Team Binary Brigade

Team Leader - Phyo Kyi
Team Member - Thant Thiri Maung

Distinguishing Burmese Male and Female Names

Team Binary Brigade

Our team

Phyo Kyi 0

Team leader

Thant Thiri Maung 1

Team Member

About Project

Our Purpose

To classify Myanmar Names by sex (male or female or both).

Impact

Useful for Myanmar Name Entity Recognition

Motivation

This can be great impact in Myanmar NER and this become our motivation

Speaker - 1

AI Ethics

- **Fairness**
  - Male and female name size nearly equal.
  - Not only on emphasize the Burmese ethnic group, but other also
  - No left for the names which can be both
- **Reliability and Safety - 93.08176100628931% accuracy score**
- **Privacy and Security - We only take name data**

Speaker - 1

AI Ethics

- Inclusiveness
  - Work on some ethnic names (Shan , Kayin), plan to include others
  - Replacing similar words help user wrong input to correct
- Transparency
  - can use Myanmar and English Language
  - Most common burmese names only
  - Unicode Only
- Accountability
  - Public Github Repo
  - Google Form for New Names

Speaker - 1

Data Collection

- **From Social Media, from GitHub, other sources like articles**

GitHub - L16H7/Myanmar_Names: Open Source List of Myanmar(Burmese) Names for Male & Female Names

- **Total Data - 6357, Male = 3191, Female = 3172**
- **English > Burmese Name changing**

https://docs.google.com/spreadsheets/d/1FRWeY1QMEsyDTOGjk4Jj5hGXQekQb0-uURDbuAKe3Cs/edit#gid=1461949965

Data Encoding

Give the label male as 0 and female as 1.

Data for Rule based Methods

- Special female two words name (eg, သီရိ, သဇ္ဇာ, သီဇ) > 105
- Special male two words name (eg, သီဟ, သူရ) > 5
- Both words (eg, သန်းဆွေ, သန်းမြင့်အောင်) > 91

Data Preparation

Tokenization

- By Segment ['ခိုင်', 'သန္တာ', 'ထွန်း']
- By Character ['ခ', '□', '□', 'င', '□', 'သ', 'န', '□', 'တ', 'ာ', 'ထ', '□', 'န', '□', 'း']
- By Syllable Tokenization ['ခိုင်', 'သ', 'န္တာ', 'ထွန်း']
- By Multilingual Semi Syllable Break ['ခို', 'င်', 'သ', 'နှ', 'တာ', 'ထွ', 'န်း']

Sources
https://github.com/swanhtet1992/ReSegment/blob/master/resegment.py
https://github.com/SaPhyoThuHtet/nlp-tool/blob/main/utilities.py

Replacing similar words

```
def clean_text(text):
    text = text.replace("\u200c","")
    text = text.replace(" ","")
    text = text.replace("ျ","ြ")
    text = text.replace("ည်","ီ")
    text = text.replace("ော်","န်")
    text = text.replace("မ်","န်")
    text = text.replace("ါ","ာ")
    text = text.replace("အူး","ဦး")
    return text
```

Data Splitting

- **train_test_split**

```
max_i=0
max_score=0
for i in range(10,70):
    x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = random_size,random_state=i)
    classifier = LogisticRegression()
    if model == "mf_logistic_regression_by_character_tokenization":
        classifier = LogisticRegression(max_iter=5000)
    classifier.fit(x_train, y_train)       #training level
    score = classifier.score(x_test, y_test)
    if score > max_score:
        max_i=i
        max_score=score
        max_classifier = classifier
        max_x_train, max_x_test, max_y_train, max_y_test = x_train, x_test, y_train, y_test
        max_y_pred = classifier.predict(max_x_test) # Predit Y Value with Test Datasets
```

Understanding on Algorithm

Rule based Systems

Logistics Regression

Dictionary

Logistics Regression

- Use in Binary/Categorical Classification
- Logistics Regression use Logistic function/ Sigmoid function

Formula

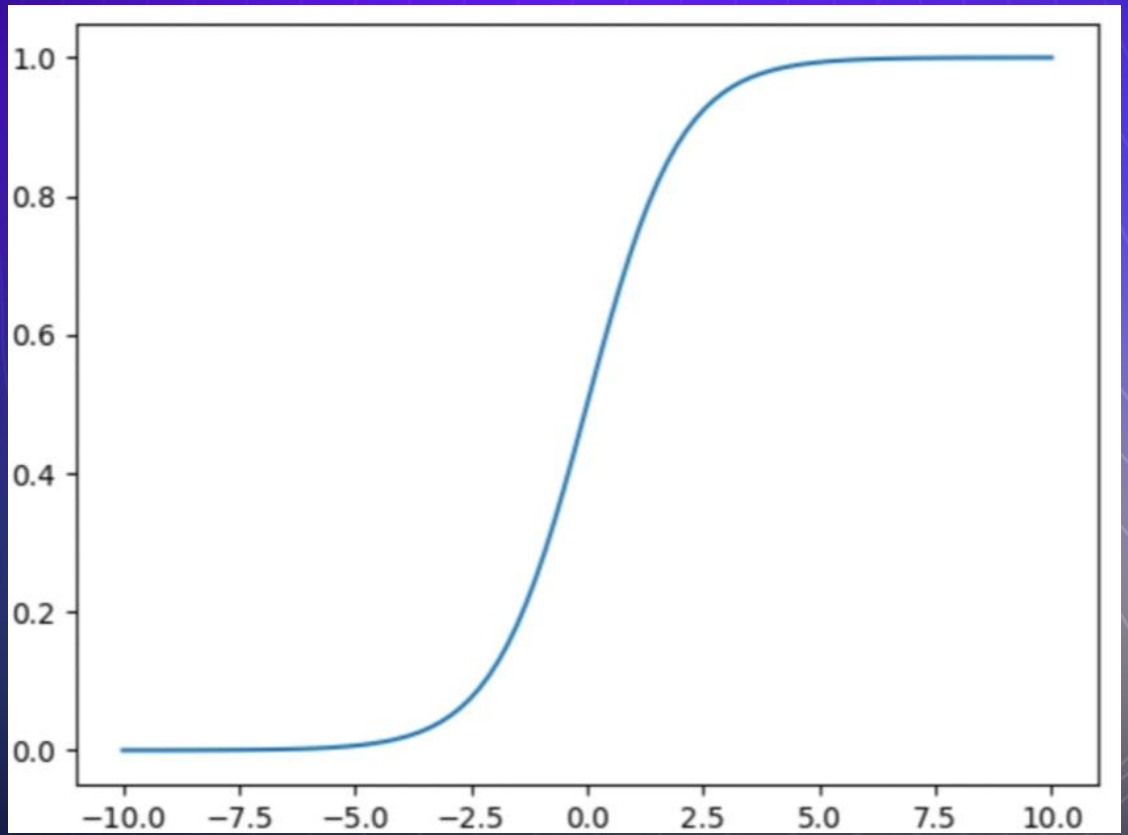
$$S(x) = \frac{1}{1 + e^{-x}}$$

$S(x)$ = sigmoid function

e = Euler's number

Speaker - 1

S-shape curve



Input values between minus and plus infinity, only values between 0 and 1 result.

Then, Sigmoid function use threshold value which can decide that a given input belongs to what type of two classes.

		ခင်	ဖိုး	မိုး	ဝင်း	နိုင်	စု	sex
ခင်ဖိုးမိုးဝင်း	female	1	1	1	1	0	0	1
ဖိုးမိုးနိုင်	male	0	1	1	0	1	0	0
မိုးမိုးစု	female	0	0	2	0	0	1	1

Syllable 1		Syllable 2		Syllable 3		sex
Male %	Female %	Male %	Female %	Male %	Female %	
13	100	30	50	20	10	1
100	20	30	20	0	0	0
40	100	40	100	10	100	1

Speaker - 0

Model and Results

By Character Tokenization - 86.949685534591 2%

By Multilingual Semi Syllable - 93.081761006289 31%

By Segment - 92.45283018867924%

By Syllable Tokenization - 93.081761006289 31%

Speaker - 0

Experiment

```
Clean Data is finished.
Word Columns by Segments is finish
Word Columns by Syllable Tokenization is finished.
Word Columns by Multilingual Semi Syllable Break is finished.
Word Columns by Character Tokenization is finished.
MF leading exclusion list is finished.
================================================
By Segment
================================================
Total is 6358. Male is 3189. Female is 3169
Model accuracy score of random state 11 is :  92.45283018867924
[[295  16]
 [ 32 293]]
================================================
By Syllable Tokenization
================================================
Total is 6358. Male is 3189. Female is 3169
Model accuracy score of random state 65 is :  93.08176100628931
[[308  24]
 [ 20 284]]
================================================
By Multilingual Semi Syllable Break
================================================
Total is 6358. Male is 3189. Female is 3169
Model accuracy score of random state 54 is :  93.08176100628931
[[304  17]
 [ 27 288]]
================================================
By Character Tokenization
================================================
Total is 6358. Male is 3189. Female is 3169
Model accuracy score of random state 54 is :  86.94968553459912
[[276  45]
 [ 38 277]]
```

```
=====test and generate for model segment=====
Right is 6056
Wrong is 304
Percentage is 95.22012578616352

Wrong Method Count
{0: 0, 1: 0, 2: 0, 3: 0, 4: 0, 5: 304, 6: 0}
Right Method Count
{0: 0, 1: 1152, 2: 31, 3: 1008, 4: 2752, 5: 1026, 6: 87}
================================================
=====test and generate for model syllable=====
Right is 6100
Wrong is 260
Percentage is 95.9119496855346

Wrong Method Count
{0: 0, 1: 0, 2: 0, 3: 0, 4: 0, 5: 260, 6: 0}
Right Method Count
{0: 0, 1: 1152, 2: 31, 3: 1008, 4: 2752, 5: 1070, 6: 87}
================================================
=====test and generate for model multi=====
Right is 6246
Wrong is 114
Percentage is 98.20754716981132

Wrong Method Count
{0: 0, 1: 0, 2: 0, 3: 0, 4: 0, 5: 114, 6: 0}
Right Method Count
{0: 0, 1: 1152, 2: 31, 3: 1008, 4: 2752, 5: 1216, 6: 87}
================================================
=====test and generate for model character=====
Right is 5494
Wrong is 866
Percentage is 86.38364779874213

Wrong Method Count
{0: 0, 1: 0, 2: 0, 3: 0, 4: 0, 5: 866, 6: 0}
Right Method Count
{0: 0, 1: 1152, 2: 31, 3: 1008, 4: 2752, 5: 464, 6: 87}
================================================
```

- Logistics Regression with Tokenization: Character, Syllable, Segment, Multi Syllable
- Rule Base (Special Word and Leading Two Words) and Dictionary
- Logistics Regression weight value on Syllable
- Tensorflow Neural Network

Speaker - 0

Data for Rule based Methods

- M6 : Both words (eg, သန်းဆွေ, သန်းမြင့်အောင်) > 91
- M0 : Special List
- M1 : Special female two words name (eg, သီရိ, သဥ္ဇာ, သီဂီ) > 105
- M2 : Special male two words name (eg, သီဟ, သူရ) > 5
- M3 : Leading words for female
- M4 : Leading words for male
- M5 : from Model

Speaker - 0

Streamlit Demo

https://myanmar-names-male-female.streamlit.app

Github

https://github.com/binarybrigade/myanmar-name-mf-classification

Speaker - 0

Thank You

Any Questions?