Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# INTENT CLASSIFICATION FOR MALAYSIAN ACADEMIC WRITERS' PROOFREADER CHATBOT USING MACHINE LEARNING

**SITI NOOR BAINI MUSTAFA[1], LAILATUL QADRI BINTI ZAKARIA[2]**

[1,2] Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

E-mail: [1]mustafa.baini@gmail.com, [2]lailatul.qadri@ukm.edu.my

**ABSTRACT**

The intent classification component is important in developing a chatbot as it helps the chatbot system to understand the meaning and purpose of conversation from the user. Earlier researchers have developed datasets for niche domains and analyzed various input representations and machine learning techniques for chatbot intent classification. However, there is no dataset and intent classification analysis for chatbot readily available in the niche of proofreading. Other than that, this study finds out the feature, input representation, and the best machine learning classifier that are suitable for intent classification analysis. This research is divided into seven main phases. The first phase is the feasibility study. The second phase is the dataset development. The third phase is the text preprocessing phase where input is cleaned and normalized. The fourth phase is the feature extraction phase whereby features are extracted using POS tagger, bag of words technique, and bigram words technique. The fifth phase is the input representation phase using Term Frequency – Inverse Document Frequency (TF-IDF) technique, Word2Vec embedding, or One-Hot encoding technique. Finally, the sixth phase is the intent classification phase using machine learning algorithms. The machine learning methods tested were Support Vector Classifier, Support Vector Machine, Stochastic Gradient Descent, and Naïve Bayes. The final phase is the testing phase. This study finds that the combination of noun and verb as features and using One-Hot encoding as input representation together with Support Vector Machine as the machine learning technique produces the best performing classifier for this research with 0.89 accuracy. This study hopes to pioneer the development of a proofreading chatbot that can help and take over a proofreader's task of answering the questions asked by Malaysian academic writers regarding the grammar corrections made.

**Keywords:** *Chatbot; Intent classification, Machine learning, Natural language processing; Proofreading corpus*

## 1. INTRODUCTION

The implementation of chatbots as a support system for organizations is widely used and accepted by consumers. Chatbot as a support system entity can be found in various industries such as telecommunications, tourism, and sales. As mentioned in the Allied Market Research[1], the chatbot market is estimated to be worth US$339.3 billion by 2027 with an annual growth rate of 27.3%. Such projection of growth is largely due to research advancement in natural language processing and machine learning [1].

Besides that, there are also many researches in the subject of using chatbot to assist language learning. A study by [2] shows that interaction between humans and chatbots for short-term learning of tenses and grammar is just as effective as human-to-human interaction. The study was conducted with Korean undergraduates whose mother tongue is not English language. On the other hand, chatbot can assist English language learning by providing a comfortable environment to ask questions repeatedly and make mistakes, which are all part and parcel of mastering a language [3].

Similarly, this study also focuses on English Language for non-native speakers. Albeit

[1] https://www.alliedmarketresearch.com/chatbot-market

proofreading applications such as Grammarly[2] and ATDEiTTM [4] can help ease writing tasks for academic writers, engagement with a professional proofreader allows the writer to ask questions about the corrections made. Counter-checking is a crucial task on the writer's part to ensure that the corrected sentence carries the intended meaning even after grammatical errors have been sorted out. Counter-checking is especially required for non-native speakers who may have less grasp in understanding the nuances of the English language [5]. The use of a chatbot may replace the role of a proofreader for answering such questions as a chatbot can be a cognitive medium for such tasks [6].

According to [7] cognitive services are integrated into one of the main components of a chatbot's architecture. The chatbot's architecture consists of three main components: user message analysis, dialogue management, and response generation. The user message analysis component analyzes user input to determine user intent. The dialogue management component regulates the conversation structure based on user intent, and finally response generation component generates appropriate responses from the knowledge base to the user.

This study focuses on the user message analysis component of a chatbot. This study aims to provide automatic intent identification of input or questions asked by academic writers after the proofreading of their writing. Ultimately, identifying user intent is a crucial task to ensure an effective chatbot response [8]. Therefore, the data received for this research was collected from a proofreading service company which needed to be prepared before it could be developed into a dataset. The raw data was annotated as questions and answers regarding the corrected sentences and labeled into a dataset of 876 lines for a proofreader chatbot.

During the analysis of the dataset, similar patterns were identified in the questions that can group them into several categories. The outcome of the research would show if manual identification of question patterns using prefixes is the best feature to use for intent classification when using bigram as feature to extract. The specific patterns of prefixes were used to then label the questions into their intent categories. This process mimics a previous study by [9] for a software development management chatbot and is considered as the chatbot pre-development phase [10] The labelled dataset was then tested under several different conditions such as text preprocessing, feature extraction, vectorizers, and embedding techniques to discover the best performing machine learning model for the proofreader chatbot dataset. The machine learning methods used for this study are Support Vector Classifier, Support Vector Machine, Naïve Bayes, and Stochastic Gradient Descent.

The major contributions of this research are as follows:

1. A list of intent categories for text classification in proofreading domain.

2. A dataset corpus of questions and answers between proofreader and academic writer (non-native English speaker) in the domain of English Language tenses and grammar error.

3. Identification of features to extract for intent classification in the proofreading domain using machine learning.

## 2. RELATED WORKS

The three contributions of this paper separate this study into two major topics which are manual dataset annotation and intent classification. The following discusses the reference studies used in materializing the contribution goals. Therefore, our references are separated into two sections namely Dataset Annotation and Intent Classification Techniques.

### 2.1 Dataset Annotation

As there is no known dataset in the domain of proofreading for Malaysian academic writer, a new corpus was created for this study. For the purpose of clarification, the dataset in this paper was created for question-answering with no relationship between one question and its conversation history [11] as this is an initial study for such corpus.

In the study by [9], archived chat messages were analyzed to determine possible intent classes. From the analysis, it was discovered that there were possibly 14 chat topics. Subsequently, within each topic there were seven possible dialog acts to further specify the intent class. As an outcome, 13 intent categories were outlined in combination of the chat topics and dialog acts. The dataset of 8030 messages was then manually labelled with 13 intent categories. Some examples of intent categories in the multi-turn conversation chatbot within the software development management domain are Greet, Plan Task, Query Plan, Schedule Meeting, Report Progress, and Query Progress. However, this study

[2] https://www.grammarly.com/

did not state the method of analyzing the chat messages in order to determine their intent categories.

Differently in the study by [11], its dataset was manually annotated by two personnel. One as the questioner and the other as the answerer. This setup was used as the Conversational Question Answering corpus created in this study which referred to various other existing datasets such as MCTest and CNN/Daily Mail. The conversational input (question) of the dataset is the outcome of possible questions from the passages in the given existing datasets. The questions were categorized based on whether or not it was dependent on the conversation history. If found to be dependent, the intent of the question is then based on an explicit marker or an explicit coreference marker such as he or she as a feature to extract. Subsequently, the conversational output (answer) was based on the rationale from the passages in the existing datasets. 100 annotated conversations were analyzed to classify the intent of the conversational output into seven categories namely Named Entity, Noun Phrase, Yes, No, Number, Data/Time, and Other.

Without relying on manual annotators, [12] used unsupervised machine learning to perform citation intent classification on an existing dataset called Citation Context Database (C2D). This study used several combinations of embedding techniques such as BERT, Glove, and Infersent that was teamed with different clustering techniques like Kmeans and DBScan. Subsequently, the context of citation intent in C2D was classified. The intent categories identified for C2D are categorized as Background, Method, and Results.

Meanwhile in our research, the dataset is manually annotated to suit the nature of questions that are commonly asked by academic writers to proofreader. Human-labeled data for natural language processing tasks results in better generalization capability for the machine learning model [13]. This is especially the case in the niche of proofreading chatbot for Malaysian academic writers.

### 2.2 Intent Classification Techniques

Five related works were selected as baseline researches for this study. These researches were selected based on the similar categories of chatbot being studied as in this paper. The categories of chatbot as per [7] are:

1. Chatbot conversation medium is text.
2. Chatbot is for a closed knowledge domain.
3. Chatbot provides intrapersonal service.
4. Chatbot is task oriented.

A study by [14] used several machine learning techniques to measure the accuracy of classification using semantic hashing representations. This study records the duration of text preprocessing and feature extraction besides measuring the statistical performance of each model. According to this study, the semantic hashing technique may improve model's performance and is an alternative to embedding technique. The dataset used in this study is the publicly available 'Travel Scheduling' Corpus and 'Ask Ubuntu' and 'Web Applications' StackExchange Corpus.

The same sets of dataset were used in the study by [15]. This study found that the SVM model in combination with embedding technique gave a better accuracy in comparison to deep learning model (LSTM). The performance of the intent classification models was measured on the entire corpus and also on each dataset separately. The purpose of this exercise was to discover if there was a separate dataset that was more efficient for text classification. This study concludes that the hierarchical embedding technique improves model's performance for intent classification. However, it can also be concluded that hierarchical embedding is only efficient on certain datasets.

In the study by [9], the data annotation efforts was done on a closed domain chatbot as well. The domain of the intent classification research was in software development management. This study used rule-based extraction for terms with special semantics within the software development domain. Intent classification is done using several combinations of text preprocessing, feature extraction, and machine learning models. This study found that feature extraction that was performed prior to intent classification gave profound impact to model's performance.

Another study on intent classification delves into citation intent in academic writing [12]. This study uses two datasets which are SciCite to train the unsupervised machine learning model and C2D for testing. The outcome of this study was the annotation of C2D dataset for its citation intent. This study experimented with Infersent, Glove, and BERT

embedding techniques and clustering machine learning techniques for citation intent classification. This study found that the BERT embedding technique provides semantic context and plays a huge role in predicting a citation intent classification. However, the evaluation of the study was limited to only statistical results and was not tested manually on new data.

## 3. METHODS

The framework of the intent classification is divided into seven phases as shown in Figure 1. The framework begins with a feasibility study to ensure that the research objective can be met. Once the research objective and boundaries are set, data preparation is commenced. It is then followed by text preprocessing and then feature extraction. The following phase is input representation and subsequently the classification phase. The final phase, which is the testing phase is where the research's finding is concluded.

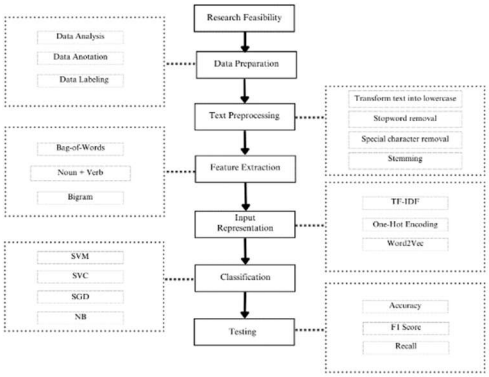


*Figure 1: Intent Classification Framework*

### 3.1 Data Collection and Annotation

Developing a new dataset for this research is crucial as there is no known publicly available dataset in the domain of English language usage questions by Malaysian professionals specifically for academic writers. In a study by [16], the LAIX corpus was developed specifically to cater Chinese learners to study English and accommodating to the learners' cultural understanding of the language. Providentially for this study, a dataset of archived text messages between a proofreader and a few academic writers is provided by a proofreading company and is available for use.

The raw data, however, was too noisy to work with as the majority of the questions were asked or answered with a sprinkling of the native language, Bahasa Melayu. Besides that, some questions posed as an open topic and will have challenges [17] such as needing to learn an adaptive decision boundary for the open topics, which deviates from the objective of this paper. Nonetheless, the raw data by the proofreading company provided solid examples of the typical questions asked by academic writers to a proofreader. Therefore, the raw data became a guide for the annotation of the new dataset.

Two annotators were tasked with constructing possible questions for 300 lines of sentences with grammatical errors. The sentences were also sourced from the same proofreading company. From the 300 lines of sentences, 876 pairs of questions and answers were generated. The annotated data was then reviewed by a certified translator. The steps to annotate the data is as shown below:

1. Review original sentence by academic writer e.g *The disadvantage is it is computational costly.*
2. Review corrected sentence by proofreader e.g *The disadvantage is its computational cost.*
3. Generate possible questions by academic writer e.g *What is the difference between it is and its? What other words can replace computational costly? How do I know when to use its or it is? Can I change computational cost to computationally expensive? I think the corrected sentence is wrong.*

Analysis was performed on the questions generated from reviewing the difference between the original sentence and the corrected sentence. One corrected sentence may have more than one correction hence more varieties of questions may be generated by annotators by a single line of corrected sentence. A sample of labeled dataset from one line of corrected sentence is as shown in Table 1.

*Table 1: Intent Category and Sample Data*

| Intent Category | Data |
|---|---|
| Information | What is the difference between it is and its? |
| Comparison | What other words can replace computational costly? |
| Confirmation | How do I know when to use its or it is? |

| | |
|---|---|
| Suggestion | Can I change computation cost to computationally expensive? |
| Request Feedback | I think the corrected sentence is wrong. |

The intent category is decided upon analysis of the questions generated by the annotators. It is found that there is a general tendency of the question prefix pattern that shows the possible intent of the question. Therefore, the questions' prefix become a guide to manually label the data into their intent categories. Table 2 describes each intent category and their general question prefix.

*Table 2: Description of Intent Categories*

| Intent Category | Description | Question Prefix |
|---|---|---|
| Information | Seeking for information on the correction made on the corrected text. | What does…?, When to use…?, Why…? How did/does…? How can I know…? |
| Comparison | Seeking for other alternatives instead of the corrected text. | What other…?, How can I make/change…? What else…?, Is there another…? |
| Confirmation | Grounding own comprehension on the corrected text. | Can I…?, Which one is…?, Is…?, Must I/we *verb*…? |
| Suggestion | Seeking for information by means of comparison between two or more terms. | Which one…?, Why is… more…?, What is… and…?, 'difference', 'between', 'than'. |
| Request Feedback | Seeking assessment on thought in regards to corrected text. | I think…, What do you think…? |

It should be noted that the annotators tasked to generating the possible questions were final year computer science students. Thus, limitation is shown in the lack of questions generated in other intent categories besides Information. This results in the number of lines for other categories being hugely imbalanced with the rest of the data. Consequently, the imbalance of Request Feedback intent category was too stark as shown in Figure 2. Consequently, it was decided that the category be dropped. The research moves forward with only four intent categories.

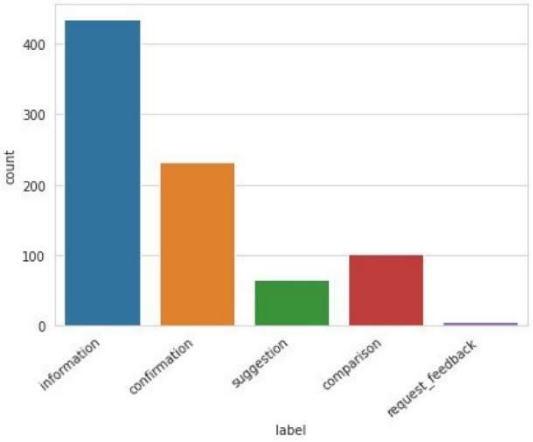


*Figure 2: Number of lines in each intent category*

### 3.2 Text Preprocessing

Text preprocessing consists of transforming input into lowercase and the removal of symbols and white spaces. Removing input that carries no meaning to the semantic of the data enables computation to be exclusively done on data that contributes directly to the intent classification. Besides the removal of stop words, removal of punctuation, words with less than two letters, and words with more than 21 letters are also done in text preprocessing. After all the unnecessary parts of an input are removed, the input goes through stemming.

Stemming is a text preprocessing technique that minimizes the inflection of a word. Normalizing words into their root forms removes the attributes carried by the words. Attributes such as grammatical or connotation create varieties of an individual word which may all imply the same meaning. Cleaning up the data off the unnecessary inflections is the reason stemming is one part of text preprocessing.

Text preprocessing allows the input to be at a bare minimum so that the intent classification model can run efficiently. Removing stop words such as 'I', 'me', 'the', 'a', 'of', 'about', and others allow the next steps to only capture the gist of the input. For example, an input 'Can you help with grammar?' will be simplified as 'help grammar'.

### 3.3 Feature Extraction

Feature extraction is important to improve the quality of machine learning classifier as not all input is needed for classification [18]. Three techniques are experiment in this research which are tagging the nouns and verbs from the input, the use of Bag-of-Words, and using bigram.

Tagging verbs and nouns as features as the dependency relationship between verbs and nouns can help identify the intent of the input [19]. For example, an input that contains the words 'change' (VB) and 'example' (NN) may indicate a suggestive purpose to 'replace' (VB) the word 'example' (NN) with a synonym such as 'instance' (NN). However, if 'the' (DET) is placed between 'change' (VB) and 'example' (NN), it may indicate a suggestive purpose to replace the specimen of the example.

Using Bag-of Words as feature assigns equal weights to each input. This way, all text input is considered important to determine intent classification. For example, the question "How do I use 'furthermore' in a sentence?" ensures that each word is taken as equally important regardless of their word class or a stop word.

The third technique used for feature extraction is bigram. Contrary to the Bag-of-Word technique, bigram considers the sequence of each input as important. This also helps to show if the question prefixes identified in Data Preparation phase is relevant for machine learning classification.

### 3.4 Input Representation

In the input representation phase, the input is vectorized using Term Frequency – Inverse Document Frequency (TF-IDF), or Word2Vec embedding, or One-Hot Encoding.

TF-IDF shows how rare or how common a word is among its corpus. The assumption made for TF-IDF is that it will help lower the value of frequently occurring words such as those defined as stop words. Thereby TF-IDF provides high value to important terms within its corpus besides forming its document representation.

Word2Vec provides association between the words in the dataset [19]. A defined relationship between the words in the corpus gives an upper hand to intent categories such as Comparison or Suggestion which inadvertently require semantic understanding. Word2Vec embeds the representation of input into a distribution.

One-Hot Encoding provides label values to data without semantic representation. However, it attributes the significance of a label from its frequency within its document. One-Hot Encoding transposes the representation of input into ones and zeros. This is particularly useful as the dataset in study is nominal data and encoding will disallow the machine learning model to assume a natural order between the intent categories.

### 3.5 Classification

Finally, the input is classified using four supervised machine learning techniques namely Support Vector Classifier (SVC), Support Vector Machine (SVM), Naïve Bayes, and Stochastic Gradient Descent (SGD). Each machine learning technique has different qualities that would process the input differently. The performance of the best model is measured based on the best accuracy from the machine learning models. These machine learning classifications have been used by previous related works such as the ones by [15], [20], and [21].

## 4. EXPERIMENT AND RESULTS

The performance of the intent classification model is tested using the dataset developed. The outcome of the intent classification models are as shown on Table 3. The best performing model is SVM with 0.89 accuracy. The winning combination uses no stop word removal for text preprocessing, extraction of noun and verb as features, and One-Hot Encoding for representation as shown in Table 3.

*Table 3: Intent Classification Model Performance*

| Classifier | NLP Technique | | | | | | | Results | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Text Preprocessing | Feature Extraction | | | Input Representation | | | Accuracy | Recall | F1 Score |
| | Stop word Removal | Noun + Verb | Bag-of-Words | Bigram | TF-IDF | W2V | 1HE | | | |
| **SVC** | | | y | | | y | | 0.88 | 0.87 | 0.86 |
| | | | | y | | y | | 0.83 | 0.81 | 0.82 |
| | y | | | y | | y | | 0.72 | 0.71 | 0.72 |
| | | y | | | | | y | 0.89 | 0.88 | 0.87 |
| | y | | y | | | y | | 0.73 | 0.71 | 0.69 |
| | y | y | | | | | y | 0.83 | 0.83 | 0.83 |
| | | | | y | | | y | 0.78 | 0.73 | 0.74 |
| | y | | | y | | | y | 0.76 | 0.75 | 0.76 |
| | | | y | | y | y | | 0.87 | 0.85 | 0.86 |
| | | | | y | y | y | | 0.87 | 0.83 | 0.85 |
| | y | | | y | y | y | | 0.78 | 0.74 | 0.76 |
| | | y | | | y | | y | 0.80 | 0.78 | 0.79 |
| | y | | y | | y | y | | 0.77 | 0.73 | 0.77 |
| | y | y | | | y | | y | 0.74 | 0.71 | 0.73 |
| | | | | y | y | | y | 0.76 | 0.74 | 0.75 |
| | y | | | y | y | | y | 0.73 | 0.73 | 0.73 |
| **SVM** | | | y | | | y | | 0.87 | 0.87 | 0.87 |
| | | | | y | | y | | 0.81 | 0.79 | 0.80 |
| | y | | | y | | y | | 0.79 | 0.78 | 0.78 |
| | | **y** | | | | | **y** | **0.89** | **0.88** | **0.89** |
| | | | | y | | | y | 0.79 | 0.75 | 0.77 |
| | y | | | y | | | y | 0.77 | 0.76 | 0.76 |
| | y | | y | | | y | | 0.74 | 0.72 | 0.74 |
| | y | y | | | | | y | 0.83 | 0.82 | 0.81 |
| | | | y | | y | y | | 0.88 | 0.87 | 0.88 |
| | | | | y | y | y | | 0.81 | 0.79 | 0.81 |
| | y | | | y | y | y | | 0.74 | 0.72 | 0.73 |
| | | y | | | y | | y | 0.81 | 0.80 | 0.79 |
| | y | | y | | y | y | | 0.80 | 0.79 | 0.79 |
| | | | | y | y | | y | 0.78 | 0.76 | 0.77 |
| | y | | | y | y | | y | 0.74 | 0.74 | 0.74 |
| | y | y | | | y | | y | 0.72 | 0.71 | 0.71 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | to be continued... | | | | | | | | | |
| | ...continuation | | | | | | | | | |
| **NB** | | | y | | | y | | 0.82 | 0.82 | 0.80 |
| | | | | y | | y | | 0.77 | 0.76 | 0.77 |
| | y | | | y | | y | | 0.70 | 0.72 | 0.71 |
| | | y | | | | | y | 0.87 | 0.86 | 0.87 |
| | y | | | | | y | | 0.66 | 0.66 | 0.66 |
| | y | y | | | | | y | 0.78 | 0.77 | 0.76 |
| | | | | y | | | y | 0.79 | 0.79 | 0.79 |
| | y | | | y | | | y | 0.73 | 0.71 | 0.72 |
| | | | y | | y | y | | 0.79 | 0.78 | 0.79 |
| | | | | y | y | y | | 0.80 | 0.80 | 0.80 |
| | y | | | y | y | y | | 0.76 | 0.74 | 0.75 |
| | | y | | | y | | y | 0.81 | 0.80 | 0.81 |
| | | | | y | y | | y | 0.77 | 0.75 | 0.76 |
| | y | | | y | y | | y | 0.71 | 0.69 | 0.70 |
| | y | | y | | y | y | | 0.74 | 0.72 | 0.73 |
| | y | y | | | y | | y | 0.69 | 0.68 | 0.67 |
| **SGD** | | | y | | | y | | 0.85 | 0.83 | 0.84 |
| | | | | y | | y | | 0.82 | 0.81 | 0.82 |
| | y | | | y | | y | | 0.67 | 0.66 | 0.67 |
| | | y | | | | | y | 0.89 | 0.87 | 0.86 |
| | y | | y | | | y | | 0.73 | 0.71 | 0.71 |
| | y | y | | | | | y | 0.82 | 0.81 | 0.80 |
| | | | | y | | | y | 0.80 | 0.80 | 0.80 |
| | y | | | y | | | y | 0.76 | 0.74 | 0.75 |
| | | | y | | y | y | | 0.88 | 0.86 | 0.85 |
| | | | | y | y | y | | 0.83 | 0.81 | 0.82 |
| | y | | | y | y | y | | 0.79 | 0.79 | 0.79 |
| | | y | | | y | | y | 0.78 | 0.76 | 0.77 |
| | y | | y | | y | y | | 0.75 | 0.74 | 0.72 |
| | y | y | | | y | | y | 0.72 | 0.71 | 0.72 |
| | | | | y | y | | y | 0.77 | 0.76 | 0.77 |
| | y | | | y | y | | y | 0.73 | 0.71 | 0.72 |

It is apt that the outcome of the study is not conventional due to the nature of the dataset. As shown from the results, the accuracy of the classifier improves when stop word removal is not applied. The definition of stop word is words that have high frequency in the English Language such as the, is, and to.

A Stop word List[3] constitutes of (but not limited to) prepositions, pronouns, and conjunctions. The importance of stop words in the proofreader chatbot intent classification is aligned with a study by [22]

[3] https://www.nltk.org/search.html?q=stop words

that found singular/plural forms and preposition as one of the top five most common grammar errors among Malaysian students. Hence, the corpus of the dataset has such inputs as questions from academic writers. Applying stop word removal to conform with the conventional text preprocessing phase would lose out on valuable input data for this study.

However, the same cannot be said with the use of noun and verb tagging for feature extraction. It is observed from the output that high accuracy results with noun and verb tagging only arise when in a set up that does not perform stop word removal. Alternatively, a set up without noun and verb tagging performs well when combined with W2V embedding.

In concern with the input representation, the best accuracy is derived from the use of One-Hot Encoding. Although shown in the study by [19] that the performance of word embeddings depends on the semantic similarity distribution of the dataset. This shows that the complexity of an embedding technique may not necessarily provide better intent classification outcomes. The better performance of One-Hot Encoding can also be said to be due to the compatibility of the data length with the representation.

Generally, the performance results show that SVM is the best machine learning classifier for this dataset. Other than that, the use of stop word removal has a negative impact on the intent classification. Furthermore, the combination of input representation in One-Hot Encoding provides the best combination setup for the intent classification framework in this study.

A detailed results of the best classifier model (no stop word removal+1HE+SVM) is as shown in Table 4. It shows that the intent category Confirmation has the best accuracy of 0.95 while intent category Suggestion has the lowest score of 0.75. This can be attributed to the fact that intent category Suggestion has a comparatively low amount of data for training in this dataset.

*Table 4: Detailed Results of Intent Classification Models*

| Intent Category | Accuracy | Recall | F1 Score |
|---|---|---|---|
| Information | 0.95 | 0.87 | 0.91 |
| Comparison | 0.81 | 0.88 | 0.85 |
| Confirmation | 0.92 | 0.92 | 0.92 |
| Suggestion | 0.75 | 0.67 | 0.71 |

As the intent classification model is built for a chatbot, the model is tested with input that is neither from the training data nor the test data. The new input query is typed in as a user naturally would test the intent classification model and the below outputs are observed. The input query used to test Information intent class is "i'm not sure how to use the". The input tested is not within the typical prefixes for class Information as stated in Table 3. However, the model can classify the input in Figure 3 into the correct intent class as Information.

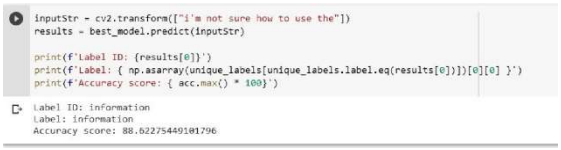


*Figure 3: Chatbot test using non dataset input (Intent: Information)*

Meanwhile in Figure 4, the non-dataset input tested includes a term that is not within the dataset i.e. *same*, *similar*. The input query used to test the Comparison intent class is "what is the difference between same and similar". Nonetheless, the model can classify the input into the correct intent class as Comparison.

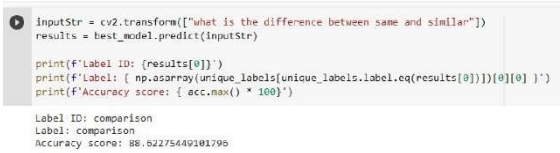


*Figure 4: Chatbot test using non dataset input (Intent: Comparison)*

Similar to the example in Figure 3, the non-dataset input tested in Figure 5 is not within the prefix listed under its intent class, Confirmation, as shown in Table 2. The input query used to test the Confirmation intent class is "are you sure i cant use furthermore here?". It is found that the model is able to classify the input into the correct intent class which is Confirmation.


*Figure 5: Chatbot test using non dataset input (Intent: Confirmation)*

Subsequently in Figure 6, the non-dataset input tested is not a term that is included within the dataset i.e. *adjoint*. The input query used to test the Suggestion intent class is "is there a different word for adjoint?". The output shows that the model can classify the input into the correct intent class which is Suggestion.


*Figure 6: Chatbot test using non dataset input (Intent: Suggestion)*

As shown in Figure 3 through Figure 6, it is important to test the trained data with new data. This is because a chatbot must be flexible in interpreting knowledge and command from user [10].

## 4. CONCLUSION AND FUTURE WORK

The outcome of this study shows that intent classification for a proofreader chatbot can have high levels of accuracy by applying different combinations of text preprocessing, feature extraction, input representation with machine learning technique. This study is able to come up with a model that performs at 0.89 accuracy.

Consideration taken about the dataset and the combination of techniques used gave a notable outcome to the performance of the intent classification model that was implemented and tested in this study.

Improvements are possible for future works by using a larger dataset and fully integrating the intent classification model into a proofreader chatbot to assist academic writers. In addition to that, it will be interesting to see if a dataset trained with a balanced amount of data for Suggestion and Comparison intent categories will improve the performance of the classifier model using Word2Vec embedding technique. This is due to the fact that these two intent categories rely more on the representation of semantic similarity which is available using Word2Vec.

Other than that, the existing dataset can be further enlarged more conveniently using unsupervised clustering techniques such as Large Language Models instead of manually labelling the annotated data. With a bigger dataset, a study on multiple intent per input instead of one intent per input can also be done in the future 23].

## ACKNOWLEDGMENT

This study was supported by Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia. We would like to thank the annotators of the dataset, Norfakhira Iman Mohd Roslan and Amirah Rasyidah Azhar from Universiti Tun Hussein Onn Malaysia.

## REFERENCES:

[1] Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: A systematic review. Vol. 2, Computers and Education: Artificial Intelligence. Elsevier B.V.; 2021.

[2] Kim NY. A Study on the Use of Artificial Intelligence Chatbots for Improving English Grammar Skills. Journal of Digital Convergence [Internet]. 2019;17(8):37–46. Available from: https://doi.org/10.14400/JDC.2019.17.8.037

[3] Ruan S, Jiang L, Xu Q, Liu Z, Davis GM, Brunskill E, et al. EnglishBot: An AI-Powered Conversational System for Second Language Learning. In: International Conference on Intelligent User Interfaces, Proceedings IUI. Association for Computing Machinery; 2021. p. 434–44.

[4] Omar N. Pengesanan ralat nahu esei Bahasa Inggeris melalui pemprosesan bahasa tabii Arabic Part of Speech Disambiguation View project Multi-lingual Sentiment Analysis View project [Internet]. 2009. Available from: https://www.researchgate.net/publication/235772409

[5] Darus S, Luin HW. Investigating teachers' use of computers in teaching English: A case study [Internet]. Available from: https://www.researchgate.net/publication/235772392

[6] Tian J, Tu Z, Wang Z, Xu X, Liu M. User Intention Recognition and Requirement Elicitation Method for Conversational AI Services. In: Proceedings - 2020 IEEE 13th International Conference on Web Services, ICWS 2020. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 273–80.

[7] Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications. Machine Learning with Applications. 2020 Dec;2:100006.

[8] Chandrakala CB, Bhardwaj R, Pujari C. An intent recognition pipeline for conversational AI. International Journal of Information Technology (Singapore). 2023 Feb 1;

[9] Hefny AH, Dafoulas GA, Ismail MA. Intent Classification for a Management Conversational Assistant. In: Proceedings of ICCES 2020 - 2020 15th International Conference on Computer Engineering and Systems. Institute of Electrical and Electronics Engineers Inc.; 2020.

[10] Muizzah Johari N, Nohuddin PN. Quality Attributes for a Good Chatbot: A Literature

Review. International Journal of Electrical Engineering and Technology (IJEET). 2021;12(7):109–19.
[11] Reddy S, Chen D, Manning CD. CoQA: A Conversational Question Answering Challenge. 2018 Aug 21; Available from: http://arxiv.org/abs/1808.07042
[12] Roman M, Shahid A, Khan S, Koubaa A, Yu L. Citation Intent Classification Using Word Embedding. IEEE Access. 2021;9:9982–95.
[13] Chen Q, Zhuo Z, Wang W. BERT for Joint Intent Classification and Slot Filling. 2019 Feb 28; Available from: http://arxiv.org/abs/1902.10909
[14] Shridhar K, Dash A, Sahu A, Pihlgren GG, Alonso P, Pondenkandath V, et al. Subword Semantic Hashing for Intent Classification on Small Datasets. 2018 Oct 16; Available from: http://arxiv.org/abs/1810.07150
[15] Schuurmans J, Frasincar F. Intent Classification for Dialogue Utterances. IEEE Intell Syst. 2020 Jan 1;35(1):82–8.
[16] Zhang H, Xu H, Lin TE. Deep Open Intent Classification with Adaptive Decision Boundary [Internet]. Available from: https://github.com/hanleizhang/Adaptive-
[17] Nogales RE, Benalcázar ME. Analysis and Evaluation of Feature Selection and Feature Extraction Methods. International Journal of Computational Intelligence Systems. 2023 Dec 1;16(1).
[18] Qiu L, Chen Y, Jia H, Zhang Z. Query Intent Recognition Based on Multi-Class Features. IEEE Access. 2018 Sep 8;6:52195–204.
[19] Ivan V, Carmona S, Riedel S. How Well Can We Predict Hypernyms from Word Embeddings? A Dataset-Centric Analysis. Vol. 2, the Association for Computational Linguistics.
[20] Muttaleb A, Zakaria LQ, Hasan AM, Qadri Zakaria L. Question classification using support vector machine and pattern matching Knowledge-Based Semantic Relatedness measure using Semantic features View project Arabic Text Analysis View project QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING. J Theor Appl Inf Technol [Internet]. 2016;20(2). Available from: www.jatit.org
[21] A COMPARATIVE STUDY OF THE ENSEMBLE AND BASE CLASSIFIERS PERFORMANCE IN MALAY TEXT CATEGORIZATION HAMOOD ALI ALSHALABI SABRINA TIUN NAZLIA OMAR. Available from: http://www.ftsm.ukm.my/apjitm
[22] Darus S, Subramaniam K. Error analysis of the written english essays of secondary school students in Malaysia: A case study [Internet]. Vol. 8, European Journal of Social Sciences. 2009. Available from: https://www.researchgate.net/publication/235772401
[23] Louvan S, Magnini B. Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. 2020 Nov 1; Available from: http://arxiv.org/abs/2011.00564