ON REDDIT.COM AND THE EASTERN AND WESTERN CHURCHES

The purpose of this survey is to evaluate the potential for accurately classifying the subreddit source of a post or writing when the subreddits in question are divergent but closely related religious traditions.

The Eastern and Western Churches (i.e. Catholicism and Orthodoxy) share much of their history and tradition, but remain divided as the result of a nine-century schism (c. 1054).

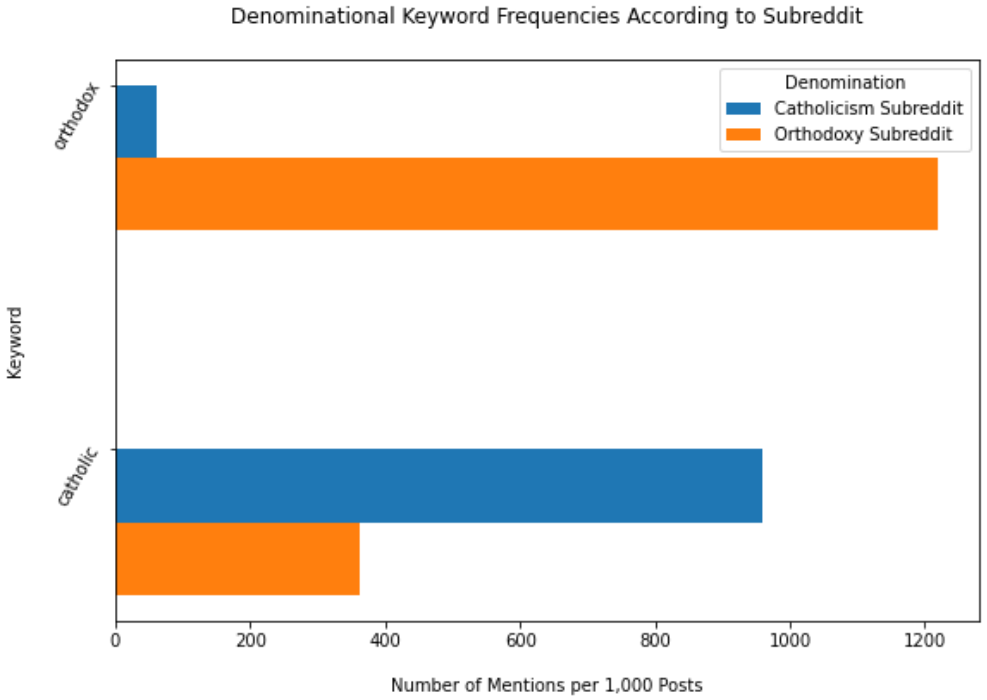
To the present day, it remains unclear whether they should be considered **two Churches divided by schism or a single Church divided by schism**.

Devout Catholics often consider their Church orthodox (small *o)* and devout Orthodox Christians still consider their Church catholic (small *c*).

SOME INITIAL FINDS

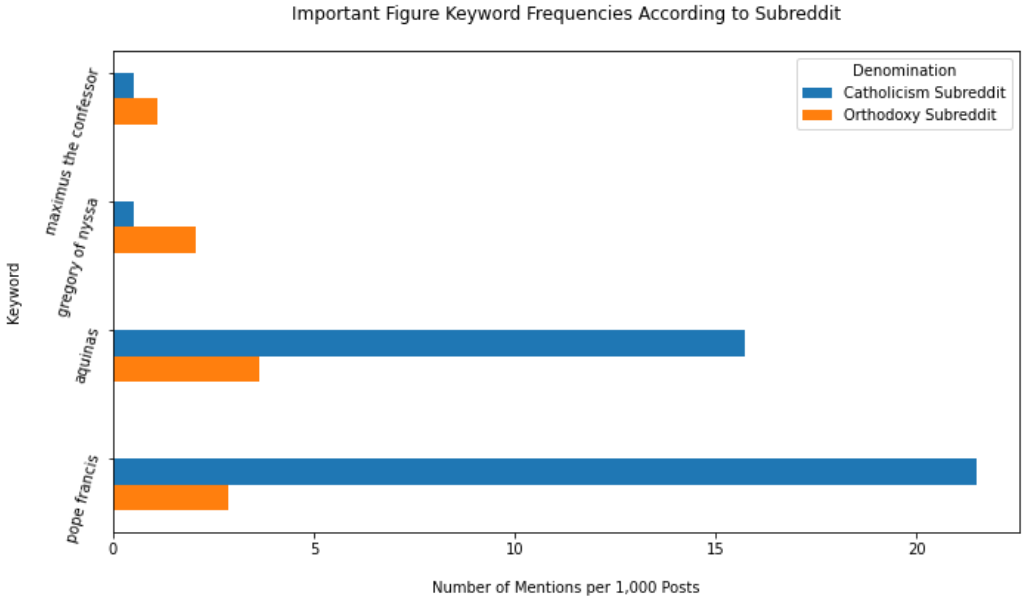
From each subreddit (r/Catholicism and r/OrthodoxChristianity), approximately 10,000 posts were gathered, parsed, and processed in several ways to maximize model interpretability.

Initial exploration of the data revealed distinctions, the most significant being the frequencies of **roots for denominational words ('catholic', 'orthodox', etc.)**



SOME INITIAL FINDS

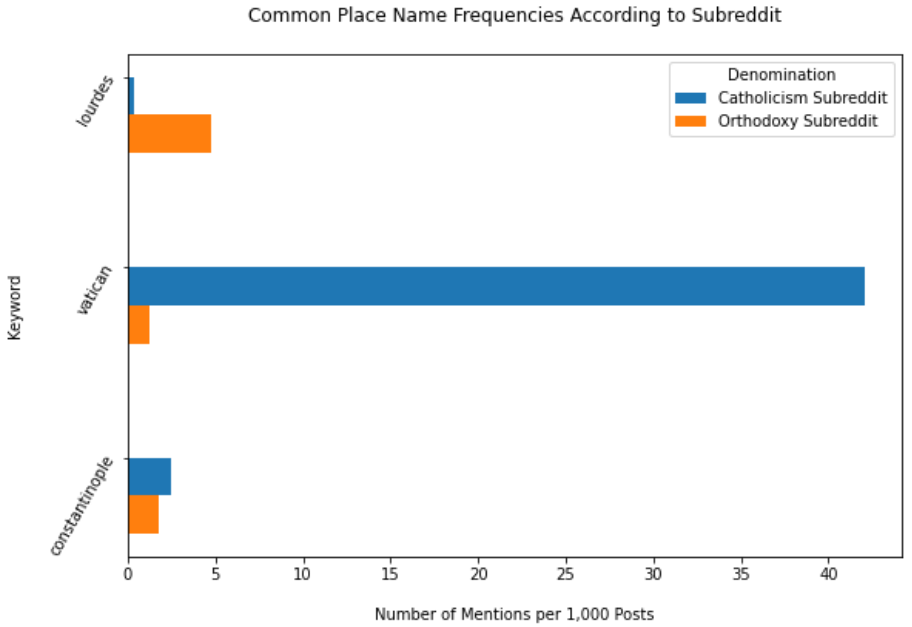
Similarly, **many leaders and other figures** appear in frequencies reflective of each denomination's respective degree of preoccupation with them.



SOME INITIAL FINDS

Similarly, **many leaders and other figures** appear in frequencies reflective of each denomination's respective degree of preoccupation with them.

Important place names also reflect a divergence of focus.

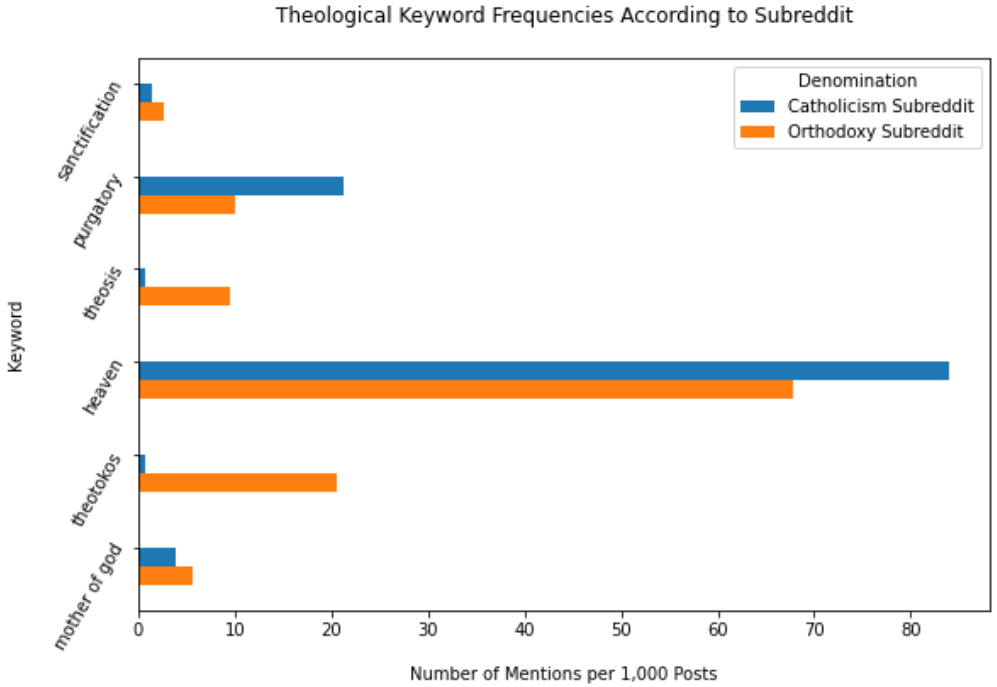


SOME INITIAL FINDS

Similarly, many leaders and other figures appear in frequencies reflective of each denomination's respective degree of preoccupation with them.

Important place names also reflect a divergence of focus.

Some differences in phrasing are more linguistic than conceptual.



THE MODELS

ON THE MODELS USED

Several different models were fit to the data in various forms (**bagging classification**, **AdaBoost**, **random forests**, and **logistic regression**).

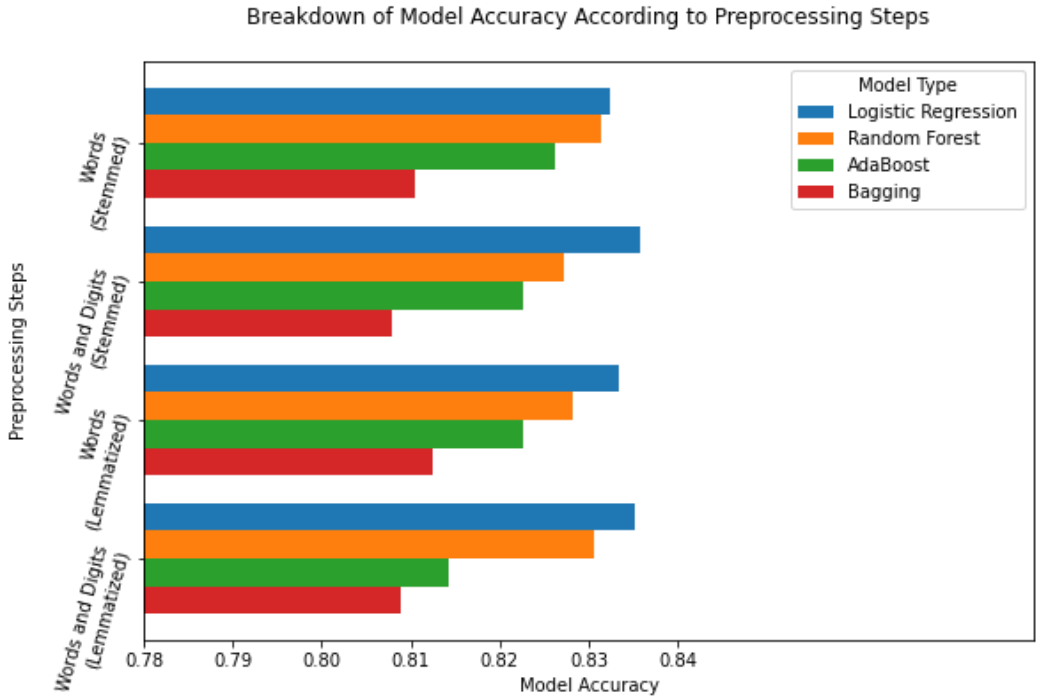
All models consistently performed well above baseline, with random forests and logistic regression performing somewhat better than the others.

ON THE MODELS USED

Ultimately, **logistic regression proved most accurate** across the entire range of text forms, optimizing especially well on stemmed words and digits that had been vectorized according to count frequency.

Random forests were a close second, and future revisions could easily bring them to an even higher degree of accuracy.

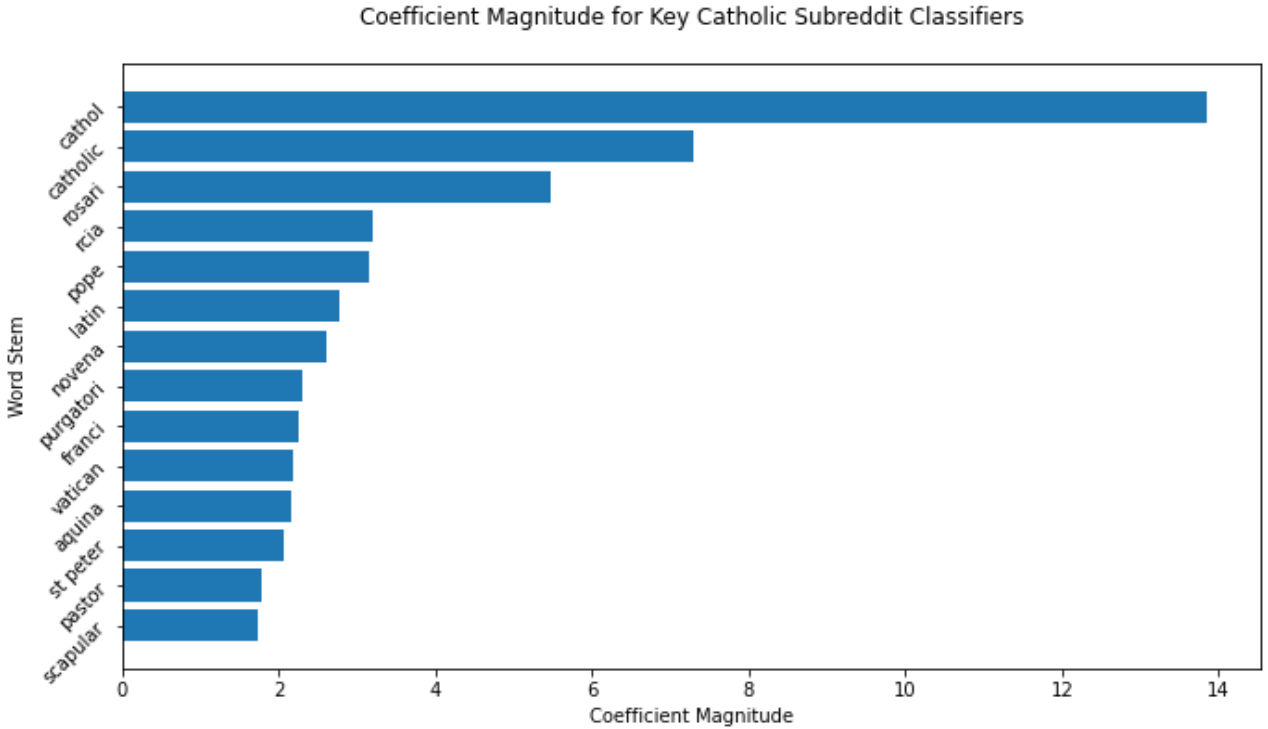
The degree of success in classification reflects somewhat the relationship between the two denominations (i.e. divided, but also joined by a shared history and tradition).



MORE FINDS

Words and word roots prioritized by the logistic regression model closely with what initial exploration of the data foreshadowed.

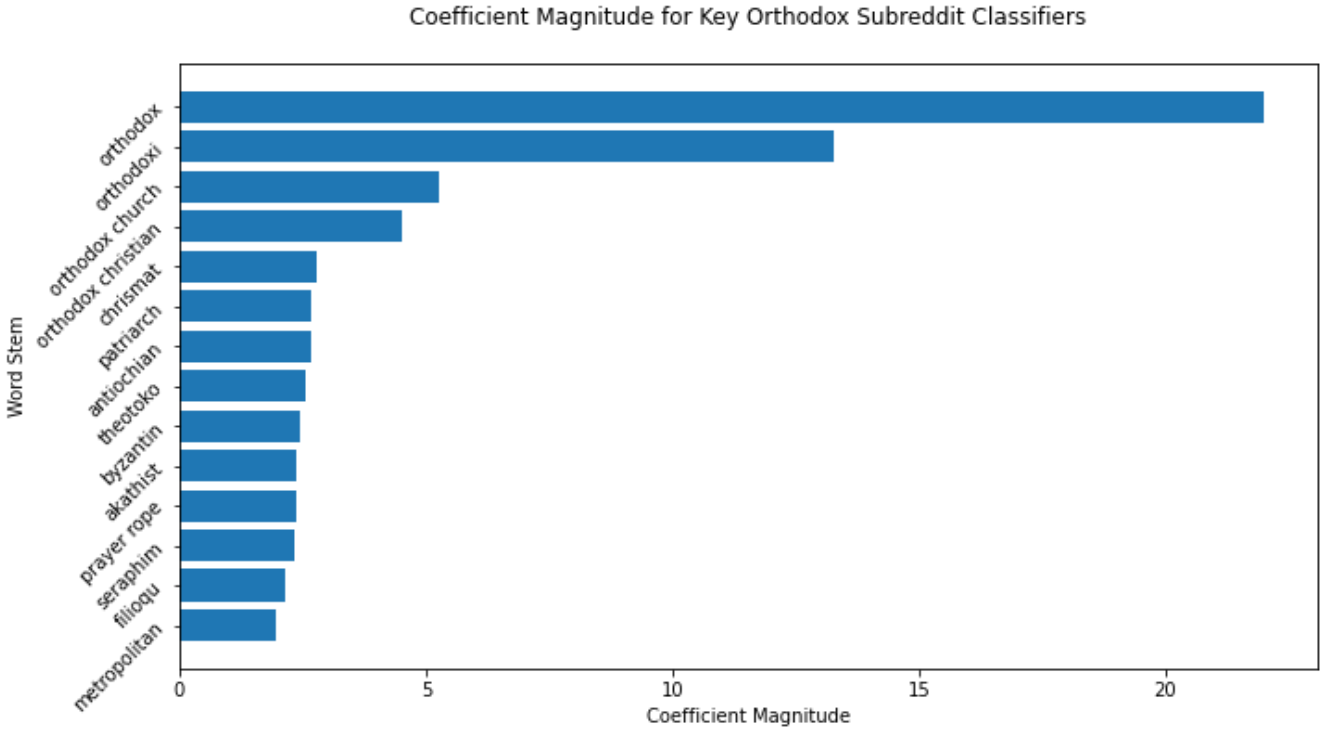
Figures, places, concepts, and linguistic nuances served as the most significant keywords.



MORE FINDS

Words and word roots prioritized by the logistic regression model closely with what initial exploration of the data foreshadowed.

Figures, places, concepts, and linguistic nuances served as the most significant keywords.



IN CLOSING

WRAPPING UP

Predictive modeling can (it seems) classify the source of at least some writings according to authorial denomination accurately, even when denominations are not entirely separate.

When models are optimized for interpretability, the results align largely with what the implications of history and doctrine would anticipate.

More exploration might allow for future classification not just between two closely related denominations, but between subgroups within a single denomination.

The use of white box classifiers might then potentially reveal the nature/tenor/points of contention that are in play and even detect reactionary events, trends, etc.

QUESTIONS OR COMMENTS?