Chapter 3

Theoretical Framework

This chapter introduces and explains the relevant concepts and theories to be used in this research.

3.1 Linguistic Features

In machine learning research, features are measurable properties or characteristics of data being processed (Bishop, 2006). In the context of readability assessment, features can be in the form of countable elements such as average length of words, total word count, or ratio of pronouns to nouns. Good input features are critical to the success of models' predictive ability trained to solve a certain task especially in classification-based approaches. Thus, it is imperative and ideal to explore a wide range of features or predictors to be able to identify the best discriminating characteristics.

In the succeeding subsections, we list down the following feature sets to be extracted as detailed by an expert (Macahilig, 2015a) for effective training of models for readability assessment in Filipino.

3.1.1 Traditional or Surface-Based Features (TRAD)

Traditional or surface-based features were famous and commonly used in old readability formulas. This feature set is considered the easiest to extract since it usually only relies on the occurrence of selected features in a text or some form of

calculation of weights. A few examples of surface-based features are average word length, average syllables per word, total number of sentences, etc. However, they are considered as shallow and do not capture the deeper structure and complexity of a language (Hancke et al., 2012). Nonetheless, they are still used as baseline for comparing new methods. Table 3.1 shows 9 traditional, surface-based predictors extracted for this study adapted from various formula-based readability indices for both English (Flesch, 1948; Thorndike, 1921; Dolch, 1936; E. Fry, 1968; McLaughlin, 1969) and Filipino (Guevarra, 2011b; Macahilig, 2015a).

Table 3.1: Traditional, surfaced-based features extracted for the study.

Notation	**Definition**	**Formula**
WC	Total number of words per document.	-
SC	Total number of sentences per document.	-
PC	Total number of phrases per document.	-
PolySyll	Total number of words with more than 5 syllables.	-
AveWL	Average word length (AveWL) is the ratio of the summation of all lengths of words over the total number of words in the document WC.	Total length of words / WC
AveSL	Average sentence length (AveSL) is the ratio of the summation of all lengths of sentences ver the total number of sentences in the document WC.	Total length of sentences / SC
AveSyll	Average syllable count of words (AveSyll) is the ratio of the summation of all syllable counts of words over the total number of words in the document WC.	Total syllables of words / WC

3.1.2 Lexical Features (LEX)

Lexical features are words or terms belonging to a specific lexical category such as nouns, pronouns, adjectives, and adverbs. These lexical categories are linguistic elements that carry meaning or content, sometimes called information-carrying words, that make up the context of a text. The works of François and Fairon (2012), Lu (2012) and Petersen and Ostendorf (2009) backed the inclusion of LEX features, stating its significant impact to the performance of readability assessment models. While some even considered it as the most important predictor (Lorge, 1944; Chall & Dale, 1995). Table 3.2 shows 9 lexical, information-carrying predictors extracted for this study along the definition of each one. Aside from the recommendation of experts (Macahilig, 2015a), the selection of lexical features to be used in the study will also be adapted from the multiple works of François and Fairon (2012); Lu (2012); Hancke et al. (2012); Dell'Orletta et al. (2014); Forti et al. (2019b) to be able to have an increased feature space to identify what predictors may also affect reading difficulty in Filipino.

Table 3.2: Lexical and information-carrying features extracted for the study.

Notation	**Definition**	**Formula**
TTR	Type-Token Ratio (TTR) measures the number of unique lexical categories or word types T by the total number of words in a document WC . This sheds light on how a sentence is packed with content-carrying words (Templin, 1957).	$T \;/\; WC$
RootTTR	Root Type-Token Ratio (RootTTR) is a variation of TTR by Guiraud (1959) to alleviate the effect of sentence length to the predictor.	$T \;/\; \sqrt{WC}$
CTTR	Corrected Type-Token Ratio (CTTR) is a variation of TTR by Carroll (1964) to alleviate the effect of sentence length to the predictor.	$T \;/\; \sqrt{2WC}$
BiLogTTR	Bilogarithmic Type-Token Ratio (BiLogTTR) is a variation of TTR by Herdan (1964) to alleviate the effect of sentence length to the predictor.	$\log T \;/\; \log WC$

N-TR	Lexical variation for nouns is the ratio of the total number of words from the noun lexical category NC to the total number of words in the document WC. This measure was developed by Hancke et al. (2012) which hypothesized that easy documents have low nominalizations of nouns compared difficult documents.	NC / WC
V-TR	Lexical variation for verbs is the ratio of the total number of words from the verb lexical category VC to the total number of words in the document WC. This measure was developed by Hancke et al. (2012) which hypothesized that easy documents have low nominalizations of verbs compared difficult documents.	VC / WC
LexDense	Lexical Density (LexDense) is measure of lexical richness which deals with the ratio of total lexical words (nouns, verbs, adjectives, and adverb) WC_Lex against the total number of words in the document WC (Ure, 1971).	WC_Lex / WC
ForeignDense	Foreign word density is the ratio of total identified foreign words WC_{For} over the total number of words in document WC. We hypothesize that the density of foreign words such as in English will occur more in harder-to-read documents than in the easier ones.	WC_{For} / WC

CompDense	Compound word density is the ratio of total identified compound words WC_{Comp} over the total number of words in document WC. We hypothesize that the density of compound words such as *bahay-bata, balat-sibuyas* will occur more in harder-to-read documents than in the easier ones.	WC_{Comp} / WC

3.1.3 Language Model or Text Structure Features (LM)

Language model features use the capability of a trained statistical language model for predicting how likely a certain passage is with training corpus used to generate the language model (Kate et al., 2010). Moreover, language models are trained using n-grams or continuous slices of words (Broder, Glassman, Manasse, & Zweig, 1997; Dunning, 1994). In the context of readability assessment, these models can be used to capture the regularities of a language as well as content information related to text difficulty (Si & Callan, 2001). The intuition behind generating language models as potential predictors for readability is described in the work of Schwarm and Ostendorf (2005) where a teacher is more likely to look for reading materials with a particular readability level in mind rather than grouping text into various categories. In this manner, language models are usually considered as one-class classifiers that can be used to determine if a given text is likely to be similar to its structure or not.

Perplexity

For language model features, the perplexity score PP will be the numeric feature value. The perplexity score of a language model is an intrinsic evaluation metric that provides how a test sentence or target document is more likely to be produced from a certain language model. Thus, the lower the perplexity score the higher the probability (Jurafsky, 2000). The perplexity score is computed as follows,

$$PP = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \tag{3.1}$$

where N is the total size of vocabulary, $P(w_i|w_{i-1})$ is the probability of a word occurring after the previous one. Table 3.3 shows the predictors extracted using the perplexity score of various language models.

Table 3.3: Statistical language model features extracted for the study.

Notation	**Definition**
L1-UniLM	Perplexity score PP of a target document given by a statistical unigram language model trained on easy or L1 documents.
L1-BiLM	Perplexity score PP of a target document given by a statistical bigram language model trained on easy or L1 documents.
L1-TriLM	Perplexity score PP of a target document given by a statistical trigram language model trained on easy or L1 documents.
L2-UniLM	Perplexity score PP of a target document given by a statistical unigram language model trained on difficult or L2 documents.
L2-BiLM	Perplexity score PP of a target document given by a statistical bigram language model trained on difficult or L2 documents.
L2-TriLM	Perplexity score PP of a target document given by a statistical trigram language model trained on difficult or L2 documents.

3.1.4 Syllable Pattern or Orthography Features (SYLL)

The orthography of a language describes the set of linguistic rules and considerations in terms of spelling, word breaks, hyphenation and punctuation of a certain language. Likewise, the orthography defines the set of symbols used in writing the language and how these symbols build up words (Seidenberg, 1992). In 2013, the Komisyon sa Wikang Filipino (KWF) or Commission on the Filipino Language, the official regulating body of the Filipino language of the Philippines, published the *Ortograpiyang Pambansa* or National Orthography which contains the new set of guidelines and documentations for writing, specifically the syllabication pattern, of Filipino words. After the release of this document, the Department of Education has mandated the use of this orthography for all the educational institutions of the Philippines under the Memorandum 34, series of 2013[1].

[1]https://www.deped.gov.ph/2013/08/14/do-34-s-2013-ortograpiyang-pambansa/

Table 3.4 shows the normalized syllable pattern or orthography-based predictors extracted for the study. No existing work in readability assessment for Filipino text has ever explored the use of these variables as predictors.

Table 3.4: Syllabication patterns or orthographic features extracted for the study.

Notation	**Definition**	**Formula**
CCDense	Consonant Cluster Density is the ratio of the frequency of consonant clusters or consonant blends CC (joined consonant words such as *bl* in *problema*, *gr* in *programa*) against the total number of words WC in the document.	CC/WC
v_density	v_density is the ratio of the frequency of words observing the syllable pattern P from the Ortograpiyang Pambansa against the total number of words WC in the document.	P/WC
cv_density	cv_density is the ratio of the frequency of words observing the syllable pattern KP from the Ortograpiyang Pambansa against the total number of words WC in the document.	KP/WC
vc_density	vc_density is the ratio of the frequency of words observing the syllable pattern PK from the Ortograpiyang Pambansa against the total number of words WC in the document.	PK/WC
cvc_density	cvc_density is the ratio of the frequency of words observing the syllable pattern KPK from the Ortograpiyang Pambansa against the total number of words WC in the document.	KPK/WC
vcc_density	vcc_density is the ratio of the frequency of words observing the syllable pattern PKK from the Ortograpiyang Pambansa against the total number of words WC in the document.	PKK/WC

cvcc_density	cvcc_density is the ratio of the frequency of words observing the syllable pattern *KPKK* from the Ortograpiyang Pambansa against the total number of words *WC* in the document.	$KPKK/WC$
ccvc_density	ccvc_density is the ratio of the frequency of words observing the syllable pattern *KKPK* from the Ortograpiyang Pambansa against the total number of words *WC* in the document.	$KKPK/WC$
ccvcc_density	ccvcc_density is the ratio of the frequency of words observing the syllable pattern *KKPKK* from the Ortograpiyang Pambansa against the total number of words *WC* in the document.	$KKPKK/WC$
ccvccc_density	ccvccc_density is the ratio of the frequency of words observing the syllable pattern *KKPKKK* from the Ortograpiyang Pambansa against the total number of words *WC* in the document.	$KKPKKK/WC$

3.1.5 Morphological Features (MORPH)

Morphological features pertain to the countable instances of morphemes, the smallest unit of a language making up the structure of words, that convey a range of grammatical meanings for a certain language (Hancke et al., 2012). Filipino, observing the same morphological properties as Tagalog, exhibits morphological phenomena such as affixation, stress shifting, consonant alternation, and reduplication to name a few. In addition, Filipino is also rich in usage of particles such as prefixes, infixes, suffixes, and circumfixes, making the language more morphosyntactically complex than English (Ramos, 2020; Nelson, 2004). In the context of readability assessment, extracting inflectional morphemes that build up the focus, aspect, and mood of verbs are often done to identify if they contribute to a text's reading difficulty (Hancke et al., 2012). Table 3.5 shows the morphological predictors extracted for the study. We adapted these features from Hancke et al. (2012) since both Filipino and German are considered morphologically-rich languages.

Table 3.5: Morphological or verb inflection features covering focus, mood, and tense extracted for the study.

Notation	Definition	Formula
VR_{actor}	Actor-focus verb ratio VR_{actor} is the ratio of verbs as inflected by *-um-, mag-, ma-, mang-* against the total number of verbs VC in a document.	V_{actor}/VC
VR_{object}	Object-focus verb ratio VR_{object} is the ratio of verbs as inflected by *-in, -an, i-* against the total number of verbs VC in a document.	V_{object}/VC
VR_{ben}	Benefactive-focus verb ratio VR_{ben} is the ratio of verbs as inflected by *i-, ipag-* against the total number of verbs VC in a document.	V_{ben}/VC
VR_{loc}	Locative-focus verb ratio VR_{loc} is the ratio of verbs as inflected by *-an, -in, pag...an* against the total number of verbs VC in a document.	V_{loc}/VC
VR_{inst}	Instrumental-focus verb ratio VR_{inst} is the ratio of verbs as inflected by *ipang-* against the total number of verbs VC in a document.	V_{inst}/VC
VR_{ref}	Referential-focus verb ratio VR_{ref} is the ratio of verbs as inflected by *pinag-* against the total number of verbs VC in a document.	V_{ref}/VC
VR_{inf}	Infinitive aspect verb ratio VR_{inf} is the ratio of verbs as inflected by *mag-, ma-, mang-, ka- , mapag- , makipag-* against the total number of verbs VC in a document.	V_{inf}/VC
VR_{past}	Perfective aspect verb ratio VR_{past} is the ratio of verbs as inflected by perfective morphemes (ex. *nahulog, kumain, pinaalis, nag-, naging*) against the total number of verbs VC in a document.	V_{past}/VC
VR_{pres}	Imperfective aspect verb ratio VR_{pres} is the ratio of verbs as inflected by imperfective morphemes (ex. *nahuhulog, kumakain, pinapaalis, nagiging*) against the total number of verbs VC in a document.	V_{pres}/VC

VR_{future}	Contemplative aspect verb ratio VR_{future} is the ratio of verbs as inflected by contemplative morphemes (ex. *mahuhulog, kakain, papaalisin, magiging*) against the total number of verbs VC in a document.	V_{future}/VC
VR_{part}	Participle aspect verb ratio VR_{part} is the ratio of verbs as inflected by past and present morphemes against the total number of verbs VC in a document.	$V_{past}+V_{pres}/VC$
VR_{recent}	Recent-past aspect verb ratio VR_{recent} is the ratio of verbs as inflected by recent-past morphemes (ex. *kahuhulog, kakakain, kapapaalis*) against the total number of verbs VC in a document.	V_{recent}/VC
VR_{aux}	Auxilliary mood verb ratio VR_{aux} is the ratio of verbs acting as modal or auxiliary verbs (ex. *kailangan, pwede, dapat, maari, gusto, ayaw, ibig, nais*) against the total number of verbs VC in a document.	V_{aux}/VC

3.2 Supervised Machine Learning Algorithms

We provide the definition and mathematical information for each machine learning algorithms to be used by the study.

3.2.1 Logistic Regression (LR)

Logistic regression is a classification algorithm derived from applying a sigmoid function σ which transforms the output of a linear regression formula to a value of 0 or 1 in the case of binary classification. The formal representation of the algorithm is show below described in the context of readability assessment for this study,

$$P(Y = 1|X; \theta) = \frac{1}{1 + e^{-\theta^T X}} \tag{3.2}$$