Lit-Shazam Know Thy Author

Tigran Poladian
School of Information
University of California, Berkeley
tpoladian@ischool.berkeley.edu

Denny Lehman
School of Information
University of California Berkeley
den@ischool.berkeley.edu

Abstract

In this paper we apply fine-tuned transformer models to perform authorship attribution on classical English texts. Research in authorship attribution typically focuses on small sized data constrained to make the learning task manageable, often relying on binary classification techniques. We propose a transformer RoBERTa model, fine-tuned to English literary lexicon, as part of an ensemble of models to achieve strong performance on the more challenging task of multi-classification of 19th and 20th century authors. Our purpose is to perform better than the baseline classical Naive Bayes performance and improve RoBERTa performance through transfer learning and ensemble methods. We choose classical authors to investigate how language models learn style and structure, motivated by use cases to understand and circumvent tweet datasets that include bot generated texts. Additional use cases are document plagiarism and literary search forensics. We explore the attribution task across non-transformer, transformer, transfer learning, and ensemble models, achieving macro average F1 scores of 0.85 on fine-tuned RoBERTa model, further improving to 0.86 with transfer learning and consistently topping at 0.87 with models in an ensemble.

1 Introduction

A variety of models and architectures have been used to perform authorship attribution, the task of classifying unknown text to the correct author (Boukhaled and Ganascia, 2017). Many of these are task-specific, but we wish to compare modeling techniques on a generic multiclass classification task as a way to make a generalized model that would perform well across multiple fields, such as forensics and plagiarism detection. Considering some data sets have been corrupted with bot generated text (Schwartz et al., 2013), we find commodity, human generated content in the form of classical English language literature from Project Gutenberg[1]. Thus, we compare various multi classification models on 3 sentence chunks from the books of 7 authors: Fitzgerald, Hemingway, Hardy, Dickens, Austen, Chesterton, and Shakespeare.

As part of our generalization strategy, we wish to have varying styles representative in our dataset. Style[2] as a form of vocabulary, sentence length, and syntactic structure have been hallmarks of authorship attribution (Singh et al., 2017), so we purposefully select authors with similar styles, like Hemmingway and Fitzgerald, to increase difficulty. Additionally, we add works by William Shakespeare as a foil to the other authors, since his unique stylometrics, play-structured works, and 16th century language will differ dramatically from the others.

Finally, from literature we identified popular models to stress test on authorship attribution. We organize them as traditional machine learning models (non-transformers), transformers, transfer learning, and ensembles. We will train each model on the 7 authors and evaluate the model performance with quantitative F1 scoring and qualitative comparisons of stylometric detection via the performance on Shakespeare classification.

2 Background

Authorship attribution has a rich history in natural language processing (Williams, 1975) thanks to the challenging task of learning unique literary styles and finding those patterns during inference. Bag-of-words models, which assume token independence, have historically performed well on text classification tasks thanks to their ability to detect useful function words (Argamon and Levitan, 2005). Paired with preprocessing steps, like tokenization and stemming, classifiers like linear models and Naive Bayes are still effective (Stamatatos, 2009). Neural networks have entered the sentence classification (Kim, 2014) and authorship attribution

[1] https://www.gutenberg.org/
[2] https://en.wikipedia.org/wiki/Stylometry

(Shrestha et al., 2017) domain with impressive results using CNNs. Pretrained large language models like RoBERTa (Liu et al., 2019) are state of the art for many language tasks. Not only has research started on authorship attribution with RoBERTa (Uchendu et al., 2020), but researchers have found that pretrained language models like RoBERTa have poor performance on multiclass classification of authors and can be surpassed by simpler models (Altakrori et al., 2021). With this background, we investigate whether pretraining transformers, transfer learning, or ensemble models show the best performance on authorship attribution.

3 Methods

3.1 Data Selection

Data was selected from Project Gutenberg. We chose 7 English language authors who had at least 3 literary works available for processing. Six of the authors were selected from 19th and 20th century periods and chosen such that some of the styles and themes would be similar. The 7th author selected was William Shakespeare. We assumed that Shakespeare's distinctive style, 16th century lexicon, and the nature of the play composition would serve to stand out from the other authors and therefore serve as a litmus test for the integrity of our model development. This indeed turned out to be the case as we will see in the results. Fitzgerald

	Author	Short Title	Title	Sentence Count	Sentence Groups	Group Counts
0	fitzgerald	gatsby,this side of paradise,beautiful and damned	The Great Gatsby,This Side of Paradise,The Bea...	15988	[I In my younger...	5330
1	hemingway	sun also rises,men without women,in our time	The Sun Also Rises,Men Without Women,In Our Time	9166	[BOO...	3056
2	hardy	mayor,jude,native	The Mayor of Casterbridge,Jude the Obscure,Ret...	18318	[I. One evening of late summer, before th...	6106
3	dickens	tale,great expectations,bleak house	A Tale of Two Cities,Great Expectations,Bleak ...	28184	[CHAPTER I. The Period It was the best...	9395
4	austen	emma,sense,pride	Emma,Sense and Sensibility,Pride and Prejudice	14695	[VOLUME I CHAPTER I Emma W...	4869
5	chesterton	wisdom brown,thursday,ball	The Wisdom of Father Brown,The Man Who Was Thu...	10083	[ONE -- The Absence of Mr Glass ...	3361
6	shakespeare	as you like it,caesar,hamlet,merchant,othello...	As You Like It,Julius Caesar,Hamlet,Merchant o...	13909	[SCENE: OLIVER'S house; FREDERICK'S c...	4637

Figure 1: Authors, works, and sentence examples from Project Gutenberg dataset

and Hemingway have similar styles from the same era, early 20th century[3]. Likewise, Hardy, Dickens, and Austen wrote comparable British stories primarily during the 19th century, Austen writing earlier than Dickens and Hardy, which we think is reflected in the results. The 6th author, G.K. Chesterton was selected to bridge the time gap between the Austen-Hardy-Dickens group and the Fitzgerald-Hemingway group, plus a writing style and themes that can arguably be characterized as elements from both groups [4]. We hoped this mix would provide a concise set and manageable set of data and enough to a challenge to produce practical results.

3.2 Data Preparation

Books and plays were collected from Project Gutenberg as text files. For each work, we manually located the start and end sections of the corpus and removed extraneous parts such as table of contents and copyright information. After this processing step, we used the NLTK sentence tokenizer[5] to split the corpus into sentences and then group sentences into contiguous sets of 3 sentence chunks. Additional processing was performed for the non-transformer models under the bag of words assumption, such that each word from the group of sentences was tokenized with NLTK's RegexpTokenizer[6] and then stemmed with PorterStemmer[7]. Finally, on all examples, we removed Named Entity Recognition (NER) tokens with spaCy[8]. To generate a balanced data for the multi classification problem, we identified the author with the fewest sentence groups and performed under-sampling for all other authors. The sentence groups/bag of words groups were then shuffled, split into Training and Test sets and assigned an integer label for the supervised training. We experimented with 80/20 and 90/10 splits, results were similar, and ultimately stayed with 90/10 for final runs. We ignore word counts or count equality because we wanted the model training on author style to include long and short sentences. Once the data was balanced, we selected a smaller subset (30%) in order to facilitate faster training and model development. Analysis using Mann-Whitney test indicated that the subset was statistically representative of the full balanced training set when comparing up to the top 70 words from the two groups.

3.3 Model Selection

A primary goal of the study was to analyze transformer architecture performance for author attri-

[3] https://www.litcharts.com/blog/analitics/what-makes-hemingway/

[4] https://www.britannica.com/biography/G-K-Chesterton

[5] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

[6] https://www.nltk.org/api/nltk.tokenize.RegexpTokenizer.html

[7] https://www.nltk.org/api/nltk.stem.porter.html

[8] https://spacy.io/api/entityrecognizer

bution and identify solutions for enhancing performance through transfer and ensemble learning. Research review indicated BERT and BERT variants as good candidates. We initially selected BERT and ultimately chose RoBERTa due to faster tuning. For non-transformer models, we analyzed model performance on Naïve Bayes, Stochastic Gradient Descent (SGD) classification, Random Forest (RF), Convolutional Neural Nets (CNN) for multiclassification cases. For its high performance on author classification and simplicity, we choose the Naive Bayes model as our baseline.

3.4 Model Development

RoBERTa base uncased was selected with transfer learning to a stacked CNN model, followed by a Dense layer and classification output. Fine-tuning top 4 layers produced the best results. Dropout was set to 0.2 to minimize bias.

The CNN model includes 4 filters, with kernel sizes ranging from 2 to 5, with the intention to capture bi-grams, tri-grams, and so on. RoBERTa with fine tuning on the training corpus provides attention to key words within each sentence group and across the author, whereas, CNN provides learning on the structural elements of the author's style in terms of long and short words and composition of the placement of word sizes. We think this approach provides a complimentary union with the transformer and deep learning models finding different elements, then the dense layer helping to tie in the various elements. Without the dense layer, we found that performance dropped.

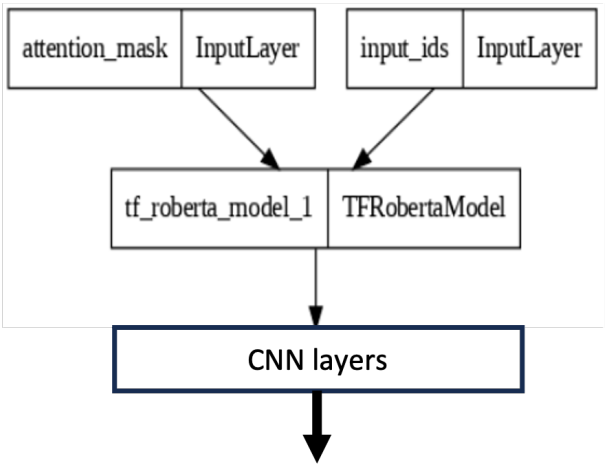


Figure 2: Transfer learning model: Roberta with Stacked CNN

3.5 Named Entity Recognition and Encoding

Prior to training, we used a named entity recognition (NER) model from spaCy to identify and remove proper names, with the idea to allow the models to learn the style and structure of the work rather than to memorize unique character names. This worked for the most part, however, due to the nature of the literary lexicon some names were missed.

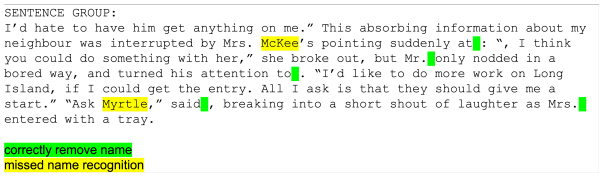


Figure 3: Example of named entity recognition, some names were missed

Names such as "McKee" and "Myrtle" were missed, as shown, or "Pip", "Clym", not shown. In future work, the NER model could use fine-tuning, particularly to train on names more common to the literary genre under study. We performed experiments without removing proper names and the models did have higher F1 scores but deemed it less practical and generalizable given that some literary works are defined and easily recognizable by character names, "Hamlet" for instance.

With NER, we had the option to remove the identified name or to replace the token with a generic word such as "PERSON". We tested both approaches and found that performance improved by removing the name entirely. We suspect, given the literary domain and genres, character names are paramount, and replacing it with "PERSON" added a lot of instances of this generic word, likely leading to an increase in noise. For RoBERTa and SGD models, stop-words were not removed. Stop words and lemmatization are tricky given that we want to understand literary style and certain authors tend to use more prepositions and conjunctions, short and direct versus long winded writing (Gatzemeyer, 2020) and any removal of words may have adverse performance effects. The CNN training layers include filters for bi and tri-grams, which tend to be primarily stop-words.

Regarding the RF model, experiments showed that the use of stop-words performed better than without. This is likely due to the learning approach that RF involves, to count words and emphasize near neighborhood weights rather than to include

word importance on a wider scale. Concretely, stop-words add noise for RF.

3.6 Evaluation

Class level F1 scores along with macro average and weighted F1 are reported as measures of multi-classification performance. Along with metrics, we sample positive and negative results to understand what may have caused the good and bad predictions.

4 Results and Discussion

We conducted 14 experiments, we termed "cases", referring to a random training data set. Each case was evaluated across different classical learning technique and some variant of a RoBERTa model. The last 3 cases were additionally evaluated for ensemble performance. Results from case 14 are presented along with average results for RoBERTa performance. Primary evaluation metric is the macro average F1 score although class scores are presented as well. Most experiments used maximum sequence of 128 tokens and learning rate of 0.0001 and our best results presented were with 256 tokens with learning rate of 1.0e-5. Across multiple runs, we found that 4 epochs worked best for RoBERTa and data presented was trained as such. As we can see in Fig. 5, after 4 epochs validation loss begins to flatten. Table 1 contains a summary of results.

4.1 Classical Learning Models

The Naive-Bayes model met expectations, performing the best of the classical models with a F1 score of 0.77 and a Shakespeare-ian class F1 score of 0.92, correctly differentiating the Bard[9] from the other authors. Additionally, it was easy to work with due to short training times and easy debugging.

Stochastic Gradient Descent (SGD) is reported to perform better for small datasets (Lipenkova, 2022). The performance did not beat baseline, although as part of an ensemble, there is utility in using it.

Random Forest (RF) did not perform well individually and produced an F1 macro average score of 0.58, the lowest of all candidates. The model is not able to adequately learn the complex authorship styles of classic literary fiction. Performance was too far below baseline to include in an ensemble.

[9] https://www.collinsdictionary.com/dictionary/english/the-bard

In the following example, RF confuses Fitzgerald for Austen:

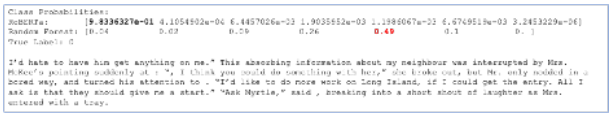


Figure 4: Random Forest model confuses Fitzgerald for Austen

4.2 RoBERTa

RoBERTa (R) performance far exceeded classical models. The attention mechanism and pre-training on a large language corpus help determine stronger word associations fundamental to literary structure. Best case performance was a 0.85 F1 score achieved after 4 layer fine-tuning. Any additional performance improvement will likely need to come from additional data engineering and fine-tuning the NER modules, however, over-training and introducing bias would be a concern.

4.3 Transfer Learning

We performed transfer learning with CNN and a dense layer neural net which improved F1 by 0.1 and 0.2, respectively. The 4-filter CNN model may be tuning to more nuanced word sizes. The Neural Net model uses a single hidden layer of size 128, chosen to correspond to the sequence token count, and may be helping performance by further coalescing attention results to better channel to the classification layer, but just as likely the dense layer improvement could be noise. As a thought experiment, and literal experiment, we asked ChatGPT3.5 to craft a 3-sentence passage about a fluffy pup in the styles of Austen and Shakespeare and in both instances the trained model failed to identify the style. In both cases, Further investigation is warranted as to the generalizability of our model.

4.4 Ensemble Method

The utility of the ensemble approach is that it can generalize well to handle small datasets (Abbasi et al., 2022) and large datasets. Our approach developed 3 aggregation mechanisms. The first 2 function element-wise from the set of model probabilities to select the maximum and average values. The selected probabilities correspond to the selected author. The third method evaluates the winner from each model and selects the author getting the most votes. We found that the averaging mechanism consistently performed better than the

Author - Label	SGD	RF	NB	R Avg	R Best	R+ CNN	R+ NN	Max	Avg	Vote
Fitzgerald - 0	0.74	0.51	0.70	0.85	0.84	0.88	0.90	0.89	0.88	0.84
Hemingway - 1	0.74	0.51	0.80	0.86	0.85	0.89	0.90	0.88	0.88	0.84
Hardy - 2	0.72	0.54	0.74	0.75	0.80	0.81	0.84	0.84	0.84	0.84
Dickens - 3	0.74	0.54	0.76	0.72	0.81	0.83	0.81	0.83	0.85	0.83
Austen - 4	0.81	0.60	0.73	0.83	0.88	0.85	0.89	0.83	0.85	0.85
Chesterton - 5	0.71	0.56	0.75	0.73	0.81	0.79	0.79	0.83	0.83	0.79
Shakespeare - 6	0.89	0.80	0.92	0.94	0.97	0.97	0.95	0.96	0.96	0.95
Accuracy	0.76	0.58	0.76	0.76	0.85	0.85	0.86	0.86	0.87	0.85
Macro Avg	0.76	0.58	0.77	0.81	0.85	0.86	0.87	0.87	0.87	0.85
Weighted Avg	0.76	0.57	0.76	0.80	0.85	0.86	0.86	0.86	0.87	0.85

Table 1: F1 score for each model across all authors

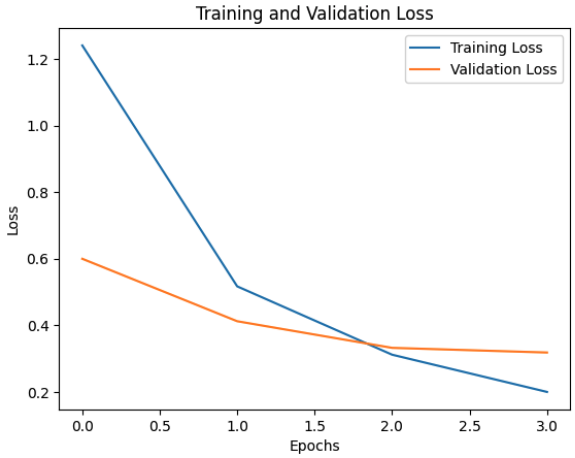


Figure 5: Transfer learning model training loss by Epoch

other two and improved RoBERTa performance by 0.02 to an F1 score of 0.87 while also improving balance across class probabilities. Ensemble performance results are shown in Table 1.

4.5 Performance Examples

We investigated examples the models mislabeled and found that sentence chunks would be near impossible to identify.

```
Class Probabilities
RoBERTa  :[0.1501432  0.20371762 0.13045247 0.20904544 0.11336076 0.11380325 0.07947725]
SGD      :[0.16080738 0.21495064 0.37985119 0.04554999 0.01501138 0.04011505 0.14371436]
RF       :[0.16080738 0.21495064 0.37985119 0.20904544 0.11336076 0.11380325 0.14371436]

True Label: 1

SENTENCE GROUP:
Yes. Yes. Yes.
```

Figure 6: An interesting mislabeled example from the test set

The sentence group would be difficult for any model and although all models miss classified this chunk, RoBERTa provided a strong second probability of choosing the right label.

5 Conclusion

We present 2 augmentations to the RoBERTa transformer model that improves performance on author attribution of classical English fiction authors. An ensemble of RoBERTa with a dense player as part of classical models, primarily naive bayes, achieved the top results. We identified that additional tuning of the NER performance could improve scores further. Additional research should be conducted with a larger set of authors and literary works from analogous genres with the aim to improve generalization towards unseen and untrained works such as for historical forensic analysis tasks. Furthermore, the ensemble prediction process can be optimized to achieve results faster, with an API layer to support literary document author discovery.

Acknowledgements

We thank Mark Butler and all the MIDS 266 instructors and TAs for guidance and support on this project.

References

Irshad Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Gadekallu, and Natalia Kryvinska. 2022. Authorship identification using ensemble learning. *Scientific Reports*, 12.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, pages 1–3.

Mohamed Amine Boukhaled and Jean-Gabriel Ganascia. 2017. 8 - stylistic features based on sequential rule mining for authorship attribution. In Bernadette Sharp, Florence Sèdes, and Wiesław Lubaszewski, editors, *Cognitive Approach to Natural Language Processing*, pages 159–175. Elsevier.

Jace Gatzemeyer. 2020. The hemingway writing technique you've never heard of. [Online; posted 13-March-2020].

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jana Lipenkova. 2022. Choosing the right language model for your nlp use case. [Online; posted 26-September-2022].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, Seattle, Washington, USA. Association for Computational Linguistics.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

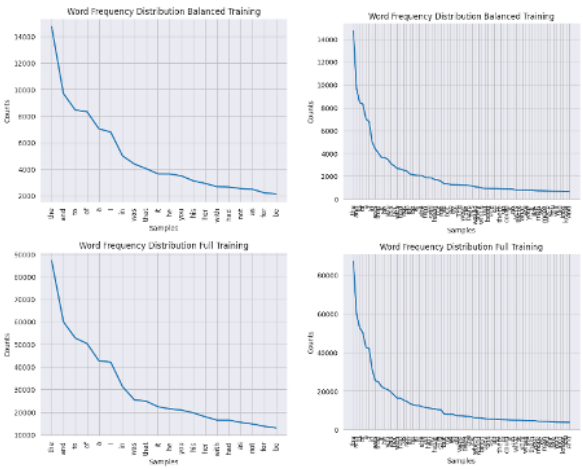
Pramod Singh, Kanaparthy Vivek, and Shirisha Kodimala. 2017. Stylometric analysis of e-mail content for author identification. pages 1–8.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

C. B. Williams. 1975. Mendenhall's studies of word-length distribution in the works of shakespeare and bacon. *Biometrika*, 62(1):207–212.

A Mann-Whitney Analysis Appendix

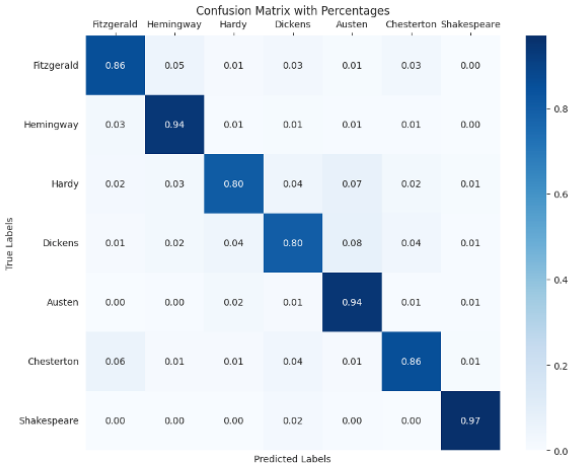


Up to the top 71 words
U Statistic: 2048.5
P-Value: 0.05428353716884162
There is no significant difference at 0.05 p-value

Top 71 words account for about 47% of the total word count for both large and small data sets

Total unique words are 20909 for small and 44813 for large, which is

B Model Performance Analysis Appendix




```
Confusion Matrix:
[[457  27   7  16   6  18   2]
 [  8 288   2   3   2   2   1]
 [ 13  16 491  27  45  14   5]
 [ 13  22  42 750  73  33   7]
 [  0   2  12   7 456   3   3]
 [ 20   4   5  12   4 289   3]
 [  0   1   1   9   2   1 450]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.86      0.88       533
           1       0.80      0.94      0.86       306
           2       0.88      0.80      0.84       611
           3       0.91      0.80      0.85       940
           4       0.78      0.94      0.85       483
           5       0.80      0.86      0.83       337
           6       0.96      0.97      0.96       464

    accuracy                           0.87      3674
   macro avg       0.86      0.88      0.87      3674
weighted avg       0.87      0.87      0.87      3674
```