Language Proficiency and Authorship Classification Strength

Léa Gainon, Kaede Johnson, Arthur Tabary
EPFL Course CS-433

***Abstract*—We investigate whether language proficiency variation has a performance impact on traditional stylometric author classification methods. Insight into this potential link could lessen the misattribution of dangerous or illegal text. In particular, we pursue multi-class author classification SVM models with Writeprints-inspired feature sets to aggregated english-language Reddit comments written by three cohorts: native English authors, non-native English authors, and a mix of the two. We find that applying SVMs to a mix of both native and non-native English authors consistently outperforms SVMs applied to either native or non-native authors alone. Additionally, after tuning n-gram collection sizes specifically for each cohort, we do not find evidence of an 'ideal' set of model parameters for a given language proficiency level - but we do find non-native authors consistently benefit more from variation in said parameters than do native authors.**

I. Introduction

Stylometry can be used to identify anonymous authors of dangerous or illegal forum posts, often in support of legal action. Classification tools should therefore be robust to variations in author type. If they are not, authors of dangerous text may escape classification efforts, and innocent authors may even be accused in their place. Unfortunately, the nature of online privacy means author characteristics can remain unobserved, complicating study of their relationship to stylometric tools. We isolate one oft unobserved characteristic - language proficiency - and ask the following: does language proficiency affect the efficacy of traditional stylometry in online forums? If so, is there an optimal parameter set to use based on target author language proficiency?

We are not aware of any research that tackles these specific questions. Al-Khatib and Al-qaoud [1] train separate author verification (not identification) models on native and non-native authors of online newspaper articles, but they do not consider online forums, nor do they measure the performance differential for a model trained on both native and non-native authors types without explicit language efficiency information in the feature set.

The absence of language proficiency data is not necessarily deleterious to model performance. Differences in language use between native and non-native English authors can be observed in standard stylometric features (Shirzadi et al. [2], Duppenthaler [3]), suggesting robust feature extraction on text alone may capture the effect of language proficiency variation. Indeed, Bergsma et al. achieve high accuracy with common stylometric features when predicting native or non-native status for authors of scientific articles, albeit with constructed language proficiency labels [4]. Nonetheless, it is feasible that authors with less command of a given language are, for example, less able to display their idiolect on the page, complicating classification.

Online forums are particularly useful in this context due to their low barrier to entry and casual nature, which help reveal the writing style of the individual author [5]. Additionally, online forum data is easy for researchers to collect. Reddit in particular houses r/languagelearning, where users are encouraged to self-report language proficiency levels, providing rare access to a usually opaque author characteristic. Age, gender, and even cross-language corpora ([6], [7]) have received stylometric attention in the Reddit domain, but we current research does not consider language proficiency.

We propose standard stylometric analysis on Reddit comments within a novel framework that isolates the impact of language proficiency. Using self-reported language proficiency data, we divide Reddit users into cohorts of native, non-native, or mixed-proficiency authors of English text. We then develop corpora from these users' Reddit comments, train SVM author identification models for each cohort, and evaluate each model's performance across our three cohorts. Our aim is to quantify differentials in the best- and worst-case outcomes for stylometric analysis based on presumed knowledge of target authors' language proficiency.

II. Data & Preprocessing

Our original dataset included 2,144 Reddit users who commented on https://www.reddit.com/r/languagelearning/ (henceforth "r/languagelearning/") between August 5 and October 27, 2022. Self-reported language proficiency flairs reflect data from r/languagelearning in the same time range, and the user comment histories across all subreddits amounted to 1,099,808 comments between July 9, 2010 and December 7, 2022. We removed users who were suspended, deleted, or lacking intelligible language proficiency flairs. After retaining the most recent twelve months of comment history by user to protect against time's effect on language skill (see [8]), we removed comments that were solely links or, to target a common behavior of bot accounts, duplicate or near-duplicate comments by the same user.

Next, we applied the language classifying Python library langdetect [9] to all comments and retained those which received one highest-probability language assignment (as opposed to multiple or none). Those which could not be classified were solely numbers, punctuation, emojis, or other

non-letter characters. Those which received multiple language assignments with equal probability were generally exceptionally short, composed of multiple languages, or both. We allowed some non-language elements (e.g. GIF references) to remain without plans to accommodate them in the feature set, thereby permitting their potential relationship to language proficiency to persist through standard stylometric features (such as punctuation n-grams) instead of Reddit-specific structural features, in line with the broader scope of our research questions.

Figure 1: Number of authors with >10K words by language. Log scale.

An overwhelming amount of the remaining comments written by users with more than 10,000 words in their corpus were written in English (Figure 1). This posed data availability problems for non-English corpora development. Thus, we limited our analysis to English-language comments and users self-reporting some level of English proficiency. Users reporting A1, A2, B1, B2, C1, and C2 English proficiency were aggregated into a set henceforth referred to as "non-native authors". Users reporting native English skills are henceforth referred to as "native authors".

We took inspiration from Overdorf and Greenstadt [10] when developing corpora from online forum text. In particular, we aggregated comments in random order at the author level to create documents (henceforth "feeds") of at least 500 space-delimited character sequences. Unlike Overdorf and Greenstadt, we did not split comments to create a uniform feed length, instead preserving the full grammar flow of complete comments and allowing feeds to range between 500 and 1,892 character sequences in length. Our features avoid absolute counts to accommodate this length variability.

Authors without enough English text for twenty such feeds (approx. 10,000 words) were excluded, leaving 354 native and 135 non-native authors available for model development and evaluation. When necessary, we filtered each remaining author's corpus down to twenty feeds at random.

III. Feature Development

Our features are time-honored in stylometry due to the nature of our research question. They are widely used, in our case sourced from Abbasi and Chun [11] as well as de Vel, Anderson, Corney, and Mohay [12], and extracted for each author feed. They can be grouped as follows (parentheses indicate name in Table I, to be introduced in Results section):

A. Lexical

- Letter, digit, special character share of characters
- Average character count of all words
- Distribution of letter 1-, 2-, and 3-grams (Let #-G)
- Distribution of digit 1-grams (Dig 1-G)
- Distribution of 1-20 character words
- Proportion of words 1-3 characters long
- Hapax legomena share of all non-stop-word tokens
- Hapax legomena share of unique non-stop-word tokens
- Unique token share of all non-stop-word tokens

B. Syntactic

- Stop words share of tokens
- Distribution of punctuation & special character 1- and 2-grams (P&S #-G)
- Punctuation, whitespace share all characters
- Distribution of POS tag 1- and 2-grams (POS #-G)
- Distribution of upper- and lowercase letters
- Distribution of all-upper-, all-lower-, first-upper-rest-lower-, and othercase words

C. Structural, Content, and Idiosyncratic

- Median comment length[1]
- Distribution of word 1- and 2- grams.
- Proportion of words that are misspelled[2]

Above, 'word' refers to the space-delimited letter sequences in a feed after all non-letter, non-space characters are removed (misspellings and slang remain). 'Token' refers to the space-delimited character sequences created by applying spaCy, a natural language processing library in Python [13], to raw comments within each feed. spaCy is also the source for our POS tags and collection of stop words.

We normalize median comment length and average word length so they, like all other features, range between 0 and 1. In share-of-characters features, whitespace is included in total character count. N-grams may not stretch across two comments within a feed.

The number of words, characters, or POS tags we pull from train data for n-grams is subject to parameter tuning as described in the next section.

IV. Models and Methods

We pursue multi-class author classification through linear SVM models. Per T. Neal et al., SVM is effective with large, sparse feature sets, and is common in stylometry and multi-class author attribution applications in particular[14].

[1]Here, length is the number of space-delimited character sequences.
[2]Misspellings are found using Python's inbuilt autocorrect package.

Table I: Optimal parameter set by cohort. From the "development" stage.

Cohort	Auth	Model	λ	Let 1-G	Let 2-G	Let 3-G	Let 4-G	Dig 1-G	P&S 1-G	P&S 2-G	POS 1-G	POS 2-G	Word 1-G	Word 2-G	**Accuracy**	**F1**
1: Native	30	Baseline	1	26	50	50	50	10	36	50	48	50	50	50	**80.00**%	**78.44**%
		Tuned	7	26	500	500	500	10	36	100	48	500	500	20	**91.33**%	**91.67**%
2: Non-Nat	30	Baseline	1	26	50	50	50	10	36	50	48	50	50	50	**63.33**%	**58.57**%
		Tuned	10	26	500	500	500	10	36	20	24	200	200	200	**83.33**%	**80.56**%
3: Mix	30	Baseline	1	26	50	50	50	10	36	50	48	50	50	50	**85.00**%	**84.56**%
		Tuned	5	26	500	500	500	10	36	500	48	200	500	20	**100**%	**100**%

Our process involves a "development" stage and an "evaluation" stage. They are outlined below and in Appendix Figure 1. Both stages involve three cohorts: native authors (cohort 1), non-native authors (cohort 2), and an equal mix of native and non-native authors (cohort 3). Importantly, cohort 3 always consists of equal-sized subsets of cohorts 1 and 2.

In the "development" stage, we populate cohorts 1 and 2 at random with thirty authors each and subsequently populate cohort 3 via random sampling. We then train a separate linear SVM model on each cohort with Python's scikit-learn library [15]. Our models pursue a 30-class classification problem. That is, we apply author-specific labels to feeds. Again following the procedure of Overdorf and Greenstadt, we reserve 10% of each author's twenty feeds for model validation. Equitable stratification ensures no single author can sway model performance.

The development stage is so-named for its emphasis on parameter tuning. We perform the following process for each cohort: first, train a baseline model with regularization parameter $\lambda = 1$ and collections sized at fifty for all n-gram features. Next, perform an eleven-degree grid search with 262,144 combinations on n-gram collection sizes. This takes us 980 core-hours using the EPFL supercomputer SCITAS. After identifying the models with the highest F1 scores, calculate the modal value of each n-gram collection size among these 'winning' models, and extract the model with the highest share of parameters equal to the modal parameter set. Last, perform a linear search for a more optimal λ, should one exist.

In the "evaluation" stage, we perform three trials according to the following process: first, populate cohorts 1 and 2 with a random sample of thirty authors each. Next, populate cohort 3 via random sampling from cohorts 1 and 2. Finally, train all tuned parameter sets from the development stage on 540 author feeds and test on 60 author feeds, ensuring equal author representation in the train and test sets.

Note that authors cannot be involved in both the development and evaluation stages, nor can they appear in more than one of the three evaluation-stage trials.

V. Results

The effect of development stage parameter tuning can be seen in Figure 2. Tuning does not change the inter-cohort ordering of model performance. However, it does improve non-native authors' F1 score by 22 percentage points, nearly ten percentage points above the improvement for native authors. Mixed-proficiency authors see about as much improvement as native authors but are bounded above by perfect classification.

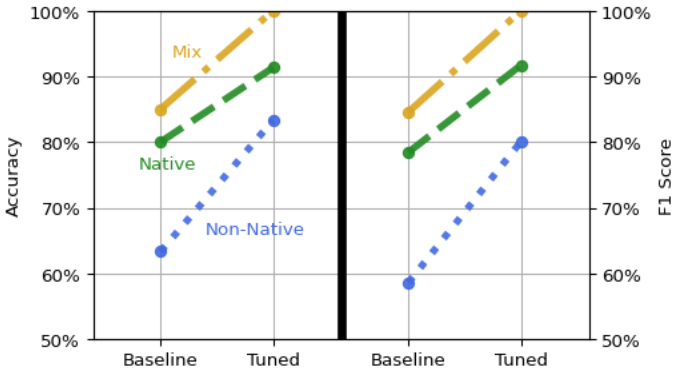

Figure 2: Increases to accuracy and F1 score due to parameter tuning. From the "development" stage.

Each cohort's optimal parameters can be found in Table I. Optimal sets for cohorts 1 and 3 share nine out of twelve parameter values, while the optimal sets for cohorts 2 and 3 share seven out of twelve. Optimal sets for cohorts 1 and 2 share six.

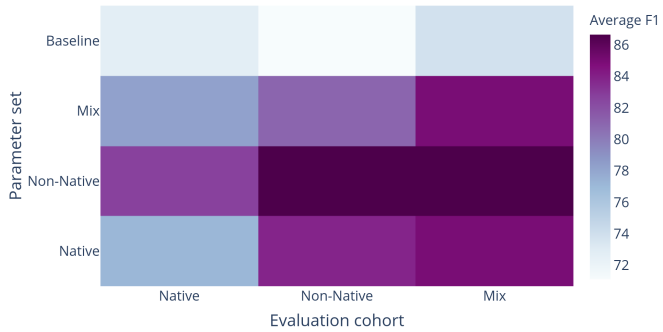

Figure 3: Average F1 score by evaluation cohort and parameter set. From the "evaluation" stage. Darker is higher.

Figure 3 portrays average F1 scores for each evaluation cohort during the evaluation stage. Mixed-proficiency authors have the highest F1 scores regardless of the development-stage parameter set applied, though non-native

authors are not far behind. Most glaring are the native author cohort's relatively low performance increases.

In Figure 4, we display the evaluation stage's average increase in F1 over the baseline model for each combination of evaluation cohort and parameter set. The non-native and mixed-proficiency author sets both see 10-14 percentage point higher F1 scores relative to the baseline regardless of which parameter sets are applied. Native authors, meanwhile, might only see a 5-7 percentage point increase in F1, depending on which model is applied. Notably, non-native parameters produce the highest F1 score for all three broad cohorts, and for the native cohort, native parameters produce the worst average increase in F1 score. Complete evaluation stage trial results are reported in Appendix Table I.

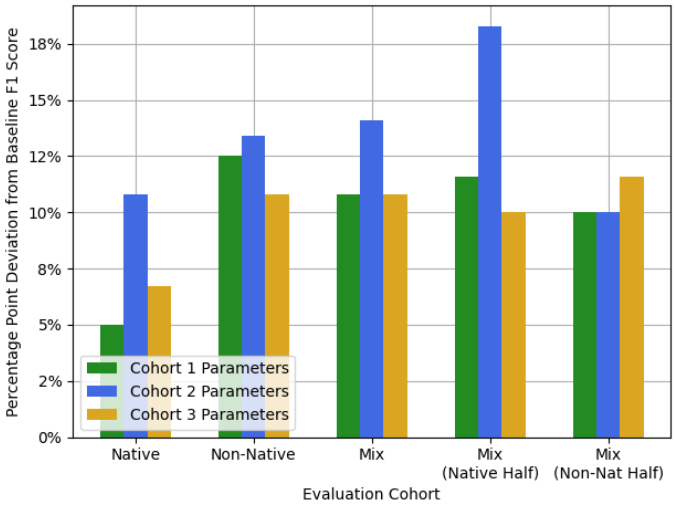

Figure 4: Average boost to F1 score by evaluation cohort and parameter set. From the "evaluation" stage.

VI. Discussion

Across the development stage (Table I) and evaluation stage (Appendix Table I) we consistently find higher F1s when native and non-native authors are classified in combination. Clearly, language proficiency variation produces a demonstrable, efficiency-based distinction in writing style that is readily visible to support vector machines. In the online forum space, traditional stylometric methods improve with greater diversity in language proficiency.

Due to the lack of compelling F1 trends in Figures 3 and 4, we do not see any evidence that there exists an optimal parameter set for native, non-native, or mixed-proficiency author sets. Different parameter sets either produce similar F1 scores (non-native and mixed cohorts) or counter-intuitive results (native cohort). Indeed, the success of non-native parameters on native authors suggests development stage tuning more likely targeted specific author idiolects than broader cohort language proficiency, especially for native authors. Lack of a similar outlier for non-native authors could mean non-native authors are more difficult to overfit. In any case, universal improvement over the baseline means any potential parameter overfitting was not too severe.

That non-native authors benefited more than native authors in the evaluation stage (Figure 4) shows non-native authors benefit more from the tuning or perhaps simple expansion of n-gram collection sizes. We conclude native authors' idiolects are more immediately observed than non-native authors' idiolects, but this can be compensated for with parameter tuning, even to the point of achieving higher performance with non-native authors than native authors.

To uncover the authors of dangerous text without inculpating the innocent, we recommend investigators include as much language proficiency variety in their author sets as possible. Expanded n-gram collection sizes will always be helpful, but the extra time and resources spent developing larger collections would be especially helpful if investigators suspected their target authors did not have native English proficiency. Unfortunately, we cannot recommend an optimal parameter set based on language proficiency alone. We therefore caution investigators against borrowing parameter sets applied to authors of similar English proficiency without doing any n-gram size and λ tuning of their own, especially if the parameter sets were tuned on native English authors.

VII. Conclusion

We sought to determine whether language proficiency affects the efficacy of traditional stylometry with online forum text, and if so, whether there is an optimal parameter set based on target author language proficiency. Through a novel framework that stratifies authors based on English level, we found that greater author variation in language proficiency improves classification performance, and that traditional stylometric features identify non-native idiolects less readily than native idiolects when n-gram collection sizes are low. Additionally, our findings suggest stylometry is more at risk of overfitting the parameter set for native authors than non-native authors, and that there does not exist an optimal classification parameter set based on language proficiency alone.

While outside the scope of our work as it is not yet considered standard to classification models, Bergsma et al. espouse tree substitution grammar as an important metric for classifying authors as native or non-native [4]. Thus, it should be considered in future work on this topic. Separately, a larger collection of comments would allow researchers to split out the non-native cohort into A-, B-, and C-level proficiency authors, provide enough authors to prevent overfitting in the parameter development stage, and allow similar analysis on non-English comments. Finally, at the risk of building models more relevant to Reddit than online forums in general, researchers could incorporate more structural features in the feature sets, in the case language proficiency correlates to these as well.

REFERENCES

[1] Mahmoud A. Al-Khatib and Juman K. Al-qaoud. Authorship verification of opinion articles in online newspapers using the idiolect of author: a comparative study. *Information, Communication & Society*, 24(11):1603–1621, 2021.

[2] Milad Shirzadi, Farzad Akhgar, Amir Rooholamin, and Sajad Shafiee. A corpus-based contrastive analysis of stance strategies in native and nonnative speakers' english academic writings: Introduction and discussion sections in focus. *International Journal of Research*, 2:31–40, 2017.

[3] Peter Duppenthaler. A comparison of essays written by native and nonnative speakers of english on the topic kokusai shakai (international society). In *American Educational Research Association (AERA) annual meeting, San Diego, CA: Pearson Education*, 2004.

[4] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, 2012.

[5] Thanh Nghia Ho and Wee Keong Ng. Application of stylometry to darkweb forum user identification. In Kwok-Yan Lam, Chi-Hung Chi, and Sihan Qing, editors, *Information and Communications Security*, pages 173–183, Cham, 2016. Springer International Publishing.

[6] Yaakov HaCohen-Kerner. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, 199:117140, 2022.

[7] Benjamin Murauer and Günther Specht. Dt-grams: Structured dependency grammar stylometry for cross-language authorship attribution, 2021.

[8] Guilherme Ramos Casimiro and Luciano Antonio Digiampietri. Authorship attribution with temporal data in reddit. In *XVIII Brazilian Symposium on Information Systems*, SBSI, New York, NY, USA, 2022. Association for Computing Machinery.

[9] M Danilak. langdetect: Language detection library ported from google's language detection. *See https://pypi. python. org/pypi/langdetect/(accessed 19 January 2015)*, 2014.

[10] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, 2016, 07 2016.

[11] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2), apr 2008.

[12] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64, dec 2001.

[13] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[14] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6), nov 2017.

[15] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

APPENDIX

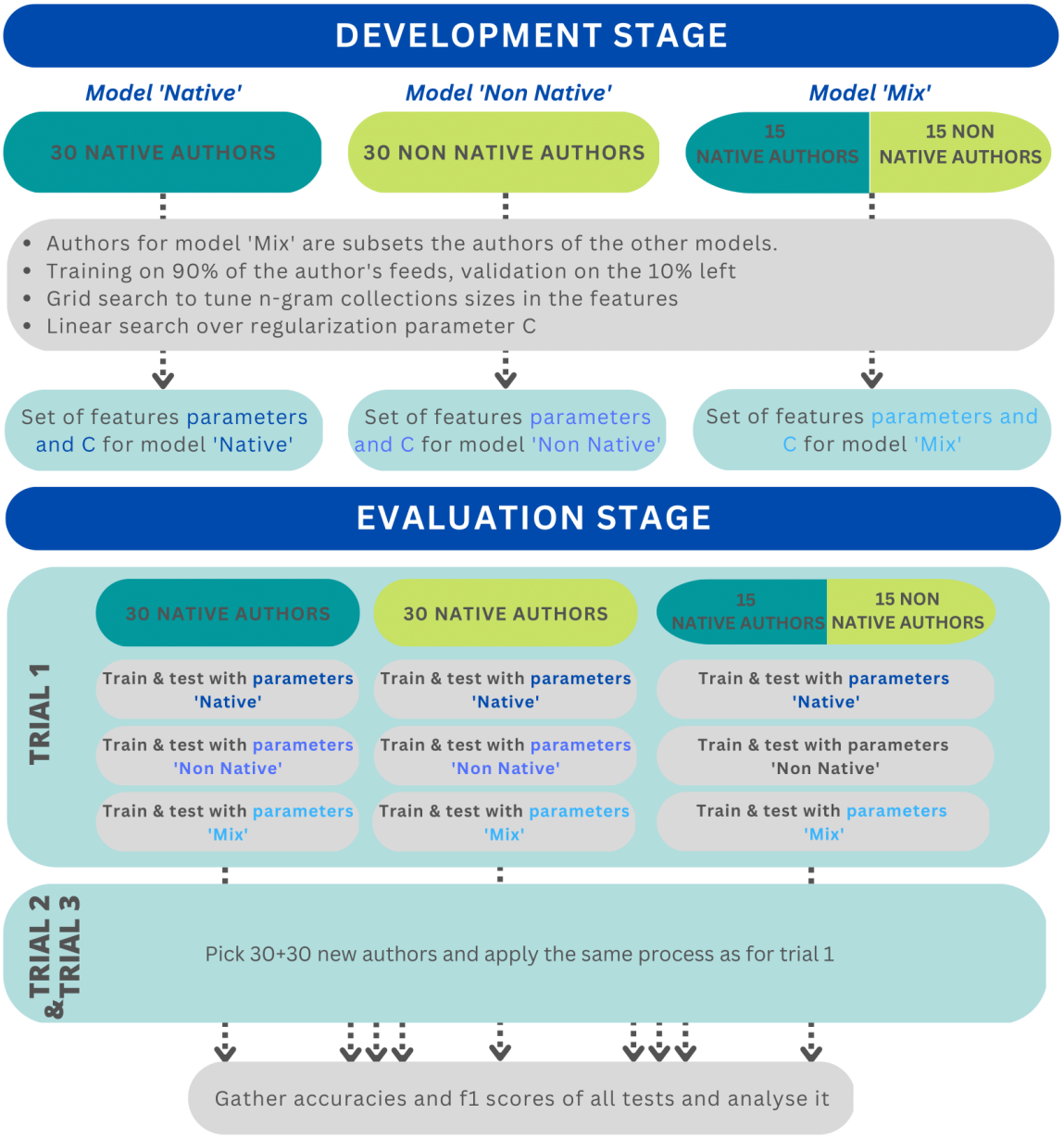

Figure 1: Architecture for the “development” and “evaluation” stages.

Table I: Complete results from the "evaluation" stage.

Model	Eval Cohort	Per Trial Counts: Authors	Tr Feeds	Te Feeds	Accuracy: T1	T2	T3	F1: T1	T2	T3	**Average**: **Acc**	**F1**
Baseline	1: Native	30	540	60	73%	68%	77%	71%	64%	73%	**72.8**%	**72.5**%
	2: Non-Native	30	540	60	68%	70%	75%	65%	67%	74%	**71.1**%	**72.5**%
	3: Mix	30	540	60	73%	68%	80%	70%	66%	79%	**73.9**%	**74.2**%
	Native Subset	15	270	30	77%	67%	77%	74%	69%	79%	**73.3**%	**71.7**%
	Non-Native Subset	15	270	30	70%	70%	83%	73%	75%	83%	**74.4**%	**76.7**%
Native, Tuned	1: Native	30	540	60	77%	75%	80%	75%	73%	78%	**77.2**%	**77.5**%
	2: Non-Native	30	540	60	82%	85%	85%	80%	84%	84%	**83.9**%	**85.0**%
	3: Mix	30	540	60	85%	87%	83%	83%	86%	82%	**85.0**%	**85.0**%
	Native Subset	15	270	30	90%	87%	80%	90%	87%	82%	**85.6**%	**83.3**%
	Non-Native Subset	15	270	30	80%	87%	87%	83%	88%	86%	**84.5**%	**86.7**%
Non-Native, Tuned	1: Native	30	540	60	82%	83%	83%	79%	80%	80%	**82.8**%	**83.3**%
	2: Non-Native	30	540	60	88%	88%	83%	86%	88%	83%	**86.7**%	**85.9**%
	3: Mix	30	540	60	83%	88%	88%	82%	88%	88%	**86.7**%	**88.3**%
	Native Subset	15	270	30	87%	90%	90%	86%	92%	92%	**88.9**%	**90.0**%
	Non-Native Subset	15	270	30	80%	87%	87%	84%	88%	87%	**84.5**%	**86.7**%
Mix Tuned	1: Native	30	540	60	77%	77%	82%	75%	73%	80%	**78.3**%	**79.2**%
	2: Non-Native	30	540	60	77%	82%	85%	75%	81%	85%	**81.1**%	**83.3**%
	3: Mix	30	540	60	85%	88%	82%	83%	88%	80%	**85.0**%	**85.0**%
	Native Subset	15	270	30	90%	87%	77%	90%	87%	79%	**84.4**%	**81.7**%
	Non-Native Subset	15	270	30	80%	90%	87%	83%	92%	86%	**85.6**%	**88.3**%