A Supplementary Materials

The countries and languages reflected in the dataset are listed in Table 10.

Country	Language
Austria	German
Germany	German
Australia	English
Ireland	English
New Zealand	English
United Kingdom	English
United States	English
Bulgaria	Bulgarian
Croatia	Croatian
Czech Republic	Czech
Estonia	Estonian
Finland	Finish
France	French
Greece	Greek
Hungary	Hungarian
Italy	Italian
Lithuania	Lithuanian
Netherlands	Dutch
Norway	Norwegian
Poland	Polish
Portugal	Portuguese
Romania	Romanian
Russia	Russian
Serbia	Serbian
Slovenia	Slovenian
Spain	Spanish
Mexico	Spanish
Sweden	Swedish
Turkey	Turkish

Table 10: Countries and Languages in Dataset

A.1 Data downsampling

We randomly downsampled the data to ensure that each class had the same number of users. To do so, we calculated the number of users tagged with each label in the data (the label can be a language, a family, or native/non native, depending on the task). We then randomly selected the minimum number of users with each label. Note that the number of chunks per label is still not equal because each user may have a different number of chunks. In order to cancel the bias caused by users that are over-represented in the texts of their country (i.e., users authoring a significant portion of their country's sentences), we used at most the median number of randomly selected chunks for each user. For native users the median is 3, for nonnative ones it is 17.

A.2 Evaluation scenarios

We defined two evaluation scenarios: *in-domain*, where training and testing is done only on chunks from the European subreddits; and *out-of-domain*, where we train on chunks from the European subreddits and test on chunks from other subreddits, making sure they were authored by different users.

In both cases, a *fold* is defined over *users*, rather than text chunks. Consider first the in-domain scenario. We only consider chunks in the European subreddits, of which there are 18,172 (after downsampling). The number of users in this dataset is 8,145, but to avoid bias, we only select the minimum number of users for each label; for the NLI task, this number is 104, so we are left with $104 \times 23 = 2392$ users. We now run 10-fold cross-validation evaluation on the set of (chunks authored by) these users, where in each fold we train on 90% of the users and train on the remaining 10%. We use the same evaluation strategy for the two other tasks.

In the out-of-domain scenario we use the much larger non-European corpus. We begin with almost 2M text chunks authored by almost 34K users, but downsampling reduces this number to about 400K chunks. As above, we are left with 2392 users. We randomly select 10% of these users in a stratified way (uniformly across L1s), and use their non-European chunks for testing. For training, we use the European chunks authored by the remaining 90% users. We repeat this process ten times and report the average of the ten runs. Again, we use the same evaluation strategy for the two other tasks.