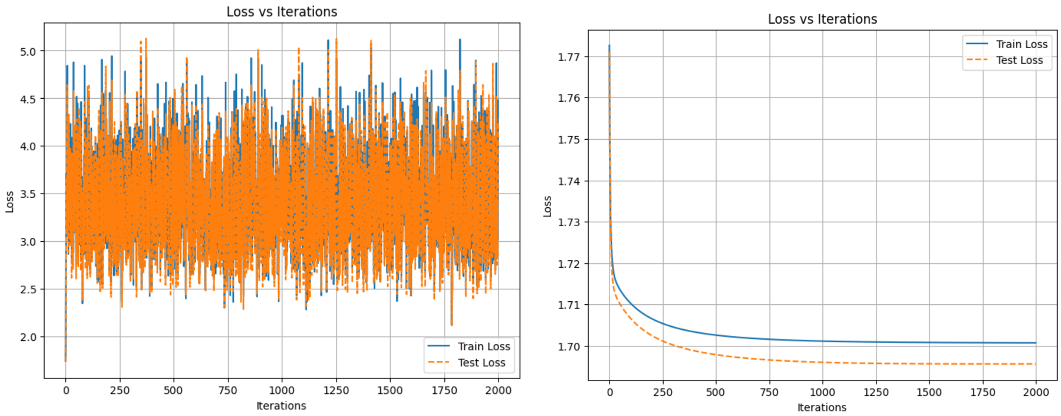
Here alpha refers to the Laplace smoothing parameter.

I did not see much improvement in Naive Bayes-related models, so I later experimented with logistic regression models. and the following are the results.

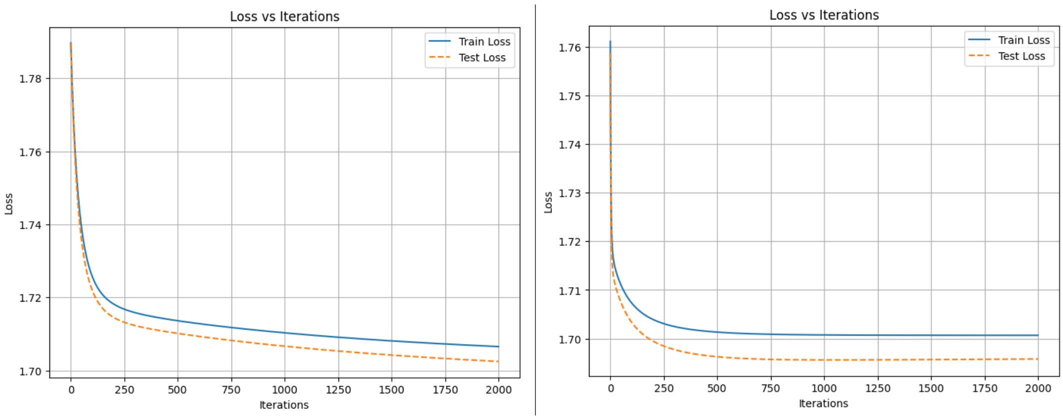
Table 3: Model Performance Summary with Learning Rate

Learning Rate	Train Accuracy	Test Accuracy
10	18%	20%
1	32.57%	33.02%
0.1	25.91%	26.24%
e-1	32.51%	32.78%
e-2	32.58%	**33.41%**

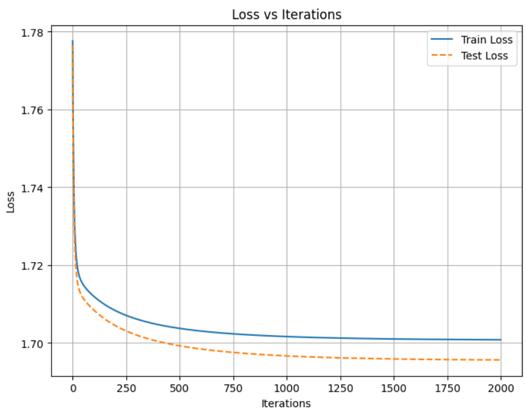
Below are the images of the train test loss curve.



Left: Learning rate=10 Right: Learning rate=1



Left: Learning rate=0.1 Right: Learning rate=e-1



Learning rate=e-2