Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

(Un)fair Mistakes on Social Media: How Demographic Characteristics Influence Authorship Attribution

Jasmin Wyss[1], **Rebekah Overdorf**[1,2]

[1]Ruhr Universität Bochum
[2]Research Center Trustworthy Data Science and Security in University Alliance Ruhr
jasmin.wyss@rub.de, rebekah.overdorf@rub.de

Abstract

Authorship attribution techniques are increasingly being used in online contexts such as sock puppet detection, malicious account linking, and cross-platform account linking. Yet, it is unknown whether these models perform equitably across different demographic groups. Bias in such techniques could lead to false accusations, account banning, and privacy violations disproportionately impacting users from certain demographics. In this paper, we audit authorship attribution for bias in three different ways with respect to gender, native language, and age. First, we evaluate how the proportion of users with a certain demographic characteristic impacts the overall classifier performance. Second, we evaluate if a user's demographic characteristics influence the probability that their texts are misclassified. Our results for these two evaluations indicate that authorship attribution does not demonstrate bias across demographic groups in the closed-world setting. Third, we evaluate the types of errors that occur when the true author is removed from the suspect set, thereby forcing the classifier to choose an incorrect author. This controls for the influence of a user's fluctuations in writing style on the mistakes made. Unlike the first two settings, our results here indicate a tendency to attribute authorship to users who share the same demographic characteristic as the true author. Our results highlight that though an NLP model may appear fair in the closed-world setting for a performant classifier, this does not guarantee fairness when errors are inevitable.

Code — https://github.com/JWYSS2/UnFairMistakes

Introduction

Pseudonymous social networks such as Reddit, Discord, and Telegram differ from social networks like Facebook in that a majority of users do not display their real names. Reddit (Reddit 2025b) and Discord (Discord 2025) explicitly allow for accounts with no real name associated with them, and Telegram recently updated its account creation protocol to allow for anonymous accounts (Shakir 2025). Because many use these platforms with no publicly visible tie to their real-world identity, users on these platforms have some expectation of anonymity.

Authorship attribution, an application of stylometry, undermines the pseudonymous protections that users rely on for privacy and anonymity. In the same way that users can leverage anonymity for malicious (e.g., harassment, disinformation) or constructive (e.g., peer support on sensitive topics, whistleblowing) purposes, authorship attribution can be applied in both harmful and protective ways. Platforms or moderators can use stylometry to strengthen platform safety and enforce platform policies and rules by finding sock puppet accounts (Sakib and Spezzano 2022), detecting coordinated malicious accounts (Kireev et al. 2025), and linking malicious users across platforms (Afroz et al. 2014). However, stylometry can also be used to deanonymise genuine users (Alonso-Fernandez et al. 2021). For example, anonymous comments about users' negative experiences working for a company or organisation are common on Reddit and are often voted to the top of comment threads. The target of such a comment could use stylometry to determine which of their employees authored the comment. Similarly, because Reddit posts can contain sensitive information (Brown et al. 2018), another user who suspects that the author of such a post is someone they know could employ stylometry to confirm their suspicion.

Stylometry has a wide range of applications, including both harmful and beneficial uses. Understanding the weaknesses of authorship attribution classifiers can help develop better authorship obfuscation methods to protect users. When authorship attribution is used to protect platform integrity, understanding its limitations allows for better interpretation of its results. For example, if the sockpuppet detection mechanism is prone to falsely identifying accounts based on topic-similarity, this would need to be mitigated and addressed in the interpretation of the results. Especially in cases where the classes are imbalanced, good performance metrics, such as a high F1 score or high accuracy, are not enough. It is crucial to test the mechanisms for biases and third variable influences.

We design a three-step audit to test if there is bias in authorship attribution related to the demographic characteristics age, gender and native language.

First, we vary the proportion of users from different demographic backgrounds in the training data and evaluate the impact on the classification score. This reflects the real-world conditions on social networks, including Reddit, Discord, and Telegram, where users are not evenly distributed by their demographic traits. If, e.g., a high percentage of

non-native English speakers in the suspect set negatively impacts classification performance, it would follow that an English-language gaming channel on Discord with many users from Southeast Asia would have a lower detection rate for sockpuppet accounts. On Reddit, subreddits such as r/europe, where many users are not native English speakers but most posts and comments are in English, would be similarly impacted.

Second, we examine who is impacted by the errors. For this, we compare the misclassification rates across different demographic groups. Consider a subreddit focused on domestic abuse where most of the users are women. If women are more likely to be correctly classified than men, this would cause more women to be subject to successful deanonymisation attacks on their sensitive posts.

These first two sets of experiments focus on closed-world classification errors, in which the classifier could predict the correct author, but instead infers another member of the suspect set. However, these errors generally occur on texts that deviate most from the author's normal writing style, making it difficult to disentangle an author's stylistic inconsistencies from real demographic bias. To address this, we evaluate if demographic characteristics influence the errors made when the true author is removed from the suspect set, thereby forcing the classifier to choose an incorrect author. This setting allows for an examination of the influence of demographic characteristics on writing style without the writing style consistency impacting the result.

Contributions

Our first contribution is a novel Reddit data collection methodology combining the Reddit API as well as the Wayback Machine API. This methodology allows for a more extensive collection that still complies with Reddit's terms of service.

Secondly, we are the first to adapt fairness metrics from the literature to study the influence of demographic characteristics on authorship attribution. The resulting experiments show that for our chosen classification scheme:

- The demographic makeup of a set does not impact the overall classifier performance.
- The demographic characteristics of a user do not impact the probability that their text is misclassified.
- In a forced misclassification setting where the true author is not present, the demographic characteristics *do* impact who is falsely accused.

Related Work

Authorship Attribution or Verification Stylometry applied to text from multiple accounts on the same platform can be used to deanonymise users or to identify sockpuppet clusters (Weerasinghe, Singh, and Greenstadt 2022). Being able to detect accounts held by the same person allows for better mitigation strategies as well as an understanding of manipulative behaviour. On Wikipedia, identifying sockpuppets helps prevent interest groups from influencing the narrative (Sakib and Spezzano 2022; Raszewski and de Kock 2025). Writing style can also be used to link users across platforms (Xu and Fung 2025), for example, to deanonymise accounts on dark web forums by linking them to clear web accounts (Arabnezhad et al. 2020).

Topic and other biases Previous work has already explored the importance of one important confounding factor on authorship classification performance: topic. Features used in stylometry have long been categorised in terms of how much style as opposed to content they appear to encode (Abbasi and Chen 2008; Stamatatos 2009). The actual impact of topic has been studied in a variety of ways. For example, Weerasinghe, Singh, and Greenstadt (2022) use feature analysis to ensure the most important features are not purely topic related. Sari, Stevenson, and Vlachos (2018) perform an ablation study to highlight the importance of content-encoding features for stylometry. Finally, Wang et al. (2023) tested the transformer-based LUAR model (Rivera-Soto et al. 2021) for its capacity to encode writing style distinct from content. All three show that topic is an important confounding factor to consider.

Other work focuses on influences on writing style, such as genre/domain (Bartelds and de Vries 2019; Barlas and Stamatatos 2020) or the dependency on artefacts in a given corpus (Bevendorff et al. 2019; Murauer and Specht 2021). Most of these papers do not frame the influence of topic, genre/domain or corpus artefacts as a question relating to bias as understood in the fairness literature, but as a threat to the validity of the technique.

To the best of our knowledge, no work has analysed the influence of authors' demographic characteristics on authorship attribution or verification. However, there have been attempts to predict demographic characteristics such as age and gender based on text (Santosh et al. 2013; Piot-Perez-Abadin, Martin-Rodilla, and Parapar 2021). These profiling classifiers generally overfit on the domain they were trained on (Chen, Roth, and Falenska 2024). To facilitate further investigations into the biases of authorship profiling and authorship obfuscation, Emmery et al. (2024) collected a dedicated Reddit dataset (SOBR) for researchers.

Datasets

To better study how a user's demographic characteristics affect authorship attribution made with Reddit data, we collect our own purpose-built dataset. For our experiments, we need a large number of users who self-declare either their native language, their gender or their age with a sufficient amount of long comments for a classification to be viable.

Data Collection

To gather such a dataset at scale, we developed a novel data collection methodology that combines the Wayback API (Archive 2025) and the Reddit API(Reddit 2025a).

The Reddit API only returns the 1k most recent threads for a particular subreddit. However, if queried directly with a *thread_id* that is older than this imposed limit, the Reddit API returns the requested resource. This is the mechanism we use to collect more data by using the Wayback API.

Our collection methodology is illustrated in Figure 1. At a high level, we collect data by first finding subreddits where

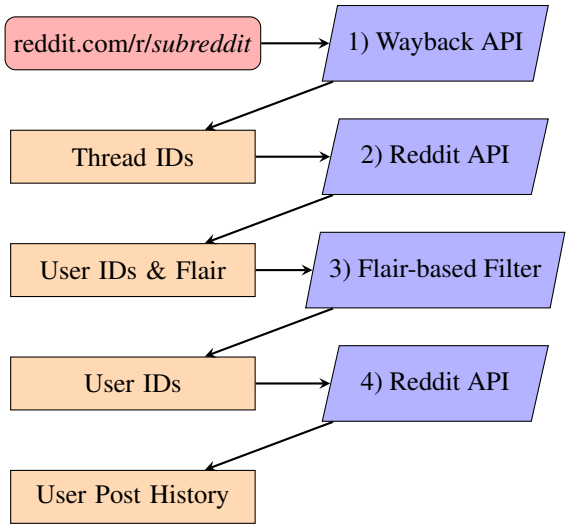

Figure 1: Diagram of our data collection methodology. The left column indicates the input and output of each step, and the right column indicates the process used at each step.

users publicly share their demographic information in their user *flair*, a subreddit-specific tag that is displayed next to the username. Then, we collect as many *thread_ids* from posts in these subreddits using the Wayback API. Using the Reddit API, we subsequently collect all *user_ids* of users who participated in those threads, including their relevant *flair* (e.g., age), and finally, we collect the post history of each user whose flair identified them as being a member of one of the demographic groups we wanted to study. Using this methodology, we collect three different datasets.

Native Language The first dataset consists of users of r/languagelearning, a subreddit dedicated to language learning. We reduce the rather complex flairs available in this community into their native language and, if available, their English skill. In total, we were able to collect 411 authors suitable for our experiments, 161 of whom indicate being native English speakers and 250 of whom indicate being non-native English speakers (see Annex Figure 7). The majority of non-native English speakers indicate a high level of proficiency in English.

Gender The Gender dataset is collected from five advice subreddits where the goal is to ask questions to people of a specific gender (r/askmen, r/askwomen, r/asktransgender, r/askmenover30, r/askwomenover30), two subreddits focused on transgender issues (r/mtf, r/ftm), and two others (r/twoXindia, r/sexover30). Based on the flair conventions in the largest of these subreddits at the moment of collection, we categorised people into three categories: woman, man, and trans*. There is no way of knowing whether somebody choosing a flair was describing their sex or their gender, especially because of varying subreddit conventions; thus, what we call gender here is a proxy variable that could indicate both. To avoid assigning users labels that they would not use to self-describe if based on their flairs, multiple labels could be assigned using our parsing; we remove the users from our dataset. A total of 6,982 users met our data requirement, of which 1,149 describe themselves as "trans*", 2,842 self-identify as "man", and 2,991 self-identify as "woman".

Age We collected the third dataset from seven subreddits focused on generation-related topics (r/generationology, r/BabyBoomers, r/GenX, r/Xennials, r/Millenials, r/Millennials, r/Zillennials). We do not assign a user a generation based only on their participation in that generational subreddit, since users are free to comment in subreddits of different generations. We instead use flair to assign users to a generation, since in all but the Millennial subreddits, the flair indicates the user's year of birth. We run our experiments on a generational basis. We compare users with a year of birth between 1968 and 1972, associated with the Generation X (GenX), to users born between 1999 and 2003, attributed to the Zoomer (GenZ) generation. We chose these two generational subsets because we had more than 160 users for each generation within this timespan, and all inter-generational age differences are larger than the intra-generational ones (see Figure 8).

Data Preprocessing We perform all our experiments at a comment-by-comment level. Before using the comments, we strip them of emojis, Reddit quotations and URLs. Furthermore, we determine the language they are written in using polyglot (Al-Rfou 2015) and only retain comments written in English. We did not have enough authors in our dataset who wrote in a different language to extend our experiments into other languages. In all our experiments, we train with between 4,950 and 5,050 words worth of comments that are at least 128 words long. We based this decision on tests with Random Forest Classifiers, Support Vector Machines and logistic regression classifiers, where our goal was "good-enough" classification with the least training data available and where we had enough authors left in our dataset to repeat experiments with suspect sets size 16 at least 10 times without repeating authors between sets. We use 10 comments of a minimal length of 128 words for testing. To mimic a real-world setting, we use a user's oldest texts for training and use their newer texts for testing.

Alternative Data Sources While many social media datasets contain authorship information, we found no publicly available dataset that was appropriate for our specific task. Most existing datasets lack demographic characteristics, and those that do have them use unreliable proxies, such as first names for gender.

A similar dataset to the one we collected is the SOBR dataset (Emmery et al. 2024), which provides texts labelled by author, gender and age, but not native language. However, we were unable to gain access to it during the course of the project.

As an alternative to the Reddit API, the Pushshift dataset (Baumgartner et al. 2020) could be used to build the dataset. However, this method requires much more processing power and storage space than the method presented in this work.

Classification Scheme

We evaluated a variety of authorship attribution schemes to determine the best-performing model (see Appendix). A reduced version of the *writeprints* features described by Abbasi and Chen (2008) in combination with logistic regression (logR) achieved the best results. This is in alignment with findings by Tyo, Dhingra, and Lipton (2023). The results displayed in the following three experiments use this combination.

However, to corroborate our results, we also ran the experiment with a logistic regression in combination with normalised character-n-grams. We noted no difference in the result, except for a decrease in classifier performance. Furthermore, we tested different machine learning algorithms in combination with the two aforementioned feature sets on the native language dataset, again with the same overarching results presented in the following sections.

We decided against using the current SOTA-classifier Rivera-Soto et al. (2021), as it was trained using Reddit data, and we did not want the presence/absence of certain authors in the training data to influence the fairness audit.

Impact of Demographic Characteristics on Classifier Performance

Experimental Setup

We first assess how the composition of demographic characteristics in the suspect set impacts the performance of authorship attribution.

For this, we vary what we refer to as *suspect set composition*, i.e., the proportion of users with one demographic characteristic (dc) compared to users with another demographic characteristic ($\neg dc$). For each $k \in \{1, ..., n\}$, we run the same experimental setup for all possible suspect sets of size $n \in \{2^2, 2^3, 2^4\}$ of k authors with characteristic dc and $n - k$ authors with characteristic $\neg dc$. Because of the limited size of our datasets and our experiments requiring repetition without an overlap in authors, we were unable to perform experiments with more extensive suspect set sizes.

If the suspect set composition impacts the classification accuracy, we would expect the accuracy to increase or decrease with the number of native speakers in the suspect set. That is, if a suspect set with more, e.g., women in it, performed worse than one with more men in it, we could conclude that the proportion of women in the suspect set does impact classifier performance. If the performance does not decrease or increase, we can conclude that the suspect set composition does not affect the classifier's performance.

We evaluate the relationship between classifier performance and suspect set composition using the first-order polynomial resulting from a linear regression. The latter is computed on the chosen performance metric (accuracy or F1) over the number of authors with characteristic dc in the suspect set. This is illustrated with an example in Figure 2.

We use two metrics to evaluate the linear regression: the mean-squared error (MSE) to assess how well the slope fits the data and the coefficient of determination (R^2) to determine how well the equation explains the observed variation. The closer the former is to zero, the better the slope fits the data; the closer the latter is to one, the better the suspect set composition explains the observed variability in the performance metric. For a given suspect set composition, we generate 10 different suspect sets with no overlap in authors.

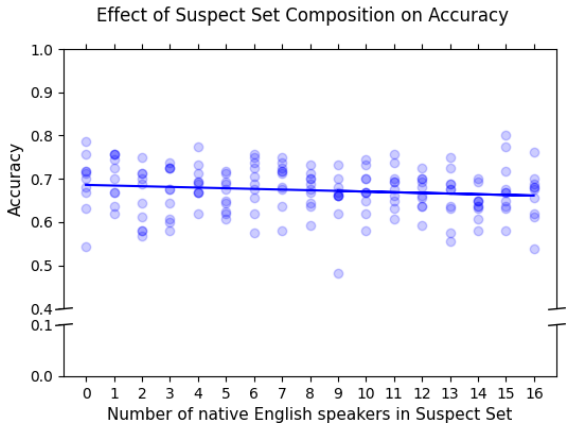

Figure 2: The first-order polynomial predicting a classifier's accuracy based on the native language composition of the suspect set. The points represent the classification accuracy of 10 experiments for each suspect set composition (0 native speakers/16 non-native speakers to 16 native speakers/0 non-native speakers), and the line represents the linear regression result. The slope of the line is almost horizontal. This indicates visually that the performance is not influenced by this type of suspect set composition. Furthermore, the line is well-centred in the point clouds, indicating that it accurately describes the data.

Results Table 1 shows the slopes, y-intercepts, and evaluation metrics describing the first-order polynomials for different suspect set sizes and different demographic factors. The slope, the mse, as well as the R^2, are all close to zero. This indicates that there is little to no influence of age, gender, or native language on classification, as the regression line is consistently horizontal. Furthermore, this horizontal line fits the data well, but there are other variables influencing the variability of the observed data.

Equity of Odds

Experiment Setup

In our second experiment, we determine how a given user's demographic characteristic influences the probability that their text is misclassified. The fairness literature refers to this metric as *equity of odds* or *predictive parity* (Barocas, Hardt, and Narayanan 2023). Fairness is achieved for this metric if all groups of a population have the same probability of having a positive or negative classification outcome (Dieterich, Mendoza, and Brennan 2016).

In our case, a positive outcome is a correct classification, and a negative outcome is a misclassification. We test the equality of odds by comparing the probability distributions of a text of an author being misclassified, given their demographic characteristic, which is noted as: $P(mis|dc) =$

	Experiment Setup				Regression Metrics			
	dc	$\neg dc$		# Users	y-intercept	slope	mse	R^2
Age	GenX	GenZ		4	0.83	-0.02	0.01	0.05
	GenX	GenZ		8	0.82	-0.01	0	0.14
	GenX	GenZ		16	0.76	-0.01	0	0.19
Gender	F	T		4	0.89	-0.01	0.01	0.03
	F	T		8	0.79	0	0	0
	F	T		16	0.72	0	0	0.04
	M	F		4	0.8	0.02	0.01	0.1
	M	F		8	0.78	0	0	0
	M	F		16	0.7	0	0	0
	M	T		4	0.82	0.02	0.01	0.07
	M	T		8	0.8	0	0.01	0.01
	M	T		16	0.73	0	0	0.04
			$\neg$N Lang. Selection					
Native Lang.	N	$\neg N$	Random NL	4	0.86	-0.01	0.01	0.02
	N	$\neg N$	Random NL	8	0.76	0	0.01	0
	N	$\neg N$	Random NL	16	0.69	0	0	0.02
	N	$\neg N$	Shared NL	4	0.84	0	0.01	0
	N	$\neg N$	Shared NL	8	0.77	0	0.01	0.03
	N	$\neg N$	Shared NL	16	0.67	0	0	0.01

Table 1: This table describes the first-order polynomial resulting from a linear regression, where, based on the number of dc in the suspect set (x), we predict the accuracy (y).The compared demographic factors are listed in the columns dc and $\neg dc$. The y-intercept (y-int) describes the mean value for the accuracy [y], whereas the slope describes the steepness of the line. The regression quality is evaluated using the mean squared error (mse) and R^2. The # Users column indicates the number of authors in the suspect set. The $\neg N\ Lang.Selection$ column indicates whether the non-native authors in the experimental setup necessarily share a native language *Shared NL* or not *Random NL*. The slopes of every experiment are near zero, indicating that the suspect set composition has no impact on the classifier's performance.

$P(mis \cap dc)/P(dc)$. In this notation, the fairness criteria translates to $P(mis|dc) \approx P(mis|\neg dc)$. We evaluate whether these distributions are statistically significantly different using a two-sided Wilcoxon rank-sum test as implemented by Virtanen et al. (2020).

The null hypothesis of this test is that the observed samples, in our case, the probabilities, are drawn from the same distribution. If the hypothesis is rejected with a given p-value, it is a clear indication that the underlying distributions are not the same for said threshold. However, if the null hypothesis is not rejected, this only indicates that the underlying probability distribution might be the same, not that it necessarily is. We choose this non-parametric test because it makes minimal assumptions about the underlying probability distribution; for example, it does not require a normal distribution, and it works with small and differently sized sample groups. We tested whether the probabilities we compared are normally distributed using the Shapiro-Wilk test, and while some sample sets passed the test, not all of them did; thus, using a statistical test that requires normality is not an option. The Wilcoxon rank-sum test requires that the compared sets of samples are independent, which we can guarantee. We test this in a setting where the prior probability of the assignment of a demographic characteristic is random chance.

Results

The probability distributions of being misclassified given a specific demographic characteristic dc are largely equivalent between dc and $\neg dc$. This can be seen in Figure 3, where the results of the experiments for gender, native language, and age are displayed. The p-values from the Wilcoxon rank-sum test comparing a $P(mis|dc)$ to $P(mis|\neg dc)$ are above the threshold of 0.05. Thus, the null hypothesis of the samples originating from the same distribution cannot be rejected. This indicates that in our datasets, an author's demographic characteristics do not influence the rate at which their text is misclassified.

Forced-Misclassification

In previous experiments, we only considered misclassifications as they occur in a closed-world setting. In these experiments, errors are relatively rare, and those that do occur are mostly explained by a deviation in the writing style of the

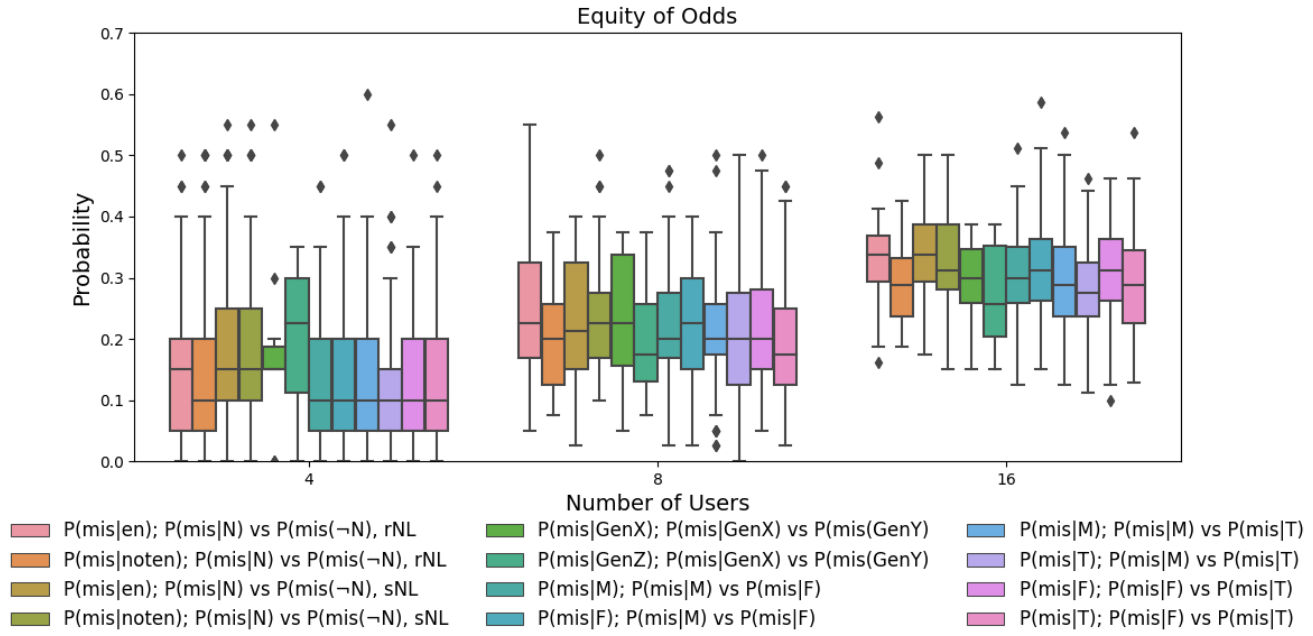

Figure 3: Boxplots showing the probability distributions of being misclassified depending on your demographic characteristic. This figure recombines multiple experiments; the different shades are labelled by the probability followed by the experimental setup. For the native language experiment, we run two experiments: *sNL*, where non-native authors share a native language, and *rNL*, where the non-native authors are randomly chosen.

author. For example, a user might employ a formal tone on r/AskHistorians, but a more informal tone on r/aww, and a more combative tone on r/Politics.

The expected behaviour of supervised machine learning models is that test samples that are close to training samples are correctly classified, while those that deviate from the training data are more likely to be misclassified. To illustrate this point with an example, we chose at random an experimental setup from the previous sections (Male/Female, 4 users) and measured the cosine distance from the mean training vector to each of the testing samples. The closer the cosine distance is to 0, the more similar the two compared vectors are. Unsurprisingly, we find that the misclassified samples have an average distance of 0.95 from the average training vector, and the correctly classified samples are much closer, with an average distance of 0.84.

Because we want to study the influence of confounding demographic factors on mistakes, we eliminate the influence of a user's writing style variability from the following experiment. To this end, we perform it in a setting we refer to as *forced misclassification*. In this setting, the true author of the tested comment is not in the training data of the classifier.

The forced-misclassification setting differs from the open-world setting found in the literature in approach as well as in its intended goal. We do not test the classifier on some labels that are in the training set and some that are not, because we do not want to compare how well the classifier is able to distinguish between known and unknown authors. Instead, we exclusively classify the text of authors not in the training set to investigate how the classifier makes mistakes. We want to establish if users who share a demographic characteristic with the true author write measurably in a similar style, thus systematically get mistaken for each other. To gain insight into the influence of the demographic characteristic on the classification in this setting, we compare the following probability distributions: The probability that an author who shares the demographic characteristic with the true author gets assigned the text $P(dc|dc)$. We refer to this as an intra-status assignment. The probability that an author with a different demographic characteristic is assigned the text, $P(dc|\neg dc)$, we call this an inter-status assignment.

We test this in a setting where the prior probability of the assignment of a demographic characteristic is random chance.

Native Language

Figure 4 displays the probability distributions for 4, 8, and 16 classes. In the experiment setting where the non-native authors are randomly chosen (the green and blue boxplots in Figure 4 and the Wilcoxon rank-sum test results in Table 2a), intra-status assignments of text occur at a higher rate for non-native speakers than inter-status assignments. For native speakers, intra-status assignments occur at higher rates for a suspect set size of 8 and 16 users. For 4 users, the hypothesis that intra- and inter-status assignments occur at the same rate cannot be rejected. The rates of intra-status assignments for native and non-native speakers for a suspect set size of 4 or 8 users are similar. However, for 16 authors,

the distributions are statistically significantly different. The same is true for the rates of inter-status assignment.

For the setup where non-native authors share a native language (the orange and red boxplots in Figure 4 and the Wilcoxon rank-sum test results in Table 2b), the results are different. The rates of intra-status assignments for texts of native and non-native speakers are statistically significantly different, independent of suspect set size. The same is true for inter-status assignments and the comparison between intra- and inter-status assignments for texts of non-native authors. For native authors, intra- and inter-status assignments for texts occur at different rates for a suspect set of 16 authors. For a suspect set size of 4 or 8, the Wilcoxon rank-sum test does not reject the null hypothesis; therefore, the probability distributions might be the same.

In brief, for sock-puppet detection on a subreddit like r/europe, where it is known that non-native authors have a diverse set of native languages, the classifier would behave similarly for native and non-native authors. However, if the classifier is deployed on a subreddit where it is assumed that non-native authors all share a native language, the tendency of the classifier to make intra-status assignments is higher for non-native authors as compared to native authors.

Gender

Intra-status assignments occur at similar rates for different genders. The same is true for inter-status assignments. This can be seen in Figure 5 by comparing $P(M|M)$ to $P(F|F)$ and $P(F|M)$ to $P(F|M)$ respectively. This is confirmed by the Wilcoxon rank-sum test results, where for these comparisons the null hypothesis is not rejected (see Table 2c).

For both women and men, intra-status assignments occur at statistically significantly higher rates than inter-status assignments. I.e., text from a woman has a higher likelihood of being assigned to another woman rather than to a man. However, this only holds for a suspect set size of 16 if we use a less strict cut-off for statistical significance of p-value < 0.05 (see Table 2c).

Age

Intra-status assignments occur at similar rates for different generations. The same is true for inter-status assignments. This can be seen in Figure 6 and is confirmed by the Wilcoxon rank-sum test results (see Table 2d).

When comparing the intra- and inter-generational assignments, the results differ depending on the number of users (see Table 2d). For a suspect set size of 16 authors, the intra- and inter-generational assignments occur at different rates. The text has a statistically significant tendency to be assigned to an author of the same generation. For a suspect set size of 8 authors, the same is true, except with the caveat that for authors belonging to GenZ, this is only the case with the less strict significance cut-off p-value < 0.05. In contrast, for the suspect set size of 4, this pattern of statistical significance is no longer given.

Discussion

In a closed-world setting, our results do not indicate any influence of demographic factors on authorship attribution. In our first experiment, we studied the impact of the demographic composition of the training data on classifier performance. Our results show that the diversity of the training set, with regard to gender, age, or native language, does not impact the performance of authorship attribution. Therefore, when building an authorship attribution-based classifier, for example, to link sockpuppet accounts on Reddit, there is no performance incentive to consider the demographic makeup of the authors in the training set. It performs equally well regardless of the demographic makeup of the suspect set. However, these results also indicate that demographic traits do not provide any natural cover for any users. For example, native English speakers are just as easy to deanonymise as non-native English speakers.

In the second experiment, we study the equity of odds: whether a user's demographic trait affects the rate at which their text is misclassified. Our results indicate that the weight of misclassification is not disproportionately carried by one group. In other words, no particular demographic group that we studied would be disproportionately impacted by model errors in automated moderation using authorship attribution *in the closed-world setting*. At the same time, these results also indicate that such demographic factors cannot aid in building defences against authorship attribution attacks on forms of legitimate anonymity (e.g., whistleblowing). They also signal that everyone is equally vulnerable to an attack against their anonymity.

In our third experiment, we study mistakes made in the *forced misclassification* setting. We present the classifier with text from users not in the training data, then measure the rate at which text is attributed to users who share the demographic trait of interest with the true author. This allows for an observation of the influence of the demographic characteristic on the classification result while controlling for an author's variability in writing style. This setting also mimics the deployment of closed-world classifiers in an open-world setting.

In this experiment, we observe a statistically significant tendency to assign text to users who share demographic characteristics with the true author of the text. For non-native authors, this is dependent on the experimental setup; if the users in the training data have a diverse set of native languages, this effect is less pronounced than if all non-native authors share a native language 2c.

In the context of using authorship attribution, this means that false positives can be unequally distributed among users who share a demographic trait with the malicious account. This result is especially relevant in the context of social networks, where the suspect pool is vast and, thus, mistakes in choosing the suspect set are easy to make.

Overall, our tests indicate that authorship attribution is fair and robust in closed-world settings with regard to the tested demographic characteristics. This may be because text misclassified in a closed-world setting is, by definition, a deviation from the author's normal writing style. Therefore, their demographic trait might not have as much of an impact on the classification. However, when we force the classifier to make an error by removing the true author from the training set, we control for a given text sample's close-

	Number of Users		
	4	8	16
$P(N\|N)$ **vs** $P(\neg N\|N)$	<0.05	<0.01	<0.01
$P(\neg N\|\neg N)$ **vs** $P(N\|\neg N)$	<0.01	<0.01	<0.01
$P(\neg N\|\neg N)$ **vs** $P(N\|N)$	>0.05	>0.05	<0.01
$P(N\|\neg N)$ **vs** $P(\neg N\|N)$	>0.05	>0.05	<0.01

(a) **Native Language**: Non-English-native suspects do not necessarily have the same native language

	Number of Users		
	4	8	16
$P(N\|N)$ **vs** $P(\neg N\|N)$	<0.05	<0.05	<0.05
$P(\neg N\|\neg N)$ **vs** $P(N\|\neg N)$	<0.01	<0.01	<0.01
$P(\neg N\|\neg N)$ **vs** $P(N\|N)$	<0.01	<0.01	<0.01
$P(N\|\neg N)$ **vs** $P(\neg N\|N)$	<0.01	<0.01	<0.01

(b) **Native Language**: Non-English-native suspects all share the same native language

	Number of Users		
	4	8	16
$P(M\|M)$ **vs** $P(F\|M)$	<0.01	<0.01	<0.05
$P(F\|F)$ **vs** $P(M\|F)$	<0.01	<0.01	<0.05
$P(F\|F)$ **vs** $P(M\|M)$	>0.05	>0.05	>0.05
$P(F\|M)$ **vs** $P(M\|F)$	>0.05	>0.05	>0.05

(c) **Gender**

	Number of Users		
	4	8	16
$P(X\|X)$ **vs** $P(Z\|X)$	<0.05	<0.01	<0.01
$P(Z\|Z)$ **vs** $P(X\|Z)$	>0.05	<0.05	<0.01
$P(X\|X)$ **vs** $P(Z\|Z)$	>0.05	>0.05	>0.05
$P(X\|Z)$ **vs** $P(Z\|X)$	>0.05	>0.05	>0.05

(d) **Generation**: X = GenX, Z = GenZ

Table 2: Summary of the Wilcoxon rank-sum test results. The column on the left lists the compared probability distributions. The values describe whether the p-values are below the indicated threshold. The *Users* column lists the number of authors in the training set.

ness to the author's average writing style. In this setting, we find that there is a statistically significant influence of the demographic characteristic. Thus, while in an ideal testing setup, there might be no demonstrable bias when using basic fairness metrics, this does not guarantee that there is an absence of bias in the underlying mechanism of the classification.

Threats to Validity

Self-reported labels and proxy labels Self-reported labels are useful because they allow a user to self-identify their demographics. However, they come with two notable drawbacks. First, nothing prevents users from lying in their flairs, thus poisoning our dataset. Second, as is made clear in our gender label, some terms can have multiple meanings. For instance, words some people use for gender are used by others to describe their sex. When we see the flairs, we cannot be sure which way the user meant the flair; thus, we are only able to measure a proxy variable.

Generalizability Reddit's user demographics are not representative of the population at large, and the text written on Reddit follows platform and subreddit-specific conventions. While we control for one demographic characteristic at a time, the other demographic characteristics in our experiments are influenced by Reddit´s user base and more specifically by the subreddit where the text is published (Cinus et al. 2025). Furthermore, we only perform our experiments on English text; it is unclear how the results change in other languages.

Selection Bias Our experimental setup requires authors to have at least 6,280 words worth of comments written in English. This criterion was met by users who are active in a variety of subreddits over a relatively long period of time (on average, one to two years). Thus, our suspect sets are mainly composed of long-term, active Reddit users. Especially in the native language dataset, our preprocessing greatly impacts the types of users selected; due to the high bar for the minimum amount of text required, the users in this dataset have high English language skills.

Intentional Changes of Writing Style If the author uses language correction tools, translation services or writing from generative AI, we assume that they do so consistently. Thus, their "measured writing style" can be impacted by this tool usage, but because we assume the use of the tools is consistently impacting their "writing style", this should not impact our classification greatly. If a user decides to mimic someone else's writing style, adopts very specific in-group speak in certain circumstances, or otherwise greatly deviates from their personal norm, we do not have any mechanism in place to detect this behaviour change. We assume that all comments made from a specific user account are equally representative of the user's writing style. Furthermore, we assume that Reddit accounts are not shared, or if they are, that the demographic attribute mentioned in the flair applies to every person using the account.

Intersectionality There exists an overlap in users between our datasets. However, it is very small, and the distribution of demographic characteristics of this intersection does not allow for experiments analysing the influence of multiple characteristics at a time. Based on our results, we cannot make any statements about which demographic characteristic has a stronger influence or what their combined influence looks like. We look to future work to address this issue.

Conclusion

In this work, we explored the ways in which demographic characteristics of the suspects impacted the results of an authorship attribution task. Using Reddit data, we found that none of the demographic attributes we tested — age, gender, and native language — had a statistically significant influence on the overall performance of the classifier. In the closed-world setting, we found the classifier to be "fair",

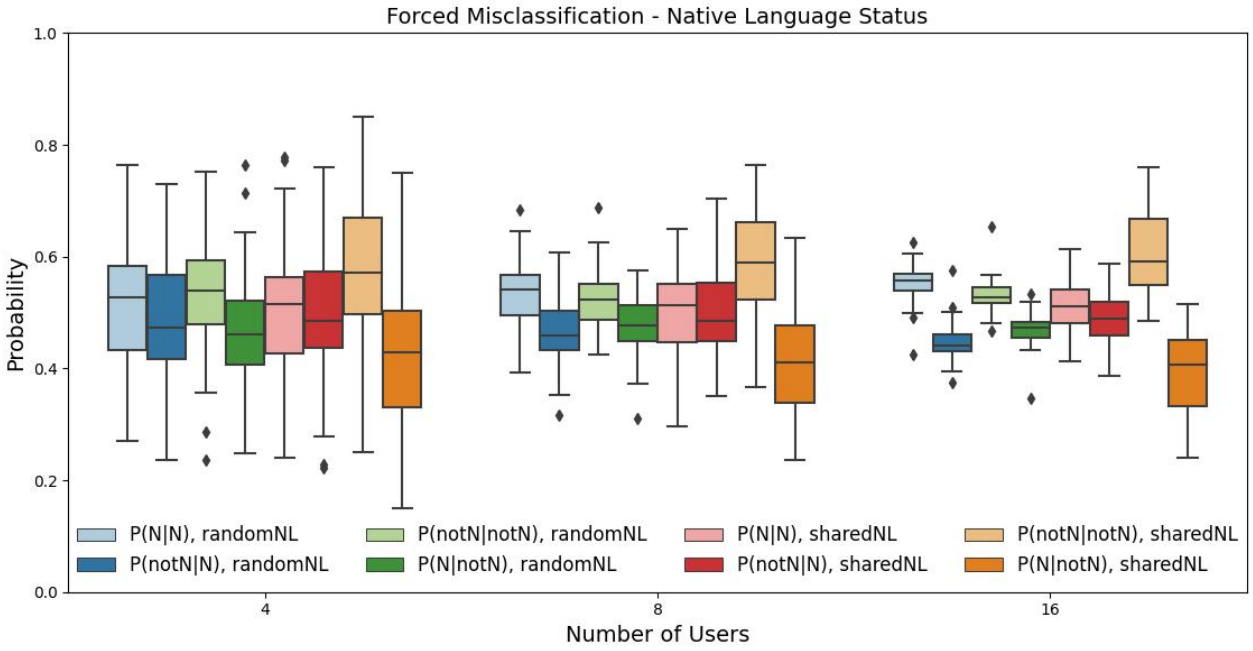

Figure 4: Boxplots displaying the probability distribution of intra- and inter-native language assignment for different native languages. N stands for native English speaker, $\neg N$ stands for non-native English speaker. In the experiment setting $randomNL$, non-native authors are randomly chosen, as opposed to the experiment setting $sharedNL$, where non-native authors have a common native language.

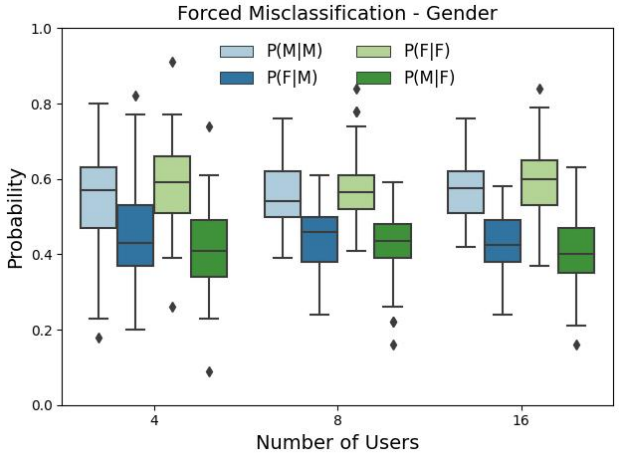

Figure 5: Boxplots of displaying the probability distributions of intra- and inter-gender assignments for different genders.

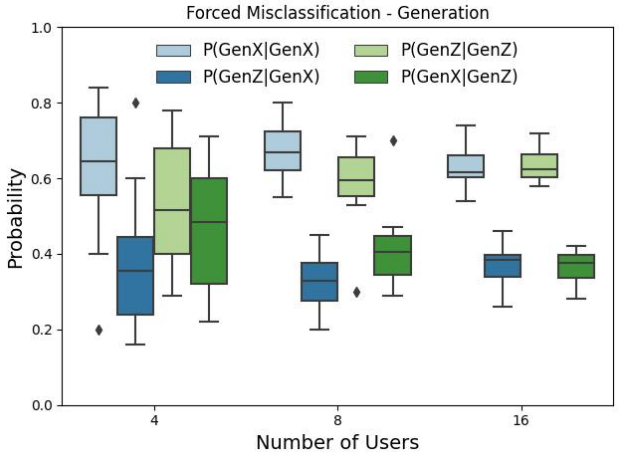

Figure 6: Boxplots of displaying the probability distributions of intra- and inter-generation assignment.

satisfying the equity of odds across all experiments. However, if texts of unknown-to-the-closed-world-classifier authors are classified, we found a statistically significant tendency to attribute text to authors with the same demographic characteristic as the unknown author. This not only reveals underlying biases of the classification, but also simulates the deployment of a closed-world classifier in an open-world setting. These results show that fairness results obtained in clean, lab settings must be interpreted cautiously. While our closed-world results indicated that the classifier is not biased, the more realistic setting in which the true author is not present demonstrates bias. This shows there might be hidden biases in the model that cannot be seen under ideal conditions. As such, bias should be tested for not only under ideal conditions before deployment, but under realistic conditions and regularly in deployment to uncover latent biases.

References

Abbasi, A.; and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*.

Afroz, S.; Islam, A. C.; Stolerman, A.; Greenstadt, R.; and McCoy, D. 2014. Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy*. IEEE.

Al-Rfou, R. 2015. *Polyglot: A massive multilingual natural language processing pipeline*. Ph.D. thesis, State University of New York at Stony Brook.

Al-Rfou', R.; Perozzi, B.; and Skiena, S. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Alonso-Fernandez, F.; Belvisi, N. M. S.; Hernandez-Diaz, K.; Muhammad, N.; and Bigun, J. 2021. Writer Identification Using Microblogging Texts for Social Media Forensics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Arabnezhad, E.; La Morgia, M.; Mei, A.; Nemmi, E. N.; and Stefa, J. 2020. A Light in the Dark Web: Linking Dark Web Aliases to Real Internet Identities. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*.

Archive, I. 2025. Wayback Machine APIs. https://archive.org/help/wayback_api.php last visited on: 2025-09-14.

Barlas, G.; and Stamatatos, E. 2020. Cross-domain authorship attribution using pre-trained language models. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*, chapter 3. MIT Press.

Bartelds, M.; and de Vries, W. 2019. Improving Cross-domain Authorship Attribution by Combining Lexical and Syntactic Features: Notebook for PAN at CLEF 2019. In *CEUR Workshop Proceedings*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In Choudhury, M. D.; Chunara, R.; Culotta, A.; and Welles, B. F., eds., *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, 830–839. AAAI Press.

Bevendorff, J.; Hagen, M.; Stein, B.; and Potthast, M. 2019. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Boettcher, N. 2021. Studies of depression and anxiety using reddit as a data source: scoping review. *JMIR mental health*.

Brown, D. K.; Ng, Y. M. M.; Riedl, M. J.; and Lacasa-Mas, I. 2018. Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations. *Social Media+ Society*.

Chen, H.; Roth, M.; and Falenska, A. 2024. What Can Go Wrong in Authorship Profiling: Cross-Domain Analysis of Gender and Age Prediction. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics.

Chi, Y.; and Chen, H.-y. 2023. Investigating substance use via Reddit: systematic scoping review. *Journal of Medical Internet Research*.

Cinus, F.; Monti, C.; Bajardi, P.; and Morales, G. D. F. 2025. Uncovering the Sociodemographic Fabric of Reddit. arXiv:2502.05049.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*.

Discord. 2025. Discord Identity Authenticity Policy. https://discord.com/safety/identity-authenticity-policy-explainerlast last visited on: 2025-09-14.

Emmery, C.; Miotto, M.; Kramp, S.; and Kleinberg, B. 2024. SOBR: A Corpus for Stylometry, Obfuscation, and Bias on Reddit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Holtgraves, T.; and Robinson, C. 2020. Emoji can facilitate recognition of conveyed indirect meaning. *PloS one*.

Khan, J. A.; Khan, N. D.; Yaqoob, M.; Yasin, A.; and Alwadain, A. 2024. Exploring reddit forum for software evolution as an alternative requirements source: An end-user discussion dataset on Google maps. *Data in Brief*.

Khemani, B.; and Adgaonkar, A. 2021. A review on reddit news headlines with nltk tool. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

Kireev, K.; Mykhno, Y.; Troncoso, C.; and Overdorf, R. 2025. A Telegram Dataset of Propaganda and its Moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Murauer, B.; and Specht, G. 2021. Developing a Benchmark for Reducing Data Bias in Authorship Attribution. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Piot-Perez-Abadin, P.; Martin-Rodilla, P.; and Parapar, J. 2021. Gender classification models and feature impact for social media author profiling. In *International Conference on Evaluation of Novel Approaches to Software Engineering*. Springer.

Raszewski, L.; and de Kock, C. 2025. Detecting Sockpuppetry on Wikipedia Using Meta-Learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Reddit. 2025a. Reddit API. https://www.reddit.com/dev/api/ last visited on: 2025-09-14.

Reddit. 2025b. Reddit Inc Privacy Page. https://redditinc.com/privacy last visited on: 2025-09-14.

Rivera-Soto, R. A.; Miano, O. E.; Ordonez, J.; Chen, B. Y.; Khan, A.; Bishop, M.; and Andrews, N. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Sakib, M. N.; and Spezzano, F. 2022. Automated Detection of Sockpuppet Accounts in Wikipedia. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Santosh, K.; Bansal, R.; Shekhar, M.; and Varma, V. 2013. Author profiling: Predicting age and gender from blogs. *Notebook for PAN at CLEF*.

Sari, Y.; Stevenson, M.; and Vlachos, A. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Shakir, U. 2025. Telegram version 9.2 adds SIM-free anonymous phone numbers, new auto-delete. *The Verge*.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*.

Tyo, J.; Dhingra, B.; and Lipton, Z. C. 2023. Valla: Standardizing and Benchmarking Authorship Attribution and Verification Through Empirical Evaluation and Comparative Analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

Wang, A.; Aggazzotti, C.; Kotula, R.; Soto, R. R.; Bishop, M.; and Andrews, N. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*.

Weerasinghe, J.; Singh, R.; and Greenstadt, R. 2022. Using Authorship Verification to Mitigate Abuse in Online Communities. *Proceedings of the International AAAI Conference on Web and Social Media*.

Weninger, T.; Zhu, X. A.; and Han, J. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*.

Xu, W.; and Fung, B. C. M. 2025. StyleLink: User Identity Linkage Across Social Media with Stylometric Representations. *Proceedings of the International AAAI Conference on Web and Social Media*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Authorship Attribution can be used, for example, for deanonymising accounts and linking accounts. Both of these applications can be used either to protect platform integrity or to harm the users. Understanding how authorship attribution works and its limits helps create better protections against it, if it is used as a mechanism to harm people, as well as guide the interpretations in its legitimate applications. If the technique is found to be too biased, this research can be used as a justification to limit the use of authorship attribution in applications with far-reaching consequences.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? We specifically choose subreddits and authors based on their demographic characteristics, thus the subset of authors we study does not necessarily reflect either the demographic composition of Reddit at large or the global population. However, we do acknowledge in the discussion that we only control for one demographic characteristic at a time, and that the other demographic characteristics are influenced by Reddit's demographic abnormalities.
   (e) Did you describe the limitations of your work? Yes, we took care to make the limitations clear in the dedicated section *Threats to Validity*
   (f) Did you discuss any potential negative societal impacts of your work? We address explicitly that the technique studied can be used to cause harm.
   (g) Did you discuss any potential misuse of your work? We mention both in the introduction as well as in the discussion, that the study's technique has both beneficial and malicious use cases.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymisation, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, in the methodology section of the "Equity of Odds" Experiment, we have a paragraph introducing the statistical test used
   (b) Have you provided justifications for all theoretical results? We do not have theoretical results
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? We do not have theoretical results
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? We use hypothesis testing to explore the influence of a third variable on the result.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, this work focuses on measuring such biases. Further, we explicitly outline potential biases in the Threats to Validity Section.
   (f) Have you related your theoretical results to the existing literature in social science? We relate it to other interdisciplinary work studying bias on authorship attribution, but not specifically to existing literature in social science.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? We relate our results to their real-world implications for pseudonymous social networks, in the introduction and the discussion.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? We do not include theoretical proofs
   (b) Did you include complete proofs of all theoretical results? We do not include theoretical proofs
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We released our code prior to publication
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? There are descriptions in the experimental setups, the appendix and the code is thoroughly commented to permit reproducability
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, in particular, we rerun the experiments on multiple author splits (not random seeds)
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? In the Annex is a paragraph, "Resources Required"
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? We specifically study/evaluate the classification outcome with respect to a third variable. We explain the logic of our approach in the corresponding "Experimental Setup" sections.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, this is what motivated us to start working on this paper.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? We reference the previous work we used
   (b) Did you mention the license of the assets? We reference the different libraries we use, but not the licence under which they are released. As for the dataset, we collected it on our own.
   (c) Did you include any new assets in the supplemental material or as a URL? To comply with the Reddit API we only publish a guideline on how to recollect the data. The created dataset will be available upon request, if the data sharing is compliant with the Reddit API.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? In the ethics section in the Appendix, we outline that we collected our dataset in compliance with the Reddit api and that users agree in the terms of service, that their text is available to request using the Reddit API.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Reddit is a pseudonymous network, and users are aware that their posts are publicly accessible. Which is why we didn´t additionally anonymise the data we collect. We also do not control whether the text posts we collect are offensive.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? The dataset we collected is inherently non-compliant with the FAIR principles, as the reddit API doesn´t allow us to publish the dataset.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, but we will not provide it at this stage, because this would jeopardize the anonymous submission.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Not applicable
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable
   (d) Did you discuss how data is stored, shared, and deidentified? Not applicable

Data Collection

Detailed Explanation Figure 1 illustrates our data collection methodology.

Step 1: Query the Wayback API to find thread_ids We query the Wayback API to get all of the visible posts from a subreddit for each snapshot stored by The Wayback Machine. These were parsed to find the `thread_ids` for the threads seen on that page.

Step 2: Query the Reddit API to find those threads Using this list, we query the Reddit API to find which users commented in the thread and their `flair_text`.

Step 3: Find relevant users by parsing their flair From these users, we used a parser to determine, depending on the dataset, either the native language, gender, or age of each user based on the indications in their flair. The flair is a free-text field filled out by the user, and, as such, there is the possibility for variety in these entries. Depending on the subreddit in question, parsing thus can be challenging.

Step 4: Query the Reddit API to collect the post histories of relevant users With these usernames, we query the Reddit API again to obtain the entire post history of each user. This results in a database of usernames, their flair on the subreddit of interest, and their entire post histories.

Generalizability Using the data collection methodology presented here, a combination of the Wayback Machine API and the Reddit API can be reused in further studies to collect a more expansive Reddit dataset. Our collection methodology could be utilised to collect Reddit data for many different types of studies in which older Reddit data or data from more than the 1k most recent posts is needed, but archives such as Pushshift are not available. However, this strategy has one caveat: the sample of posts acquired does not necessarily follow a specific strategy, nor is it guaranteed to be a representative slice of an activity in a given subreddit. The crawls completed by `archive.org` are done by multiple independent crawlers that follow different, changing crawling strategies. This results in a dataset that lacks a discernible overall collection strategy. For example, in January 2023, `reddit.com/r/languagelearning` was archived 10 times, in March it was archived 15 times, and in August 2023 it was archived just once. If crawling coverage is of importance, this strategy is not the most suitable. For the present paper, this caveat is not important because we are only looking to find more users with flair and are not concerned with studying the content of the subreddit itself.

Dataset Preprocessing

Before we can analyse the Reddit comments we collected, we must handle the fact that they are text collected "in the wild" and are therefore not clean and ready for analysis. Before cleaning the dataset, we remove all users with fewer than 30 comments. This threshold was conservatively chosen with the goal of not running every comment through the expensive cleaning process.

We developed a three-step process to clean the dataset. First, we retracted elements from the text that were not relevant to our task, including non-text elements and quoted text. In the second step, we detected language of the comment and only retained English-language texts. Finally, we discarded comments that are shorter than 128 words after the pretreatment.

Removing URLs, Emojis, and Cited Text For the first step, we removed URLs from the comments because URLs are not reflective of someone's writing style. In contrast, emojis convey meaning (Holtgraves and Robinson 2020). That is, a writer can choose among different emojis that share a similar meaning, making this choice part of their writing style. However, we choose to remove emojis from our text to prevent information leakage, such as certain emojis taking on different meanings in different languages. As an example, the cabbage emoji can be used in French to replace the word 'cute' because cabbage and cute are homophones in French. Finally, we also removed quoted text from the comments. Reddit has a specific quotation style delineated with $>$, for users to cite other posts or quote text. We remove all of these quotes to avoid mixing the 'writing styles of users. Other types of quotes were not removed.

Potential Problem in our text pre-processing While we have retracted Reddit quotes from the text, there is no guarantee that we have retracted all quoted text. Our preprocessing does not account for quotes that are not marked in the Reddit quotation style or for unmarked copy-pasted text. We also do not distinguish between edited Reddit comments and unedited Reddit comments.

Filtering Non-English Texts We then used the language detection module polyglot (Al-Rfou', Perozzi, and Skiena 2013) to assess the language of each comment. Polyglot offers an indication of confidence in its assessment. Only comments assessed with a confidence of 99 or 100 out of 100 were added to the dataset. Furthermore, we manually validated the assessment of 100 randomly chosen comments per language for languages that at least one of the authors was proficient enough in to label (German, English, French, Spanish, and Italian). We found no false positives for this confidence level. However, this does not guarantee that our dataset is only composed of English text. In particular, comments that contain multiple languages may pass the language filtering step if the majority of the beginning of the comment is in English. All examples of these mixed texts we came across originated from the r/languagelearning subreddit itself. Thus, we removed all text from this subreddit from our dataset to minimize the amount of non-English text. This does not guarantee that no mixed-language comments are in the dataset.

Minimal comment length In order to detect writing style, a minimum amount of text is necessary. The feature extraction results of a two-word comment, such as 'Thank you', is limited and conveys little of an author's habitual linguistic choices. As most of our comments are very short, we needed to determine a minimum comment length where we could start to measure 'writing style' consistently. We tested

minimum comment length requirements of 16, 32, 64, 128, and 256 words. We determined that a minimum comment length of 128 words was sufficient to achieve consistent classification results on the Native Language Dataset by testing different minimal comment lengths concerning the achieved accuracy. The longer the minimal comment length, the better the accuracy. With a minimal comment length shorter than 128 words, the achieved accuracy varies greatly depending on the authors in the suspect set. For example, for a cut-off at 64 words certain suspect set compositions only achieved random-chance performance, whereas others achieved an accuracy of 0.81 for a suspect set size of 8 authors. Due to the limited data available, we did not choose the cut-off value where the classifier performance stopped increasing, but rather the cut-off value where the classification performance was no longer highly dependent on the suspect set composition. In other words, we did not choose the minimal length of comment in regard to achieving the highest performance. We chose the first tested comment length, where we achieved a stable classification result, e.g. for different suspect sets we achieve similar accuracies.

Train-Test Split Before classification, we split the data into training and testing sets. To mimic a real-world setting, we use a user's oldest texts for training and use their newer texts for testing. Because the users in our dataset have different time frames of activity, this cut-off point between training and testing is set on a per-user level, and we do not have one global point in time where we use all text before as training data and all text after as testing data.

Dataset Description

Native Language

Our native language dataset has a majority of native English speakers (see Figure 7). The second most prevalent group is German speakers, followed by Spanish and French speakers. All languages for which we could construct suspect set sizes of size 16 are Indo-European.

Generation

The years of birth chosen to represent Generation Y are between 1968 and 1972. There are roughly the same number of authors in each year of birth. Generation Z is represented by users born between 1999 and 2003 in our dataset, where the majority of users are born in 1999 (see Figure 8).

Dataset Access

In compliance with the Reddit API, we cannot publicly share the collected data. If you are interested in our dataset, please contact us directly.

Authorship Attribution Comparison

Feature-Sets For our experiment, we compared three different feature extraction methods: normalized char-1,3-grams, a reduced version of the *writeprint* feature set as well as sentence embeddings. The style_featureset (see Table 3) differs from the first implementation by Abbasi and

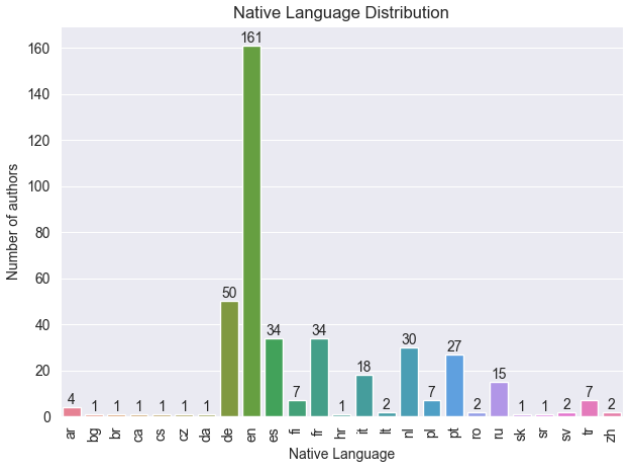

Figure 7: Native Language Distribution of authors for whom we collected at least 6280 words worth of text. Native English speakers (en) represent the majority. Most non-native speakers have an Indo-European first language.

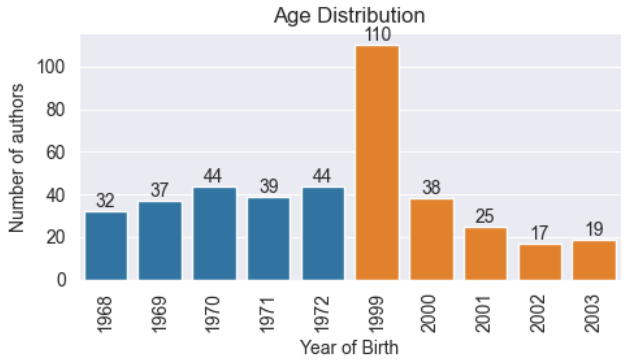

Figure 8: Distribution of authors of Generation Y (in blue) and Generation Z (in orange) for whom we collected at least 6280 words worth of text.

Chen (2008). We only consider features that are comparable between languages and apply to Reddit comments. Thus, structural features were retracted as well as features that are difficult to compare between languages, such as misspelled words. Our feature set is not language-independent, as we need to know which language the text is in before extracting the features; however, the extracted features should be comparable for Indo-European languages. For the sentence embeddings, we initially wanted to use the state-of-the-art LUAR (Rivera-Soto et al. 2021) sentence embedding. However, it is trained on Reddit data, and there seems to be an overlap between their training data and our dataset. Because the goal of our study is to measure the influence of third variables, and not data contamination, we instead chose to use a more general model where we made sure it was not

Feature	*#features*
Lexical features	
Average length of words	1
Median length of words	1
Distribution of word length	1
total number of characters	1
char-n-grams (1-grams to 3-grams)	1000
Content-related features	
word-n-grams (1-grams to 3-grams)	1000
Yules-K (vocabulary richness)	1
Type-to-Token-Ratio (vocabulary richness)	1
Syntactic Features	
pos-n-grams (1-grams to 3-grams)	1000

Table 3: List of features composing the style_feature set. This is a reduced version of the feature-set proposed by Abbasi and Chen (2008)

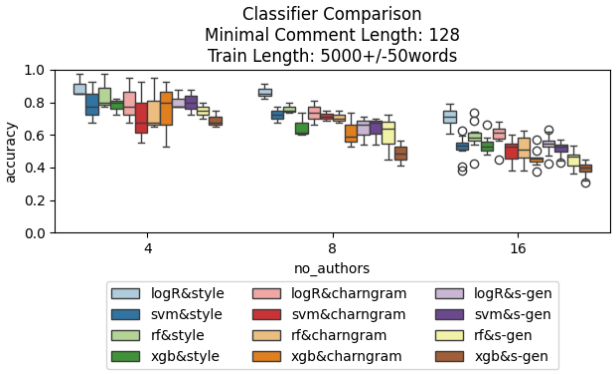

Figure 9: Comparison of classifiers, made with between 4950 and 5050 comments worth of training data. Trained and evaluated with comments that are at least 128 words long.

fine-tuned using Reddit data[1].

Classifiers We compared random forest (rf), XGBoost (xgb), support vector machines (svm) as well as logistic regression (logR) classifiers.

Amount of Training Data The goal was to create an authorship attribution classifier that performed relatively well and with consistent performance. The less training data we require, the more users were able to be included in our experiment. The more training data available, the better the authorship attribution model performs. We decided that 5000 words were the best trade-off between the number of users available for the experiment and the performance of the resulting classifiers.

Comparison Figure 9 shows clearly that the style features in combination with a logistic regression outperform the other classification schemes. This is in accordance with the results of Tyo, Dhingra, and Lipton (2023) who found that logistic regression outperforms transformer-based models if less data is available.

Resources Required

We ran our experiments locally on a Windows computer with 32 GB of RAM, a 128 MB Intel(R) Graphics Card, and Intel(R) Core(TM) Ultra 7 165U (1.70 GHz) Processor.

Ethics

This study only uses public data provided by the Wayback and Reddit APIs. Reddit data collected through the API has been extensively used in prior works (Khemani and Adgaonkar 2021; Weninger, Zhu, and Han 2013; Boettcher 2021; Khan et al. 2024; Chi and Chen 2023). We comply with the Terms of Service of both APIs that we utilize, including respecting rate limits. In general, Reddit accounts are not tied to real-world identities, and we make no attempt to link user accounts to real people. We will also only release the dataset on request with sanitized usernames. However, for reproducibility, all code used in this project is available at $https://github.com/JWYSS2/UnFairMistakes.git$.

[1]specifically this one: https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1 (last accessed: 14.09.2025)