**Predicting Reading Ages**

**Abstract**

This paper suggests a hybrid model of GloVe, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN), alongside traditional machine learning algorithms to perform the multi-classification of the reading ages to predict the reading age. The reading ages are numerical classes which range from 6 to 18, where a reading age of 6 means a 6th grader will understand the text and a text with a reading age above 17 will need advanced knowledge and understanding, usually above a college graduate. The data was created by splitting the speech of the Ex-British Prime Minister -Theresa May at a 2018 conference[[1]](#footnote-1) and was then pre-processed to remove unnecessary information. Normalization and data augmentation was used to balance the data and to convert it into a structure understandable by the model. The Evaluation considered each model's precision, Recall and F1-Score with GloVe-CNN-LSTM performing the best with a Precision, Recall of 0.95 and F1-Score of 0.94. The experiment showed that Natural Language Processing (NLP) can be helpful in predicting the reading age of text.

Keywords: Long Short-Term Memory (LSTM); Convolutional Neural Network (CNN); Word Embeddings; GloVe; Deep Learning; Multiclass text classification; Natural Language Processing (NLP), Hybrid Models, Traditional Machine learning algorithms

**1.0 Introduction**

In general, literacy rate refers to the percentage of people in a given age group who can read and write. An adult literacy rate corresponds to those 15 and older, a youth literacy rate to those 15 to 24, and an elderly literacy rate to those 65 and older. In standard measurements, comprehension is determined by how well one comprehends a simple statement from their daily lives. As a general rule, literacy is a term that encompasses numeracy as well, and measurement may also include assessing arithmetic ability. (Unesco Institute for Statistics, n.d.).

Reading is important as it can be used as one of the tools to attain your personal goals or academic success. The reading age of an individual is the index estimated in the years of education an individual requires to assimilate a text upon reading it for the first time. Knowing the reading age of an individual, especially children can be helpful to understand the child more and allows the educator or education body to develop methods to assist the child.

Reading is an essential skill that is necessary for academic and personal success. However, not all individuals possess the same reading abilities. Some people may read very well at a young age, while others may struggle even in old age. Predicting the reading ages of individuals is important for the following reasons:

* Early Identification of Reading Difficulties: This will help in identifying children with reading difficulties or disabilities (Eberhard et al., 2021).
* Educational planning: This will help schools or educators plan and allocate resources for children in need (Eberhard et al., 2021).

The UK government also published a report examining children's and young adults' enjoyment of reading. According to the report, 49% of 8-16-year-old children enjoyed reading to some extent in 2010, with 22% reporting that they did so greatly (Gov.uk, 2012). Hence, developing a love of reading in children may be a key factor in supporting their reading development.

There have been technological advancements to aid the prediction of the reading ages of individuals. One of these technologies is a Python library known as Textstat, which helps determine the readability, complexity, and grade level of text (Python Package Index, n.d.).

In addition to the main aim of this study, which is to predict the reading ages of text, we will also investigate and analyze the effectiveness of using natural language processing techniques to predict an individual’s reading age based on their written text without the use of already made Python’s library Textstat. This will be analyzed by comparing the results of the different machine learning algorithms and deep learning models and evaluating the best-performing model.

**1.1 Related Works**

Research has shown that there is little or no related study as regards predicting reading ages using machine learning or AI. The closest study found was published in 2021 by Sinclair et al. (2021). The authors used linguistic features such as vocabulary, syntax, and semantics to predict reading comprehension levels in children, this was accomplished using a file (either audio or text) which uploaded, and the linguistic features were extracted and output as a single-row table in a new document, which was finally converted to the dataset used. The study found that machine learning models trained on extracted linguistic features could accurately predict children's reading comprehension scores and concluded that support vector machine (SVM), especially popular for solving binary classification problems, where data points are classified into one of two groups, performed better when comparing the mean absolute error against other algorithms like gradient boosting and random forest, with an MAE of 17.21 on text elicitation data while gradient boosting performed better with an MAE of 18.09 on oral elicitation data.

Natural language processing algorithms like GloVe and Long Short-Term Memory (LSTM), deep learning models, and machine learning algorithms have also been applied in analyzing text such as text classification analysis. In businesses, it is used to categorize and extract valuable insights from unstructured text data, such as emails, news articles, customer interactions, or social media content for sentiment analysis (Sadekov, 2023). Ismail & Yusoff (2022) used data gotten from twitter to construct a Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and bidirectional LSTM model. CNN had the highest accuracy of the four algorithms with a score of 0.98 for test accuracy, 0.95 for precision, 0.94 for recall, and 0.95 for the f1-score. Due to their competitive performance and computational efficiency, CNNs are a leading candidate when classes are relatively balanced, according to Lu, et al. (2022).

Moreso, the study by Harris (2019) aimed to develop a method for automatically identifying student frustration and difficulty in online courses using discussion forum postings to provide timely and targeted support to students in online courses. Using different algorithms like SVM, Decision tree, Random Forest, Naïve Bayes and so on, he concluded that SVM performed the best with an average accuracy of 81.86% and an average f1-score of 0.82.

The study by Banerjee, et al. (2019) aimed to compare the performance of two deep learning architectures, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), for classifying radiology text reports and concluded that deep learning models, CNN Word - GloVe and Domain Phrase Attention-Based Hierarchical Recurrent Neural Network (DPA-HNN), outperformed the state-of-the-art rule-based system and traditional machine learning methods (SVM and Adaboost) when trained on a single institutional dataset with an F1-score of 0.99, indicating high accuracy in identifying Pulmonary Emboli (PE) cases.

Furthermore, according to the findings of Asudani, et al (2023), domain-specific word embedding together with the LSTM model can be used to improve overall text analytics performance. and achieves an accuracy ranging from 90.00% - 99.50% based on the work cited in the study.

There are other types of neural networks that can be used for text analysis or classification tasks such as predicting the reading age such as Recurrent Neural Networks (RNNs) which are useful for sequential data and MLP (Multilayer Perceptron) networks which are useful for non-sequential data but Convolutional neural network (CNN) beats all these methods in terms of speed (Brownlee, 2022).

Overall, the data and requirements of the task determine the effectiveness of a model. Hence, in this study, we will be testing the effectiveness of using NLP techniques like LSTM, GloVe-LSTM and GloVe-CNN-LSTM on our dataset and comparing the performance against traditional machine algorithms like support vector machines (SVM), gradient boosting, XGboost etc. The NLP Models will leverage the long-term memory of the LSTM, the pre-trained word embedding of GloVe and CNN to extract local features from input text to improve the performance of the models.

**1.2 Data**

Although no specific location is available to get the data needed for this project including listing the text and the corresponding reading age for the text. We obtained the sample text from the British political speech[[2]](#footnote-2), an archive for political speeches. The speech used was given by Theresa May, the Ex-British Prime Minister and leader of the conservative party at the 2018 Conservative Party Conference, a speech containing over seven thousand words (7074). The text was then split into sentences using Python into a list, the reading age was also calculated using the Fog index[[3]](#footnote-3) of Textstat[[4]](#footnote-4), to get the baseline of our data. “The index estimates the years of formal education a person needs to understand the text on the first reading” as described by (Wikipedia, n.d.), The index ranges from 6 to 17, where 6 is for a 6th grader and 17 is for a college graduate. The result was then used to get this project’s dataset. The dataset comprises of 533 samples and 2 features listing the Sentence and the fog index which is the reading age of the text. Figure 1.0 below shows the dataset:

A screenshot of a computer

Description automatically generated

**Fig. 1.0: The Dataset**

The figure above shows the structure of the dataset, having 533 samples and two features: Sentence and Fog Index. The sentence contains the political speech separated per sentence, while the Fog Index contains the reading age of each sentence.

**Data Utilization**

As part of our research, we used the data to predict the reading ages of the text based on the fog index of our dataset, we utilized a variety of machine learning and data analytics techniques to extract meaningful insights and develop predictive models based on the data. Different machine learning techniques were used to determine which technique is best for predicting the reading ages. This will be emphasized more in subsequent sections of this paper.

**Data Cleaning/Pre-Processing**

Different cleaning processes were considered to make sure the data was cleaned optimally. The steps are listed below:

* Handling missing values: Since the data was gathered from the text of political speeches and the fog index of Textstat, the dataset had no missing values available.
* Feature selection: The dataset is comprised of just the features we require, which are the text and the fog index of the text.
* Data transformation: The features required transformation to improve distributional properties and normalization. We Tokenized the sentences and padded the sequences created from the tokenization to prepare our input for training.
* Handling outliers: During the cleaning process we noticed that the target variable “Fog Index” had outlier(s) that could affect the outcome of our model, this was handled using transformation techniques by converting reading ages above 17 to 18 and converting reading ages lower than 6 to 6. The outliers can be spotted in Figure 1.1 below:

A graph with a blue rectangle

Description automatically generated

**Fig. 1.1: Outliers**

The figure above shows the outliers present in the dataset, with reading ages going above 17 which were taken as outliers and treated using transformation techniques as stated earlier.

**Balance of the data**

In terms of class distribution or target variable (Fog Index) distribution, the dataset was unbalanced; this can be seen in Figure 1.2 below. When a dataset is balanced, it indicates that the classes or target variable categories are distributed fairly and equally, whereas when it's unbalanced, it indicates that they're not distributed evenly.

Techniques such as Randomoversampler from imbalanced learn[[5]](#footnote-5) were used to balance the dataset so that the model will not be biased towards just one class alone. Figure 1.3 below shows the balanced dataset:

A graph of different colored bars

Description automatically generated

**Figure 1.2: Unbalanced Dataset**

The Figure above shows the level of unbalance each class has compared to the other, with the reading age 6.0 having the highest count while the reading age 17.0 having the lowest which was balanced using Randomoversampler in Python as pointed out earlier.

**2.0 Methodology**

To achieve the purpose of this research, which is to be able to predict the reading ages of text, we adopted a combination of different machine learning algorithms and deep learning models. The selected methodology is designed to effectively process and extract insights from the data, as well as develop predictive models and at the end be able to select the model which is best in predicting the reading age of text.

The architecture used below was gotten from related works done on text analysis which suggests that deep learning and NLP techniques like Long Short-Term Memory (LSTM), GloVe and Convolutional Neural Networks increase the performance level of a model when trained using them, this has been highlighted in the related works section of this study. The architecture of the deep learning models used can be seen in Figure 2.0 below:

A diagram of a layer

Description automatically generated

**Figure 2.0: Deep Learning Model Architecture**

The figure above shows the approach taken to analyse the dataset using deep learning models. The dataset was cleaned and pre-processed into acceptable formats for the model, and then after splitting, the training dataset is used to train the model and the test is used to test the accuracy of a new set of data, before we achieve our results.

The methodology considered followed the following steps:

* **Data gathering and preprocessing**: The data was gathered from an archive of political speeches[[6]](#footnote-6) and by classifying the speech into sentences and using Textstat[[7]](#footnote-7) to get the reading ages for the sentences, the dataset used for this research was created. This has been described in detail in the Data section of this paper.
* **Model Development:** The dataset was split into training and testing data using a percentage of 80:20. i.e., 80% for the training data and 20% for the test data. We selected the following algorithms and deep learning models based on their suitability for our research objectives and their proven effectiveness in similar studies:
  1. **LSTM:** Learning and retaining information over a long-time lag between relevant input events and target signals is challenging for traditional RNNs. LSTM resolves this issue by capturing dependencies over extended periods and retaining relevant information for predicting outputs (Browniee, 2017). This is the purpose of using LSTM as one of the NLP models.
  2. **GloVe-LSTM:** In GloVe, word vectors of different dimensions (e.g., 50d, 100d, 200d, 300d) are generated from large texts such as Wikipedia, which are pre-trained using large text. Using these pre-trained word embeddings, machine-learning models can improve their NLP performance (Lendave, 2021). These pre-trained embeddings can be used as inputs to an LSTM model to leverage the knowledge encoded in them, enabling faster convergence and improved generalization. For this paper, the 100d word vectors will be used.
  3. **GloVe-CNN-LSTM:** A hybrid model combining GloVe word vectors with CNN and LSTM networks can better represent input text, capturing local and global semantic information. It has been successfully applied to tasks like sentiment analysis, text classification, and gender violence categorization, achieving high accuracy, precision, recall, and f1-scores (Ismail & Yusoff, 2022).
  4. **Traditional Algorithms:** Traditional machine learning algorithms like gradient boosting classifiers, Random Forest classifiers, Decision tree classifiers, XG boosting classifiers and an ensemble model of gradient boosting classifiers, Random Forest classifiers and Decision tree classifiers were all used to classify our classes to predict the reading ages for each text.

After the development of the model using the algorithms and deep learning methods listed above, and after observing that the dataset was unbalanced, metrics like the F1-score and precision were considered more compared to the accuracy to measure the model's performance. The performance was also assessed by comparing the result against the baseline generated from Textstat. This should in turn produce or develop predictive models that accurately capture the data's underlying patterns and relationships; models that will help classify and predict the reading ages based on the text provided. Also, the model will be able to provide insights into identifying meaningful features, significant variables, or relevant clusters that shed light on the factors influencing the prediction of reading ages.

**Note**: The algorithms chosen were based on the inference that the research was seen as a classification problem. If the problem was seen as a regression problem the traditional algorithms might change and the results different.

**3.0 Results**

The performance of the models created using different algorithms and deep learning models was assessed using the following evaluation metrics:

**F1-Score:** It is a machine learning metric that emphasizes class-wise performance rather than overall performance like accuracy assesses a model's predictive skill. A model's performance is measured through the F1 score, which combines precision and recall metrics (Kundu, 2022).

**Precision:** The precision of a prediction is a measure of how many positive predictions are made correctly. Therefore, Precision calculates the accuracy of minorities (Brownlee, 2022).

**Recall:** Out of all the positive predictions that could have been made, recall represents how many correct positive predictions were made. When compared to precision, recall provides an indication of missed positive predictions compared to all positive predictions (Brownlee, 2022).

The evaluation results for the algorithms considered (described in detail in the methodology section) can be found in Figure 3.0 below:

A screenshot of a graph

Description automatically generated

**Figure 3.0: Evaluation Results**

The table above shows the results of the evaluation metrics used to know the performance of each model with the f1-score considered more since it is the harmonic mean of both the precision and the recall. The table was sorted in descending order using the f1-score to get the algorithm with the highest f1-score. According to Figure 3.0, it was seen that the hybrid model of GloVe-CNN-LSTM performed best as it has the highest f1-score of approximately 0.94 compared to other models and a recall of approximately 0.95, meaning it correctly identified 95.00% of the positive instances and the precision also of approximately 0.95, which shows that 95.00% of the predicted positives were correct.

Other models did slightly lower in terms of the evaluation metrics shown above compared to the GloVe-CNN-LSTM with the least having an f1-score of approximately 0.90 which is still a good performance.

Further evaluation was done on the models by checking the predictions of these models against the baseline prediction from TextStat. This can be seen in Figure 3.1 below:

A table with numbers and text

Description automatically generated

**Figure 3.1: Results of The Actual Values and Predicted Values**

The table above shows the results obtained from each model after training. We can see that overall, all the models did well with just a few instances where the prediction was off by small numerical differences.

We went further to see the explanation for these values that are off using the best model GloVe-CNN-LSTM according to the f1-score by checking the confusion matrix and this can be seen in the Figures below:

A grid of numbers and letters

Description automatically generated

**Figure 3.2: Confusion Matrix**

**A graph with blue squares

Description automatically generated**

**Figure 3.3: Confusion Matrix Heatmap**

Figures 3.2 and 3.3 show the confusion matrix of the GloVe-CNN-LSTM model. The model classified 16 cases successfully as a reading age of 6, 10 cases as a reading age of 8 when it is 6, 2 cases as a reading age of 9 when it is 6, 1 case as a reading age of 10 when it is 6, and 1 case as a reading age of 18 when it is 6. All the other classes had good classifications except for the readings ages 6,8,9 and 12 which were off but slightly.

The confusion matrix indicates that the model achieved good classification performance, with all cases classified almost 94.00% correctly in their classes.

Using the results from our models, we can gain valuable insights into their performance, showing that the GloVe-CNN-LSTM model achieves the desired results more efficiently. Figure 3.4 below shows the performance of the model when used to predict the reading age of an entirely new data.

A table of information

Description automatically generated

**Figure 3.4: Reading Age Prediction**

The table above shows a list of text and their corresponding reading ages, it shows that according to the maximum reading age of 12, the texts can be understood generally by a wide range of audiences.

**4.0 Discussions**

The study shows that the LSTM and CNN model using GloVe word embedding pre-trained model (GloVe-CNN-LSTM) performs best overall compared to other models. When comparing the three (3) deep learning models used i.e., the LSTM model, the GloVe-LSTM model and the GloVe-CNN-LSTM, the LSTM model performed the least, but the performance was greatly improved when GloVe word embedding was introduced, before arriving at the best model of combining CNN to the GloVe-LSTM Model. This supports the study by (Ghelani, 2019) which suggest that CNNs are capable of classifying sentences and paragraphs, this makes them ideal for our NLP tasks to predict reading ages.

Following the study by Sinclair et al. (2021) which concluded that support vector machine and gradient boosting perform better compared to other traditional machine learning algorithms, the present study showed that although gradient boosting performed well, it was not as good as the deep learning models that had GloVe word embedding.

Furthermore, when checking the performance of the traditional machine learning algorithms, we noticed that they had perfect accuracy, precision, recall and f1-score on the training data with a score of 1.00, this could be because of over-fitting. To avoid this, methods like ensemble were used by combining the predictive capability of different models, the outcome was also a perfect score of 1.00. We then went ahead to test the performance on the test data and this yielded a good result ranging from 0.92 – 0.94 for the traditional algorithms which signifies that there was no overfitting since it performed well on a new set of data.

After the data was balanced, we could see from our results that CNN performed better compared to other models when trained on the dataset. These results are similar to those of Lu, et al. (2022) and even more on par with that of Ismail & Yusoff (2022), who concluded that CNN performed the best compared to the other deep learning models used. Although in our study, we noticed that the model didn’t classify the reading ages 6,8,9 and 12 too well, with 6 and 8 having the lowest f1-score of 0.63 and 0.79 respectively, while 9 and 12 still have good classifications of 0.90 and 0.93 respectively. Despite these small differences, overall, we can conclude that deep learning models can be quite effective when used in analysing text and the requirement of predicting the reading age of text.

**Limitations and Conclusions**

This study aimed to see the effectiveness of NLP techniques in predicting the reading age of text. During the experimentation and analysis, the following limitations were observed:

* Low classification of the reading ages/classes 6 and 8; better hyperparameter tuning can be done to increase the classification of these classes.
* Storage issue that might affect the model due to GloVe containing large word embeddings which will require space, this might affect or slow down deployment. Utilizing cloud storage for the storage of word embedding can be useful here.

The following conclusions were drawn:

* Pre-trained word embeddings like GloVe show effectiveness as the performance improved when trained with GloVe.
* GloVe-CNN-LSTM performed the best hence proving the effectiveness of NLP techniques in predicting the reading ages of text.

**Future Works**

Further hyperparameter tuning can be done to improve the performance of predicting the classes with lower f1-scores. The model could be then improved by creating an interface to expose the model to the public to test its performance in real-life and real-time scenarios.

# **References**

Alawneh, E., Al-Fawa'reh, M., Jafar, M. T. & Fayoumi, M. A. 2021. Sentiment analysis-based sexual harassment detection using machine learning techniques. *2021 International Symposium on Electronics and Smart Devices (ISESD),* pp. 1-6.

Asudani, D. S., Nagwani, N. K. & Singh, P. 2023. Impact of word embedding models on text analytics in deep learning environment: *Artificial Intelligence Review,* 56:10345–10425.

Banerjee, I. et al. 2019. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial Intelligence in Medicine,* 97:79-88.

Browniee, J. 2017. *A Gentle Introduction to Long Short-Term Memory Networks by the Experts.*    
Available at: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/  
[Accessed 16/07/2023].

Brownlee, J. 2022. *How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification.*    
Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/  
[Accessed 24/07/2023].

Brownlee, J., 2022. When to Use MLP, CNN, and RNN Neural Networks. [Online]   
Available at: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/  
[Accessed 17/08/2023].

Eberhard, A., Jost, L. B., Daum, M. M. & Maurer, U. 2021. Predicting Reading From Behavioral and Neural Measures – A Longitudinal Event-Related Potential Study. *Frontiers in Psychology,* 12:1-14.

Ghelani, S., 2019. Text Classification — RNN’s or CNN’s?. [Online]   
Available at: https://towardsdatascience.com/text-classification-rnns-or-cnn-s-98c86a0dd361  
[Accessed 17/08/2023].

Gov.uk, 2012. Research evidence on reading for pleasure, s.l.: Education standards research team.

Harris, S. 2018. *Identifying student difficulty and frustration from discussion forum postings,* s.l.: Athabasca University.

Imbalanced learn, n.d. Over-sampling methods. [Online]   
Available at: https://imbalanced-learn.org/stable/references/generated/imblearn.over\_sampling.RandomOverSampler.html  
[Accessed 16/07/2023].

Ismail, A. A. & Yusoff, M., 2022. An Efficient Hybrid LSTM-CNN and CNN-LSTM with GloVe for Text Multi-class Sentiment Classification in Gender Violence. *International Journal of Advanced Computer Science and Applications,* 13(9):853-863.

Kundu, R. 2022. *F1 Score in Machine Learning: Intro & Calculation.* [Online]   
Available at: https://www.v7labs.com/blog/f1-score-guide#:~:text=F1%20score%20is%20a%20machine%20learning%20evaluation%20metric%20that%20measures,prediction%20across%20the%20entire%20dataset.  
[Accessed 24/07/2023].

Lendave, V. 2021. *Hands-On Guide To Word Embeddings Using GloVe.*    
Available at: https://analyticsindiamag.com/hands-on-guide-to-word-embeddings-using-glove/  
[Accessed 16/07/2023].

Lu, H., Ehwerhemuepha, L. & Rakovski, C., 2022. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology,* 22(181):1-12*.*

Python Package Index, n.d. Textstat. [Online]   
Available at: https://pypi.org/project/textstat/   
[Accessed 17/07/2023].

Sadekov, K. 2023. *8 Most Important Natural Language Processing (NLP) Applications in 2023.*    
Available at: https://mindtitan.com/resources/blog/natural-language-processing-applications/  
[Accessed 31/07/2023].

Sinclair, J., Jang, E. E. & Rudzicz, F. 2021. Using machine learning to predict children’s reading comprehension from linguistic features extracted from speech and writing. *Journal of Educational Psychology,* 113(6):1088-1106.

Unesco Institute for Statistics, n.d. *Literacy rate.*    
Available at: https://uis.unesco.org/en/glossary-term/literacy-rate  
[Accessed 03/08/2023].

Wikipedia, n.d. Gunning fog index. [Online]   
Available at: https://en.wikipedia.org/wiki/Gunning\_fog\_index  
[Accessed 17/07/2023].

1. <http://www.britishpoliticalspeech.org/speech-archive.htm?speech=370> [↑](#footnote-ref-1)
2. <http://www.britishpoliticalspeech.org/speech-archive.htm?speech=370> [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/Gunning_fog_index> [↑](#footnote-ref-3)
4. <https://pypi.org/project/textstat/> [↑](#footnote-ref-4)
5. <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html> [↑](#footnote-ref-5)
6. <http://www.britishpoliticalspeech.org/speech-archive.htm?speech=370> [↑](#footnote-ref-6)
7. <https://pypi.org/project/textstat/> [↑](#footnote-ref-7)