Transformer-based Amharic Complexity Classification and Simplification

Abstract

Amharic is a Semitic family language widely spoken in Ethiopia. Based on expertise recommendation, some of the document organized using this language contains complex texts that need further simplification. Such complexity is the level of difficultness of the text for understanding by the target readers. In addition to humans, this complex text challenges different NLP applications such as machine translation. To address this issue, we have developed three sequential models such as complexity classification, complex term detection, and simple text generation models. For the first model, we have used the pre-trained transformer-based models such as BERT and XLNET to train these models. 33.9k Amharic sentences are used, and for building the detection model 1002 complex terms are used. Lastly, 91k Amharic sentences are used to build the simple text generation model such as Word2Vec, Fastext, and Roberta. As the experimental result shows, the classification models such as BERT and XLNET score an accuracy of 86.1% and 70% respectively. For the specific complex term detection and to generate the simple equivalent text, the Word2Vec model has better prediction and ranking results. This Word2Vec generates the most similar simple terms with a cosine similarity of 0.91, while the Fastext scores 0.85 and Roberta 0.57. Addressing the syntactic complexity of Amharic text is our recommendation in this work for future research.

1 Introduction

Text documents like academic textbooks utilize a wide variety of vocabularies when organizing them. Some of the vocabularies in the document may not be familiar for readers which leads to text complexity, because vocabulary and prior knowledge are well-known determinants for reading comprehension ability (Speech et al., 2021). One of the reasons for the occurrence of such text complexity is due to the existence of unfamiliar words in a document, which is lexical complexity (Ide et al., 2023). Lexical complexity is one of the main reasons leading to overall text complexity and thus results in reading comprehension difficulty for readers who have low literacy in the language (Pan et al., 2021). When the word is frequently accessible, it becomes familiar to the readers so the document organized based on such frequent words is easily understandable by readers who have low levels of knowledge on the language. Such familiarity level of document content is used to estimate the readability of text (Nation & Snowling, 2000). Recently, scholars have conducted works that indicate give emphasize the need for increased attention to text document organization related to its lexical complexity level in primary-level classrooms (Read, 2019). The benefit of having complex terms such as non-frequent, scientific, and mathematical terms in a textual document is to extend the scientific concepts of the readers during their study (Prof & Akba, 2016). However, for young students who are still developing literacy skills, as well as academic vocabulary need to teach these terms by generating more meaning were suggested. The reason for giving attention to the lexical complexity of such text documents is it will help to improve student's understandability, problem-solving strategies, and dispositions toward academic reading (Arya et al., 2011). Furthermore, generating equivalent simpler meanings for such challenging terms in the academic text will assist teachers, curriculum planners, and textbook authors in countering poor performance in the subject (Mulwa, 2015). When the vocabulary of science texts is dense and complex it is criticized for being inaccessible because it introduces the reader to many unfamiliar words and the teachers may fails to explain them in ways that connect with the student's prior knowledge and experiences (Snow, 2010). This lexical complexity is the primary reason for text complexity because researchers claim that simplifying a text does not necessarily improve understanding, unless, increases individual terms that a learner can understand (Shirzadi, 2014). Amharic is a Semitic family language and it is one research area for many NLP applications such as text classification (Kelemework, 2013), machine translation (Sulem et al., 2018), and complexity classification (Nigusie & Tegegne, 2022). To minimize such text complexities in some academic concerns documents for resource-rich languages have guidelines (Solution, 2021). Recently researchers have also shifted towards developing deep learning models to address these text complexities dynamically and more convenient way such as a neural network model for the evaluation of text complexity in the Italian language (Lo Bosco et al., 2018) and predicting lexical complexity in English texts (Shardlow et al., 2022). However, for the Amharic language, there is no standardized guideline to minimize such text complexity. Our main concern in this paper is classifying Amharic text complexity and generating its simpler equivalent using transformer-based as well as unsupervised embedding models. This helps to make information more accessible to low-literacy readers including children, and non-native speakers (Bott et al., 2012). Furthermore, it has a valuable reprocessing stage for different NLP applications, such as machine translation (Sulem et al., 2018).

2 Related Work

Texts containing highly challenging vocabularies and complex sentence structures are likely to dimension the learners' reading comprehension because various factors impact learners' reading comprehension. Some of these factors involve the learners' vocabulary knowledge, grammar knowledge, reading strategies, and motivation (Gilakjani & Sabouri, 2018). Identifying those words that can cause difficulty for a reader is an important step in the lexical simplification process for assessing text readability (Qiang et al., 2019). It also helps to enhance the reading and understanding capability of text for low literacy readers and second language learners. Nowadays due to the abundance of large textual documents and the emerging of machine learning algorithms, classifying text to its target using models trained by large text data becoming a popular technique (Gasparetto et al., 2022). Measuring the appropriateness of text to particular readers widely in the education field to organize text based on learner's understanding level and to support educators in drafting textbooks is one application area of these machine learning algorithms (Review, 2021). Texts containing unfamiliar terms and complex structures are likely to decrease readability and understandability by low literacy readers, therefore identifying complex words and sentences is an important step towards assessing text readability (Qiang et al., 2019). Lexical complexity is the first and highly impacted type of complexity, thus various scholars suggest increasing the number of familiar vocabulary to increase the readability and understandability of the document (Young, 1999). This raises the issue for determining the relationship between the number of easily understandable words and overall text comprehension (Hu & Nation, 2000). To address this text complexity problem number of works are conducted for different languages, such as measuring the complexity of a text using a supervised classification model to evaluate the language abilities of non-native speakers of Italian (Santucci et al., 2020), using 692 sentences collected from certification materials. The experiment is conducted on classical machine learning models and they have achieved better classification using SVM and RF with an accuracy of 72.5% and 71.7% respectively. Detecting such text complexities using Multinomial Naive Bayes archives an accuracy of 84% using TF-IDF weighting and 10-fold cross-validation (Hidayat, 2019). Recently the research on text complexity classification shifted towards using deep neural network models, for large dataset sizes and to handle semantics and sophisticated features of the dataset (Gasparetto et al., 2022). Due to such reasons, the latest works are using neural network models such as LSTM (Lo Bosco et al., 2018). Different pre-trained versions of this neural network model such as BERT (Kenton et al., 2019), RoBERTa (Pan et al., 2021), ALBERT, and ERNIE are also used to predict the complexity level of biomedical texts using 9476 annotated datasets and RoBERTa shows better performance than other models with MAE 0.0715 and MSE 0.0085. Addressing the lexical simplification process is their recommendation. Similar to other languages there are benchmark works for Amharic language complexity classification using machine learning. The work was conducted using 5126 sentences for binary classification of Amharic text complexity, experimenting with SVM (87.1%), NB (83%), and RF (80.4%) (Nigusie & Tegegne, 2022). In this work we have addressed such Amharic text complexity classification and simplification problems using recently emerged pre-trained models.

3 Amharic Text Dataset

Sources such as low-grade students' textbooks, published news, academic social media pages, and blogs are used for dataset collection. These academic concern sources are used as the main data source for measuring text complexity (Sen & Fuping, 2021). We have collected 33.8k sentences to build the classification models, the dataset is distributed through half-complex and half-noncomplex sentences. The sentences are labeled based on their lexical unfamiliarity. The sentences that contain unfamiliar words are labeled as complex and the sentences formed from familiar words are labeled as noncomplex sentences. To confirm the sentence complexity, we have used sentences containing complex words that are identified from academic textbooks (Belete et al., 2015; Alemu et al., 2015). Furthermore, to accurately label the sentence to its target, we have provided 10 pages of the document and distributed it to three Amharic language literatures, as we have evaluated individual responses and inter-annotation agreements they have identified a total of 126 sentences as complex. From such sentences, three of them contain phrase-level complexity. The rest 123 sentences are identified as complex that contain unfamiliar words. For the detection model we have collected 1002 complex terms and 91k sentences (complex sentences with their meaning) are collected for the simplification model.

4 Complex Words and their Meaning 235 Part-of-speech Tagging

Part-of-speech (POS) tagging is the process of classifying words into its lexical categories or word classes (Gambäck et al., 2009). In this work, this tagging process helps to identify which POS class of Amharic words have more complex terms (see Table 1). So in this stage, we are trying to identify the complex words POS with their equivalent simple meaning. To get the words POS we have used HornMorpho Amharic morphological analyzer (Mulugeta & Gasser, 2012).

Part-of-speech	**Complex words**	**Simple equivalents**
Noun	464	1815
Verb	236	1100
Adverb	2	0
Uncategorized	300	480

Table 1: Complex terms with their meaning Part-of-speech

5 Complex Terms Distribution

The complex word distributions across the training dataset are visualized in Figure 1. The graph is generated using maximum sentence frequency that contains a single complex term and minimum frequency of the existence of one complex term in the dataset.

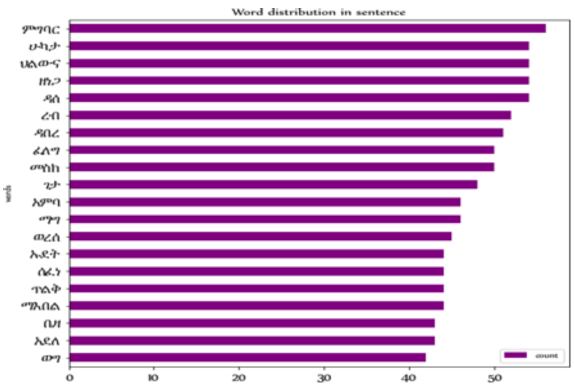

Figure 1: Complex term distribution

6 Dataset Preprocessing

Amharic language is one of the morphologically inflected and have contains different special characters (Mulugeta & Gasser, 2012). So data pre-processing is a critical step towards building an optimized machine learning model for the Amharic language. Cleaning and transforming the raw data to useful features for building transformer-based models to solve the desired problem is our target in this dataset pre-processing because the performance of these models depends upon the quality of data (Kenton et al., 2019; Pandey et al., 2020).

Tokenization splits the sentence into a list of tokens and removes special characters like ',', '።', '!'. Stop-word removal is necessary because in natural language processing applications, an appropriate stop-word extraction technique is required (Kaur, 2018), and the Amharic language text dataset has words such as "to" (ወደ) and "this" (ይህ) that need to be filtered accurately. Normalization for the Amharic phoneme such as /h/ can be represented by the <ሀ>, <ሐ>, <ኀ>, and <ሃ> series of graphemes (Zupon, 2021). To reduce such Fidel variation in Amharic words, we have applied this normalization. Finally, we have used morphological analysis. Amharic is one of the most morphologically complex and inflected languages (Goebel, 2014). Due to this, we have reduced such morphological variation of Amharic tokens to their representative root form by removing affixes.

7 Dataset Preprocessing

To convert the Amharic text dataset to its numeric vector for building the pre-trained classification models we have used both BERT and XLNET embeddings. For building these embeddings, we have used a dataset with 25,143 unique vocabularies. The reason for selecting these embedding techniques rather than previously used ones such as word2vec is that the BERT-based embedding is contextualized embedding and it has higher correlation with the human-annotated word importance scores (Amin et al., 2022). Furthermore, it has the ability to understand a complicated text context. Figure 2 shows the embedding of BERT architecture that we have used for Amharic text.

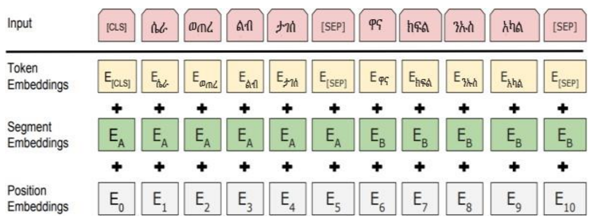

Figure 2: BERT embedding for Amharic text

8 Transformer-based Complexity Classification

Large-scale pre-trained models have recently achieved great success and become a milestone in the field of artificial intelligence (Han et al., 2022). These large-scale pre-trained models can optimally capture knowledge from massive labeled and unlabeled data. The models such as BERT (Kenton et al., 2019), XLNET, and RoBERTa (Pan et al., 2021) are trained on unlabeled text in both left and right contexts of the layer and fine-tune these models for developing pre-trained language models for specific downstream tasks for morphologically-rich and medium-resourced languages (Seker et al., 2022).

In this work, we have used these pre-trained language models for Amharic text complexity classification and simplification by fine-tuning the pre-trained layers of the model by adding a new layer on top of the pre-trained layer. This helps to reduce the problem that comes with data limits, such as underfitting issues (Zhang et al., 2021). The initial studies of fine-tuned encoders have shown state-of-the-art performance on benchmark suites (Merchant et al., 2020). BERT is an auto-encoding language model that can work with bidirectional context (Kenton et al., 2019).

The model has a self-attention mechanism, which is pre-trained on large data corpora, and it is a state-of-the-art model used to address several NLP problems. It allows fine-tuning the base model for a specific task (Koroteev, 2021).

Furthermore, we have used the generalized autoregressive pre-trained model XLNet (Yang et al., 2019), which maximizes the expected likelihood over all permutations of the factorization to overcome the limitations of BERT.

Fine-tuning: Training these transformer-based models from scratch using a small dataset size may not be appropriate to build an optimized model. So, fine-tuning and adding a fully connected layer on top of these pre-trained models achieve state-of-the-art results with minimal task-specific arrangements for a wide variety of tasks (Sun et al., 2019). Due to this, we have fine-tuned the layer of these transformer-based models for Amharic text complexity classification problems.

9 Amharic Text Complexity Classification Experiments

The Amharic text complexity classification models, such as BERT and XLNET, are built using an 80/10/10 dataset distribution. For this experiment, we have used a total of 33.8k Amharic sentences collected from different sources. 25,143 unique vocabularies are extracted from the total dataset size for building pre-trained embedding. Optimal hyperparameter selection and setting help in building a better machine learning model (Panda, 2020), so that the maximum input length, activation function, dense layers, optimizer, and other related parameters are arranged and selected to train these models.

When we fine-tune the layers of the model, we have added two hidden layers and one output layer on top of the base pre-trained model. In the first dense layer, we used 64 fully connected neurons, and in the second hidden layer, we used 32 neurons with a dropout rate of 0.2. The RELU activation function is applied on these hidden layers of the model. The last output layer has two neurons for binary classification (complex and non-complex text). The sigmoid activation function is used in this output layer because it is a common activation function for binary classification problems.

The first experiment is conducted using the BERT model, and it scores a classification accuracy of 86.1%. The second experiment is on the XLNET model using similar parameter configurations to the BERT model. Furthermore, the model is trained through hybrid features with BERT. When we evaluate the experimental result of this XLNET model, it scores a classification accuracy of 70%.

10 Simple Text Generation Experiment

Our final goal is to detect the specific complex term from the sentence, generate a simpler equivalent of the complex term, and reformulate the sentence. For the detection model, we have built Word2Vec using 1,002 Amharic complex terms collected from academic books and related sources (Belete et al., 2015).

To build the simple text generation model, we have conducted three experiments on unsupervised embedding models and transformer-based models to compare the results and select the optimal one. The first experiment is conducted on the Word2Vec model (Mikolov et al., 2013a), which considers a single word per context by predicting one target word on given contextual words (Rong, 2016). The second model is fastText, which is based on continuous skip-grams; each word is represented as the character of n-grams. The last model we have used in this experiment is the RoBERTa model, which tries to handle the context by randomly masking 15% of the sentence during training. The hyperparameter configuration of these simplification models is presented in Table 2.

Word2Vec and Fastext		**Roberta**	
Parameters	Size	Parameters	Size
window	5	Epoch	30
Mini_count	1	max_position_embeddings	514
Epoch	25	num_attention_heads	12
		num_hidden_layers	6

Table 2: Simplification models hyperparameter

As we evaluated the individual models' accuracy in predicting the simplest equivalent of complex terms in a sentence, the Word2Vec model demonstrated better simplification generation accuracy compared to the other two models. This is because it learns contextual words to predict a target word using the CBOW (Continuous Bag of Words) architecture, where the distributed representations of context words are combined to predict the target word by calculating the cosine similarity between word vectors. In contrast, fastText represents a single word as a combination of sub-character n-grams, which may reduce accuracy for certain terms.

For example, in the sentence "የነዳጅ ክምችት በስፋት በመካከለኛው ምሥራቅ ይገኛል" ("Oil reserves are widely found in the Middle East"), the Word2Vec model represents the context as የነዳጅ,ከምችት,በመካከለኛው,ምሥራቅየነዳጅ, ከምችት, በመካከለኛው, ምሥራቅየነዳጅ,ከምችት,በመካከለኛው,ምሥራቅ to predict the target word በሰፊት ("widely"). Meanwhile, the fastText model decomposes the word በመካከለኛው into sub-character n-grams such as በመካከ,መካከለ,ካከለኛ,ከለኛውበመካከ, መካከለ, ካከለኛ, ከለኛውበመካከ,መካከለ,ካከለኛ,ከለኛው, and the RoBERTa model learns the context by randomly masking 15% of the sentence during training to predict masked words.

11 Experimental Result

In this section, we have discussed the experimental result of the complexity classification, detection, and simplification models of Amharic text. For the classification of Amharic text to its target (complex or noncomplex), we have trained transformer-based pre-trained models, namely BERT and XLNET. A total of 33.8k sentences are used to train, validate, and test the model.

Then, based on such dataset distribution, we have conducted two separate experiments. The first experiment is on the BERT model, and it scores a classification accuracy of 86.1%, while the second XLNET-based experiment scores 70% classification accuracy.

As the result shows in Table 3 Row 2, the BERT model has better classification accuracy than the XLNET model because the maximum length of the sentence used in the experiment does not exceed the maximum length that the BERT model can handle (Ding et al., 2020). The reason for the reduced length of sentences in the dataset that we have used is that it passes through different preprocessing stages and some unwanted tokens are reduced.

Furthermore, BERT is easily trainable with a limited-size dataset for specific tasks and addresses long-term information dependence (Jang et al., 2020). See the detailed experimental result analysis of these two classification models in Table 3.

Model	**Precision**	**Recall**	**Test Accuracy**
BERT	86%	86%	86.1%
XLNET	72%	70%	69.9%
BERT+ XLNET	73%	69%	69%

Table 3: Classification models experimental result.

To detect specific complex terms from the sentence classified as complex by transformer-based models, we have built Word2Vec embedding using 1002 terms. Finally, a simple text generation model is built on Word2Vec, Fastext, and Roberta using 91k Amharic sentences. The simple equivalent of complex terms is collected from Amharic dictionaries organized by Aleka Kidanewold Kflie in 1948. As we evaluated the prediction result of these models based on cosine similarity, the Word2Vec (CBOW) model has more accurate prediction than the Fastext and Roberta models, it predicts simple text up to cosine similarity of 0.91 while the Fastext and Roberta models score up to 0.86 and 0.54 respectively. The predicted simple equivalent terms for the detected complex term based on its cosine similarity is visualized in Figure 3 and 4. The reason for the RoBERTa model has less accurate prediction is that it is not trained well due to resource limit and the masked words that we have used are less replicative words on the training document, which is masked so very few times in the taring time of the RoBERTa.

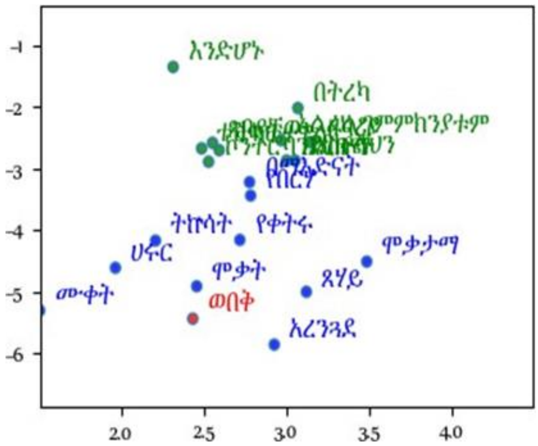

Figure 3: Word2Vec simple term prediction

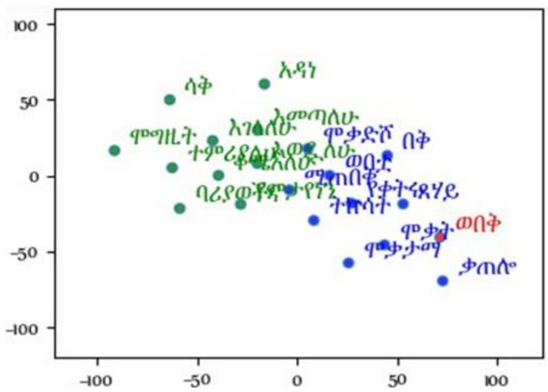

Figure 4: Fastext simple term prediction

12 Discussion

The Amharic complexity classification model's experimental result is analyzed based on precision, recall, and accuracy. For this experiment, we have used 33.9k sentences collected from different sources and the pre-trained transformer-based models namely BERT and XLNET are trained using 80/10/10 dataset split. To victories the dataset we have used BERT and XLNET embedding layers using 25143 unique vocabularies. The classification performance of these models is evaluated using 3390 test dataset. Based on the confusion matrix result analysis the BERT model predicts 2919 sentences correctly (1417 complex and 1502 noncomplex) the rest 471 sentences are predicted falsely by the model. While the XLNET model predicts 2370 sentences correctly from the total test data (942 complex and 1428 noncomplex). As the accuracy result of these transformer-based models shows the BERT has better classification performance which scores an accuracy of 86.1%, The model can be easily fine-tuned for small datasets and it considers long-term information dependence (Rong, 2016), while XLNET scores an accuracy of 70% for such Amharic text complexity classification problem.

The next experiment we have conducted in this work is to detect the specific complex term from the sentence classified as complex by these transformer-based models. We have trained this detection Word2Vec model using 1002 complex terms. Finally, for simple text generation, the Word2Vec, Fastext, and Roberta models are used for training. To build these models we have used a total of 91k Amharic sentences (complex text with their simpler equivalent). The result comparison of the models shows that Word2Vec predicts more accurate simple text for the detected complex term than the other models. Based on the sample test data this model predicts cosine similarity of 0.91, 0.63, 0.62, 0.59, and 0.59 for five ranked simpler equivalent texts.

13 Conclusion

In this study, we have developed a transformer-based complexity classification model for Amharic text. Furthermore, we have built two sequential models for specific complex term detection and simple text generation processes. We are motivated to work on such Amharic text complexity because there are numerous Amharic terms identified by authors that are not frequent and unfamiliar to low-literacy readers. To address such complexity issues in Amharic texts previously teachers and scholars used dictionaries to find their meaning and elaborate them for the readers. Recently due to the emerging of machine learning models basically the pre-trained models, build transformer-based complexity classification and simplification model helps to address the issue accurately than the previous methods, because these models can work well for sentence and document level detection and simplification processes.

The mining for complex terms can also be handled dynamically using such machine learning models. For this work total of 33.9k sentence for BERT and XLNET classification models, 1002 complex terms for the Word2Vec detection model, and 91k sentences for simplification Word2Vec, Fastext, and Roberta models is used. The classification and simplification performance of the models is evaluated based on precision, recall, accuracy, and cosine similarity. For the classification task BERT model scores better accuracy (86.1%) and for the simple text generation Word2Vec scores better accuracy (0.91) than other models. Syntactic and morphological complexity of the Amharic text are the other types of complexity that need to be studied in the future.

14 Limitations

The primary limitations of this work unavailability of large annotated data for Amharic, which hinder the model's ability to learn complexity patterns across various types of text. This constraint could impact the generalizability of the model. The other major limitation is the computational cost of training large models such as BERT which limiting the model we have fine-tuned for optimal performance as we have used free versions colab for training. Furthermore, for this study we have used educational and limited number of complex terms which will be improved for by considering more complex terms in future studies.

References

Speech, L., Dahl, A., Carlson, S., Renken, M. D., & Mccarthy, K. S. *2021. Materials Matter: An Exploration of Text Complexity and Its Effects on Middle School Readers' Comprehension Processing.* Language, Speech and Hearing Service in School, Page 1-9.

Ide, Y., Mita, M., Nohejl, A., Ouchi, H., & Watanabe, T. *2023. Japanese Lexical Complexity for Non-Native Readers: A New Dataset.* Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 477–487.

Nation, K., & Snowling, M. J. *2000. Factors influencing syntactic awareness skills in normal readers and poor comprehenders.* Applied Psycholinguistics, 21 (2000):229–241.

Pan, C., Song, B., Wang, S., & Luo, Z. *2021. DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach.* Proceedings of the 15th International Workshop on Semantic Evaluation, Bangkok, Thailand, pages 578–584.

Read, M. *2019. Reading for Ethiopia's Achievement Developed Monitoring.* USAID. 1-20.

Prof, A., & Akba, S. *2016. The Effect of Reading Comprehension on the Performance in Science and Mathematics.* Journal of Education and Practice, 7(16):108–121.

Arya, D. J., Hiebert, E. H., & Pearson, P. D. *2011. The effects of syntactic and lexical complexity on the comprehension of elementary science texts.* International Electronic Journal of Elementary Education, 4(1), 107–125.

Mulwa, E. C. *2015. Difficulties Encountered by Students in the Learning and Usage of Mathematical Terminology: A Critical Literature Review.* Journal of Education and Practice, 6(13):27–38.

Snow, C. E. *2010. Academic Language and the Challenge of Reading for Learning About Science.* Science. American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC, 328 ,450.

Shirzadi, S. *2014. Syntactic and lexical simplification: The impact on EFL listening comprehension at low and high language proficiency levels.* Journal of Language Teaching and Research, Volume 5, pages 566–571.

Kelemework, W. *2013. Automatic Amharic text news classification: A neural networks approach.* Ethiop. J. Sci. & Technol. 6(2) 127-137.

Sulem, E., Abend, O., & Rappoport, A. *2018. Semantic structural evaluation for text simplification.* NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Pages 685–696.

Nigusie, G., & Tegegne, T. *2022. Amharic Text Complexity Classification using Supervised Machine Learning.* 10th EAI International conference on Advancement of Science and Technology Bahir Dar, Ethiopia, Page 12-23

Solution, P. *2021. Guidelines for Minimizing the Complexity of Text Prepared by the Center for Literacy & Disability Studies Department of Allied Health Sciences.* School of Medicine University of North Carolina at Chapel Hill, Pages 1–9.

Lo Bosco, G., Pilato, G., & Schicchi, D. *2018. A Neural Network model for the Evaluation of Text Complexity in Italian Language: A Representation Point of View.* Procedia Computer Science. Volume 145. Pages 464–470. https://doi.org/10.1016/j.procs.2018.11.108.

Shardlow, M., Evans, R., & Zampieri, M. *2022. Predicting lexical complexity in English texts: the Complex.* In Language Resources and Evaluation (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s10579-022-09588-2

Bott, S., Rello, L., Drndarevic, B., & Saggion, H. *2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish ¿ Puede ser el Español más simple?* LexSiS: Simplificación Léxica en Español. Proceedings of COLING 2012: Technical Papers, Mumbai, Pages 357–374.

Sulem, E., Abend, O., & Rappoport, A. *2018. Semantic Structural Evaluation for Text Simplification.* Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, Pages 685–696.

Gilakjani, A. P., & Sabouri, N. *2018. Learners Listening Comprehension Difficulties in English Language Learning.* A Literature Review, 9(6).

Qiang, J., & Wu, X. *2019. Unsupervised Statistical Text Simplification.* IEEE Transactions on Knowledge and Data Engineering, Page 1.

Gasparetto, A., Marcuzzo, M., & Zangari, A. *2022. A Survey on Text Classification Algorithms: From Text to Predictions.* Information 2022. 13(83):1–39.

Review, S. *2021. Levels of Reading Comprehension in Higher Education: Systematic Review and Meta-Analysis.* Systematic Review and Meta-Analysis, Volume 12.

Young, D. N. *1999. Linguistic Simplification of SL Reading Material.* Modern Language Journal, Volume 83, Pages 350–366.

Hu, H.-C., & Nation, P. *2000. Unknown Vocabulary Density and Reading Comprehension.* Reading in a Foreign Language, Volume 13, Pages 403–30.

Santucciz, V., Santarelli, F., Forti, L., & Spina, S. *2020. Applied Sciences Automatic Classification of Text Complexity.* Applied Sciences, 10(20):1-19.

Hidayat, M. F. *2019. Using K-Means Clustering and Multinomial Naive Bayes.* 2019 International Seminar on Application for Technology of Information and Communication (ISemantic), Pages 163–170.

Gasparetto, A., Marcuzzo, M., & Zangari, A. *2022. A Survey on Text Classification Algorithms: From Text to Predictions.* Information. 13(83):1–39.

Kenton, M. C., Kristina, L., & Devlin, J. *2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv:1810.04805v2 [cs.CL], Pages 1-16.

Pan, C., Song, B., Wang, S., & Luo, Z. *2021. DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach.* Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1–16.

Sen, Y., & Fuping, Y. *2021. Chinese Automatic Text Simplification Based on Unsupervised Learning.* 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Volume 2021, Pages 1-8028.

Belete, Z., Mlkt, Z., Bezabh, E., & Chekol, T. *2015. Amharic Teacher Guide Grade-7.* FDRE Minister of Education and ABKME Education Bureau, Pages 1-247.

Alemu, D., Aklilu, S., & Mengstie, Y. *2015. Amharic Teacher Guide Grade-9.* FDRE Minister of Education. Pages 1-185.

Gambäck, B., Olsson, F., Argaw, A. A., & Asker, L. *2009. Methods for Amharic Part-of-Speech Tagging.* Ethiopian Parliament Projections in December 9 2008 Based on the Preliminary Reports from the Census of May 2007, Volume 104.

Mulugeta, W., & Gasser, M. *2012. Learning Morphological Rules for Amharic Verbs Using Inductive Logic Programming.* Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012).

Pandey, N., Patnaik, P. K., & Gupta, S. *2020. Data Pre-Processing for Machine Learning Models using Python Libraries.* Volume 4, Pages 1995–1999. https://doi.org/10.35940/ijeat.D9057.049420

Kaur, J. *2018. Stopwords Removal and Its Algorithms Based on Different Methods.* International Journal of Advanced Research in Computer Science,9(5), Page 81–88.

Zupon, A. *2021. Text Normalization for Low-Resource Languages of Africa.* arXiv:2103.15845v1 [cs.CL], Page 1-10.

Goebel, R. *2014. Advances in Natural Language.* 9th International Conference on NLP, PolTAL 2014 Warsa, Poland, 2014 Proceedings.

Amin, A. Al, Hassan, S., Alm, C. O., Huenerfauth, M. *2019. Developing an Amharic Text-to-Speech System Using Machine Learning Approaches.* Applied Sciences 2019, Volume 9, Pages 14–35.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. *2022. Pre-trained models: Past, present and future.* AI Open, Volume 2, Pages 225–250.

Seker, A., Bandel, E., Bareket, D., & Brusilovsky, I. *2022. AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level.* Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1, Pages 46–56.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. *2021. Revisiting Few-Sample Bert Fine-Tuning.* Pages 1–22.

Yang, Z., Dai, Z., Yang, Y., & Carbonell, J. *2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding.* NeurIPS, Pages 1–18.

Sun, C., Qiu, X., Xu, Y., & Huang, X. *2019. How to Fine-Tune BERT for Text Classification?* Lecture Notes in Computer Science, 11856 LNAI(May), Pages 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Panda, B. *2019. A survey on application of Population Based Algorithm on Hyperparameter Selection.* Advanced Topics in Artificial Intelligence, Pages 1–9.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. *2013. Efficient estimation of word representations in vector space.* 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, Pages 1–12.

Rong, X. *2016. word2vec Parameter Learning Explained.* arXiv:1411.2738v4 [cs.CL], Pages 1–21.

Ding, M., Zhou, C., Yang, H., & Tang, J. *2020. [2020-NeurIPS] CogLTX: Applying BERT to Long Text.* 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, NeurIPS.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. *2020. Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism.* Applied Sciences, Switzerland, Volume 10(17), . https://doi.org/10.3390/app10175841