Text Simplification using Hybrid Semantic Compression and Support Vector Machine for Troll Threat Sentences

Juhaida Abu Bakar[1], Nooraini Yusoff[2], Nor Hazlyna Harun[3], Maslinda Mohd Nadzir[4], Salehah Omar[5]
Data Science Research Lab, School of Computing, Universiti Utara Malaysia, Kedah, Malaysia[1, 3, 4]
Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kota Bharu, Kelantan, Malaysia[2]
Department of Information Technology and Communication, Sultan Abdul Halim Mu'adzam Shah Polytechnic, Kedah, Malaysia[5]

***Abstract*—Text Simplification (TS) is an emerging field in Natural Language Processing (NLP) that aims to make complex text more accessible. However, there is limited research on TS in the Malay language, known as Bahasa Malaysia, which is widely spoken in Southeast Asia. The challenges in this domain revolve around data availability, feature engineering, and the suitability of methods for text simplification. Previous studies predominantly employed single methods such as semantic compression, or machine learning with the Support Vector Machine (SVM) classifier consistently achieving an accuracy of approximately 70% in identifying troll sentences—statements containing threats from online trolls notorious for their disruptive online behavior. This study combines semantic compression and machine learning methods across lexical, syntactic, and semantic levels, utilizing frequency dictionaries as semantic features. Support Vector Machine and Decision Tree classifiers are applied and tested on 6,836 datasets, divided into training and testing sets. When comparing SVM and Decision Tree with and without semantic features, SVM with semantics achieves an average accuracy of 92.37%, while Decision Tree with semantics reaches 91.21%. The proposed TS method is evaluated on troll sentences, which are often associated with cyberbullying. Furthermore, it is worth noting that cyberbullying has been reported to be a significant issue, with Malaysia ranking as the second worst out of the 28 countries surveyed in Asia. Therefore, the outcomes of the study could potentially offer means, such as machine translation and relation extraction, to help prevent cyberbullying in Malaysia.**

Keywords*—Text simplification; semantic compression; machine learning; natural language processing; cyber bullying

I. Introduction

Natural Language Processing (NLP) represents a branch of artificial intelligence dedicated to enabling both machines and humans to comprehend, interpret, and deduce significance from human languages [1]. In the contemporary landscape, NLP encounters its most noteworthy challenges in the complexity of human communication. The process of deciphering and manipulating language is highly intricate, hence the common practice of employing diverse techniques to address a multitude of challenges.

This area of research encompasses numerous expanding and valuable applications. Natural Language Processing (NLP) encompasses a wide spectrum of tasks, ranging from straightforward ones like spell checking, keyword search, synonym identification, data extraction, classification, summarization, and text simplification, to more complex tasks like machine translation. In the future, NLP holds the potential to revolutionize task assistance. In this chapter, we will delve into past research related to a specific NLP task—text simplification.

Text simplification involves the transformation of a sentence into one or more straightforward sentences, making it more understandable for both machines and humans while preserving the original context and content. Additionally, text simplification serves as a valuable application that can improve various Natural Language Processing (NLP) tasks. Study by [2] highlights that text simplification tasks encompass several operations, including theoretical simplification to streamline content and structure, elaborate modification to clarify key points, and text summarization to remove peripheral or irrelevant information. The primary goal of text simplification is to enhance the accessibility of information for individuals with disabilities [3-4], those with low literacy levels [5-6], and non-native speakers [7].

In the Malay language, the exploration of text simplification is a relatively new area of study. Recent years have witnessed extensive research in Malay language studies, particularly in the domains of text summarization and sentence compression [8-11]. Researchers have been keen on enhancing the quality and cohesiveness of generated summaries. Sentence compression, a technique that involves eliminating non-essential details while preserving sentence grammar patterns, has garnered significant attention. This process identifies and removes frequently occurring sequences of adjacent words across a collection of documents, resulting in heuristic knowledge for sentence compression with an 85% confidence value [8].

Study by [8] primarily focuses on the Frequent Pattern growth tree, which stores compressed and critical information related to frequent patterns in large databases. However, it's worth noting that this study of text summarization does not encompass semantic compression, potentially leading to issues of ambiguity. Existing literature suggests that Malay language studies primarily concentrate on text summarization, specifically sentence compression, without delving into semantic comprehension. Fig. 1 illustrates the distinction between text simplification and text summarization.

Example sentence:
Google began in January 1996, as a research project by Larry Page, who was soon joined by Sergey Brin, when were both PhD students at Stanford University in California.
text simplification:
Google **was started** in January 1996~~, as a research project~~ by Larry Page~~, who was soon joined by~~ **and** Sergey Brin~~, when were both~~ **two** ~~PhD~~ students at Stanford University ~~in California,~~ **USA**.
text summarization:
Google began in ~~January~~ 1996~~, as a research project by Larry Page, who was soon joined by Sergey Brin, when were both PhD students at Stanford University in California.~~

Fig. 1. Text simplification versus text summarization [31].

In Text Simplification (TS), information extraction stands as a pivotal phase. The primary output of the information extraction process is the Syntax Tree, which illustrates the sentence's structure [12]. However, the syntax tree can become ambiguous when a sentence adheres to multiple grammar rules. To address this issue, machine learning techniques are commonly employed. These methods encompass Support Vector Machine (SVM) (e.g., [13-14]), Maximum Entropy (e.g., [15]), Decision Tree (DT) (e.g., [16]), and Conditional Random Field (e.g., [17]).

Among these techniques, the Support Vector Machine (SVM) has been recognized as the most effective classifier for text simplification, achieving an accuracy of approximately 70% [18]. It is important to note that studies employing SVM for text simplification have predominantly concentrated on the English language. In contrast, there is a lack of research on text simplification in the Malay language.

Furthermore, within the domain of Text Simplification, the primary objective is to condense a given sentence. This task necessitates a process of comprehending the inherent meaning of the sentence, commonly referred to as semantic compression.

In many text simplification approaches, a singular method is typically employed, whether it's a machine learning method or semantic compression. Studies solely focused on machine learning methods tend to overlook the significance of sentence structure properties crucial for semantic interpretation. Conversely, research exclusively centered on semantic compression may encounter challenges in predicting syntax trees, leading to potential ambiguity problems. Therefore, there is a growing recognition of the necessity to combine machine learning methods and semantic compression. In this hybrid approach, machine learning is applied to identify ambiguous sentence structures, while semantic compression is employed to simplify sentences based on relevant semantic content.

Troll is a prime example necessitating text simplification, as it often comprises sentences laden with concealed meanings. Originally, trolling involved the use of deceptive posts as bait to elicit responses from other online community members, often luring them into engaging with a fabricated story. Trolling encompasses various forms, and the term "trolling" has been broadly applied to describe various malicious or harassing activities on the internet. These activities may include instigating contentious discussions, targeting individuals or groups with harassment, sharing offensive content, vandalizing community-contributed pages, defacing memorial pages, and even being used interchangeably with cyberbullying. As a result, this study focuses on trolls associated with cyberbullying as the domain for testing a proposed text simplification method.

The motivation by engaging in TS research in a minority language offers the opportunity to develop language-specific techniques and tools, enriching the broader NLP field while deepening insights into the unique linguistic features and challenges of that language. This paper introduces a hybrid approach for text simplification in the Malay language. The model effectively distinguishes between complex and non-complex words, offering a potential solution to combat cyberbullying in Malaysia through means like machine translation and relation extraction. The key steps involve developing text simplification features that emphasize semantic aspects. Additionally, lexical features, including stemmed words, are incorporated into the study. Subsequently, hand-crafted features encompassing lexical, syntactic, and semantic attributes are organized and classified using machine learning techniques to attain the highest accuracy results.

II. Related Works

The NLP components employed in TS encompass five levels: lexical, syntactic, semantic, discourse, and pragmatic. According to [19], the TS process primarily involves the lexical and syntactic levels. However, it's worth noting that semantic considerations play a crucial role in both the lexical and syntactic approaches to ensure the preservation of word and sentence meanings.

The lexical level, referred to as lexical simplification (LS), concentrates on replacing complex words with simpler synonyms. For instance, it involves substituting "facile" with "easy." Previous research in psycholinguistics has shown that such substitutions of complex terms within a sentence, as done by comprehensive lexical simplification, have significant potential to enhance sentence readability [20]. LS involves altering the intricate or unusual phrasing within a sentence by replacing it with a synonymous word that is more straightforward and comprehensible [21].

In the realm of syntactic simplification, it encompasses distinct elements like idiomatic phrases, apposition, coordination, subordination, and voice. Study by [22] employ the typed dependency representations provided by the Stanford Parser. They argue that these formatted dependencies offer a high level of precision, facilitating the creation of straightforward standards and the automation of corporate acquisition processes.

Recent research demonstrates that the semantic approach has been applied in text simplification tasks, as evidenced by studies such as [22-27]. Study by [28] also highlights that

semantic compression can serve as a valuable technique for intelligently generalizing terms while minimizing information loss. To address structural mismatches, study by [29] suggests employing semantic parsing to rephrase sentences.

There are various approaches employed for text simplification (TS) tasks. Recent research has shown a growing interest in hybrid approaches that integrate multiple techniques for simplification such as deep semantic and monolingual machine translation have been combined in the hybrid approach, as demonstrated by [30], structural semantics and neural methods are another focus in recent studies, exemplified by [27], hybrid approaches may involve a combination of hand-crafted transformation rules, machine learning (ML) techniques, and semantic parsers, as explored by [31], these hybrid approaches often merge natural language processing (NLP) components with machine learning techniques. The research conducted by [20] advocate for the use of Machine Learning (ML) techniques as a means to achieve more reliable solutions in text simplification. These hybrid methods represent a multifaceted approach to text simplification, leveraging various techniques to enhance the quality and effectiveness of simplification processes.

As a relatively new language within the field of text simplification, a more comprehensive investigation of each feature is essential to achieve higher accuracy. The study in [32] involved the utilization of all relevant features, with a subsequent comparison of results to identify the most effective features for future use. Thus, the primary objective of this study is to combine the strengths of semantic compression and machine learning methods through hybridization. This approach aims to leverage the benefits of both techniques to enhance the practice of text simplification.

III. METHODOLOGY

The research methodology of the study can be segmented into five distinct phases: a literature review phase, a phase dedicated to defining data sets and specifications, a phase focused on designing text simplification features for the TS model, a phase involving the construction of the TS model based on SVM classifier and selected features, and finally, a phase dedicated to performance evaluation.

A. Datasets

In this study, the primary data sources include news articles, online resources, and existing datasets for the Malay language. Additionally, a corpus from previous studies, including [33-36], covering Parts of Speech (POS) and Noun Phrases, was used to create the Malay Text Simplification Dataset (Malay TS Dataset) with 6,836 instances categorized as complex or non-complex.

The work begins by utilizing the state-of-the-art corpus developed by [34], known as the Malay corpus. This corpus comprises 18,387 tokens, each of which is accompanied by word category information and is written using the Rumi script. It includes 21 word categories for part-of-speech (POS) tagging, following the standard provided by the Dewan Bahasa dan Pustaka (DBP). You can find the Malay part-of-speech tagset within the corpus in Table I.

TABLE I. PART-OF-SPEECH DBP TAGSET IN MALAY CORPUS [34]

Tag Set	Description	Example in Malay language with English gloss	Number of tokens
KN	Noun	chair (*kerusi*)	6108
KK	Verb	eat (*makan*)	2539
ADJ	Adjective	black, beautiful, deep (*hitam, cantik, dalam*)	1623
KSN	Preposition	at, to, from, to (*di,ke, dari, kepada*)	1409
KB	Auxiliary verb	will, not yet, can (*akan, belum, boleh*)	390
KG	Pronoun	me, you (*saya, awak*)	496
KH	Conjunction	which, and, or (*yang, dan, atau*)	1608
ADV	Adverb	perhaps (*bahasawanya, barangkali*)	817
KT	Question	what, how much (*apa, berapa*)	49
KBIL	Cardinal	one, two (*satu, dua*)	258
KPM	Narrator	is (*adalah, ialah*)	100
KP	Command	don't, please (*jangan, sila*)	5
KAR	Direction	in, up, down (*dalam, atas, bawah*)	48
PW	Discourse mark	even, then (*hatta, maka*)	9
KEP	Short form	*UNCR, PBB*	179
#E	Clitic *lah*	try it (*cubalah*)	31
KN@	Clitic *nya*	His/her book (*Bukunya*)	235
KNF	Deny	No, it's not (*tidak, bukan*)	171
KNK	Proper noun	*Allah, Muhammad*	236
SEN	List number	*(i), (ii), (iii), etc*	3
SYM	Any symbol or punctuations	*. , " - + etc*	2073

The study in [37] established a process for identifying complex words in three languages. This study follows the same process developed by Yimam, known as Complex Word Identification (CWI). In this process, a survey was conducted using 10 TS control samples and 10 TS non-control samples from the Malay corpus. For instance, the study focuses on TS users, who are non-native speakers. Therefore, 10 non-native speakers of the language were selected as a control sample, along with 10 native speakers. Native speakers are individuals who learned their first language in childhood, often referred to as their mother tongue [38]. Non-natives are individuals who learned a different language as their first language in childhood. Respondents were provided with texts from the Malay corpus and asked to annotate each word based on its complexity.

The results of the answers provided by the 10 native speakers and the 10 non-native speakers will determine whether a word is classified as complex or not. The label assigned to the target word is based on the responses of these 10 native and 10 non-native speakers. If at least one annotator marks the word as complex, the label will be "COMPLEX" (1); otherwise, it will be "NOT COMPLEX" (0).

Afterward, data cleaning is an integral part of this study, which involves removing punctuation and converting all letters to lowercase. This is done to address data sparsity within the dataset. The dataset comprises original sentences, target word indices, counts of annotations by native and non-native speakers for the sentences, counts of markings by native and non-native speakers for the target words, and binary and classification labels for the target words. Subsequently, a dataset consisting of 6,836 instances with labels indicating complexity or non-complexity is created. The detailed description of the Malay TS Dataset, including complexity information after data cleaning, is provided in Table II.

TABLE II. PART-OF-SPEECH DBP TAGSET IN MALAY TS DATASET WITH THE COMPLEX INFORMATION

Tag Set	Description	Number of tokens	Complex word	Non-complex word
KN	Noun	2459	299	2160
KK	Verb	1103	88	1015
ADJ	Adjective	687	82	605
KSN	Preposition	591	3	588
KB	Auxiliary verb	136	1	135
KG	Pronoun	210	6	204
KH	Conjunction	735	14	721
ADV	Adverb	332	12	320
KT	Question	22	2	20
KBIL	Cardinal	112	2	110
KPM	Narrator	*None*	*None*	*None*
KP	Command	*None*	*None*	*None*
KAR	Direction	*None*	*None*	*None*
PW	Discourse mark	*None*	*None*	*None*
KEP	Short form	6	6	0
#E	Clitic *lah*	1	1	0
KN@	Clitic *nya*	10	10	0
KNF	Deny	*None*	*None*	*None*
KNK	Proper noun	3	3	0
SEN	List number	*None*	*None*	*None*
SYM	Any symbol or punctuations	*None*	*None*	*None*

B. Proposed Method

Generally, the method begins by importing the raw Malay text dataset. The proposed approach encompasses three stages before obtaining the output of text simplification. Initially, the raw Malay text Part-of-Speech (POS) dataset is converted into feature extractions. Two types of feature extractions are employed: semantic compression features and lexical features. Text compression is achieved by using a semantic network and information on term frequencies from a frequency dictionary. Subsequently, lexical features are constructed based on Part-of-Speech (syntactic), vowels (lexical), characters (lexical), and syllables (lexical). Handcrafted features combine semantic compression and lexical features. Finally, machine learning classifiers, specifically Decision Tree (DT) and Support Vector Machines (SVM), are used to identify complexity patterns in the Malay language. This hybrid method is configured for these two machine learning classifiers using the frequency dictionary. Additionally, the study evaluates this method on previously unseen troll sentences. Fig. 2 illustrates the proposed method during this phase.

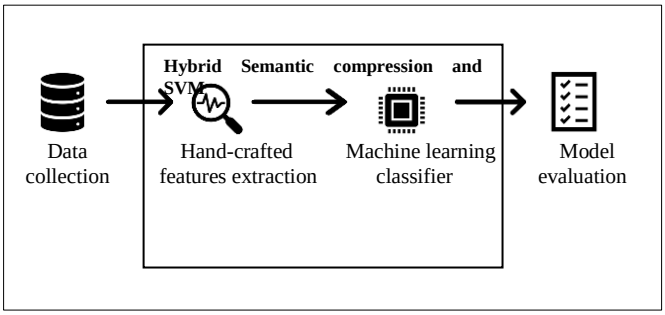

Fig. 2. Proposed method.

This phase assessed the validity of the hypotheses derived from the literature review. It primarily involved the preparation and development of lexical, syntactic and semantic features based on the findings from the preceding step. The experimental aspect of this phase focused on extracting features related to factors like length, frequency, lexical, syntactic, and semantic characteristics. Additionally, base words in the Malay language were extracted and incorporated as features. The Part-of-Speech (POS) tags present in the Malay corpus were also employed as syntactic features. To align with semantic requirements, a frequency dictionary was generated. The lexical features, as presented in Table III, were ultimately adopted for this study. Subsequently, each feature in token form underwent a normalization process to facilitate the development of learning models based on Decision Trees (DT) and Support Vector Machine (SVM) classifiers.

TABLE III. LEXICAL, SYNTACTIC AND SEMANTIC FEATURES IN MALAY TS DATASET

Type	*Features*	*Abbreviation*
Lexical	Number of syllables	SYL
	Length of word	CHAR
	Base word	STEM
	Frequency of word	FREQ
	Number of token (not stem)	VOW
	Number of token (after stem)	Vow
Syntactic	Part-of-speech Tagging	POS
Semantic	Frequency dictionary	DF

Algorithm 1 outlines the features for constructing the Malay TS method for the Malay language.

Algorithm 1: Malay TS method

```
1: Input: text T, word_feature W, gaps G, discard_empty D, flags F
2: read T sequence,
   read word_feature W,
     feature_type1: Syllable feature,
     feature_type2: Character feature,
     feature_type3: Stem feature,
     feature_type4: Frequency feature,
     feature_type5: Part-of-Speech tag feature,
     feature_type6: Vowel feature,
     feature_type7: Frequency distribution feature,
   read gaps G, read discard_empty D, read flags F,
3: If feature_type3 exists in T sequence
4:  Enhance with the modification rules and steps
5:    If not
6:          Continue to machine learning algorithms (SVM, DT)
7:    Fit to gaps G, discard_empty D, flags F
```

As a result of the above works, two classifiers were utilized, specifically the SVM and DT classifiers. The experiment is partitioned into two segments: one that takes semantic features into account and one that does not take semantic features. Data was divided using k-fold cross-validation (k=10), and subsequently, the average outcomes are computed. These results will be analyzed and discussed in the Experiment and Results section.

C. Performance Evaluation

In the domain of machine learning, particularly in the context of statistical classification, a confusion matrix, alternatively referred to as an error matrix, is a structured table format that provides a means to visually assess the effectiveness of an algorithm, often in the context of supervised learning. Fig. 3 illustrates the configuration of the confusion matrix. Its primary purpose is to evaluate the performance of a classification algorithm. In this study, four metrics were employed: accuracy, precision, recall, and F1-measures, to gauge the performance of the classification algorithm.

		Predicted	
		Negative (N) -	*Positive (P)* +
Actual	*Negative (N)* -	True Negative (TN)	False Positive (FP) **Type 1 error**
	Positive (P) +	False Negative (FN) **Type II error**	True Positive (TP)

Fig. 3. Confusion matrix.

IV. Experiments and Results

The hybrid method proposed in this study was employed on a dataset comprising 6,836 instances. This original dataset encompasses lexical details, syntactic information, sentences, base words, and semantic information, as illustrated in Table IV.

TABLE IV. Feature Engineering for Malay TS Dataset

POS	CHAR	VOL	SYL	Sentence	STEM	DF	Vow	Binary	Class
1	4	2	2	Asid Alfa Li	asid	4	2	1	Complex
1	4	2	2	Asid Alfa Li	alfa	3	2	1	Complex
1	6	3	3	Asid Alfa Li	lipoik	3	3	1	Complex
1	7	3	3	Asid Alfa Li	manfaat	3	3	1	Complex
2	5	2	2	Asid Alfa Li	untuk	73	2	0	Not complex
1	5	2	2	Asid Alfa Li	saraf	13	2	0	Complex
3	4	2	2	Saya menga	saya	9	2	1	Complex

Subsequently, a frequency dictionary and vowel characteristics dictionary are constructed for base words, contributing to the generation of semantic features. The frequency dictionary tallies the occurrences of base words within the corpus, while the vowel characteristics dictionary calculates the count of vowels in each base word. The POS features encompass a set of 31 labels, including nouns, prepositions, pronouns, verbs, denies, and more, as shown in Table V. The stem feature is then removed from the final dataset, leaving the DF (frequency dictionary) and Vow (vowel characteristics) features as representations of the stem word, as illustrated in Table VI.

TABLE V. Part of Speech Features with 31 Labels

Tagset	Labeling	Numbering
Noun	kn	1
Preposition	ksn	2
Pronoun	kg	3
Verb	kk	4
Deny	knf	5
Conjunction	kh	6
Adjective	adj	7
Adverb	adv	8
Question word	kt	9
Verb with clitics -nya	kk@	10
Auxiliary verb	kb	11
Narrator	kpm	12
Short form	kep	13
Cardinal	kbil	14
Proper noun	knk	15
Noun with clitic -lah	kn#	16
Adjective with clitics -nya	adj@	17
Adverb with clitics -nya	adv@	18
Pronoun with clitics -lah	kg#	19
Noun with clitics -nya	kn@	20
Verb with clitics -lah	kk#	21

Tagset	Labeling	Numbering
Direction	kar	22
Command	kp	23
List number	sen	24
Adjective with clitics -lah	adj#	25
Auxiliary verb with clitics -lah	kb#	26
Adverb with clitics -lah	adv#	27
Pronoun with clitics -nya	kg@	28
Explanation word	kkt	29
Deny with clitics -lah	knf#	30
Deny with clitics -nya	knf@	31

TABLE VI. FEATURE ENGINEERING FOR MALAY TS DATASET WITH SEMANTIC FEATURE

POS	CHAR	VOW	SYL	DF	Vow	Class
1	4	2	2	4	2	Complex
1	4	2	2	3	2	Complex
1	6	3	3	3	3	Complex
1	7	3	3	3	3	Complex
2	5	2	2	73	2	Not complex
1	5	2	2	13	2	Complex
3	4	2	2	9	2	Complex

In the realm of machine learning, data normalization is employed to reduce the impact of feature scales on model training. The preparation of data for machine learning in this study involves the utilization of numerical data, ensuring that our model converges to optimal weights and, ultimately, resulting in a more precise model. To achieve this, min-max normalization has been implemented. Regarding class labels, they are assigned values of 0 (indicating simplicity or non-complexity) or 1 (indicating complexity), as illustrated in Table VII. Subsequently, the dataset has been divided into training and testing sets using a 10-fold cross-validation approach, denoted as Tr:Te dataset.

TABLE VII. DATA AFTER NORMALIZATION PROCESS

	0	1	2	3	4	5	6
0	0.000	0.1667	0.250	0.086957	0.010714	0.33333	1.0
1	0.000	0.1667	0.250	0.086957	0.007143	0.33333	1.0
2	0.000	0.2778	0.375	0.130435	0.007143	0.50000	1.0
3	0.000	0.3333	0.375	0.130435	0.007143	0.50000	1.0
4	0.033	0.2222	0.250	0.086957	0.257143	0.33333	0.0

The learning process is subsequently executed using DT and SVM classifiers. To ensure a robust evaluation, the dataset has been split into an 80% training set and a 20% testing set, denoted as 80Tra:20Test. For the SVM classifier, the RBF kernel and class weighting have been applied, particularly beneficial for handling imbalanced datasets. Following the completion of the experiment table containing semantic features, the most effective classifier was determined. This optimal classifier is then saved as a "pickle" file, enabling it to be used for testing new data. In the context of this study, the aim is to classify troll data as either complex or non-complex.

After completing the feature engineering process, the training datasets undergo several performance evaluations. Two algorithms are employed to predict text simplification, distinguishing between complex and non-complex words. To ensure the suitability of the chosen model, a score test model is utilized. The algorithms in use are Decision Tree classifiers and Support Vector Machine (SVM). The modeling is implemented in a Jupyter notebook using Python code, and both datasets, one with semantic features and one without, are tested. The Decision Tree classifier achieves its highest accuracy of 92.98% when using semantic feature information. On the other hand, the SVM achieves its highest accuracy of 93.20% with or without the semantic feature information. This suggests that semantic features may or may not be necessary for the SVM classifier, but there is a significant difference for the Decision Tree classifiers.

The average accuracy of both classifiers indicates that SVM outperforms the DT classifier by a margin of 0.6%. Table VIII provides a performance comparison between the two classifiers, revealing that the frequency dictionary does not significantly impact the results. Both cases, with and without a frequency dictionary, yield similar accuracy levels. The presence or absence of the frequency dictionary doesn't result in a noticeable difference in average accuracy in this experiment.

However, when examining each production of the classifier model individually, the significance of semantic features in the training dataset becomes evident. Table IX and Table X present precision, recall, and F1-score for the best models of SVM and DT, respectively, highlighting the importance of semantic features in improving these metrics.

TABLE VIII. PERFORMANCE OF TWO CLASSIFIERS WITH TWO DIFFERENCE FEATURES

Data Split / ML classifier	Frequency distribution			
	With frequency distribution		*Without frequency distribution*	
	DT (%)	SVM (%)	DT (%)	SVM (%)
90Tr:10Te	**92.98**	92.40	92.69	92.40
80Tr:20Te	92.62	**93.20**	92.91	**93.20**
70Tr:30Te	91.96	92.30	91.61	92.30
60Tr:40Te	91.15	92.07	91.55	92.07
50Tr:50Te	90.46	92.22	91.72	92.22
40Tr:60Te	90.59	92.52	91.83	92.52
30Tr:70Te	90.76	92.35	91.98	92.35
20Tr:80Te	89.78	92.10	91.17	92.10
10Tr:90Te	90.56	92.17	90.44	92.17
Average	91.21	**92.37**	91.77	**92.37**

TABLE IX. SVM LEARNING MODEL

	Precision (%)	Recall (%)	F1-score (%)	Support
0	94	100	97	1279
1	50	3	6	89
accuracy			93	1368
macro avg	72	52	51	1368
weighted avg	91	93	91	1368
0	94	100	97	1279

TABLE X. DT LEARNING MODEL

	Precision (%)	Recall (%)	F1-score (%)	Support
0	94	98	96	632
1	58	27	37	52
accuracy			93	684
macro avg	76	63	67	684
weighted avg	92	93	92	684
0	94	98	96	632

The top-performing model from the Malay TS Dataset, as determined by the research conducted by [40], is utilized to categorize unannotated troll threat sentences. The research materials comprise vlogs, which are video content sourced from the YouTube platform. The study scrutinizes 30 videos recorded by Mat Luthfi between 2011 and 2014. This investigation delves into the use of sarcastic language in YouTube videos, utilizing modern technology as the primary medium of contemporary society. Sarcasm is the examination of employing irony to ridicule or express disdain. On the other hand, "trolling" refers to a predominantly indirect form of communication. The term "trolling" is widely used to describe various malicious or harassing activities on the internet, such as initiating inflammatory discussions, among others, as noted by [39]. To the best of the researcher's knowledge, there is no publicly accessible Malay language troll dataset, so the work by [40], which examines sarcasm, serves as a suitable substitute for a troll dataset.

Before classifying unannotated troll threat sentences as either complex or non-complex words, these sentences (unseen data) must undergo a feature extraction process. This study investigates three different types of sarcasm: Irony Sarcasm, Sarcastic Sarcasm, and Sinise Sarcasm. There are 173 instances in 11 scripts for Irony Sarcasm, 101 instances in seven scripts for Sarcastic Sarcasm, and 303 instances in 10 scripts for Sinise Sarcasm, totaling 578 instances used for testing the Malay TS model.

The initial step involves data cleaning, which includes removing punctuation, converting words to lowercase, and applying the stemming process. Subsequently, a Malay Part-of-Speech tagging system, developed based on the ID3 algorithm by [41], is employed. Table XI provides an overview of the unseen dataset and its preparation process.

TABLE XI. UNSEEN DATASET

INPUT	POS	CHAR	VOW	SYL	STEM	DF	Vow
Test	4	4	1	1	*test*	1	1
Ke	2	2	1	1	*ke*	1	1
Facebook	1	8	4	2	*facebook*	1	4
Dalam	2	5	2	2	*dalam*	2	2
hidup	4	5	2	2	*hidup*	1	2
aku	3	3	2	2	*aku*	3	2
tak	5	3	1	1	*tidak*	13	2
da	4	2	1	1	*ada*	8	2
sapa-sapa	9	9	4	4	*siapa*	1	3

Table XII displays the proportions of complex and non-complex sentences in the troll threat dataset. Language experts have thoroughly evaluated the test results on these troll threat sentences. Table XIII presents the marks assigned by expert analysts to each test data sample generated by the Malay TS model.

TABLE XII. PROPORTION OF COMPLEX AND NON-COMPLEX TROLL SENTENCE

Sarcasm types	Non-complex	Complex
Irony	151	22
Sarcastic	88	13
Sinise	284	19

TABLE XIII. EXPERT RESULT FOR TROLL SENTENCE BASED ON SVM

Test sample	Irony	Sarcastic	Sinise
Total token	173	101	303
Token wrongly label	22	13	19
Token correctly label	151	88	284
Accuracy (%)	87.28	87.13	93.73
Average accuracy (%)	**89.38**		

As indicated in Table XIII, the SVM model effectively recognizes only non-complex words. It encountered difficulties in identifying complex words within this unseen dataset, resulting in a low success rate for complex words. When testing with unseen data using SVM, it shows that there are no instances of Type II errors, but Type I errors are present. The SVM model struggles to predict the complex class in three separate unseen datasets.

According to Table XIV, the Decision Tree (DT) model demonstrates success in identifying both non-complex and complex words. However, it occasionally misclassifies words, leading to a lower accuracy percentage compared to the SVM model. Testing on the unseen data reveals the presence of both Type I and Type II errors in the predictions made by the DT model. Notably, the DT model can predict complex classes in the Sarcastic and Sinise datasets, although the number of accurate predictions in these cases is relatively small.

TABLE XIV. EXPERT RESULT FOR TROLL SENTENCE BASED ON DT

Test sample	Irony	Sarcastic	Sinise
Total token	173	101	303
Token wrongly label	41	11	26
Token correctly label	132	90	277
Accuracy (%)	76.30	89.11	91.42
Average accuracy (%)	**85.61**		

V. DISCUSSIONS

In this project, a novel dataset called the Malay TS Dataset has been introduced. Additionally, a new Malay TS method has been developed by integrating three levels of NLP components with ML classifiers. The proposed method combines lexical, syntactic, and semantic features with an SVM classifier. To assess the classifier model, a comparison has been made between SVM and DT classifiers, and the findings of this comparative study are presented.

Based on the readings, the SVM classifier exhibits the highest accuracy in identifying troll sentences. The experiment involved utilizing K-fold cross-validation to split the data. To assess the method's effectiveness, the outcomes of the proposed approach were compared with another classifier, specifically DT. The proposed approach demonstrates promising results with a robust classifier model. The findings indicate that the SVM classifier, utilizing an 80-20 split of training and test data, performs as the best classifier model. However, when applied to troll data, the developed SVM model struggles to predict complex words. In contrast, the DT model, while encountering fewer complex words, exhibits better performance in predicting them.

In this research, an automated Malay TS model has been successfully developed. A novel approach, referred to as the Hybrid Semantic Compression-SVM method, has been introduced. This method aims to identify complex words within text. The research utilizes a dataset extracted from the Malay corpus by [33], containing a total of 6,836 instances. Previous studies have typically employed these two methods independently, while this study seeks to combine them for enhanced accuracy. The primary objective of this research is to hybridize semantic compression and Support Vector Machine to enhance text simplification performance. This overarching goal is complemented by three sub-objectives. Firstly, the creation of a Malay TS lexical dataset is undertaken. Secondly, the design of text simplification features for the TS model is carried out, drawing from prior work by [42]. Lastly, the results of the proposed method are evaluated against an existing Python-based classifier.

VI. CONCLUSION

Text simplification is a subfield of NLP that has seen significant development in recent years. While research in English has been extensive, tackling simplified text in other languages presents challenges due to limited resources and associated data. This study focuses on analyzing lexical, syntactic, and semantic features to identify troll threat sentences in the Malay language, and the development of resources marks the beginning of this effort.

In summary, this study exclusively incorporates frequency dictionary features within the semantic compression method. Looking ahead, there are several avenues for enhancing this project. Malay, being a minority language, has limited potential for leveraging semantic information. Semantic compression is a component of semantic analysis and comprises two crucial stages: the frequency dictionary and the semantic network. In this research, to the best of our knowledge, only the frequency dictionary has been implemented, as the code is available for development alongside existing features (lexical and syntactic). However, due to the constraints in accessing tools freely for building syntactic information based on dependencies and constituent trees, the discussion of semantic networks is omitted in this study.

To enhance the application of this project, it can be extended with three additional stages in the development of Complex Word Identification (CWI). These stages encompass Substitution Generation, Substitution Selection, and Substitution Ranking, constituting the second, third, and fourth steps in CWI. The second step involves generating potential substitutions for the target words identified in the initial step. Subsequently, the system selects the most appropriate replacement, and the final step entails organizing the hierarchy of replacement options that can be applied to the previously identified target word.

Exploring higher-level Natural Language Processing (NLP) components, such as syntactic analysis, proves more suitable for analyzing social media data compared to mere word-level comprehension. Lexical feature analysis, on the other hand, aligns better with users facing language difficulties (e.g., dyslexia, aphasia) and non-native speakers. Investigating patterns in troll sentences as compared to standard Malay sentences could yield valuable insights if developed further.

Social network datasets necessitate a distinct approach from conventional language sentences. There are additional preprocessing steps required to analyze such data effectively. Handling text abbreviations, dialects, slang, and other variations is essential before arriving at the base words within the text. Techniques like lemmatization are more appropriate for word recognition than stemming. Furthermore, resources like WordNet Bahasa should be considered in this analysis. A comprehensive study integrating social network analysis and data analytics is essential for identifying troll threat sentences.

ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers (RACER/1/2019/ICT02/UUM/1).

REFERENCES

[1] M. J. Garbade, “A Simple introduction to Natural Language Processing,” Becoming Human: Artificial Intelligent Magazine, 2018.

[2] A. Siddharthan, "A survey of research on text simplification," ITL - International Journal of Applied Linguistics, 165, pp. 259-298, 2014.

[3] Y. Canning, J. Tait, J. Archibald, and R. Crawley, "Cohesive generation of syntactically simplified newspaper text," In Proceedings of the Third International Workshop on Text, Speech and Dialogue (TSD), 2000.

[4] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura, "Text simplification for reading assistance: A project note," In Proceedings of the 2nd International Workshop on Paraphrasing, 2003.

[5] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. d. M. Fortes, T. A. S. Pardo, and S. M. Aluísio, "Facilita: reading assistance for low literacy readers," In Proceedings of the 27th ACM International Conference on Design of Communication, 2009.

[6] D. J. Belder, and M. Moens, "Text simplification for children," In Proceedings of the SIGIR Workshop on Accessible Search Systems, pp. 19–26, 2010.

[7] A. Siddharthan, "Preserving discourse structure when simplifying text," In Proceedings of European Workshop on Natural Language Generation (ENLG), 2002.

[8] S. Alias, S. K. Mohammad, G. K. Hoon, and T. T. Ping, "A Malay text corpus analysis for sentence compression using pattern-growth method," Jurnal Teknologi, 78(8), 2016.

[9] S. Alias, S. K. Mohammad, G. K. Hoon, and M. S. Sainin, "Understanding Human Sentence Compression Pattern for Malay Text Summarizer," In 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP) (pp. 1-6). IEEE, 2018.

[10] D. Gerz, I. Vulić, E. Ponti, J. Naradowsky, R. Reichart, and A. Korhonen, "Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction," Transactions of the Association of Computational Linguistics, 6, pp. 451-465, 2018.

[11] S. A. M. Noah, N. M. Ali, and M. S. Hasan, "*Penjanaan Ringkasan Isi Utama Berita Bahasa Melayu berdasarkan Ciri Kata* (Generation of News Headline for Malay Language based on Term Features)," GEMA Online® Journal of Language Studies, 18(4), 2018.

[12] D. Jurafsky and J. H. Martin, Speech and language processing (Vol. 3). London: Pearson, 2014.

[13] F. Ali, D. Kwak, P. Khan, S. H. A. Ei-Sappagh, S. R. Islam, D. Park and K. S. Kwak, "Merged ontology and SVM-based information extraction and recommendation system for social robots," IEEE Access, 5, pp. 12364-12379, 2017.

[14] M. N. Ayyaz, I. Javed, and W. Mahmood, "Handwritten character recognition using multiclass svm classification with hybrid feature extraction," Pakistan Journal of Engineering and Applied Sciences, 2016.

[15] P. Ficamos, Y. Liu, and W. Chen, "A naive bayes and maximum entropy approach to sentiment analysis: Capturing domain-specific data in weibo". In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 336-339). IEEE, 2017.

[16] Z. Yan, D. Tang, N. Duan, S. Liu, W. Wang, D. Jiang, ... and Z. Li, "Assertion-based QA with Question-Aware Open Information Extraction," In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[17] D. Thai, S. H. Ramesh, S. Murty, L. Vilnis, and A. McCallum, "Embedded-State Latent Conditional Random Fields for Sequence Labeling," arXiv preprint arXiv:1809.10835, 2018.

[18] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on Twitter." *Computers in Human Behavior*, *89*, pp. 258-268, 2018.

[19] M. Shardlow, "Lexical simplification: optimising the pipeline," The University of Manchester (United Kingdom), 2015.

[20] G. Paetzold and L. Specia, "Lexical simplification with neural ranking," In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 34-40, 2017.

[21] S. Stajner and H. Saggion, "Data-Driven Text Simplification". In Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts, 19–23. Association for Computational Linguistics, 2018.

[22] A. Siddharthan, A. Nenkova, and K. McKeown, "Information status distinctions and referring expressions: An empirical study of references to people in news summaries," Computational Linguistics, 37(4), pp. 811–842, 2011.

[23] H. V. Jagadish, T. N. Raymond, C. O. Beng, and K. H. Anthony, "ItCompress: An Iterative Semantic Compression Algorithm," In Proceedings. 20th International Conference on Data Engineering, 2004.

[24] A. Omri and A. Rapport, "Universal Conceptual Cognitive Annotation (UCCA)," In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,1, pp. 228–238, 2013.

[25] S. Narayan, and C. Gardent, "Unsupervised Sentence Simplification Using Deep Semantics". In Proceedings of the 9th International Natural Language Generation conference, pp. 111–120, 2016.

[26] A. Omri and A. Rapport, "The State of the Art in Semantic Representation," 2017.

[27] E. Sulem, O. Abend, and A. Rappoport, "Simple and effective text simplification using semantic and neural methods," In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 1, pp. 162–173, 2018.

[28] D. Ceglarek, "Semantic compression for text document processing," Trans. Computational Collective Intelligence, 2014.

[29] B. Chen, S. Le, H. Xianpei, and A. Bo, "Sentence rewriting for semantic parsing," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1, pp. 766–777. 2019.

[30] S. Narayan and C. Gardent, "Hybrid simplification using Deep Semantics and Machine Translation," In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 435–445, 2015.

[31] C. Niklaus, B. Bermeitinger, S. Handschuh, and A. Freitas, "A sentence simplification system for improving relation extraction," In Proc. of COLING'16, 2016.

[32] G. Smolenska, *Complex Word Identification for Swedish*. MS thesis, Dept. of Linguistics and Philology, Uppsala University, Sweden, 2018.

[33] J. A. Bakar, Development of a Malay Part-of-Speech Tagging for Jawi Characters based on Maximum Entropy Model with Contextual Information (in Malay language). Ph.D. Thesis, Universiti Kebangsaan Malaysia, 2016.

[34] H. Mohamed, N. Omar, and M. J. Ab Aziz, "Statistical malay part-of-speech (POS) tagger using Hidden Markov approach," In 2011 International Conference on Semantic Technology and Information Retrieval, pp. 231-236, IEEE, 2011.

[35] N. Saphra and A. Lopez, "Language Models Learn POS First," In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 328-330, 2018.

[36] S. Ab Rahman, N. Omar, and M. J. Ab Aziz, "The effectiveness of using the dependency relations approach in recognizing the head modifier for malay compound nouns," In 2014 International Conference on Computer and Information Sciences (ICCOINS), pp. 1-5, IEEE, 2014.

[37] S. M. Yimam, H. M. Alonsso, S. Stajner, M. Riedl, and C. Biemann, "Learning Paraphrasing for Multiword Expressions," In *Proceedings of the 12th Workshop on Multiword Expressions*, Association for Computational Linguistics, pp. 1–10, 2016.

[38] A. Davies, Native speakers and native users: Loss and gain. *Cambridge University Press*, 2013.

[39] T. Jussinojo, Life-Cycle of Internet Trolls. Master's Thesis, University of Jyväskylä, 2018.

[40] S. N. M. A. Rashid and N. A. Yaakob, "*Jenis bahasa sindiran dalam ujaran Vlog* (The type of sarcastic language in Vlog speech)," International Journal of Language Education and Applied Linguistics, 2017.

[41] M. F. Salim, *Text simplification using Syntactic Simplification Approach* [Unpublished manuscript]. School of Computing, Universiti Utara Malaysia, 2022.

[42] S. M. Yimam, S. Stajner, M. Riedl, and C. Biemann, "CWIG3G2 – ComplexWord Identification Task across Three Text Genres and Two User Groups," In Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2, Taipei, Taiwan, Asian Federation of Natural Language Processing, pp. 401– 407, 2017.