Readability Score Classification of Computer Science Publications using Neural Network

Team 13
Md Shahriar Forhad
Amit Das
Alfredo J. Pena

Supervisor
Dr. Dongchul Kim

October 6th, 2021

Summary of the Proposal

Readability is an important variable when writing a cohesive scholarly paper. We often take for granted and will neglect this aspect when writing a focused paper replacing readability for word jargon and deep long-winded explanations. Computers have long been used in the analysis of text syntax. We plan to take it a step further and use Deep Learning to quantify readability in a specific type of writing. Our goal is to develop a deep learning model that analyses scholarly papers' readability and returns a usable score, using the publications' data from Scopus®.

Background

The unparalleled computing capabilities of Artificial Intelligence (AI) and Machine Learning (ML) have introduced new dimensions in many fields. A majority of those fields being Science and Technology. The expansion of these technologies into fields unrelated to its birth are untapped and brimming with potential. Here, we will use that computing potential to explore the readability of publications.

The most common algorithms for scoring readability is Flesch Reading Ease Score, Flesch-Kincaid Grade Level test. In Flesch-Kincaid score [1], it rated the text on a U.S. school grade level, e.g, a score of 8.0 means that an eighth grader can understand the document. For the most published articles, aim for a score of approximately 7.0 to 8.0. The grading formula for the Flesch-Kincaid –

$$(0.39 * ASL) + (11.8 * ASW) - 15.59 \qquad (1)$$

ASL = Average Sentence Length
(the number of words divided by the number of sentences)

ASW = Average Number of Syllables per Word
(the number of syllables divided by the number of words)

Flesch Reading Ease, tested the rates of text on a 100-point scale. The higher the score, the easier it is to understand the document. For most standard files, the score ranges from 60 to 70 using the formula -

$$206.835 - (1.015 * ASL) - (84.6 * ASW) \qquad (2)$$

ASL = Average Sentence Length
(the number of words divided by the number of sentences)

ASW = Average Number of Syllables per Word
(the number of syllables divided by the number of words)

From US DOD to insurance policies, reading scores have a big impact on policymakers as well as common citizens. For example, Florida state mandates 45 or higher Flesch Reading Ease score for its insurance companies policy documents [2].

There are some wonderful papers on a similar topic exploring the readability of languages in general writing or even exploring the readability of code which may also be a lot more simple for a computer to understand and may lead to the development of self written or self maintaining code in the future. As well as reduce costs for development overall. [3]

A paper from the Karolinska Institutet in Sweden suggests that the readability of scientific texts is trending down overall and being filled with more and more jargon every year.[4]
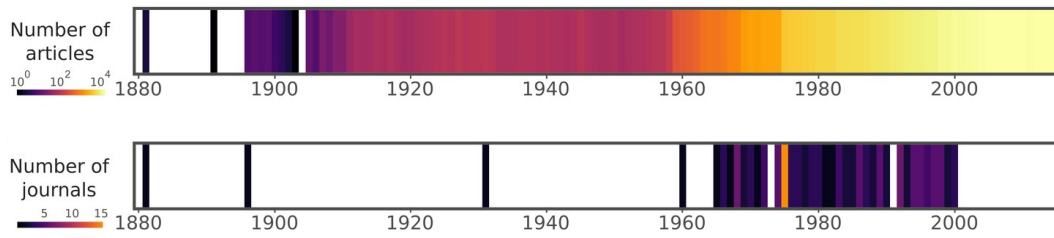

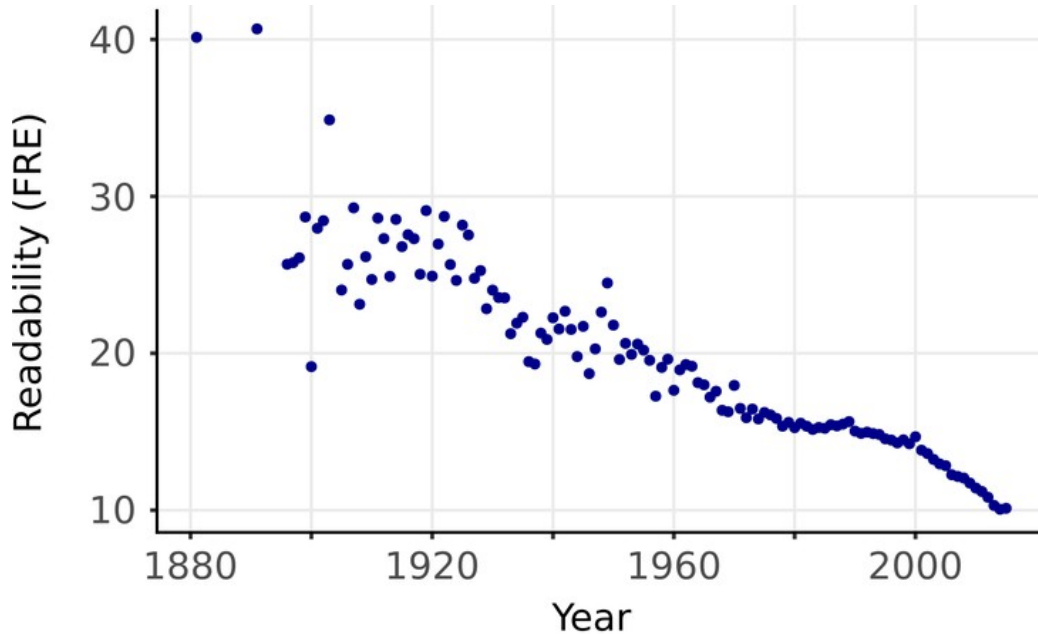Figure 1: Articles and Journals from 1880 to 2017. [4]

Figure 2: Mean Flesch Reading Ease (FRE) readability for each year. [4]

This type of trend may be dangerous to the availability of research papers to the general public and even to inter field research. The less we can understand one another the less progress we make. Although the advancement of technology over the years may suggest that there is no need to fix this readability problem, we can also assume we have simply not hit the point where we can no longer understand ourselves with ease.

Readability being such an important attribute for a well-written paper in any field makes study with Deep Learning practices more enticing. Passing a paper through machine eyes to quantify readability will prove a useful asset for future development.

Goal and Objectives

We seek to develop an application that can produce an accurate readability score for a Computer Science research paper. Taking traditional DL and ML methods and implementing a usable application to quantify readability.

Data and Methods

In this research we will create a novel dataset from the Scopus database, the largest repository of scientific publications. We will collect all the publications available in the fields of Machine Learning till now from the database. A big chunk of time will be spent in generating and cleansing this data. We found approximately 300k publications in the database. We are downloading and cleaning this in a group of less than 2k papers because of the limitation set by the SCOPUS database. Then we will use traditional reading scoring algorithms to quantify those abstracts. After that, we will use the dataset and newly acquired score, to train a deep learning model. First, we are building a pre-model for testing purposes, with around 10k papers data, and measuring the accuracy with testing data.

Timeline

OCT Week-1	Understanding Our Goals and Objectives with Project
OCT Week-1	Begin Data Gathering from SCOPUS
OCT Week-1	Work on Building Initial Model for Readability Analysis
OCT Week-2	Finding the Accuracy of our Predictions
OCT Week 3	Add more Data and Build Expanded Model
OCT Week 4	Use Model to Test Single Paper for Readability
NOV Week 1	Build Mock Production Interface for App
3 Weeks Before Submission	Final Presentation
Last week of FALL'21	Complete Project Report Submission

References

[1] R. Flesch, "A new readability yardstick.," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948. DOI: 10.1037/h0057532. [Online]. Available: http://dx.doi.org/10.1037/h0057532.

[2] *Insurance rates and contracts*, The 2021 Florida Statutes, 2021. [Online]. Available: http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0600-0699/0627/Sections/0627.4145.html.

[3] Q. Mi, J. Keung, Y. Xiao, S. Mensah, and Y. Gao, "Improving code readability classification using convolutional neural networks," *Information and Software Technology*, vol. 104, pp. 60–71, 2018, ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2018.07.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584918301496.

[4] P. Plavén-Sigray, G. J. Matheson, B. C. Schiffler, and W. H. Thompson, "The readability of scientific texts is decreasing over time," *eLife*, vol. 6, Sep. 2017. DOI: 10.7554/elife.27725. [Online]. Available: http://dx.doi.org/10.7554/eLife.27725.