Text readability

Abstract

The proposed project in this paper will aim to determine how easily readable a financial document is, using Natural Language Processing and Machine Learning. In today’s information-rich world automatically evaluating the level of readability of a certain document should be a priority because it would increase productivity substantially, especially in the financial domain.

The target financial document for this project will be annual reports published by companies. Annual reports are documents that are manufactured by a company to deliver important corporate information to its shareholders. If an annual report’s readability is easy, the shareholders are less prone to error when they are analyzing it. The motivation for this project is to improve the communications between shareholders and financial analysts by predicting accurately the readability of a financial document, namely annual reports, using state-of-the-art Machine Learning methods, Natural Language Processing and Computational Linguistics.

The machine learning models will be trained on outputs by CFIE-FRSE tool which extracts sections from annual reports and will make it possible to examine critical information from the reports.

1. Introduction

Producing a readability score for a given document would indicate the probability of extracting the correct information from text by the reader [8].

Research has shown that easy to read texts improve comprehension, reading speed, and reading persistence [1]. In the finance domain, if shareholders knew the percentage of readability of a given annual report, they would have a better idea how to approach the document. For example, shareholders and potential investors will be able to assess the documents more accurately and will be more likely to invest in companies whose financial disclosures are not buried in legal jargon and difficult language [2].

It is suggested that the management of companies that are not performing well tend to hide bad news by decreasing the readability of financial reports [3]. In his paper [4], Feng Li discovered that lower earnings firms’ annual reports are harder to read. Not only that but firms with annual reports that are easier to read have more overall persistent positive earnings. The lower earnings reports tend to have a higher Fog index and are generally longer [4]. If shareholders know the percentage of readability of the report beforehand, they would know if they must be extra careful.

The proposed project will aim to produce a readability score for a given financial document with high percentage of accuracy which goal is to allow shareholders and others to formulate more educated decisions. There are many tools that can be applied to measuring readability of text, but some are better than others. The Coh Metrix 2.0 version-Flesch Reading Ease Score Formula has been proved to measure text cohesion and text difficulty with higher accuracy than, for example, Flesch-Kincaid [3]. All readability formulas include a relatively simple mathematical equation that is tuned using a small set of documents on different reading levels [5]. However, an accurate readability analysis requires a more complex model. In addition, this project will attempt to use Machine Learning and Computational Linguistics to utilize linguistic features such as named entities, part of speech words and grammatical structure.

The project proposal will be split into the following sections: background, the proposed project, the program of work, the required resources and references. The background section will contain a description about related projects and existing systems involved in analyzing readability of text. The proposed project section will contain the project’s overall aim, the software development life cycle, the main objectives, the dataset and the potential machine learning models that will be attempted. The program of work will describe the project plan, breaking down the schedule of tasks for the year. The resources required section will detail the resources needed within the project and the refences section will contain references to any resources and papers used in this proposal.

1. Background

Background literature for this paper comes from a couple of areas of research: measurement of text readability in general texts, measurement of text readability in financial disclosures, fraud detection and manual (tuned mathematical formulas) versus automatic analysis (using CL and ML) of business and general documents.

The variety of projects that examine general text readability use the Fog index which has shown poor results when used in the financial domain [6]. Hence, this project will not be including the Fog index in any of the calculations or training of the models. One critique regarding Computational Linguistics is that algorithms are not able to interpret natural language accurately. However, this critique is not credible because other research has shown that algorithms *can* decipher context in sentences [7].

Sentence-length and word-complexity are two features that have been widely used to produce a readability score. However, in financial context, there are many words that have more syllables but should be easily understood [6]. In addition, sentences in financial disclosures can be lengthy but that does not necessarily mean it is hard to read because it could be enlisting different quantitative outcomes.

Many related papers use either Machine Learning or Computational Linguistics methods but rarely combine them. This project will take the approach of combining methods from both fields with the goal of seeing how effective it would be to analyze financial reports. [5] concludes that using natural language processing and machine learning proved to be more useful than using standard formulas alone.

The benefits of using Computational Linguistics methods instead of static and tuned formulas are clearly outlined in many previous projects. Automated text/content analysis offers the following advantages: reduced data collection costs, increased statistical power via large sample sizes, generalizability and improved objectivity and replicability [7].

The proposed project will use similar methods outlined in the listed sources in [3] because fraud prevention and assessing text readability require similar, if not the same, approaches. Previous projects use similar feature categories which can be summarized to lexico-semantic, morphological, cognitive, syntactic, semantic, and discourse [8]. Each of these features is responsible for different aspect of text analyzation – some are responsible with the parsing of the text as an action, while others are responsible for measuring the text cohesion (Coh-Matrix is one such example).

1. The proposed project
   1. Aims and Objectives

The goal of this project is to create a tool that manages to analyze the text readability of annual reports from companies with high accuracy. The following objectives will need to be met:

* The annual report to be analyzed need to be tokenized. That means separating a piece of text into smaller units called tokens. In addition, there can be word-level, character-level and subword-level (n-gram) tokenization. The problem with word-level tokenization is the Out of Vocabulary (OOV) words. These are words that refer to new words which are encountered at testing but not at training. The other issue with this approach the vocabulary size would become enormous. It seems that the most optimal way of tokenizing large texts (such as financial reports) is using subword (n-gram)-level tokenization. This tokenization technique does not transform most common words and it decomposes rare words in meaningful units.
* To train the machine learning models, corpora is required. Preferably, the corpora should consist of annual reports and not any other financial documents to get as accurate results as possible.
* The CFIE-FRSE tool [9] will be used to summarize sections of reports. This will help speed up the process of the training because the model can be trained on specific parts of the reports and not the whole text.
* A best-accuracy machine learning model will need to be determined. The summarization of the annual reports will not only be used to train the model but also to evaluate which one performs better than the others.
  1. Methodology

**Software Development Approach**

The exercised software development approach will be the spiral method/approach for two reasons. First, given the nature of the project, there will be need for constant prototyping and assessing used tools, algorithms and models. Second, the risk analysis step (which is the most significant development stage for this approach) will be extremely useful because a variety of models will be trained, and they need to be carefully analyzed and thoroughly tested. Even if a model proves to be faster than others, it needs to be analyzed what would be the tradeoffs. This unpopular software development life cycle method is almost identical with the waterfall model but with a very high emphasis on risk analysis which is much desired for the proposed project.

**Dataset**

When the dataset has been loaded, there will be 3 main parts that will be separated from it that will have different purposes:

1. *Train data* for training the model
2. *Validating data* for evaluation and hyperparameters tuning
3. *Test data* for final evaluation of the models

**Libraries**

There are a couple of libraries that will be used:

* NLTK
* TensorFlow
* Pandas and NumPy
* SpaCy

It is important to mention that during the implementation of the project, more libraries may be used. The above-listed are included in the proposal because they will most certainly be used.

**Deep Learning Models**

Before specifying the models that will be considered for implementing, it is important to briefly describe what Deep Learning is. Deep learning is a class of machine learning and uses machine learning methods that are based on artificial neural networks whose learning can be 3 types: supervised, semi-supervised or unsupervised [1-]. Thus, there is a wide range of neural networks that can be used such as DNN (deep neural network), RNN (recurrent neural network) and CNN (convolutional neural network).

RNN is good at understanding and classifying textual data. However, the main use cases for this type of artificial network are sentiment analysis and text generation which will not be useful for the proposed project and therefor will not be considered for implementation.

CNN is mostly used in the image and video processing, but it would be interesting to see how effective it would be in textual context. Theoretically, it would be able to predict if there are hard-to-read words in each sentence and output them. Therefore, this type of neural networks will be considered for implementation.

DNN has been used numerous times for contextual entity linking, writing style recognition and text classification. All those characteristics will be useful for text readability analysis. Thus, one of the models that will certainly be tested is DNN.

**Machine Learning Models**

Decision Tree (DT) algorithms seem to cope well with high dimensional data but requires a large amount of data to be trained on. Luckily, the proposed project will work with large amount of available data and therefore can capitalize on that by considering DT algorithms.

Regression models are used for predicting a real value. In the scope of this project, this value would represent the readability score. Regression techniques vary from Linear Regression to SVR and Random Forests Regression. The regression models that will be considered for implementation will be:

* Simple Linear Regression
* Polynomial Regression
* Support Vector for Regression
* Decision Tree Classification
* Random Forest Classification

1. Program of Work

The proposed project will begin late October 2020, running until March 2020 and it will be split into the following parts:

* Analysis and design – this step will involve designing an application that meets the project goals and evaluating what the system should do. In addition to that, a data set needs to be selected. This stage will take 1 week.
* Application Development and Testing – this step will involve developing the application based on the design. The models will be thoroughly tested to be able to evaluate which works best. All the scores and trade-offs of the tested models will be documented. This stage will take 7-8 weeks.
* Evaluation of models – during the Christmas break (4 weeks), the documentation of the scores and trade-offs of models will be carefully analyzed and will be decided upon which model to use.
* Integration of new chosen model – after the model has been chosen, it will be integrated into the program and tested again. This stage will take 1 week.
* Risk analysis – a careful review of the system and identification of potential risk is done. This stage will take 2-3 weeks.
* Model improvement – new parameters will be attempted to possibly improve the current results of the model. This stage will take 1-2 weeks.
* Risk analysis – will involve reviewing the model and determine if there is overfitting after the new fine-tuned parameters. This stage will take 1 week.

1. Resources Required

Access to financial reports published by companies. The annual reports will be in English and they are extracted from UK firms listed on the London Stock Exchange (LSE).

References:

1. *New Readability Measures for Bangla and Hindi Texts*- <https://www.aclweb.org/anthology/C12-2111.pdf> - Accessed 29th September 2020
2. *Measuring Readability in Financial Text -* <http://securitieseditor.com/wp/wp-content/uploads/2014/05/Plain_English_v5.pdf> Accessed 30th September
3. *Text readability and fraud detection* - <https://ieeexplore.ieee.org/abstract/document/6422890> - Accessed 30th September 2020
4. *Annual report readability, current earnings, and earnings persistence* - <https://reader.elsevier.com/reader/sd/pii/S0165410108000141?token=C74F7263D3051FCCA0FF2F21CB0600965BAB5CDEB62013387FDA73B08AF192862DD15613C0BA67BB67573D44B9A419D4> – Accessed 1st October 2020
5. *Readability Classification: Combining the Power of Machine Learning and Natural Language Processing* - <https://www.wizenoze.com/language/en/readability-classification/> - Accessed 1st October 2020
6. *Measuring Readability in Financial Disclosures* - <https://onlinelibrary.wiley.com/doi/full/10.1111/jofi.12162?saml_referrer> – Accessed 2nd October 2020
7. *In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse -* [*https://onlinelibrary.wiley.com/doi/full/10.1111/jbfa.12378#*](https://onlinelibrary.wiley.com/doi/full/10.1111/jbfa.12378) *-* Accessed 3rd October 2020
8. *Automated readability prediction -* <https://medium.com/voice-tech-podcast/automated-readability-prediction-71d40b677563> - Accessed 3rd October 2020
9. *CFIE*-FRSE tool - <https://github.com/drelhaj/CFIE-FRSE> - Accessed 4th October 2020
10. *Different Types of Deep learning models Explained -* [*https://roboticsbiz.com/different-types-of-deep-learning-models-explained/*](https://roboticsbiz.com/different-types-of-deep-learning-models-explained/) *-* Accessed 7th October 2020