A Machine-Learning Based Text Classification and Machine Translator for Selected Under-Resourced Languages

Aniedi Bernard **Oboho-Etuk,** Patience U. **Usip,** Olufemi S. **Adeoye,** Ikechukwu **Ollawa**

Department of Computer Science, University of Uyo, Uyo, Nigeria

Abstract

Machine Learning, which aims at removing language barrier, uses the performance of computers to achieve efficient translation of any language text. For non-native Efik/Ibibio speakers, the cost of learning this language may be high and very difficult to reach the level of free communication. In this paper, we have used machine learning technique for translation of Efik/Ibibio to English Language. The Stochastic Gradient Descent approach is the machine learning algorithm adopted for this work. The resulting machine translator is able to translate text in selected under-resourced languages to English and vise-versa.

Keywords

Machine Learning, Text Classification, Machine Translator, Under-Resourced Languages

1. Introduction

Text classification is an analytical technique in machine learning in which a text is assigned to a predetermined label. A text classifier is used for labeling unstructured texts into already defined text groupings or categories. Orza [1], explains that users would have had to review and analyze vast amounts of information to understand the text context, whereas text classification helps one derive the relevant insights faster. Classifiers are algorithms at the heart of a machine learning process; they may be SVM (Support Vector Machine), Naïve Bayes, or even a Neural Network. They can also be defined as a 'collection of rules' on how you wish to group or categorize your data.

Under-resourced languages in this context are languages that have low-resource and are difficult to process, however there are well-resourced languages that can also be low-resource. Social media has been providing most of the text data researchers and developers often use for machine learning models. [2]. We need these languages to sustain the culture and history of their speakers and clearer understanding of concepts. It is estimated that of the roughly 7,000 languages spoken on the planet today, 50 to 90 percent are considered vulnerable to extinction by the end of the century. [3]. Hence, the need for classification and translation tools [4].

This project is aimed at designing a language classification tool that can identify selected under-resourced languages as well as provide a tool for interpretation into the English language. Hence, the encouragement needed for speakers to socialize, bond and educate themselves in their language

Joint Proceedings of Second International Workshop on Semantic Reasoning and Representation in IoT (SWIoT-2023) and Third International Workshop on Multilingual Semantic Web (MSW-2023), November 13-15, 2023, University of Zaragoza, Zaragoza, Spain

aniedibobohoetuk@uniuyo.edu.ng (Aniedi Bernard Oboho-Etuk)patienceusip@uniuyo.edu.ng (Patience. U. Usip);

Text data is a common communication type on social media and oftentimes we read through texts and wonder what language they are in. It is difficult to obtain insights from text data because they do not always come structured. The misclassification of these texts can be another major challenge; "is Annang a subset (dialect) of Ibibio or is it another language in its right?"

The dialects/languages, Ibibio, Efik, Annang, Oro and Ekid, have their variants in each community. Spoken Ibibio, Efik or Annang are well understood amongst all groups of speakers, however, they are different when presented in text format. The variants "esiere"(Efik), "asiere"(Ibibio) and "achiere"(Annang) should be properly identified and should return the same translation in the English language. The aim of this work is to develop a text classification and machine translator for selected under-resourced languages.

2. Related works

When working on Text Classification we assign predefined categories to free-text documents. According to Devopedia [5], Machine Learning is the main tool used to extract keywords from text and classify them into categories. Text classification can be implemented using the several supervised algorithms, Naïve Bayes, SVM and Deep Learning being common choices.

Text classification is mostly useful in Natural Language Processing (NLP), used for detecting spam, sentiment analysis, subject labeling or analyzing intent. Automating mundane tasks makes search, analysis and decision making faster and easier. For an effective Text classification process we need vast amounts of historical data, even though real time data is used to improve the model.

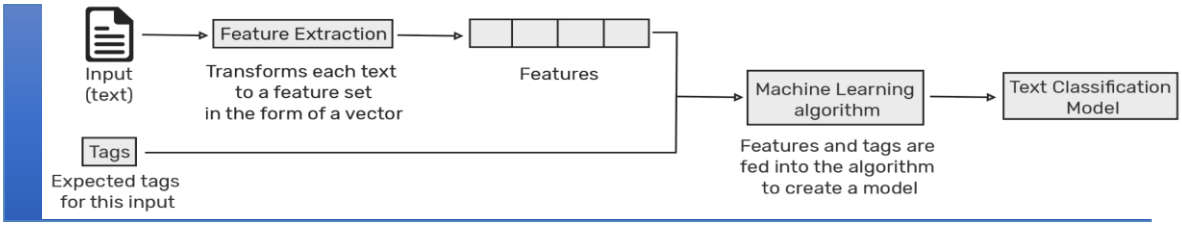

Input
(text)
Feature Extraction
Transforms each text to a feature set in the form of a vector
Features
Machine Learning algorithm
Text Classification Model
Tags
Expected tags for this input
Features and tags are fed into the algorithm to create a model

Figure 1. Text Classification Pipeline

Text classification process consist the following steps, Documents ↦ Preprocessing ↦ Indexing ↦ Feature Selection ↦ Classification Algorithm ↦ Performance Measure. [6]

There are about four approaches to machine translation;

a. Statistical Machine Translations: Ordóñez [7], wrote on Statistical machine translation, known as SMT or StatMT, as an approach to machine translation that yields the most probable output (translation) of each element that makes up a sentence. StatMT is based on the use of statistical models that analyze and search for relationships between two texts with the same content: one in the source language and the other in the target language.

   According to Koehn [8, 9], Statistical Machine Translation (SMT) can learn how to translate by analyzing existing human translations (which he calls bilingual text corpora). This, he said, is different from the Rules-Based Machine Translation (RBMT) approach that is usually word-based, most modern SMT systems are phrase-based and assemble translations using overlap phrases. The idea of using phrase-based translation is to reduce the limitations inherent in word-based translation by translating whole sequences of words, where the lengths may differ. Each sequence of words is called a phrase. Though they may not be linguistic phrases, but phrases found using statistical methods from bilingual text corpora.

   We can analyze bilingual text corpora (source and target languages) and monolingual corpora (target language) and generate statistical models that can transform text from one language to another by giving statistical weights that decide the most likely output of the text.

a. Rules-Based Machine Translation: Koehn also described the Rules-Based Machine Translation (RBMT) systems as the first commercial machine translation systems and are based on linguistic rules that allow the words to be put in different places and to have different meanings depending on the context. RBMT technology applies to large collections of linguistic rules in three different phases: analysis, transfer, and generation. The rules, he said, are developed by human language experts and programmers who have deployed extensive efforts to understand and map the rules between two languages. RBMT would rely on manually built translation lexicons, some of which can be edited and refined by its users to improve their translations.

b. Hybrid Machine Translation: the hybrid machine combines both the Statistical and the Rules-Based Machine Translations.

c. Neural Machine Translation: this applies neural networks in learning from existing translations and its previous translations to improve results without human inputs.

Stochastic Gradient Descent: Stochastic gradient descent is an optimization algorithm that uses a binary comparison approach to predict an output. It is often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. The results may not be accurate, however, it makes a powerful technique. Stochastic gradient descent described in equation (1). Considering the minimization of an average of functions:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} \ell_i(w),$$

w is a d-dimensional vector (or the feature dimension is d).

Minimizing the negative of a log-likelihood function of the full gradient descent is given in equation (1).

$$w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i\left(w^{(t)}\right). \qquad (1)$$

Assuming a computational cost O(dn). When reducing the cost, a subset of all samples is used to approximate the full gradient. The revised gradient descent step as given in equation (2).

$$w^{(t+1)} = w^{(t)} - \eta \cdot \nabla \ell_{I_t}\left(w^{(t)}\right), \qquad (2)$$

Let's say I_t is randomly chosen within {1, 2, ..., n} with equal probabilities. We can then have our stochastic gradient descent (SGD) with the computational cost of a single step now reduced to O(d).

In order to identify which language a text belongs to and also provide its translation we adopt the machine learning-based approach to our language classification and machine translator. This project will provide a simple tool for further works to document more texts in selected Akwa Ibom languages/dialects. There is a **rule-based** approach that can tell the system to classify text into a particular category based on the content of a text by using semantically relevant textual elements.

In our case, the **machine learning-based** system would learn the mapping of the input data (raw text) with the labels (also known as target variables). This is similar to non-text classification problems where we train a supervised classification algorithm on a tabular dataset to predict a class, with the exception that in text classification, our input data is raw text rather than using numeric features. Because we are working on classifying more than two languages, the Stochastic Gradient Descent classifier presents a good algorithm for multi-class classification despite being a binary classifier.

Why do we care? There is a need to have a record of all the world's languages and Akwa Ibom State being my homeland should be represented too. Language is one the the driving force of civilization, everything begins with communication; in 1997, Philip Parker provided a detailed statistical analysis of more than 460 language groups in 234 countries. He illustrated issues connecting linguistic cultures to nine areas of concern (which he listed as economics, cultural resources, demography) with key variables for each area (railways, water, telecommunications).

The significance of this project centers on the preservation of selected Akwa Ibom languages/dialects through deliberate inclusion in contemporary technology, contributing to linguistic research and improved translation to the English language. As stated above, this project will be valuable for the development of Akwa Ibom State by providing a tool which can be built upon for future indigenous language classification projects.

This project explores machine learning-based in attempting to classify selected Akwa Ibom languages and provide a translation tool for English language users. The Under-Resourced Languages used in this study are Ibibio, Efik, Annang and Oro dialects. The project will not cover the accuracy of the model used and data will not be fine-tuned for better performance.

- Collect sample data of selected texts and their translations written in the language of interest
- Develop a Stochastic Gradient Descent (SGD) algorithm to translate texts to English
- Use same SGD algorithm to classify the language each text belongs
- Launch algorithms as web application

3. Machine Translator

The machine translator is defined by following the steps described in the following sub-sections.

3.1 Data Collection

We would create a register of words, sentences in the selected under-resourced Akwa Ibom languages (Ibibio, Efik, Oro) and their English translations for the study.

The data will be stored in three columns and saved in a .csv (comma separated values) file format as represented in the table 1.

Table 1
Text Translation Table

Dialect	Translation	(language/dialect)
nta akara	i am eating akara	(generic Ibibio)
Ikpong	Cocoyam	(generic Ibibio)
Ye	And	(generic Ibibio)
Ye	With	(generic Ibibio)
Mme	And	(generic Ibibio)
Mme	With	(generic Ibibio)
Ekese	Several	(Efik-Ibibio)
Ekese	Others	(Efik-Ibibio)

Here is the link (https://docs.google.com/spreadsheets/d/1PgaKArdAFo-4fRTdbgCepZs4rudxp55wEw-nzluovl0/edit?usp=sharing) to our dataset. It is a table with three columns namely *dialect, translation and (language/dialect)* [11].

Dialect: this column contains texts read in Akwa Ibom dialects.
Translation: this column contains the English language translation of each text.
(language/dialect): this column contains the dialect in which the text was written in.

3.2 System Architecture

There are six key items in our system; User input, User interface, Classifier, Translator, Text Corpus and System Output. The User interface handles input and output, the processing unit handles the algorithms (classifier and translator) while the text corpus are stored in the system database.

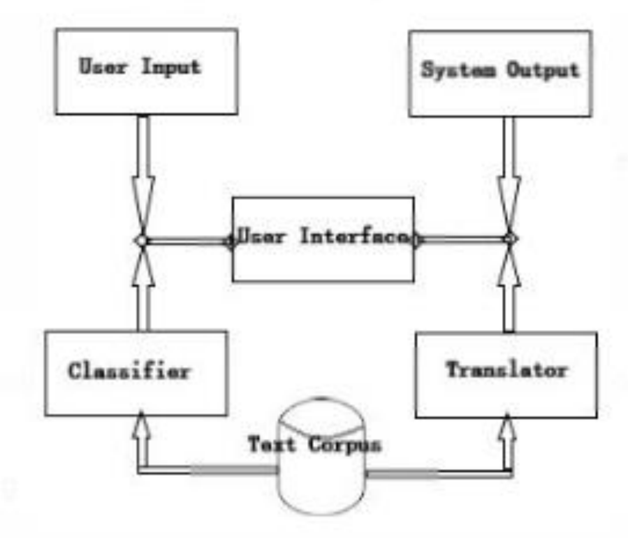


Figure 3. System Architecture

The system takes input through the User Interface, classifies and translates it by comparing with the Text Corpus in the database, then outputs the result through the user interface too.

User Input

The user inputs a text in any of the selected under-resourced languages for the system to provide a translation for the text

System Output

The system takes a user input and provides the most probable output form the text corpus (database)

User Interface

The user interface interacts with the user of the web application, it allows for inputs and outputs the result to the user.

Classifier

The classifier assigns the language label to the input text and returns an output on the user interface

Translator

The translator assigns the probable translation to the input text and returns an output on the user interface

Text Corpus

The text corpus is the database of texts written in the selected under-resourced languages with their English translations.

3.3 Python Scripts

Three python scripts were written: the *classify.py, ibom.py* and *ibibio_app.py*.

Classify.py:

The classify.py file identifies which dialect the text reads in.

Ibom.py:

The next python script is the *ibom.py*, and it follows a similar algorithm as the *classify.py*. The *ibom.py* is for text translation, translating the text into English language.

Ibibio_app.py:

The *ibibio_app.py* is designed to run as an interactive Streamlit application.

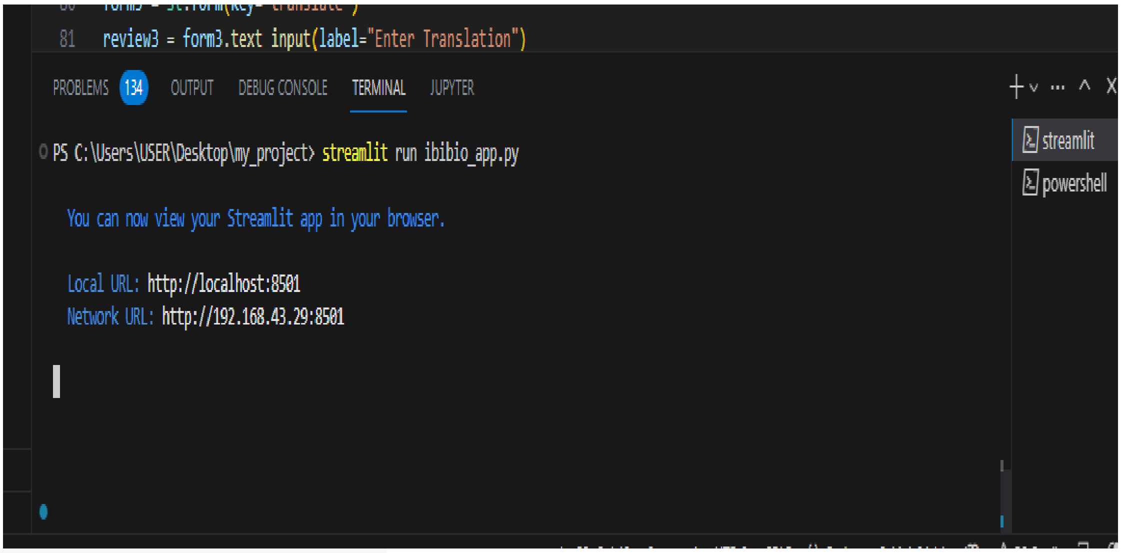
4. System Implementation

There are two approaches to launching the Web App; the local and the remote approaches. To launch locally one will need a computer with at least 2 gigahertz processing speed and 6 gigabyte RAM. If one wants to launch remotely, login to github.com and fork the repo https://github.com/imanibom/Akwa-Ibom-Language-Classification-and-Machine-Translation and open the github workspace and run the command **streamlit run ibibio_app.py** (remember to press enter).

Launching the Streamlit Application Locally

The dataset (ibibio.csv) and the three python scripts (*classify.py, ibom.py* and *ibibio_app.py)* must be in the same folder (you can create an empty folder for this). Then follow the following steps;

1. Go to the Start Menu on your computer and launch the Command Prompt (or GitBash, Anaconda Prompt etc.)
2. Navigate to the file location
3. Type the command **streamlit run ibibio_app.py** and press enter



```
review3 = form3.text_input(label="Enter Translation")
```

```
PS C:\Users\USER\Desktop\my_project> streamlit run ibibio_app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.43.29:8501
```

Figure 4. Input Screen Screenshot

Running the script in figure 4 will cause the interface to load on your browser page, ready for use!

Web App Interface

The web app interface has the following features;

- Welcome message: “You are Welcome. Enter Akwa Ibom Word or Sentence of Your Choice”
- User Input: Textbox
- Translation: Output
- Dialect: Classifies word into a language group (dialect)

5. Conclusion

Language classification and machine translation are large fields in machine learning. Several large language models (LLM) have been developed to handle specific areas ranging from machine virtual assistance to machine translation. Haystack, Langchain, Bert, Chat-GPT etc., provide system ready platforms to launch LLM apps, however in this project we developed a simple machine translation and language classification app for selected under-resourced Akwa Ibom languages. In this research we have created a web application where we can both classify and translate texts written in selected Akwa Ibom dialects. This would also open up further AI researches such as,

- Real-time machine translation
- Audio machine translation etc.

With the 10th generation laptop computers, it is possible to work on large datasets and use text annotation techniques to scrape text documents. For further works I would recommend;

a. Implementing text scraping and annotation for the dataset
b. Linking to a databank of Akwa Ibom language texts
c. Implementation of Web Scraping tools on online and social media posts written in Akwa Ibom dialects

References

[1] Orza, P. (2022). Text Classifiers in Machine Learning: A Guide. levity.ai

[2] Firsanova, V. (2022). A Quick Guide to Low-Resource NLP. mlops.community

[3] Riehl, A. (2019). Why are Languages Worth Preserving. sapiens.org

[4] Usip, P. U., & Ekpenyong, M. E. (2018). Towards Ontology-Driven Application for Multilingual Speech Language Therapy. *Human Language Technologies for Under-Resourced African Languages: Design, Challenges, and Prospects*, 85-101.

[5] Efik Language Encyclopedia, Science News & Research Reviews: Classification of Language

[6] Rani, N., Sharma, A. & Pathak, S. (2018). Text Classification Using Machine Learning Techniques: A Comparative Study, pp. 551-555

[7] Ordóñez, M. B. (2022). What Is Statistical Machine Translation?. Blog.pangeanic.com

[8] Koehn, P. (2023). What is Statistical Machine Translation (SMT)?. Omniscien.com

[9] Koehn, P. (2023). What is Rules-Based Machine Translation (RBMT)?. Omniscien.com

[10] Tibshirani, R. (2018). Stochastic Gradient Descent. www.stat.cmu.edu

[11] Oboho-Etuk, A. (2022). South African Language Identification. Kaggle.com
https://github.com/imanibom/Akwa-Ibom-Language-Classification-and-Machine-Translation