FinRead: A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms

Sohom Ghosh[†], **Shovon Sengupta**[†], **Sudip Kumar Naskar**[⋆], **Sunny Kumar Singh**[‡]
[†]Fidelity Investments, Bengaluru, India
[⋆]Jadavpur University, Kolkata, India
[‡]BITS, Pilani, Hyderabad, India
{sohom1ghosh, ssg.plabon, sudip.naskar, sunnysingh.econ}@gmail.com

Abstract

Simplified definitions of complex terms help learners to understand any content better. Comprehending readability is critical for the simplification of these contents. In most cases, the standard formula based readability measures do not hold good for measuring the complexity of definitions of financial terms. Furthermore, some of them works only for corpora of longer length which have at least 30 sentences. In this paper, we present a tool for evaluating readability of definitions of financial terms. It consists of a Light GBM based classification layer over sentence embeddings (Reimers et al., 2019) of FinBERT (Araci, 2019). It is trained on glossaries of several financial textbooks and definitions of various financial terms which are available on the web. The extensive evaluation shows that it outperforms the standard benchmarks by achieving a AU-ROC score of 0.993 on the validation set.

1 Introduction

The notion of readability assumes a central position in the emerging financial literature on textual analysis. Readability as a concept is difficult to define precisely. Readability can be broadly defined as a measure of how easy a text document is to read. In this exercise we aim to examine the readability measure for terms present in a financial glossary and compare various formula based approaches for measuring readability. These include "Automated Readability Index (ARI)" (Smith and Senter, 1967), "Flesch Reading Index (FRI)" (Flesch, 1948), "Dale-Chall formula (DCF)" (Chall and Dale, 1995) and "SMOG Index Score (SIS)" (Mc Laughlin, 1969). At the very outset, all these measures calculate a readability score based on U.S education system's grade level or years of education a reader might require to understand a text content. We further explore the limitations of these measures in the context of financial terms and develop a transfer learning-based system to measure readability of their definitions.

Inspired by the approach followed by (Chakraborty et al., 2021), we collect financial terms and their definitions from seven different sources. We crawl the data dictionary of a popular financial website, Investopedia[1]. The other six sources are text books related to finance. We extract financial terms and their definitions from glossaries of these books by transforming them to HTML format. The data distribution is: NCERT-"Introductory Macro-economics" (149 records), Investopedia (6204 records), (Samuelson and Nordhouse, 2009) (350 records), (Brealey et al., 2012) (177 records), (Hull, 2003) (531 records), (Bodie and Kane, 2020) (525 records), (Mishkin and Eakins, 2006) (465 records).

Among these sources, we assign a readability score of 1 to the definitions present in NCERT, Investopedia and (Samuelson and Nordhouse, 2009). For the remaining sources we assign a readability score of 0. We do this because the content of the former three sources are simple. They are read by school going kids and people in general. The latter four sources constitute of complex graduate level textbooks. We use 80% data for training and the remaining for validation. Finally, we extract ARI, FRI, DCF and SIS scores for each of the definitions using the textstat[2] library.

2 Model Development & Results

In this section, we discuss various approaches we explored and their performances. Firstly, using standard methods we calculate Area under Receiver Operating Characteristic curve (AU-ROC) which

[1]https://www.investopedia.com/financial-term-dictionary-4769738 accessed on 1st Oct 2021

[2]https://pypi.org/project/textstat/ accessed on 1st Oct 2021

Proceedings of the 18th International Conference on Natural Language Processing, pages 658–659
Silchar, India. December 16 - 19, 2021.

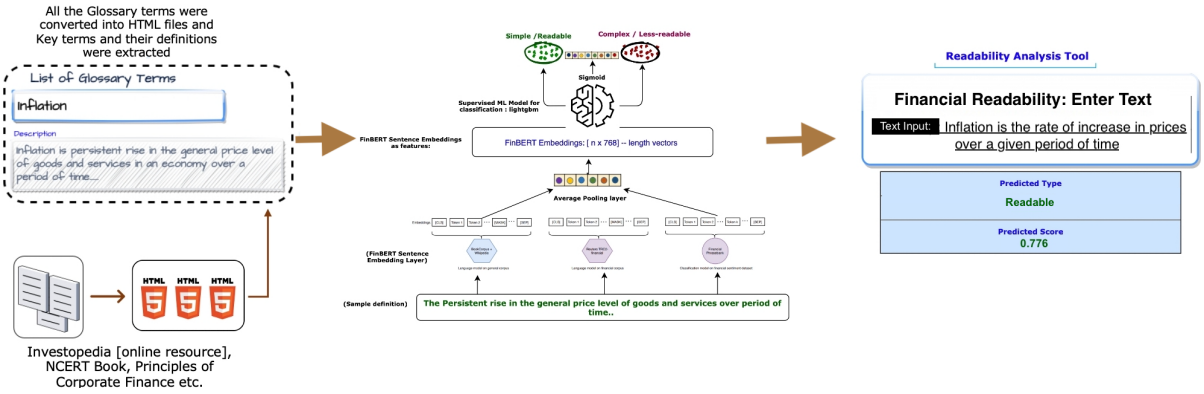

Figure 1: Financial Readability Flow Chart and Tool

is 0.742 for ARI, 0.435 for FRI, 0.413 for DCF and 0.730 for SIS. After that, we represent the definitions numerically using TF-IDF (ngrams: 1 to 4) and train several machine learning based models for classification like Logistic Regression, Random Forest and so on. We also experiment by replacing TF-IDF with "sentence-transformers" based embeddings (Reimers et al., 2019) with FinBERT (Araci, 2019) (768 dimensions). We also try other classification based approaches like XGBoost, CatBoost and lightGBM. We have summarised the model performances on the validation set in Table 1. We perform these experiments on Google Colab. Analysing the results we conclude that a lightGBM (20 min-child samples, 31 num-leaves) based classifier trained over sentence transformers embeddings (Reimers et al., 2019) having FinBERT (Araci, 2019) gives the best performance (AU-ROC 0.993). Moreover, it outperforms all the standard methods of measuring readability (like ARI, FRI, DCF and SIS) in terms of AU-ROC on the validation set.

Our contributions: a) Preparation of a corpus comprising glossaries of financial terms and their definitions b) *FinRead*- a tool to assess the readability of such definitions as shown in Figure 1.

In the future, we want to improve the overall quality of the system by increasing the size and quality of the corpora.

Model	AU-ROC	Model	AU-ROC
T+LR	0.940	F+RF	0.983
T+RF	0.930	F+XG	0.988
F+LR	0.992	**F+LG**	**0.993**

Table 1: Model performance on validation set. T: TF-IDF vectors, F: FinBERT embeddings, LR: Logistic Regression, RF: Random Forest, XG: XGBoost, LG: lightGBM

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Zvi Bodie and Alex Kane. 2020. Investments.

Richard A Brealey, Stewart C Myers, Franklin Allen, and Pitabas Mohanty. 2012. *Principles of corporate finance*. Tata McGraw-Hill Education.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. Simple or complex? learning to predict readability of bengali texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12621–12629.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

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

John C Hull. 2003. *Options futures and other derivatives*. Pearson Education India.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Frederic S Mishkin and Stanley G Eakins. 2006. *Financial markets and institutions*. Pearson Education India.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Samuelson and V Nordhouse. 2009. Economics: a textbook.

E A Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.