

Ancient Texts NLP & Classification Project

Insights from Balance and Preparation

All code at https://github.com/izsolnay/Ancient_NLP

OVERVIEW

Objective

Develop a reliable machine learning model which can classify ancient texts into genres to reveal insights

Problem

Target variable 'is_hymn' was unbalanced

- 'is_hymn=0' had on average over twice as many words per text (718/266)
- 'is_hymn=1' had almost twice as many texts (250/131)

	count	mean	std	min	25%	50%	75%	max
is_hymn								
0	131.0	718.0	853.0	8.0	150.0	359.0	1052.0	5305.0
1	250.0	266.0	326.0	4.0	108.0	175.0	305.0	2836.0

PROJECT STATUS

Steps taken

- Merged columns from 'all_texts_sentiment.csv' and 'top_word_df.csv' into a new DataFrame
- Introduced new feature 'sentence_count' and binary target variable 'is_hymn'
- Balanced target variable creating 'summary' and 'lemmatized_summaries' features
- Generated visuals

Result: a more balanced data set

	count	mean	std	min	25%	50%	75%	max
is_hymn								
0	154.0	396.0	294.0	37.0	152.0	310.0	630.0	1192.0
1	210.0	257.0	217.0	31.0	114.0	185.0	337.0	1169.0

NEXT STEPS

Move forward with model building

A binomial logistic model (rather than linear) is selected because it is a statistical technique that models the probability y of an event ('is_hymn') based on one (or more) independent variables.

Steps

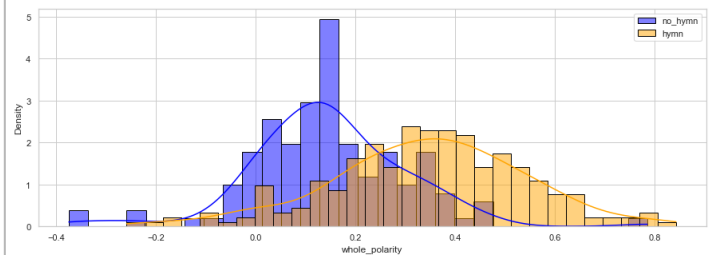
- Consider upscaling 'is_hymn=0' to balance target variable counts
- Check data correlations using VIF and Pearson's
- Build and validate random forest and XGBoost models
- Compare models, test winning model
- Get feature importances
- Create visual evaluations

KEY INSIGHTS

Density Histograms

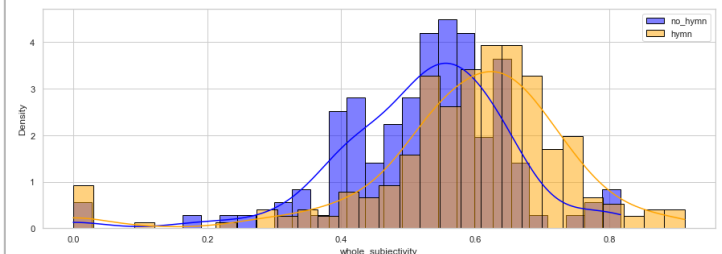
Density histograms provide visual representations of the relationships between 'is_hymn' variable and text polarity and subjectivity, highlighting trends in positivity and truthfulness based on assigned values.

Whole polarity scores after balancing data set



- The curve for 'hymn=0' shows a concentration of lower polarity scores, indicating a tendency towards less positive texts
- The curve for 'hymn=1' exhibits a higher polarity score predominance, suggesting a greater likelihood of more positivity

Whole subjectivity scores after balancing data set



- The curve for 'is_hymn =0' shows a concentration of lower subjectivity scores, indicating a tendency towards more objective (truthful) texts
- The curve for 'is_hymn =1' exhibits a shift towards higher subjectivity scores, reflecting a greater likelihood of subjectivity (less truthfulness)