## Testing

## Ukwebalana corpus

### Set 1

Training dataset 262 286

Testing data set

Unique words 11869 of 25820

Threshold 0.003

| | |
|---|---|
| 9413 true positive | 4 false positive |
| 2456 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 9357 true positive | 3 false positive |
| 2512 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 8580 true positive | 3 false positive |
| 3289 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 7368 true positive | 2 false positive |
| 4501 false negative | 44 true negative |

### Set 2

Training dataset 261 301

Testing data set

Unique words 14545 of 26805

Threshold 0.003

| | |
|---|---|
| 13541 true positive | 4 false positive |
| 1004 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 13047 true positive | 3 false positive |
| 1498 false negative | 43 true negative |

Threshold 0.005

12321 true positive 3 false positive

2224 false negative 43 true negative

Threshold 0.006

9314 true positive 2 false positive

5231 false negative 44 true negative

**Set 3**

Training dataset 255623

Testing data set

Unique words 14453 of 32483

Threshold 0.003

9641 true positive 4 false positive

4812 false negative 42 true negative

Threshold 0.004

8267 true positive 3 false positive

6186 false negative 43 true negative

Threshold 0.005

7741 true positive 3 false positive

6712 false negative 43 true negative

Threshold 0.006

6697 true positive 2 false positive

7756 false negative 44 true negative

**Set 4**

Training dataset 262 604

Testing data set

Unique words 11485 of 25502

Threshold 0.003

9139 true positive 4 false positive

2346 false negative 42 true negative

Threshold 0.004

| | |
|---|---|
| 9073 true positive | 3 false positive |
| 2412 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 7829 true positive | 3 false positive |
| 3656 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 6375 true positive | 2 false positive |
| 5110 false negative | 44 true negative |

### Set 5

Training dataset 263 551

Testing data set

Unique words 12737 of 24555

Threshold 0.003

| | |
|---|---|
| 10158 true positive | 4 false positive |
| 2579 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 9957 true positive | 3 false positive |
| 2780 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 7743 true positive | 3 false positive |
| 4994 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 6030 true positive | 2 false positive |
| 6707 false negative | 44 true negative |

### Set 6

Training dataset 240818

Testing data set

Unique words 18150 of 47288

Threshold 0.003

| | |
|---|---|
| 15090 true positive | 4 false positive |
| 3060 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 14351 true positive | 3 false positive |
| 3799 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 10607 true positive | 3 false positive |
| 7543 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 9405 true positive | 2 false positive |
| 8745 false negative | 44 true negative |

**Set 7**

Training dataset 248757

Testing data set

Unique words 16158 of 39349

Threshold 0.003

| | |
|---|---|
| 13155 true positive | 4 false positive |
| 3003 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 12017 true positive | 3 false positive |
| 4141 false negative | 43 true negative |

Threshold 0.005

9349 true positive 3 false positive
6809 false negative 43 true negative

Threshold 0.006
8758 true positive 2 false positive
7400 false negative 44 true negative

**Set 8**
Training dataset 260 091
Testing data set
Unique words 11524 of 28015
Threshold 0.003
7248 true positive 4 false positive
4276 false negative 42 true negative
Threshold 0.004
5290 true positive 3 false positive
6234 false negative 43 true negative
Threshold 0.005
4635 true positive 3 false positive
6889 false negative 43 true negative
Threshold 0.006
4112 true positive 2 false positive
7412 false negative 44 true negative

**Set 9**
Training dataset 255018
Testing data set
Unique words 13617 of 33088
Threshold 0.003
11404 true positive 4 false positive

| | |
|---|---|
| 2213 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 10905 true positive | 3 false positive |
| 2712 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 9626 true positive | 3 false positive |
| 3991 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 8161 true positive | 2 false positive |
| 5456 false negative | 42 true negative |

**Set 10**

Training dataset 282 901

Testing data set

Unique words 3529 of 5205

Threshold 0.003

| | |
|---|---|
| 3006 true positive | 4 false positive |
| 523 false negative | 42 true negative |

Threshold 0.004

| | |
|---|---|
| 2835 true positive | 3 false positive |
| 694 false negative | 43 true negative |

Threshold 0.005

| | |
|---|---|
| 2731 true positive | 3 false positive |
| 798 false negative | 43 true negative |

Threshold 0.006

| | |
|---|---|
| 2520 true positive | 2 false positive |
| 1009 false negative | 44 true negative |

Corpus compiled by Prof. Langa

Training dataset

538 732 raw words

Testing dataset

Unique words 33034

**Set 1**

Testing data

3202 unique words

Threshold 0.003

2257 true positive 5 false positive

945 false negative 41 true negative

Threshold 0.004

2146 true positive 5 false positive

1056 false negative 41 true negative

Threshold 0.005

1824 true positive 4 false positive

1378 false negative 42 true negative

Threshold 0.006

1790 true positive 3 false positive

1412 false negative 43 true negative

**Set 2**

Testing data

3202 unique words

Threshold 0.003

1824 true positive 5 false positive

1378 false negative 41 true negative

Threshold 0.004

1690 true positive 5 false positive

1512 false negative 41 true negative

Threshold 0.005
1513 true positive 4 false positive
1541 false negative 42 true negative
Threshold 0.006
1273 true positive 3 false positive
1929 false negative 43 true negative

**Set 3**
Testing data
3202 unique words
Threshold 0.003
1624 true positive 5 false positive
1578 false negative 41 true negative
Threshold 0.004
1490 true positive 5 false positive
1712 false negative 41 true negative
Threshold 0.005
1312 true positive 4 false positive
1890 false negative 42 true negative
Threshold 0.006
1290 true positive 3 false positive
1912 false negative 43 true negative

**Set 4**
Testing data
3202 unique words
Threshold 0.003
1913 true positive 5 false positive
1289 false negative 41 true negative
Threshold 0.004

1799 true positive     5 false positive

1344 false negative     41 true negative

Threshold 0.005

1744 true positive     4 false positive

1458 false negative     42 true negative

Threshold 0.006

1670 true positive     3 false positive

1532 false negative     43 true negative

**Set 5**

Testing data

3323 unique words

Threshold 0.003

2322 true positive     5 false positive

1001 false negative     41 true negative

Threshold 0.004

1976 true positive     5 false positive

1347 false negative     41 true negative

Threshold 0.005

1777 true positive     4 false positive

1546 false negative     42 true negative

Threshold 0.006

1482 true positive     3 false positive

1841 false negative     43 true negative

**Set 6**

Testing data

2467 unique words

Threshold 0.003

1521 true positive 5 false positive

946 false negative 41 true negative

Threshold 0.004

1307 true positive 5 false positive

1160 false negative 41 true negative

Threshold 0.005

1078 true positive 4 false positive

1389 false negative 42 true negative

Threshold 0.006

797 true positive 3 false positive

1670 false negative 43 true negative

**Set 7**

Testing data

3323 unique words

Threshold 0.003

2178 true positive 5 false positive

1145 false negative 41 true negative

Threshold 0.004

1912 true positive 5 false positive

1411 false negative 41 true negative

Threshold 0.005

1356 true positive 4 false positive

1967 false negative 42 true negative

Threshold 0.006

834 true positive 3 false positive

2489 false negative 43 true negative

**Set 8**

Testing data

2112 unique words

Threshold 0.003

876 true positive 5 false positive

1236 false negative 41 true negative

Threshold 0.004

736 true positive 5 false positive

1376 false negative 41 true negative

Threshold 0.005

699 true positive 4 false positive

1413 false negative 42 true negative

Threshold 0.006

514 true positive 3 false positive

1598 false negative 43 true negative

**Set 9**

Testing data

3123 unique words

Threshold 0.003

2434 true positive 5 false positive

689 false negative 41 true negative

Threshold 0.004

2306 true positive 5 false positive

817 false negative 41 true negative

Threshold 0.005

1883 true positive 4 false positive

1240 false negative 42 true negative

Threshold 0.006

| | |
|---|---|
| 1712 false negative | 3 false positive |
| 1411 true positive | 43 true negative |

**Set 10**

Testing data

5878 unique words

Threshold 0.003

| | |
|---|---|
| 4468 true positive | 5 false positive |
| 1410 false negative | 41 true negative |

Threshold 0.004

| | |
|---|---|
| 3144 true positive | 5 false positive |
| 2734 false negative | 41 true negative |

Threshold 0.005

| | |
|---|---|
| 2846 true positive | 4 false positive |
| 3032 false negative | 42 true negative |

Threshold 0.006

| | |
|---|---|
| 1738 true positive | 3 false positive |
| 4140 false negative | 43 true negative |

New items corpus

**Set 1**

Training dataset 18000

Testing dataset

Unique words 1347 of 2000

Threshold 0.003

| | |
|---|---|
| 1140 true positive | 5 false positive |
| 207 false negative | 41 true negative |

Threshold 0.004

| | |
|---|---|
| 1091 true positive | 5 false positive |
| 256 false negative | 41 true negative |

Threshold 0.005

| | |
|---|---|
| 948 true positive | 4 false positive |
| 399 false negative | 42 true negative |

Threshold 0.006

| | |
|---|---|
| 760 true positive | 3 false positive |
| 587 false negative | 43 true negative |

**Set 2**

Training dataset 18000

Testing dataset

Unique words 1278 of 2000

Threshold 0.003

| | |
|---|---|
| 1111 true positive | 5 false positive |
| 167 false negative | 41 true negative |

Threshold 0.004

| | |
|---|---|
| 987 true positive | 5 false positive |
| 291 false negative | 41 true negative |

Threshold 0.005

| | |
|---|---|
| 932 true positive | 4 false positive |
| 346 false negative | 42 true negative |

Threshold 0.006

| | |
|---|---|
| 880 true positive | 3 false positive |
| 398 false negative | 43 true negative |

**Set 3**

Training dataset 18000

Testing dataset

Unique words 1335 of 2000

Threshold 0.003

1059 true positive 5 false positive

276 false negative 41 true negative

Threshold 0.004

878 true positive 5 false positive

457 false negative 41 true negative

Threshold 0.005

836 true positive 4 false positive

499 false negative 42 true negative

Threshold 0.006

812 true positive 3 false positive

523 false negative 43 true negative

**Set 4**

Training dataset 18000

Testing dataset

Unique words 1365 of 2000

Threshold 0.003

974 true positive 5 false positive

391 false negative 41 true negative

Threshold 0.004

831 true positive 5 false positive

534 false negative 41 true negative

Threshold 0.005

720 true positive 6 false positive

645 false negative 40 true negative

Threshold 0.006

594 true positive 3 false positive

771 false negative 43 true negative

**Set 5**

Training dataset 18000

Testing dataset

Unique words 1392 of 2000

Threshold 0.003

1314 true positive 7 false positive

78 false negative 39 true negative

Threshold 0.004

1295 true positive 7 false positive

97 false negative 39 true negative

Threshold 0.005

1225 true positive 6 false positive

167 false negative 40 true negative

Threshold 0.006

969 true positive 6 false positive

423 false negative 40 true negative

**Set 6**

Training dataset 18000

Testing dataset

Unique words 1376 of 2000

Threshold 0.003

1052 true positive 5 false positive

324 false negative 41 true negative

Threshold 0.004

887 true positive 5 false positive

489 false negative 41 true negative

Threshold 0.005

829 true positive 4 false positive

547 false negative 42 true negative

Threshold 0.006

789 true positive 3 false positive

587 false negative 43 true negative

**Set 7**

Training dataset 18000

Testing dataset

Unique words 1363 of 2000

Threshold 0.003

1292 true positive 5 false positive

71 false negative 41 true negative

Threshold 0.004

1271 true positive 5 false positive

92 false negative 41 true negative

Threshold 0.005

1131 true positive 4 false positive

232 false negative 42 true negative

Threshold 0.006

1016 true positive 3 false positive

347 false negative 43 true negative

**Set 8**

Training dataset 18000

Testing dataset

Unique words 1402 of 2000

Threshold 0.003

1170 true positive 7 false positive

232 false negative 39 true negative

Threshold 0.004

1153 true positive 7 false positive

249 false negative 39 true negative

Threshold 0.005

1101 true positive 6 false positive

301 false negative 40 true negative

Threshold 0.006

| | |
|---|---|
| 1076 true positive | 6 false positive |
| 307 false negative | 40 true negative |

**Set 9**

Training dataset 18000

Testing dataset

Unique words 1332 of 2000

Threshold 0.003

| | |
|---|---|
| 1010 true positive | 5 false positive |
| 322 false negative | 41 true negative |

Threshold 0.004

| | |
|---|---|
| 881 true positive | 5 false positive |
| 451 false negative | 41 true negative |

Threshold 0.005

| | |
|---|---|
| 842 true positive | 4 false positive |
| 490 false negative | 42 true negative |

Threshold 0.006

| | |
|---|---|
| 723 true positive | 3 false positive |
| 609 false negative | 43 true negative |

**Set 10**

Training dataset 18000

Testing dataset

Unique words 1894 of 2000

Threshold 0.003

| | |
|---|---|
| 890 true positive | 5 false positive |
| 1004 false negative | 41 true negative |

Threshold 0.004

| | |
|---|---|
| 654 true positive | 5 false positive |
| 1240 false negative | 41 true negative |

Threshold 0.005

| | |
|---|---|
| 470 true positive | 4 false positive |

1424 false negative    42 true negative

Threshold 0.006

331 true positive    3 false positive

1563 false negative    43 true negative

## Second phase of the experiment
## Testing Ukwebalana corpus with the other two corpora

### Results for Prof. Langa corpus

Unique words: 33020

Threshold 0.003

16450 true positive    4 false positive

16570 false negative  42 true negative

Threshold 0.004

16122 true positive    3 false positive

16898 false negative   43 true negative

Threshold 0.005

15248 true positive    3 false positive

17772 false negative  43 true negative

Threshold 0.006

14341 true positive    2 false positive

18679 false negative   44 true negative

### Results for new items corpus

Unique words: 9587

Threshold 0.003

6601 true positive    4 false positive

2986 false negative    42 true negative

Threshold 0.004

5414 true positive    3 false positive

4173 false negative    43 true negative

Threshold 0.005

5359 true positive 3 false positive

4228 false negative 41 true negative

Threshold 0.006

4801 true positive 2 false positive

4786 false negative 44 true negative

### Testing new items corpus with the other two corpora

Results for Prof. Langa corpus

Unique words: 33020

Threshold 0.003

29131 true positive 5 false positive

3889 false negative 41 true negative

Threshold 0.004

28009 true positive 5 false positive

5011 false negative 41 true negative

Threshold 0.005

27811 true positive 4 false positive

5209 false negative 42 true negative

Threshold 0.006

27453 true positive 3 false positive

5567 false negative 43 true negative

Results for testing with Ukwebalana corpus

Unique words: 87033

Threshold 0.003

23244 true positive 5 false positive

63789 false negative 41 true negative

Threshold 0.004

19692 true positive 5 false positive

67341 false negative 41 true negative

Threshold 0.005

18469 true positive 4 false positive

68564 false negative 42 true negative

Threshold 0.006

17357 true positive 3 false positive

69676 false negative 43 true negative

## Testing Prof. Langa corpus with the other two corpora

Results testing with new items corpus

Unique words: 9587

Threshold 0.003

8227 true positive 5 false positive

1360 false negative 41 true negative

Threshold 0.004

7097 true positive 5 false positive

2490 false negative 41 true negative

Threshold 0.005

7064 true positive 4 false positive

2523 false negative 42 true negative

Threshold 0.006

6327 true positive 3 false positive

3260 false negative 43 true negative

Results for testing with Ukwebalana corpus

Unique words: 87033

### Threshold 0.003

35833 true positive 5 false positive

51200 false negative 41 true negative

### Threshold 0.004

30432 true positive 5 false positive

56601 false negative 41 true negative

### Threshold 0.005

24322 true positive 4 false positive

62711 false negative 42 true negative

### Threshold 0.006

20710 true positive 3 false positive

66323 false negative 43 true negative