www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 873-880

e-ISSN : 2583-1062

Impact Factor : 5.725

DIALEXA AN ONLINE DIALECT CLASSIFIER USING NAÏVE BAYES ALGORITHM

Kirk Edja B. Accion[1], Karla Nadhine G. Enero[2], Christy Mar S. Orfano[3]

[1,2,3]College of Information and Computing (CIC), University of Southeastern Philippines (USeP), Davao City, Davao Del Sur, Philippines

[1]ORCID Number: 0009-0003-1981-3677

[2]ORCID Number: 0009-0007-8922-0198

[3]ORCID Number: 0009-0007-9062-0789

DOI: https://www.doi.org/10.58257/IJPREMS31350

ABSTRACT

Dialects are evolving through the years, and there has been a slight variation between dialects of the same language, such as the Visayan language. Because of dialectal variation, this led to dialect misclassification. This paper aimed to classify the dialect of a particular word or sentence using the Naive Bayes algorithm. This algorithm was used to classify the dialect by getting and comparing the probability of every dialect. The choices used for classification were Cebuano, Hiligaynon, and Waray. This paper also provided a repository for each dialect in which to preserve and evade dialect extinction. Consequently, using the Naive Bayes algorithm for classifying gave a high accuracy for dialect classification.

Keywords: Dialect Classification, Naive Bayes Algorithm, Machine Learning, Text Classification, Dialect, Visayan Language, Dialect Extinction, Dialectal Variation

1. INTRODUCTION

Dialect is assigned to languages for two principal reasons: linguistic and socio-political. Linguistic dialects are a variety of languages that are deemed mutually intelligible, meaning that speakers of one language can comprehend speakers of the other language without teaching the other. Socio-political dialects are languages that, for social, political, or cultural purposes, are less important than another, more normal language. [1]. Dialect is often used to characterize a way of speaking that differs from the standard variety of language. Dialectal variation refers to changes in language due to various influences. These include social, geographic, individual, and group factors. Dialects have a certain level of variation. The variability can be noted in distinct dialects and registers everywhere in English at all levels. Many researchers recognize that language variation involves variations that might have personal meaning, such as the speech behavior of certain social groups (communities) and socially significant elements of individual speaker performance [2].

Clopper and Pisoni performed an auditory-free classification experiment in which listeners in the United States were asked to classify speakers by regional dialect based on brief, phrase-long statements. Listeners are classified into two (2) - mobile listeners refer to those who have lived in more than one (1) region with varying dialects, and non-mobile listeners are those who have lived in one (1) region only with one known dialect. The speakers came from six United States dialect regions: New England, Mid-Atlantic, North, Midland, South, and West. The results indicated that the general output of listeners measured in terms of the precision of the classification of speakers based on the phrases spoken was only 28%. In addition, it was discovered that more talker groups were formed on average by mobile listeners than non-mobile listeners. Residential history refers to geographic mobility and region of origin impacts the classification outcome. Those with little experience with distinct regional varieties will have a higher percentage of dialectal classification failure. Mainly, the study implies that most respondents need help classifying the different talkers using a particular dialect [3,4].

According to the latest edition of the Ethnologue, the Philippines has approximately 187 dialects with different levels of dialectal variation, including Filipino Sign Language (FSL). The dialect variety of Tagalog and Kapampangan is very moderate. Dialects of Bicol, however, there is a considerable variation in the dialect. Bikol is an instance of a macrolanguage: a set of associated languages or dialects that must be strongly linked to each other and that some domain recognizes a single language identity. Some people consider the dialects spoken in the cities of Cebu, Iloilo, and Tacloban as dialects of the same language (they call this the Visayan language); others consider Cebuano, Hiligaynon, and Waray as three dialects belonging to the Visayan language. The different levels of dialectal variation challenge individuals in performing correct and reliable classification, especially if the variation between dialects is very moderate. Classifying dialects involved two primary methods - linguistic and traditional. The linguistics methods include Morphology, Phonology, Lexis/Semantics, Pragmatics, Syntax, Discourse Structure, and Graphology. At the same

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 05, May 2023, pp : 873-880

e-ISSN : 2583-1062

Impact Factor : 5.725

time, the traditional way was mainly based on reducing the integrated process of language use into subsets of discrete skills and areas of knowledge. It is essentially a functional procedure focusing on skills and areas of expertise in isolation [6]. In the linguistic approach, the classification of languages and dialects was through the comparison of linguistic descriptions and intelligibility [7]. Only a few have tried to employ the linguistic approach of identifying or classifying dialects [4]. The majority of the individuals, primarily those who have no prior knowledge and training in the linguistic approach, would likely use the traditional way, which was found to be not precise, reliable, and accurate.

In Davao City, the proponents conducted a survey to know the proficiency of the participants in the Cebuano dialect. The survey instrument was intentionally composed of 30 phrases or sentences from three different dialects, including Cebuano, Hiligaynon, and Waray. The results of the survey have shown that 71% of the respondents can correctly classify the sentences for Cebuano, while the majority of them also misclassified the sentences in Hiligaynon and Waray to Cebuano. The misclassification was attributed to the small variation between the identified dialects, thus making it more difficult for the respondents to provide the correct classification.

In addition, language preservation is important in one's culture. Grammar was presumed to be a fixed, unchanging scheme in the previous linguistic tradition. However, various authors and speakers use this scheme differently, which has led to the evolution and extinction of the dialect. Facts show that 183 dialects are living in the Philippines, 187 and 4 are extinct. Of the living languages, 41 were institutional, 73 developed, 45 were vigorous, 13 were in difficulty, and 11 were dying. [5]. Language is a significant component of any culture as it allows individuals to interact and express themselves. Future generations lose a crucial portion of culture when a language dies out, which is essential to fully comprehend it.

Due to the challenges mentioned, the researchers developed Dialexa: Online Dialect Classifier using Naïve Bayes Algorithm. The system provides a reliable classification of dialects employing the linguistic approach [8]. The system will accept text input representing a particular dialect and perform the classification using Naive Bayes Classifier. A repository of words for the three dialects was included to address the challenge of language preservation.

2. CONCEPTUAL FRAMEWORK

Figure 1 shows the input, process, and output of the proposed system. The system will use an internet connection to ease development from the ground up (scratch), and it will also be easier to improve and deploy. The user will input text into the system then it will be processed through NLP for word count on how many times each word from the input occurred to the database. After getting the word frequency, Naïve Bayes will take part and simplify the classification task. With the implementation of Naive Bayes, this will get the accuracy and compares the results of the 3 dialects and classifies the dialect of the inputted string. The following dialects are Cebuano, Hiligaynon, and Waray. The system will then show the output, which is the classified dialect.

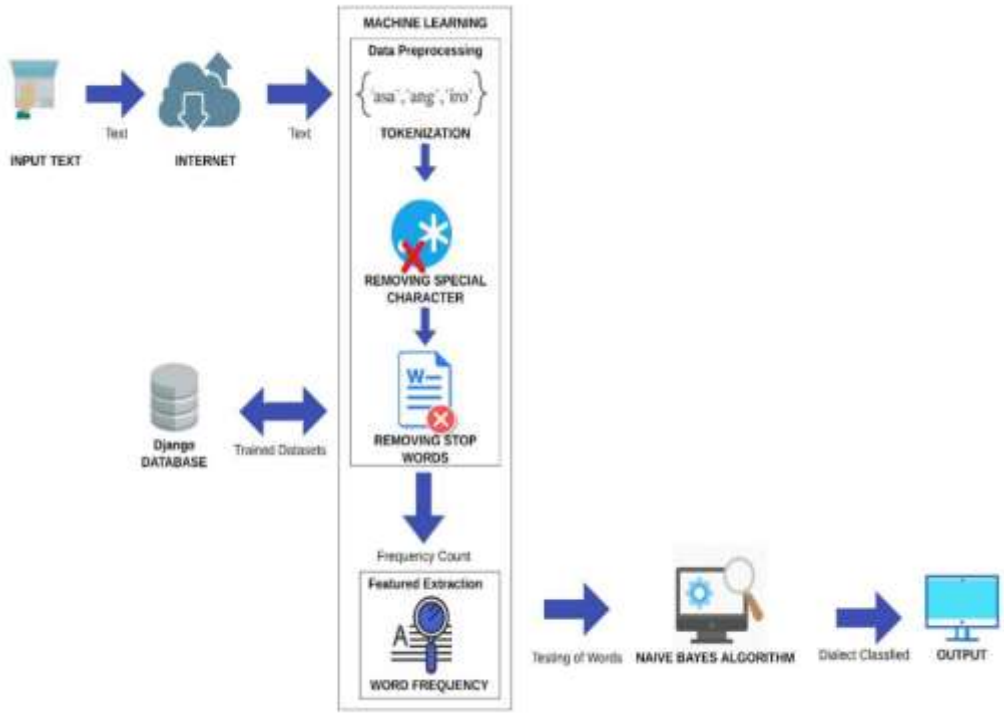

Figure 1: Conceptual Framework

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 873-880

3. METHODOLOGY

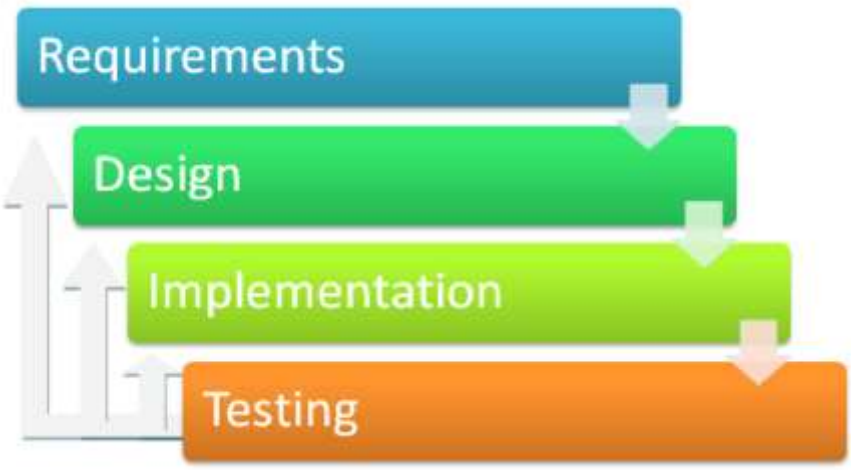

Figure 2: Methodology

Figure 2 shows that the researchers will use the Modified Waterfall Model in this study. The model uses the Software Development Life Cycle (SDLC) approach, which illustrates the software development process in a linear sequential flow. This means that any phase in the development process begins only if the previous phase is complete. Movement through this model is in phases, and the next phase cannot be undertaken without finishing the prior phase. In this model, the phases do not overlap. There are four phases of the Modified Waterfall model: Requirements, Design, Implementation, and Testing.

3.1 Requirements

Table 1 shows all the requirements captured and documented in this phase. The proponents need the dataset of all the words that belong to the three (3) selected Visayan Dialects, namely Cebuano, Hiligaynon, and Waray. The dataset serves as the reference for the classification. Below is a table showing the list of requirements of the system:

Table 1. Requirements of the System

Requirement ID	Description
R01	The system must be able to train the inputted word/sentence.
R02	The system must be able to classify the dialect of the inputted text.
R03	The system must be able to display the repository of the 3 Visayan dialects (Cebuano, Waray, and Hiligaynon).
R04	The system must be able to show the results of the searched word in the dictionary.

3.2 Implementation

In this phase, with the inputs from the system design, the system was developed in small programs called units, which were joined up or integrated into the next phase. All the requirements in the requirement table have been achieved in this phase. Each unit was developed and tested for its functionality, referred to as Unit Testing. Combining the input from system design, technical implementation began.

3.3 Testing

Once the prototype was developed. Testing will show up in this phase. Each unit is developed and tested for its functionality which is referred to as unit testing. This phase measures the quality, performance, and reliability of the prototype. The testing phase was also used to ensure that whatever errors or failures will happen or be detected, they can be fixed. The web-based dialect classifier was developed by combining the input from the system design, and technical implementation began. The researchers have tested and used constant software testing to find out if there are any flaws, code bugs, or errors.

3.4 Integration Testing

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 873-880

In this test, the trained datasets and the classifier were combined and tested. Integration testing involved the exposure of the faults in the interaction between the integrated units and the implementation of Naive Bayes. This testing involved debugging and looking for errors. This was done to ensure that the individual models of the system work as intended. The proponents performed tests that reflect under requirement ID R01 were met. The system can now input user values.

3.5 System Testing

Once the developed system was completed and well-integrated, the proponents conducted system testing. This test checked if the system was able to comply with the objectives that were set previously. The proponents performed tests that reflect under requirement IDs R02-R04 were met. The system can now display the repository of every dialect, display the output of the searched word, and classify inputted text.

4. RESULTS AND DISCUSSION

This section discusses how this paper met its set objectives.

The proponents used Machine Learning for training datasets and also for the inputted string in order to filter the stopwords, the most common words in a language, e.g., “sa”,” ang” etc., and other special characters that are not useful for the classification of. The process for training words/sentences started with the user input.

Figure 3 shows the process after the input; the entered string will be split word by word or converted into words (Tokenization). The unnecessary characters such as “!”, “?”,” *,” etc. will be removed as also the stop words. For the word frequency, the system sent a query to the database to count the occurrence of every word in the dataset. Last will be the implementation of Naive Bayes for the classification of dialects.

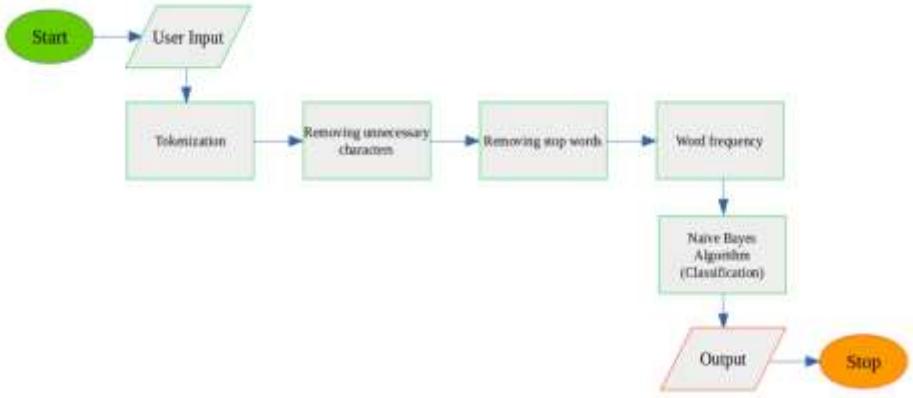

Figure 3: Training Process

Table 2 shows the data pre-processing; the sample input was “Palihug hinaya ang paghambal.” During the tokenization process, the input was split or converted into words {‘Palihug’, ‘hinaya’, ‘ang’, ‘paghambal’, ’.’}. After the tokenization, removing unnecessary characters takes part. Stop words have also been removed, such as ‘ang’, which belongs to the most commonly used word.

Table 2. Machine Learning Data Pre-Processing

ML Training (Data Pre-processing)	Sample Dialects
Input	Palihug hinaya ang paghambal.
Tokenization	{‘Palihug’, ‘hinaya’, ‘ang’, ‘paghambal’, ‘,’}
Removing of unnecesary characters	{‘Palihug’, ‘hinaya’, ‘ang’, ‘paghambal’}
Removing of stopwords	‘Palihug’, ‘hinaya’, ‘paghambal’}

Table 3 shows the final output from data pre-processing used in word extraction (word frequency). In this process, every word was sent to the database to count the occurrence of each word in every dialect. The first word sent was the word ‘palihug’ shows that Waray: 1; Cebuano: 1; Hiligaynon: 1; this means that ‘palihug’ exists in every dialect. The next word was ‘hinaya’, which shows that Waray: 0; Cebuano: 0; Hiligaynon: 1; this means that the word only exists in Hiligaynon. Last was the word ‘paghambal’; it also shows that Waray: 0; Cebuano: 1; Hiligaynon: 1.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 873-880

e-ISSN : 2583-1062

Impact Factor : 5.725

Table 3. Machine Learning Data Pre-Processing

Word Frequency			
palihug	hinaya	paghambal	**Dialects**
1	0	0	**Waray**
1	0	1	**Cebuano**
1	1	1	**Hiligaynon**

Table 4 shows that the proponents conducted multiple tests in order to determine if the Training Process of the system is functional. From the 10 tries conducted, 9 passed. The only system error during the process was when the inputted text did not exist in the database due to a lack of words in the dataset for each dialect. In conclusion, the training has 90% accuracy.

Table 4. Training (User Input) Process (Integration testing results)

Requirement ID	Description	Tries	Pass
R01	The system must be able to train the inputted word/sentence.	10	9

Classify dialect using the Naive Bayes Algorithm

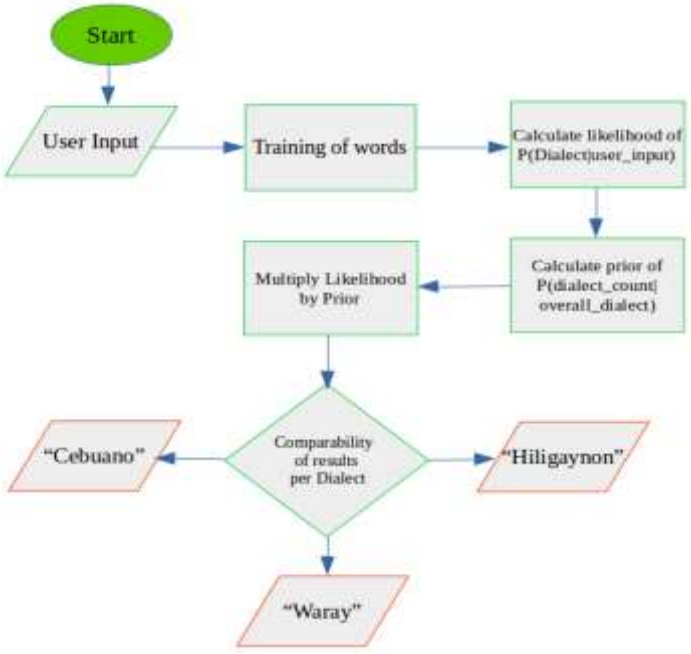

Figure 4. Dialect Classification Process

Figure 4 shows that proponents created a system using the Naive Bayes algorithm for dialect classification. The algorithm was used to find the probability of each dialect in a string, then compares the results and classify the dialect of the inputted string by the user. As the figure is shown from Figure (indicate number), after training the word/s from the user input, the algorithm for the classification is applied.

Table 5 shows that proponents conducted multiple tests in order to determine if the Training Process of the system was functional. From the ten tries conducted, 9 passed. The only system error during the process was when the inputted text did not exist in the database due to a lack of words in the dataset for each dialect. In conclusion, the training has 90% accuracy.

Table 5. Classification (System testing results)

Requirement ID	Description	Tries	Pass
R02	The system must be able to classify the dialect of the inputted text.	10	9

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 873-880

e-ISSN : 2583-1062

Impact Factor : 5.725

Table 6 shows sample testing results; it is necessary to convert the probability we wish to calculate into a form that can be calculated using word count. Naive Bayes is useful for dealing with conditional probabilities. Since the system will only try to find out which dialect has the highest probability, it makes sense to remove the divisor and compare only. First, the system calculates P(Dialect|user_input). In order to avoid wiping out all the information in the other probabilities, the proponents used a technique for smoothing categorical data to avoid having a probability of zero, called Laplace smoothing. Second, the system calculates P(dialect_count|overall_dialect), the probability of each dialect (Cebuano, Hiligaynon, and Waray). Third, it gets the product of likelihood and prior. Lastly, the system will collect the probability results of every dialect and compare them. The dialect that has the highest probability result will be the classified output.

Table 6. Sample Testing Results

Sample Dialects	Answer	Testing Result
Palihug hinaya ang pag-hambal	Hiligaynon	√
Hu-o, jutay lang	Hiligaynon	√
Palihog sang madali lang	Hiligaynon	√
Damo nga salamat	Waray	√
Tag pira ini?	Waray	√
Matud nila	Cebuano	√
Unsa Imung kahimtang?	Cebuano	√
Wala sang anuman	Hiligaynon	√
Diha sa pikas walo kilometro gikan diri	Cebuano	×
Di-in ka gaistar?	Waray	√

Figure 5 shows that the computational error during the process were computational bugs that caused the wrong output from the system.

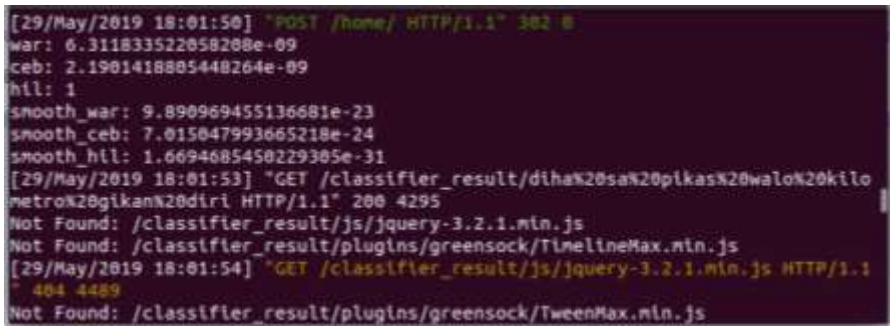

Figure 5. Dialect Classification Process

In the system, the proponents also created a dictionary as a repository for the three dialects: Cebuano, Hiligaynon, and Waray. The system also contains a search feature for word search, and the output will show the results for the searched word by the user together with its meaning and dialect.

Table 7. Sample Testing Results

Requirement ID	Description
R03	The system must be able to display the repository of the 3 Visayan dialects (Cebuano, Waray, and Hiligaynon).
R04	The system must be able to show the results of the searched word in the dictionary.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 873-880

e-ISSN : 2583-1062

Impact Factor : 5.725

Figures 6, 7, & 8, shown below, are screenshots of the repository of the three dialects.

Figure 6: Cebuano Dictionary

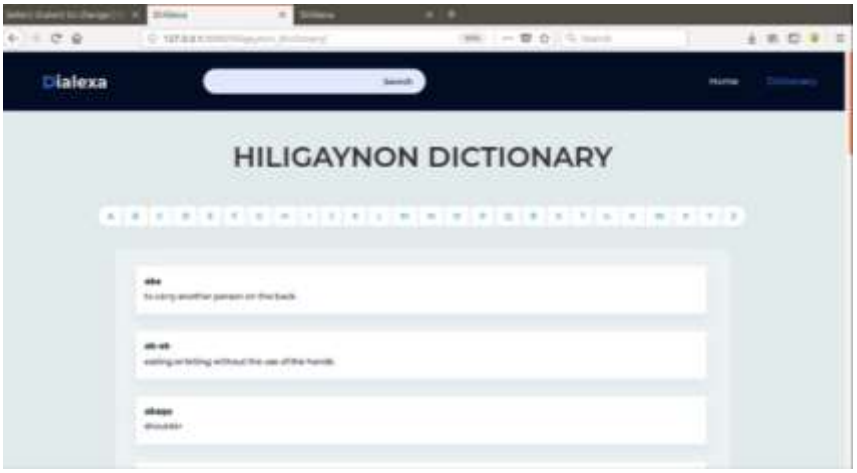

Figure 7: Hiligaynon Dictionary

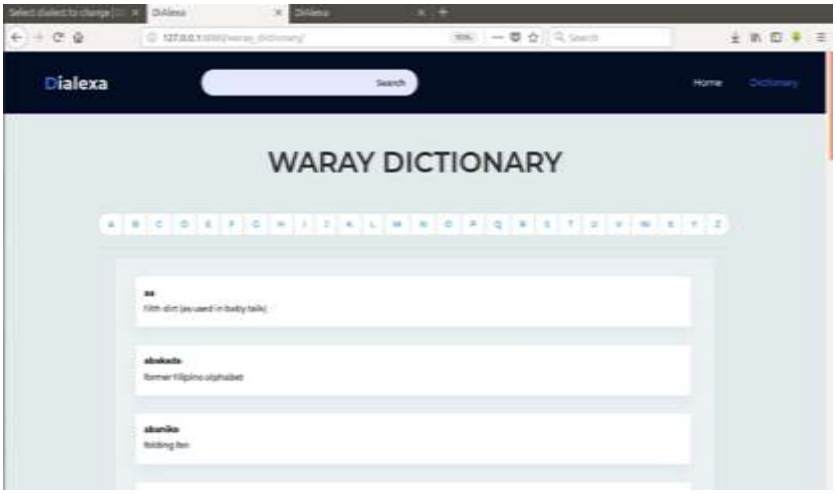

Figure 8: Waray Dictionary

5. CONCLUSIONS AND RECOMMENDATIONS

This section presents the conclusions drawn based on the output of the tests conducted on the system. This section will also include recommendations for future modifications that could be done for this study.

5.1 Conclusions

Based on the analysis of the test results that were conducted on the system, the proponents have concluded the following:

- The system could train words or phrases in Cebuano, Hiligaynon, and Waray dialects using machine learning. The training involves data pre-processing and word extraction, where an input goes through tokenization, and removal of special characters, figures, and stop words.

- The system could classify the dialect of words or phrases the user gives using the Naïve Bayes algorithm. The results showed 90% accuracy of classification. The margin of error is attributed to computational errors in which the total number of words of each dialect was uneven, and it only prioritized the first words in the database. Also, it was revealed that the number of words in the datasets is proportional and significant for achieving a higher accuracy rate.
- A repository of words in Cebuano, Hiligaynon, and Waray was successfully created. The repository kept a total of 23,210 words comprised of 19,525 Cebuano, 1,722 Hiligaynon, and 1,963 Waray.

5.2 Recommendations

To further improve this system, the proponents strongly recommend the following:

- Conduct more research about text classification to make the Naive Bayes Algorithm more accurate.
- Suggest Lemmatization and Stemming to achieve the root forms of the derived words. Lemmatization removes the inflection by determining the part of speech and utilizing a detailed dialect database. Stemming is also used to reduce the words into a root by removing inflection by dropping unnecessary characters, usually suffixes. Using these methods can help to improve the accuracy of dialect classification.
- Develop a mobile app version of the system to improve user engagement and accessibility. Also, to increase exposure across mobile devices.
- Add other dialects that also have dialectal variations with the three dialects used by the proponents to widen the scope of the system and for the user to be able to classify other dialects.

6. REFERENCES

[1] T. Mohana Priya, Dr. M. Punithavalli & Dr. R. Rajesh Kanna, Machine Learning Algorithm for Development of Enhanced Support Vector Machine Technique to Predict Stress, Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 20, Issue 2, No. 2020, pp 12-20

[2] Ganesh Kumar and P.Vasanth Sena, “Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit,” International Journal of Computer Science and Network Security, Vol. 15, issue 9, Sep. 2015, pp. 222-234

[3] Gyusoo Kim and Seulgi Lee, “2014 Payment Research”, Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[4] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, Hao Fu, “Financial Fraud Detection Fluid: Based on Random Forest,” International Journal of Economics and Finance, Vol. 7, Issue. 7, pp. 178-188, 2015.

[5] Hitesh D. Bambhava, Prof. Jayeshkumar Pitroda, Prof. Jaydev J. Bhavsar (2013), “A Comparative Study on Bamboo Scaffolding and Metal Scaffolding in Construction Industry Using Statistical Methods”, International Journal of Engineering Trends and Technology (IJETT) – Volume 4, Issue 6, June 2013, Pg.2330-2337.

[6] Boumová, Viera (2008), English Language and Literature, Traditional vs. Modern Teaching Methods: Advantages and Disadvantages of Each https://is.muni.cz/th/f62v8/MgrDiplomkaBoumova.pdf

[7] Tagalog: A History of the Language of the Philippines. Retrieve from: https://www.livinglanguage.com/blog/2014/11/25/tagalog-a-history-of-the-language-philippines/

[8] Accredited Languages Services (2019). Cebuano. Retrieve from: https://www.alsintl.com/resources/languages/Cebuano/