Comparing the Efficiency of Machine Learning and Deep Learning Models in Source Code Authorship Attribution

Richard Čerňanský

March 2, 2025

Abstract

1 Introduction

Introduction

2 Methods

Methods

3 Results

Models' Author Attribution Accuracy for Different Author Set Sizes				
Model	**Author Set Size**			
	110	27	11	3
Random Forests (TF–IDFs w/o comments)	66.24% | 3.63 min	70.60% | 0.23 min	80.56% | 0.08 min	85.84% | 0.05 min
BERT (source code w/o comments)	72.47% | 70.33 min	82.72% | 32.49 min	89.84% | 9.34 min	90.70% | 1.82 min
BERT (AST pre_order traversal)	24.07% | 73.53 min	35.66% | 30.00 min	40.62% | 5.69 min	83.72% | 1.22 min
AttentionNN (AST paths)	—	—	—	—

Table 1: Comparison of models across author attribution accuracy and training time for different author set sizes.

4 Discussion

Here we discuss the implications of our results.

5 Conclusion

We conclude that this document compiles successfully. If you can read all sections, view the figure, and see the mathematical expressions, your LaTeX environment is correctly set up.

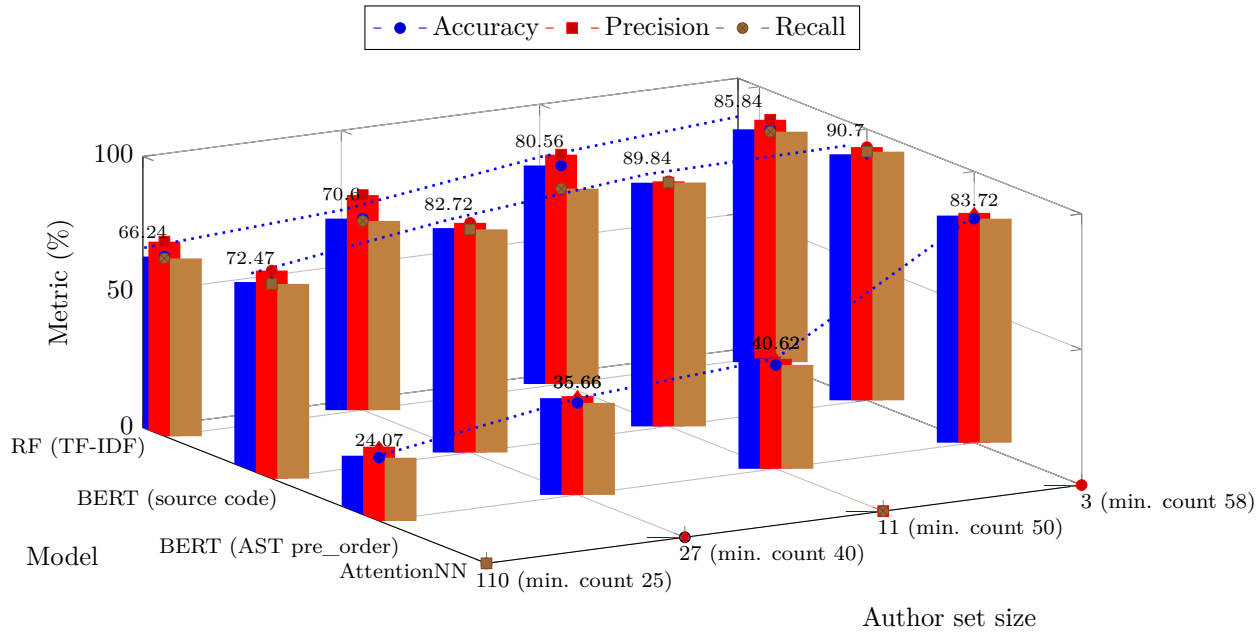

Figure 1: Bar chart comparing Accuracy (with accuracy values labeled), Precision, and Recall across models and author set sizes with minimum function count per author.