Shreayan Chaudhary
February 7, 2023
Homework #1 Information Retrieval 601.666

Models I experimented with:

1. Logistic Regression
2. Gaussian Naive Bayes
3. Support Vector Classifier
4. Decision Trees
5. Random Forest
6. K Nearest Neighbours
7. XGBoost

Features:

Note: Let context words be the 3 words before the period and 2 words after the period.

1. Number of words to the left of the period before the next reliable sentence delimiter.
2. Number of words to the right of the period before the next reliable sentence delimiter.
3. Number of spaces following the period.
4. Capitalization of the first letter of the next word.
5. If the previous word is in the common abbreviations list. (example: m.p.h)
6. If the previous word is in the titles list. (example: mr.)
7. If the previous word is in the abbreviated time list. (example: a.m.)
8. If the previous word is in the abbreviated places list. (example: u.s.a)
9. Total number of periods in the context words.
10. Length of the previous word.
11. Length of the next word.
12. If the previous word has a period.
13. If the 2nd last word has a period.
14. Total count of context words that are unlikely to be proper nouns.
15. If the previous word is in upper case.
16. If the next word is upper case.

I used some more features that were useful but removed them since external libraries are not allowed:

1. Part of speech of the 3rd last word.
2. Part of speech of the 2nd last word.
3. Part of speech of the last word.
4. Part of speech of the next word.
5. Part of speech of the next to next word.

Some features that did not work well:

1. Total number of characters in the context words.
2. Length of the all the context words.

Performance and Metrics:

Baseline accuracy (using Decision Trees, without any additional feature): 93.338%

Model	Accuracy on training set	Accuracy on test set
Logistic Regression	95.83	94.04
Gaussian Naive Bayes	96.64	95.91
Support Vector Classifier	93.87	92.11
Decision Trees	99.57	99.13
Random Forest	**99.75**	**99.68**
K Nearest Neighbours	94.65	95.23

Result:

We can see that the Random Forest is working the best for the given problem. It gives an accuracy of 99.75 on the training data and 99.68 on the test data, which means it is not overfitting and is able to generalize well on new data.