A3: Perceptron Classifier for Native Language Identification

COSC-572
Yuan-Yao Chang yc704

1.

Label	Label Count	Probability
ZHO	593	0.110510622438
TUR	504	0.0939247111442
FRA	473	0.0881475959747
SPA	450	0.0838613492359
ARA	494	0.0920611256057
DEU	337	0.06280283265
HIN	352	0.0655982109579
ITA	516	0.0961610137905
TEL	533	0.0993291092061
KOR	577	0.103801714499
JPN	577	0.103801714499

training set

Label	Label Count	Probability
ARA	51	0.0852842809365
TUR	57	0.0953177257525
FRA	53	0.0886287625418
TEL	62	0.103678929766
SPA	52	0.0869565217391
ITA	53	0.0886287625418
DEU	34	0.056856187291
KOR	60	0.100334448161
HIN	47	0.0785953177258
ZHO	69	0.115384615385
JPN	60	0.100334448161

dev set

The majority class baseline accuracy of the dev set is the language ZHO with an accuracy of 0.115384615385.

2. ´The learning process suspend at 34th iteration(33), which indicates the training data has been separated when the train accuracy is equal to 1. Also on the 11th iteration(10), the dev set owns the highest accuracy of 0.6588628762541806.

3. Base on lower and bigram, with lemmatize as an additional feature.

Features	Number of iteration to separate	Test accuracy	Dev accuracy
Lowercase	20	0.6688741721854304	0.6538461538461539
Bigram	13	0.6821192052980133	0.6722408026755853
Lowercase and Bigram	14	0.7003311258278145	0.7090301003344481
Lowercase, Bigram, Lemmatize	16	0.6804635761589404	0.7006688963210702

The model with the features of Lowercase and Bigram performed the best with the test accuracy of 0.7003311258278145.

4.

a) Confusion Matrix

ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR	ZHO
41	1	2	1	1	2	4	4	3	0	1
0	31	0	0	1	2	0	3	1	3	0
2	2	37	1	3	1	1	1	1	1	1
1	1	0	16	0	0	0	0	8	3	1
2	1	6	0	33	1	0	4	2	3	2
5	1	0	0	0	46	7	0	1	0	2
2	2	0	2	0	10	39	1	1	0	4
7	0	2	1	1	1	0	41	0	7	1
2	0	0	8	0	1	0	2	50	0	1
2	1	1	0	1	1	6	0	1	41	1
2	0	2	2	0	4	2	4	0	1	48

b)

ARA 10 highest-weighted:
[('alot of', 53), ('every thing', 30), ('the right', 29), ('many reasons', 28), ('that will', 26), ('. some', 25), ('BIAS', 25), ('not do', 24), ('reasons .', 24), ('. also', 23)]

ARA 10 lowest-weighted:

[('during the', -23), ('the statement', -22), ('. if', -22), ('. a', -21), (', it', -21), ('a lot', -20), (', who', -20), ('future .', -20), ('have enough', -19), ('people do', -19)]

BIAS: 25

DEU 10 highest-weighted:

[(', that', 52), ('important to', 29), (', because', 28), ('the statement', 24), ('. furthermore', 23), ('younger people', 23), ('able to', 22), ('. but', 22), ('. another', 22), ('a broad', 21)]

DEU 10 lowest-weighted:

[(', and', -31), ('ideas and', -24), (', we', -23), ('a person', -19), ('example ,', -19), ('of them', -18), ('such as', -18), ('like to', -18), ('we can', -18), ('life is', -18)]

BIAS: -5

FRA 10 highest-weighted:

[('for instance', 36), ('even if', 36), ('to conclude', 33), ('. indeed', 33), ('think that', 31), ('more and', 29), ('fact that', 26), ('and more', 26), ('nowadays ,', 25), ('indeed ,', 25)]

FRA 10 lowest-weighted:

[('this is', -27), ('not only', -26), ('the people', -25), (', because', -23), ('in life', -23), ('there are', -23), ('with the', -23), ('the young', -22), ('has a', -21), ('. another', -20)]

BIAS: -13

HIN 10 highest-weighted:

[('and concept', 34), ('of life', 33), ('old age', 26), ('increase in', 25), ('in this', 25), ('in todays', 25), ('according to', 24), ('number of', 23), ('concept and', 23), ('them to', 23)]

HIN 10 lowest-weighted:

```
[('and the', -31), ('i think', -22), ('academic subjects', -21), ('people ,', -20), ('people can', -19), ('. when', -19), ('as
an', -19), ('that is', -19), ('things that', -18), ('the main', -18)]
BIAS: 12

ITA 10 highest-weighted:
[('in fact', 35), ('my opinion', 35), (', for', 35), ('i think', 35), ('possibility to', 30), ('and so', 29), ('is important', 26),
('that in', 26), ('you can', 26), ('way to', 25)]
ITA 10 lowest-weighted:
[('. but', -29), ('when you', -24), ('. because', -23), ('now .', -21), ('and concepts', -21), ('tour guide', -21), (', i', -20), ('.
also', -20), ('we do', -19), ('subject .', -19)]
BIAS: 17

JPN 10 highest-weighted:
[('in japan', 71), ('i agree', 57), ('japan ,', 43), ('i think', 34), ('i disagree', 32), ('. however', 30), ('do .', 28), ('. it', 28),
('reasons .', 28), ('opinion that', 28)]
JPN 10 lowest-weighted:
[('is a', -46), ('all the', -38), (', that', -35), ('in my', -29), ('their products', -29), ('are the', -28), ('in our', -27), ("it 's", -
27), ('their own', -26), ('for a', -26)]
BIAS: -11

KOR 10 highest-weighted:
[('in korea', 49), ('however ,', 39), ('. however', 37), ('even though', 35), ('learn about', 35), ('korea ,', 33), ('such as',
31), ('. also', 30), ('. even', 29), ('these days', 28)]
KOR 10 lowest-weighted:
[('in japan', -33), ('may be', -29), (', because', -26), ('for me', -24), ('as a', -23), ('the life', -23), ('a lot', -22), ('is to', -
22), ('have the', -22), ('with this', -22)]
BIAS: -4

SPA 10 highest-weighted:
[('that you', 33), ('their lives', 31), ('the city', 31), (', etc', 30), (', is', 28), ('have a', 25), ('every day', 24), ('going to',
24), ('think that', 24), ('other hand', 24)]
SPA 10 lowest-weighted:
[('. so', -25), ('their life', -25), ('people .', -24), ('from the', -23), ('and so', -22), ('to use', -20), ('about it', -20),
('successful people', -19), ('in fact', -19), ('you might', -19)]
BIAS: -4

TEL 10 highest-weighted:
[('may not', 32), ('and also', 29), ('the subject', 29), ('with out', 27), ('academic subjects', 27), ('i conclude', 26), ('the
statement', 26), ('about the', 24), ('the above', 24), ('the concept', 24)]
TEL 10 lowest-weighted:
[('i think', -37), ('do not', -34), ('. however', -28), ('students to', -27), ('however ,', -25), ('think that', -24), ('and
concept', -22), (', they', -21), ('want to', -21), ('the most', -21)]
BIAS: 10

TUR 10 highest-weighted:
[('. because', 47), ('can not', 36), ('as a', 29), ('sum up', 28), ('to sum', 27), ('in turkey', 26), ('. moreover', 25), ('lots
of', 25), ('of this', 24), ('according to', 24)]
TUR 10 lowest-weighted:
[(', and', -41), (', i', -29), ('i agree', -27), (', but', -24), ('the statement', -23), ('agree with', -23), ('to know', -22), ('his
life', -21), ('i think', -21), ('enjoy their', -21)]
BIAS: 3

ZHO 10 highest-weighted:
```

```
[('do the', 40), ('kinds of', 35), ('time on', 33), ('. people', 32), ('how to', 31), ('and more', 28), ('enjoy the', 27), ('.
take', 26), ('but not', 26), ('get the', 24)]
ZHO 10 lowest-weighted:
[('try to', -33), ('even if', -31), ('BIAS', -30), ('and i', -28), ('know about', -25), ('. even', -23), ('going to', -23), ('is
important', -23), ('it was', -22), ('can be', -22)]
BIAS: -30
```

c)

Language	Precision	Recall	F1	BIAS weight
ARA	0.62121	0.68333	0.65079	25
DEU	0.77500	0.75610	0.76543	-5
FRA	0.74000	0.72549	0.73267	-13
HIN	0.51613	0.53333	0.52459	12
ITA	0.82500	0.61111	0.70213	17
JPN	0.66667	0.74194	0.70229	-11
KOR	0.66102	0.63934	0.65000	-4
SPA	0.68333	0.67213	0.67769	-4
TEL	0.73529	0.78125	0.75758	10
TUR	0.69492	0.74545	0.71930	3
ZHO	0.77419	0.73846	0.75591	-30

ZHO, the language owns the lowest BIAS weight which is the most common language in the training and dev set. This is unusual because ZHO have the highest probability in the first question, the reason why may can be when classifying the training set ZHO has not been classified. Also in the JPN and KOR top10 weight, interestingly mention their own country relatively more. When looking at precision and recall probability, ITA behave most significantly; the difference between precision and recall is slightly more than 0.2. ARA own the largest BIAS weight, which can be consider as the language have more samples in the training data.