Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 1–9
Marseille, 24 June 2022

FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability

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

Abstract

In today's world, the advancement and spread of the Internet and digitalization have resulted in most information being openly accessible. This holds true for financial services as well. Investors make data driven decisions by analysing publicly available information like annual reports of listed companies, details regarding asset allocation of mutual funds, etc. Many a time these financial documents contain unknown **`financial terms`**. In such cases, it becomes important to look at their **`definitions`**. However, not all **`definitions`** are equally readable. Readability largely depends on the structure, complexity and constituent terms that make up a **`definition`**. This brings in the need for automatically evaluating the readability of **`definitions`** of **`financial terms`**. This paper presents a dataset, **`FinRAD`** (Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh, 2022), consisting of **`financial terms`**, their **`definitions`** and embeddings. In addition to standard readability scores (like "Flesch Reading Index (FRI)", "Automated Readability Index (ARI)", "SMOG Index Score (SIS)","Dale-Chall formula (DCF)", etc.), it also contains the readability scores (**`AR`**) assigned based on **`sources`** from which the terms have been collected. We manually inspect a sample from it to ensure the quality of the assignment. Subsequently, we prove that the rule-based standard readability scores (like "Flesch Reading Index (FRI)", "Automated Readability Index (ARI)", "SMOG Index Score (SIS)","Dale-Chall formula (DCF)", etc.) do not correlate well with the manually assigned binary readability scores of **`definitions`** of **`financial terms`**. Finally, we present a few neural baselines using transformer based architecture to automatically classify these definitions as readable or not. Pre-trained FinBERT model fine-tuned on **`FinRAD`** corpus performs the best (AU-ROC = 0.9927, F1 = 0.9610). This corpus can be downloaded from `https://github.com/sohomghosh/FinRAD_Financial_Readability_Assessment_Dataset`.

Keywords: Readability, Financial Texts, Natural Language Processing, Financial Dataset

1. Introduction

Nowadays investors prefer to avail themselves of financial services online. This saves time as well as money. While making decisions relating to investments, they tend to read relevant content online. All financial content is not easy to comprehend due to the presence of unknown terms. In such cases, they have to look for definitions of these terms. Interestingly, not all definitions are easy to understand. Thus, it is extremely important to aid financial content writers to assess how readable are the definitions which are being written by them. Figure 1 depicts the same.

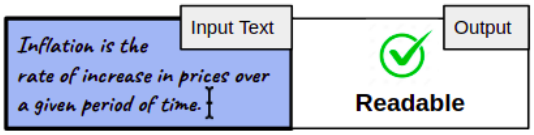

Figure 1: Readability of definition of "inflation"

We presented a basic tool, **`FinRead`** for demonstrating such a system in the 18th International Conference on Natural Language Processing (ICON-2021)[1] (Ghosh et al., 2021). It was trained using **`definitions`** of 8,401 **`financial terms`**. In this paper, in addition to extending this dataset to 13,112 **`definitions`** of **`financial terms`**, we release it publicly. Subsequently, we present several enhancements to the baseline architectures.

Our contributions

- We created a dataset comprising more than thirteen thousand **`definitions`** of **`financial terms`** along with their embeddings, standard formula based readability scores and assigned readability (**`AR`**) scores. We released it under the CC BY-NC-SA 4.0 license. To the best of our knowledge, we are the first to study readability in this context and provide the first dataset on financial terms and a proposed readability measure. A sample dataset can be downloaded from here[2]
- We showed that standard rule-based readability scores (like ARI, FRI, DCF, SMOG etc.) do not work well for financial texts.
- We proposed baseline architectures to automatically classify definitions of financial terms as readable or not.

[1] `http://icon2021.nits.ac.in/coloc_events.html`

[2] `https://github.com/sohomghosh/FinRAD_Financial_Readability_Assessment_Dataset`

The overall process flow is summarised in Figure 2. The rest of the paper is structured as follows. Section 2 states the prior works and their connection with this work. In section 3 we narrate the process we followed to collect, clean and label the data. Subsequently, we discuss various exploratory data analysis that we have performed. In section 4 we formally describe the task of assessing readability. We present various neural baseline architectures and their performances in section 5. Section 6 concludes the paper and provides some future directions of research.

2. Related Works

In this section, we discuss the prior works. Firstly, we narrate applications of readability in general and in the context of the financial domain. We then explore some of the related works and datasets.

2.1. Readability in general

For Natural Language Processing (NLP) practitioners, understanding readability of texts has always been an active area of research. Some of the standard readability scores include: "Flesch Reading Index (FRI)" (Flesch, 1948), "Automated Readability Index (ARI)" (Smith and Senter, 1967),"SMOG Index Score (SIS)" (Mc Laughlin, 1969) and "Dale-Chall formula (DCF)" (Chall and Dale, 1995). Flesh was one of the pioneers in this area. He proposed FRI which uses the ratio of total words to sentences and that of total syllables to total words as a measure of the readability. Smith et al. (Smith and Senter, 1967) defined ARI based on characters to words and words to sentences ratio. This score was used to assign the readability of a text to one of the fourteen predefined grade levels ranging from kindergarten to college student. Another new formula SIS for calculating readability was proposed by Mc Laughlin. It comprised of calculating the ratio between the number of polysyllables and sentences. However, it was only applicable for texts having at-least 30 sentences. In the paper, (Rush, 1985), Rush criticised these scores as they only dealt with the syntactic aspect of the texts and did not consider the aspect of the reading process which was interactive. Other papers which criticized these formulas include (Bruce et al., 1981) and (Anderson and Davison, 1986). Zamanian et al. (Zamanian and Heydari, 2012) presented a more detailed review of these formulas along with their advantages and disadvantages. Some of the papers which used language models to estimate readability include (Si and Callan, 2001), (Collins-Thompson and Callan, 2004), (Schwarm and Ostendorf, 2005) and (Heilman et al., 2007). In his recently published study of readability of "Policy Documents on the Digital Single Market of the European Union", Ruohonen(Ruohonen, 2021) argued that a PhD level eductaion would be required to study and understand the Digital Single Market (DSM) laws and policy documents. He further observed that there are critical differences in terms of the degree of agreement in various standard readability scores. The study also demonstrated, how the readability grades across time had evolved for the laws and policy documents in DSM as well. This in turn also indicates that the existing readability scores may fail to capture domain specific nuances for the different types of documents.

2.2. Readability in Financial Domain

Readability of financial texts has been widely explored. Most of these texts include Financial Disclosures (Loughran and McDonald, 2014), (Gosselin et al., 2021), Annual Reports and Management Discussions and Analysis (MD&A) (Arora and Chauhan, 2021), (Schroeder and Gibson, 1990), (Smith and Smith, 1971), (Lo et al., 2017). In addition to general features, Bonsall et al. (Bonsall IV et al., 2017) used the file size of 10-K documents to measure their readability. Bonsall et.al (Bonsall IV et al., 2017) proposed a new index "Bog Index" as a "plain English measure of financial reporting readability". It served as one of the standard approaches for the readability of financial reports. Loughran et al. (Loughran and McDonald, 2010) proposed a new method of measuring readability based on recommendations made by the U.S. Securities and Exchange Commission (SEC) in the year 1988. Readability scores were used for various downstream tasks like fraud detection (Othman et al., 2012), Stock Price Crash Risk prediction (Kim et al., 2019), etc. Readability of financial text books has been studied in (Chiang et al., 2008), (Plucinski and Seyedian, 2013) and (Plucinski et al., 2009). They also argued on the limitations of these popular scores as a measure of readability due to their inherent shortcomings to deal with domain specific language and jargon. Loughran et al. (Loughran and McDonald, 2014) also highlighted the need for alternative measures of readability for the financial documents like disclosures. Pitler (Pitler and Nenkova, 2008) proved that surface level standard readability scores do not correlate with the human assigned readability scores on the Wall Street Journal corpus. They further showed that a combination of entity coherence and discourse relations are the best features for assessing readability.

2.3. Related datasets

Related financial datasets on which readability has mostly been explored include 10-K SEC filing reports (Loughran and McDonald, 2009), disclosures (Ganguly et al., 2019), (Hoffmann and Kleimeier, 2021), and accounting textbooks (Chiang et al., 2008), (Plucinski et al., 2009).

2.4. Difference with prior works

To the best of our knowledge, we are the first ones to create a dataset consisting of definitions of financial terms along with their readability scores based on their complexity. We also propose transformer based neural baselines to automatically assess the readability of such definitions.

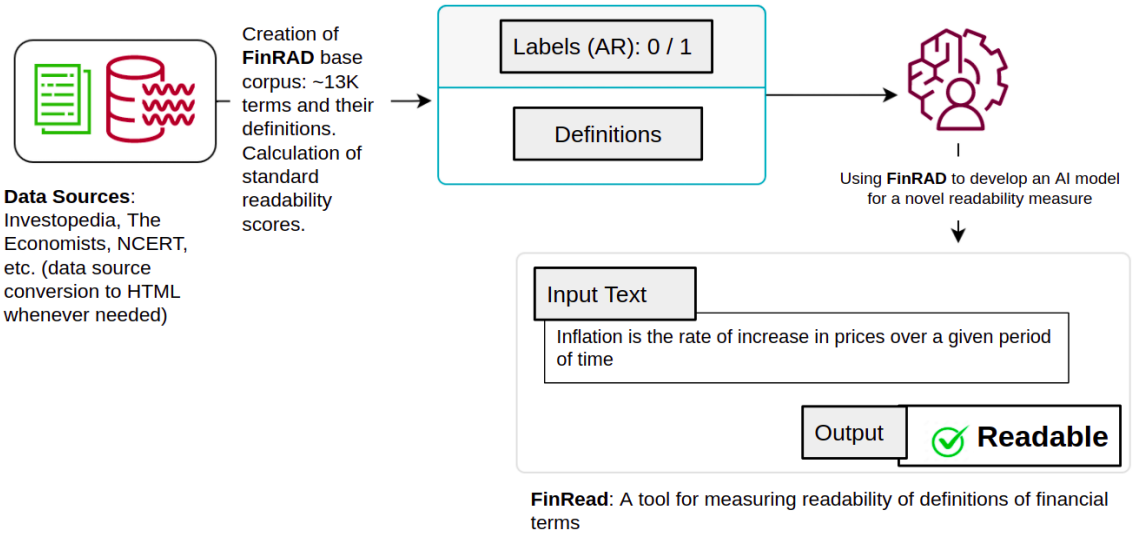

Figure 2: Overall process flow for **FinRAD**

3. Dataset

In this section, we narrate how we collected the data, cleaned and annotated it.

3.1. Data collection

Our dataset consists of 13,112 **financial terms** and their **definitions** written by experts across multiple sources. These sources include glossaries, dictionaries from financial websites, school and graduate-level textbooks relating to economics and finance. We collected the terms from 13 different sources and removed the duplicated terms during preprocessing. Source wise distribution of the dataset is presented in Table 1.

3.2. Data extraction and cleaning

Only three of the data sources considered were available as web-pages which we scraped directly. They include websites of The Economists, Federal Reserve Bank of St. Louis, and Investopedia. Other datasets were available in Portable Document Format (PDF). We tried extracting the terms and definitions directly from these PDFs first. However, we found that in most of the cases we were losing out on the structure. Thus, separating the terms from the definitions was challenging. Subsequently, we converted these PDF documents to the Hypertext Markup Language (HTML) format. For this, we used various freely available online services. We removed irrelevant texts like page numbers, the word "glossary", and texts which were mistakenly identified as terms. We removed the extra spaces and manually checked the final dataset to ensure that it is of high quality.

3.3. Data Annotations

Inspired by the method followed by Chakraborty et al. (Chakraborty et al., 2021), we consulted several professional financial experts. Subsequently, we decided to assign readability scores (**AR**) to the definitions of financial terms based on their sources. This was done since readability is subjective and manually annotating the entire dataset is expensive. Definitions from the following sources were assigned a readability score of 1.

- school-level textbooks (like NCERT textbooks, economics textbooks for begineers (Samuelson and Nordhouse, 2009))
- public websites suitable for masses (like Investopedia and The Economist).

The reason behind this is that the information from these sources is mostly consumed by beginners, school students, and by the masses. To understand the definitions which were obtained from other sources one needs to have at-least under graduate level knowledge specific to the financial domain. Thus, they were assigned a readability score of 0. This gave us 7,604 and 5,508 instances with readability scores of 1 and 0 respectively. An **AR** score of 1 represents the terms' definitions that are easily readable and 0 represents the definitions that are comparatively complex in nature or less readable. To validate this assumption we identified 112 additional terms and extracted their definitions from both kinds of sources (i.e. with **AR** = 0 and 1). We manually inspected each of the definitions and assigned them a readability score (0 or 1). In 79.91 % of the cases the manual assignment was in agreement with the assumption.

3.4. Exploratory Data Analysis

In this section, we present an overview of the **FinRAD** dataset and its contents. The dataset consists of 4 key fields:

Tag	Source Description	AR	# Terms/Definitions
prin	*Principles of Corporate Finance* by Richard A. Brealey, Stewart C. Myers, Franklin Allen (Brealey et al., 2019)	0	177
zvi	*Investments* by Zvi Bodie Alex Kane Alan J. Marcus (Bodie and Kane, 2020)	0	492
palgrave	*The Palgrave Macmillan Dictionary of Finance*, Investment and Banking by Erik Banks (Banks, 2010)	0	3925
opod	*Options, Futures, and Other Derivatives*, Global Edition by John C. Hull (Hull, 2003)	0	527
fmi	*Financial Markets and Institutions* by Frederic S. Mishkin Stanley Eakins (Mishkin and Eakins, 2006)	0	387
ncert_keec111	*NCERT Indian Economic Development Economics Class 11*[3]	1	95
ncert_kest	*NCERT Statistics for Economics Class 12*	1	53
ncert	*NCERT Introduction to MacroEconomics Class 12*	1	115
ncert_class12_econ	*NCERT Introduction to MicroEconomics Class 12*	1	41
investopedia	*Investopedia* Data Dictionary[4]	1	5946
economist	*The Economist* terms dictionary[5]	1	457
6_8_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis[6]	1	342
9_12_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis	1	188
pre_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis	1	36
sam	*Economics Textbook* by Paul Samuelson and William Nordhaus (Samuelson and Nordhouse, 2009)	1	331

Table 1: Source wise distribution. AR: Assigned Readability, #: Count

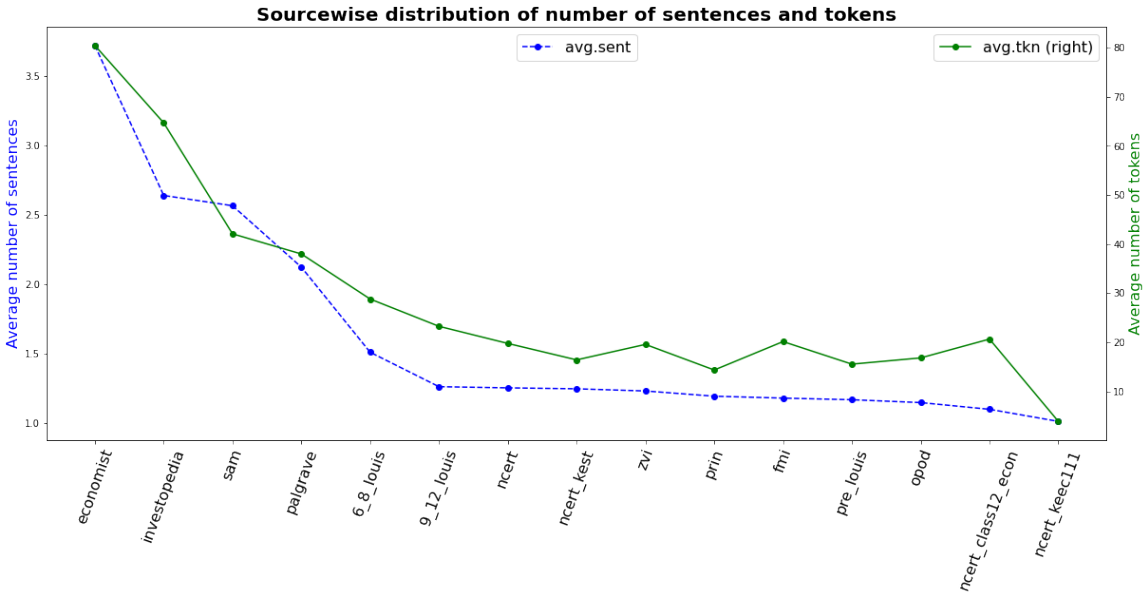

Figure 3: Source-wise distribution of the average number of sentences and tokens per definition

- **`financial terms`** (i.e. the terms that have been collected from different sources)
- **`definitions`** (i.e. the descriptions or definitions of these terms)
- **`source`** (i.e. the sources from which these terms have been collected)
- assigned readability (**`AR`** i.e. the annotated readability)

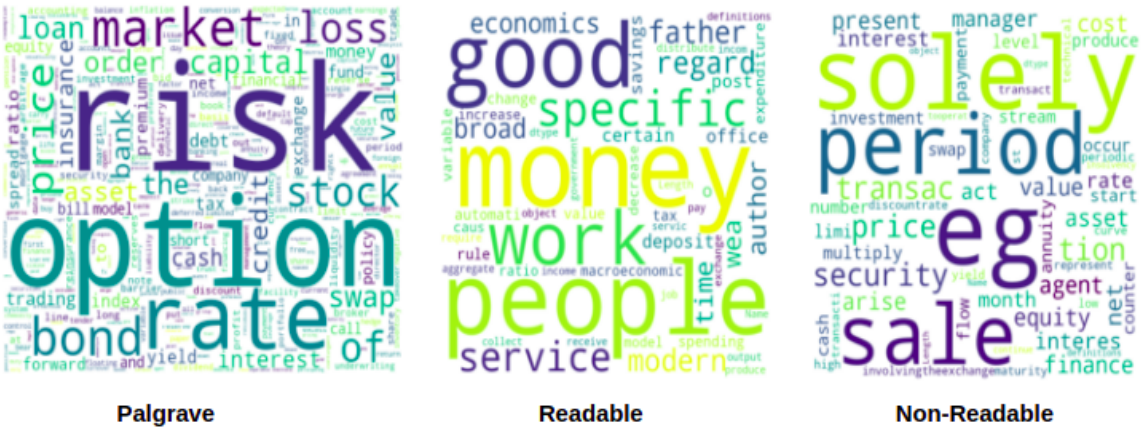

Figure 4: Word clouds of definitions from "Palgrave", readable and non-readable sources

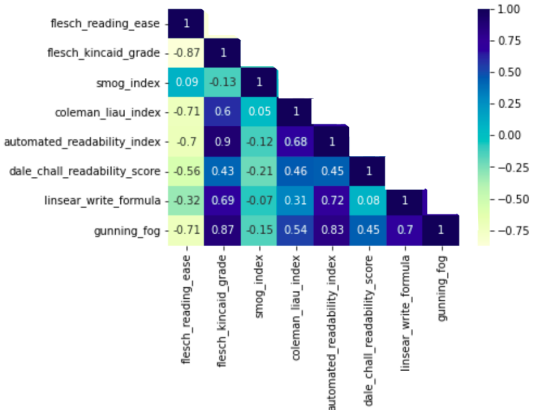

Figure 5: Correlation between standard readability scores

Apart from these 4 fields, the dataset also includes readability scores extracted using traditional methods. So far, 8 different scores have been provided for the definitions of the financial terms: Flesch Reading Ease (FRE) Score(Flesch, 1948), Flesch-Kincaid Grade Level (FKGL) Score(Kincaid et al., 1975), SMOG Index(SI) Score(Mc Laughlin, 1969), Coleman – Liau Index(CLI) Score(Coleman and Liau, 1975), Automated Readability Index(ARI) Score(Smith and Senter, 1967), Dale – Chall Readability (DCR) Score(Chall and Dale, 1995), Linsear write Formula and Gunning's Fog Index (FOG) Readability Formula. For all the definitions, these scores have been calculated using the textstat[7] library.

We started by studying the distribution of the number of sentences in the definitions across different sources. Figure 3 summarizes the distribution of the average number of sentences per definition used to define the terms across various sources. As evident from this

[7] https://pypi.org/project/textstat/

Readability type	Avg. sentences	Avg. tokens
Non-readable (0)	1.8529	32.2912
Readable (1)	2.5494	59.7701

Table 2: Average number of sentences and tokens per definition

plot, "The Economist" have definitions with the highest average number of sentences (approximately 4 sentences). We further compared the average number of sentences per definition across assigned readability segments in Table 2. It is quite interesting to note that the average number of sentences per definition in the readable set is higher than that of the non-readable set. Moreover, the average sentence length (i.e. number of tokens per sentence) for the readable set is 24.03 and that for the non-readable set is 17.22. This is because authors tend to use more words and shorter sentences to simplify concepts.

Subsequently, we studied the distribution of the average number of tokens present in the **`definitions`** across different sources. Figure 3 illustrates this. The average number of tokens per definition are approximately 80 and 64 for the definitions obtained from the readable sources "The Economist" and "Investopedia" respectively. This reconfirms our previous findings that authors tend to explain more to simplify concepts. In addition to this, we compared the average number of tokens across different readability segments. We observe that readable definitions have around 27 tokens more than that of non-readable ones. We provide more details and exact numbers in Table 2.

Word clouds are quite helpful to generate meaningful insights about text data. They offer an interesting option to visually represent the frequency of different words present in a corpus.

For ease of exposition, we have presented the word clouds of terms for one of the key sources ("Palgrave") in Figure 4. It accounts for almost 30% of the en-

tire dataset of terms. Furthermore, for effective comparison we also present word clouds of non-readable and readable definitions of financial terms in the same figure. Quite evidently, the frequent terms present in the non-readable definitions (**AR**=0) are more complex than those of the readable ones (**AR**=1).

Lastly, we study the correlation between the standard readability scores and present them in Figure 5. Now, it is apparent that all the scores can not be directly compared as they are generated using different mathematical principles. However, for a few scores which are comparable like Flesch Reading Ease formula (Flesch, 1948) and The Flesch-Kincaid Grade Level (Kincaid et al., 1975), the positive correlation is high. Similar conclusions can be drawn for other scores as well.

4. Task

Given a set $\mathfrak{D} = \{d_1, d_2, d_3, \ldots, d_n\}$ of **definitions** of **financial terms** and a set $\mathcal{R} = \{r_1, r_2, r_3, \ldots..., r_n\}$ of readability scores where r_i is the assigned readability (**AR**) corresponding to the definitions of financial term d_i and $r_i \in \{0, 1\}$. **AR**=0 denotes non-readable and **AR**=1 denotes readable. The task is to develop a system capable of classifying a definition as readable or not. Furthermore, it shoud be able to automatically compute readability score r_t for **definition** of any unknown **financial term** d_t. Note: $0 \leq r_t \leq 1$. We use Area Under the Receiver Operating Characteristic curve (AU-ROC) score as the evaluation parameter.

5. Models and Results

We divided the dataset into two parts keeping the event rate same - the training set (67%) and the validation set (33%). Firstly, we studied how standard readability scores (like FRI, ARI, SIS, DCF, etc.) performed in a domain-specific setting like this. Most of these scores provided grade levels as outputs. We calculated the AU-ROC, F1 and Accuracy considering readability of grade level higher than 12 as 0 and rest as 1. This was done following our assumption stated in section 3.3. The performance of these standard scores in measuring readability on the validation set are presented in Table 3. The performance on the validation set which was calculated using these scores was not up to the mark. The best AU-ROC was only 0.4986 using the Flesch Reading Index. Thus, we trained machine learning based classifiers to assess the readability of the **definitions**.

We represented **definitions** of the terms numerically using a Term Frequency - Inverse Document Frequency (TF-IDF) matrix. We trained various machine learning based classifiers over it such as Logistic Regression, Random Forest (Ho, 1995) and Gradient Boosting Machine (Friedman, 2001) and the results of these models are presented in Table 4. Furthermore, we experimented by replacing TF-IDF with sentence embeddings (Reimers and Gurevych, 2019) created using BERT (Devlin et al., 2019) and FinBERT (Araci, 2019). In addition to this, we tried using other machine learning based classifiers like LightGBM (Ke et al., 2017) and XG-Boost (Chen and Guestrin, 2016). This improved the AU-ROC on the validation set to 0.969. Finally, we fine-tuned the financial domain-specific language model FinBERT (768 dimensions) (Araci, 2019) for the downstream task of classifying definitions. It was trained for 20 epochs with a batch size of 256, maximum sequence length of 64 and a learning rate of 0.00002. This model out-performed all the other algorithms (**AU-ROC** = 0.9927, **Matthews Correlation Coefficient** = 0.9063, **Accuracy** = 0.9540 and **F1 Score** = 0.9610) on the validation set. The corresponding ROC curves are presented in Figure 6.

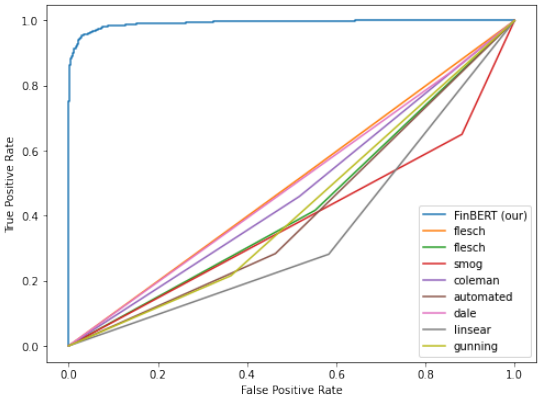

Figure 6: ROC curves

6. Conclusion

In this paper, we presented a new dataset **FinRAD** for the task of evaluating the readability of **definitions** of **financial terms**. We explored the limitations of various standard formula based readability scores which were developed to assess the readability of English texts in general. Finally, we proposed a neural architecture that outperformed all such scores in terms of AU-ROC.

There are several directions in which this research can be extended in future. We present some of the research questions (RQ) here.

- **RQ1:** *Do the predicted readability scores correlate with human judgements?*
  To understand this, we need to perform a qualitative analysis of the predicted readability scores generated automatically using machine learning algorithms. This may need additional manual tagging which is subjective and expensive. If the correlation is less, it would be essential to manually tag more definitions before developing any machine learning based classifier.

Readability Score (RS)	RS Description	AU-ROC	F1	Accuracy
flesch_reading_ease	The Flesch Reading Ease formula (Flesch, 1948)	0.4986	0.5516	0.5034
flesch_kincaid_grade	The Flesch-Kincaid Grade Level (Kincaid et al., 1975)	0.4320	0.4573	0.4296
smog_index	The SMOG Index (Mc Laughlin, 1969)	0.3841	0.5661	0.4250
coleman_liau_index	The Coleman-Liau Index (Coleman and Liau, 1975)	0.4710	0.4995	0.4691
automated_readability_index	Automated Readability Index(Smith and Senter, 1967)	0.4100	0.3494	0.3906
dale_chall_readability_score	Dale-Chall Readability Score (Chall and Dale, 1995)	0.4922	0.6793	0.545
linsear_write_formula	Linsear Write Formula[8]	0.3492	0.3295	0.3388
gunning_fog	The Fog Scale (Gunning FOG Formula)[9]	0.4259	0.2908	0.3936

Table 3: Performance of standard readability scores

Algorithms	Validation AU-ROC
TF-IDF vectors + Logistic Regression	0.9038
TF-IDF vectors + Random Forest	0.8866
TF-IDF vectors + Gradient Boosting Classifier	0.9116
BERT ST embeddings + Logistic Regression	0.9544
BERT ST embeddings + Random Forest	0.8801
BERT ST embeddings + Gradient Boosting Classifier	0.9063
FinBERT ST embeddings + Logistic Regression	0.9691
FinBERT ST embeddings + Random Forest	0.9434
FinBERT ST embeddings + Gradient Boosting Classifier	0.9523
FinBERT ST embeddings + Light GBM Classifier	0.9640
FinBERT ST embeddings + XGBoost Classifier	0.9626
FinBERT (fine-tuning [CLS] token)	**0.9927**

Table 4: Performance of models trained using Machine Learning

- **RQ2:** *Can we have better metrics to measure the performances of the models?*
  Presently, we use the Area Under the Receiver Operating Characteristic curve (AU-ROC) to measure the performance of the models. An interesting direction would be to develop a new metric that correlates more with human judgements.
- **RQ3**: *Can we develop unsupervised formulae based readability scores specific to the financial domain?*
  Machine learning based supervised models are computationally expensive and needs lots of data. Thus, it would be nice to explore if we can generate unsupervised formulae based readability scores specifically for the financial domain.
- **RQ4**: *Can we use Natural Language Generation methods to simplify definitions?*
  We removed duplicate terms while creating the `FinRAD`. A dataset consisting of readable as well as non-readable definitions for a given term would complement this. Simplification of complex definitions using Natural Language Generation techniques could be a new dimension to this research.

7. Bibliographical References

Anderson, R. C. and Davison, A. (1986). Conceptual and empirical bases of readability formulas.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

Arora, S. and Chauhan, Y. (2021). Do earnings management practices define the readability of the financial reports in india? *Journal of Public Affairs*, n/a(n/a):e2692, 05.

Banks, E. (2010). *The Palgrave Macmillan Dictionary of Finance, Investment and Banking*. Palgrave Macmillan, London, UK.

Bodie, Z. and Kane, A. (2020). Investments.

Bonsall IV, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357.

Brealey, R., Myers, S., and Allen, F. (2019). *Principles of Corporate Finance*. Economia e discipline aziendali. McGraw-Hill Education, USA.

Bruce, B., Rubin, A., and Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, PC-24(1):50–52.

Chakraborty, S., Nayeem, M. T., and Ahmad, W. U. (2021). Simple or complex? learning to predict readability of bengali texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12621–12629, May.

Chall, J. and Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, USA.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chiang, W.-C., Englebrecht, T. D., Phillips Jr, T. J., and Wang, Y. (2008). Readability of financial accounting principles textbooks. *The Accounting Educators' Journal*, 18:47–80.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Collins-Thompson, K. and Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Ganguly, A., Ganguly, A., Ge, L., and Zutter, C. (2019). Shareholder litigation and readability in financial disclosures: Evidence from a natural experiment.

Ghosh, S., Sengupta, S., Naskar, S. K., and Singh, S. (2021). Finread: A transfer learning based tool to assess readability of definitions of financial terms. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON): System Demonstrations*, Silchar, India, December. NLP Association of India (NLPAI).

Gosselin, A.-M., Le Maux, J., and Smaili, N. (2021). Readability of accounting disclosures: A comprehensive review and research agenda*. *Accounting Perspectives*, 20(4):543–581.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York, April. Association for Computational Linguistics.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282 vol.1.

Hoffmann, A. O. and Kleimeier, S. (2021). Financial disclosure readability and innovative firms' cost of debt. *International Review of Finance*, 21(2):699–713.

Hull, J. C. (2003). *Options futures and other derivatives*. PearsonPrentice Hall, Boston, USA.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

Kim, C., Wang, K., and Zhang, L. (2019). Readability of 10-k reports and stock price crash risk. *Contemporary accounting research*, 36(2):1184–1216.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Lo, K., Ramos, F., and Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, 63(1):1–25.

Loughran, T. and McDonald, B. (2009). Plain english, readability, and 10-k filings.

Loughran, T. and McDonald, B. (2010). Measuring readability in financial text.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *the Journal of Finance*, 69(4):1643–1671.

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

Mishkin, F. S. and Eakins, S. G. (2006). *Financial markets and institutions*. Pearson Prentice Hall, Boston, USA.

Othman, I. W., Hasan, H. H., Tapsir, R., Rahman, N. A., Tarmuji, I., Majdi, S., Masuri, S. A., and Omar, N. (2012). Text readability and fraud detection.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Plucinski, K. J. and Seyedian, M. (2013). Readability of introductory finance textbooks. *Journal of Financial Education*, 39(1/2):43–52.

Plucinski, K. J., Olsavsky, J., and Hall, L. (2009). Readability of introductory financial and managerial accounting textbooks. *Academy of Educational Leadership Journal*, 13(4):119.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Ruohonen, J. (2021). Assessing the readability of policy documents on the digital single market of the european union.

Rush, R. T. (1985). Assessing readability: Formulas and alternatives. *The Reading Teacher*, 39(3):274–283.

Samuelson, P. and Nordhouse, V. (2009). Economics: a textbook.

Schroeder, N. and Gibson, C. (1990). Readability of management's discussion and analysis. *Accounting Horizons*, 4(4):78–87.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 523–530, USA. Association for Computational Linguistics.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 574–576, New York, NY, USA. Association for Computing Machinery.

Smith, E. A. and Senter, R. (1967). Automated readability index.

Smith, J. E. and Smith, N. P. (1971). Readability: A measure of the performance of the communication function of financial reporting. *The Accounting Review*, 46(3):552–561.

Zamanian, M. and Heydari, P. (2012). Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1):43–53.

8. Language Resource References

Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh. (2022). *FinRAD: Financial Readability Assessment Dataset - 16,000+ Definitions of Financial Terms for Measuring Readability*. distributed via GitHub: `https://github.com/sohomghosh/FinRAD_Financial_Readability_Assessment_Dataset`.