Naive Bayes Networks predict authorship by Pronoun-Frequency, establishing novel trends in Contemporary Language

Michael W Roffo[1] and Prof. Steven Harris[1]

[1] *University of Massachusetts at Amherst, Massachusetts, United States of America*

July 14, 2020

In the burgeoning field of Literary Statistical Analysis, Machine Learning techniques reveal numerical trends in literary works, opening doors towards insights not available to previous generations. We compile literary works of diverse styles including: canonized poet Robert Frost, "Instagram" poet Rapi Kaur, and English songwriter Dodie Clarke. Then we implement a Naive Bayes Classifier that uses the relative frequency of particular pronouns to predict authorship. We remark on how each author's style accounts for observed pronoun-frequency. By observing cases of false predictions, we note remarkable stylistic parallel between authors typically considered disparate. Finally, even the most accurate models occasionally confuse Frost with Kaur or Dodie, perhaps helping legitimize a new school of contemporary poetry, one raised by social media.

1 Introduction & Methodology

In the 2018 book, *Macroanalysis: Digital Methods for the Humanities*, Matthew Jockers, PhD[1] shows how recent advances in computer algorithms make it possible to statistically measure—with respect to a mathematical model—qualities of literary works that were previously considered subjective. Using an Artificial Intelligence technique called Supervised Learning, we attempt to describe an author's style, theme, or even nationality in terms of choice characteristics, or "features", that describe their writing. Some features' usefulness seems self-evident, while others seem insignificant yet can be used reliably to distinguish poets from the literary canon, from poets of a new school.

One key cultural artifact of the Instagram era is its impact, not just on how we access written media, but also on the media itself. For example, consider the rising popularity of so-called "Instagram poetry"—often simply an observation about relationships, but with line-breaks. Rupi Kaur, for examples, writes—always in all-lowercase—"you have to stop / searching for why at some point / you have to leave it alone".

For ages, such literary debates have ended in stalemate because we consider them a matter of personal taste. But in Jockers' groundbreaking book, he suggests why some questions in literature may not be so subjective as we thought. One might object, "What does a mathematical model understand about literature?" Another may answer: whereas a scholar may study

[1] Associate Professor of English, University of Nebraska-Lincoln. Fellow, Center for Digital Research in the Humanities.

for a week and still guess authorship incorrectly, these statistical measurements enable machines to predict authorship, author-gender, and even nationality with high accuracy in mere seconds. So maybe, through their "distant reading" perspective, computers can understand literary works in ways that humans cannot. As a result, maybe they can guide our intuition about what really constitutes poetry, and especially, what surprising features might distinguish poets from one another.

2 Corpus

Like any scientific paper, we must introduce a dataset to support our explorations. Any digital humanities research paper presents a dataset called a "corpus"—a body of texts, indexable, and often enhanced with labels or "tags" for particular grammatical structures. This formal data structure for a text allows researchers to extract key statistics from a text using computer algorithms. Again: through such "distant reading", we gain insights into a text that a "close reading" human could not detect.

For our particular question, we consider a corpus of poems by poets Robert Frost, Rupi Kaur, and Dodie Clark. These texts were variously web-scraped and written into distinct text files for each author. Then, each author file was read into a python script which splits the text at some key character which indicates the end of one poem and the beginning of another. Finally we are left with a directory of text-files, each named according to author and id number. We include this corpus alongside this paper.

3 Techniques

We analyze this corpus using several standard textual-analysis techniques and measures. A key metric in style analysis is "pronoun-frequency", a measure of how often in a poem an author tends to use a particular pronoun. As Jockers convincingly argues in Macroanalysis, this deceptively-simple measurement is surprisingly informative for distinguishing one author from another. Initializing in the Python programming language the WordFreq object from the Natural Language ToolKit (NLTK [2]), we define the function

```
get_fdist(fileid)
```

, which takes the fileid of a particular poem, and returns NLTK's WordFreq object. Next a

```
pronoun_feature_extractor(fileid, pronouns)
```

function takes a particular poem's fileid along with a list of prospective pronouns, and returns data describing the frequency of each given pronoun. Next, we run each poem through the feature extractor in order to represent each poem in terms of only the frequencies of the given pronouns. Now we have a list of all the poems, but with their titles removed, and only distinguishable where these pronoun frequencies differ.

Table 1: *Tested Accuracy of Each Classifier*

Classifier	Accuracy
Dodie / Frost	84.61%
Kaur / Frost	98.15%
Dodie / Kaur	86.44%

Finally we borrow from the Machine Learning literature a statistical technique called "supervised classification": given enough examples of some pronoun-frequency object paired with its correct author, can a computer infer what distinguishes the authors in terms of such pronoun-frequency data? In particular, can it correctly classify a poem it has not seen before, based solely on that new poem's pronoun-frequency data? To find out, we again leverage the NLTK. First we run all poems in our corpus through our feature-extractor, then pair each pronoun-frequency set with its author-label. We separate four fifths of the features-label pairs into one set—namely the "training set"—and withhold the last fifth for the "test set". We withhold the test set because properly assessing the accuracy of our classifier demands that we test it on examples it has not been "trained" on. In other words, if it correctly labels the author of poems it has never seen before, we begin to speak of such a machine as "intelligent". Perhaps it may even reveal textual insights that human close-reading cannot perceive! Having a training set and a test set, we finally run the python code 'classifier = nltk.NaiveBayesClassifier.train(trainingset)'. [3] Note that a Naive Bayes Classifier is simply a common variety of supervised classifier which is built into NLTK and easily adaptable to any dataset that we may provide.

4 Analysis of Results

We introduce three such classifiers, report their tested accuracy, then reflect on their outputs. In particular we have a Kaur-Frost classifier, a Dodie-Frost classifier, and a Frost-Dodie classifier. Using NLTK's accuracy tool, we run each classifier against our test set, ultimately showing that, based solely on the respective per-poem frequency of the pronouns "I", "you", "he", and "she", all three classifiers identify authorship of a poem with between 84% and 98% accuracy! See accuracy remarks in Figure 1.

For each classifier, we can then ask NLTK to reveal which pronoun's frequency was the most helpful—statisticians say "informative"—in distinguishing

[2] See https://www.nltk.org/

[3] See https://www.nltk.org/book/ch06.html for detailed instructions.

Table 2: *Dodie / Frost Most Informative Features*

Pronoun-Freq Feature	Ratio
he = 0.0	dodie : frost = 2.8 : 1.0
you = 0.0	frost : dodie = 2.1 : 1.0
she = 0.0	dodie : frost = 1.9 : 1.0

Table 3: *Kaur / Frost Most Informative Features*

Pronoun-Freq Feature	Ratio
he = 0.0	kaur : frost = 4.3 : 1.0
you = 0.0	kaur : frost = 2.2 : 1.0
she = 0.0	kaur : frost = 1.5 : 1.0

one author's poems from another. Observe these results in Figure 2, Figure 3, and Figure 4. Observe in Figure 2 that the pronoun "he" is 2.8 times more likely to occur 0.0 times in a Dodie Clark poem, than in a Robert Frost poem. Similarly, Figure 2 indicates that Frost is 2.1 times more likely not to use "you", while Dodie is 1.9 times more likely not to use "she".

We interpret Figure 3 and Figure 4 in a similar manner. Kaur is 4.3 times more likely than Frost not to use "he". Most remarkably, Kaur is 6.2 times more likely than Dodie not to mention herself—i.e. to use the pronoun "I".

5 Conclusions

Perhaps an engaging first reflection is why Dodie is nearly three times more likely than Frost not to use the "he" pronoun. Firstly let's consider her audience. Dodie's most loyal concert-going, YouTube-following, and carry-book-everywhere-type fans are mostly other young women, especially those in the age range between high-school and college. Based on her fanbase's demographic, then, one might suggest that Dodie's focus on a female audience accounts for less use of the male pronoun. Yet her poems' concern with intimate relationships with men seems to counter this intuition. It may simply be Dodie's characteristic deep concern for trending young women's issues such as body-positivity steers her diction away from male subjects, and towards female ones. This is further suggested by her using the "she" pronoun almost twice as often as Frost.

Even more stark is Rupi Kaur's trend for not using the 3rd-person male pronoun, being 4.3 more likely than Frost not to refer to a "he" in a poem. Similarly to Dodie, we might account this to her writing about women's issues towards a female audience. But Kaur's higher zero-occurrence ratio must also be accounted for. It may be that Dodie's style of songwriting leads her to use choruses, which would skew our data, making our occurrence rates artificially high for pronouns in choruses that appear several times throughout a song. Alternatively, even a brief survey of Kaur's work reveals an even more intense focus away from relationships with men. While Dodie has several songs celebrating a relationship with a man, Kaur's celebratory poems remark on a female's individual strength, rarely mentioning a "he". Naturally, then, Kaur would be even less likely to use the "he" pronoun than Dodie.

Table 4: *Kaur / Dodie Most Informative Features*

Pronoun-Freq Feature	Ratio
i = 0.0	kaur : dodie = 6.2 : 1.0
you = 0.0	kaur : dodie = 5.1 : 1.0
she = 0.0	kaur : dodie = 1.3 : 1.0

But most drastic of all is the difference between Dodie and Kaur's use of "I": Kaur is 6.2 times more likely not to mention herself directly. Since Dodie is a songwriter and lyric-poet, she writes in the first person more often. Note further that by contrast, Kaur's poems sometimes read like self-help advice, which naturally has a second-person—not first-person—audience: "if you are not good enough for yourself / you will never be enough / for someone else", or "you have to stop / searching for why at some point / you have to leave it alone". Therefore while Dodie is distinguished by first-person lyric poetry, Kaur tends to address a second-person audience.

But perhaps it is most interesting of all to reflect not merely on how accurate our classifiers are, but on what examples were classified inaccurately, and whether the model may be revealing shared habits of distinct authors. Consider the poem "Good Hours", under file ID "frost016.txt", by Robert Frost. According to our classifier, this poem is by Rupi Kaur. Figure 6 shows that the classifier made this decision based upon the fact that "he", "she", and "you" do not appear in this poem, while "I" appears ten times in sixteen lines. If we recall our earlier observation that Kaur is identifiable by comparatively less use of the "I" pronoun, we may be surprised to find that the model attributes an "I"-heavy poem like "Good Hours" to Kaur. But it must be recalled as well that this observation distinguished Kaur from Dodie, not Kaur from Frost. In fact since "I" is not among the "most informative features" reported for the Kaur-Frost classifier, we conclude that the model's decision was not strongly informed by "I"-frequency at all. We might note next what other pronoun-frequencies the model may be using to distinguish Kaur from Frost. Also by Figure 6, the Kaur-Frost classifier depends most heavily not on "I", but on the frequencies of "he", "she", and "you". In "Good Hours", all three of these pronouns are absent. In particular, "he" is absent, which strongly indicates a focus away from male themes, which we would intuitively associate with Kaur. Indeed if the poem were by Frost we would not expect to find "he" missing. The model, then, naturally supposes that Kaur is the author.

Finally, even as our model makes incorrect guesses

Table 5: *Dodie/Frost Relative Pronoun-Frequency in Select Examples*

Filename	Predicted Author	True Author	“I"	“You"	“He"	“She"
frost016.txt	Dodie	Frost	0.07	0.0	0.0	0.0
dodie022.txt	Dodie	Dodie	0.01	0.01	0.0	0.0
frost004.txt	Dodie	Frost	0.04	0.03	0.0	0.0

Table 6: *Kaur/Frost Relative Pronoun-Frequency in Select Examples*

Filename	Predicted Author	True Author	“I"	“You"	“He"	“She"
frost006.txt	Kaur	Frost	0.07	0.0	0.0	0.0
kaur143.txt	Kaur	Kaur	0.05	0.0	0.0	0.0
kaur064.txt	Kaur	Kaur	0.11	0.00	0.0	0.0

about authorship, we discover unexpected similarities between distinct authors. As a songwriter and lyric-poet, Dodie is characterized by the first person perspective. Rupi Kaur, being concerned with women’s issues, tends to not use “he”. As a dramatist, perhaps Robert Frost is not so distinguishable by the lack of some pronoun because he describes broad arrays of characters, whereas Dodie and Kaur both relate experiences from their own lives. Perhaps most interestingly, our model, despite being 98% accurate for Kaur / Frost, in several cases confuses our “instagram" poet with a canonized poet. Perhaps, then, digital analysis not only reveals insights beyond human reading, but helps legitimize and usher in this new genre in modern poetry.

References

[1] Matthew Lee Jockers. (2013). *Macroanalysis: digital methods and literary history*. Urbana (Ill.)

$Etc.$

University Of Illinois Press.

Table 7: *Kaur/Dodie Relative Pronoun-Frequency in Select Examples*

Filename	Predicted Author	True Author	“I"	“You"	“He"	“She"
dodie001.txt	Kaur	Dodie	0.09	0.0	0.0	0.0
kaur087.txt	Kaur	Kaur	0.0	0.13	0.0	0.0
kaur010.txt	Kaur	Kaur	0.07	0.07	0.0	0.0