Reddit NLP Analysis

Andres Aguilar

Problem Statement

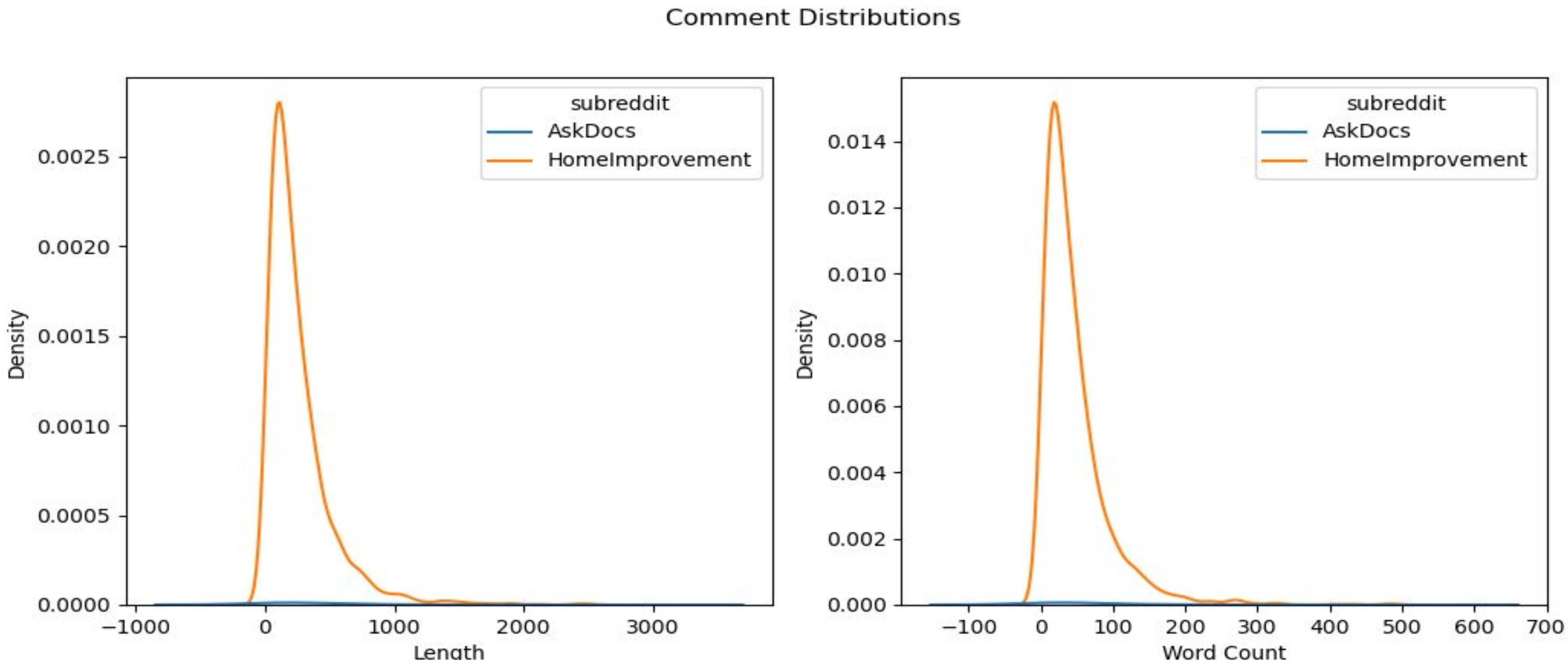
- Explore the classification of health and home questions and creating a model that can determine to which subreddit a post belongs.
- This model can have multiple use cases, such as redirecting questions from the internet to proper medical outlets.

- An accuracy of at least 99% to reduce the amount of misclassified medical questions.

Reddit API

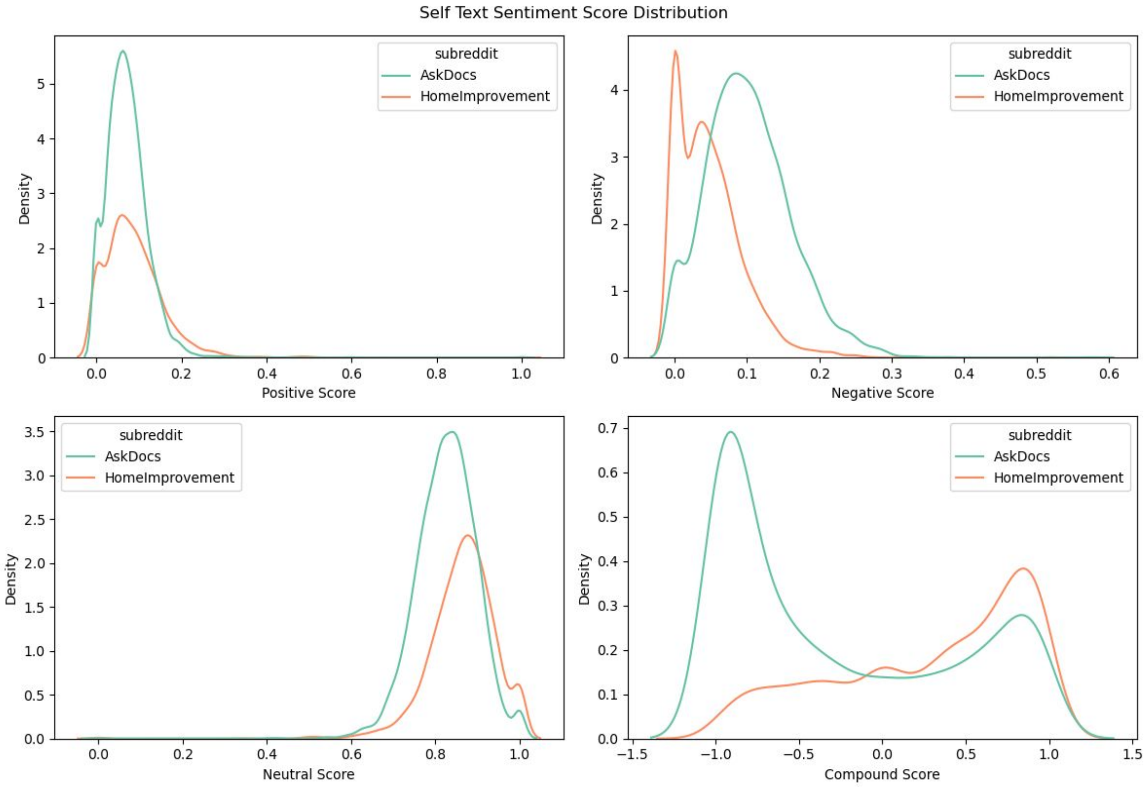
- PRAW was used to extract new submissions’ data from r/HomeImprovement and r/AskDocs
- This included:
    - Post Title
    - Post Self Text
    - Comments
    - Post and Comment IDs
- The data was collected with a ratio of 60:40, with majority for r/AskDocs

Text Analysis



Sentiment Analysis

Compound Sentiment score was only feature with significant difference in distribution.


Self Text Sentiment Score Distribution

Word Vectorizing

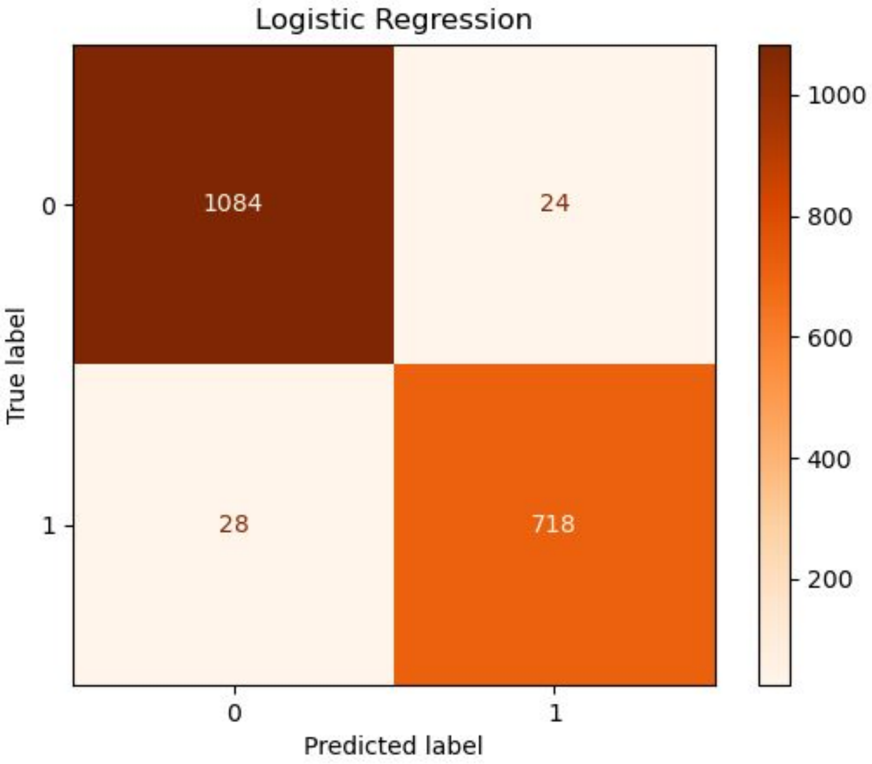
- Used two sets of TfidfVectorizers, one with single word vectors and the other with 2 to 5 word vectors.
  - Each only used words that were not commonly in 20% and 30% of the dataset.
  - Used to gain insights on more domain specific words and phrases.

Modeling

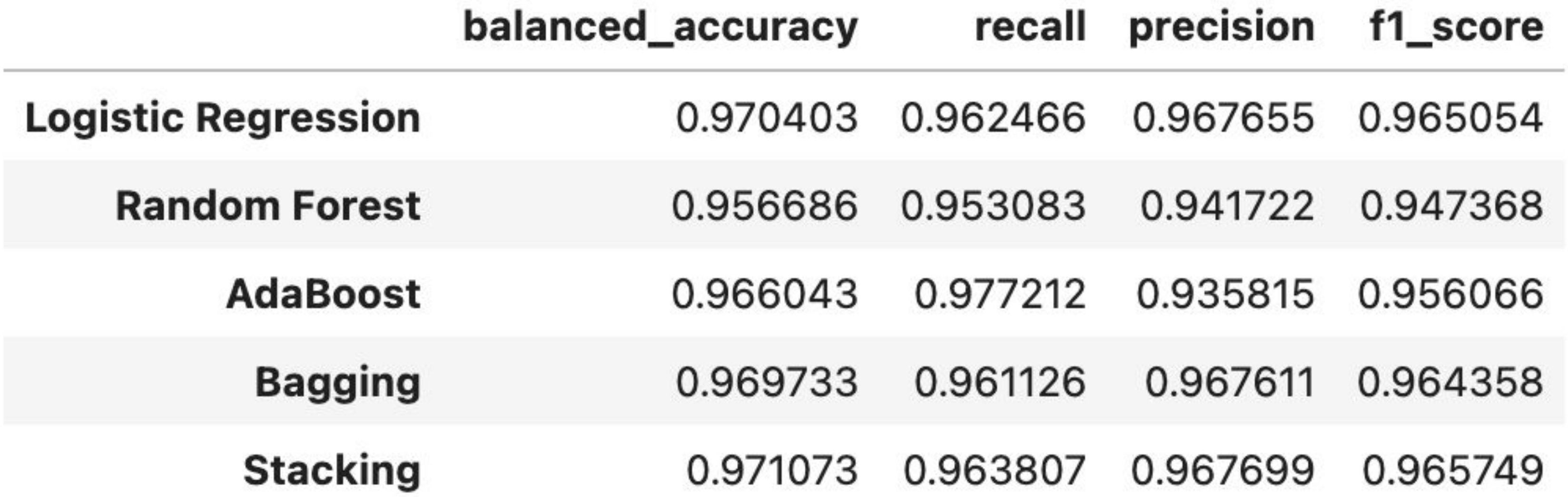
HomeImprovement -> 1

AskDocs -> 0

Had achieved an accuracy of 97%



Model Scores



	balanced_accuracy	recall	precision	f1_score
Logistic Regression	0.970403	0.962466	0.967655	0.965054
Random Forest	0.956686	0.953083	0.941722	0.947368
AdaBoost	0.966043	0.977212	0.935815	0.956066
Bagging	0.969733	0.961126	0.967611	0.964358
Stacking	0.971073	0.963807	0.967699	0.965749

Conclusion

The best models are the Logistic Regression and StackingClassifier. While they are both ideal, I would recommend the use of the Logistic Regression because it allows for more interpretability with its' coefficients.

- Did not achieve the 99% required for used case with medical question classification
- Can perhaps create improvements with additional word vector usage