JURIST Digital Scholars Report

Agasha Ratam

September 27, 2020

1 Project Description

1.1 Research Question

How does the writing style of an online article affect user engagement?

1.2 Background

One of the core missions of legal news providers is to make news more accessible to the general public. **Readability**, one measurement that describes an author's **writing style**, is often an important factor that determines the suitable audience for the text. English articles on Wikipedia have counterparts written in "Simple English", which emphasizes the use of simple grammar and basic English words when possible, with the intention of making the article accessible to more readers. A study finds that indeed, articles on the Simple English Wikipedia are more readable than the English Wikipedia, when measured using the Flesch Reading Ease scale [1]. This gives rise to the question of whether improving aspects of an author's writing style, such as readability, would prompt a higher level of **user engagement**.

Therefore, this project seeks to use JURIST's legal news article archive to examine the relationship between writing style and user engagement with an article. Understanding this relationship would help authors create a new way to assess their writing style and how it may impact the level of user engagement with their articles. The outputs could then help writers who intend to contribute to JURIST adopt writing techniques that are likely to capture higher user engagement, as predicted using a neural network trained on past articles.

1.3 Measurements

The **exit rate** of the page is one way of capturing user engagement with an article. The exit rate of a page is defined as the number of exits from the page (i.e. the number of times a user leaves the JURIST website from this page) divided by the number of times this

page is viewed. A page having a lower exit rate implies that upon viewing the page, users are more likely to browse other pages on JURIST (e.g. by clicking a hyperlink embedded in the article or by using the navigation bar), presumably out of interest developed after encountering the article. Websites typically aim to lower exit rates, as it would imply that users are more engaged by the website's contents.

For this project, an author's writing style comprises decisions that the author makes in expressing information in the text. While writing style as a whole may not be easy to quantify, there are concrete features that can be identified in a given piece of text.

One aspect of writing style, for instance, is the readability of the text. The **Flesch Reading Ease Scale** is one of several measurements of readability for texts written in English. It takes into consideration the average number of syllables per word and the average number of words per sentence. The Flesch Reading Ease score is a real value: on one end, a text scoring 10.0 or lower is categorized as "professional" and suitable for graduate students, while on the other end, a text scoring 90.0 or above is categorized as "fifth grade" and is easily understood by an average 11-year-old student.

The **voice** of sentences is another feature of an author's writing style. In writing, authors are often encouraged to avoid the use of the passive voice, which may take away from the clarity of the sentence, as opposed to the active voice, which is typically a more natural way to express thoughts and information. There are many exceptions, however, where ideas are in fact better conveyed when the sentence is written in passive voice. For a given article, the frequency at which an author uses the passive voice can be measured.

Sentence structure is another stylistic choice that the author has control of. Starting sentences with a conjunction such as "and" or "but" or ending a sentence with a preposition such as "in" or "to" may appear more informal but can help maintain flow in transitioning between sentences. The frequency of these sentence structures is measured as another aspect of writing style.

The overall **frequencies of parts of speech** — conjunctions, adverbs, and adjectives — are also measured. Some authors choose to be very descriptive in expressing information by using adverbs and adjectives, while others prefer not to in order to make the text more concise and easily understandable. Similarly, conjunctions are typically used to maintain a certain flow of logic when transitioning between ideas, but overusing them may make the text less clear.

The choice to use these variables are discussed in the Project Development section.

1.4 Data

The data consists of two parts: an archive of all JURIST articles from 2000 to mid-May 2020 and Google Analytics data from 2014 to 2020.

There are two types of JURIST articles: news and commentary. The articles are stored as XML files exported from Wordpress. In this format, each item contains components that would be displayed on the website, including the title, text, and time of the article's

publishing. In addition, each item also contains additional information, such as a unique ID as well as tags and categories associated with the article. Commentary articles also contain data on the authors of an article and any organizations affiliated with them. The XML files of the news articles lack this information, despite this information being displayed on their web pages.

The Google Analytics data contains user engagement data of every page on the JURIST website for each year since 2014. The data is stored in CSV format, with each page identified by their title, and page metrics including page views, average time spent on page, and exit rate for the given year.

2 Project Development

Developments in this project can be categorized into five steps.

1. Data merging and cleaning
2. Data processing
3. Feature selection
4. Neural network training
5. Web tool development

Developments were made in this general order, but there were many instances where it was deemed necessary to return to the previous steps to make adjustments. For instance, during feature selection, potential features could be thought of and added in the data processing step. Once a features is found to be unnecessary for the neural network, it would no longer be computed in future iterations of data processing.

2.1 Data cleaning and merging

The goal is to create a table where every row is a unique article, and where the columns are information related to the article, from both the article archive and aggregated Google Analytics data.

First, the article archive files are parsed using an XML parser, keeping all information needed including the article's title, text, tags, and authors. The text of a commentary article usually features one paragraph or more containing the author's bio, as well as extraneous information such as suggested citation and footnotes. A script was introduced to detect and discard these blocks of text, as they are not the main content of the article.

Next, the articles are matched to the Google Analytics data. The Google Analytics files identify articles by the title of the web page, rather than the title of the article. In 2018, the JURIST website underwent a change in page title naming. For instance, a news article

would be posted on a page titled "JURIST – [Article title]" prior to 2018 but would be on a page titled "[Article title] – JURIST – News – Legal News & Commentary" afterwards. Nonetheless, articles could still be matched to both versions since in both cases, the article title is contained in the page title.

Many articles have analytics data spanning several years. When this is the case, data is then aggregated. For instance, the **exit rate** is defined as the number of page exits divided by the number of page views. Since page views are also given for each year, the number of page exits for a given year could then be calculated. Finally, the aggregated exit rate would be the total number of page exits across all years divided by the total number of page views across all years.

Altogether, the final table consists 10,587 news articles and 960 commentary articles, totaling 11,547 articles.

2.2 Data processing

After cleaning and merging, the data undergoes processing in which features of interest are extracted.

The first step is to obtain basic features: the number of words, sentences, paragraphs, and links embedded in the text. Heuristics are introduced to identify the start and end of sentences, also accounting unusual circumstances. For instance, articles may contain a list of long sentence fragments; these would be treated as individual sentences since they visually appear as separate entities, otherwise this could lower the readability scores significantly. Python segmenter modules that can separate sentences demonstrates similar performance to the self-developed heuristics and may still miss sentences with formatting mistakes, which the heuristics can address.

Other similar features include the number of words in the title and in the first two paragraphs.

Readability scores are calculated for different components of the text: the entire text, the first two paragraphs of text, and the title. Eight different measurements of readability were used: Kincaid Grade Level, Automated Readability Index (ARI), Coleman-Liau Index, Flesch Reading Ease, Gunning-Fox Index, Lix, SMOG Index, and Rix. These measurements were calculated using the Python module **readability** [cite].

Self-developed measures of writing style as discussed in the Project Description are also calculated: percentage of sentences that are in passive voice, start with a coordinating conjunction (e.g. "and", "or"), and end with an adposition (e.g. "in", "to"); percentage of words that are coordinating or subordinating conjunctions, adverbs, and adjectives. Part-of-speech tagging was done with the assistance of the Python module **syntok** [cite].

The ten most common tags across all articles were identified: domestic, international, human rights, US Supreme Court, UN, immigration, Donald Trump, gay rights, death penalty, and abortion. Then, for each of these tags, a binary indicator variable is created to indicate whether an article is associated with the tag.

Table 1 shows the mean statistics of the articles.

	News	Commentary
n	10,587	960
Links	9.32	7.74
Words	262.76	1114.39
Kincaid Grade Level	14.60	14.38
Flesch Reading Ease	41.81	42.11
Page views	280.01	1,067.65
Entrances	207.24	885.25
Average Time on Page (seconds)	182.04	263.65
Bounce Rate	84.11%	88.74%
Exit Rate	68.31%	80.94%

Table 1: Mean statistics

2.3 Feature selection

In the final table produced from data processing, each item now has 61 variables. The goal of feature selection is then to only select relevant features as well as reduce redundant features, since the effects of similar features could be unclear.

The first step is to analyze the many measures of readability. An analysis of the covariance of pairs of these measurements shows that most of these measurements are highly correlated. To eliminate unnecessary redundancy, dimensionality reduction using Principal Component Analysis (PCA) was implemented. PCA reorients a multidimensional data set and orders the axes or principal components by the variance of data points with respect to each axis. This way, the first principal component contains captures the most variation in the data, followed by the second principal component, and so on. The first few principal components are usually selected to represent the data, while the rest are discarded.

It was found that the Flesch Reading Ease has a 0.975 correlation with the first principal component, which by itself accounts for 89% of the variance. Since the Flesch Reading Ease manages to represent the other measures very well, the Flesch Reading Ease was henceforth used as the only measure of readability.

Linear regression is used to see whether features are relevant to the outcome. The features were used to predict the three candidates for the dependent variables of interest. It is important to keep in mind that these features will be the inputs of a neural network. Hence it is better to be lean towards being lenient in deciding which variables matter, since the irrelevance of a variable could eventually be detected through the training of the neural network.

Two results of linear regression were of interest: statistical significance at the 0.1 level and the R-squared value. A variable that is statistically significant for one of the dependent variables suggests its relevance to user engagement. Also, as features are added and removed, the R-squared value indicates how well the whole selection of features explains the variance in the dependent variables.

There were 25 features selected as inputs to the neural network. Some of the selected features are shown in Table 2. Note that some variables are not statistically significant to any of the three dependent variables yet were still selected as inputs to the neural network. This is because lenience is preferred, as the effect of some variables might be obscured by other variables that are strongly correlated with them.

	Dependent Variables		
	df[[dv]]		
	Avg. Time on Page	Bounce Rate	Exit Rate
	(1)	(2)	(3)
paragraphs	−2.211***	0.253***	−0.119***
	(0.258)	(0.026)	(0.043)
text_words	0.082***	0.001	0.010***
	(0.004)	(0.0004)	(0.001)
text_flesch	−0.022	0.014*	0.023**
	(0.071)	(0.007)	(0.012)
sentences_passive_rate	−2.001	0.476	−0.386
	(5.226)	(0.517)	(0.870)
sentences_start_with_cconj_rate	43.431*	0.265	−4.712
	(25.677)	(2.539)	(4.275)
sentences_end_with_adp_rate	−10.330	−0.839	−7.071
	(35.414)	(3.502)	(5.896)
words_conjunctions_rate	58.892	35.555***	42.336***
	(45.948)	(4.544)	(7.649)
words_adv_rate	−5.508	14.128**	1.656
	(55.834)	(5.522)	(9.295)
words_adj_rate	67.032**	−8.365***	−11.473**
	(31.762)	(3.141)	(5.288)
Constant	176.325***	82.682***	68.580***
	(6.407)	(0.634)	(1.067)
Observations	11,547	11,547	11,547
R^2	0.127	0.122	0.141
Adjusted R^2	0.125	0.120	0.139
Residual Std. Error (df = 11524)	66.329	6.560	11.042
F Statistic (df = 22; 11524)	76.157***	72.758***	85.915***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Linear regression showing some of the selected features.

The final goal is to determine which output to prioritize, that is, to select which measure of user engagement will be considered the dependent variable. The three major candidates are directly available from the Google Analytics data: average time on page, bounce rate, and exit rate.

The **average time on page** was first eliminated, as there were two problems. The first was that it intuitively seemed bound to error. For instance, if an article is viewed once for an incredibly long time, for instance when a person stays on the page and leaves the computer, then this might skew the average time on page upwards. This effect is especially concerning since many articles do not have a lot of views, which renders potential fixes such as eliminating the outliers less helpful. Second, interpreting the meaning of a higher average time on page is not straightforward. One argument is that more time spent on the page shows that users are more engaged and willing to read the article, while another argument is that it shows how users need a longer time to comprehend the article, decreasing user engagement. Another measurement ? the average time per word? was tested, but this was found to also be prone to error.

The **bounce rate** of a web page is defined as how often users leave the website immediately, given that they first entered the website from this page. This is different from the exit rate, which considers all instances where the page is viewed, regardless of whether the user first entered the website from the page or another. Then, between the bounce rate and the exit rate, the exit rate was chosen since the bounce rate has a smaller sample size by definition, as it only considers instances where the user visits exactly one page on the website. The exit rate is therefore considered more representative of user behavior, as well as intuitive.

2.4 Neural network

The implementation of the neural network was done using Keras 2.3.0 with Theano backend.

The neural network has a feedforward structure. It contains an input layer with 25 neurons, some of which are real-valued (e.g. percentage of sentences that are passive) and some binary (e.g. tag indicator variables). There are 5 hidden layers, each with 10 neurons. The output layer contains 1 neuron for the predicted exit rate. All neurons have a ReLU activation function. The loss function used was Mean Squared Error.

The data set were separated into two equal halves for training and validation. Training was done for 200 epochs. The loss throughout the training process is displayed in Figure 1. The final loss is 119.35.

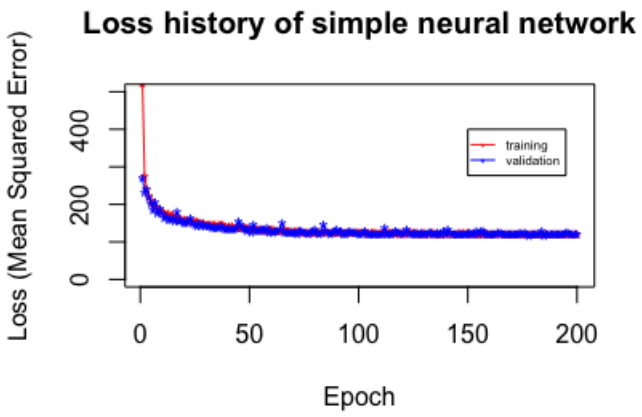

Figure 1: Loss of neural network over 200 epochs

2.4.1 Benchmark comparison

To evaluate how well the neural network performs, a benchmark neural network was tested. One common practice to analyze text data using neural networks is by using a recurrent neural network with long short-term memory. The input layer consists of 500 neurons, meaning that only the first 500 tokens (words or punctuations) are considered in a text. This was not a problem, since most articles are short enough to fit in their entirety.

The training was done with equivalent computational time for 20 epochs. The loss throughout the training process is displayed in Figure 2. The final loss is 135.64.

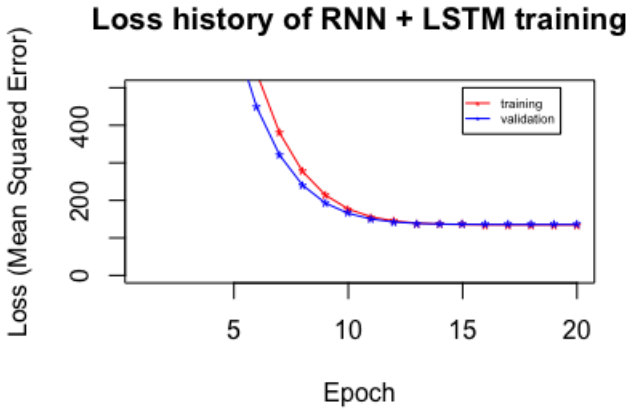

Figure 2: Loss of recurrent neural network over 20 epochs

The primary neural network therefore slightly outperforms the benchmark comparison.

2.5 Web tool development

A final web tool was developed using Flask and D3 for visualization, with the intended users being writers who intend to contribute to JURIST. The goal of this project is to help these writers learn how their writing style affects user engagement. The tool therefore attempts to communicate the findings and outputs of this project directly to writers.

A snapshot of the web tool is displayed in Figure 3 and contains 3 main panels.

The left panel contains buttons and text boxes where the user inputs the article. Even though an input in the format of WordPress-formatted text was considered, the final version of the tool accepts plain text so that it can be used by JURIST non-staff writers who do not incorporate WordPress into their workflow. The top-right panel shows summary statistics for the article as well as mean statistics for similar articles, which with the same tags (considering only the 10 most popular tags) to that of the inputted article. The bottom-right panel contains two figures. The first shows the Flesch Reading Ease of the text in the article and in similar articles. The second shows the histogram of exit rates of similar articles. A vertical line then shows the predicted exit rate for the inputted article. This prediction is made using the trained neural network.

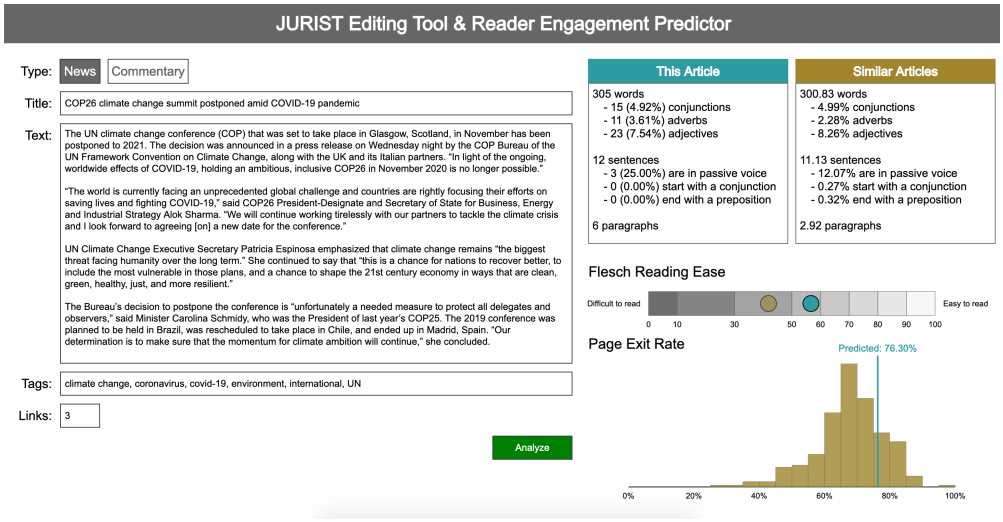

Figure 3: A snapshot of the web tool. The input is a JURIST article from April 2020. The tool shows a predicted exit rate of 76.30%.

3 Discussion

Throughout its development, the scope of the project became wider than what was envisioned in the initial proposal. The initial proposal suggested using linear regression to study the effects of text readability on user engagement, after controlling for variables such as article content, authors, time of publishing. It only suggested neural networks as a possibility. Moreover, a tangible product such as the web tool was developed as opposed to just a report.

Many obstacles were encountered throughout the project which encouraged improvements. One such obstacle was discovering that readability alone does not predict user engagement very well, even after controlling for many possible confounding variables. The project took a major turn as the independent variable is expanded to the author's writing style rather than just readability. The steps of feature exploration and selection thus became necessary.

In retrospect, the order or prioritization of the various steps in the project could be reconsidered. Data cleaning and merging required a lot of effort, but this was necessary and hard to avoid since the data is specific to JURIST and came in various formats.

But the steps of data processing and feature selection, which involved a lot of trial and error, could have been de-prioritized for two reasons. The first one is that the methodology of trial and error is simply not effective. When a lot of variables are created and tested in a linear regression, it is bound that some will show statistical significance. Since the range

of variables to test was unrestricted and involved too much exploration, this methodology could seem questionable. The second one resulted from the lack of foresight for the following steps. Since the selected features are inputs to a neural network, it should matter little if some inputs prove to be irrelevant, since the training would be able to set their weights to 0. As long as not too many such inputs are included, the performance of the neural network should remain considerably acceptable.

Instead, more time could have been focused towards learning about and designing an appropriate structure for the neural network. Many improvements could be made, for instance by selecting the appropriate number of hidden layers, number of neutrons in these layers, or studying the effects of different activation and loss functions. Using the current neural network is one way to perform regression or real-valued prediction, but does not guarantee by any means that there are no better ways to structure it.

Some other improvements could be done for the project. First, since many of the data cleaning, merging, and processing functions were written from scratch, it would be useful to run more careful testing to detect bugs or mistakes in handling edge cases. Another improvement would be to filter out data points. For instance, since the Google Analytics data only dates back to 2014, articles before then should be excluded, since the user engagement recorded for those articles may not reflect its true values throughout the article's entire history. For instance, one would expect that a new article has highest user engagement soon after publishing, when the content is more relevant to readers.

Finally, it is important to note that while the relationships between writing style and exit rate may be generalizable to legal news articles and perhaps more general news articles, the neural network trained is specific to JURIST. This is because the exit rate of web pages can vary significantly between different websites. The web pages of JURIST articles all share the same many features including presentation and site interactivity, which affect the exit rate. However, articles from different websites may differ in exit rates for reasons beyond the content of the article. Therefore, what is considered a low exit rate for one website may be considered high for another.

References

[1] Jatowt, A., & Tanaka, K. (2012). Is Wikipedia Too Difficult? Comparative Analysis of Readability of Wikipedia, Simple Wikipedia and Britannica. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 4.