Developing a machine learning-based grade level classifier for Filipino children's literature

Joseph Marvin Imperial[1], Rachel Edita Roxas[2], Erica Mae Campos, Jemelee Oandasan, Reyniel Caraballo, Ferry Winsley Sabdani
College of Computing and Information Technologies
National University
Manila, Philippines
{jrimperial[1], reoroxas[2]}@national-u.edu.ph

Ani Rosa Almario
Adarna House, Inc.
Quezon City, Philippines
ani@adarna.com.ph

***Abstract*—Reading is an essential part of children's learning. Identifying the proper readability level of reading materials will ensure effective comprehension. We present our efforts to develop a baseline model for automatically identifying the readability of children's and young adult's books written in Filipino using machine learning algorithms. For this study, we processed 258 picture books published by Adarna House Inc. In contrast to old readability formulas relying on static attributes like number of words, sentences, syllables, etc., other textual features were explored. Count vectors, Term Frequency–Inverse Document Frequency (TF-IDF), n-grams, and character-level n-grams were extracted to train models using three major machine learning algorithms—Multinomial Naïve-Bayes, Random Forest, and K-Nearest Neighbors. A combination of K-Nearest Neighbors and Random Forest via voting-based classification mechanism resulted with the best performing model with a high average training accuracy and validation accuracy of 0.822 and 0.74 respectively. Analysis of the top 10 most useful features for each algorithm show that they share common similarity in identifying readability levels—the use of Filipino stop words. Performance of other classifiers and features were also explored.**

Keywords—Filipino, storybook, readability, machine learning, classification

I. INTRODUCTION

"*Pagbasa…pag-asa.*" *(In reading, there is hope.)*

– Reading Association of the Philippines

Readability is the difficulty or complexity of a published work [1, 2, 3, 4]. Over the years, researchers have studied and used different techniques to measure readability. Determining the readability of texts ensures that a certain book matches the reading ability of the target reader.

At present, over a hundred readability formulas have been developed for the English language. Traditional readability formulas use vocabulary and sentence difficulty as their basis in calculating text difficulty yet those proved to misleading since other factors should be considered too [5, 6, 7]. Researchers are now trying out new approaches by applying different advancements in the field of Natural Language Processing and Computational Linguistics. Most of the existing readability measurements are used solely for the English language and when used in other languages, results yielded were invalid as shown by Kumiko, Satoshi and Hiroshi [8].

There is still no known and approved scale for measuring the readability of Filipino texts. Since the Filipino language has a structure different from English, different characteristics should be considered and explored in determining how to measure its readability.

In this paper, the researchers developed a baseline prototype model for leveraging the readability of Filipino texts from an extensive experimentation of extracting various text-level features of picture book data such as word frequencies, character and word grams, and TF-IDF scores using machine learning algorithms. This study would benefit authors, publishers, educators, and readers as it aims to ensure that a certain reading material is appropriate for the intended users.

II. REVIEW OF RELATED LITERATURE

Text Readability Indexing (TRI) is the process where texts are evaluated and categorized into grade levels or scores based on their difficulty [2]. Readability assessment has been widely used for many purposes, primarily to determine how difficult a passage is to be read [9]. To be able to compute the readability of a certain text makes it easier for authors and publishers to ensure that they would be able to deliver ideas effectively to their target audience [5, 3, 10, 11]. Along with this, the reading ability of the reader should also be kept in mind to ensure effective measurement of text complexity [12]

It is important for the complexity of the learning material to match the reading ability of the student [1,7]. When a student is given text with a difficulty level that does not match his reading level, the student may struggle in understanding and processing the content of the given text. But if his readability level matches that of the text, he can focus more and would probably perform better in school [10].

Without an accurate and approved system of measuring readability, some texts may not be applicable to students from a certain grade level. Because of this, teachers and educators learned to improvise for the benefit of their students [1]. Readability indexing systems should consider the learners' reading abilities to ensure accuracy when used in academic reading and instructional materials [2]. Effective readability assessments would help educators prepare useful and proper reading materials for learners [7].

A. Factors Affecting Readability

According to Kumiko, Satoshi and Hiroshi [8], different features and indicators affect readability. Moreover, it is dependent on two factors: the reader (language capability, background knowledge of subject matter, and motivation) and the text (syntactic and semantic rules) as shown by Al-Tamimi et. al [3]. In semantic complexity, word familiarity, word difficulty, number of syllables per word, and number of words per sentence is a great contribution to existing

978-1-7281-5014-7/19/$31.00

formulas. Certain text variables such as format, typography, content, literary style, vocabulary difficulty, sentence complexity, concept density, and cohesiveness all play a role as well [10, 12].

Text-related factors (e.g. word length, word frequency, vocabulary load, number of difficult words, average sentence length, sentence complexity, clarity of idea, use of topology or metaphors, and grammatical structure complexity) are considered in the calculation of readability [3]. According to Beinborn et. al [7], the complexity of a sentence's syntactic structure affects its difficulty. In conjunction with this, a sentence is said to be more readable when common words and terms are used in its construction.

While certain studies state that the longer the sentence and word, the more complex the text is, it can also be the other way around [4]. Graesser et al. [13] supported the above statement and said that texts with longer words and sentences are more difficult to read than shorter ones.

B. *Traditional vs. Modern Approaches in Readability Identification*

Despite the popularity of traditional readability formulas, these have received criticism on how they were unable to consider other factors such as text-based processing, situation levels, and sentence cohesion [5]. Their reliability and validity have been questioned since they only focused on structural characteristics and / or surface-based factors which can be misleading [4,5,6,7]. Features used by traditional readability formulas do not really analyze text at a deeper level, but have been popularly used.

Recent studies regarding readability greatly progressed from using traditional or superficial language properties to taking advantage of natural language processing tools to further analyze texts. The modern age gave way to improvements in the calculation of readability of texts. Newly developed readability formulas were proven to perform better [14]. Algorithms, machine learning techniques, frameworks, language modeling, and more recently developed technologies are creeping its way in computational linguistics and NLP.

Cognitive models are starting to be used in many researches as an approach to assessing texts [4]. Machine learning approaches in determining text readability allow for broader features to be considered [15]. Data training may lead machines to predict the complexity of unseen texts [16].

A study by Si and Callan [17] made use of the Expectation Maximization (EM) algorithm to compute for the weight value of their proposed models. The research produced two models: the unigram language model, which made use of words in text, and the sentence length distribution model. Based on the experiments performed, sentence length proved to be a useful predictor for readability since its mean values increases alongside the text's readability level while syllable count doesn't exhibit the same performance.

Advances in Natural Language Processing, Machine Learning, and language psychology delivered the field of research a huge range of linguistic features; the ability to process large amounts of information; and awareness regarding deeper-lying text features that may have a great influence on the readability of a certain text [4].

C. *On Readability of Filipino Texts*

Readability formulas used for assessing texts in English proved to be unsuccessful in determining the readability of texts written in other languages. This is due to the different linguistic characteristics across the languages [1]. There is a high possibility that using a readability formula for a language would yield invalid and disappointing results when used for another language. There are other factors and elements to be considered that aren't specified in existing formulae [10].

The Philippines is a multilingual country that uses both English and Filipino as official languages. Both languages are formally taught in schools and used in different instructional materials and in literature. There are countries that have developed their own language's formula in assessing the readability of texts written in it. In the present time, there is no known scale that is extensively used for evaluating the difficulty of Filipino texts [1]. Since there is no formula that is recognized to determine the readability of Filipino texts yet, books and reading materials written in the language are categorized via subjective judgment [16]. There are Filipino researchers who have attempted to develop their own readability formulas [1].

In a research by Gutierrez [10], she used Fry and SMOG readability formulae in processing English and Filipino passages. The dataset was inputted into an online readability calculator with the application of the said formulas. The testing yielded invalid results during the Fry formula testing. Invalid results weren't shown after the SMOG formula testing. There may be a possibility that SMOG could be a suitable formula for Philippine English if certain adjustments are considered by Filipino authors. In the testing of texts written in Filipino, no valid result was produced by the two formulas which implies the need for a readability formula intended for Filipino texts.

III. Methodology

Described in this section is the step-by-step process of the study.

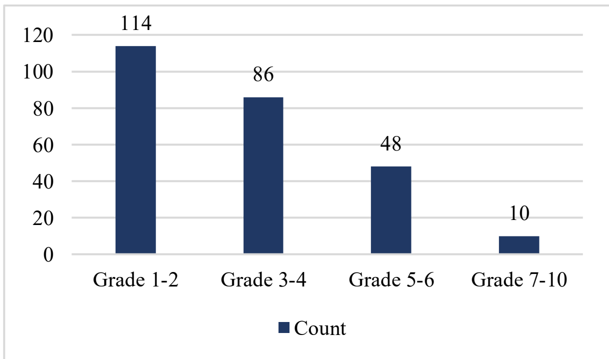

Fig. 1. Grouping categories per grade level

A. Data Collection and Annotation

The researchers collected a total of 258 pre-labelled picture book data from Adarna House Inc. from grade levels 1 to 10. The categorization of picture books according to reading age (category) was done by reading experts. The data was then partitioned to 80:20 ratio for training and testing respectively. All picture book data were written in Filipino. Due to the imbalanced dataset gathered, the researchers decided to group the classes to reduce the complexity of the data and to improve generalization. The clustered picture book grade levels are shown in Figure 2. Therefore, the final label for each test and validation data will be one of the four grouped classes.

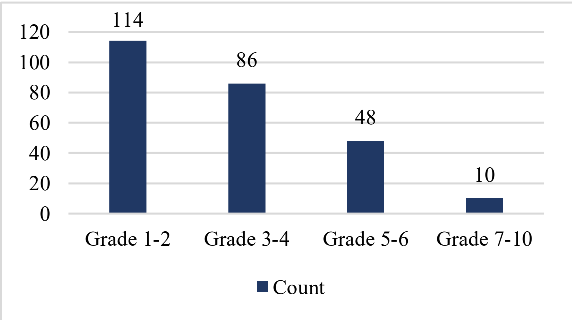

Fig. 2. Grouped picture book grade levels

In observance with the machine learning-based research methodology, the training set for each classifier was processed using select machine learning algorithms discussed below. On the other hand, the testing set was used to leverage the learning of the algorithms after training. Accuracy of the models was computed for training and testing. Lastly, the validation set will identify the reliability of the models in predicting unseen data.

B. Preprocessing

Before directly processing the raw data for the machine learning algorithms, several cleaning processes, including removal of numeric characters, symbols, extra spaces were performed to structure the data into uniform format and remove noise which may negatively affect the overall performance of the model.

C. Feature Extraction

Feature extraction is an essential step in machine learning-based research. It refers to the extraction of linguistic and textual elements from a given corpus to provide a representative, reduced sample of the content. Distinctive vocabulary "features" found in a document are assigned to the different categories by measuring the importance of those elements to the document content [26].

Count Vectors - a type of feature which converts text data into a matrix of token counts. It is a vocabulary construction feature where for each document in the dataset represented per row and each word in the document represented per column, the number of occurrences or frequency count of words are listed. The total size of the matrix or dictionary is the product of the number of documents and the number of unique words present in all documents.

Term Frequency – Inverse Document Frequency - TF-IDF is a common feature used in text analysis. It shows how important a word is to a document in a collection or corpus [21]. The TF-IDF value increases to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, this caters to frequently occurring words by giving them lower weights than rare words.

Word Trigrams - Trigram are three slices of words or characters. For example, the trigram or three-word slice of the Filipino sentence "*Maganda ang bulalak sa hardin*" (The flowers in the garden are beautiful) would be "*Maganda ang bulaklak*," "*ang bulaklak sa*," "*bulalak sa hardin*." For this study, the trigram TF-IDF feature represents the frequency of trigrams occurring in the document multiplied by the logarithm of the total number of documents divided by the number of trigrams appearing in it.

Character Trigrams - On the other hand, character trigrams are n-slices of characters of words. For example, the character trigram slice of the Filipino word "*mabuhay*" (greetings) would be "*_ma,*" "*mab,*" "*abu,*" "*buh,*" "*uha,*" "*hay,*" "*ay_*." This feature represents the frequency of character trigrams occurring in the document multiplied by the logarithm of the total number of documents divided by the number of character trigrams appearing in it.

The proponents noted the use of trigrams instead of bigrams or unigrams because trigrams were often used in building language models of Filipino as shown by Oco, Syliongka, and Roxas [25].

D. Training

For this study, common machine learning algorithms commonly used for document classification and sparse datasets were experimented to see how they perform in identifying readability of Filipino texts. The algorithms are described below.

Multinomial Naïve-Bayes

The Naive Bayes classifier is based from the Bayes theorem. It has a 'naive' assumption of that a feature in a class is unrelated to any other feature present [22]. Given a sample story text file to be classified, represented by a vector $x = (x_1, \ldots x_n)$ representing some n features (independent variables), it assigns to this instance probabilities $p(C_k|x_1, \ldots x_n)$ for each of K possible outcomes in the set of classes C_K determined. In the present study, the researchers used Multinomial Naive Bayes in training the converted data vectors (both count vectors and TF-IDF vectors). Multinomial Naïve Bayes implements the Naive Bayes algorithm for multinomially distributed data [24]. The formula for the multinomial Naïve Bayes is shown below:

$$C_{NB} = argmax\,(\log P(C_k) + \sum_{i=1}^{n} x_i\, logp_{ki})$$

where p_{ki} is the probability of a story belonging to class K and x_i is the features of the story.

K-Nearest Neighbors

For K-Nearest Neighbors or KNN, the textual features are plotted on a plane and its K nearest neighbors are searched using a given distance formula, in this study the Euclidean distance was used, where K is the number of features nearest to the feature being identified. The formula for getting the distance between two feature points is shown below. The researchers set the value of k to 5 for each feature word.

$$D(u,v)^2 = \|u = v\|^2 = (u-v)^T(u-v) = \sum_{i=1}^{d}(u_i - v_i)^2$$

where u and v are sample real-valued features of a story being compared in terms of distance.

Random Forest

A Random Forest classifier creates multiple decision trees from a random subset of the training data. The classifier then gets the average of all decision trees generated to improve accuracy and control overfitting [24], this term is also called 'bagging'. The researchers set the n-estimators (number of trees in the forest) to 100 and the random state (seed used by the random number generator) to 0. After training, predictions for an unseen story $\acute{x}$ can be made by averaging the predictions from all the individual decision trees on the said story or by taking the 'majority' vote in the case of classification trees. The mathematical equation for this process is shown below.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\,(\acute{x})$$

where B is the number of bagging iterations during training and f_b is the current decision tree being trained.

IV. Results and Discussion

The proponents conducted an extensive experimentation by pairing each feature set extracted (count vectors, TF-IDF, trigram, and character trigram) to the select machine learning algorithms (Multinomial Naïve Bayes, KNN, Random Forest) for document classification. The following tables, Table I, II, and III, describes the performance of each algorithm with varying features on identifying readability of story book test data. Cross fold validation during was performed with five instances (k = 5) to avoid bias and overfitting. Insights from the analysis of the experiments are described below.

For the experiment with Multinomial Naïve-Bayes, the best model came from using trigrams as features with an accuracy of 0.61. From this result, the researchers can infer that there are common trigram combinations in the Filipino language used for constructing storybooks that are prominent enough for the Naïve-Bayes algorithm to recognize what readability class they belong.

For K-Nearest Neighbors, the best model is using the count vectors as features. The researchers noted the stable value of accuracy in the training folds, with one instance reaching 0.857. The researchers note that the occurrence of words for each story is enough for the KNN algorithm to categorize which readability label a test storybook belongs to just by getting the first five (set value) nearest features (neighbors) for each word of the test story book.

On the other hand, for Random Forest, the best model achieved a high accuracy of 0.942 using count vectors as features, same with KNN described above. The Random Forest algorithm outperformed both Multinomial Naïve-Bayes and KNN in terms of performance during training. The researchers expected this result, as Random Forest is an ensemble and decision-tree based algorithm which effectively reduces biases, control overfitting, and improves classification performance by aggregating the results from a collection of decision trees. This shows that the frequency of words for each storybook is enough for Random Forest to differentiate which grade level a storybook belongs to.

TABLE I. Result of Classification with Multinomial Naïve Bayes (MNB) and Various Features

Algorithm & Feature	Training Folds					Average
	K=1	K=2	K=3	K=4	K=5	
MNB + Count Vectors	0.342	0.421	0.297	0.314	0.2	**0.31**
MNB + TF-IDF	0.5	0.368	0.514	0.514	0.486	**0.48**
MNB + Trigrams	0.605	0.474	0.622	0.629	0.657	**0.6**
MNB + Char Trigrams	0.421	0.421	0.432	0.429	0.429	**0.43**

TABLE II. Result of Classification with KNN and Various Features

Algorithm & Feature	Training Folds					Average
	K=1	K=2	K=3	K=4	K=5	
KNN + Count Vectors	0.857	0.828	0.742	0.742	0.7428	**0.7828**
KNN + TF-IDF	0.371	0.6	0.4	0.314	0.3714	**0.4114**
KNN + Trigrams	0.457	0.628	0.45	0.371	0.514	**0.485**
KNN + Char Trigrams	0.342	0.485	0.314	0.285	0.314	**0.348**

TABLE III. Result of Classification with Random Forest and Various Features

Algorithm & Feature	Training Folds					Average
	K=1	K=2	K=3	K=4	K=5	
RF + Count Vectors	0.942	0.857	0.685	0.885	0.685	**0.811**
RF + TF-IDF	0.742	0.742	0.771	0.8	0.54	**0.72**
RF + Trigrams	0.742	0.742	0.657	0.828	0.6	**0.714**
RF + Char Trigrams	0.857	0.74	0.685	0.685	0.714	**0.737**

With count vectors as a prominent feature for KNN and Random Forest, the researchers can infer from this result that the occurrence of words and its frequency is a great factor in determining readability of books. For instance, inflected and lengthy Filipino words like "nanggagalaiti" (furious), "sumusubaybay" (tracking or following), "nagtutulog-tulugan" (pretending to be asleep) were often found on higher grade levels (6-10) than lower grade levels. In addition, from using count vectors as feature, the researcher observe that higher grade levels have higher word count than storybooks from lower grade levels. All these patterns are taken in consideration by KNN and Random Forest, thus achieving a high accuracy score during training.

A. Applying a Voting Mechanism for Classification Enhancement

In the hopes of achieving a higher accuracy, the researchers implemented an additional 'voting' feature to combine the classification effectivity of the two highest accuracy models utilizing the KNN and RF algorithm. A voting classifier model combines multiple different models into a single model, which is (ideally) stronger than any of the individual models alone as described in [24]. For soft voting, we predict the class labels based on the predicted probabilities p for a classifier. The mathematical formula for soft voting mechanism is shown below:

$$\hat{y} = \arg\max \sum_{j=1}^{m} w_j \, p_{ij}$$

where w_j is the weight that can be assigned to the jth classifier. Meanwhile in hard voting, we predict the class label $\hat{y}$ via majority (plurality) voting of each classifier C_j [24]. The mathematical formula for hard voting mechanism is shown below:

$$\hat{y} = \arg\max \sum_{j=1}^{m} w_j \, X_A(C_j(x) = i)$$

where X_A is the characteristic function $\left[C_j(x) = i \in A\right]$ and A is the set of unique class or grade levels.

TABLE IV. RESULT OF CLASSIFICATION OF KNN AND RF WITH VOTING MECHANISM

Algorithm & Feature	Training Folds					Average
	K=1	K=2	K=3	K=4	K=5	
Soft Voting (KNN + RF with Count Vectors)	0.91	0.86	0.77	0.86	0.71	**0.822**
Hard Voting (KNN + RF with Count Vectors)	0.86	0.77	0.742	0.86	0.69	**0.628**

By using a soft-voting classifier combining the power of KNN, RF and count vectors, a high average accuracy of 0.822 was achieved. The proponents note a high score of 0.914 at the first random fold. It is also noted that incorporating a soft-voting classifier increased the current single-algorithm model accuracy by small 0.11 fraction. Using a hard-voting classifier produced a mediocre result of 0.628 in accuracy. With this, the proponent selected the model using a voting classifier combining KNN, RF and count vectors as the model to be used for predicting unseen story book data, or the validation set.

B. Validating on Unseen Data

The table below shows how the selected model performed on unseen data not included in the training. The voting-enhanced classifier identified the grade level of 23 books out of 31 from the validation set of 10% from the total storybook data count, equal to a percentage of 0.7419 against its 0.822 average accuracy on the testing data. The proponents deem this result as the highest validation score the model can achieve without overfitting. Overfitting is the result of model memorizing the training data too much instead of generalizing the shared features across categories.

TABLE V. PERFORMANCE ON UNSEEN, VALIDATION DATA

Algorithm & Feature	Books Classified Correctly	Total Number of Books	Percentage
Soft Voting (KNN + RF with Count Vectors)	23	31	**0.7419**

C. Analysis of Top Storybook Feature Words

While achieving a high accuracy score for training and validation is good, it is also essential to analyze the top features used by the algorithms to know how much they contributed in identifying the readability of the data and how it can be improved for future work.

Table VI shows the top performing and most useful in terms of identifying and discerning patterns from different grade levels. For each level, more than 80% of the words are Filipino stop words. Stop words refer to the common, connecting words used in the language. Usually these words are removed before preprocessing since they do not contribute in context according to Rajaraman and Ullman [21]. But for this research, it is discovered that Filipino stop words serve as a very important feature and can be used to identify the readability of Filipino story books.

V. CONCLUSION

The main purpose of research is to create a new method of automatically identifying the readability of Filipino literature, in the hopes of helping children find books suited for their reading ability. With data provided by Adarna House Inc., the researchers were able to explore various classification models using machine learning algorithms (Naïve-Bayes, KNN, and Random Forest) and features (count vectors, TF-IDF, trigrams, and character trigrams) present in the data. From the training, the best performing models were the combination K-Nearest Neighbors and Random Forest enhanced by a voting mechanism and using count vectors feature which obtained an average validation accuracy score of 74%. Analysis of features used show that Filipino stop words contribute greatly in terms of identifying readability. For future work, the researchers recommend the use of a larger data set with each category having the balanced data count to produce better predictions by thoroughly representing every grade level in the semantic space. With a larger set of data, more powerful methods like Deep Learning can be used for automatic feature extraction. Application of the same algorithms and features may be used for other Philippine languages, however, retraining may be needed.

ACKNOWLEDGMENT

We would like to thank Adarna House Inc. for providing the much-needed storybook data which became the backbone of this study. We acknowledge the use of National University - Human Language Technology lab during the whole conduct of the study. And lastly, we are highly indebted to all the people who gave their assistance for this research.

TABLE VI. TOP FEATURE WORDS FOR EACH GRADE LEVEL

Level 1		Level 2		Level 3		Level 4	
isa	0.003266	*nang*	0.003989	*itinatag*	0.003764	*mga*	0.003837
lamang	0.003335	*higit*	0.004027	*lahat*	0.004349	*din*	0.003839
dati	0.003391	*noon*	0.004041	*kay*	0.004356	*isang*	0.003933
dahil	0.003561	*ni*	0.004157	*lamang*	0.004376	*ni*	0.004033
noon	0.003608	*wala*	0.004157	*ni*	0.004414	*ilang*	0.0042
pang	0.003617	*nito*	0.004169	*din*	0.004787	*bayan*	0.004207
lahat	0.003743	*tatay*	0.00461	*ito*	0.004834	*bansa*	0.004385
mga	0.003811	*mula*	0.004731	*panahon*	0.004888	*rin*	0.004466
ito	0.003824	*siya*	0.004871	*hanggang*	0.004931	*may*	0.004485
kung	0.004086	*maynila*	0.005024	*na*	0.005016	*ay*	0.004715

REFERENCES

[1] H.B. Macahilig. 2014. A Content-Based Readability Formula for Filipino Texts. The Normal Lights: Journal on Teacher Education, vol.8, no.1, ICEAP-17 (2014).

[2] A.R. Razon, P.C. Naval,Jr., M.S. Bosque, R.R. Ledesma, and H.M. Loberas. 2015. A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features. Proceedings of Recent Advances in Natural Language Processing (September 2015), 521-528.

[3] A. Al-Tamimi, M. Jaradat, N.Aljarrah, and S. Ghanim. 2014. AARI: Automated Arabic Readability Index. The International Arab Journal of Information Technology, vol. 11, no.4 (July 2014), 370-378.

[4] P.van Oosten, D. Tanghe, and V. Hoste. 2010. Towards an Improved Methodology for Automated Readability Prediction. LREC 2010: Seventh Conference on International Language Resources and Evaluation (2010), 775-782.

[5] D.S. McNamara, M.M. Louwerse, and A.C. Graesser. 2005. Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension. (2005).

[6] D. Gemoets, G. Rosemblat, T. Tse, and R. Logan. 2004. Assessing Readability of Consumer Health Information: An Exploratory Study. MEDINFO 2004 (2004), 869 – 873.

[7] L. Beinborn, T. Zesch, and I. Gurevych. 2012. Towards fine-grained Readability Measures for Self-Directed Language Learning. Proceedings of the SLTC 2012 Workshop on NLP for CALL 80 (2012), 11 – 19.

[8] K. Tanaka-Ishii, S. Tezuka, and H. Terada. 2010. Sorting texts by Readability. Association for Computational Linguistics, vol. 36, no. 2 (2010), 203 – 228.

[9] S. Tonelli, K.T. Manh, and E. Pianta. 2012. Making Readability Indices Readable. NAACL-HLT 2012, Workshop on Predicting and Improving Text Readability for Target Reader Populations (2012), 40 – 48.

[10] M. M. Gutierrez. 2014. The Suitability of the Fry and SMOG Readability Formulae in Determining the Readability of Filipino Texts. The Normal Lights, vol. 8, no. 1 (2014).

[11] J. Hancke, S. Vajjala, and D. Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. Proceedings of COLING 2012: Technical Papers (2012), 1063 – 1080.

[12] E. Pitler and A. Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (2008), 186 – 195.

[13] A.C. Graesser, D.S. McNamara, M.M. Louwerse, and Z. Cai. 2004. Coh Metrix: Analysis of Text Cohesion and Language. Behavior Research Methods, Instruments, and Computers, vol. 36, no. 2 (2004), 193 – 202.

[14] J.O. dela Rosa and C.F. Genuino. 2018. Correlating Writer's Block and ESL Learners' Writing Quality. Indonesian Journal of Applied Linguistics, vol. 7, no. 3 (2018), 604 – 612.

[15] M. Xia, E. Kochmar, and T. Briscoe. 2016. Text Readability Assessment for Second Language Learners. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (2016), 12-22.

[16] A.R. Razon, P.C. Naval,Jr., M.S. Bosque, R.R. Ledesma, and H.M. Loberas. 2011. A New Approach to Automated Text Readability Indexing using Concept Indexing with K-Means Algorithm. The 11th Philippine Computing Science Congress (March 2011).

[17] L. Si and J. Callan. A Statistical Model for Scientific Readability. Proceedings of the 2001 ACM CIKM, 10th International Conference on Information and Knowledge Management. 2001.

[18] M.J. Heilman, K. Collins, and J. Callan. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. Proceedings of NAACL HLT. 2007.

[19] K.M. Sheehan, I. Kostin, Y. Futagi, and M. Flor. 2010. Generating Automated Text Complexity Classifications that are Aligned with Targeted Complexity Standards. ETS RR-10-28 (2010).

[20] A.M. Villamin and E.S. de Guzman. Pilipino Readability Formula: The Derivation of a Readability Formula and a Pilipino Word List. Language Study Center: Philippine Normal University. 1979.

[21] A. Rajaraman and J.D. Ullman. (2011). "Data Mining" (PDF). Mining of Massive Datasets. pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

[22] Sci-Kit Learn Org. 2007-2018. 1.9. Naive Bayes. Retrieved from https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes.

[23] Sunil Ray. 2017. 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R). Retrieved from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[24] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[25] Oco, N., Ilao, J., Syliongka, L., & Roxas, R. (2013). Measuring language similarity using trigrams: Limitations of language identification. 2013 International Conference on Recent Trends in Information Technology (ICRTIT). doi:10.1109/ICRTIT.2013.6844

[26] Antonina Durfee. (2008). Text Mining. In G. Garson, & M. Khosrow-Pour, D.B.A. (Eds.), Handbook of Research on Public Information Technology (pp. 592-603). Hershey, PA: IGI Global. doi:10.4018/978-1-59904-857-4.ch054