Evaluation of Text Summarisation and Simplification Algorithms on Fictional Texts, Measured by Common Natural Language Processing Metrics and Human Judgement Metrics

UNIVERSITY OF BIRMINGHAM

By

Ethan Chang Liu

MSc project report submitted to the University of Birmingham for the degree of MSc in Data Science.

Student ID: 1799234

Supervised by **Jizheng Wan**

School of Computer Science

University of Birmingham

September 2022

Table Of Contents

Contents

List of Figures

List of Tables

ABSTRACT

Reading is not an innate skill, and therefore the consumption of published content could be inaccessible for individuals due to factors such as language barriers and learning based disabilities. From a social and economic perspective, one may ask if there is a way to automate commercial fictional text, using the current NLP technologies. Alice In Wonderland was chosen as the fiction dataset along with the 2005 Azores Subtropical Wikipedia page as the non-fiction dataset as the control group. SIMPLE guide implemented human simplification was compared with three competitive NLP processes of BART, PEGASUS and MUSS and were evaluated in three perspectives of machine level, human level and text statistics. It was observed that human simplification remains the best with MUSS a close second in both fiction and nonfiction, as well as overall low correlation between human and machine level metrics. Results however remain inconclusive due to overall data size and volunteer size because of time and resource constraint.

CHAPTER 1: Overview

1.1 Introduction and Motivation

Literacy, the physical alternative method of communication via reading and writing, is an ability deeply intertwined with humanity's progression through time.[1]

Manifestation of literacy has allowed humanity to thrive. From practical applications such as simple written contracts that legitimises trades; to giving birth to the study of history as a pursuit of knowledge; to enriching mankind through sharing of ideas and perspectives that resists against time. [2-5] The importance and complexity of literacy is never ending, and this is especially significant in the modern world convoluted by the 'information-age'.

Comparable to the ethos of medicine. An altruistic part of modern-day literacy progress is to help individuals that face difficulties in leveraging literacy for personal efficacy, thus directly affecting their health such as life expectancy. [6-8] In a wider scale, literacy efficacy is an important part that affects a country's economic status such as economic competitiveness. [9-11]

For project motivation, a U.K specific example is presented to demonstrate the social, economic significances and relevancies of this project. The filmography industry is an invaluable asset to the United Kingdom's economic and social development throughout the years, with more than £6.3 billion pounds in contribution to the U. K's GDP in 2016. [12,13] There exists an intrinsic positive link between the film industry and the book industry due to adaptations of literature work onto the big screens, and such statements that have been positively academically assessed before. [14]

Due to the easier accessibility of the film median, related publishing products do not fully benefit from the influx of potential consumers for when a new box-office hit adaptation has been released. One of the big contributors of this behaviour is due to the relatively high accessibility wall presented by the English language and its related semantics. For consumers that are non-native English speakers or have dyslexia, any publishing products related to a film they've experienced remains daunting.

Currently there are software-less production solutions to solve the problem stated above, in that different reading level versions of the same books have been released. [15] These re-written books are simplified by hand and therefore the production process is long and only mainstream works are worth the resources.

The area of natural language processing (NLP), the study of interactions between computers and human language, presents a candidate perspective into enhancing current manual solutions.

Whilst the focus of the project is towards fictional texts, the processing of non-fictional texts provides a foundation in processing of fictional texts. Presently, machine translation of non-fictional texts aims to simplify lexical and syntactical features. Therefore, the complexity of non-fictional texts is language bound and usually semantics are clear, concise and have little room for alternative meaning. As a result, there have been many successes in applying non-fiction NLP processes in real life scenarios such as automatic summarisation methods for notes and record summaries. [16,17]

An evolution of text processing problem now presents itself to simplify features beyond-sentence-level, high-level semantics that are abstractive in nature of which are all commonly found in fictional texts. Requiring prior context knowledge, experience, culture and other resources for the reader to be able to understand.

The significance of this project is to explore, analyse and evaluate current NLP machine learning process that could potentially contribute to fictional text processing. By gaining an understanding of

the current landscape in fictional text summarisation and simplification, potential directions could be formulated in closing the gaps between machine and human fictional text processing. Therefore partial / fully automation of simplification publishing products could be one step closer to be achieved, increasing market share of publishing products, and as well as altruistically helping those who wish to read literature works but otherwise cannot do so easily.

1.2 Aims and Objectives

The aim of this project is to analyse the performance of abstractive text summarisation, text simplification algorithms on fictional texts from both human and machine perspectives. Thereby gaining insight to the current NLP progress in fully automating text processing for fictional publishing products.

Project Objectives in Order to Fulfil Project Aims:

1. Exploration of the lexical, syntactical and beyond-sentence semantical features that defines fictional texts from non-fiction, hence deriving appropriately constrained datasets for machine learning processes.

2. Exploration of the natural language processing landscape, to derive representative algorithms to process datasets and common machine learning metrics to evaluate resulting output.

3. Exploration of text complexity landscape to derive human metric systems to evaluate obtained results.

4. Implementation of appropriate dataset as input to representative algorithms, followed by three level analysis to elucidate findings.

CHAPTER 2: Literature Review

2.1 Psychological Component

The aim of this section of the literature review is to comprehensively understand the complex mechanisms behind the actions of reading and writing, and to which how learning difficulties affect literacy efficacy.

By knowing the psychological component of literacy. This format is not only advantageous in, providing a deeper insight into the psychological design and effects of NLP solutions, but also forms a more cohesive narrative in how different disciplines intertwine to solve profound problems.

2.1.1 The Mechanisms Behind Reading and Writing

Unlike spoken language which is innate in humans, writing and reading are taught skills and therefore can by decomposed into sequential steps. [18] The current mechanisms of literacy therefore focuses within the area of child development to leverage their growing cognitive abilities, such as better non-native language sound distinction than their older peers. [19,20]

One of the most historically significant scientific theory to framework literacy is called the 'Simple View of Reading' (SVR), introduced by the psychology field 1986. [21] Since then SVR has been used widely in education and linguistics, from becoming the standard to which how reading is taught in classrooms to aiding research educators in professional development. [22-25]

Formalised by equation 1 below. Reading comprehension (RC), the ability to understand the meaning of written words, is a function of both decoding (D) and language comprehension LC.

$$(\boldsymbol{Eqn1})\ D \times LC = RC$$

Where:

D and LC are percentages between 0% and 100% inclusively.

Whilst it has been commented that the SVR equation form was only used to promote legitimacy.[26] The true value of the SVR is that it provides a sequential step to achieve reading comprehension. Firstly decoding, which is the ability to recognise or decode written words via phonics (sound). Secondly with language comprehension, the ability to understand recognise or decoded words. Understanding written texts cannot be accomplished without either step.

In summary with Figure 1 below, for an individual to understand a written sentence. The individual must first recognise each individual word written via phonics (decoding), and then to be able to understand what those words mean individually and sentence wise from their sounding (language comprehension).

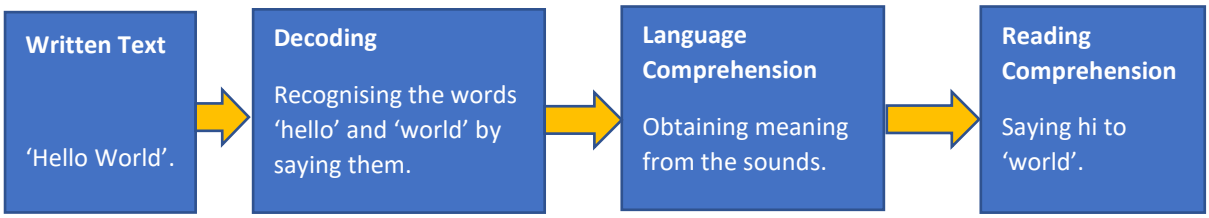


Figure 1. An example flow chart in how the written sentence 'Hello World' is processed via SVR

SVR proceeded over the decades to become a popular platform to which the Science of Reading (SOR) field has advanced from. Contributing to its prevalence, SVR has been tested and used in many studies that strengths its credibility. From factor analysis studies that determined decoding to be a distinct factor in young English-speaking children, [27,29] to live trials such as a metric evaluation of phonemic awareness program, reading comprehension diagnosis and more. [30-33]

One of the main directions of SOR in recent times is to address the main criticism of SVR, in that the SVR is a high-level framework that does not suggest what factors affect the SVR variables in theory or real life. Psychologically how does one practically influence decoding and language comprehension to affect reading comprehension?

Examples of SVR expansion includes; challenge of independent relationship between decoding and comprehension, explained by increase in reading variance through shared variance of the two variables [34-36]; additional psychological factors such as 'self-regulation' to manage the reading process and retain engagement [37]; and also linguistic factors such as morphological awareness, where meaning of un-seen words can be derived from similar words along with roots, prefixes and suffixes (such as drive and driver).[38]

Scarborough's Reading Rope Model (SRRM) was specifically used this report, to describe and explain the interactions of computer science solutions towards literacy problems.[39] Based upon SVR, SRRM is a comprehensive expansion that incorporated many additional factors created over the SOR field progression, thus providing a more applicable and practical but also verified perspective into reading comprehension.

Noticeably as seen in figure 2 below, many factors that affect both language comprehension and word recognition (decoding) are represented by real life physical features.

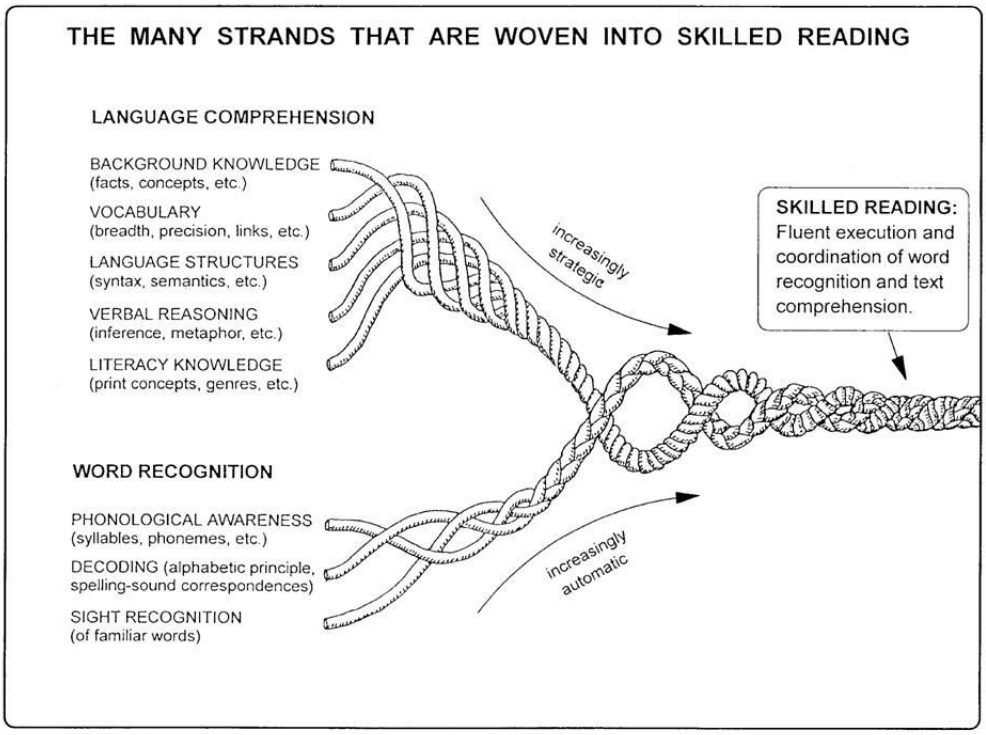


Figure 2 An infographic of Scarborough's Reading Rope Model.[39]

2.1.2 Language-Based Learning Disability Described Via SRRM

Dyslexia is a common language-based learning disability and is estimated that 1 in every 10 people in the U.K are affected by it.[40] One of the man ways to explain dyslexia as described by the British Dyslexia Association (BDA), is that individuals have difficulties in processing and remembering information they see and hear.[41] This definition is then further expanded upon by the famous independent Rose report in 2009 to which specified difficulties in phonological awareness (the awareness of sounds in a language) , verbal memory and verbal processing speed.[42]

By describing dyslexia via SRRM, there is a link between physical properties of writing and reading and their influences on the psychological theory of literacy. For an overview, dyslexia can be considered as difficulties in the word recognition and language comprehension variables within the SRRM.

Demonstrated by table 1, many common examples of symptoms of dyslexia can then be considered as factors that affects the SRRM variables. It is through this physical and psychological framework that this report can describe and explain NLP solutions. Specifically, how NLP solutions aim to rectify the physical symptoms of dyslexia in literacy, and their effects on the psychological model that defines reading and writing.

Physical Symptoms of Dyslexia in Literacy	Literacy Psychological SRRM Variables Influenced? Word Recognition (WR) or Language Comprehension (LC)	Specific Factors That the Physical Symptoms Belong to That Contributes to the SRRM Variables.
Confusion of similar looking letters or words such as 'b' and 'd'.	WR	Decoding and Sight Recognition.
Read and write very slowly.	Both	Individual could struggle in decoding from WR that causes them to read very slowly but with no problems in LC where they understand the meaning. Individuals could struggle in language structure from LC where they know what to write but struggle to implement it constrained to English grammar.

Table 1 . An example of how several dyslexia symptoms can be described via SRR. Symptoms take from the official NHS website.[43]

The study of dyslexia is still a thriving field, such as advances in neurology to determine which part of the brain is affected, [44,45] and neurobiology concepts such as describing dyslexia via attention.[46] For this project, only the literacy symptomatic aspect of dyslexia is considered. Specifically, as validated by many previous studies to influence reading comprehension of dyslexic individuals, the lexical, semantical, syntactical and language complexity of texts. [47-51]

2.2 Computer Science Component

The aim of this section is to provide a guided tour of the currently NLP landscape in improving reading comprehension amongst individuals. From the previous psychological component of the literature review, NLP approaches towards this problem is vast and numerous, covering all aspects of the SSRR model.

2.2.1 Natural Language Processing Landscape

NLP is a popular branch of computer science that seeks to enhance a computer's ability to understand, and to process the human language to solve specific problems.[52] Therefore, using the SRR model, the aim of NLP in project scope is to maximise reading comprehension by changing the associated SRRM variables and therefore physical factors.

There are many factors of which affects SRRM reading comprehension, and therefore many approaches to maximise RC. One example is the exploration of text complexity which is a language structure factor that is a function of language comprehension within the SRRM. Examples within this direction includes the construction automatic text complexity models such; via classification models that classify at a sentence level within a corpus; or even graph-based models that breaks down text complexity into different nodes, such as background knowledge, and using network representations. [53-56] The aim of these approaches allows further techniques to be developed to minimise text complexity which increases reading comprehension. However, language structure is only one factor out of the many SRRM factors, and therefore higher-level approaches are required to fully maximise reading comprehension.

Rather than exploring SRRM factors individually such only concentrating on language structure. A wider scope of NLP solutions consists of text summarisation and text simplification. These approaches consider the entire text as a whole and therefore more of the SRRM factors are considered in their optimisation problems.

2.2.2 Automatic Extractive Text Summarisation

The aim of automatic text summarisation (ATS) is to minimise the length of an original text whilst keeping the important information, hence producing a summary.[57] The consequences of shorter reads can increase language comprehension by reducing background knowledge needed, the variety of vocabulary needed, more common syntax and less literacy knowledge needed. Furthermore, from the word recognition SRRM variable, less words can also decrease the phonological awareness needed, less likely to have unfamiliar words (hence increasing sight recognition) and make decoding less bloated.

ATS can be split into two major approaches. The first is extractive summarisation whereby the produced summary is a subset of the original text, no new sentences are produced. The general method to achieve this is to transform the texts into a form that could be analysed. This form is then analysed and scored using certain criteria's such as a form of importance or relevancy, and finally depending on tolerance the best scored sentences will be used.

Examples of extractive ATS includes topic focused metrics that is used to process news [58]. As part of information retrieval approaches for search engines that have big data. Using probabilistic methods such as SumBasic system or frequency approaches such as Term Frequency Inverse Document Frequency (TFIDF) to determine importance of sentences. [59,60]

Extractive ATS has been consistently used as in the world but only in a non-fiction capacity. Even then, it suffers problems such as; understanding of original text is omitted which is very important in fiction texts; sometimes lack of coherency and loss of information. [61,62]

2.2.3 Automatic Abstractive Text Summarisation

Abstractive ATS, unlike extractive ATS considers the meaning of the original text. The output of abstractive ATS is generated from the original text, new and with deeper analysis of meaning of which is closer to humans.[63]

Methods used by different abstractive ATS models can significantly vary. From tree-based methods that creates similar important sentences into tree-like structure. This then can be analysed in different ways such as by theme of content, but however it has been showed to miss important phrases. [64-66]. Template based methods that use parts of the original text to create template-based summaries with the advantage of more coherency but can only be pre-defined, such as the GISTEXTER system for multi-template document summaries. [67,68]. Graph-based methods of which is a data structure that allows sentences to be fused together for the abstractive output. Due to graph-based methods popularity, there are many further implementations within this approach such as inclusion of paraphrasing, prioritising information rather than linguistic quality and elimination of redundant information. [69-71]

For this project, the focus is within the deep learning abstractive ATS approach in the form of the Transformer architecture.[72] Specifically large-scale pre-trained models such as BERT and PEGASUS. The transformers architecture is currently the leading architecture for NLP tasks and has shown significant leaps compared to other neural models such as convolution and recurrent neural networks. Examples includes in text classification, machine translation and more. [73,74]

Transformers, BART and PEGASUS will be discussed in more detail in the technical background of the report.

2.2.4 Automatic Text Simplification

Unlike text summarisation but closely related, text simplification seeks to simplify both the semantics and linguistic features of a text, whilst retaining its original meaning and content.[75] Simplification, being a more complex task that involves incorporating meaning behind a text, has progressed in many learning-based disabilities projects such as relating to autism and even physical deafness. [76,77]

There are numerous different models towards text simplification such as the HTSS that combines abstractive summarisation with simplification.[78] HTSS is based on the 'Pointer-generator network' that combines both abstractive and extractive models, to address the common weaknesses of transformers of repeating outputs.[79] Whilst simplification methods have produced better machine learning results, each individual model is fundamentally different and thus general performance are not guaranteed. One of the main draw backs of simplification methods are the specificity of the models which lack scalability and generality of use. This is further highlighted when compared to abstractive ATS that are built upon the transformer's architecture. [80-82]

For this project scope, MUSS will be used and will be discussed in more detail in the technical background of the report.

CHAPTER 3: Technical Background

The aim of this section is to describe the specific NLP processes used in this project, allowing a deeper insight into why they are chosen in the methodology chapter.

3.1 Transformers

Transformers from the landmark paper 'Attention is all you need',[72] is an architecture that incorporates the concept of 'self-attention', which differs from other deep learning methods such as recurrent neural networks.

A transformer is of a typical stacked encoder-decoder architecture but with several introductions such as 'multi-head attention 'layers. Using figure 3 below, this section will give an overall simplified explanation of the key points of how a transformer work. This is to facilitate the understanding of why they are chosen in the methodology, and to ease the explanation of BART and Pegasus transformer models.

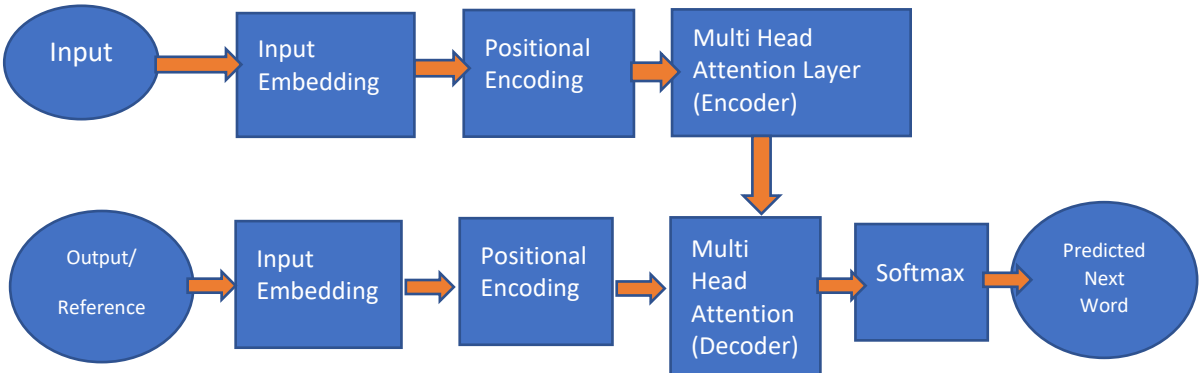


Figure 3 Simplified overview of the transformer's architecture. For the full diagram which includes layers such as forward feed, the original paper is recommended. [72]

As an example, for translating English to French. First, the English input undergoes imbedding to convert text into sequences to allow computational processing. As the meaning of a word in a sentence can be altered by its position, positional encoding is then implemented so that the resulting vectors for English words now contains contextual information.

Secondly, within the multi-head attention layer. Correlations of each word with other words in the same sentence are computed and this results in attention vectors for each word. Step one and two also occurs for the reference input which in this example is French.

Thirdly, by using the attention vectors of both input and reference, mapping of English to French can occur. This is because similar words will have attention vectors that are close together in the embedded space, the output of the multi-head attention layer is another attention vector of both reference and input. One of the methods used in the paper to measure similarity during mapping of input-output is demonstrated by figure 4 and equation 2 below, where similarity is computed as an angle between attention vectors.

$$(\boldsymbol{Eqn2})\ Similairty(a,b) = \cos(\theta) = \frac{a.b}{|a||b|}$$

Where:

a,b are attention vectors of different words with contextual and syntactical information.

θ, the similarity between a,b computed as an angle.

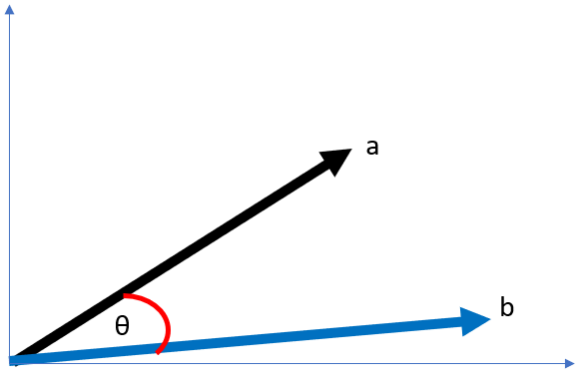


Figure 4 A method of how similarity is quantified. Note that a, b are two different 3D attention vectors in the same 3D spaces.

Finally, the attention vectors are converted into a probability distribution via the softmax layer. The final output of the transformer will be the predicted next French word when writing in English.

3.1.1 BART

BART (Bi-directional and Auto regressive) by Facebook is a NLP task competitive transformer that contains the key characteristics of denoising autoencoder, bidirectional encoder and autoregressive decoder.

Denoising autoencoders is a technique like masked language modelling, this is because neural networks can suffer from identity function and therefore generate inputs as outputs. [83.] The denoising autoencoder sets some input node to 0 and therefore introduce noise into the data to prevent overfitting. In the case of BART within text summarisation, some tokens in the input texts are masked. [84,85]

BART is advantageous in that using the bidirectional encoder, processing of the text can be done in both directions. Therefore because of a word's meaning is a function of their position, during the embedding of words into vectors, there are additional vectors containing sentence level information.

Together with the Auto-regressive encoder which seeks to predict the next token based on the previous token, BART is very good in text generation and therefore abstractive text summarisation. [86,87]

3.1.2 PEGASUS

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarisation) by Google, is also a state-of-the-art transformer based Abstractive ATS model. PEGASUS still uses the encoder-decoder models and leveraging massive training copra with masked tokens approach.

What separates PEGASUS and BART is the focus on pre-training of the model with new self-supervised objectives. Specifically, rather than individual words, the model masks at a sentence level from the input, and then the resulting sentence gaps are generated as part of abstractive summarisation. It was hypothesised that the self-supervised objective gap sentence generation (GSG), improved contextual understanding of input text as a whole and contributed to its performance in this NLP task.[88]

3.2 MUSS

MUSS was developed in 2021 and has the full name of Multilingual Unsupervised Sentence Simplification by Mining Paraphrases.[89] Described as unsupervised as it can be trained without

labelled simplification data, it is a slight departure from the previously described transformer-based models.

The main mechanism behind MUSS is that simplification is done via sentence-level paraphrasing, whereby sentences can have the same meaning but phrased differently. Paraphrases are firstly mined from a corpus after pre-processing, and then they are embedded as sequences using a popular mass-scale embedding technique called LASER.[90] The resulting sequences are then indexed in such a way that each sequence can be used as a query against the entire index. Finally, by using unsupervised method of nearest neighbours clustering, querying a phrase will return similar paraphrases which are closest neighbours in the index.

It is noted that the actual simplification is done by a traditional BART model with the output of the paraphrasing mining training. [89]

3.3 Metrics for Evaluation of Automatic Text Summarisation or Simplification

The landscape of evaluation of automatic text summarisation or simplification is complicated and wide. Therefore, only three machine learning metrics are considered in this project: BLEA, SARI and METEOR. This section will briefly describe how they work whilst their efficacy as metrics will be evaluated in the methodology section instead.

3.3.1 BLEU

BLEA (Bilingual Evaluation Understudy) was developed in 2002 and is a metric that originated in multilingual machine translation. Like many other ATS metrics, BLEA compares the outputs of NLP processes and compares them with human references. BLEA takes a precision approach whereby it measures how many n-grams in the human reference sentences are produced by the candidate sentences. [91,92]

3.3.2 SARI

SARI (system output against references and against the input sentence), was developed in 2016 and is a text simplification specific metric. SARI considers multiple simplification references as many sentences can be written different, in the example of paraphrasing. What further separates SARI from BLEA is that it is lexically sensitive, specifically SARI measures how well words are added, deleted and retained by a simplification model.[93]

3.3.3 METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) was developed in 2005 and is also a machine translation metric. METEOR was specifically created to address the weaknesses of BLEU, in that METEOR uses the harmonic mean of precision and recall. Other more advanced features not found in BLEU also includes the consideration of stemming and synonyms matching with the goal to produce good correlations with human references. [94]

CHAPTER 4: Methodology

This section will give an overview of the methodology process of this project as well as reasoning of the key decisions made during each of the key stages.

4.1 Overview

Figure 5 below shows an overview of the methodology implemented (left to right), different stages are colour coded.

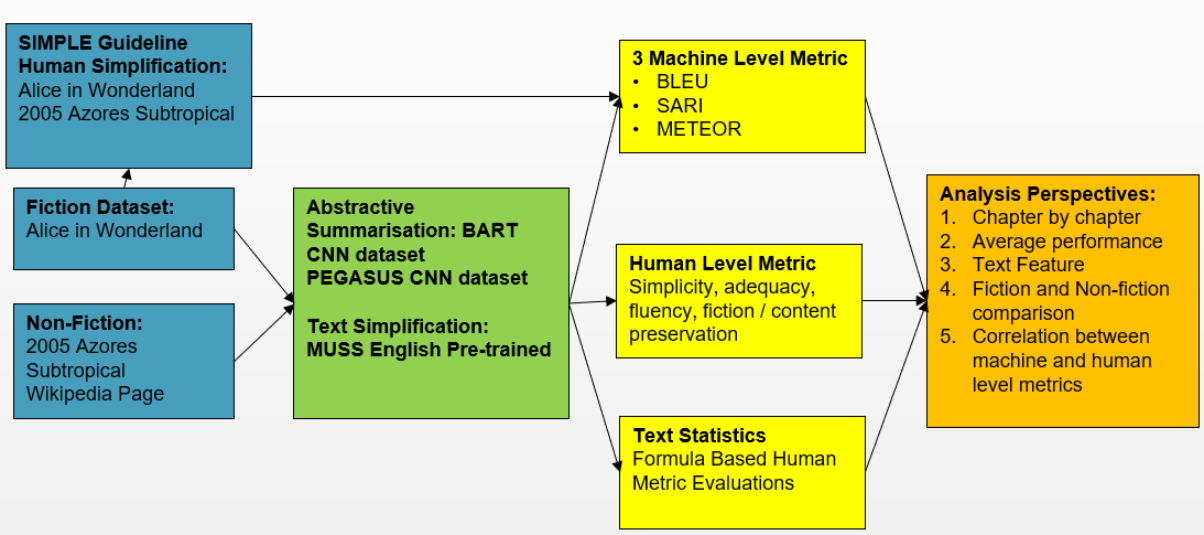


Figure 5. Displays the methodology processes with blue blocks as data processing stage, green block as the NLP implementation stage, yellow block metrics evaluation stage and finally orange block as the anlaysis stage.

Methodology Chronological Description:

1. **Data Processing Stage (blue blocks).**
    a. Two datasets 'Alice in Wonderland' (for non-fiction) and '2005 Azores Subtropical Wikipedia Page' (for fiction) were cleaned as txt files.
    b. Alice in Wonderland had 12 chapters in total but only 5/12 chapters were used due to time constraint. 2005 Azores Subtropical was a single Wikipedia page.
    c. Using the SIMPLE guideline, human simplification was performed on both datasets by hand to result in two new text files that represented human simplification for fiction and nonfiction.
2. **NLP Implementation Stage (green block).**
    a. BART trained by CCN, and PEGASUS trained by CCN were implemented in Jupyter Notebook with the *transformers* library that allowed pre-trained models to be loaded and used immediately. Performed on python 3.10.6.
    b. MUSS was taken from the official GitHub page and was implemented in LINUX in a virtual environment, using Python 3.8 due to compatibility issues of the code.
3. **Evaluation Stage (**yellow blocks).
    a. Three perspectives of evaluation were used.
    b. Machine level metric were carried out using the *transformers* library via Jupyter Notebook.
    c. Human level metric were simply questionnaires where 8 volunteers read the different text versions and rated categories from 0-4.
    d. Text statistics were used by using an online readability checker.

4. **Analysis Stage (**orange block).
    a. All scores were scaled so that they can be compared.
    b. Analysis was carried out from chapter-by-chapter comparisons as well as overall average across chapters.

Since this is a research and analysis project, many systems were taken from other papers but were implemented in novel ways. Table 2 is shown below to summarise what was implemented from other papers and what were project novelty implementations and contributions.

External Tools Used	Project Novelty and Contributions
• **Pre-trained NLP processes such as MUSS, BART and PEGASUS.** • **Machine level metrics of BLEU, SARI and METEOR.** • **Human level metric of simplicity, adequacy and fluency.** • **Scarborough's Reading Rope Model.**	• Usage of MUSS, BART and PEGASUS on fiction texts. • Human level evaluation on BART and PEGASUS. • Addition of fiction preservation category in human level metrics to accommodate for fiction features. • Usage of Scarborough's Reading Rope Model to explain the effects of NLP processes on reading. • Chapter by chapter analysis approach to observe NLP performance across chapters. • Evaluation framework to analysis NLP process via three different perspectives, text statistics, human and machine level metrics. None of these have been used in conjunction before. • Finding the correlation between project specific human metrics and machine level metrics in a fiction setting.

Table 2. A summary of external tools used and project contributions.

4.2 Dataset

In terms of datasets, one of the hypotheses is that fiction datasets are more difficult to read than non-fiction datasets due containing more beyond-sentence level semantics. Some clearer examples can be the usage of literature techniques such as metaphors that are more prevalent in fiction than in non-fiction. Evidence can be found within the machine translation NLP tasks where fictional and non-fictional texts produce statistically significantly different machine translation scores by the same processes, as well as different scores from non-fiction texts that contains metaphors compared to non-fiction texts that do not. [95,96]

For the fiction dataset, 'Alice in Wonderland' was used from the public domain. [97] It was specifically chosen as it is a well-known story and therefore volunteers will find it much easier to spot discrepancies in meanings and content in the human metric questionnaires, its accessibility from being a public domain content was also of benefit.

A non-fiction dataset was used as a control group, specifically the '2005 Azores subtropical storm' Wikipedia page was used due to its public domain nature. [98] One of the major reasons to use non-fiction text as a control group is because many NLP processes are trained and developed in the non-fiction setting, this is mostly due to the absence of beyond-sentence level semantics. Therefore, the simpler non-fiction dataset was used to see if NLP performance variations are due to the text type or some other unknown variables. [96]

Both datasets were then used to produce human simplifications under the SIMPLE guidelines. There are two main reasons for this. Firstly, literature has shown that human simplification remains the best simplification approach and therefore to assess the commercial potential of automated fictional text simplification, human simplification then must be the benchmark to compare to other NLP processes. Secondly, all machine level metrics requires a human simplification reference as they compute the differences between the NLP processes outputs compared to human references. Therefore, it is implied that human simplification remains to be the target standard.

For the human simplification guideline, SIMPLE was used as it was a government guideline used to simplify instructions and government related documents for individuals with learning disabilities. As a result, literature has shown that it indeed improves human understanding of those kind of texts. Currently there are no public domain simplification guidance, and most are done by experts with literature experiences, the simplified publishing content then varies from company to company. [99,100] Some of the guidance can be summarised in table 2 below, and examples of implemented changes are shown in table 3 below.

Even though a guideline was used, there are still degrees of bias due to the manual implementation. Therefore, in order to preserve unquantifiable fiction text features such as writer style and character attitude, most of the speeches were preserved in terms of wording but sentence structure were altered to allow better fluency.

SIMPLE Guidance
1. Simplest words in simplest way. 2. Avoid abstract concepts 3. Short words of every day spoken English. 4. Use many personal words. 5. Use short sentences mostly. 6. One main idea per sentence. 7. Use words consistently. 8. Keep punctuation simple. 9. Avoid jargon, abbreviations and initials.

Table 3. A summary of the important guidelines from SIMPLE.

Example Original Text	Example SIMPLE Guideline Changes	Guidance Followed
Though she very seldom followed it.	Though she very rarely followed it.	• Using more common words.
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into	Alice began to get very tired sitting by her sister on a bank as she had nothing to do. She peeped into the book her sister was reading several times. The book had no pictures or speeches. Alice thought "and	• Keeping punctuation simple. • Using short sentences. • Clearer sentence structure that doesn't intertwine speech with narration.

the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice *"Without pictures or conversations?"*	what is the use of a book without pictures or conversations?".	

Table 4. A sample of the implemented changes of the fiction text. Note the full version can be viewed in Gitlab.

Unfortunately, the mandatory manual implementation of human simplification was a conflict with the project timeline and therefore compromises had to be done. As a result, only 5 out of the 12 chapters of 'Alice in Wonderland' was used and for the non-fiction dataset, only the mentioned Wikipedia article was chosen, and it was significantly less difficult and smaller to read than the fiction dataset. Furthermore, 'Alice through the looking glass' the sequel written by the same author was also disregarded.

4.3 Algorithms

In terms of algorithms that best represented the current abstractive summarisation's frontier, transformer-based models were chosen over other deep learning such as recurrent neural networks (RNN) and long-short term memory models (LSTM) due to the advantage of sequential processing. [72] This is especially important in fiction texts due to the text being longer and that context is affected as a result, transformer-based models do not need to process one word at a time like LSTM and RNN which requires sequential steps, where processing of one word needs the hidden status of a previous word. As a result, transformers can perform parallel processing of all the words in a sentence at the same time via the attention mechanism. [101-103]

As with all deep learning models, they are data dependent and the CNN Daily Mail pretrained version of both BART and PEGASUS was chosen. The CNN Daily Mail dataset is specifically for text summarisation tasks and contains news stories from both CNN and Daily Mail websites. Whilst there are many other datasets such as WikiSummary, X-SUM and WikiHow, CNN Daily Mail was used due to its popularity in developing core transformers such as BART. [104]

CNN Daily Mail Data Set Information	
FEATURE	**VALUE**
Training Pairs	286,817
Validation Pairs	13,368
Test Pairs	11,487

Table 5. Some general information of the dataset used.

One of the major downsides of the transformer models used is that whilst it can process words in parallel, there is a fix sentence length it can compute within a text and therefore during post processing this was by passed by grouping the entire text as one entire string which may have affected the final results. One solution for this was to use Transformer-XL which has a higher context length and claimed to have up to 450% longer context than vanilla transformers, this was not used in the project to do its absence in the *transformer* library and therefore there is no CNN Daily Mail pre-trained version.[105]

MUSS was chosen as it is one of the few un-supervised methods towards text simplification and within the paper, it has demonstrated significant results to other established transformer models as shown in table 6 below.

	English					
	ASSET		TurkCorpus		Newsela	
	SARI ↑	FKGL ↓	SARI ↑	FKGL ↓	SARI ↑	FKGL ↓
Baselines and Gold Reference						
Identity Baseline	20.73	10.02	26.29	10.02	12.24	8.82
Truncate Baseline	29.85	7.91	33.10	7.91	25.49	6.68
Gold Reference	44.87±0.36	6.49±0.15	40.04±0.30	8.77±0.08	—	—
Unsupervised Systems						
BTRLTS (Zhao et al., 2020)	33.95	7.59	33.09	8.39	37.22	3.80
UNTS (Surya et al., 2019)	35.19	7.60	36.29	7.60	—	—
RM+EX+LS+RO (Kumar et al., 2020)	36.67	7.33	37.27	7.33	**38.33**	2.98
MUSS	**42.65**±0.23	8.23±0.62	**40.85**±0.15	8.79±0.30	**38.09**±0.59	5.12±0.47
Supervised Systems						
EditNTS (Dong et al., 2019)	34.95	8.38	37.66	8.38	39.30	3.90
Dress-LS (Zhang and Lapata, 2017)	36.59	7.66	36.97	7.66	38.00	4.90
DMASS-DCSS (Zhao et al., 2018)	38.67	7.73	39.92	7.73	—	—
ACCESS(Martin et al., 2020)	40.13	7.29	41.38	7.29	—	—
MUSS	**43.63**±0.71	6.25±0.42	**42.62**±0.27	6.98±0.95	**42.59**±1.00	2.74±0.98
MUSS (+ mined data)	**44.15**±0.56	6.05±0.51	**42.53**±0.36	7.60±1.06	**41.17**±0.95	2.70±1.00

Figure 6. Results of MUSS performance taken from the original paper.[89]

The mentioned pointer-generator network was not used due to its unmaintained code and therefore there were discrepancies with the original paper performance and code performance.[106]

4.4 Machine Level Metrics

The general landscape is that no machine level metric is the best, indeed many of the currently used metrics such as BLUE and SARI have been scrutinised. One example is that all ATS metrics are heavily dependent on the manual references selected and the type of simplification process that occurs, therefore the scores are not very generalised. [107] However, for the scope of this project, only BLEU, SARI and METEOR were used. A list of their positives and negatives can be summarised in the table below.

Metric Type	Advantages	Disadvantages
BLEU	• Shown to correlate well with human assessments of grammaticality and meaning preservation. [109,110]	• Preliminary evidence had poor judgements of overall simplicity when using multiple operation manual efferences [108]
SARI	• Specifically designed for text simplification evaluation • Uses multiple human references	• Preliminary evidence had poor judgements of overall simplicity when using multiple operation manual efferences [14]
METEOR	• High Pearson's R correlation with human judgements of translation quality in machine translation.[94]	• Not widely used outside of machine translation.[94]

Table 6. Summary of the advantages and disadvantages of the three metrics used.

It is noted that there are some other metrics available such as SAMSA and variations of BLEU, however these metrics are specialised and expansions, and therefore the 'vanilla' versions of each metrics were used instead because of their accessibility and ease of implementation in coding. [107]

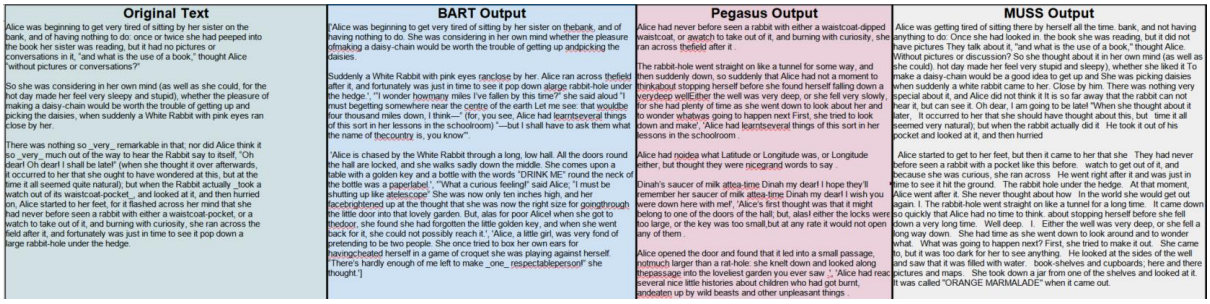
4.5 Human Level Metrics

Since from literature there is a large gap between human and machine level understanding of texts, it is therefore necessary to incorporate human level judgement metrics to evaluate the NLP's potential of commercial fictional text simplification.

8 volunteers were used in the end. Each of them was asked to read the different texts that are grouped by chapters, then they are to rate the texts from 0-4 in the different categories in relative to the original text. The definitions for the categories and a sample of the questionnaire are shown below.

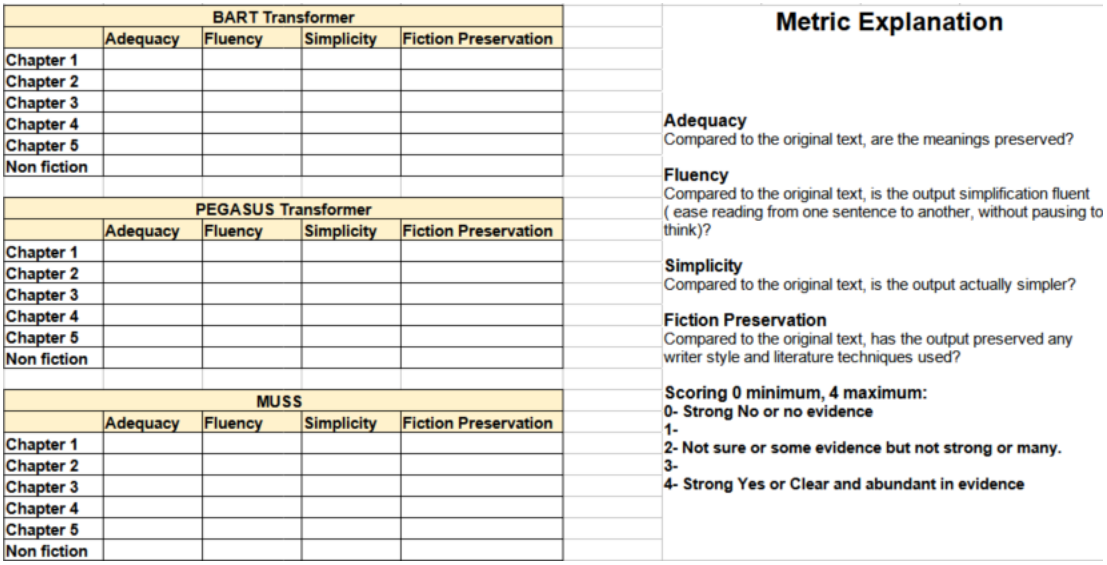
Category	Definition
Simplicity	Compared to the original text, is the output simpler?
Adequacy	Compared to the original text, are the meanings preserved?
Fluency	Compared to the original text, is the output simplification fluent? Ease of reading from one sentence to another, without pausing to think?
Fiction / Content Preservation	Compared to the original text, has the output preserved much of the writer style and literature techniques used?

Table 7. Definitions of the human level metrics categories.



Original Text	BART Output	Pegasus Output	MUSS Output
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice "without pictures or conversations?" So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her. There was nothing so _very_ remarkable in that, nor did Alice think it so _very_ much out of the way to hear the Rabbit say to itself, "Oh dear! Oh dear! I shall be late!" (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually _took a watch out of its waistcoat-pocket_, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.	"Alice was beginning to get very tired of sitting by her sister on thebank, and of having nothing to do. She was considering in her own mind whether the pleasure ofmaking a daisy-chain would be worth the trouble of getting up andpicking the daisies. Suddenly a White Rabbit with pink eyes ranclose by her. Alice ran across thefield after it, and fortunately was just in time to see it pop down alarge rabbit-hole under the hedge. ', "I wonder howmany miles I've fallen by this time?" she said aloud "I must begetting somewhere near the centre of the earth Let me see: that wouldbe four thousand miles down, I think—" (for, you see, Alice had learntseveral things of this sort in her lessons in the schoolroom) "—but I shall have to ask them what the name of thecountry is, you know". 'Alice is chased by the White Rabbit through a long, low hall. All the doors round the hall are locked, and she walks sadly down the middle. She comes upon a table with a golden key and a bottle with the words "DRINK ME" round the neck of the bottle was a paperlabel,' "What a curious feeling!" said Alice; "I must be shutting up like atelescope" She was now only ten inches high, and her facebrightened up at the thought that she was now the right size for goingthrough the little door into that lovely garden. But, alas for poor Alice! when she got to thedoor, she found she had forgotten the little golden key, and when she went back for it, she could not possibly reach it," "Alice, a little girl, was very fond of pretending to be two people. She once tried to box her own ears for havingcheated herself in a game of croquet she was playing against herself. "There's hardly enough of me left to make _one_ respectablepersonl" she thought.']	Alice had never before seen a rabbit with either a waistcoat-dipped waistcoat, or awatch to take out of it, and burning with curiosity, she ran across thefield after it . The rabbit-hole went straight on like a tunnel for some way, and then suddenly down, so suddenly that Alice had not a moment to thinkabout stopping herself before she found herself falling down a verydeep well.Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her and to wonder whatwas going to happen next First, she tried to look down and make', 'Alice had learntseveral things of this sort in her lessons in the schoolroom . Alice had noidea what Latitude or Longitude was, or Longitude either, but thought they were nicegrand words to say . Dinah's saucer of milk attea-time Dinah my dear! I hope they'll remember her saucer of milk attea-time Dinah my dear! I wish you were down here with me!', 'Alice's first thought was that it might belong to one of the doors of the hall; but, alasI either the locks were too large, or the key was too small,but at any rate it would not open any of them . Alice opened the door and found that it led into a small passage, notmuch larger than a rat-hole: she knelt down and looked along thepassage into the loveliest garden you ever saw .', 'Alice had read several nice little histories about children who had got burnt, andeaten up by wild beasts and other unpleasant things .	. Alice was getting tired of sitting there by herself all the time. bank, and not having anything to do: Once she had looked in. the book she was reading, but it did not have pictures They talk about it, "and what is the use of a book," thought Alice. Without pictures or discussion? So she thought about it in her own mind (as well as she could). hot day made her feel very stupid and sleepy), whether she liked it To make a daisy-chain would be a good idea to get up and She was picking daisies when suddenly a white rabbit came to her. Close by him. There was nothing very special about it, and Alice did not think it It is so far away that the rabbit can not hear it, but can see it. Oh dear, I am going to be late! "When she thought about it later, It occurred to her that she should have thought about this, but time it all seemed very natural), but when the rabbit actually did it He took it out of his pocket and looked at it, and then hurried . Alice started to get to her feet, but then it came to her that she They had never before seen a rabbit with a pocket like this before. watch to get out of it, and because she was curious, she ran across He went right after it and was just in time to see it hit the ground. The rabbit hole under the hedge. At that moment, Alice went after it. She never thought about how In the world she would get out again. I. The rabbit-hole went straight on like a tunnel for a long time. It came down so quickly that Alice had no time to think. about stopping herself before she fell down a very long time. Well deep. I. Either the well was very deep, or she fell a long way down. She had time as she went down to look around and to wonder what. What was going to happen next? First, she tried to make it out. She came to, but it was too dark for her to see anything. He looked at the sides of the well and saw that it was filled with water. book-shelves and cupboards; here and there pictures and maps. She took down a jar from one of the shelves and looked at it. It was called "ORANGE MARMALADE" when it came out.

Figure 7. A sample of the questionnaire format, note the FULL volunteer questionnaire (raw data) is in Gitlab.



BART Transformer				
	Adequacy	Fluency	Simplicity	Fiction Preservation
Chapter 1				
Chapter 2				
Chapter 3				
Chapter 4				
Chapter 5				
Non fiction				

PEGASUS Transformer				
	Adequacy	Fluency	Simplicity	Fiction Preservation
Chapter 1				
Chapter 2				
Chapter 3				
Chapter 4				
Chapter 5				
Non fiction				

MUSS				
	Adequacy	Fluency	Simplicity	Fiction Preservation
Chapter 1				
Chapter 2				
Chapter 3				
Chapter 4				
Chapter 5				
Non fiction				

Metric Explanation

Adequacy
Compared to the original text, are the meanings preserved?

Fluency
Compared to the original text, is the output simplification fluent (ease reading from one sentence to another, without pausing to think)?

Simplicity
Compared to the original text, is the output actually simpler?

Fiction Preservation
Compared to the original text, has the output preserved any writer style and literature techniques used?

Scoring 0 minimum, 4 maximum:
0- Strong No or no evidence
1-
2- Not sure or some evidence but not strong or many.
3-
4- Strong Yes or Clear and abundant in evidence

Figure 8. A sample of the questionnaire format, specifically the part to be filled in.

Simplicity, adequacy and fluency were used in the original MUSS paper as human judgement metrics in a non-fiction setting. Therefore, to incorporate the unquantifiable features of fiction texts such as writer style, 'fiction preservation' categories were used to make the human judgement more adapted to fiction text more. This is because for fiction text, the quality of the read is important in commercial settings, and one obvious example is writer style.[89]

4.6 Text Statistics

Since the physical changes of each output can vary a lot and due to time constraints. Text statistics was used to help keep track of the physical changes of each output, in order to facilitate the viewing of each NLP output.

The texts were processed in an online automatic text readability checker and the meanings behind each text statistics are summarised below.[111] Note that from literature, many formula-based reading levels can be biased such as favouring longer texts, hence an average of 'readability consensus was used to give an overall score.[112]

Text Statistics Explanation		
Variable Name	**Explanation**	**SRR Model Variable Affected**
Readability Consensus U.S grade level	A general U.S grade reading level required to read the text based on 7 readability formulas.[111]	Language comprehension and word recognition.
Average Reader Age	Average age required to read the text.	Language comprehension and word recognition.
Unique Words Percentage	Percentage of words in the text that are unique.	Word recognition, specifically sight recognition.
Average Words Per Sentence	Average words found in a sentence.	Word recognition, specifically phonological awareness.
Average Character Per word	Average characters found in a sentence.	Word recognition, specifically decoding.
Single Syllable Words in the Text Percentage	Percentage of words that are single syllable that make up the text.	Word recognition, specifically decoding.

Table 8. Explanation of the text statistics used and their relevance to the reading model.

Whilst there are many more kinds of text statistics, an online checker was used to streaming the methodology process. In the future more bespoke analysis could be done via python such as term frequency inverse document frequency. [113]

CHAPTER 5: Results

Analysis of the data collected take in three main directions. Firstly, from the text feature perspective to show how SRRM factors and variables change with NLP outputs, from chapter to chapter. Then both machine and human level metrics which considers both chapter by chapter analysis as well as average across chapter analysis. Comparisons between nonfiction and fiction datasets were also carried out where available. The advantage of the three directional analysis is that each allows a deeper insight of the next.

5.1 Text Feature Analysis

This section of the analysis focuses on the characteristics of both fiction and non-fictional texts across chapters, as well as averages over all chapters.

5.1.1 Readability Consensus Across Chapters

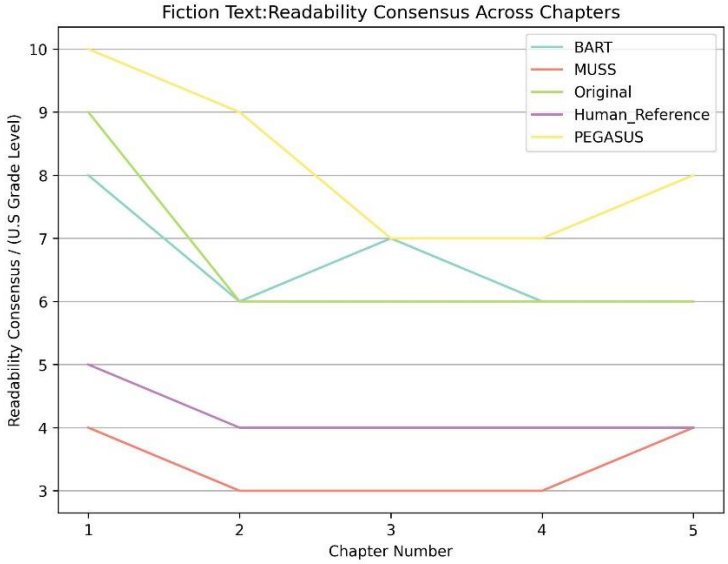


Figure 9. Reading consensus trend across chapters.

From figure 9, it was observed that chapter 1 in the original text is the most difficult to read chapter and that RC remains the same beyond it. The decrease in RC, and therefore decrease in reading difficulty, from chapter 1 to 2 can be observed in all output types.

The human reference output is the most consistent and direct improvement in RC in that it shares the exact same shape as original text, whilst improving RC for each chapter by at least two grades.

BART ATS managed to improve RC from by one grade level in chapter one but however, it increased RC by one grade level in chapter 3. Chapter 2,4 and 5 followed the same RC trend as the original text.

PEGASUS ATS was observed to have the most sporadic RC trend across chapters. At chapter 1, it managed to increase RC difficulty compared to the original text by one grade. Unlike the other output types, PEGASUS gradually improved its own RC from chapter 1 to 4 but however did not improve on the original text. Surprisingly, RC on chapter 5 was an increase.

For MUSS, it was observed to be the best at improving RC overall and by at least 3 grades per chapter. It was also the second most exact simplification of RC difficulty as it shared the same trend shape as original text, from chapters 1 to 4. However surprisingly RC difficulty was an increase at chapter five which is the same behaviour as PEGASUS.

Overall, for RC across chapters, human referencing is most simplified and consistent output with MUSS second, BART third, and lastly PEGASUS which was not an improving in RC or captured the original texts behaviour.

5.1.2 AWPS and ASSW Across Chapters

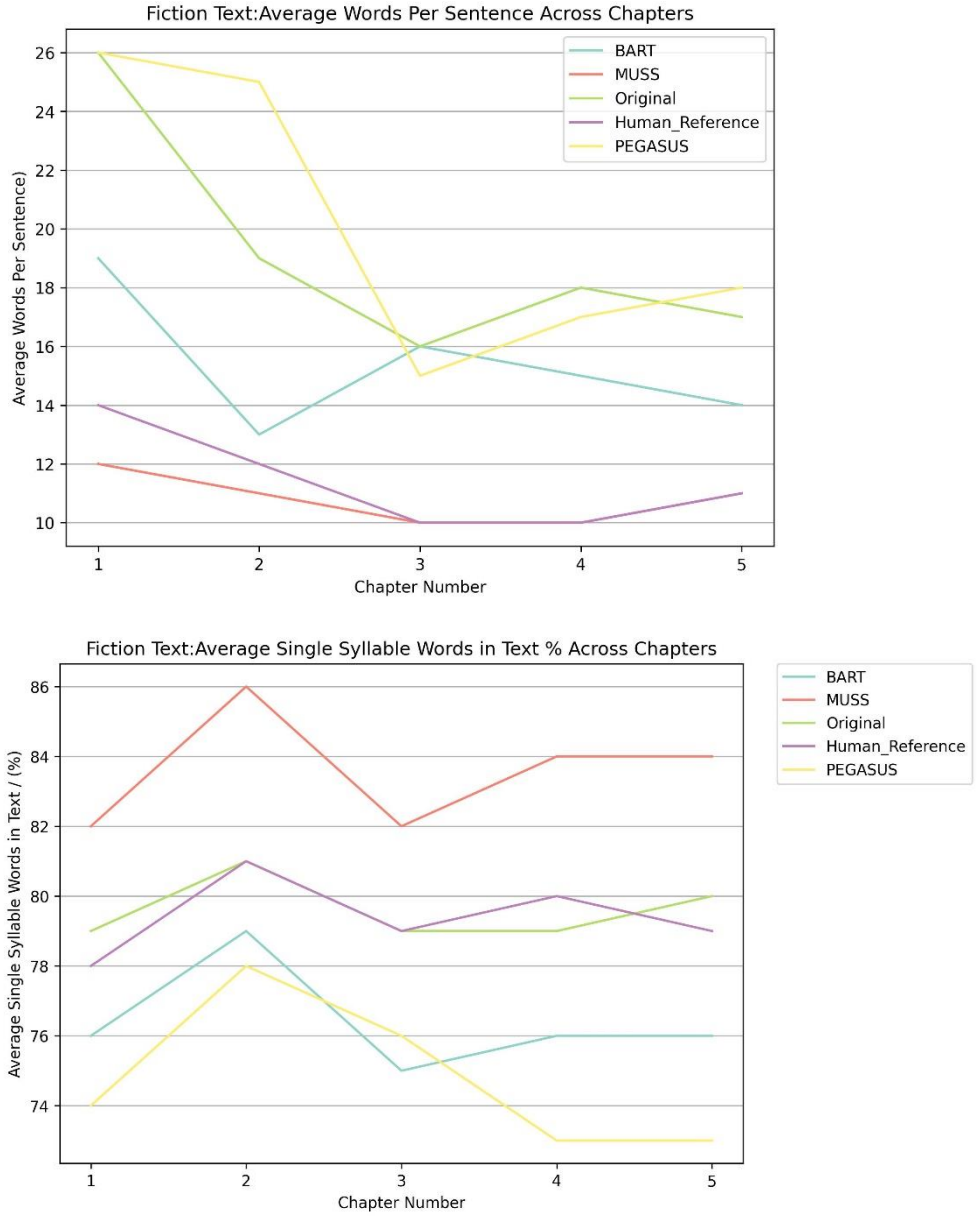


Figure 10. (A) Above shows the average sentence per word across chapters. (B) Below shows the average single syllable words % across chapters.

Since reading consensus is a derived feature which takes in an average. Specific text statistics were analysed to further look at how each chapter differs in terms of the original and other outputs.

There are numerous different words statistics, specifically average words per sentence (AWSP) and average single syllable words (ASSW) in text were chosen to analyse. The reason for this is that both are interpretable physical features of the SRRM, in that AWSP affects the language structure factor of

language comprehension variable, whilst ASSW affects the decoding factor of the word comprehension variable. Multi syllable words and high AWSP decreases reading comprehension.

From Figure 10A The original text has the highest AWPS in chapter 1, before gradually decreasing from chapter 2 to 3. Then there is an increase from chapter 3 to 4, followed by a decrease from chapter 4 to 5.

Both human reference output and MUSS share the exact same shape, except that MUSS reduced AWSP by 2 more words than human reference. Both follow the sharp decrease in AWSP from chapter 1 to chapter 3 which is the same as the original text. However, from chapter 3 onwards, whilst the original text is still decreasing in AWSP, both human reference and MUSS has increased in AWSP.

For BART, it did reduce AWPS in chapter one significantly by at least 4 words and continued to decrease in chapter 2. It managed to increase in AWSP in chapter 3 before resuming reduction from chapter 4 to 5.

For PEGASUS, it did not reduce AWPS in chapter one but did exhibit the sharp AWPS decrease observed in the original text from chapter 1 to 3. AWPS continued to increase from chapter 3 onwards.

Overall, for AWPS, human reference and MUSS were the best at reduction with BART third and PEGASUS last. It was observed that many of the change in reduction diverged at chapter 3, and that PEGASUS output was the most similar to the original text for AWPS across chapters.

From Figure 10B, the original text had a high ASSW of around 85% to 75% across the chapters, with the highest at chapter 2. The oscillation from chapter 1 to chapter 3 can be observed in all outputs, therefore all outputs where able to capture this characteristic.

Human reference remains the most consistent with the original text in terms of ASSW but however did not improve it. MUSS was the only output that increased AWSS whilst the other output except for human reference decreased ASSW and therefor potentially decreased reading comprehension.

5.1.3 Chapter Averaged Fiction Text Features

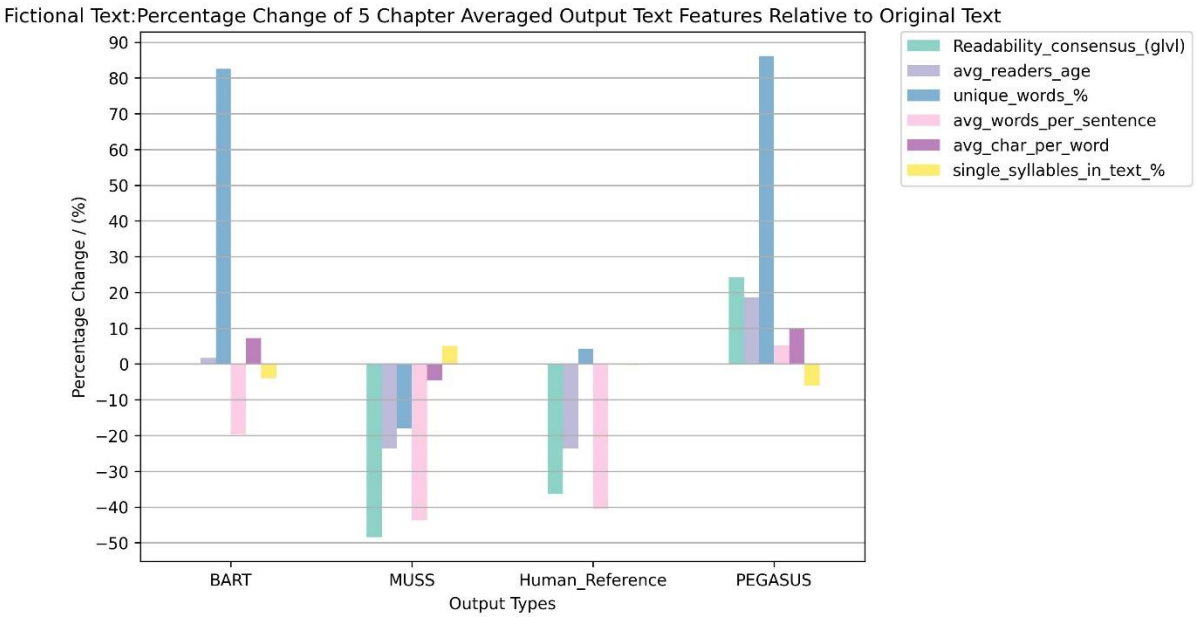


Figure 11. Displays all text features as percentage changes relative to human references in fiction texts.

Overall, in terms of readability consensus for fictional texts, MUSS provides the highest improvement in readability consensus with around 48% decrease in grade level. Followed by Human reference of around 36% decrease. BART had no impact on RC with 0% and PEGASUS increased RC grade level, hence making the text more difficult to read by this metric. Average readers age followed similar trends as RC.

For percentage of unique words in the text, the ATS outputs of BART and PEGASUS improved the repetitiveness of the original text by at least 80% each. This suggests that BART and PEGASUS have increased word comprehension as there are less variety in vocabulary requirement.

For AWPS, MUSS was observed to be the best at reducing sentence length around 44% decrease, although human reference is not far behind with around 40% decrease. BART had a decrease of around 20% whilst PEGASUS increased sentence length by around 5%.

In terms of ASSW, the effects of all outputs are small with the maximum of around 10% increase by PEGASUS. Human reference had no effect on ASSW, MUSS had negative effect of around 5% and BART increased ASSW by around 8%. This suggests that ASSW maybe an insignificant physical factor that does not contribute to the language comprehension variable in the SSR, hence it does not affect reading comprehension much.

5.1.3 Chapter Averaged Non-Fiction Text Features and Comparison Against Fiction

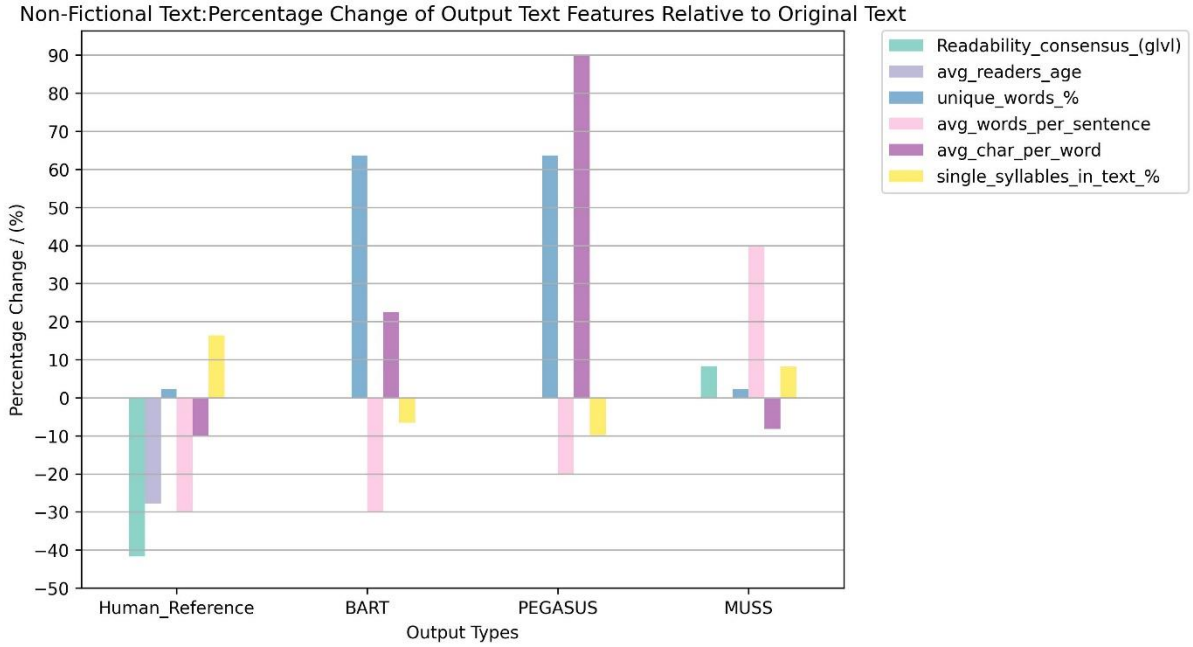


Figure 12. Displays all text features as percentage changes relative to human references in fiction texts.

From figure 12, it was observed that readability consensus improved by around 40% from human reference only. BART and PEGASUS models had no impact on the readability consensus and MUSS increased grade level requirement by around 8%.

Human reference output was also the only output that decreased the average readers age by around 28% whilst all the other outputs had no effect. Furthermore, human reference and MUSS were the only outputs that decreased average character per word of around 10% and 8% respectively.

PEGASUS increased average character per word by 90% and therefore it could suggest PEGASUS increased the language comprehension requirement of the text significantly. PEGASUS, along with BART, was also the most effective at increasing unique words in the text with around 60%.

In comparison with fictional text. Human reference stood out as the more effective output based on text features, with highest percentage difference improvement in readability consensus and average reader age. Furthermore, unlike fictional text, machine learning outputs did not appear to affect many of the text features with 0% change.

5.2 Machine Level Metric Analysis

5.2.1 Fiction Text: ML Scores Across Chapters

This section is an analysis of machine level metrics patterns within each chapter and across chapters.

5.2.1.1 BLEU Scores

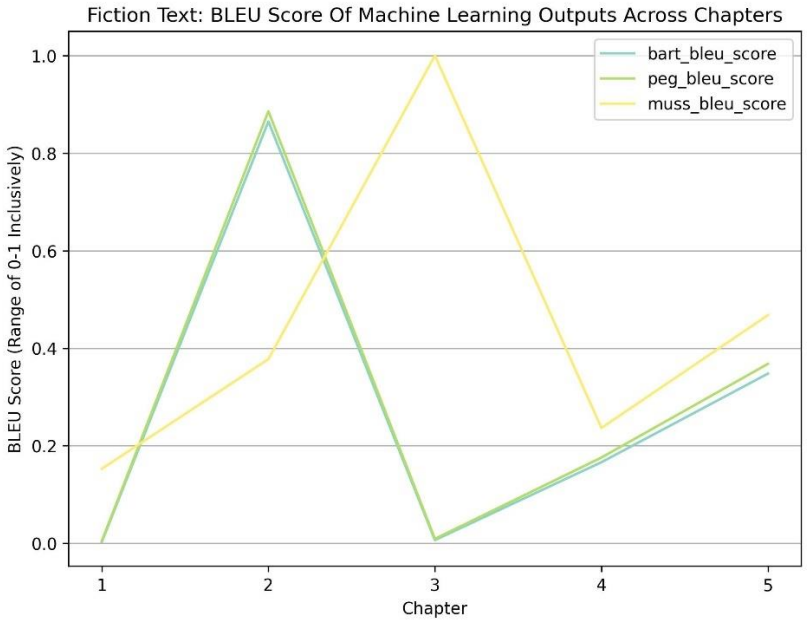


Figure 13.BLEU scores for each machine learning output across chapters. Note that BLEU scores can range from 0-1. Where 1 is a perfect scoring compared to human referencing.

From figure 13, it was observed that BART and PEGASUS had very similar trends across the 5 chapters. Both scored 0 on chapter one and maximised their BLEU scores of around 0.9 at chapter 2, before scoring 0 again for chapter 3 and consequently improving from chapter 3-5 steadily.

Insignificantly, PEGASUS could be observed to generally score a few decimal points higher than BART, and this similarity begins to diverge from chapter 3 onwards.

For MUSS, it generally performed better across all 5 chapters but still exhibits chapter to chapter performance variations similar to BART and PEGASUS. MUSS had scored around 0.18 in chapter 1 maximised its BLEU score of a perfect 1 in chapter 3, before dipping in chapter 4 and consequently improving in chapter 5.

Interestingly, chapter 2 did seem to influence the performance of MUSS with a seemingly increase in score rate.

5.2.1.2 SARI Scores

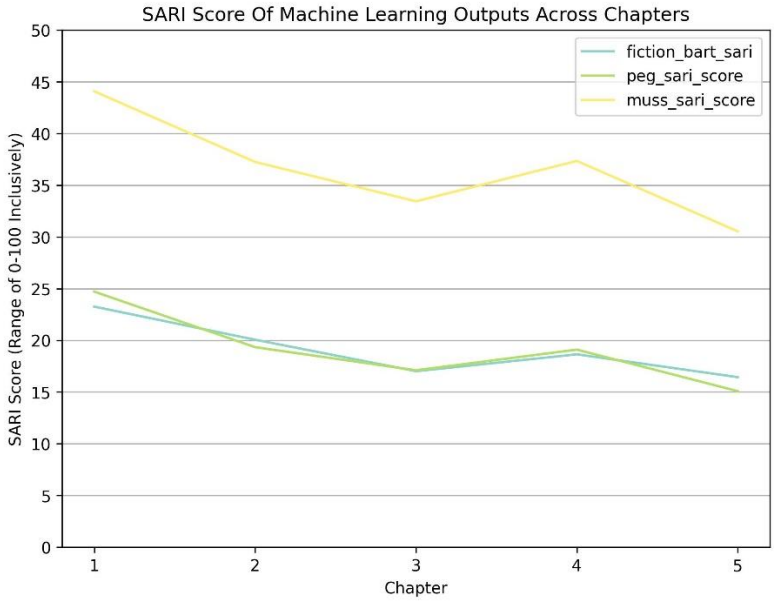


Figure 14.SARI scores for each machine learning output across chapters. Note that SARI scores range from 0 – 100 where 100 is a perfect score relative to human reference.

From figure 14, it was observed that all NLP process had very similar trends across all five chapters, however MUSS generally scored higher than BART and PEGASUS by at least 20 SARI points

BART, PEGASUS MUSS scored their maximum SARI score of around 24, 25 and 44 respectively in chapter 1. Then decline continued from chapter 2 to 3 for all three. Afterwards SARI scores improved from chapter 3 to 4 but however this was more noticeable in MUSS than in BART or PEGASUS.

Chapter 5 caused a decrease in SARI score across all three outputs; however, PEGASUS was more affected than BART despite both having nearly the exact same trends and scores. All three processes had their minimum SAIR scores in chapter 5.

5.2.1.1 METEOR Scores

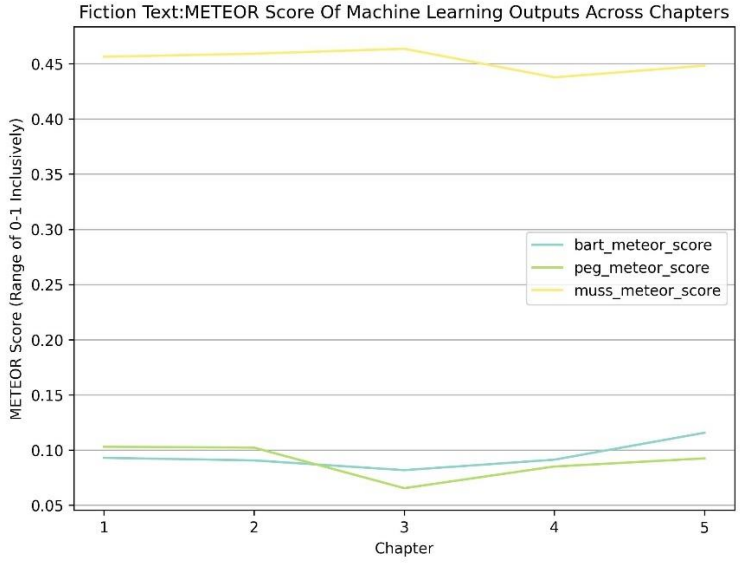


Figure 15. METEOR scores for each machine learning output across chapters. Note that METEOR scores range from 0 – 1 where 100 is a perfect score relative to human reference.

From figure 15, it was observed that METEOR scored the highest per chapter, followed by BART and then PEGASUS.

At chapter 1, both MUSS and PEGASUS had high METEOR score of around 0.46 and 0.11 each whilst BART achieved its maximum METEOR score of around 0.12 in chapter 5.

All three NLP processes exhibits a dib in their scores. For MUSS the dip occurred in chapter 4 with a score of around 0.44, whilst for BART and PEGASUS the dip occurred in chapter 3 of scores of around 0.85 and 0.75 respectively.

Its worth noting that MUSS maximised its METEOR score around chapter 3 of around 0.47 whilst in the same chapter, both BART and PEGASUS dipped.

BART overtook PEGASUS in METEOR scores from chapter 3 onwards and the diverging trend continued.

5.2.2 Fiction Text: ML Scores Averaged Grouped by NLP Processes

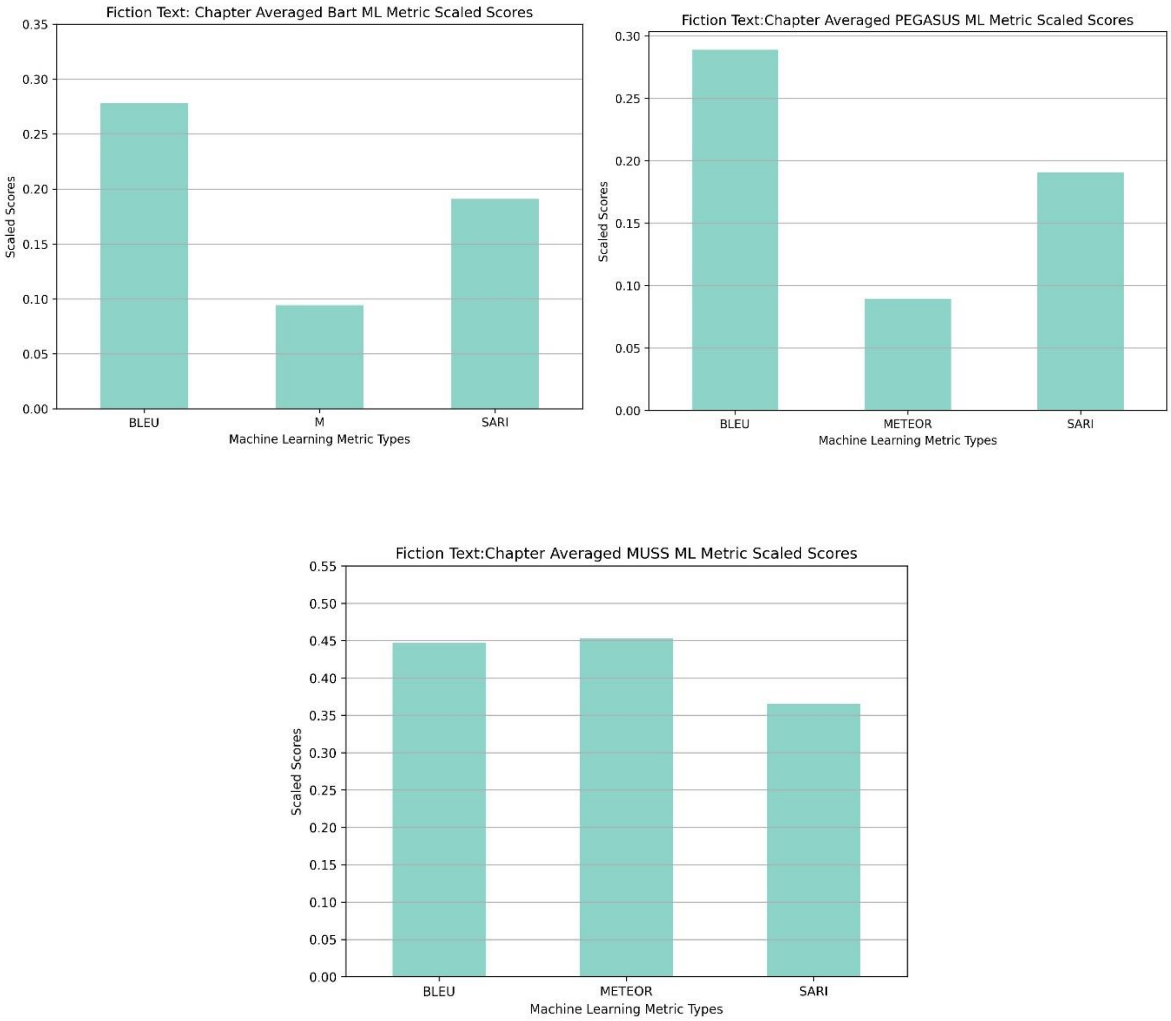


Figure 16. Across chapter average ml scores grouped by the NLP processes used. (A) above is BART, (B) top right is PEGASUS and (C) bottom left is MUSS. Note that all scores have been scaled for better data visualisation.

By considering the across chapter average performance of NLP processes, from Figure 16A. It was observed that BART scored the highest in BLEU with a score of around 0.375, then SARI of around 0.19 and then METEOR of around 0.09.

As well as sharing very similar average scaled scores, from Figure 16A and 16B PEGASUS and BART share the same machine learning metric trends. PEGASUS had the highest score of around 0.375 in BLEU as well, with 0.19 in SARI and around 0.09 in METEOR.

For MUSS from figure 16C, it scored the highest in METEOR with around 0.455, the second highest of around 0.445 within BLEU and the lowest in SARI with a score of around 0.36. Unlike BART and PEGASUS, MUSS did equally well in both BLUE and SARI, and the difference between each ML metric had a lower range of around 0.1 points. Unlike BART and PEGASUS with ranges as high as 0.2.

Overall MUSS performed the best in all three ml metrics by at least 0.1 points, with both BART and PEGASUS joint second.

5.2.3 Non-Fiction Text: ML Scores Averaged Grouped by NLP Processes and Comparison with Fiction Text

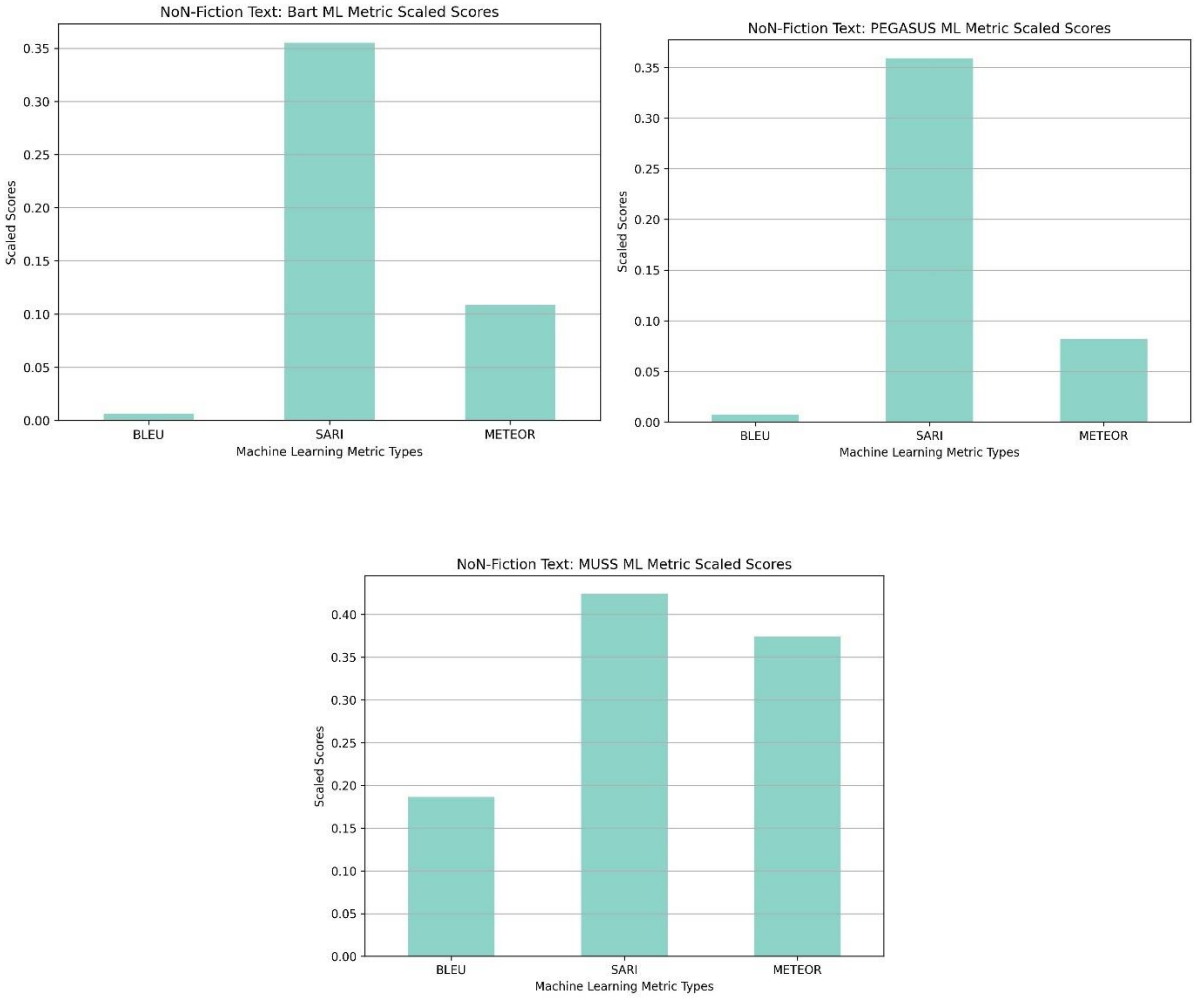


Figure 17. Across chapter average ml scores grouped by the NLP processes used. (A) above is BART, (B) top right is PEGASUS and (C) bottom left is MUSS. Note that all scores have been scaled for better data visualisation.

In terms of the ml metric performance for non-fiction text, from figure 17A. It was observed that BART scored the worst in BLEU with score of around 0.01, then around 0.36 in SARI and finally around 0.11 in METEOR.

PEGASUS from figure 17B, shared very similar trends with BART. It scored about the same in terms of both BLEU and SARI, however it scored lower than BART in METEOR with a score of around 0.07.

For MUSS from figure 17C, it scored the highest in SARI with a score of around 0.425, then around 0.375 in METEOR, followed by a score of around 0.18 in BLEU.

Overall, for non-fiction text, MUSS scored the highest in all three ml metrics. However, whilst both BART and PEGASUS scored minimally in both BLEU and METEOR, it was clearly observed that both had favourability in SARI.

In comparison with fictional, both datasets had MUSS scoring the best in all three metrics. For BART and PEGASUS, surprisingly both only had good BLEU scores in fictional texts rather than non-fiction text. Furthermore, both models showed favourability to SARI in non-fiction whilst had modest scores in non-fiction.

A final point in fiction and non-fiction ml metric comparison, METEOR scores for all three NLP processes remained relatively the same no matter the dataset.

5.3 Human Level Metric Analysis

5.3.1 Fiction Text: HL Metric Across Chapters

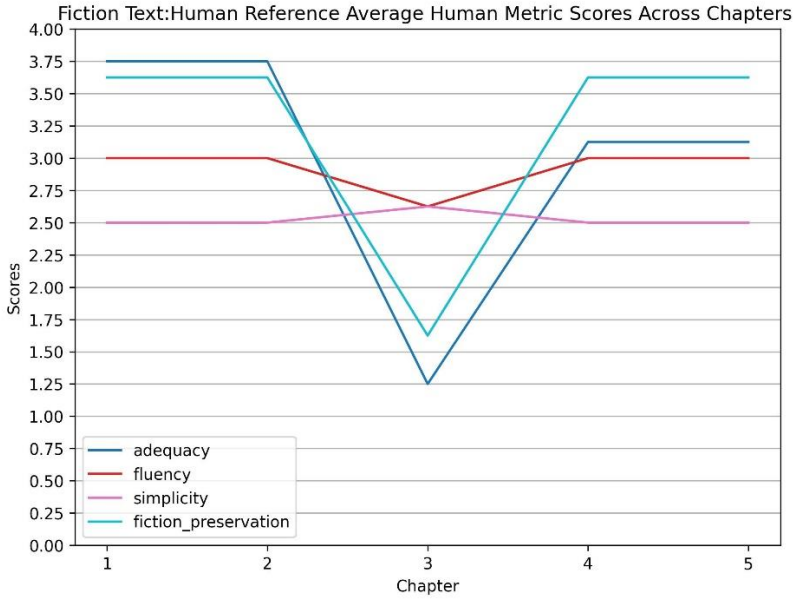


Figure 18.Average human level metric scores chapter by chapter analysis for SIMPLE human reference outputs.

From figure 18 above, it was observed that human reference output had the same HLM scores for both chapter 1 and 2. HLM scores dived into their lowest in chapter 3 except for simplicity which maximised around 2.65. Finally, all HLM scores resumed their constant trend from chapter 4 and 5.

Human reference outputs HLM performance also varies chapter by chapter. It was seen that initially from chapter 1 to 2, human reference output favoured in the order of adequacy, fiction preservation fluency and simplicity. However, this order drastically changed in chapter 3 where adequacy and fiction preservation became the worst HLM scores with fluency and simplicity having the same score of around 2.65. It is towards the end from chapter 4 to 5 that adequacy and fiction preservation were above the fluency and simplicity again, however this time with fiction preservation being the best HLM score.

For human reference overall, simplicity and fluency were consistent across chapters ranging from 3.00 to 2.50. Whilst adequacy and fiction preservation were less consistent with scores from 3.75 to 1.25.

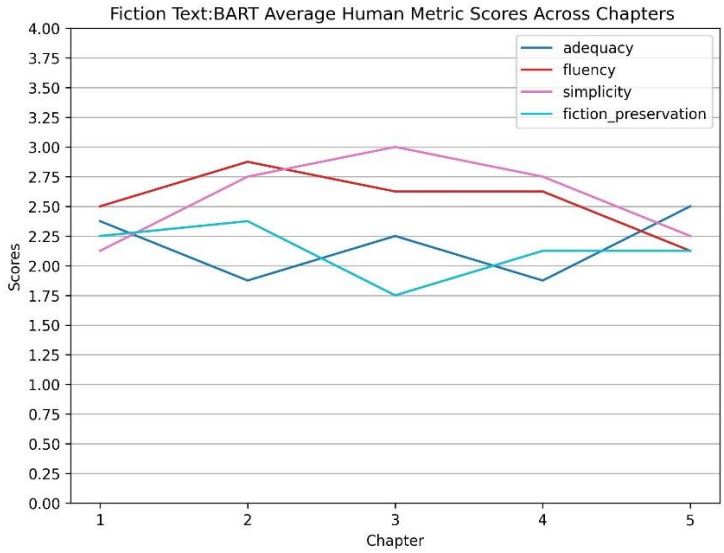


Figure 19. Average human level metric scores chapter by chapter analysis for BART transformer outputs.

HML score trends for BART were very sporadic as observed in figure 19 above. Generally, from chapter 1 to 2, improvements were made with fluency, simplicity and fiction preservation. Adequacy was the only HLM score that dipped from chapter 1 to 2. From chapter 2 to 3, only simplicity out of the 3 original HLM scores continued to improve, whilst adequacy increased instead. Then from chapter 3 to 5, simplicity and fluency continued to decrease whilst fiction preservation stayed consist, and adequacy continued to improve.

Overall, for the HLM performance of BART, simplicity score was maximised at chapter 3 with a score of 3.00 and minimised in chapter 1 with a score of around 2.15. Fluency score was maximised in chapter 2 with a score of 2.85 and minimised in chapter 5 with a score of 2.25. Adequacy was minimised in chapter 2 and 4 with scores around 1.85. Finally, fiction preservation was maximised in chapter 2 with a score of 2.35 and was minimised in chapter 3 with a score of 1.75

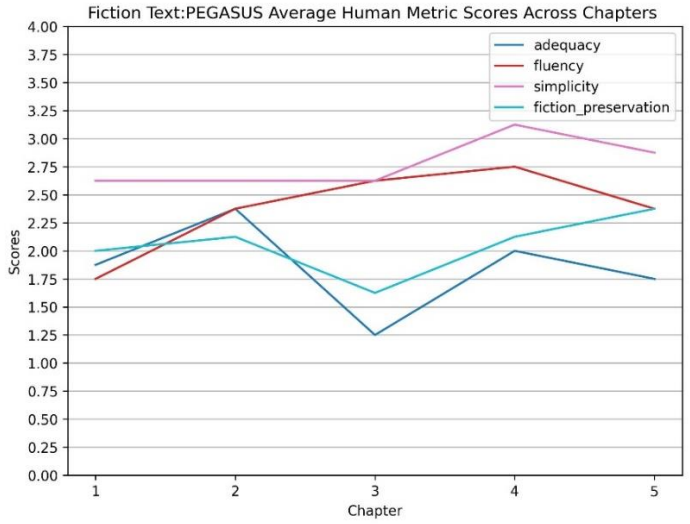


Figure 20. Average human level metric scores chapter by chapter analysis for PEGASUS transformer outputs.

For PEGASUS as shown in figure 20 above, from chapter 1 to chapter 2 there are clear trends of increase in HLM scores of fiction preservation, adequacy and fluency. Simplicity remained constant from chapter 1 to chapter 3.

Within chapter 3, only fluency increased out of the original 3 increasing HLM scores, with adequacy and fiction preservation reaching their minimal scores of 1.25 and 1.65 respectively.

All four HLM scores increased from chapter 3 to chapter 4, and all HLM scores decreased from chapter 4 to 5 except for fiction preservation which had its highest score of 2.35 in chapter 5.

Overall, for the HLM performance of PEGASUS, simplicity score was maximised at chapter 4 with a score of 3.00 and was constant from chapter 1 to 3 with a score of around 2.65. Fluency score was maximised in chapter 4 with a score of 2.75 and minimised in chapter 1 with a score of 1.75.

Adequacy was minimised in chapter 3 with a score of 1.25. Finally, fiction preservation was maximised in chapter 5 with a score of 2.35 and was minimised in chapter 3 with a score of around 1.65.

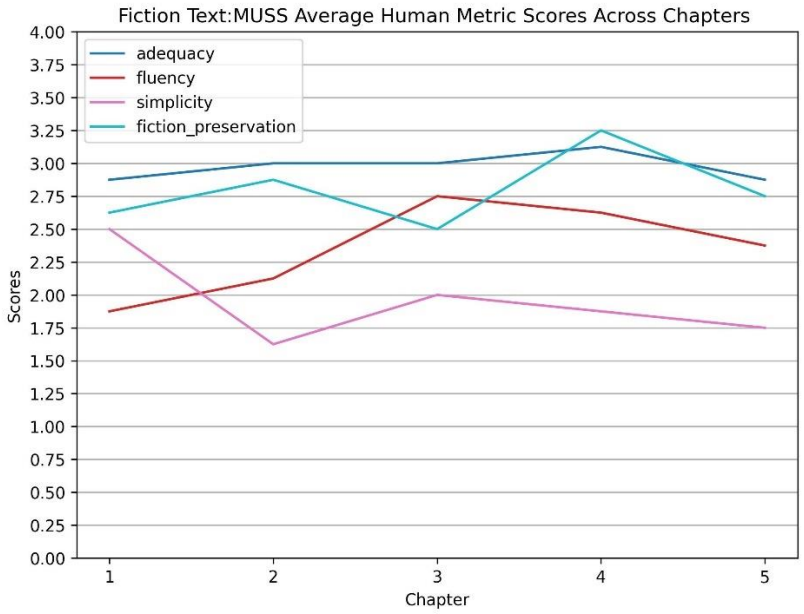


Figure 21. Average human level metric scores chapter by chapter analysis for MUSS outputs.

For MUSS, as shown in figure 21 above. HLM scores of adequacy, fiction preservation and fluency generally increased from chapter 1 to 2, whilst simplicity decrease. Then from chapter 2 to 3, adequacy stagnated, fluency continued to increase whilst fiction preservation decreased, and simplicity began to increase. From chapter 3 – 5 onwards, adequacy, fluency and simplicity generally decreased whilst fiction preservation overall increased.

Overall, for the HLM performance of MUSS, simplicity score was maximised at chapter 1 with a score of 2.50 and minimised in chapter 2 with a score of 1.65. Fluency score was maximised in chapter 3 with a score of 2.75 and minimised in chapter 1 with a score of 1.85.

Adequacy was very consistent throughout the five chapters with score range of about from 3.15 to 2.85. Finally, fiction preservation was maximised in chapter 4 with a score of 3.35 and was minimised in chapter 3 with a score of around 2.50

5.3.2 Fiction Text: HML Scores Averaged Grouped by NLP Processes

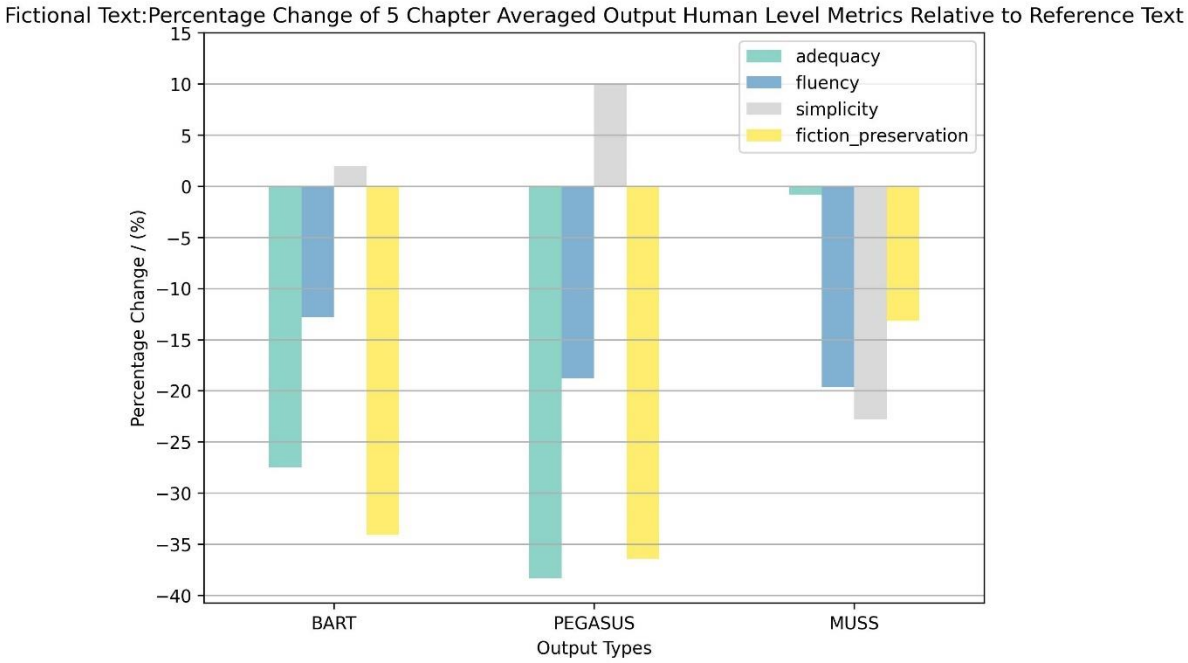


Figure 22. Chapter averaged HML scores grouped by NLP processes, percentage change relative to reference text.

Overall, human reference text scored the best in all HML across most of the chapters, and therefore human reference was chosen to be the benchmark.

Comparing the overall performance of each NLP processes relative to the human reference text. It was observed that generally all three NLP processes did not improve in any HML scores except for BART and PEGASUS on simplicity. BART scored better in simplicity than human reference text by about 2.5% whilst PEGASUS scored better by 10%.

For the transformer-based models, adequacy was poorly scored with a lowered scoring of about 27% for BART and about 38% from PEGASUS. Followed by adequacy, fiction reservation was also poorly scored with BART's scoring around 34% lower than human reference text, and PEGASUS scoring around 36% lower.

Comparing the 4 HLM scores, both BART and PEGASUS still had lower scoring of negative around 17% and 18% each in fluency, however fluency was the least bad scored behind simplicity. It was also observed that generally PEGASUS did worse than BART in fiction preservation, adequacy and fluency by at least 5%.

In terms of the performance of MUSS, although it did not improve in any HML scores relative to the human metric score. MUSS had the least negative scoring percentage change compared to BART and PEGASUS and therefore it could be argued that MUSS was the closest to human reference text. One example of this is that MUSS lowered human reference text's adequacy scoring by a negligible -1%.

For MUSS's fluency scoring, it had greater decreased compared to BART and PEGASUS of around 19.5%. MUSS however also had the lowest decrease of around 23% in simplicity, as well as around 13% in fiction preservation.

Discounting human referencing, it can be argued then MUSS significantly better in adequacy, fluency and fiction preservation. Whilst PEGASUS scored the worst in all HLM's except for simplicity which it scored the best in. BART firmly remained in the middle.

5.3.3 Non-Fiction Text: HML Scores Averaged Grouped by NLP Processes and Comparison with Fiction Text

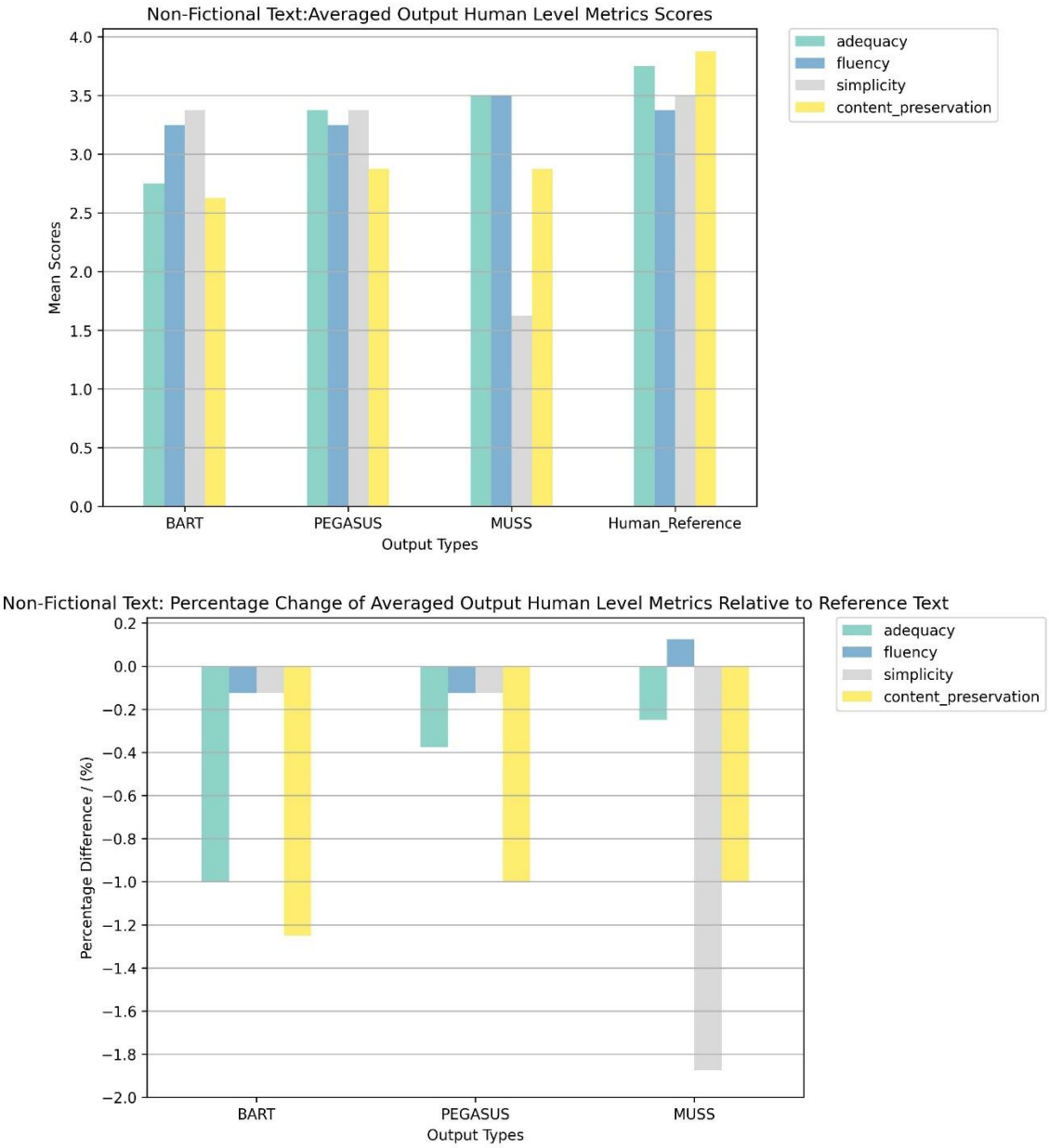


Figure 23. (A) Above, averaged output for HML in non-fiction text. (B) Below, average output percentage change compared to human reference text.

Analysis of HML in non-fiction text from figure 23A has shown that, human reference remains the best output in adequacy, simplicity and content preservation, followed by MUSS, PEGASUS and then BART.

It was observed that the general score for each NLP outputs on average was higher in non-fiction than in fiction dataset. Moreover, it appears that content preservation was the HLM that all NLP process scored in, except for MUSS's simplicity score.

From figure 23B, it was shown that whilst generally all NLP processes did not improve on the human reference text, those changes were minuscular and ranged a maximum of around 1.82% difference. This was different to fiction dataset where the maximum difference ranged of around 37% difference.

More differences when compared to fiction dataset includes; both PEGASUS and BART did not improve simplicity; BART appears to have done worse in adequacy than PEGASUS, this is also the same in terms of content preservation; BART and Pegasus both have the same fluency and simplicity scores percentage difference where as in fiction text dataset, both had different percentage changes; finally, MUSS was observed to improve the human text fluency score by 0.1%.

5.4 Human and Machine Level Metric Correlations

For each machine level and human level metric results, the Shapiro-Wilk tests were used to determine if the data were normally distributed, and it was found that they were. Hence, paired sample t-test were done for all pairs of human level and machine level metric, with results finding that all of the p-values were less than 0.05 and so correlations are within 95% confidence interval.

5.4.1 Fiction Text: By NLP Process

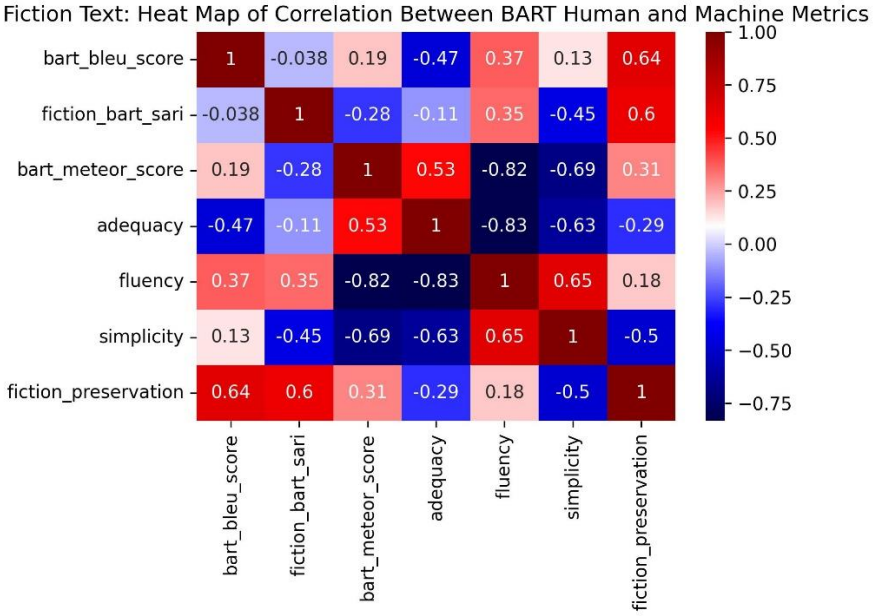


Figure 24. Displays the Pearson's correlation matrix for human and machine level metrics for BART transformer outputs.

By using figure 24 above, it was found that BART transformer's BLEU score had relatively strong positive correlation, of 0.64, with fiction preservation. BART BLEU scores also had moderate negative correlation with adequacy of -0.47, and moderate positive correlation with fluency of 0.37. Finally, BART BLEU had a small correlation with simplicity of around 0.13.

For BART SARI scores, it had strong positive correlation of 0.64 with fiction preservation, a moderate negative correlation with simplicity of -0.45, a moderate positive correlation of 0.35 with fluency. It had a small negative correlation of -0.11 with adequacy.

For BART METEOR, it had moderate positive correlation of 0.31 with fiction preservation, strong positive correlation of 0.53 with adequacy, strong negative correlation with fluency of -0.82 and finally, strong negative correlation with simplicity of -0.69.

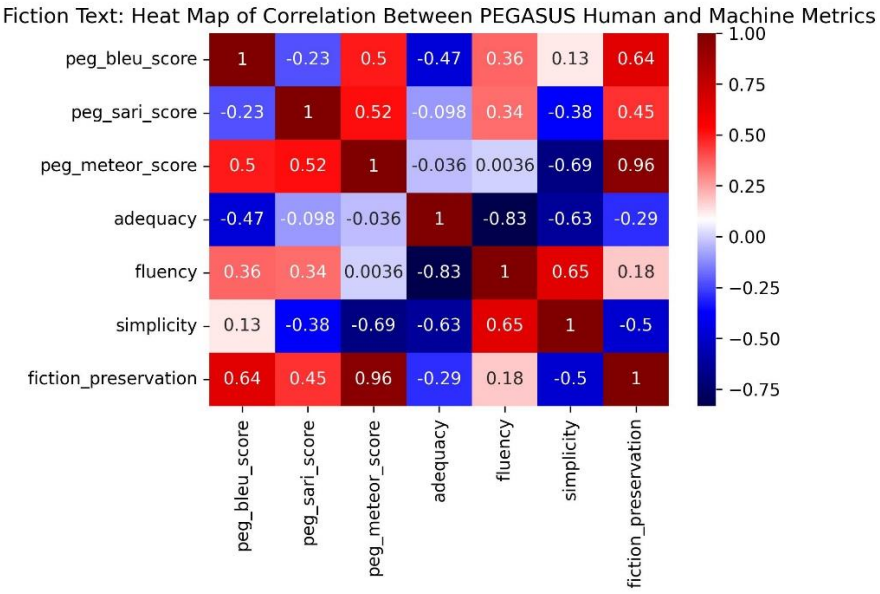


Figure 25. Displays the Pearson's correlation matrix for human and machine level metrics for PEGASUS transformer outputs.

By using figure 25 above, it was found that PEGASUS transformer's BLEU score had strong positive correlation of 0.64 with fiction preservation. Low positive correlation of 0.13 with simplicity, moderately positive correlation of 0.36 with fluency and finally moderately negative correlation of 0.47 with adequacy.

For PEGASUS SARI scores, it had moderately positive correlation with friction preservation of 0.45, moderately negative correlation of -0.38 with simplicity, moderately positive correlation of 0.34 with fluency. Finally, no correlation with adequacy.

For PEGASUS METEOR scores, it had strong positive correlation of 0.96 with fiction preservation. Strong negative correlation of -0.69 with simplicity. Finally, zero correlations with fluency and adequacy.

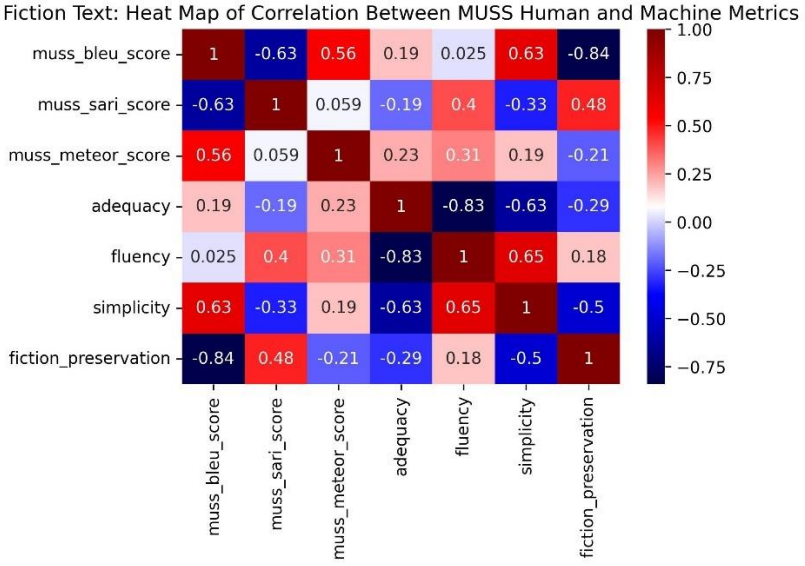


Figure 26.Displays the Pearson's correlation matrix for human and machine level metrics for MUSS transformer outputs.

By using figure 26 above, it was found that MUSS'S BLEU score had strong negative correlation of -0.84 with fiction preservation. strong positive correlation of 0.63 with simplicity, no correlation with fluency and finally weak positive correlation of 0.19 with adequacy.

For MUSS SARI scores, it had moderately positive correlation with friction preservation of 0.48, moderately negative correlation of -0.33 with simplicity, moderately positive correlation of 0.4 with fluency. Finally, weak negative correlation with adequacy of -0.19.

For MUSS METEOR scores, it had weak negative correlation of -0.21 with fiction preservation. Weak positive correlation of 0.19 with simplicity. Weak positive correlation of 0.31 with fluency and finally weak positive correlation with adequacy of 0.23

NLP Process Average Pearson's Correlation Between NLP Processes and HL metrics				
ML/HLM	Fiction Preservation	Simplicity	Fluency	Adequacy
BART	0.53	-0.22	0.00	0
PEGASUS	0.68	-0.31	0.23	0.16
MUSS	-0.19	0.16	0.24	-0.1

Table 9.Displays the average machine level metric correlations between the NLP processes and human level metrics

From table 8, it was observed that for fiction preservation. PEGASUS had the best performance correlation of 0.68, followed by BART and then MUSS with negative performance correlation of -0.19

For simplicity, both the transformer models of BART and PEGASUS had negative performance correlation of -0.22 and -0.31 respectively. MUSS was the only process with positive performance correlation of 0.16.

For fluency, BART had no performance correlation whilst PEGASUS and MUSS had very similar performance correlations of 0.23 and 0.24 respectively.

For adequacy, BART and MUSS had around no performance correlation whilst PEGASUS had weak positive correlation of 0.16

5.4.2 Fiction Text: By ML Metrics

Average Pearson's Correlation Between ML and HL metrics				
ML/HLM	Fiction Preservation	Simplicity	Fluency	Adequacy
BLEU	0.15	0.30	0.24	0.25
SARI	0.52	-0.27	0.36	0.10
METEOR	0.35	-0.40	-0.17	0.25

Table 10.Displays the average correlations between the human level and machine level metrics.

In terms of average correlation of human and machine level metrics, it was found that SARI had the highest correlation with fiction preservation of 0.52, followed by METEOR of 0.15 and then BLEU of 0.15.For simplicity, only BLEU had positive correlation of 0.30 whilst SARI and METEOR had negative correlation of -0.27 and -0.40 respectively.

For fluency, SARI had the highest positive corelation of 0.36, followed by BLEU of 0.24 whilst METEOR had a negative correlation of -0.17.

For adequacy, all machine level metrics were weak positive correlations, with METEOR and BLEU having the same correlation of 0.25, followed by SARI of 0.10.

CHAPTER 6: Results Discussion and Evaluation

6.1 Text Feature Analysis Discussion and Evaluation

6.1.1 Key Results Discussion and Evaluation

The choice of chapter-based analysis did indeed provide a deeper perspective into how the various outputs performed. Overall, all text features varied from chapter to chapter with all outputs matching the trend, this was especially true for reading comprehension of chapter 1 to 2.

Due to the NLP landscape in favouring of non-fictional texts where chapters are irrelevant. [114,115] From this project one can hypothesis that future NLP fictional texts could benefit from delving into chapter-based analysis rather than whole text analysis, as observed in the text feature analysis where text features vary by chapter and the observed chapter performance of each output also varies as a result.

By text features alone, MUSS was the most successful output followed by human reference, BART and PEGASUS. It was expected that human reference to be the most successful due to the complexity of text simplification, [116] however the differences only differ by a few percentages and therefore MUSS dominated results could be statistically insignificant.

In conjunction with the benefits of chapter-based analysis and the negligible difference between MUSS and human reference text feature results. Future iterations of this project would benefit from a larger non-fiction dataset.

Due to resources and time constraints, only 5/12 of the total chapters were used and therefore useful chapter trends would have been lost. Another suggestion for improvement is the usage of more than one book, specifically ‘Alice Through the Looking Glass’ could also be used as it was written by the same author, therefore unquantifiable features such as writing style would have been consistent with project fiction dataset. [117]

As expected for BART and PEGASUS, the two processes were dominant in increasing unique word count. Whilst this result could be considered as merits, it is most likely not due to the common weakness of transformer-based models producing repetitive outputs. [118] A key lesson learned then, is that future iterations of the project should include as many text features as possible. Good performances in one text feature is insufficient in determining the overall effects of text simplification processes, as demonstrated by the overall low performance of BART and PEGASUS. Further text features could be more literature technique aligned such as including counts of metaphors, idioms, sarcasm and word rarity. [119,120]

6.1.2 Fiction and Non-Fiction Results Comparison Discussion and Evaluation

Comparing fiction and non-fiction texts from a text feature perspective. It was found that despite the favourability of non-fictional text of NLP processes, human reference had the most effect on text features like reading consensus in the non-fiction text feature analysis.

BART and PEGASUS had no effect on readability consensus and average readers age in the non-fiction text, unlike their fiction performances. A reason for this may be due to the naturally low complexity of the non-fictional text used. Therefore, the transformer-based models focusing on word and sentence level approaches, failed to achieve any substantial outputs from a text feature perspective. [121,122]

Future iterations of this should include a bigger non-fiction dataset, this will allow more variety in complexity and topic richness. Since it was expected that no-fictional text would favour NLP processes more, inclusion of bigger non-fiction dataset should display the status quo.

6.2 Machine Level Metric Analysis Discussion and Evaluation

6.2.1 Key Results Discussion and Evaluation: Best Performing Metric

Overall, in terms of machine level metrics, paraphrasing method of MUSS performed the best across all three, as expected given the established metric scores mentioned in the original paper. [98] Despite this, chapter-based analysis did show that MUSS performance had variability and trends across chapters. This can be partially explained by the difference of text features that defines each chapter, however only partial explanation can be suggested due to only few text features were used. Therefore, as an expansion to the improvement of adding more text features as mentioned in the previous section, another benefit would be better explanation for the performance variability and trends for the NLP outputs.

The limitation of the constrained fiction dataset can also be partially observed in both averaged and chapter-based results. For the overall results, both PEGASUS and BART shared very similar trends and scores, this could be explained by the fact that both are transformer based and therefore would share similar performance. However, given the sentence level attention in PEGASUS, one would expect PEGASUS to share similar trends to BART, for example weaker ml scores in the same chapters, but would not expect PEGASUS to share similar metric results. As mentioned in the original paper, it was proven to hold more leverage over BART in several tests. [97]

Therefore, a potential benefit in increase in fiction dataset would be a clearer performance difference between the two transformer-based models. Some observations to support this hypothesis can be found in the diverging trends of BART and PEGASUS from chapter 4 to 5 in all three machine learning metrics.

6.2.2 Key Results Discussion and Evaluation: Metric Trends in Relation with NLP Processes

When observing the across chapter average scores in fiction texts, it was clear to see that transformer-based models scored very well in BLEU, then SARI and then the worst in METEOR.

As explained in the methodology section, there have been studies on the errors of using BLEU for ATS evaluation. [123] In that, BLEU does not correlate with simplicity, therefore this could be the reason to why BLEU scores were very high compared to SARI and METEOR for both PEGASUS and BART.

Interestingly, transformer models always performed the worst in METEOR in the fiction texts. This could be explained as METEOR includes more text features in its scoring such as roots of words and paraphrasing, therefore since BART and PEGASUS are words and sentence level focuses. Their text summarisation objectives do not wholly align with METEOR parameters.

In terms of SARI, both transformer models scored average when compared to the other two metrics. A reason for this may be due to implementation as only one human reference was provided in the scoring, SARI's special feature is that a single input can have multiple references. Therefore, it may be argued that the project SARI scoring could be ineffective at evaluating the three NLP processes. For future iterations of this project, SARI implementation could be improved upon by using multiple human references that are from different simplification guidelines. Due to resource and time constraint, only SIMPLE guideline human reference was used.

Overall, when using metrics to evaluate the NLP outputs as mentioned in the methodology section. Each metric has its own flaws and omissions due to the NLP metric landscape providing numerous

incrementally different metrics. Therefore, for this project when considering the performance across all three ml metrics, MUSS was deemed to be the best performing NLP process.

One of the major reasons for this is that, as well as MUSS scoring the highest in all three metrics. MUSS scores for all three metrics were relatively close to each other, therefore showing no favourability to any individual metric. Indeed, from a machine learning metric perspective paraphrasing method of MUSS maybe the best approach in text simplification.

A counter argument for MUSS's performance as the best approach may lie in the fact that all three metrics require human referencing. Therefore, because the total scores are similarity comparisons to human referencing, it is difficult to say if MUSS is better than human referencing.

The performance difference between transformer-based models of PEGASUS and BART have so far shown to be significantly less effective than MUSS. That is, word and sentence level attention approach appear to be less effective than paraphrasing. One of the ways to perhaps challenge this question would be to use another type of NLP processes that improves on the abstractive summarisation nature of BART and PEGASUS.

Indeed, a pointer-generator model such as the HTSS would perhaps performed better as it seeks to improve the repetitiveness of abstractive summarisation models, as observed in the text feature analysis section. [87]

6.2.3 Fiction and Non-Fiction Results Comparison

MUSS was observed to be the favourable NLP process in both fiction and non-fiction results with the highest scores in all three metrics again. However, unlike in fiction results where metric scores were consistent, MUSS had relatively low scores in BLEU which was shared by PEGASUS and BART.

One of the reasons for this could be a combination of how the BLEU is computed and the relatively small non-fiction dataset. Specifically, because BLEU incorporates brevity penalty to try and include recall in its computations, the naturally short non-fiction dataset and the subsequently shorter outputs makes BLEU scores tend towards zero. An increase in non-fiction data size as a mentioned project improvement will correct this to a certain extent, however there is little project evidence to suggest MUSS would not remain the best NLP process in the new non-fiction dataset.

Interestingly, SARI and BLEU scores significantly varies between the two datasets. For fiction, BLEU was maximised whilst for non-fiction SARI was maximised. This suggests that ATS metrics heavily depends on the dataset applied to, more so than the NLP processes they're evaluating. A reason for this is due to the fact that human reference is derived from the datasets, which is the main factor in how an NLP process scores. Similar research has been carried out to show that there are favourable metrics depending on the type of dataset been used. [123]

Another hypothesis drawn from the dataset comparison would be that maybe BLEU scores and SARI scores have negative correlations with each other, therefore they should not be used together. However, literature has shown to contradict this and empirically demonstrated the positive correlation between BLEU and SARI scores.[124]

There are several reasons that could be suggested to explain the contradiction above. One is the constraint of the relatively small sized non-fiction dataset compared to the fiction-dataset, perhaps more data points for both fiction and non-fiction could show positive correlation. Another is that the chosen non-fiction dataset was small in terms of content such as sentence count and word count, this as mentioned above possibly made BLEU scores purposely small.

6.3 Human Level Metric Analysis Discussion and Evaluation

6.3.1 Key Results Discussion and Evaluation: HLM Trends Across Chapters

One of the key observations from the results section was that many of the chapter trends that HLM exhibited significantly varies at different chapters. Examples includes; for the human reference scores, both adequacy and fluency hit their minimum in chapter 3; for PEGASUS, adequacy and fiction preservation hit their minimum also in chapter 3; and finally for MUSS, simplicity, fluency and fiction preservation all changed their previous correlations within chapter 3.

Reasons for this may be found in the text feature analysis section. Specifically, whilst readability consensus was constant from chapter 2 to chapter 3 in the original text, suggesting that at least by readability RC perspective, chapters 2-5 were same in reading difficulty level. Chapter 3 had the lowest AWSP and ASSW out of all the other chapters.

Using the SRR model as mentioned before, it could be described that chapter 3 relatively high word recognition requirements due to a more difficult level of decoding, caused by low ASSW. However, in contrast, chapter 3 also had relatively low language comprehension due to its simpler language structure of the lower AWPS.

Since the project defined simplicity and fluency can be described as functions of language comprehension and word recognition. It could be suggested that because language comprehension and word comprehension values were more extreme when compared to other chapters, therefore more susceptible to change provided by the NLP processes. Then extreme changes in simplicity and fluency were more likely to occur in chapter 3.

One counter argument for the hypothesis above is that although chapter 3 did have more trend departures occurring. The constant RC from chapter 3-5 should've indicated trends for AWPS and ASSW to be relatively constant, however this is not true.

Therefore, perhaps some other text feature factors may be involved in the extreme departure of HML trends in chapter 3, or that readability consensus is ineffective at describing the text features of texts. For further iterations of this project, perhaps derived text features such as RC should not be used due to the skewing effects of averaging, therefore more direct text features should be used such as counting as explained in machine level metric discussion section.

Another key point in chapter-by-chapter analysis was the range to which each NLP processes scored. Human reference HLM scores in simplicity and fluency were of consistent range, therefore suggesting resistance in text feature change across chapters for those two scores.

BART was very sporadic in all 4 HML scores, just like its PEGASUS. It could be suggested that PEGASUS and BART were more sensitive in capturing the variance of text features across chapters, and therefore scores were heavily influenced and could be more susceptible to noise within the dataset.

MUSS's paraphrasing approach which has been stated to be closer to human reference also had a smaller range in all 4 HML scores similar to human reference text. From the perspective of consistency of HML scores then, MUSS could be argued to be closer to human reference than BART and PEGASUS.

6.3.2 Key Results Discussion and Evaluation: Best HML Scored NLP Process

A key observation found was that PEGASUS and BART improved the simplicity score when compared to the human reference benchmark. It could then be argued from observation that transformer-based models are good increasing simplicity of an original text. However, in wider context this is not a good result.

Simplicity is only one of the 4 HML metrics, and from observation transformer models scored very badly in fiction preservation and adequacy scores when compared to human metrics. As mentioned before using the SRR model, simplicity is a function of language comprehension and from the text feature analysis, transformer models excel at decreasing AWPS and increasing ASSW, therefore reducing the language structure factor of the language comprehension SRR variable. This in turn then decreases the reading comprehension requirement and therefore a sign in increasing in simplicity.

However, when considering simplicity with fiction preservation and adequacy, one can see that BART and PEGASUS, sacrificed quality of the translation in terms of content. Almost 50% of fiction preservation and adequacy is lost when compared to the human reference. As a result, simplicity (or decrease in language comprehension requirement) is further inflated due to decreases in Language comprehension factors such as background knowledge, vocabulary and literacy knowledge. All factors that are heavily related to the content of the text.

This unfortunately, shows the difficulty in creating precise HLM that measures something abstract such as simplicity. For future iterations of this project, better human level metrics could be implemented. In that rather than just measuring simplicity which could be achieved through decrease in information, the quality of simplicity could be measured instead. One example could be feature engineering of fusing adequacy and simplicity metric together to make an adequacy-simplicity percentage metric, however one downside is that it would be difficult to interpret.

6.3.3 Key Results Discussion and Evaluation: Fiction and Non-Fiction Results Comparison

In comparison of HLM performance between fiction and non-fiction datasets, as expected the non-fiction datasets had higher HML scores overall. As explained in the literature review and methodology section, this is due to the favourability of the NLP uses as they have been created and trained with non-fiction dataset in mind.

Therefore, by using the SRRM model, language and word comprehension factors within non-fiction datasets were less complex and therefore the natural reading comprehension for non-fiction datasets were lower. This could be the reason of why BART and PEGASUS did not increase in simplicity like in their fiction dataset, as the content was less complex, BART and PEGASUS had less information and complexity to produce repetitive outputs and therefore did not influence the simplicity score.

Another interesting point to focus on is the small percentage difference range of HLM scores in non-fiction than fiction. This observation could be explained again, by the favourability of non-fiction data sets, however another reason could also be the constraint of the dataset size. As suggested in the sections previously, non-fiction dataset was much smaller than fiction in both length and text density. As a result, this small percentage difference may vary if for future iterations of this project, bigger and more complex non-fiction dataset was to be used. However, it is still expected that non-fiction dataset would have a more consistent range of HML score for each NLP process.

A final locus of discussion would be the improvement of fluency in MUSS output over human text reference, even though that improve was very small of around 0.1%. It would be interesting to do a chapter-by-chapter analysis of fluency with MUSS and human text reference in future iterations with a bigger non-fiction dataset. This could provide insight to see if MUSS had significant improvement of fluency per non-fiction dataset and that the resulting percentage change could be bigger.

By using the SRRM, project defined fluency is also a function of language comprehension and word recognition. Fluency then is affected by all of the SRRM factors and therefore, it could be argued that MUSS's distance-based paraphrasing could not have considered all of the SRRM factors to affect fluency. Hence, the low percentage difference of 0.1% and what is likely to be statistically insignificant.

A more optimistic argument for MUSS's fluency improvement would be that the more higher level approach of paraphrasing rather than sentence or word level, can take into account more of the SRRM factors when dealing with HLM metrics that affects both language and word comprehension, therefore MUSS could have potential in being favourable at improving fluency of a text. Again a bigger non-fiction dataset is needed to pursuit this line of argument.

6.4 Machine and Human Level Metric Correlation Analysis Discussion and Evaluation

6.4.1 Key Results Discussion and Evaluation: Fiction Text Correlation

Using the raw results from the result section, it was observed that different metrics can have different correlations within the same NLP process. This is an example of the variety of ATS metrics that exists and how each of them only captures one perspective of texts. Therefore, this observation seems to agree with existing view that no-one ATS metric should be used in evaluating NLP text summarisation and simplification tasks. [125]

For the performance correlation of each of the NLP processes. It was surprising to learn the PEGASUS had the best fiction preservation and adequacy, a contrast in the HL metric analysis where transformer-based modes underperformed significancy in those areas when compared to MUSS.

One of the reasons that may explain the results above would be the interpretation of the HL metrics used by the volunteers, even though explanation of the metrics were included in the questionnaires. Using the attention mechanism that affects reading comprehension in the SRR model, it maybe that volunteers could only remember the important parts of the texts that are considered fiction, since the text was indeed very long, bias maybe introduced when volunteers try and recall important fictional content that appears in both the original text and NLP output texts. It is unlikely that a volunteer would remember all of the fictional content and therefore the summarisation nature of PEGASUS and BART skewed their answers away from MUSS which is paraphrasing.

One potential solution that could be implemented in future iterations of this project, is to expand on the fiction preservation metric to be more precise and countable. That is, volunteers could be asked to track a countable number of metaphors, ideas or idioms in the original text, and evaluate if their presence exist in the NLP output texts. One could hypothesis that this could lead to MUSS having dominated scores again as presented by the previous text feature, machine level and human level metric analysis.

Another key idea to discuss would be the overall low performance correlation between HL metrics and NLP processes. Whilst it could be interpreted as that overall, the three chosen NLP processes are not effective at affecting the human level metrics of a text. The original sample size of 8 volunteers were not ideal, even with a statistical significance test passed. Given the law of central limit theorem and literature to show that population-based variables exist as standard distributions, there is relative confidence that the chosen paired t-test was appropriate. Perhaps also using the law of large numbers, by increasing sample size further would make the correlations different.

Bigger sample size was considered and implemented but however due to the intensity of the questionnaires, dropout rates were among 50%. Therefore, a further practical improvement for the future is to use a focus group which were more varied. Due to GDPR, information about volunteers could not be mentioned but however this was advantageous in the context of commercial texts as the target audience were of laymen. If experts were to be used, then results would've been skewed and perhaps not of true market representation.

Control groups then could also be used in the future and compared against specific spectrums of learning-based disabilities to see how human level metric performs and their efficacy in describing texts.

Overall, for the fiction text perspective of how ML and HL correlate with each other. It appears that SARI was best with fiction preservation, adequacy, and second best in fluency. This was partly expected as SARI was created as a metric specifically for sentence simplification. Even though only one reference was provided as a human reference, the implementation of how an NLP process adds, removes and modifies input proved to be effective at correlating how humans interpret fiction preservation, adequacy and fluency.

By using the SRR model, the reason why SARI had the highest correlation with fluency and adequacy could be that. For fluency and adequacy, both are heavily influenced by word and language comprehension factors, such as sentence structures to dictate flow of speech, and similar words to dictate the ease of speech. SARI's lexical simplicity of tracking modifications of inputs relative to outputs could consider more of the mentioned factors, therefore closer to human judgement. This is consistent with other literature.[125]

BLEU and METEOR had similar correlations overall, this was partly expected as METEOR was an extension of BLEU to improve its recall capabilities beyond the brevity penalty. This then explains why BLUE and METEOR had very similar adequacy. One except to this was that METEOR had negative correlation with simplicity.

6.4.2 Key Results Discussion and Evaluation: Lack of Non-Fiction Comparison

Due to the time and resource constraint of the project as mentioned from the previous sections. One of the downsides of the significantly smaller non-fiction dataset is that correlations of human level and machine level metrics could not be carried it.

Therefore, another benefit to the bigger non-fiction dataset is to be able to compare the difference in human level and machine level correlations across both non-fiction and fiction-datasets. From literature it is also expected that SARI most correlated with the four human level metrics, based on the definition that fiction preservation, adequacy, simplicity and fluency are subsets of human judgement.

Since SARI has been shown to correlate highly with human judgement, further investigation maybe needed to see the correlations between the four human level metrics with the SARI paper definition of human judgement. [113]

CHAPTER 7: Limitations and Future Work

7.1 Improving the Execution of Current Project

One of the first aspects for future work is to improve and mitigate the current project constraints and limitations. Table 11 below provides a summary of the suggestions and reasons.

Summary of Project Execution Improvement	
Limitations and Their Consequences	**Improvements and Explanation**
Time constraints which lead to: • **Only 5/12 chapters were used which made BART and PEGASUS performance similar.** • **Unbalanced non-fiction dataset as the control group which contributed to NLP processes doing better in non-fiction.**	**More time will allow:** • All chapters for 'Alice in Wonderland' to be used, as well as its sequel 'Through the Looking Glass' by the same author to keep consistency. Allows more time for SIMPLE guideline to be used to produce human simplifications. • More balanced non-fiction dataset counterpart so more datapoints of comparison between each NLP process.
Volunteers used in questionnaires: • **Small dataset size for both fiction and non-fiction, as well as NLP processes explored. This is because the questionnaires required a heavy amount of reading. A week were given to the volunteers for this project but dropout rate was about 60%.**	**Dedicated Participants:** • Since volunteer dropout will not be a problem, more NLP processes could be used as well as the entire fiction and non-fiction dataset. • Controlled participants could be used to see how different skilled participants score. An in-depth analysis of the characteristics of participants in relative to the different metrics could be done.
Literature Human Level Metrics (simplicity, adequacy and fluency) • **Volunteers had different definitions of the human level metrics even though explanations were provided. This meant that the results were less interpretable, especially in human and machine level correlations.**	**Using Countable Human Level Metrics** • For example, in terms of fiction perseveration. Literature techniques of metaphors could be counted in the original text, and then counted again in the output texts. Therefore, because this is countable there is no bias in the human level metric of fiction preservation.
Fixed Attention Context Length of Transformers • **Since transformers were pre-trained, these can only have a maximum of around 50-word sentence length to compute attention. This would've significantly affected the results in fiction where contexts span in paragraphs.**	**Using Transformers-XL** • Transformers-XL is a recent expansion of normal transformers with a longer attention sentence length. • With more time, more coding could've been done to integrate Transformers-XL

Table 11. Summary to show project limitations, their effects and how improvements may mitigate those effects.

7.2 Using Transfer Learning to Expand Project Scope

Since the motivation of project was to evaluate the potential of automated commercial fiction text simplification. An expansion of this would be to move attempt to improve the current NLP technologies by training them with fiction dataset and then compare them with this project (with implemented changes as mentioned above).

One of the hypotheses for this would be potentially better scoring in fictional texts and perhaps the performance gap between fiction and non-fiction in all three analysis types would be smaller. The major reason for this is the consideration of '*transfer learning*', where NLP processes trained with fiction text would perform better if they were also trained on non-fiction text, therefore extra knowledge from related tasks may help. [126,127]

It could then be expanded where three groups of NLP processes could be compared using the analysis and evaluation framework provided by this project. That is NLP processes trained on non-fiction only, fiction only and both, could be compared to further provide evidence for difference between non-fiction and fiction text as shown in figure 27 below.

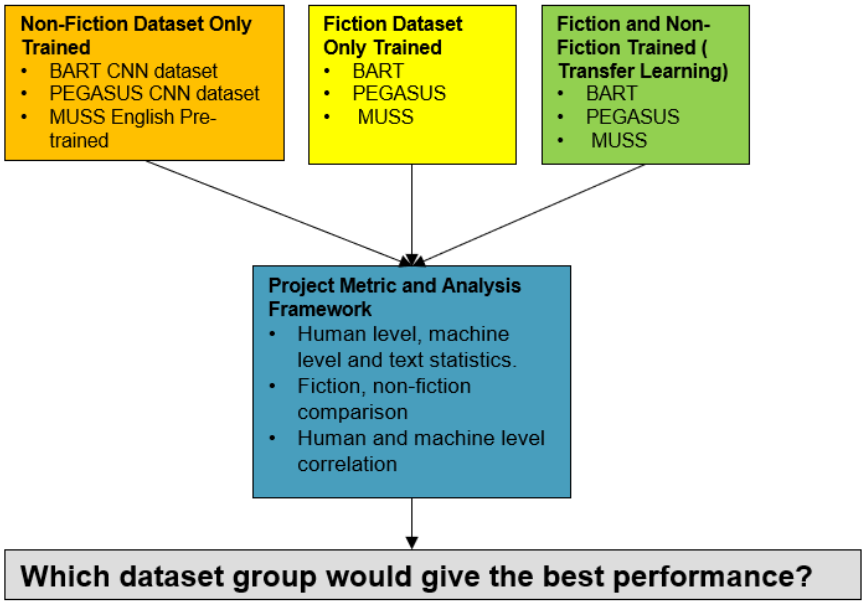


Figure 27. Sample of future work overview

CHAPTER 8: Conclusion

Many of the world's most famous film franchises originated from published content and with there being a direct economical link between the two industries. Individuals who may wish to explore the published contents that inspired their films may find them inaccessible due to factors such as language barriers and learned based difficulties.

Currently solutions do exist in that published content has to be written again which is time and resource consuming. Therefore, from an enterprise perspective, this project is technology market research and the evaluation of the potential of automated fiction text simplification, using current NLP technologies.

Three popular and developed NLP algorithms were chosen in the form of BART, PEGASUS transformer models for abstractive summarisation, and MUSS for text simplification. Alice in Wonderland was chosen for the fiction dataset and 2005 Azores Subtropical Wikipedia page was chosen as the non-fiction dataset for control group, as all NLP processes were trained on non-fiction dataset. Evaluations were carried out in the three perspectives of machine level metrics, human level metrics and text statistics.

The Overall Main Findings Were That:

1. SIMPLE guideline human simplification remains the best type of simplification, this is especially true in human metrics.
2. MUSS was the closes to human simplification in all analysis types of by text feature, machine level and human level, and in both fiction and non-fiction.
3. Chosen machine level and human level metrics appears to have low correlations in fiction text.
4. Performance gap between NLP and human simplification results in all analysis types widens in fiction compared to non-fiction.

Other Findings And Ideas Found Were That:

1. **Text Feature Analysis**
    a. Chapter-based analysis for fiction texts provided a deeper perspective into how the various NLP performed across chapters. Overall, there were some chapters that were consistent with all text outputs.
    b. Human simplification had the most effect on text features like reading consensus in non-fiction than in fiction.
2. **Machine Level Metric Analysis**
    a. All metrics performed differently on the same NLP processes, therefore agrees with literature that there is no uniform metric.
    b. NLP processes struggle with METEOR scoring the most and BLEU was the easiest to score in.
3. **Human Level Metric Analysis**
    a. Transformer based models maybe falsely best at increasing simplicity compared to human simplification reference, evidenced by up to 80% increase in unique words percentage compared to original text in text feature analysis.
4. **Fiction vs Non-Fiction Analysis**
    a. NLP processes generally performed closer to human simplification in both metrics in non-fiction than in fiction.
5. **Human and Machine Level Metric Correlations.**
    a. SARI had the best correlation with friction preservation of 0.52 and fluency of 0.36.

REFERENCES

1.
Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering [Internet]. 2010 Oct [cited 2019 Jun 4];22(10):1345–59. Available from: http://home.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf
2.
Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big Data. 2016 May 28;3(1).
3.
Mohammad SM. NLP Scholar: A Dataset for Examining the State of NLP Research [Internet]. ACLWeb. Marseille, France: European Language Resources Association; 2020 [cited 2022 Aug 22]. p. 868–77. Available from: https://aclanthology.org/2020.lrec-1.109/
4.
Xu W, Callison-Burch C, Napoles C. Problems in Current Text Simplification Research: New Data Can Help. Transactions of the Association for Computational Linguistics. 2015 Dec;3:283–97.
5.
Kauchak D, Mouradi O, Pentoney C, Leroy G. Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text [Internet]. IEEE Xplore. 2014 [cited 2022 Aug 22]. p. 2616–25. Available from: https://ieeexplore.ieee.org/abstract/document/6758930?casa_token=kAe72S_Wkn8AAAAA:wfPyo20vT7kgQ7w8ZoHsbQmPz0LKkr1L9jFTCSGK2uGeyZugaN5f78uKMv1S13tTT3Cqdaf1
6.
Carroll L. Through the Looking-Glass [Internet]. Project Gutenberg. 2008 [cited 2022 Aug 22]. Available from: https://www.gutenberg.org/ebooks/12
7.
Brasoveanu AMP, Andonie R. Visualizing Transformers for NLP: A Brief Survey. 2020 24th International Conference Information Visualisation (IV). 2020 Sep;
8.
Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F. Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts [Internet]. IEEE Xplore. 2017 [cited 2020 May 4]. p. 703–9. Available from: https://ieeexplore.ieee.org/document/8079931
9.
Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data. 2008 Jul 1;2(2):1–25.
10.
Alikaniotis D, Raheja V. The Unreasonable Effectiveness of Transformer Language Models in Grammatical Error Correction. arXiv:190601733 [cs] [Internet]. 2019 Jun 4; Available from: https://arxiv.org/abs/1906.01733
11.
Kovaleva O. Transformer Models in Natural Language Understanding: Strengths, Weaknesses and Limitations. 2021 [cited 2022 Aug 22]; Available from: https://www.proquest.com/docview/2583136463?pq-origsite=gscholar&fromopenview=true
12.
Barlow R, Vissandjée B. Determinants of National Life Expectancy. Canadian Journal of Development Studies / Revue canadienne d'études du développement. 1999 Jan;20(1):9–29.
13.
Lutz W, Cuaresma JC, Sanderson W. ECONOMICS: The Demography of Educational Attainment and Economic Growth. Science. 2008 Feb 22;319(5866):1047–8.
14.
Ranis G. Human Development and Economic Growth [Internet]. papers.ssrn.com. Rochester, NY; 2004. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=551662
15.
Cambridge Dictionary. LITERACY | meaning in the Cambridge English Dictionary [Internet]. Cambridge.org. 2019. Available from: https://dictionary.cambridge.org/dictionary/english/literacy
16.
Reder S. Adult Literacy Development and Economic Growth [Internet]. ERIC. National Institute for Literacy; 2010 [cited 2022 Jul 27]. Available from: https://eric.ed.gov/?id=ED512441
17.

Bulled NL, Sosis R. Examining the Relationship between Life Expectancy, Reproduction, and Educational Attainment. Human Nature. 2010 Oct;21(3):269–89.
18.
Meletis D, Dürscheid C. Writing Systems and Their Use. De Gruyter; 2022.
19.
Napier CJ. Accounting history and accounting progress. Accounting History. 2001 Nov;6(2):7–31.
20.
Lerner F. The Story of Libraries, Second Edition: From the Invention of Writing to the Computer Age [Internet]. Google Books. Bloomsbury Academic; 2009 [cited 2022 Jul 27]. Available from: https://books.google.co.uk/books?hl=en&lr=&id=ULlIKmxQc7EC&oi=fnd&pg=PR7&dq=invention+of+writing&ots=48hVgGH2Zz&sig=W2mCFZx4ercG3RT64qsEYNLipnQ&redir_esc=y#v=onepage&q=invention%20of%20writing&f=false
21.
Deshpande N, Kumar A, Ramaswami R. The Effect of National Healthcare Expenditure on Life Expectancy. Gatechedu [Internet]. 2014; Available from: https://smartech.gatech.edu/handle/1853/51648
22.
Senner WM. The Origins of Writing [Internet]. Google Books. U of Nebraska Press; 1991 [cited 2022 Jul 27]. Available from: https://books.google.co.uk/books?hl=en&lr=&id=Kc4xAlunCSEC&oi=fnd&pg=PP9&dq=invention+of+writing&ots=CqsS0GRgAo&sig=w_R4Po3dJntTYR0C3t2L3bS5rG8&redir_esc=y#v=onepage&q=invention%20of%20writing&f=false
23.
BFI Research & Statistics Units,. THE UK FILM ECONOMY [Internet]. 2017. Available from: https://www2.bfi.org.uk/sites/bfi.org.uk/files/downloads/bfi-uk-film-economy-2019-01-30.pdf
24.
Wikipedia Contributors. List of highest-grossing films [Internet]. Wikipedia. Wikimedia Foundation; 2019. Available from: https://en.wikipedia.org/wiki/List_of_highest-grossing_films
25.
Clark K, Stern L. Synergies between the Publishing and Film Production Industries: Where does the Consumer Fit In? [Internet]. 2006 [cited 2022 Jun 20]. Available from: https://www.stern.nyu.edu/sites/default/files/assets/documents/con_043271.pdf
26.
Bloomsbury. HARRY POTTER: Accessible Editions and Other Languages: Books: Bloomsbury Publishing (UK) [Internet]. www.bloomsbury.com. Available from: https://www.bloomsbury.com/uk/harry-potter/accessible-editions-and-other-languages/
27.
Moen H, Peltonen L-M, Heimonen J, Airola A, Pahikkala T, Salakoski T, et al. Comparison of automatic summarisation methods for clinical free text notes. Artificial Intelligence in Medicine. 2016 Feb;67:25–37.
28.
Scott D, Hallett C, Fettiplace R. Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories. Patient Education and Counseling. 2013 Aug;92(2):153–9.
29.
Lawler A. ARCHAEOLOGY: Writing Gets a Rewrite. Science. 2001 Jun 29;292(5526):2418–20.
30.
Gopnik A, Griffiths TL, Lucas CG. When Younger Learners Can Be Better (or at Least More Open-Minded) Than Older Ones. Current Directions in Psychological Science. 2015 Apr;24(2):87–92.
31.
Kuhl PK. Early language acquisition: cracking the speech code. Nature Reviews Neuroscience [Internet]. 2004 Nov;5(11):831–43. Available from: https://www.nature.com/articles/nrn1533
32.
Hoover WA, Gough PB. The simple view of reading. Reading and Writing. 1990 Jun;2(2):127–60.
33.
Luft Baker D, Al Otaiba S, Ortiz M, Correa V, Cole R. Chapter Ten - Vocabulary Development and Intervention for English Learners in the Early Grades. Advances in Child Development and Behavior. 2014;46:281–338.
34.
Lonigan CJ, Burgess SR, Schatschneider C. Examining the Simple View of Reading With Elementary School Children: Still Simple After All These Years. Remedial and Special Education. 2018 Sep;39(5):260–73.

35.
Snow CE. Simple and Not-So-Simple Views of Reading. Remedial and Special Education. 2018 Sep;39(5):313–6.
36.
Education MD of E and S. What Is the Simple View of Reading? - Evidence Based Early Literacy [Internet]. www.doe.mass.edu. 2021. Available from: https://www.doe.mass.edu/massliteracy/skilled-reading/simple-view.html
37.
Duke NK, Cartwright KB. The Science of Reading Progresses: Communicating Advances Beyond the Simple View of Reading. Reading Research Quarterly. 2021 May 7;56(S1).
38.
Hoover WA, Tunmer WE. The Simple View of Reading: Three Assessments of Its Adequacy. Remedial and Special Education. 2018 Sep;39(5):304–12.
39.
Kendeou P, Savage R, Broek P. Revisiting the simple view of reading. British Journal of Educational Psychology. 2009 Jun;79(2):353–70.
40.
Rose J. Independent review of the teaching of early reading : final report. Nottingham: Dept. For Education And Skills; 2006.
41.
Byrne B, Fielding-Barnsley R. Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. Journal of Educational Psychology. 1995;87(3):488–503.
42.
Johnston TC, Kirby JR. The Contribution of Naming Speed to the Simple View of Reading. Reading and Writing. 2006 Jun;19(4):339–61.
43.
Malatesha Joshi R, Aaron PG. THE COMPONENT MODEL OF READING: SIMPLE VIEW OF READING MADE A LITTLE MORE COMPLEX. Reading Psychology. 2000 Apr;21(2):85–97.
44.
Catts HW, Adlof SM, Weismer SE. Language Deficits in Poor Comprehenders: A Case for the Simple View of Reading. Journal of Speech, Language, and Hearing Research. 2006 Apr;49(2):278–93.
45.
Taboada Barber A, Cartwright KB, Hancock GR, Klauda SL. Beyond the Simple View of Reading: The Role of Executive Functions in Emergent Bilinguals' and English Monolinguals' Reading Comprehension. Reading Research Quarterly. 2021 Mar 2;
46.
Quinn JM, Wagner RK, Petscher Y, Roberts G, Menzel AJ, Schatschneider C. Differential codevelopment of vocabulary knowledge and reading comprehension for students with and without learning disabilities. Journal of Educational Psychology. 2020 Apr;112(3):608–27.
47.
Foorman BR, Petscher Y, Herrera S. Unique and common effects of decoding and language factors in predicting reading comprehension in grades 1–10. Learning and Individual Differences. 2018 Apr;63:12–23.
48.
Ho ESC, Lau K. Reading engagement and reading literacy performance: effective policy and practices at home and in school. Journal of Research in Reading. 2018 Nov;41(4):657–79.
49.
Tong X, Deacon SH, Kirby JR, Cain K, Parrila R. Morphological awareness: A key to understanding poor reading comprehension in English. Journal of Educational Psychology. 2011 Aug;103(3):523–34.
50.
Scarborough HS. Connecting Early Language and Literacy to Later Reading (Dis)abilities: Evidence, Theory, and Practice. Neuman SB, Dickinson DK, editors. Handbook of Early Literacy Research. 2001;1:98.
51.
NHS. Dyslexia [Internet]. NHS. 2018. Available from: https://www.nhs.uk/conditions/Dyslexia/
52.
British Dyslexia Association. What Is dyslexia? [Internet]. British Dyslexia Association. 2009. Available from: https://www.bdadyslexia.org.uk/dyslexia/about-dyslexia/what-is-dyslexia
53.

Rose J. Identifying and teaching children and young people with dyslexia and literacy difficulties. Nottingham Dcsf Publications; 2009.
54.
NHS. Symptoms - Dyslexia [Internet]. NHS. 2019. Available from: https://www.nhs.uk/conditions/dyslexia/symptoms/
55.
Peyrin C, Lallier M, Démonet JF, Pernet C, Baciu M, Le Bas JF, et al. Neural dissociation of phonological and visual attention span disorders in developmental dyslexia: FMRI evidence from two case reports. Brain and Language. 2012 Mar;120(3):381–94.
56.
Beneventi H, Tønnessen FE, Ersland L, Hugdahl K. Working Memory Deficit in Dyslexia: Behavioral and fMRI Evidence. International Journal of Neuroscience. 2010 Jan;120(1):51–9.
57.
Shaywitz SE, Shaywitz BA. Paying attention to reading: The neurobiology of reading and dyslexia. Development and Psychopathology. 2008;20(4):1329–49.
58.
Hyona J, Olson RK. Eye fixation patterns among dyslexic and normal readers : effects of word length ¨ and word frequency. Journal of Experimental Psychology : Learning, Memory, and Cognition,. 1995;21(6):1430–40.
59.
Blazely AM, Coltheart M, Casey BJ. Semantic impairment with and without surface dyslexia: Implications for models of reading. Cognitive Neuropsychology. 2005 Sep;22(6):695–717.
60.
Caramazza A, Miceli G, Silveri MC, Laudanna A. Reading mechanisms and the organisation of the lexicon: Evidence from acquired dyslexia. Cognitive Neuropsychology. 1985 Feb;2(1):81–114.
61.
Behrmann M, Bub D. Surface dyslexia and dysgraphia: dual routes, single lexicon. Cognitive Neuropsychology. 1992 Jun;9(3):209–51.
62.
WARRINGTON EK, SHALLICE T. SEMANTIC ACCESS DYSLEXIA. Brain. 1979;102(1):43–63.
63.
IBM Cloud Education. What is Natural Language Processing? [Internet]. Ibm. 2020. Available from: https://www.ibm.com/cloud/learn/natural-language-processing
64.
Vajjala S, Meurers D. Readability assessment for text simplification. Recent Advances in Automatic Readability Assessment and Text Simplification. 2014 Dec 31;165(2):194–222.
65.
Ambati BR, Reddy S, Steedman M. Assessing Relative Sentence Complexity using an Incremental CCG Parser. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016;
66.
Vajjala S, Meurers D. Assessing the relative reading level of sentence pairs for text simplification. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014;
67.
Štajner S, Hulpuş I. Automatic Assessment of Conceptual Text Complexity Using Knowledge Graphs [Internet]. ACLWeb. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018 [cited 2022 Aug 13]. p. 318–30. Available from: https://aclanthology.org/C18-1027/
68.
Radev DR, Hovy E, McKeown K. Introduction to the Special Issue on Summarization. Computational Linguistics. 2002 Dec;28(4):399–408.
69.
Harabagiu S, Lacatusu F. Using topic themes for multi-document summarization. ACM Transactions on Information Systems. 2010 Jun;28(3):1–47.
70.
Vanderwende L, Suzuki H, Brockett C, Nenkova A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management. 2007 Nov;43(6):1606–18.
71.

Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing & Management. 2003 Jan;39(1):45–65.
72.
Hurtado C, Gutiérrez C. Reasoning about summariz-ability in heterogeneous multidimensional schemas. ACM Trans Database Syst [Internet]. 1996 [cited 2022 Aug 13];30(3):185–96. Available from: https://cs.uwaterloo.ca/~jimmylin/publications/Lin_SummarizationEntry2009.pdf
73.
Cheung J. Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection. undefined [Internet]. 2008 [cited 2022 Aug 13]; Available from: https://www.semanticscholar.org/paper/Comparing-Abstractive-and-Extractive-Summarization-Cheung/09cbab69d589058ed1ad6511a0f881aaa1a1efdd
74.
Kartheek Rachabathuni P. A survey on abstractive summarization techniques. 2017 International Conference on Inventive Computing and Informatics (ICICI). 2017 Nov;
75.
Sunitha C, Jaya A, Ganesh A. A Study on Abstractive Summarization Techniques in Indian Languages. Procedia Computer Science. 2016;87:25–31.
76.
Barzilay R, McKeown KR. Sentence Fusion for Multidocument News Summarization. Computational Linguistics. 2005 Sep;31(3):297–328.
77.
Yousif-Monod M, Prince V. Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening. International Conference on Computational Linguistics. 2008;2008.
78.
Alshaina S, John A, Nath AG. Multi-document abstractive summarization based on predicate argument structure. 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES). 2017 Aug;
79.
Carenini G, Cheung JCK, Pauls A. MULTI-DOCUMENT SUMMARIZATION OF EVALUATIVE TEXT. Computational Intelligence. 2012 Apr 23;29(4):545–76.
80.
Barzilay R, Lee L. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment [Internet]. ACLWeb. 2003 [cited 2022 Aug 13]. p. 16–23. Available from: https://aclanthology.org/N03-1003/
81.
Filippova K. Multi-Sentence Compression: Finding Shortest Paths in Word Graphs [Internet]. ACLWeb. Beijing, China: Coling 2010 Organizing Committee; 2010 [cited 2022 Aug 13]. p. 322–30. Available from: https://aclanthology.org/C10-1037/
82.
Mehdad Y, Carenini G, Tompa F, Ng RT. Abstractive Meeting Summarization with Entailment and Fusion [Internet]. ACLWeb. Sofia, Bulgaria: Association for Computational Linguistics; 2013 [cited 2022 Aug 13]. p. 136–46. Available from: https://aclanthology.org/W13-2117/
83.
Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need [Internet]. 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
84.
Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context [Internet]. arXiv.org. 2019. Available from: https://arxiv.org/abs/1901.02860
85.
Lample G, Conneau A. Cross-lingual Language Model Pretraining. arXiv:190107291 [cs] [Internet]. 2019 Jan 22; Available from: https://arxiv.org/abs/1901.07291
86.
Siddharthan A. A survey of research on text simplification. Recent Advances in Automatic Readability Assessment and Text Simplification. 2014 Dec 31;165(2):259–98.
87.

Evans R, Orasan C, Dornescu I. An evaluation of syntactic simplification rules for people with autism. Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). 2014;
88.
Inui K, Fujita A, Takahashi T, Iida R, Iwakura T. Text simplification for reading assistance. Proceedings of the second international workshop on Paraphrasing -. 2003;
89.
See A, Liu PJ, Manning CD. Get To The Point: Summarization with Pointer-Generator Networks. arXiv:170404368 [cs] [Internet]. 2017 Apr 25; Available from: https://arxiv.org/abs/1704.04368
90.
Zaman F, Shardlow M, Hassan S-U, Aljohani NR, Nawaz R. HTSS: A novel hybrid text summarisation and simplification architecture. Information Processing & Management. 2020 Nov;57(6):102351.
91.
Vinyals O, Fortunato M, Jaitly N. Pointer Networks [Internet]. Vol. 28, Neural Information Processing Systems. Curran Associates, Inc.; 2015 [cited 2022 Aug 14]. Available from: https://papers.nips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html
92.
Surya S, Mishra A, Laha A, Jain P, Sankaranarayanan K. Unsupervised Neural Text Simplification. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019;
93.
Nisioi S, Štajner S, Ponzetto SP, Dinu LP. Exploring Neural Text Simplification Models [Internet]. ACLWeb. Vancouver, Canada: Association for Computational Linguistics; 2017 [cited 2022 Aug 14]. p. 85–91. Available from: https://aclanthology.org/P17-2014/
94.
Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Translational Vision Science & Technology [Internet]. 2020 Jan 28;9(2):14–4. Available from: https://tvst.arvojournals.org/article.aspx?articleid=2762344
95.
Vincent P, Ca P, Larochelle H, Toronto L, Edu, Lajoie I, et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion Yoshua Bengio Pierre-Antoine Manzagol. Journal of Machine Learning Research [Internet]. 2010;11:3371–408. Available from: https://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf
96.
Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv.org. 2018. Available from: https://arxiv.org/abs/1810.04805
97.
Lu H, Wu Z, Wu X, Li X, Kang S, Liu X, et al. VAENAR-TTS: Variational Auto-Encoder based Non-AutoRegressive Text-to-Speech Synthesis. arXiv:210703298 [cs, eess] [Internet]. 2021 Jul 7 [cited 2022 Aug 14]; Available from: https://arxiv.org/abs/2107.03298
98.
Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:191013461 [cs, stat] [Internet]. 2019 Oct 29; Available from: https://arxiv.org/abs/1910.13461
99.
Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:191208777 [cs] [Internet]. 2020 Jul 10; Available from: https://arxiv.org/abs/1912.08777
100.
Martin L, Fan A, de la Clergerie É, Bordes A, Sagot B. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. arXiv:200500352 [cs] [Internet]. 2021 Apr 16 [cited 2022 Aug 16]; Available from: https://arxiv.org/abs/2005.00352#:~:text=We%20introduce%20MUSS%2C%20a%20Multilingual
101.
Artetxe M, Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics. 2019 Mar;7:597–610.
102.
Papineni K, Roukos S, Ward T, Zhu W-J. BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02 [Internet]. 2001; Available from: https://dl.acm.org/citation.cfm?id=1073135

103.
Daniel J, Martin J. Speech and Language Processing [Internet]. Available from: https://web.stanford.edu/~jurafsky/slp3/4.pdf
104.
Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics. 2016 Dec;4:401–15.
105.
Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments [Internet]. ACLWeb. Ann Arbor, Michigan: Association for Computational Linguistics; 2005 [cited 2022 Aug 16]. p. 65–72. Available from: https://aclanthology.org/W05-0909/
106.
Salimi J. Machine Translation Of Fictional And Non-fictional Texts: An examination of Google Translate's accuracy on translation of fictional versus non-fictional texts. DiVa [Internet]. 2014 [cited 2022 Aug 30]; Available from: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A737887&dswid=1413
107.
Clausen Y, Nastase V. Metaphors in Text Simplification: To change or not to change, that is the question [Internet]. ACLWeb. Florence, Italy: Association for Computational Linguistics; 2019 [cited 2022 Aug 30]. p. 423–34. Available from: https://aclanthology.org/W19-4444/
108.
Carroll L. Alice's Adventures in Wonderland [Internet]. Project Gutenberg. 2008 [cited 2021 Sep 7]. Available from: https://www.gutenberg.org/ebooks/11
109.
2005 Azores subtropical storm [Internet]. Wikipedia. 2022 [cited 2022 Aug 30]. Available from: https://en.wikipedia.org/wiki/2005_Azores_subtropical_storm
110.
HARRY POTTER: Accessible Editions and Other Languages: Books: Bloomsbury Publishing (UK) [Internet]. www.bloomsbury.com. Available from: https://www.bloomsbury.com/uk/harry-potter/accessible-editions-and-other-languages/
111.
Accessible Publishing [Internet]. www.accessiblebooksconsortium.org. Available from: https://www.accessiblebooksconsortium.org/publishing/en/
112.
Suleiman D, Awajan A. Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges [Internet]. Mathematical Problems in Engineering. 2020. Available from: https://www.hindawi.com/journals/mpe/2020/9365340/
113.
http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1321076 [Internet]. DIVA. [cited 2022 Aug 30]. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1321076&dswid=-4061
114.
Song S, Huang H, Ruan T. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications. 2018 Feb 16;78(1):857–75.
115.
Widyassari AP, Affandy A, Noersasongko E, Fanani AZ, Syukur A, Basuki RS. Literature Review of Automatic Text Summarization: Research Trend, Dataset and Method [Internet]. IEEE Xplore. 2019 [cited 2022 Aug 30]. p. 491–6. Available from: https://ieeexplore.ieee.org/abstract/document/8938454?casa_token=LsqtYL4QeKYAAAAA:J9iEokuQVeRKglN5ZGSrga0lDKxCe09a_TyjliJ3ofZtmncwXG9WtQKJy_wppyM3q7S_hUoG
116.
Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context [Internet]. arXiv.org. 2019. Available from: https://arxiv.org/abs/1901.02860
117.
See A. abisee/pointer-generator [Internet]. GitHub. 2022 [cited 2022 Aug 30]. Available from: https://github.com/abisee/pointer-generator
118.
Sulem E, Abend O, Rappoport A. Semantic Structural Evaluation for Text Simplification [Internet]. Available from: https://www.cis.upenn.edu/~eliors/papers/samsa.pdf
119.

Papineni K, Roukos S, Ward T, Zhu W-J. BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02 [Internet]. 2001; Available from: https://dl.acm.org/citation.cfm?id=1073135
120.
Wubben S, van den Bosch A, Krahmer E. Sentence Simplification by Monolingual Machine Translation [Internet]. ACLWeb. Jeju Island, Korea: Association for Computational Linguistics; 2012 [cited 2022 Aug 30]. p. 1015–24. Available from: https://aclanthology.org/P12-1107/
121.
Štajner S, Mitkov R, Saggion H. One Step Closer to Automatic Evaluation of Text Simplification Systems [Internet]. ACLWeb. Gothenburg, Sweden: Association for Computational Linguistics; 2014 [cited 2022 Aug 30]. p. 1–10. Available from: https://aclanthology.org/W14-1201/
122.
My Byline Media. AUTOMATIC READABILITY CHECKER, a Free Readability Formula Consensus Calculator [Internet]. readabilityformulas.com. Available from: https://readabilityformulas.com/free-readability-formula-tests.php
123.
Bilgiler S, Dergisi E, Solnyshkina M, Zamaletdinov R, Gorodetskaya L, Gabitov A. Journal of Social Studies Education Research Evaluating Text Complexity and Flesch-Kincaid Grade Level [Internet]. Available from: https://files.eric.ed.gov/fulltext/EJ1162266.pdf
124.
Wikipedia Contributors. tf–idf [Internet]. Wikipedia. Wikimedia Foundation; 2019. Available from: https://en.wikipedia.org/wiki/Tf%E2%80%93idf
125.
Cer D, Manning CD, Jurafsky D. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization [Internet]. ACLWeb. Los Angeles, California: Association for Computational Linguistics; 2010 [cited 2022 Aug 31]. p. 555–63. Available from: https://aclanthology.org/N10-1080/
126.
Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics [Internet]. 2016 [cited 2022 Aug 31];2016(4):401–15. Available from: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00107/43364/Optimizing-Statistical-Machine-Translation-for
127.
Alva-Manchego F, Scarton C, Specia L. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. Computational Linguistics. 2021 Dec;47(4):861–89.

APPENDIX

Repository

Project Gitlab link: https://git-teaching.cs.bham.ac.uk/mod-msc-proj-2021/ecl734

MUSS Github link: https://github.com/facebookresearch/muss

- For MUSS, instructions and dependencies are in the Github page. Use desired virtual environment or a LINUX machine with python version that is less than 3.9.

Gitlab Repository Description and Technical Documentation

Table below shows what each file / directory is.

File / Directory	Description
Main.ipynb	Jupyter notebook file containing the following sections: 1. Pre-processing. 2. BART, PEGASUS implementation. 3. Metrics implementation. 4. Data analysis
MUSS_output_post_processing.ipynb	Jupyter notebook file containing post processing of MUSS results.
Dataset/	Contains all raw and processes fiction and non-fiction data including: • All NLP outputs • Human simplification outputs
Graphs/	Contains all the graphs used produced from data analysis.
Analysis_data/	Contains all the raw and post processed questionnaire data for both fiction and non-fiction.
readme	Contains all the dependencies and library versions used in this project