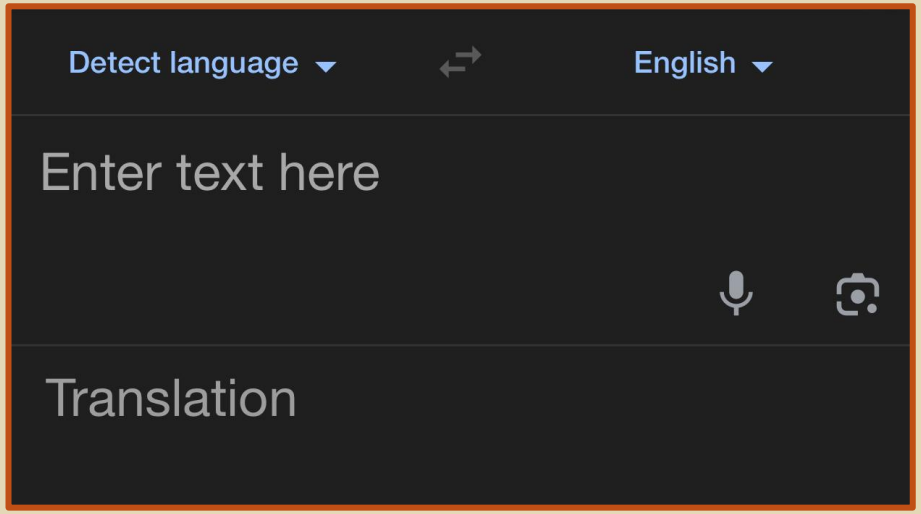
2024 AI Final Project

- Language Detection -

Group 37 – 111550037 嚴偉哲
111550117 黃毓為
111550127 郭穎達

Why?

- Translate an unknown language
- Similar languages (e.g. Spanish vs Portuguese)
- Chatbot generating response



Overview

- Given a line of text, determine the language it is written in
- Only a basic dataset with sentences and their languages is allowed to use in training

Goals

- Using classification models and techniques to detect the language

- Get a 90%+ accuracy for our model

[1]

Any reference regarding this page is shown here ↑

Related work

Language Detection For Short Text Messages In Social Media

Ivana Balažević[1], Mikio Braun[1], Klaus-Robert Müller[1, 2]

1 Machine Learning Group, Technische Universität Berlin, Berlin, Germany
2 Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

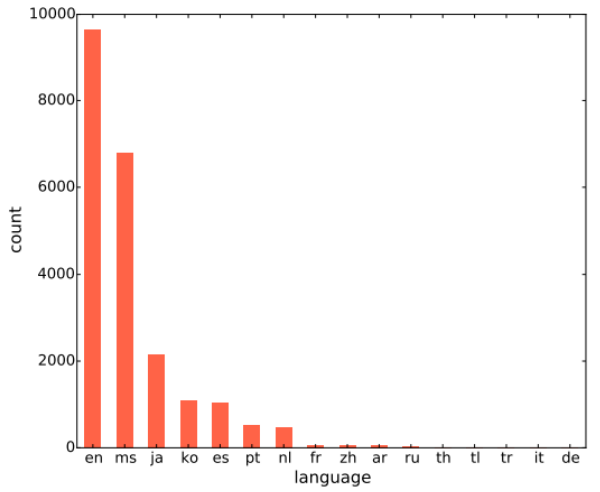


Figure 1: Distribution of languages in the dataset

- Uses SVM, Logistic Regression as the base models
- A much unbalanced language distribution

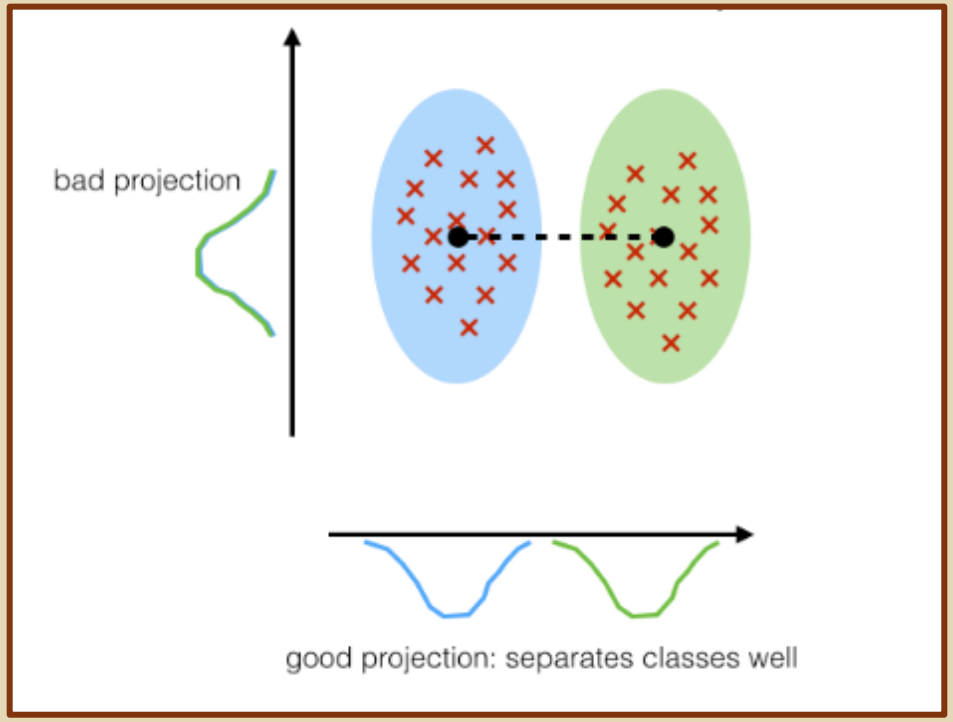
Dataset

- 20 languages, 70000 rows of data

labels string · *classes* 20 values	text string · *lengths* 2 2.42k
pl	Biały kot stojący na dywanie.
sw	Speaking injeera , anasema williams , hii ni sumbua
fr	J'ai acheté cette rallonge pour l'utiliser avec l'Oculus Rift S. Malheureusement, le débit est insuffisant et perturbe beaucoup trop le suivi des mouvements du casque (nombreuses saccades).
pl	Dwóch hokeistów walczących na lodzie.
nl	Een man maait een gazon.
ja	ソニーのイヤホンが断線してきてるようなのでウォークマンA50用に初めての中華イヤホンを購入です。 気になる点はKZ SE3なのにKZ ZSRのイヤホン ケーブルが入ってましたが(謎)断線したときのためにケーブルをもう一本買ってたのでそっちの方にリケーブルしました。 (イヤホン本体には問題はないようなので交換や返品や返金は望みません)
ar	- و تقسيم ثلاث دورات القانون الجنائي التركي في الاخاديد , والمرتفعات الجنوبية , والممار الرئيسي , وهو القانون الجنائي التركي في سمرلين .
el	Ένα παράδειγμα είναι η νομοθεσία του 1981 που εδραίωση πολλές μικρές κατηγορηματική επιχορηγήσεις σε μεγαλύτερες επιχορηγήσεις , τα κονδύλια για τα οποία θα μπορούσαν να δαπανηθούν με πολύ ευέλικτο τρόπο .
ru	Им нравится в , низким атмосфера и галльского стиль этого места .
pt	Novo Assassino em Série do Medo Francês após Assassinatos
el	Εδώ ήταν ένας ανθρωπιστής της Αναγέννησης που θεωρούσε τους συνανθρώπους του ηλίθιους και παράφρονες , σάκους για φαγητό , και πληρωτικά λεκάνες , και απολάμβανε τη σκέψη της καταστροφής της ανθρωπότητας σε έναν παγκόσμιο κατακλυσμό .

Baseline 1 – LDA (Linear Discriminant Analysis)

Find a projection such that points in different classes separate the most



[3][4]

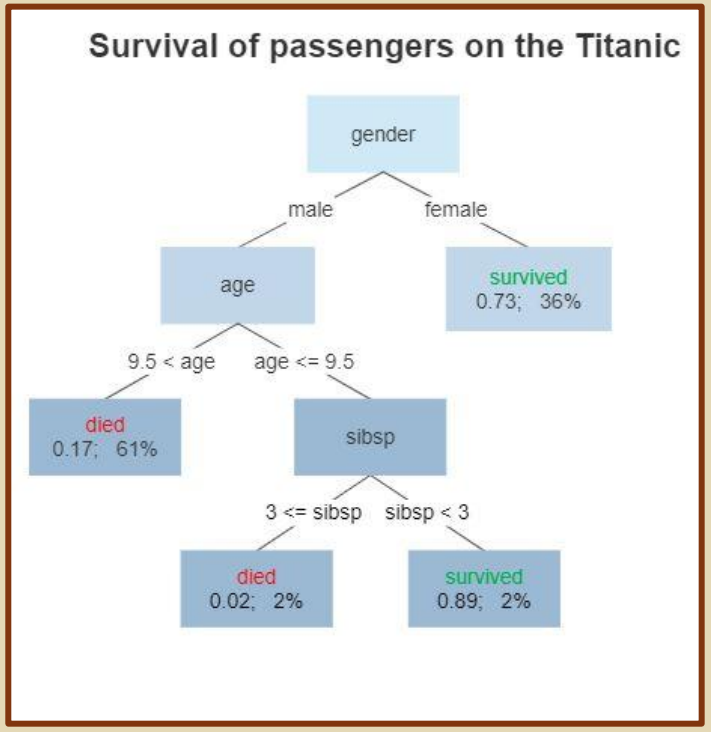
Baseline 1 – LDA (Linear Discriminant Analysis)

- LDA assumes the data in each class is normally distributed
- Use QDA (Quadratic Discriminant Analysis) instead

Baseline 2 – Decision Tree

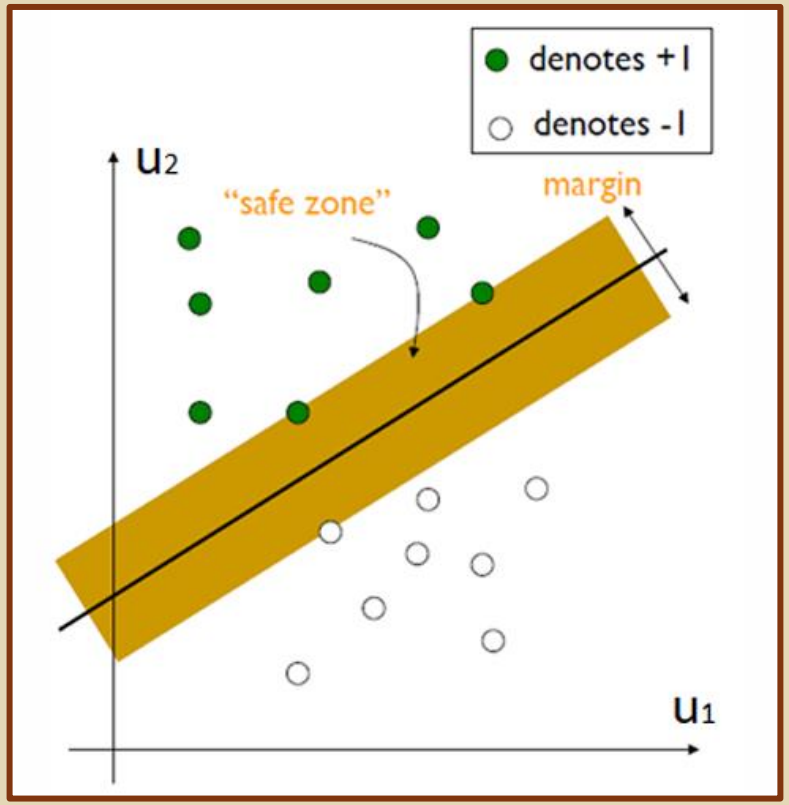
Taking multiple decisions based on variables

- One of the more interpretable method



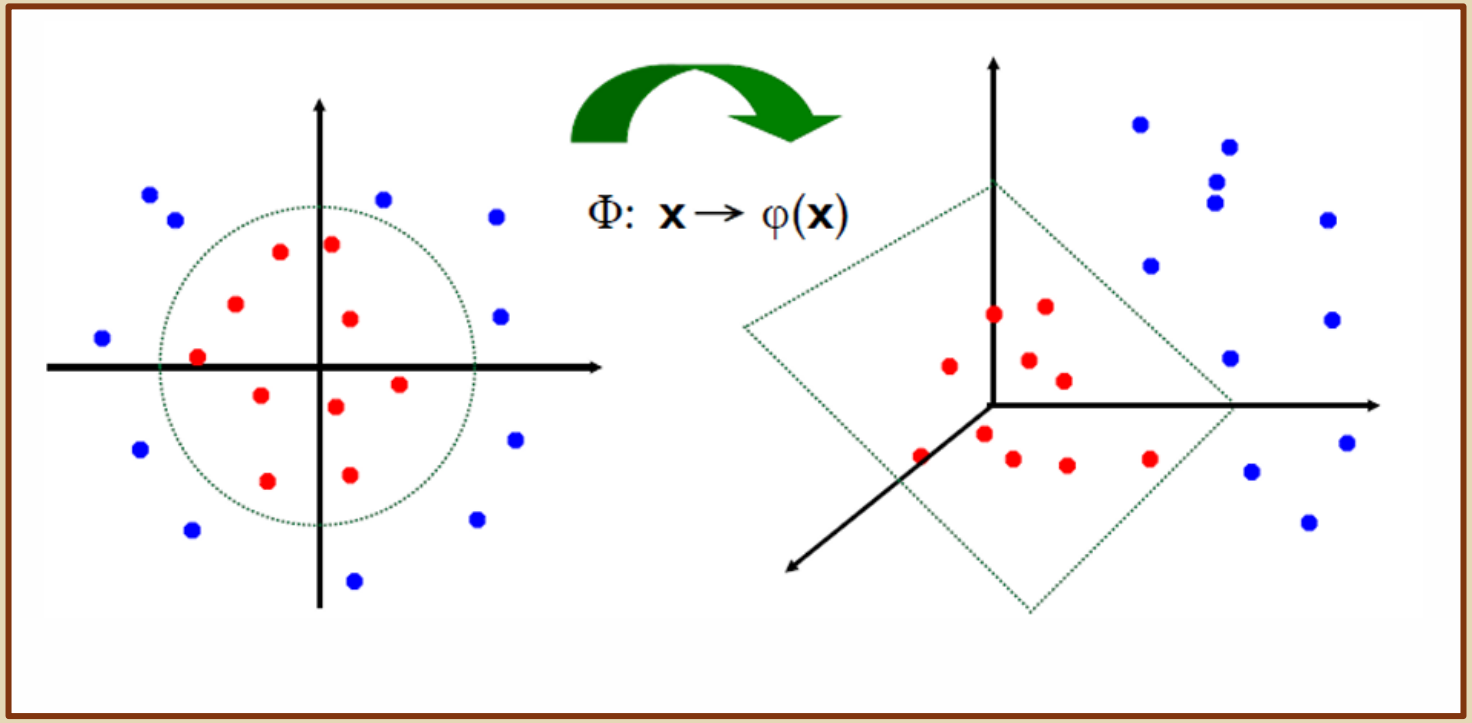
Baseline 3 – Support Vector Machine

Find a line that separate two classes with the largest margin



Baseline 3 – Support Vector Machine

- Can also deal with nonlinear classification



Step 1 – Feature Selection

Method A: Dictionary Search

- Store all discovered letters in a dictionary
- When given a text later on, search the dictionary to see if the letter is in there

Language identification is the problem of…

Four scores and seven years ago our fathers…

The quick brown fox jumps over the lazy dog

English

a, b, c, d, e, f, g, h, i, j, k, …

Step 1 – Feature Selection

Method A: Dictionary Search

Pros:

- Simple and intuitive

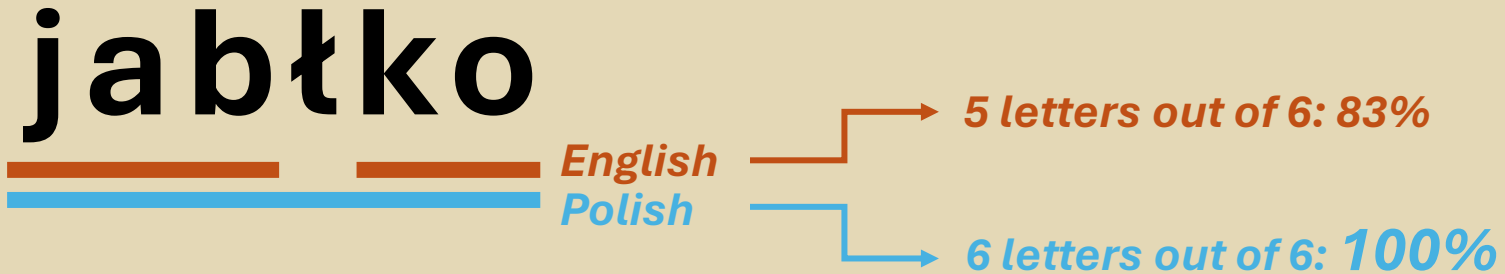
Cons:

- Languages with similar or identical char set
- Dirty dataset may affect the performance

Step 1 – Feature Selection

Method A: Dictionary Search - Quantify

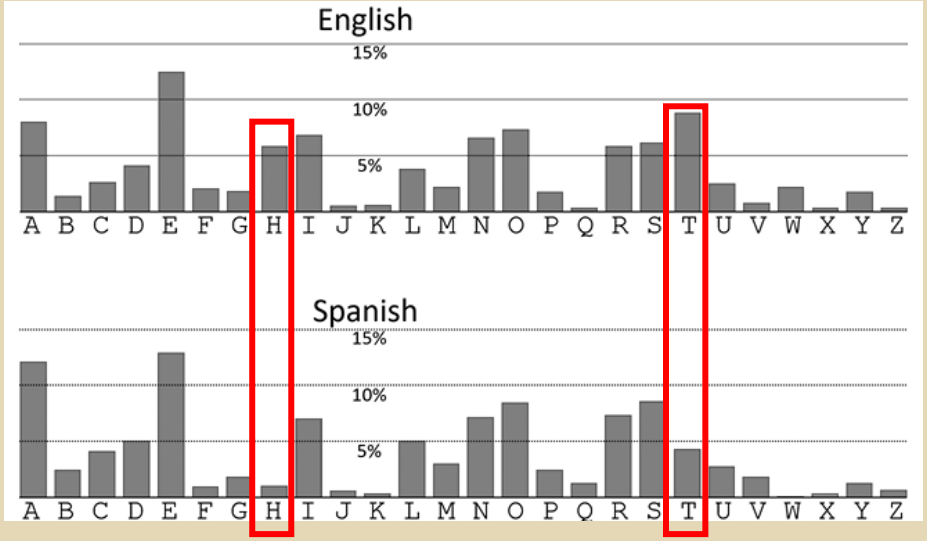
- A simple yes/no checkbox may be too strict
- Proportion of letters in the dictionary of languages



Step 1 – Feature Selection

Method B: Frequency Analysis

- Similar to dictionary search, but with letter count
- Sentences in the same language should show similar frequency table

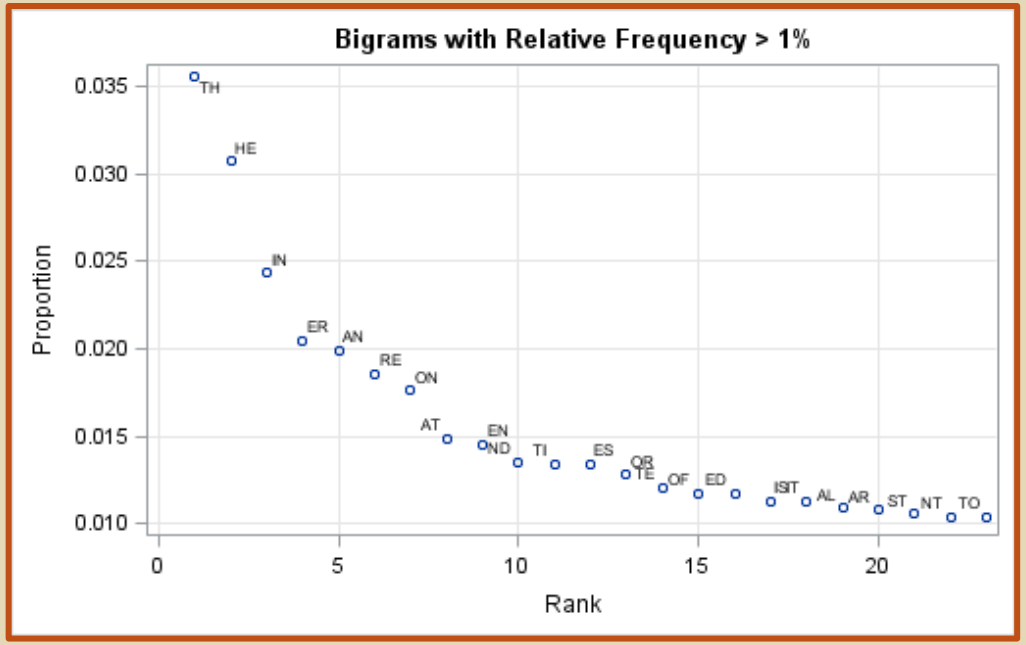


[8][9]

Step 1 – Feature Selection

Method B: Frequency Analysis

- Can be generalized to n letter combinations (n-gram)



Step 1 – Feature Selection

Method B: Frequency Analysis

Pros:

- Can further differentiate languages with same char set

Cons:

- Prone to short sample size (e.g. single word)
- Computation time

Step 1 – Feature Selection

Method B: Frequency Analysis - Quantify

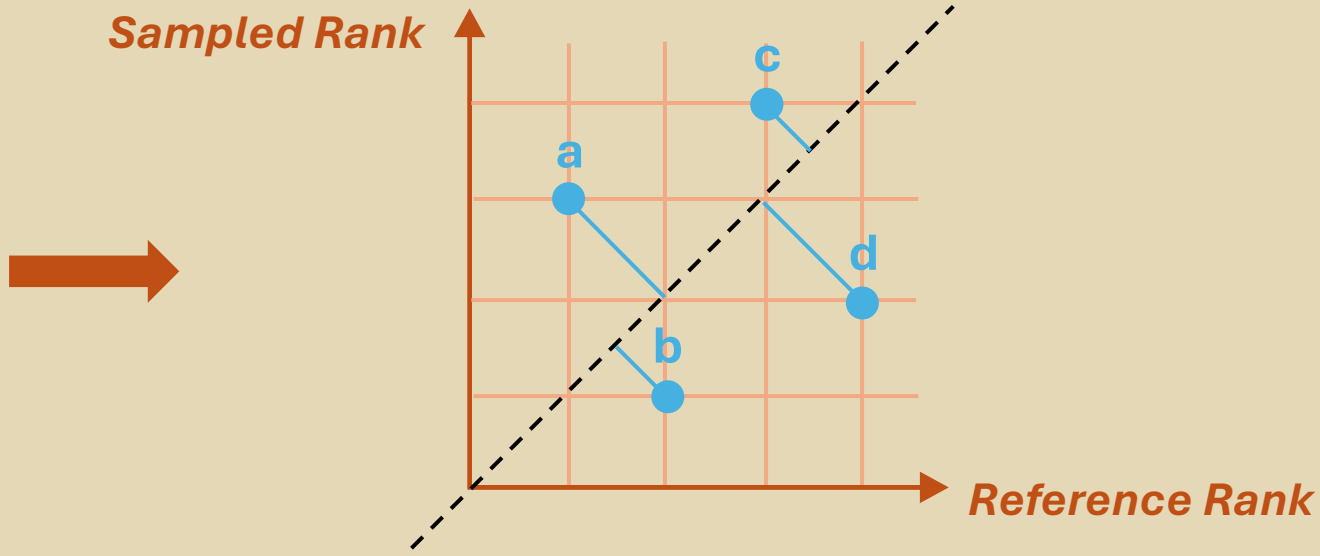
- Plot points using the rank of reference and sampled table
- Sum the distance of points to y=x line as the error

Reference Frequency Table

Letter	a	b	c	d
Rank	1	2	3	4

Sampled (From Input) Frequency Table

Letter	b	d	a	c
Rank	1	2	3	4



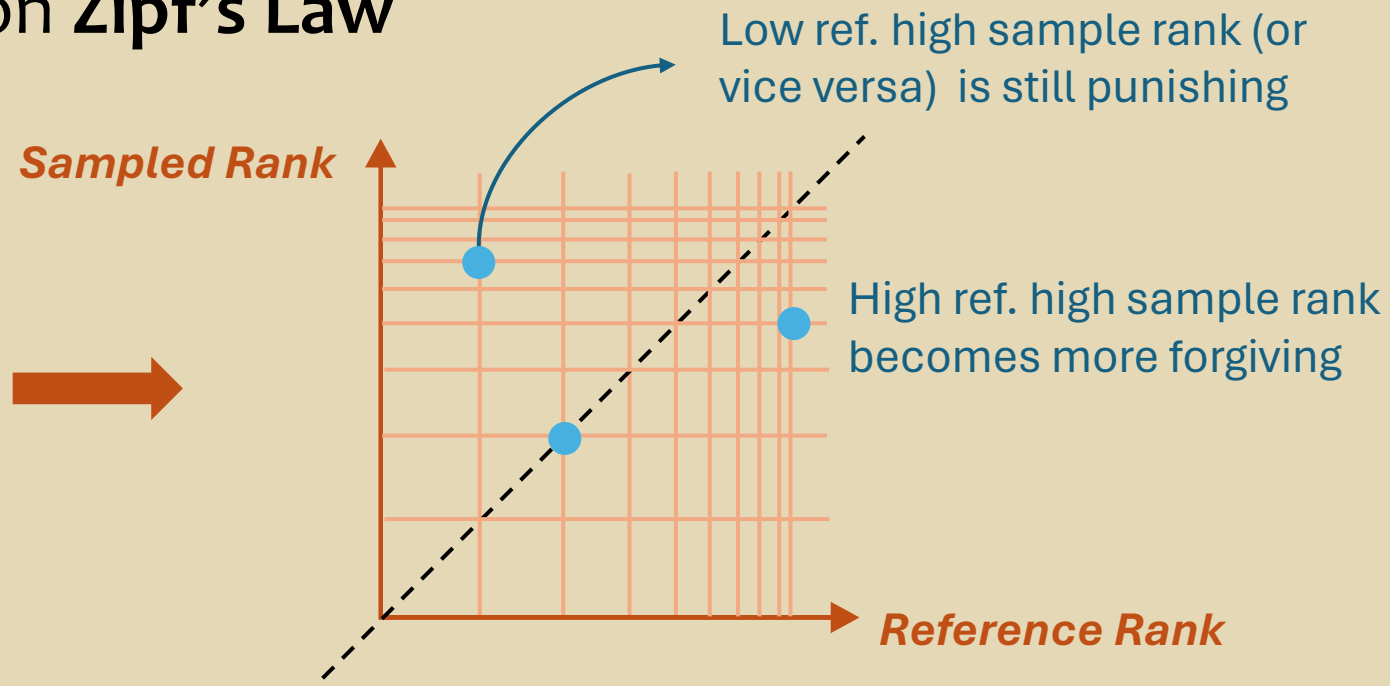
Step 1 – Feature Selection

Method B: Frequency Analysis - Quantify

- The higher rank should be less significant
- "Scale" the rank based on **Zipf's Law**

$$[Zipf's\ Law]\quad word\ freq. \propto \frac{1}{word\ rank}$$

$$scaled\ rank = \sum_{n=1}^{rank} \frac{1}{n} \approx \ln(rank)$$



Step 1 – Feature Selection

Method C: Information Density

- Different languages have different sentence/word lengths when describing the same thing
- The average information a word/letter have is different in languages

English **Pneumonoultramicroscopicsilicovolcanoconiosis**

Chinese 火山矽肺症

Step 1 – Feature Selection

Method C: Information Density

Pros:

- Simple Computation

Cons:

- No proper translation for all languages, all sentences in our dataset
- Also prone to short sample size

Step 1 – Feature Selection

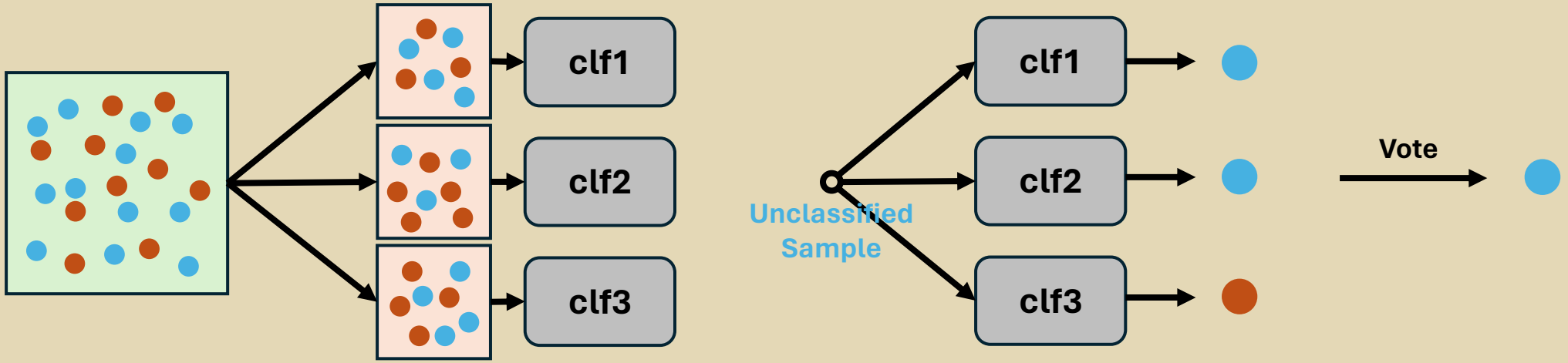
Method C: Information Density - Quantify

- Averaging out the length of the word/sentence
- Ignore words that are too short

Step 2 – Model Training

Technique A: Bagging

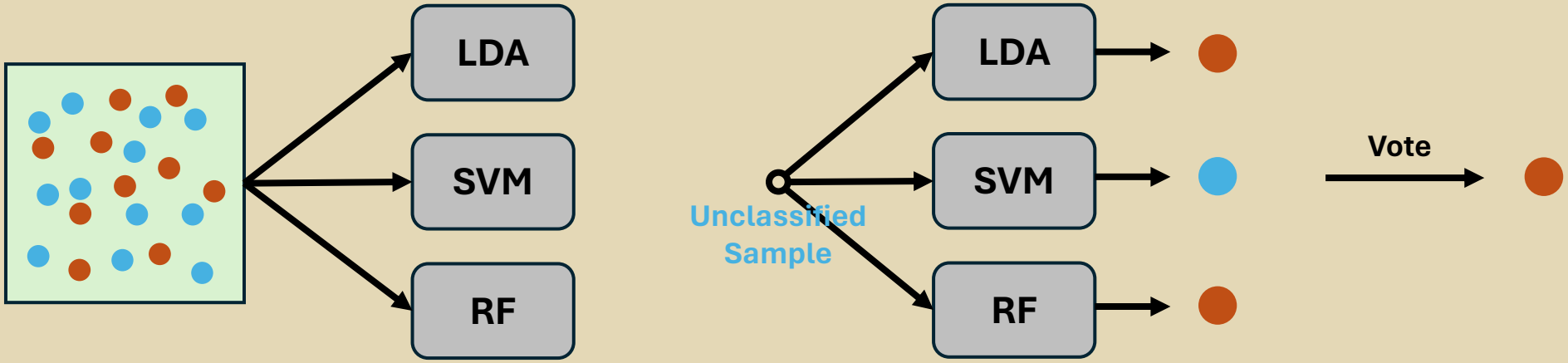
- Sample the training set multiple times to build several “weak” classifiers
- Take majority vote from these classifiers and obtain the result



Step 2 – Model Training

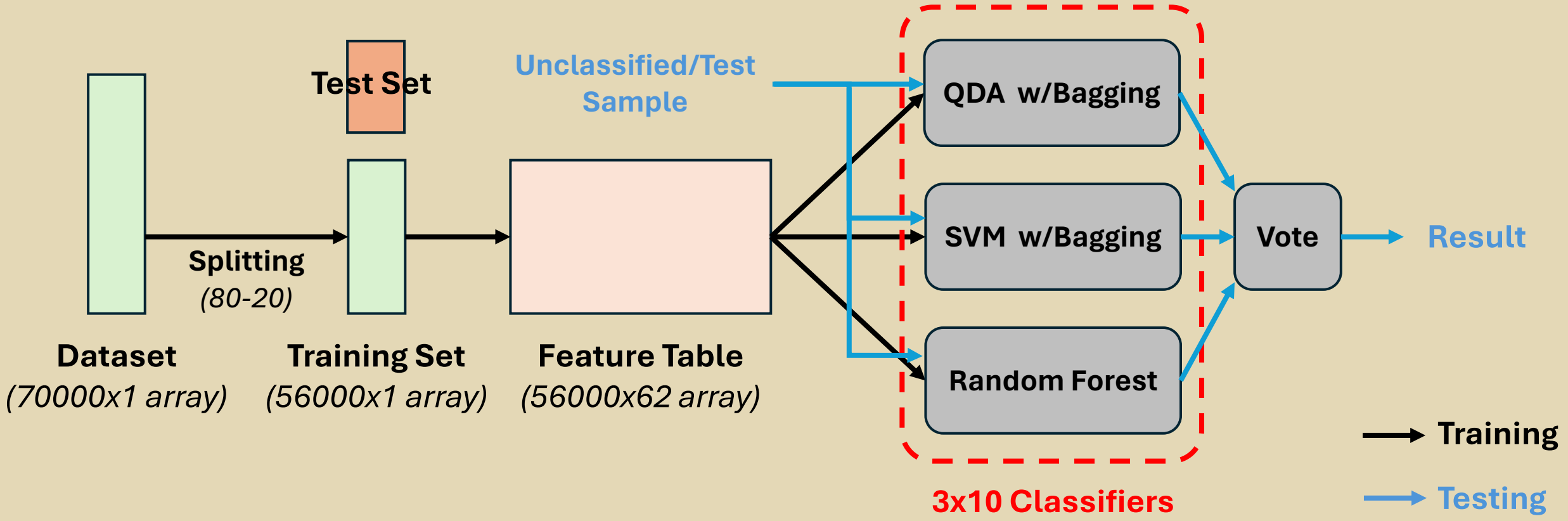
Technique B: Voting

- Building **different** classification models with the **same** training set
- Also uses majority vote



Step 2 – Model Training

Model Structure



Step 3 – Results & Analysis

Evaluation Metric

- For a classification model, **accuracy score** is used

Real	A	A	B	A	B	B	A	B	A
Predict	B	A	B	A	A	B	B	A	A
T/F	**F**	**T**	**T**	**T**	**F**	**T**	**F**	**F**	**T**

$$Acc = \frac{n(T)}{n(T) + n(F)} = \frac{5}{5 + 4} \approx 0.556$$

Step 3 – Results & Analysis

Testbench A – Test Set

- The split set from the original dataset

Testbench B – Single Word

- Texts only contain one word

Step 3 – Results & Analysis

Testbench A – Test Set // Results

QDA –

Average (10 clfs), single: 0.790

Bagged, 10 classifiers: 0.809

Support Vector Machine –

Average (10 clfs), single: 0.933

Bagged, 10 classifiers: 0.934

Decision Tree –

Average (10 clfs), single: 0.889

Bagged, 10 classifiers: 0.921

Voting –

3 classifiers (Bagged): 0.927

Step 3 – Results & Analysis

Testbench A – Test Set // Analysis

- Among the three, **QDA** has an overall worse performance (~0.8 acc), most likely due to multiple variables being collinear
- Bagging improves the accuracy by **0.02~0.04**
- The addition of voting does little effect, but it seems to stabilize the accuracy

Step 3 – Results & Analysis

Testbench B – Single Word // Results

QDA –

Average (10 clfs), single: 0.314

Bagged, 10 classifiers: 0.32

Decision Tree –

Average (10 clfs), single: 0.318

Bagged, 10 classifiers: 0.36

Support Vector Machine –

Average (10 clfs), single: 0.39

Bagged, 10 classifiers: 0.4

Voting –

3 classifiers (Bagged): 0.42

Step 3 – Results & Analysis

Testbench B – Single Word // Analysis

- The performance for all classifiers are poor
  - Frequency analysis and information density are heavily affected
  - The model was not trained with these special cases
- Voting seems to help in this case

Step 4 – Modifications

Experiment A – Bag Size

- Trying different number of weak classifiers in bagging
- Check size = 2, 4, 6, 8, 10

Voting –

2 weak clfs: 0.913 **4 weak clfs:** 0.921 **6 weak clfs:** 0.927 **8 weak clfs:** 0.925

10 weak clfs: 0.926

Step 4 – Modifications

Experiment B – Feature Selection

- Some features may be unnecessary
- Remove dictionary search, information density

Voting (10 clfs each)–

Remove Dict. Search (42 features): 0.912 **Remove Info. Density (60 features):** 0.925

Remove Freq. Analysis (22 features): 0.795

Step 5 – Restrictions & Improvements

- Time for feature computing is long
  - Could implement parallel processing or further code optimization
- Cleaner dataset
- Multilayer classification (e.g. classify language family first, then sublanguages)

- Thank You -

Contributions

嚴偉哲 – Research (90%), Coding (80%), Report Slides (100%)

黃毓為 – Research (10%), Coding (5%), Recording (50%)

郭穎達 – Coding (15%), Recording (50%)

GitHub Link

https://github.com/Mars-1114/2024-AI-FinalProject

References

[1] - https://arxiv.org/pdf/1608.08515
[2] - https://huggingface.co/datasets/papluca/language-identification
[3] - https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/
[4] - https://www.geeksforgeeks.org/quadratic-discriminant-analysis/
[5] - https://en.wikipedia.org/wiki/Decision_tree_learning
[6] - https://tinyurl.com/mry43vyd
[7] - https://en.wikipedia.org/wiki/Letter_frequency
[8] - https://blogs.sas.com/content/iml/2014/09/26/bigrams.html
[9] - https://en.wikipedia.org/wiki/N-gram
[10] - https://en.wikipedia.org/wiki/Zipf%27s_law
[11] - https://tinyurl.com/3e2nvax6
[12] - https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

References

[13] - https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier
[14] - https://ithelp.ithome.com.tw/articles/10228941