Classification of Reddit posts

GA DSI Project 3 by Suma Karanam, 13 September 2021

The subreddits

1. **r/languagelearning**

   This is a subreddit for anybody interested in the pursuit of languages. Whether you are just starting, a polyglot or a language nerd, this is the place for you!

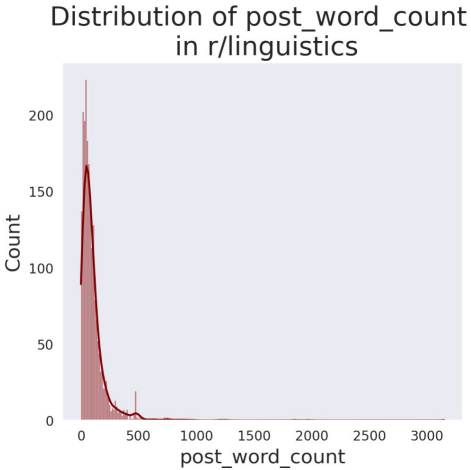
2. **r/linguistics**

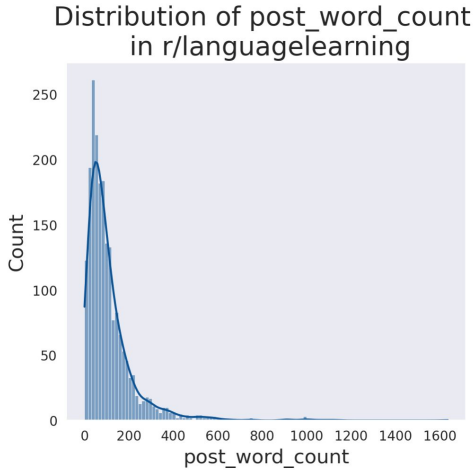
   lin · guis · tics: the scientific study of human *language* * what form does it take? * how is meaning constructed? * how is it structured? * how is it produced?

Data

- Titles, authors, number of comments and content of 2000 posts (posted before 1 Sep 2021) from each subreddit
- Did not include any posts that were removed
- Did not include posts that had no content besides the title

Data



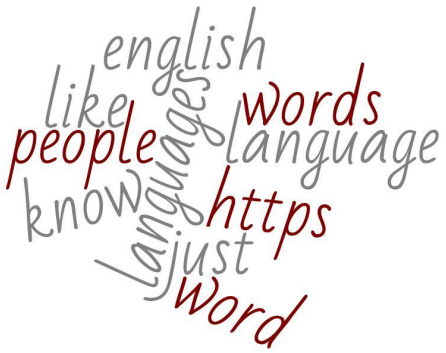


- Distributions of numerical features are very similar for both the subreddits
- Current analysis focuses on distinguishing the subreddits by using the text

Baseline model

- For each subreddit, a list of frequent words that occur in it but not frequently in the other is constructed
- Model 1: If a post contains at least one word from this list for r/languagelearning, it is classified to be from r/languagelearning, else r/linguistics
- Model 2: If a post contains at least one word from this list for r/linguistics, it is classified to be from r/linguistics, else r/languagelearning

Baseline model



/r/linguistics

/r/languagelearning

Top ten words

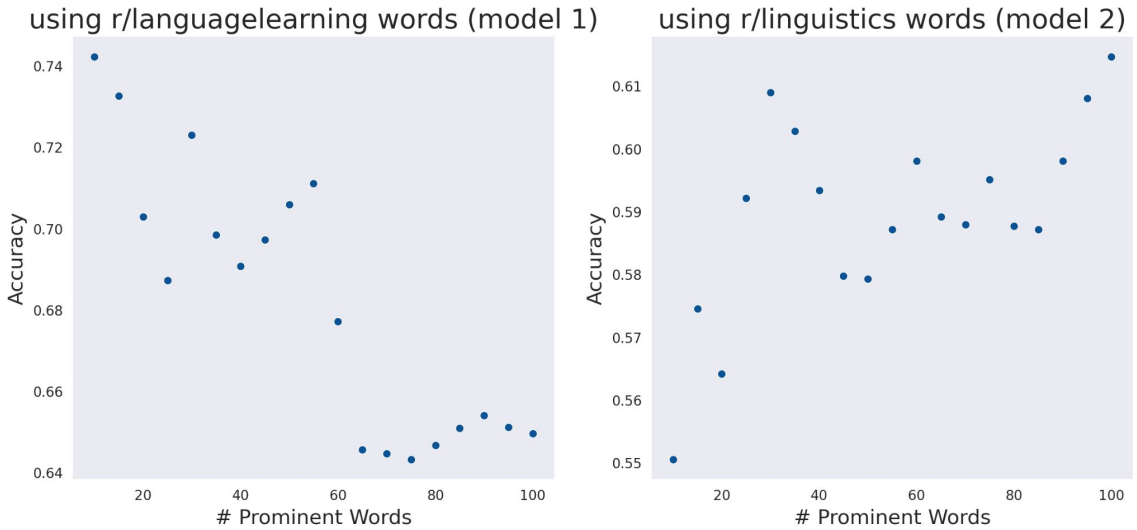
Baseline model

/r/linguistics

/r/languagelearning

Sample of distinct words in the top fifty words

Baseline model



- The baseline models were iterated for various lengths of the frequent words list
- Best Accuracies of these models were observed to be as follows:
  - Model 1: 0.74
  - Model 2: 0.61

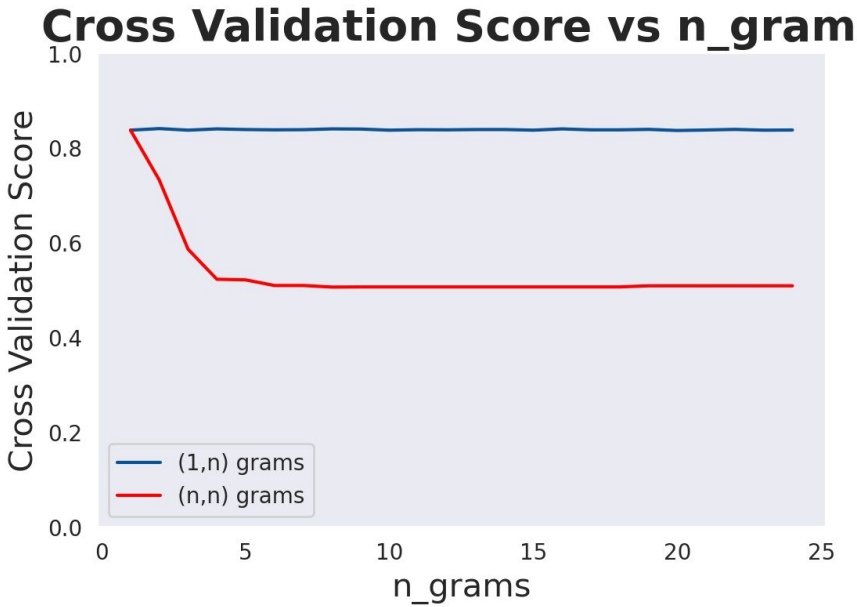
Classification models - performance

		Overall Accuracy		Accuracy of r/languagelearning		Accuracy of r/linguistics	
		Training data	Test Data	Training Data	Test Data	Training Data	Test Data
Random Forest	Title	0.79	0.72	0.69	0.62	0.88	0.83
	Content	0.84	0.8	0.76	0.73	0.93	0.88
	Content and title	0.86	0.82	0.78	0.74	0.94	0.9
Gradient Boosting	Title	0.79	0.73	0.68	0.61	0.9	0.85
	Content	0.86	0.81	0.79	0.73	0.93	0.89
	Content and title	0.88	0.83	0.82	0.77	0.95	0.89
Voting Classifier	Title	0.8	0.74	0.69	0.61	0.9	0.86
	Content	0.87	0.82	0.8	0.75	0.94	0.88
	Content and title	0.89	0.83	0.82	0.77	0.95	0.89

Guess which subreddit this is from

About different number of words in languages
I was wondering how many more words are in the romance languages as Spanish than in English.\n\nThis doubt comes to me because this languages have a lot of words if you consider all conjugations, gender, etc.\n\nI think/guess this is not reflected when associations make a total count of words in their language and I cannot find the real count of ALL words (in Spanish, at least).\n\nAny thoughts?

Further Analysis of Gradient boosting



- 1, 2 and 3 grams have a strong impact on model performance

Conclusions

- ➔ The best performing model (a voting classifier) in the analysis led to a lift of about 9% in accuracy (from 74% to 83%) from the baseline model
- ➔ All the classification algorithms had a better accuracy in r/linguistics compared to r/languagelearning
- ➔ Using the text from both the title and content of the post resulted in better performance of all the models

TODO

- ➔ Include numerical features - length of post, character count, number of comments etc in the analysis
- ➔ Preprocess the text - stemming, lemmatizing, augmenting the stopwords
- ➔ Try other kinds of vectorization of the text (TfidfVectorizer)
- ➔ Analyze larger samples of data
- ➔ More hyperparameter tuning
- ➔ Use other classification algorithms