Mode Effects' Challenge to Authorship Attribution

Haining Wang[1], **Allen Riddell**[1], and **Patrick Juola**[2]

[1]Indiana University Bloomington, Bloomington, Indiana
[2]Duquesne University, Pittsburgh, Pennsylvania
{hw56|riddella}@indiana.edu juola@mathcs.duq.edu

Abstract

The success of authorship attribution relies on the presence of linguistic features specific to individual authors. There is, however, limited research assessing to what extent authorial style remains constant when individuals switch from one writing modality to another. We measure the effect of writing mode on writing style in the context of authorship attribution research using a corpus of documents composed online (in a web browser) and documents composed offline using a traditional word processor. The results confirm the existence of a "mode effect" on authorial style. Online writing differs systematically from offline writing in terms of sentence length, word use, readability, and certain part-of-speech ratios. These findings have implications for research design and feature engineering in authorship attribution studies.

1 Introduction

That authorship attribution techniques work as reliably as they do has been attributed to the fact that each individual has a distinctive writing style. Texts written by the same author can be recognized by analyzing lexical and syntactic features in documents (Juola, 2006). This principle is practically successful in a variety of settings (Abbasi and Chen, 2008; Overdorf and Greenstadt, 2016; Afroz et al., 2014). In some cases, however, authorial style is challenging to detect. For example, authorial style fades as time goes by (Glover and Hirst, 1996; Baayen et al., 2002), varies considerably in collaborative environments (Graham et al., 2005; Kestemont et al., 2018; Zangerle et al., 2019), and drifts depending on document genre (Stamatatos, 2018; Koppel et al., 2007; Sapkota et al., 2016).

The aforementioned changes are conspicuous due to the fact that there are certain markers indicating that a document may have been written in a fashion that will lead to stylistic variation. For instance, the presence of genre-specific words indicates a document may present a puzzle for standard authorship attribution techniques. Sometimes, however, documents which may challenge an analysis of authorial style can be unannounced. This paper shows that authorial style changes with respect to sentence length, word use, readability, and certain part-of-speech ratios when the writing environment switches from traditional word processing software to an input box of a web browser.

2 Mode Effects

Originally developed in survey research and educational testing, a "mode effect" describes the following phenomenon: a respondent may answer the same question differently depending on how a survey is administered (e.g., online vs. phone) (Hochstim, 1967; Leeson, 2006). Although discussion continues about mode effects' underlying mechanism (Kreuter et al., 2008; Sidi et al., 2017), contextualized magnitude (Carpenter and Alloway, 2019; Washburn et al., 2017), and adjustment methods (Kolenikov and Kennedy, 2014), a consensus has been reached that mode effects can impair survey validity. For instance, Tourangeau et al. (2000) compiled six studies investigating illicit drug use with self-administered and interviewer-administered surveys. The results showed that illicit drug use was reported at higher rates when questions were administered without an interviewer present.

Functionally, writing modality resembles survey modality: the style observed in an individual's writing may vary depending on how the writer composes the document. A document written by hand may vary from a document composed using traditional word processing software. Further variation may be observed if the document is typed into a

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1146–1155
April 19 - 23, 2021.

text box in a web browser. Therefore, this research uses "mode effect" to label such differences.

3 Data

Participants in this experiment were recruited on Amazon Mechanical Turk (MTurk). Two distinct types of writing were collected from each of the 18 participants: (1) ca. 6,500 words of pre-existing formal writing and (2) a short ca. 500-word open-ended response to an essay prompt. For the pre-existing writing samples, participants were asked to "Submit at least 6500 words total from multiple documents of your own writing that was done for a formal purpose (school essays, grant proposals, etc)." For the 500-word essay, respondents were asked to describe [their] neighborhood to someone who has never been there before as part of a college application.[1] Respondents also completed a demographic questionnaire, reporting their gender and age bracket.

Responses that were not in English or which seemed very likely to be inauthentic were excluded. (Kennedy et al. (2018) discusses challenges dealing with MTurk surveys). We also excluded one response which appeared to contain writing copied (without attribution) from online sources. The pre-existing writing samples were further processed in order to remove personally identifying information. Lengthy quotations, headings, tables, and figures were also removed.

The data for this experiment are a subset of data collected as part of research seeking to replicate results in Brennan et al. (2012). In the full replication experiment, respondents were randomly given one of four essay prompts. In this paper, we only used the responses by respondents randomly assigned to the "control" condition. These respondents provided pre-existing writing samples and a response to the essay prompt mentioned above. Responses were collected between March 29th and June 1st, 2019. 14 out of 18 respondents reported their age as "18-34". Self-reported gender was also collected. Ten of the respondents were men and eight were women.

The overwhelming majority of pre-existing writing samples collected were essays written for undergraduate courses. Many essays discussed films and literary works. Many appeared to be written for political science and business courses.

The writing prompt generally elicited the desired response: respondents wrote about their neighborhood using formal or semi-formal prose.

To check that all writing collected exhibited approximately the same degree of formality, we compared the formality of the writing in the offline corpus with the formality of the writing in the online corpus using a formality score developed by Heylighen and Dewaele (1999). We found that the formality scores in each corpus were similar.[2]

4 Method

This paper focuses on stylistic differences introduced by seemingly innocuous variation in the mode used to enter a text—offline composition vs. online typing into a text box. We are interested in whether a mode effect occurs in writing. To the extent that it is observed, we wish to know if its impact on an author's style is predictable. Does writing mode induce similar changes in the writing style of different individuals?

To answer these questions, we extract linguistic features from the documents written by the 18 participants. We then use a Bayesian hierarchical model to estimate differences in the rates at which the linguistic features appear in texts written using different writing modes.

5 Modeling Authorial Style

5.1 Feature Selection

We use a set of high-level, familiar linguistic features in our study. Our "Comparative Style" ("CS") feature set aims to capture word-, sentence-, and chunk-level features. All features are described in Table 1. For sentence-level features, white space between words is not counted as a character. Punctuation includes periods, exclamation marks, question marks, commas, semicolons, colons, and apostrophes. For function words we use the list of 512 words from Koppel et al. (2005). The Voice of America (VOA) Special English word list contains 1,512 frequently-used words which are used

[1] The full prompt reads: "TOPIC: You are asked as part of a college application to describe your neighborhood to someone who has never been there before. Discuss the houses, people, stores, parks. Anything you think is relevant." The prompt is taken from Brennan et al. (2012).

[2] The mean formality score and standard deviation for the offline corpus were 59.2 and 17.3 respectively. The mean formality score and standard deviation for the online corpus were 57.9 and 17.6 respectively. The formality scores were calculated for each sentence using the following formula: $(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} + \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} - \text{adverb freq.} - \text{interjection freq.} + 100)/2$ (Heylighen and Dewaele, 1999).

Table 1: This table describes features in the Comparative Style feature set. The last column mentions an example sentence and an example text chunk. The first sentence of the abstract is the example sentence. The abstract is the example text chunk.

Level	No.	Feature	Abbreviation	Explanation/Example
Word	1	Word length in syllables	WordLenSyll	E.g. the word "mode" has one syllable while the word "effect" has two.
	2	Word length in characters	WordLenChar	E.g. the word "mode" has four characters while the word "effect" has five.
Sentence	3	Sentence length in syllables	SentLenSyll	E.g. the example sentence has 35 syllables.
	4	Sentence length in words	SentLenWord	The example sentence has 16 words.
	5	Sentence length in characters	SentLenChar	E.g. the example sentence has 99 characters.
	6	Punctuation to character ratio	PuncChar	E.g. the PuncChar ratio of the example sentence is 1/99.
	7	Function word to word ratio	FuncWord	E.g. the FuncWord ratio of the example sentence is 6/16.
	8	Special English ratio	SplEng	E.g. the SplEng ratio of the example sentence is 6/16, because "on", "the", "to", "individual," and two "of" are in the sentence.
	9	Common word ratio	CommWord	E.g. the CommWord ratio of the example sentence is 12/16, because "success", "on", "the", "presence", "linguistic", "feature", "specific", "to", "individual", "author," and two "of" are in the sentence.
Chunk	10	Adjective to noun ratio	AdjNoun	E.g. the AdjNoun ratio of the abstract is 18/44.
	11	Verb to noun ratio	VerbNoun	E.g. the VerbNoun ratio of the abstract is 16/44.
	12	Pronoun to noun ratio	PronNoun	E.g. the PronNoun ratio of the abstract is 2/46.
	13	Adverb to adjective ratio	AdvAdj	E.g. the AdvAdj ratio of the abstract is 3/18.
	14	Flesch–Kincaid grade level	FleschKincaid	E.g. the FleschKincaid score of the abstract is 13.11.
	15	Gunning fog index	GunningIdx	E.g. the GunningIdx ratio of the abstract is 50.00.

in VOA Special English reporting (Voice of America, 2007). The 8,013-word "common word" list is taken from the College English Test Band 4 and 6 (CET-4/6), the nationwide English proficiency test used in mainland China. The lists are used in the three sentence-level measures of vocabulary richness.

A chunk is defined as a sequence of consecutive sentences containing at least 150 words. The criterion "150 words" was arbitrarily chosen to balance meeting the length requirement of readability tests and the desire to extract as many chunk-level observations as possible in order to better estimate feature variability within writing modes. We experimented with different chunk lengths (e.g., 100, 150, 200) and found that our results did not depend strongly on chunk length.

The Flesch–Kincaid grade level Flesch (1948) and Gunning fog index (Gunning, 1968) were calculated for every chunk. More challenging texts are associated with lower Flesch-Kincaid levels and higher Gunning fog indexes.

Thirteen of the 15 CS features were transformed by taking the square root so that the feature distributions would be approximately Gaussian. The Flesch-Kincaid level and the Gunning fog index are left on the original scale as their distributions were already approximately Gaussian. Although the hierarchical model uses the transformed features, in subsequent visualizations and tables, parameter estimates are reported on the original scale.

5.2 Setup

We divide the documents into two groups: the "offline" documents, the pre-existing writing samples from the 18 subjects (authored using word processing software) and the "online" documents, written in a web browser in response to the essay prompt asking for a description of the writer's neighborhood.

To compare features across modes and individuals, we use a hierarchical model. Within each mode-specific group of documents, feature observations associated with an individual are modeled using a normal sampling model with an individual-specific mean and scale. The individual-specific means and scales are, in turn, modeled using a normal distribution and gamma distribution.

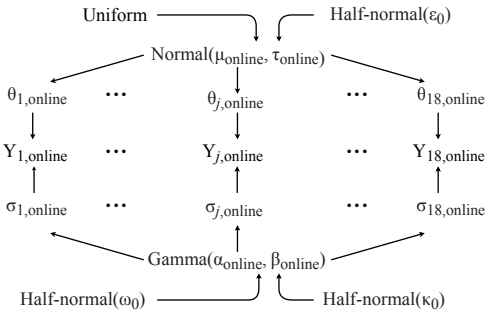


Figure 1: The model for CS feature observations for the online documents. The model for the CS feature observations for the offline documents is the same.

Figure 1 shows the hierarchical model for observations of CS features for the documents composed online. The model can be rendered in symbols as

$$\begin{aligned}
\theta_{j,\text{online}} &\overset{i.i.d.}{\sim} \text{Normal}(\mu_{\text{online}}, \tau_{\text{online}}) \\
\sigma_{j,\text{online}} &\overset{i.i.d.}{\sim} \text{Gamma}(\alpha_{\text{online}}, \beta_{\text{online}}) \\
Y_{j,\text{online}} &\overset{i.i.d.}{\sim} \text{Normal}(\theta_{j,\text{online}}, \sigma_{j,\text{online}}), \\
&\quad j \in \{1, 2, ..., 18\}
\end{aligned}$$

where $Y_{j,\text{online}} = \{y_{1,j}, y_{2,j}, ..., y_{n_j,j}\}_{\text{online}}$ are observations for the jth subject in the online mode. These observations are drawn from a normal sampling distribution $\text{Normal}(\theta_{j,\text{online}}, \sigma_{j,\text{online}})$. The individual-specific standard deviation $\sigma_{j,\text{online}}$ comes from a gamma distribution parameterized by shape α_{online} and rate β_{online}. The individual-specific mean $\theta_{j,\text{online}}$ is drawn from a normal distribution with a location μ_{online} and a scale τ_{online}. In addition, the μ_{online} was assigned a uniform prior distribution, while weakly informative priors were given to τ_{online}, α_{online}, and β_{online}. Each feature is modeled separately. The models for CS features in the offline mode mirrors those for the online mode.

In a pilot study, we considered using a Student-t distribution instead of a normal distribution as the sampling distribution. We found that the estimated degrees of freedom for these sampling distributions were sufficiently large (>30). Hence we concluded it was safe to use the simpler normal distribution as the sampling distribution.

Additionally, for simplicity, η, the group-level scale coming from $\text{Gamma}(\alpha, \beta)$, will be reported rather than the original parameterization. The effect size between modes was calculated with $\frac{\mu_{\text{online}} - \mu_{\text{offline}}}{\sqrt{(\eta^2_{\text{online}} + \eta^2_{\text{offline}})/2}}$ (Kruschke, 2014).

5.3 Decision Rule

Posterior differences between the two group means (μ_{online} and μ_{offline}) and group scales (η_{online} and η_{offline}) will be characterized using 95% highest posterior density (HPD) intervals and regions of practical equivalence (ROPE) (Kruschke and Liddell, 2018). The 95% HPD interval describes an interval in which a parameter is likely to be found. The ROPE specifies a region of practical equivalence around a null value.

When the 95% HPD for a parameter falls outside the ROPE, the null value is rejected, and the parameter is considered to be different from the null value. If the HPD falls entirely inside the ROPE, the null value is accepted. Otherwise, we withhold judgment.

The data can be analyzed using different ROPEs. Given the goals of this investigation, we use ROPEs associated with a "small" effect (Cohen's d=0.2, according to Cohen (1988)), that is, calculating ± 0.1 standard deviations for every posterior difference as the upper and the lower ROPE limits around zero, as suggested by Kruschke (2018). There are many ways to calculate the effect size. We follow Kruschke (2014) in our calculation and refer to this effect size as Cohen's d.

6 Results

Before summarizing differences between online and offline writing across all individuals and features, we first consider how a single feature varies in writing from the 18 subjects. This analysis demonstrates how we use Bayesian methods to infer collective tendencies in the data.

6.1 A Close Look at Sentence Length

Do sentence lengths vary systematically by writing mode? We consider three measures of sentence length, one of which is "sentence length in characters." Even before performing any modeling, the individual and collective tendencies can be identified in a visualization of the data (Figure 2).

The box plots on the left-hand side of Figure 2 show that all but one of the 18 participants tended to use longer sentences in their offline documents. The variability in sentence lengths also tended to be greater. The right-hand side histograms of aggregated observations further confirm this characterization.

Figure 3 shows the 95% HPD intervals for the parameters of interest related to "sentence length in

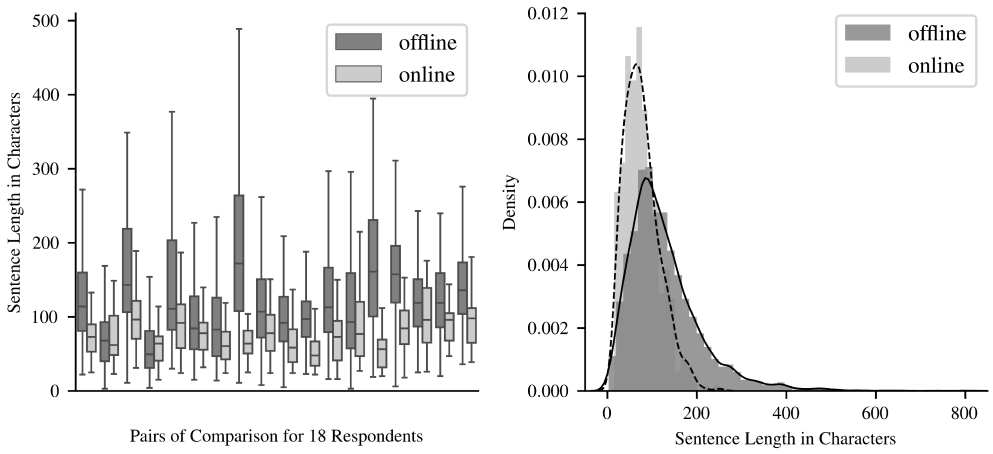


Figure 2: The grouped box plots and histograms with kernel density estimates for sentence length in characters.

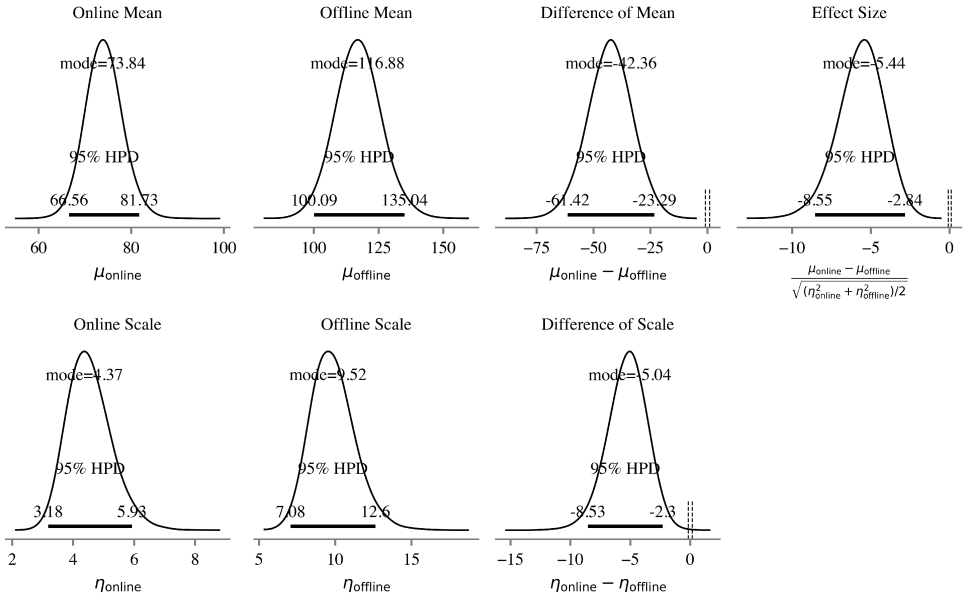


Figure 3: Posterior distribution of μ_{online}, $\mu_{offline}$, $\mu_{online} - \mu_{offline}$, η_{online}, $\eta_{offline}$, $\eta_{online} - \eta_{offline}$, and effect size for the "sentence length in characters" feature. The ranges between two dotted reference lines are ROPEs.

characters", μ_{online}, $\mu_{offline}$, η_{online}, $\eta_{offline}$, $\mu_{online} - \mu_{offline}$, $\eta_{online} - \eta_{offline}$, and effect size.

In the first two upper panels, posteriors indicate that individuals' sentences were typically 73.84 characters when typing into a text box in a web browser but were typically 116.88 characters with traditional word processing software. Those writing online tended to use shorter sentences (42.36 characters fewer, $\mu_{offline} - \mu_{online}$). Note that the 95% HPD falls far away from the ROPE, indicating a non-negligible difference. Therefore, we conclude that people wrote shorter sentences when writing online.

The lower panels of Figure 3 show another difference: the standard deviation in the online setting was 5.04 characters fewer than that in the offline mode, indicating a relative lack of variability in sentence length when individuals wrote online.

The estimated effect size was -5.44 with a 95% HPD interval between -8.55 and -2.84 (Figure 3 upper-right). A effect size of greater than 2 (in absolute value) counts as "huge" (Sawilowsky, 2009). One way of comprehending the magnitude of an effect size is the following: with the naked eye, one can barely detect a "small" effect (e.g., Cohen's d=0.2) but would have no difficulty in seeing a "large" one (e.g., Cohen's d=0.8).

The preceding analysis looked closely at a specific feature. We considered both the raw data and posterior estimates. In the remainder of the

paper, for the sake of brevity, only differences between group means ($\mu_{\text{online}} - \mu_{\text{offline}}$), scales ($\eta_{\text{online}} - \eta_{\text{offline}}$), and effect sizes will be reported.

6.2 How Writing Style Varies by Mode

In Figure 4, 12 out of 15 posterior differences of group means ($\mu_{\text{online}} - \mu_{\text{offline}}$) are credibly nonzero, leaving the rest undecided.

The posterior differences between word lengths and sentence lengths are negative. Participants tend to write shorter sentences and use shorter words in the online condition. Relative to offline writings, a positive difference in mean Flesch-Kincaid levels and a negative difference in mean Gunning fog index scores indicate that individuals simplify their writing style when they are entering prose in the web browser. Similar patterns also appear when examining the percentages of function words, Special English, and common words. Individuals tended to use simpler vocabulary in the online condition. For ratios of parts of speech, the adjective to noun and pronoun to noun ratios show credibly positive differences.

Differences in feature standard deviations are shown in Figure 5, where five features indicated nonzero differences. Sentence length, measured in three different ways, varies less in online writing than in offline writing. That is, individuals tend to use a wider range of sentence lengths in offline writing than in online writing. Two readability scores also show less variations in the online mode.

Effect size. The posterior distributions for effect size are shown in Figure 6. Eleven out of 15 features have nonzero effect size. One effect size counts as "medium", five count as large "large", one counts as "very large," and four count as "huge" (using levels defined in Cohen (1988) and Sawilowsky (2009)). Measures regarding word length both manifest "large" effect sizes, and features related to sentence length all have "huge" effect sizes. Likewise, the Flesch-Kincaid reading ease and Gunning's fog index differences are "huge" and "very large." Function words and special English show "large" effects while common words display "medium" effects. This confirms that individuals used simpler words online. The pronoun to noun ratio is the only part-of-speech ratio that shows a credibly nonzero effect (a "large" effect).

7 Mode Effects and Authorship Attribution Accuracy

Another way to understand the magnitude of mode effects is to check if standard authorship attribution techniques have a harder time identifying an individual's writing when presented with the same individual's writing composed in a different mode. That is, we can compare the rate at which an authorship attribution model identifies the correct individual when presented with an unsigned document written offline with the rate at which the model identifies the correct individual when given a document written offline. (The model is trained using writing samples written in the offline mode.) This approach has the virtue of allowing us to answer the question we began with: Does the mode effect make authorship attribution more difficult?

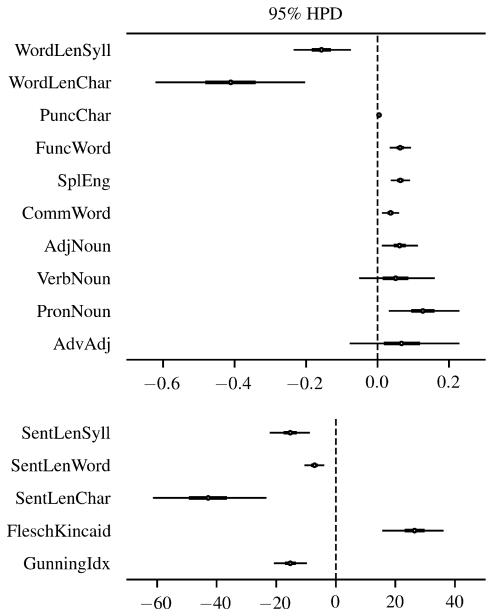
For this experiment, we use two authorship attribution models featured in Brennan et al. (2012). One is extremely basic, making use of nine features and a simple feed-forward neural network. The second model uses a larger feature set ("Writeprints Static") and a support vector machine classifier with a linear kernel.[3] Authorship attribution accuracy is calculated in the following steps:

1. Choose a random subset from the 18 authors, starting from two and gradually increasing.

2. Calculate authorship attribution accuracy with five-fold cross-validation using the authorship attribution model (SVM or neural network) with pre-existing writings.

[3] We re-engineered the Writeprints Static" and the "Basic-9" feature sets. Our re-implementation of the Writeprints-Static feature set mirrored the original set with three exceptions. First, we applied another widely-used function word list (Koppel et al., 2005) (including 512 words in total) in lieu of the original word list because we could not locate the original list. Second, we used the Penn Treebank tagset (39 non-punctuation tags in NLTK 3.4.5) rather than the original maximum entropy tagset (22 tags). We expect to see very minor differences between the two implementations. Third, we used a linear kernel instead of the polynomial kernel mentioned in Brennan et al. (2012) because the linear kernel performed far better. Indeed, we suspect Brennan et al. (2012) may have used a linear kernel (despite reporting having used a polynomial kernel). A subsequent paper, Overdorf and Greenstadt (2016), which shares an author with Brennan et al. (2012) describes the Brennan et al. (2012) as having used SVM with a linear kernel.

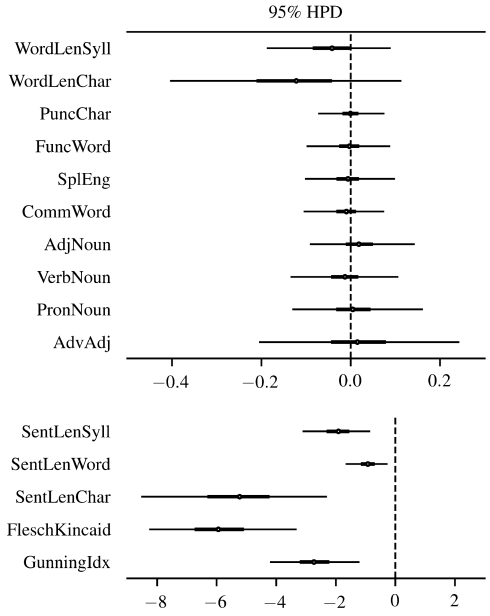
The SVM with a linear kernel used a maximum iteration of 100,000. The neural network used half of the sum of the author count and feature size (which is nine) as the hidden layer count, 100 neurons per layer, and a maximum iteration of 100. All experiments were performed using Scikit-Learn 0.22.1.



Difference of Means ($\mu_{\text{online}} - \mu_{\text{offline}}$)

Feature	Estimate (Posterior Mode)	95% HPD	ROPE	Decision
WordLenSyll	-0.16	(-0.235, -0.074)	(-0.004, 0.004)	Reject Null
WordLenChar	-0.41	(-0.621, -0.202)	(-0.011, 0.011)	Reject Null
PuncChar	0.00	(-0.001, 0.009)	(0.000, 0.000)	Undecided
FuncWord	0.06	(0.034, 0.094)	(-0.002, 0.002)	Reject Null
SplEng	0.07	(0.037, 0.091)	(-0.001, 0.001)	Reject Null
CommWord	0.04	(0.012, 0.061)	(-0.001, 0.001)	Reject Null
AdjNoun	0.06	(0.012, 0.113)	(-0.003, 0.003)	Reject Null
VerbNoun	0.05	(-0.051, 0.161)	(-0.005, 0.005)	Undecided
PronNoun	0.13	(0.031, 0.230)	(-0.005, 0.005)	Reject Null
AdvAdj	0.07	(-0.078, 0.229)	(-0.008, 0.008)	Undecided
SentLenSyll	-15.10	(-22.174, -8.712)	(-0.342, 0.342)	Reject Null
SentLenWord	-7.11	(-10.557, -3.872)	(-0.170, 0.170)	Reject Null
SentLenChar	-42.40	(-61.418, -23.287)	(-0.968, 0.968)	Reject Null
FleschKincaid	26.70	(15.538, 36.181)	(-0.517, 0.517)	Reject Null
GunningIdx	-15.30	(-20.844, -9.693)	(-0.281, 0.281)	Reject Null

Figure 4: 95% HPD, ROPE, and Decisions for Differences of Group Mean ($\mu_{\text{online}} - \mu_{\text{offline}}$)



Difference of Scales ($\eta_{\text{online}} - \eta_{\text{offline}}$)

Feature	Estimate (Posterior Mode)	95% HPD	ROPE	Decision
WordLenSyll	-0.03	(-0.188, 0.090)	(-0.007, 0.007)	Undecided
WordLenChar	-0.11	(-0.404, 0.114)	(-0.013, 0.013)	Undecided
PuncChar	0.00	(-0.073, 0.075)	(-0.004, 0.004)	Undecided
FuncWord	0.00	(-0.099, 0.088)	(-0.005, 0.005)	Undecided
SplEng	-0.01	(-0.102, 0.099)	(-0.005, 0.005)	Undecided
CommWord	-0.01	(-0.105, 0.075)	(-0.005, 0.005)	Undecided
AdjNoun	0.02	(-0.091, 0.144)	(-0.006, 0.006)	Undecided
VerbNoun	-0.01	(-0.135, 0.107)	(-0.006, 0.006)	Undecided
PronNoun	0.01	(-0.131, 0.161)	(-0.008, 0.008)	Undecided
AdvAdj	0.01	(-0.205, 0.243)	(-0.011, 0.011)	Undecided
SentLenSyll	-1.85	(-3.118, -0.848)	(-0.058, 0.058)	Reject Null
SentLenWord	-0.91	(-1.667, -0.260)	(-0.036, 0.036)	Reject Null
SentLenChar	-5.04	(-8.532, -2.301)	(-0.160, 0.160)	Reject Null
FleschKincaid	-6.06	(-8.262, -3.314)	(-0.126, 0.126)	Reject Null
GunningIdx	-2.79	(-4.212, -1.202)	(-0.076, 0.076)	Reject Null

Figure 5: 95% HPD, ROPE, and Decisions for Differences of Group Scale ($\eta_{\text{online}} - \eta_{\text{offline}}$)

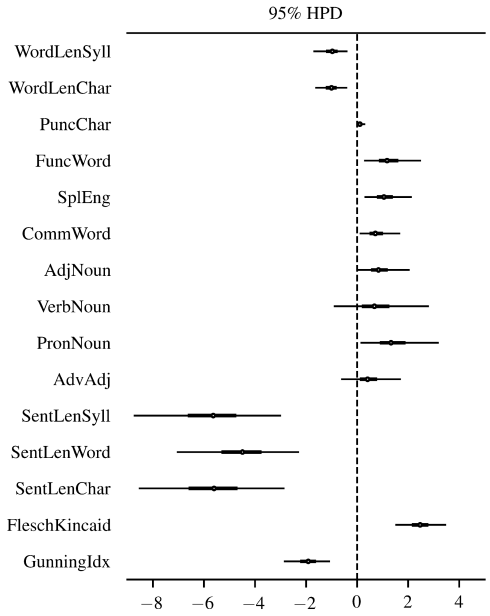
3. Repeat the previous steps 1,000 times using a different author subsets. Calculate the average accuracy over these replications.

Calculating attribution accuracy for the online writing samples follows similar steps. Step 2 differs. The model is trained on pre-existing documents and asked to predict the authorship of a document written online.

Figure 7 shows our results. It is clear that the accuracy suffers when applying offline-writing-trained classifiers to online writings. That is, changes in authorial style are big enough to confuse the classifiers. Writing mode differences make authorship attribution more difficult.

8 Discussion

This study investigated whether individuals' writing style varies by "mode": Does the mode used to



Effect Size

Feature	Estimate (Posterior Mode)	95% HPD	Decision	Most Credible Level
WordLenSyll	-0.90	(-1.715, -0.377)	Reject Null	Large
WordLenChar	-0.95	(-1.640, -0.391)	Reject Null	Large
PuncChar	0.09	(-0.028, 0.323)	Undecided	-
FuncWord	0.99	(0.274, 2.505)	Reject Null	Large
SplEng	0.95	(0.289, 2.146)	Reject Null	Large
CommWord	0.62	(0.103, 1.692)	Reject Null	Medium
AdjNoun	0.72	(0.029, 2.061)	Undecided	-
VerbNoun	0.61	(-0.915, 2.810)	Undecided	-
PronNoun	1.17	(0.133, 3.202)	Reject Null	Large
AdvAdj	0.35	(-0.629, 1.720)	Undecided	-
SentLenSyll	-5.44	(-8.750, -2.978)	Reject Null	Huge
SentLenWord	-4.31	(-7.061, -2.273)	Reject Null	Huge
SentLenChar	-5.44	(-8.552, -2.84)	Reject Null	Huge
FleschKincaid	2.49	(1.492, 3.487)	Reject Null	Huge
GunningIdx	-1.88	(-2.870, -1.059)	Reject Null	Very Large

Figure 6: 95% HPD and the most credible levels for effect sizes ($\frac{\mu_{\text{online}}-\mu_{\text{offline}}}{\sqrt{(\eta^2_{\text{online}}+\eta^2_{\text{offline}})/2}}$). The ROPEs are set with (-0.1, 0.1) associated with Cohen's "small" effect (d=0.2).

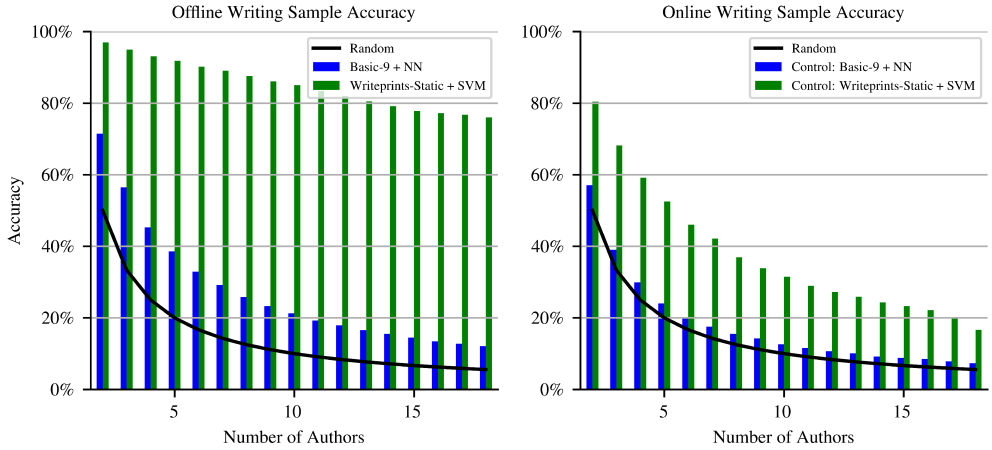


Figure 7: Authorship attribution accuracy by writing mode. The authorship attribution model is trained on pre-existing writing samples which were composed offline.

compose a document (word processor (offline) versus web browser text entry (online)) affect measurements of individuals' writing style? Our findings confirmed the existence of mode effects. In online writing, respondents tend to use shorter sentences, shorter words, more adjectives (relative to nouns) and pronouns (relative to nouns). Sentence lengths exhibit lower variability as well.

Therefore, we suggest authorship attribution researchers should exercise caution when dealing texts written using different modes.

For example, in Brennan et al. (2012), the authors attributed lower accuracy in an authorship attribution task to the fact that writers employed authorship attribution circumvention techniques. Our research suggests that this lower accuracy may be due in part to differences in writing mode. (The circumvention techniques were only used in online writing.)

Systematic differences in writing associated with different modes may complicate a broader range of experiments. Just as researchers appropriately anticipate genre-dependent stylistic differences in individuals' writing (e.g., fiction vs. non-fiction

prose), experiments should also anticipate mode-dependent differences.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1814425. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy*, pages 212–226. IEEE.

Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6th JADT*, volume 1, pages 69–75.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22.

Rachel Carpenter and Tracy Alloway. 2019. Computer versus paper-based testing: Are they equivalent when it comes to working memory? *Journal of Psychoeducational Assessment*, 37(3):382–394.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Angela Glover and Graeme Hirst. 1996. Detecting stylistic inconsistencies in collaborative writing. In *The new writing environment*, pages 147–168. Springer.

Neil Graham, Graeme Hirst, and Bhaskara Marthi. 2005. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4):397–415.

Robert Gunning. 1968. *The technique of clear writing*, revised edition edition. McGraw-Hill, New York, N.Y.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4.

Joseph R Hochstim. 1967. A critical comparison of three strategies of collecting data from households. *Journal of the American statistical Association*, 62(319):976–989.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. 2018. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, pages 1–16.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at pan-2018: Cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25.

Stanislav Kolenikov and Courtney Kennedy. 2014. Evaluating three approaches to statistically adjust for mode effects. *Journal of survey statistics and methodology*, 2(2):126–158.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(Jun):1261–1276.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217. Springer.

Frauke Kreuter, Stanley Presser, and Roger Tourangeau. 2008. Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public opinion quarterly*, 72(5):847–865.

John Kruschke. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

John K Kruschke. 2018. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

John K Kruschke and Torrin M Liddell. 2018. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206.

Heidi V Leeson. 2006. The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1):1–24.

Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, 2016(3):155–171.

Upendra Sapkota, Thamar Solorio, Manuel Montes, and Steven Bethard. 2016. Domain adaptation for authorship attribution: Improved structural correspondence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2226–2235.

Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599.

Yael Sidi, Maya Shpigelman, Hagar Zalmanov, and Rakefet Ackerman. 2017. Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction*, 51:61–73.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press.

Voice of America. 2007. *VOA Special English word book: a list of words used in Special English programs on radio, television, and the Internet.* Voice of America, Washington, D.C. OCLC: 761196573.

Shannon Washburn, James Herman, and Randolph Stewart. 2017. Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams. *Advances in physiology education*, 41(4):548–555.

Eva Zangerle, Michael Tschuggnall, Günther Specht, M Potthast, and B Stein. 2019. Overview of the style change detection task at pan 2019. In *CLEF*.