Drexel University

Automatic Essay Scoring (AES) System Using NLP and Deep Learning

DSCI691-001
Luke Chesley
Lauren Miller
Caleb Miller
Hashim Afzal

Automatic Essay Scoring (AES) System
Using NLP and Deep Learning

Overview

Data and Task

Task: Automated Essay Scoring (AES)

Data: Our data is comprised of 12,978 essays written by students in grades 7 - 10.

These essays cover a range of topics and are graded on different scales depending on the topic (set).

Essay Set	Min Score	Max Score
1	1.0	6.0
2	1.0	6.0
3	0.0	3.0
4	0.0	3.0
5	0.0	4.0
6	0.0	4.0
7	2.0	24.0
8	10.0	60.0

Data and Task

The primary task associated with AES is to develop models that can predict essay scores based on various linguistic features and patterns.

Challenges to consider:

Human graders introduce variability in scoring - different experiences, subjectivity, and personal biases

Tendency by graders to round up

Automated systems offer faster grading but require substantial computational resources and time investment.

The essays were masked for privacy, removing personally identifying information. Substantially reducing vocabulary size which impacts the model complexity and smoothing manual features.

Preprocessing

Score Scaling

Target variable was on different scales for each essay set

To make this data work for our classification model, standardizing was necessary

The grades were normalized 0-4 to represent A, B, C, D, and F.

Summary statistics were useful for determining scaling

Distributation of Scores

Grade	Percentage
A	0.14
B	0.36
C	0.34
D	0.14
F	0.02

Evaluation Method

Quadratic Weighted Kappa (QWK) is a measure that quantifies the agreement between two raters.

The two raters are the ground truth of the test set and predictions

The score ranges from -1 to 1, where 1 is perfect agreement, 0 indicates no better than random, and -1 is perfect disagreement.(5)

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Handcrafted Features

Explanation:

The following lexical features were extracted to capture various aspects of the essays' linguistic characteristics and structure.

Creating these features gives additional context for prediction. Combining handcrafted features with learned word embeddings is a technique commonly found in AES.

Handcrafted Features

Feature Name	Description
Number of Correct Words	Counts the number of correctly spelled words in each essay.
Number of Nouns	Counts the number of nouns in each essay.
Number of Adjectives	Counts the number of adjectives in each essay.
Average Number of Characters per Word	Calculates the average number of characters per word in each essay.
Average Number of Characters per Sentence	Calculates the average number of characters per sentence in each essay.

Average Number of Punctuation Marks per Word	Calculates the average number of punctuation marks per word in each essay.
Average Number of Punctuation Marks per Sentence	Calculates the average number of punctuation marks per sentence in each essay.
Average Number of Words per Sentence	Calculates the average number of words per sentence in each essay.
Average Number of Unique Words per Essay	Calculates the average number of unique words per essay.

Modeling:
Overview

Naive

Handcrafed features -> SVM

TF-IDF -> SVM

Llama2 + random embeddings

Llama2 + bert embeddings

Llama2 + handcrafted features

Llama2 + bert embeddings + handcrafted features

Hypothesis: Increasing complexity in model will result in more accurate scoring prediction results

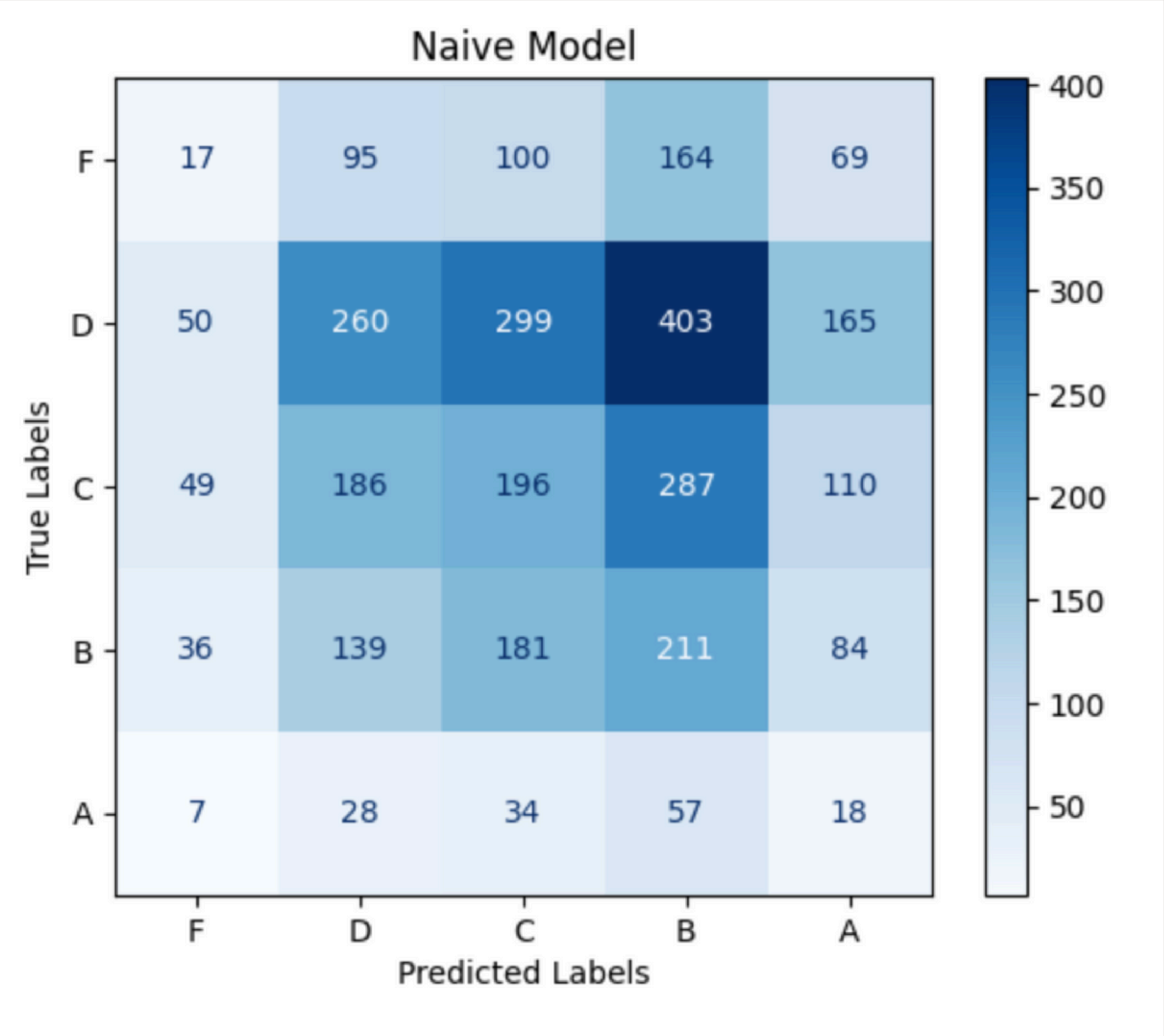
Modeling:
Naive Model

We split the dataset into training and testing sets

Then, we randomly predicted class labels on the test set based on the distribution of labels in the training set

This results in a QWK very close to 0 - in line with expectations.

Modeling:
Naive Model



Classification report:

```
              precision    recall  f1-score   support

           0       0.11      0.04      0.06       445
           1       0.37      0.22      0.28      1177
           2       0.24      0.24      0.24       828
           3       0.19      0.32      0.24       651
           4       0.04      0.12      0.06       144

    accuracy                           0.22      3245
   macro avg       0.19      0.19      0.17      3245
weighted avg       0.25      0.22      0.22      3245


QWK: 0
```

Modeling:
Handcrafted Features - SVC

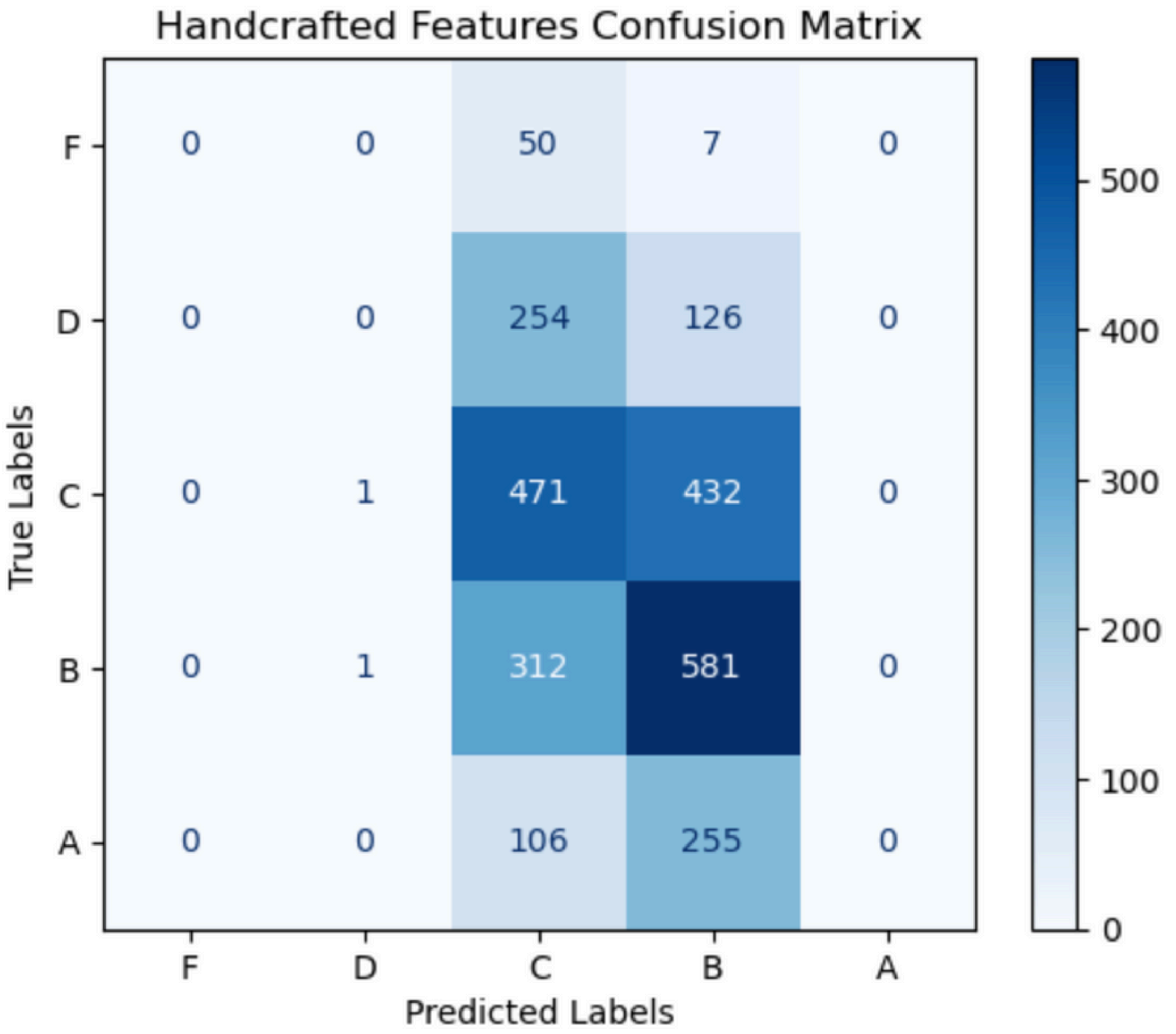
We trained an SVC model with handcrafted features to predict essay scores.

This gave insight of how the handcrafted features are in essay score prediction.

This model produced a QWK of 0.219

Modeling:
Handcrafted Features - SVC

Classification Report:



	precision	recall	f1-score	support
f	0.00	0.00	0.00	57
d	0.00	0.00	0.00	380
c	0.39	0.52	0.45	904
b	0.41	0.65	0.51	894
a	0.00	0.00	0.00	361
accuracy			0.41	2596
macro avg	0.16	0.23	0.19	2596
weighted avg	0.28	0.41	0.33	2596

qwk: 0.219

Modeling:
TF-IDF

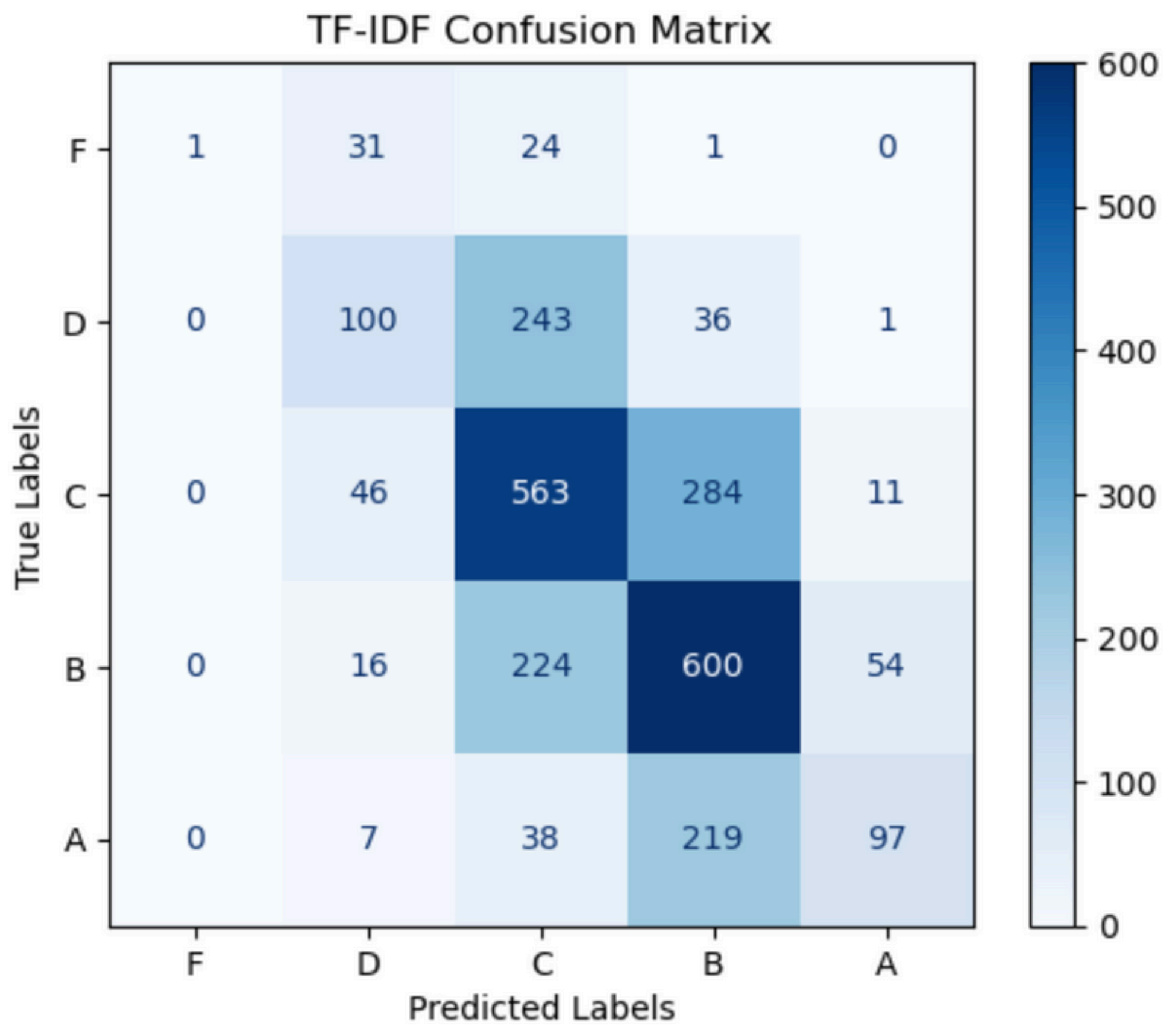
Term Frequency Inverse Document Frequency (TF-IDF) was used on the essay text to predict scores.

We trained a Support Vector Classifier (SVC) with a linear kernel on the vectorized training data and generated predictions.

This model produced moderate results with a QWK of 0.563

Modeling:
TF-IDF

Confusion Matrix:



Classification Report:

```
              precision    recall  f1-score   support

           f       1.00      0.02      0.03        57
           d       0.50      0.26      0.34       380
           c       0.52      0.62      0.56       904
           b       0.53      0.67      0.59       894
           a       0.60      0.27      0.37       361

    accuracy                           0.52      2596
   macro avg       0.63      0.37      0.38      2596
weighted avg       0.54      0.52      0.50      2596

QWK: 0.563
```

Modeling:
Deep Learning Methods
Llama 2 Model

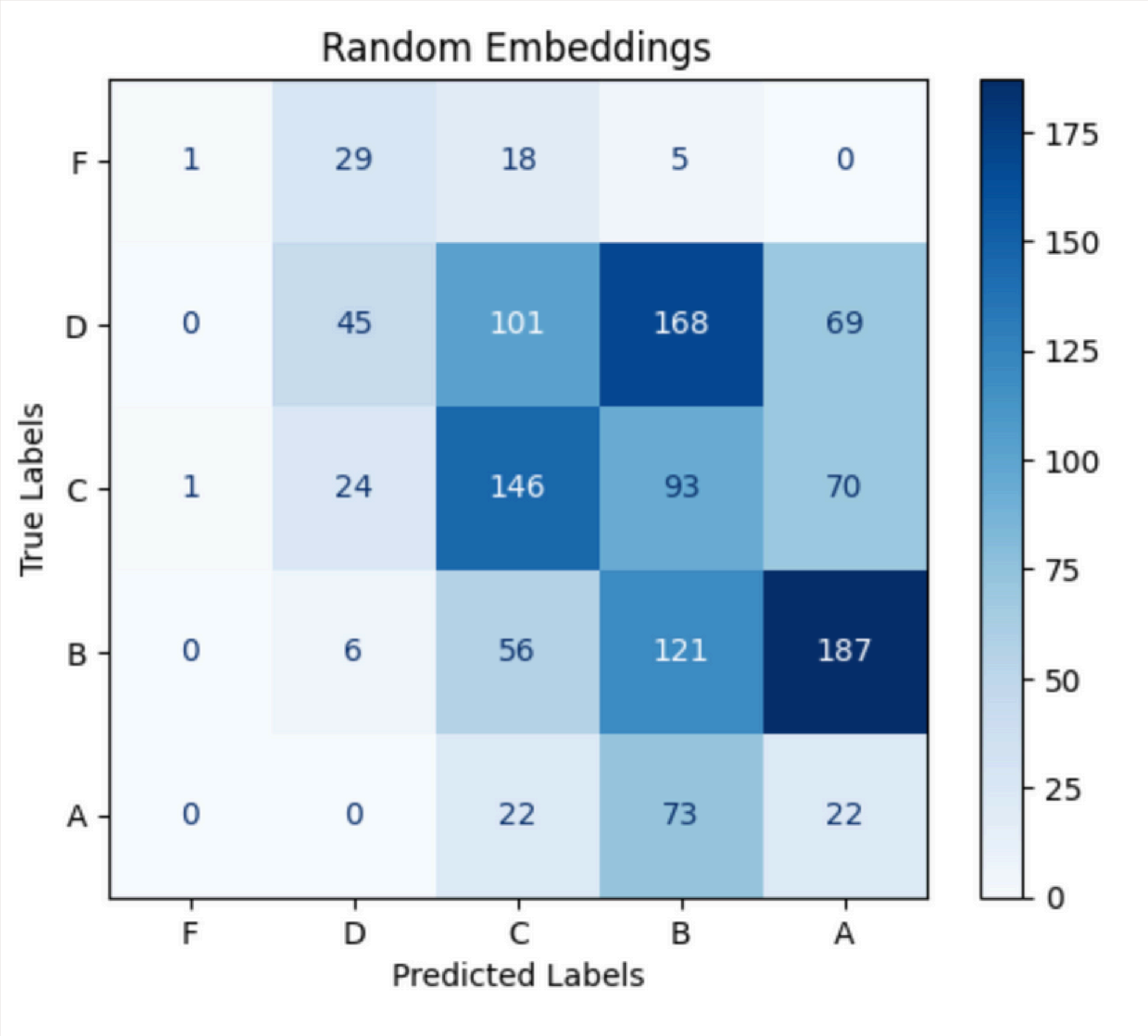
For the deep learning approach we trained a modified Llama 2 model from scratch.

The main modification was reducing the hidden dimension from 4096 to 768.

An average pooling layer and a linear layer were added to the end of the base model to adapt it for the sequence classification task.

This model used randomly initialized token embeddings and produced a QWK of 0.26 - slight improvement from handcrafted features

Modeling:
Deep Learning Methods
Llama 2 Model



	precision	recall	f1-score	support
f	0.50	0.02	0.04	53
d	0.43	0.12	0.18	383
c	0.43	0.44	0.43	334
b	0.26	0.33	0.29	370
a	0.06	0.19	0.09	117
accuracy			0.27	1257
macro avg	0.34	0.22	0.21	1257
weighted avg	0.35	0.27	0.27	1257

QWK: 0.26

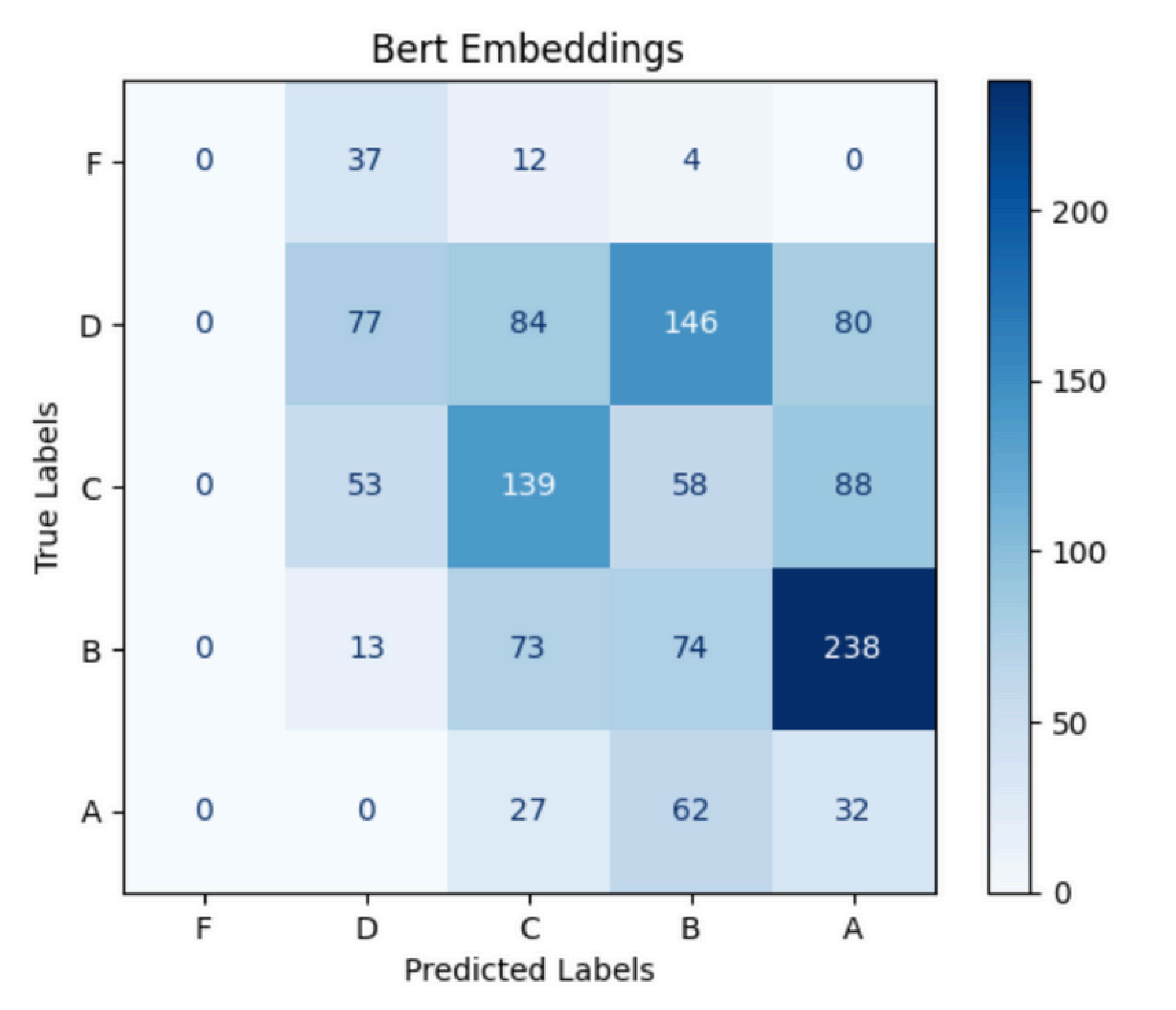
Modeling:
Deep Learning Methods
Llama 2 Model with Bert Embeddings

We trained a model with frozen pretrained embeddings, created with a bert tokenizer and model.

These embeddings were then passed to the same model, bypassing the embedding layer.

The results for this were mixed, the QWK was 0.28 but a lower test accuracy was 0.25.

Modeling:
Deep Learning Methods
Llama 2 Model with Bert Embeddings



	precision	recall	f1-score	support
f	0.00	0.00	0.00	53
d	0.43	0.20	0.27	387
c	0.41	0.41	0.41	338
b	0.22	0.19	0.20	398
a	0.07	0.26	0.11	121
accuracy			0.25	1297
macro avg	0.23	0.21	0.20	1297
weighted avg	0.31	0.25	0.26	1297

QWK: 0.28

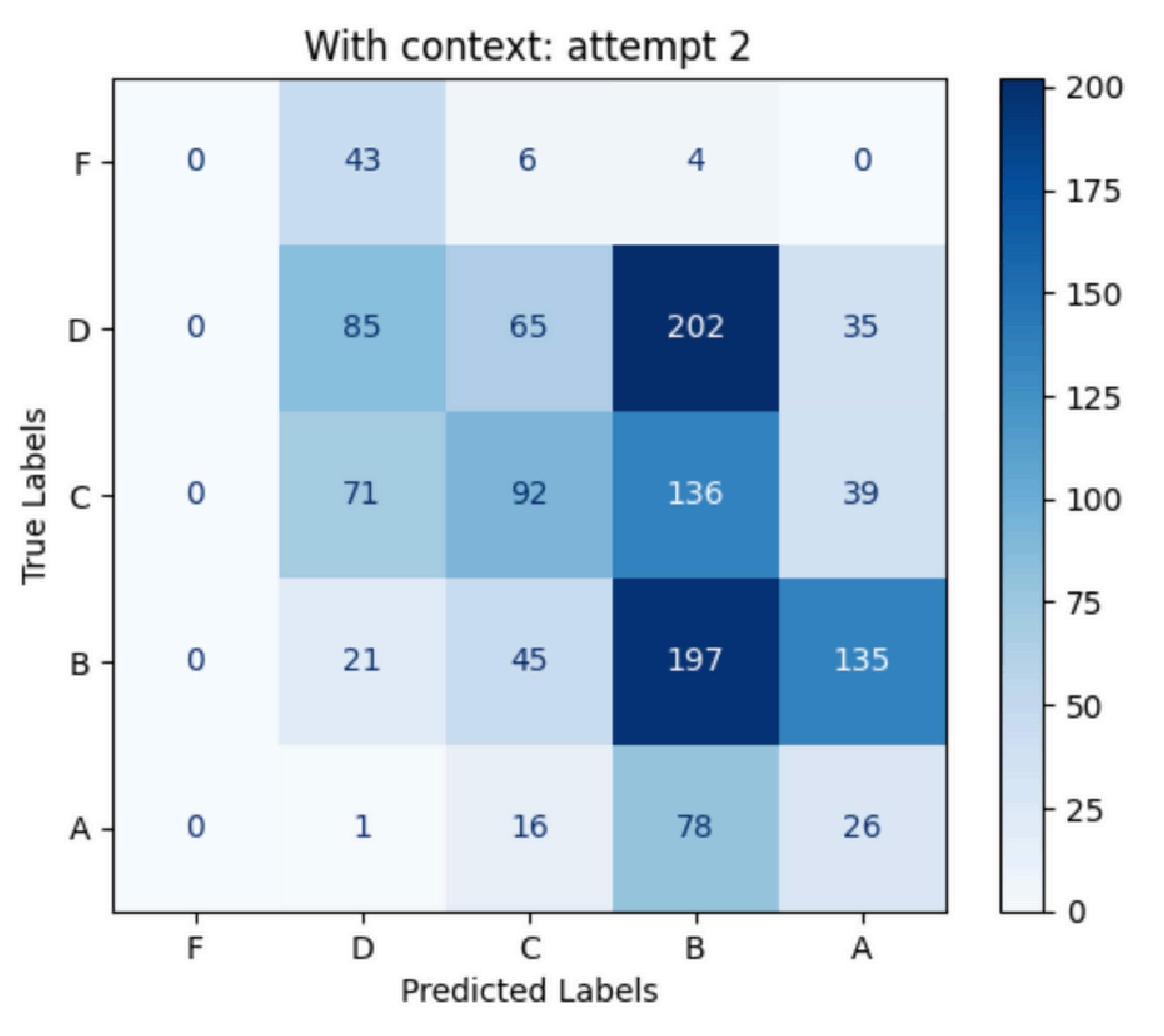
Modeling:
Deep Learning Methods
Llama 2 Model with Manual Features

Our Llama 2 Model was fed manual features.

The features were normalized, mapped to the hidden dimension with a linear layer, then added to the pre-attention word embeddings.

The results slightly improved with a QWK of 0.324

Modeling:
Deep Learning Methods
Llama 2 Model with Manual Features



	precision	recall	f1-score	support
f	0.00	0.00	0.00	53
d	0.38	0.22	0.28	387
c	0.41	0.27	0.33	338
b	0.32	0.49	0.39	398
a	0.11	0.21	0.15	121
accuracy			0.31	1297
macro avg	0.25	0.24	0.23	1297
weighted avg	0.33	0.31	0.30	1297

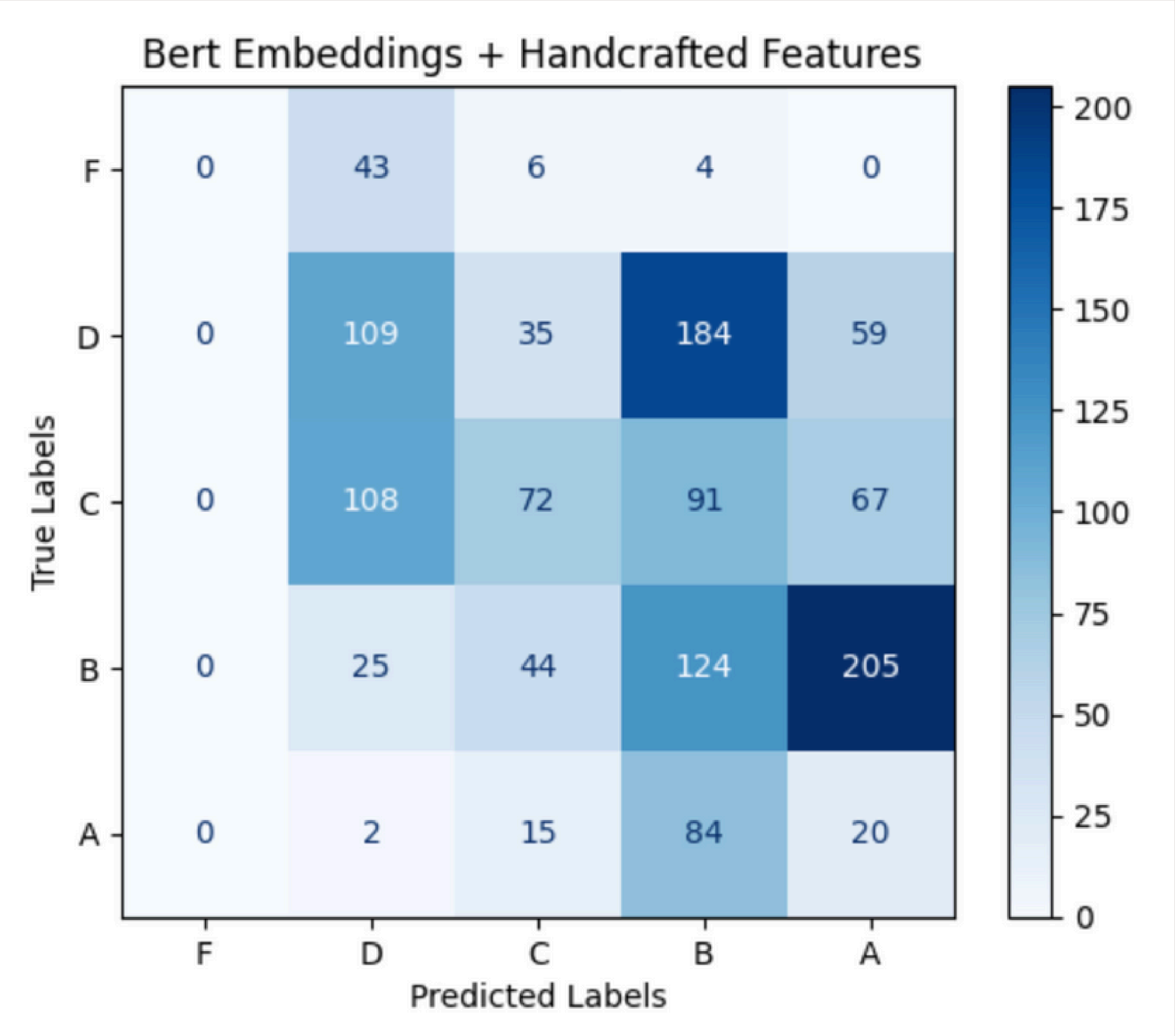
QWK : 0.324

Modeling:
Deep Learning Methods
Llama 2 Model with Manual Features and Bert Embeddings

To tie it all together, we fed our Llama 2 Model manual features and the Bert pretrained embeddings

This model produced a QWK of 0.31

Modeling:
Deep Learning Methods
Llama 2 Model with Manual Features and Bert Embeddings



	precision	recall	f1-score	support
f	0.00	0.00	0.00	53
d	0.38	0.28	0.32	387
c	0.42	0.21	0.28	338
b	0.25	0.31	0.28	398
a	0.06	0.17	0.08	121
accuracy			0.25	1297
macro avg	0.22	0.19	0.19	1297
weighted avg	0.31	0.25	0.26	1297

qwk: 0.310

Modeling:
Summary of Results

MODEL	QWK
Naive	0
Handcrafted-Features -> SVM	0.219
TF-IDF -> SVM	0.563
llama2 random embeddings	0.26
llama2 bert embeddings	0.28
llama2 + handcrafted features	0.324
llama2 + bert embeddings + handcrafted features	0.310

Results

The only model able to achieve ‘moderate’ strength was the TF-IDF vectorized model

The rest of the model fall in the ‘fair’ category

Results from deep methods are mixed

We saw slight improvement with complexity until the Bert embeddings + handcrafted features model

Results

All four of the deep models we trained showed some similar patterns in the distribution of the scores and seemed to have similar challenges.

Differentiating between A's and B's was very difficult - predicting ‘A’ when the true label was a ‘B’ was one of the most common mistakes.

There was a reluctance to predict ‘F’

No model was able to confidently differentiate between B’s, C’s, and D’s.

Conclusion

Although we were not able to achieve great results with the deep learning methods, we can still draw some valuable conclusions.

One of the most surprising findings was the strength of the TF-IDF model. This representation of the essays was robust enough that the SVC model was able to out preform all deep models.

Conclusion

Our exploration of manipulating hand crafted features included feeding raw values (number of nouns, etc.) then normalizing them by column.

Normalizing offers a more relative representation of each feature and performed better in testing.

These models could be optimized with hyperparameter fine tuning, using trainable pretrained embeddings, and using one hot encoding rather than raw values for hand crafted features.

Conclusion

Further Development/Refinement

Future work includes defining a more robust and fine-tuned scoring system.

This would entail scoring each essay in different categories, which are weighted and combined to calculate a final score.

Our Team

Luke Chesley

Caleb Miller

Hashim Afzal

Lauren Miller

References

(

(1) França, P. M., Marques, H. R., Qian, C., Monebhurrun, V., Debbah, M., & Gross, J. (2023). Reconfigurable intelligent surfaces: Bridging the gap between scattering and reflection. arXiv. https://arxiv.org/pdf/2310.05191

(2) Glymour, C. (1980). Theory and Evidence. The Journal of Philosophy, 77(11), 646-664. Retrieved May 22, 2024, from https://www.jstor.org/stable/2529310?origin=crossref

(3) Kaggle. (2023). Feedback Prize - English Language Learning. Retrieved May 22, 2024, from https://www.kaggle.com/competitions/feedback-prize-english-language-learning

(4) Kaggle. (n.d.). Automated Student Assessment Prize - Automated Essay Scoring. Retrieved May 22, 2024, from https://www.kaggle.com/c/asap-aes

(5) Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

(6) PyTorch. (2023). Model.py. In torchtitan/models/llama. GitHub. Retrieved May 22, 2024, from https://github.com/pytorch/torchtitan/blob/main/torchtitan/models/llama/model.py

Thank You