Authorship Attribution Leveraging Idiolect Analysis

Omkar Yadav, Michael Jiang, Hua Shen, Kenan Alkiek, David Jurgens

BAB LAB LAB

UNIVERSITY OF MICHIGAN

Research Objective

Can we encode an author's unique idiolect purely from linguistic features?

- Compute Biber linguistic features on each text sample to represent an author's style
- Determine if learning an author's idiolect impacts authorship attribution

Background

- Style is crucial in interpreting a message from a given text
- Methods of representing style have proven to be impactful in areas such as forensic linguistics
- We leverage Biber linguistic features as a representation of an author's style

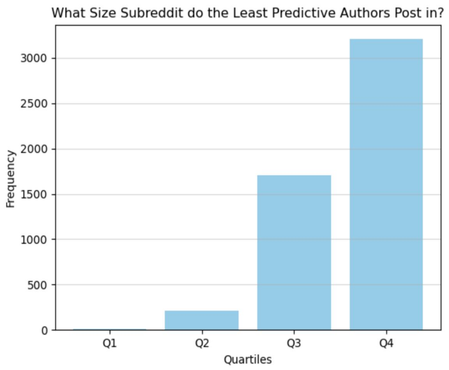
Dataset Creation

- Downloaded ~ 1.4 TB of compressed reddit posts including relevant metadata
- Filtered out posts to have a minimum of 250 words and ensured each author had at least 10 posts
- Computed linguistic features on each post and aggregated on author

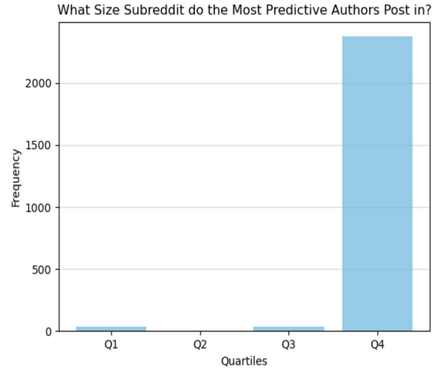
Next Steps

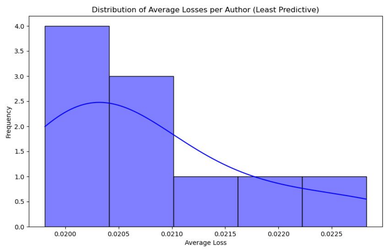
- Preprocess more datasets such as wiki and books
- Determine which settings an author is more likely to write in their own unique style
- Train a classifier to determine whether a given piece of text is indicative of an author's style
- Apply method to authorship attribution

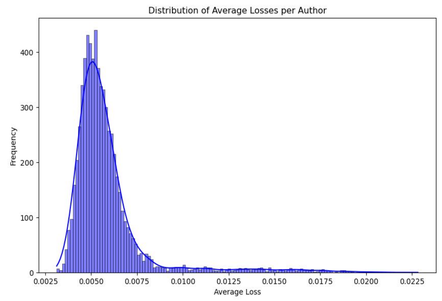
Results/Takeaways

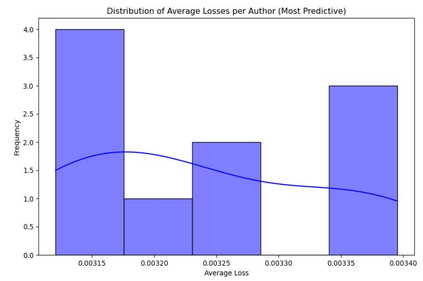


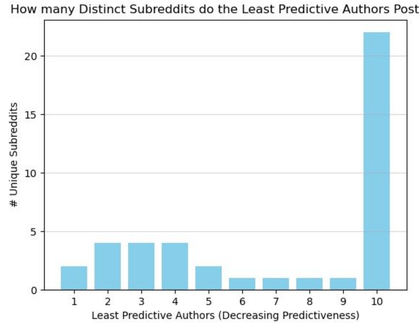
- More predictive authors seem to exclusively come from larger subreddits where less predictive authors have more variance in subreddit size



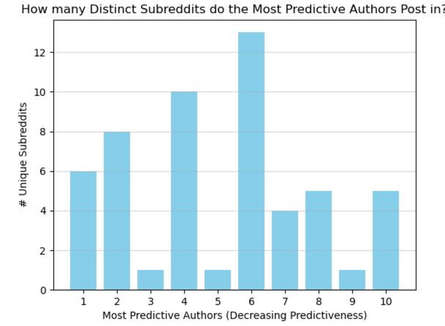








- Generally the least predictive authors post in a more concentrated number of subreddits compared to more predictive authors



Data Example

Post: I hate to sound like typical reddit but I don't care, dump your boyfriend. You've known him for 8 months whereas the nanny had been there since before your husband died. She was a second mother to them since you (understandably) were busy taking care of the business he left behind…
Biber Encoding: [0.5, 0.0, 1.0, 1.0, 0.5, 2.75, 2.0, 4.0, 2.0, 0.8291561976…]