Author Attribution with Linear SVC

Hudson Strauss, Helen Wang, JoJo Torres
Team 6

Motivation and Background

- Author classification is useful in forensic science
- Real world application of what we were learning in class.
- Enron Corpus widely used for email analysis
- Prior research used to detect spam emails.

Challenges:

- Pre-processing the data and filtering the emails
- Choosing our models
- Optimizing parameters
- Training speed

Overarching Goal

To train a classification model that can accurately detect who wrote an email based on its contents.

Project Pipeline

Preprocessing

- Go user by user and extract the emails in the sent_items folder.
- Clean the text from commas and special characters.
- Lemmatized and Stemming
- Tokenize the text
- TF-IDF vectorization on the text
- Split into test and training

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

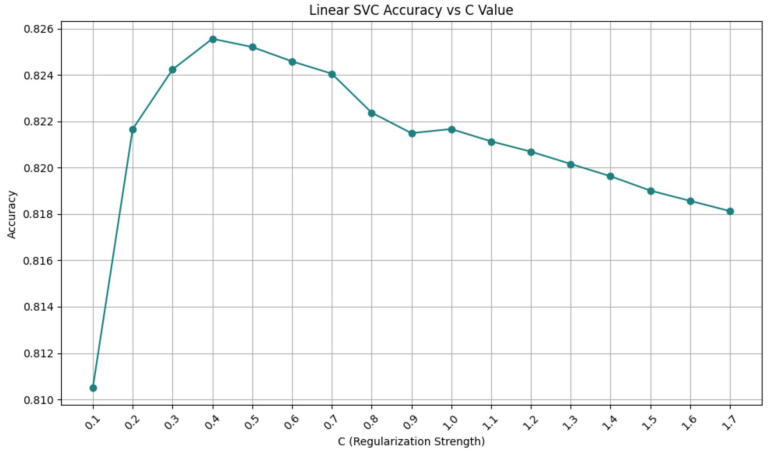
$$idf(t,D) = log \frac{N}{|\{d \in D : t \in d\}|}$$

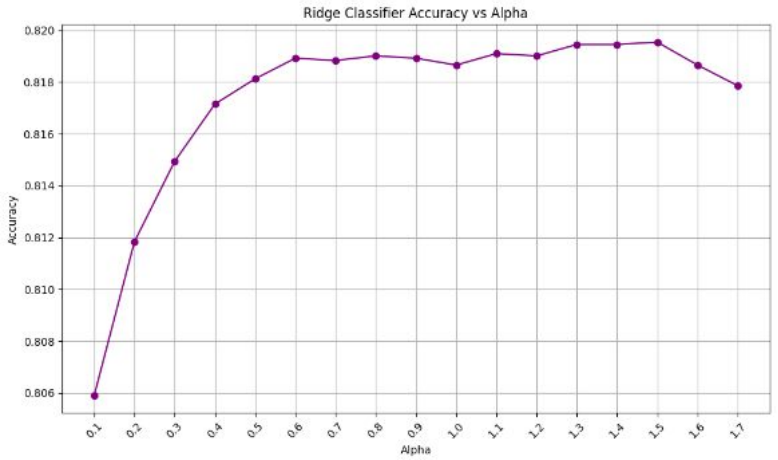
Attempted Models

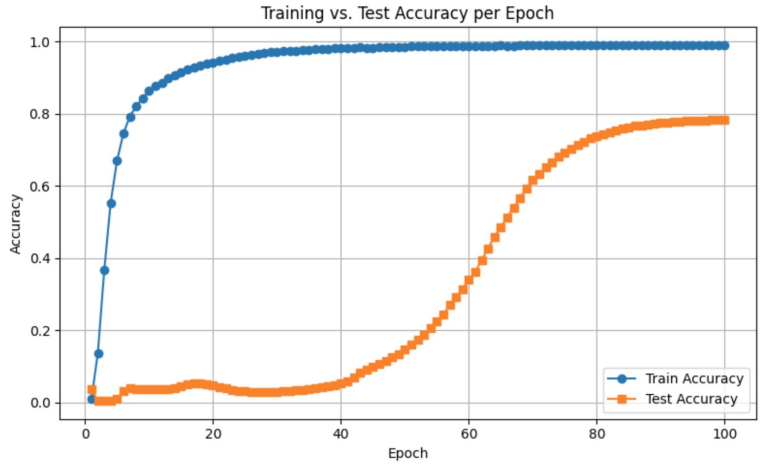
- Linear SVC
- Multinomial NB
- Ridge Classifier
- SGD Classifier
- Passive Aggressive
- Decision Trees
- KNN (Cosine)
- Linear SVC
- Neural Network

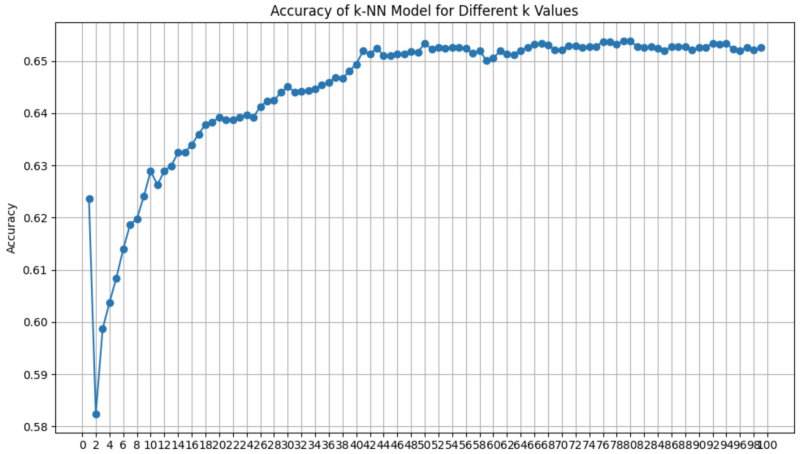
Evaluation

- F1 score
- Accuracy
- Testing different parameters







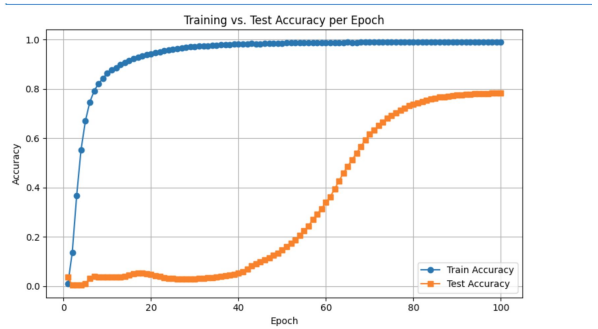


Optimization

Linear SVC
C=0.1 → Accuracy: 0.8095
C=0.2 → Accuracy: 0.8210
C=0.3 → Accuracy: 0.8233
C=0.4 → Accuracy: 0.8245
C=0.5 → Accuracy: 0.8241
C=0.6 → Accuracy: 0.8236
C=0.7 → Accuracy: 0.8226
C=0.8 → Accuracy: 0.8210
C=0.9 → Accuracy: 0.8207
C=1.0 → Accuracy: 0.8204
C=1.1 → Accuracy: 0.8202
C=1.2 → Accuracy: 0.8195
C=1.3 → Accuracy: 0.8190
C=1.4 → Accuracy: 0.8181
C=1.5 → Accuracy: 0.8171
C=1.6 → Accuracy: 0.8165
C=1.7 → Accuracy: 0.8158

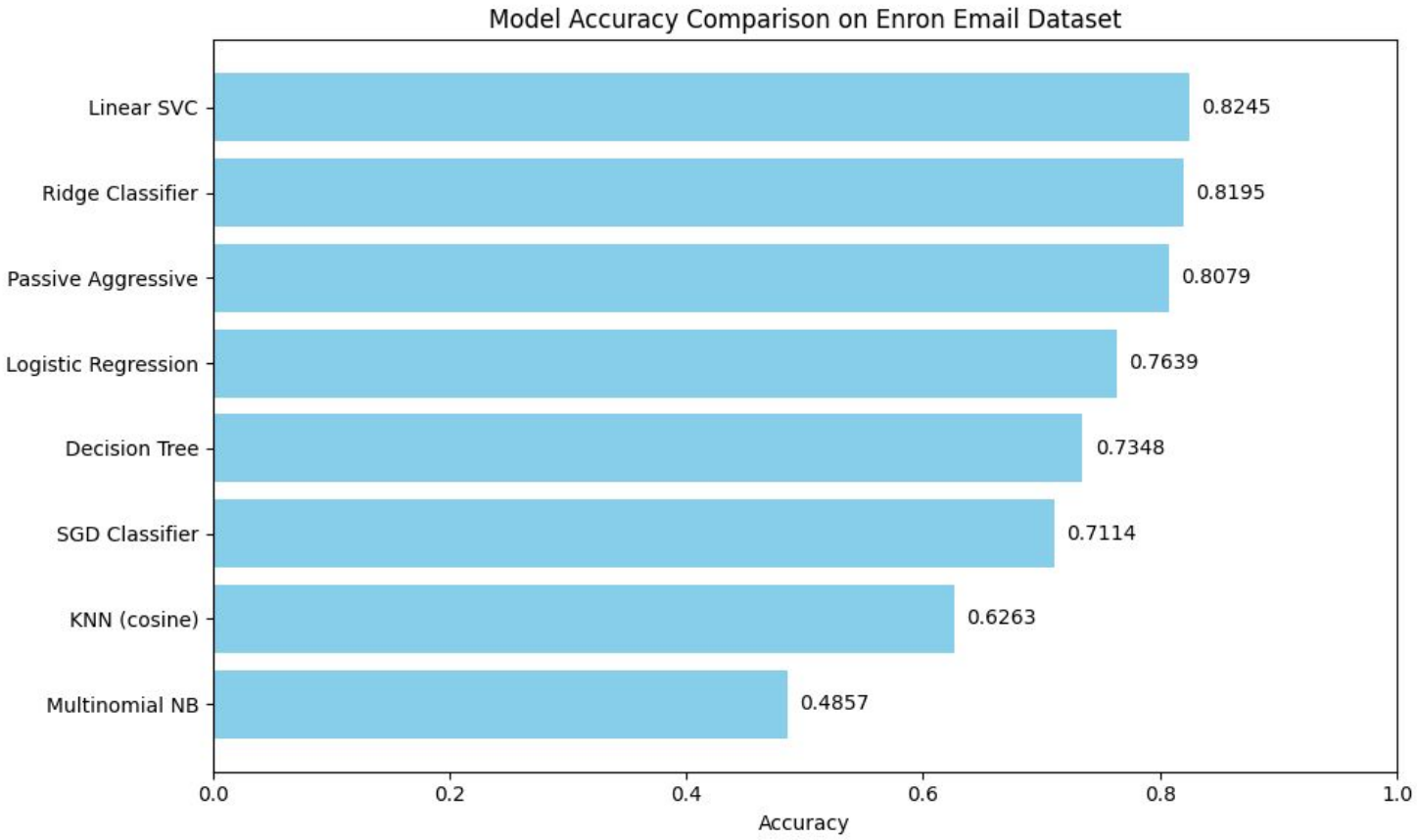
Ridge Classification
Alpha: 0.10 → Accuracy: 0.8059
Alpha: 0.20 → Accuracy: 0.8118
Alpha: 0.30 → Accuracy: 0.8149
Alpha: 0.40 → Accuracy: 0.8172
Alpha: 0.50 → Accuracy: 0.8181
Alpha: 0.60 → Accuracy: 0.8189
Alpha: 0.70 → Accuracy: 0.8188
Alpha: 0.80 → Accuracy: 0.8190
Alpha: 0.90 → Accuracy: 0.8189
Alpha: 1.00 → Accuracy: 0.8187
Alpha: 1.10 → Accuracy: 0.8191
Alpha: 1.20 → Accuracy: 0.8190
Alpha: 1.30 → Accuracy: 0.8195
Alpha: 1.40 → Accuracy: 0.8195
Alpha: 1.50 → Accuracy: 0.8195
Alpha: 1.60 → Accuracy: 0.8187
Alpha: 1.70 → Accuracy: 0.8179

```
vect = TfidfVectorizer(
    stop_words='english',
    ngram_range=(1, 2),
    max_df=0.9,
    min_df=5,
    max_features=20000
)
```



KNN
k nearest neighbors k=20→0.290999203
k nearest neighbors k=21→0.390477033
k nearest neighbors k=220.492255952
k nearest neighbors k=23→0.535622621
k nearest neighbors k=24→0.560580582
k nearest neighbors k=25→0.57713072
k nearest neighbors k=26→0.586423577
k nearest neighbors k=27→0.59323834
k nearest neighbors k=28→0.599699088
k nearest neighbors k=29→0.600495619

Comparisons



Constraints

- We are only testing our model with one dataset.
- We are using traditional DS/ML methods which is limited.
- We truncated the data-set for simplicity.

Ethical considerations

- Potential misuse for surveillance.
- Could be used in courts as evidence
- Although Enron's email database is public, this does infringe on the privacy of the users.

Future Development

- Improve Data Preprocessing
- Larger/smaller dataset collection.
- Better feature selection/dimensionality reduction
- Testing newer methods of vectorization
- Testing with deeper neural networks

Thank you for listening!

References

Iqbal, K., & Khan, M. S. (2022). *Email classification analysis using machine learning techniques*. *Applied Computing and Informatics*, 630–635. https://doi.org/10.1108/ACI-01-2022-0012 (ijisae.org)

[dev25a] scikit-learn developers. Linear models. https://scikit-learn.org/stable/modules/linear_model.html, 2025. Accessed: 2025-06-11.

[dev25b] scikit-learn developers. sklearn.feature extraction.text.tfidfvectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, 2025. Accessed: 2025-06-11.

[dev25c] scikit-learn developers. sklearn.linear model.ridgeclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html, 2025. Accessed: 2025-06-11.

[KY04] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. Machine Learning: ECML 2004. Lecture Notes in Computer Science, 3201:217–226,

[sci25a] scikit-learn Developers. sklearn.svm.linearsvc — linear support vector classification. Online documentation, https://scikit-learn.org/stable/modules/generated/sklearn.