Testing

Ukwebalana corpus

Set 1

Training dataset 262 286

Testing data set

Unique words 11869 of 25820

Threshold 0.003

10018 true positive 10 false positive

1851 false negative 36 true negative

Threshold 0.004

9873 true positive 10 false positive

1996 false negative 36 true negative

Threshold 0.005

8740 true positive 9 false positive

3129 false negative 37 true negative

Threshold 0.006

7543 true positive 9 false positive

4326 false negative 37 true negative

Set 2

Training dataset 261 301

Testing data set

Unique words 14545 of 26805

Threshold 0.003

13630 true positive 10 false positive

915 false negative 36 true negative

Threshold 0.004

13074 true positive 10 false positive

1471 false negative 36 true negative

Threshold 0.005

12389 true positive 9 false positive

2156 false negative 37 true negative

Threshold 0.006

9395 true positive 9 false positive

5150 false negative 37 true negative

Set 3

Training dataset 255623

Testing data set

Unique words 14453 of 32483

Threshold 0.003

9733 true positive 10 false positive

4720 false negative 36 true negative

Threshold 0.004

6108 true positive 10 false positive

8345 false negative 36 true negative

Threshold 0.005

6652 true positive 9 false positive

7801 false negative 37 true negative

Threshold 0.006

7682 true positive 9 false positive

6771 false negative 37 true negative

Set 4

Training dataset 262 604

Testing data set

Unique words 11485 of 25502

Threshold 0.003

10251 true positive 10 false positive

1234 false negative 36 true negative

Threshold 0.004

9163 true positive 10 false positive

2322 false negative 36 true negative

Threshold 0.005

7917 true positive 9 false positive

3568 false negative 37 true negative

Threshold 0.006

6586 true positive 9 false positive

4899 false negative 37 true negative

Set 5

Training dataset 263 551

Testing data set

Unique words 12737 of 24555

Threshold 0.003

11259 true positive 10 false positive

1478 false negative 36 true negative

Threshold 0.004

10174 true positive 10 false positive

2563 false negative 36 true negative

Threshold 0.005

7841 true positive 9 false positive

4896 false negative 37 true negative

Threshold 0.006

6152 true positive 9 false positive

6585 false negative 36 true negative

Set 6

Training dataset 240818

Testing data set

Unique words 18150 of 47288

Threshold 0.003

15671 true positive 10 false positive

2479 false negative 36 true negative

Threshold 0.004

14567 true positive 10 false positive

3583 false negative 36 true negative

Threshold 0.005

12560 true positive 9 false positive

5590 false negative 37 true negative

Threshold 0.006

11220 true positive 9 false positive

6930 false negative 37 true negative

Set 7

Training dataset 248757

Testing data set

Unique words 16158 of 39349

Threshold 0.003

14412 true positive 10 false positive

1746 false negative 36 true negative

Threshold 0.004

12458 true positive 10 false positive

3700 false negative 36 true negative

Threshold 0.005

9656 true positive 9 false positive

6502 false negative 37 true negative

Threshold 0.006

8955 true positive 9 false positive

7203 false negative 37 true negative

Set 8

Training dataset 260 091

Testing data set

Unique words 11524 of 28015

Threshold 0.003

7448 true positive 10 false positive

4076 false negative 36 true negative

Threshold 0.004

5548 true positive 10 false positive

5976 false negative 36 true negative

Threshold 0.005

4788 true positive 9 false positive

6736 false negative 37 true negative

Threshold 0.006

4253 true positive 9 false positive

7271 false negative 37 true negative

Set 9

Training dataset 255018

Testing data set

Unique words 13617 of 33088

Threshold 0.003

11886 true positive 10 false positive

1731 false negative 36 true negative

Threshold 0.004

11056 true positive 10 false positive

2561 false negative 36 true negative

Threshold 0.005

9885 true positive 9 false positive

3732 false negative 37 true negative

Threshold 0.006

8412 true positive 9 false positive

5205 false negative 37 true negative

Set 10

Training dataset 282 901

Testing data set

Unique words 3529 of 5205

Threshold 0.003

3314 true positive 10 false positive

215 false negative 36 true negative

Threshold 0.004

2985 true positive 10 false positive

544 false negative 36 true negative

Threshold 0.005

2885 true positive 9 false positive

644 false negative 37 true negative

Threshold 0.006

2654 true positive 9 false positive

875 false negative 37 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

2479 true positive 8 false positive

723 false negative 38 true negative

Threshold 0.004

2279 true positive 8 false positive

923 false negative 38 true negative

Threshold 0.005

2040 true positive 6 false positive

1162 false negative 40 true negative

Threshold 0.006

1904 true positive 5 false positive

1298 false negative 41 true negative

Set 2

Testing data

3202 unique words

Threshold 0.003

2045 true positive 8 false positive

1257 false negative 38 true negative

Threshold 0.004

1819 true positive 8 false positive

1483 false negative 38 true negative

Threshold 0.005

1761 true positive 6 false positive

1541 false negative 40 true negative

Threshold 0.006

1373 true positive 5 false positive

1929 false negative 41 true negative

Set 3

Testing data

3202 unique words

Threshold 0.003

1522 true positive 8 false positive

1780 false negative 38 true negative

Threshold 0.004

1571 true positive 8 false positive

1731 false negative 38 true negative

Threshold 0.005

1273 true positive 6 false positive

2029 false negative 40 true negative

Threshold 0.006

1373 true positive 5 false positive

1929 false negative 41 true negative

Set 4

Testing data

3202 unique words

Threshold 0.003

1550 true positive 8 false positive

1752 false negative 38 true negative

Threshold 0.004

1958 true positive 8 false positive

1344 false negative 38 true negative

Threshold 0.005

1110 true positive 6 false positive

2192 false negative 40 true negative

Threshold 0.006

2342 true positive 5 false positive

960 false negative 41 true negative

Set 5

Testing data

3323 unique words

Threshold 0.003

2423 true positive 8 false positive

900 false negative 38 true negative

Threshold 0.004

2100 true positive 8 false positive

1223 false negative 38 true negative

Threshold 0.005

1824 true positive 6 false positive

1499 false negative 40 true negative

Threshold 0.006

1532 true positive 5 false positive

1791 false negative 41 true negative

Set 6

Testing data

2467 unique words

Threshold 0.003

1632 true positive 8 false positive

735 false negative 38 true negative

Threshold 0.004

1423 true positive 8 false positive

1044 false negative 38 true negative

Threshold 0.005

1233 true positive 6 false positive

1234 false negative 40 true negative

Threshold 0.006

934 true positive 5 false positive

1533 false negative 41 true negative

Set 7

Testing data

3323 unique words

Threshold 0.003

2356 true positive 8 false positive

956 false negative 38 true negative

Threshold 0.004

1926 true positive 8 false positive

1397 false negative 38 true negative

Threshold 0.005

1441 true positive 6 false positive

1882 false negative 40 true negative

Threshold 0.006

900 true positive 5 false positive

2423 false negative 41 true negative

Set 8

Testing data

2112 unique words

Threshold 0.003

1123 true positive 8 false positive

989 false negative 38 true negative

Threshold 0.004

923 true positive 8 false positive

1189 false negative 38 true negative

Threshold 0.005

745 true positive 6 false positive

1367 false negative 40 true negative

Threshold 0.006

623 true positive 5 false positive

1489 false negative 41 true negative

Set 9

Testing data

3123 unique words

Threshold 0.003

2563 true positive 8 false positive

560 false negative 38 true negative

Threshold 0.004

2356 true positive 8 false positive

767 false negative 38 true negative

Threshold 0.005

1942 true positive 6 false positive

1181 false negative 40 true negative

Threshold 0.006

1526 true positive 5 false positive

1597 false negative 41 true negative

Set 10

Testing data

5878 unique words

Threshold 0.003

4502 true positive 8 false positive

1376 false negative 38 true negative

Threshold 0.004

3256 true positive 8 false positive

2622 false negative 38 true negative

Threshold 0.005

2914 true positive 6 false positive

2964 false negative 40 true negative

Threshold 0.006

1942 true positive 5 false positive

3936 false negative 41 true negative

New items corpus

Set 1

Training dataset 18000

Testing dataset

Unique words 1347 of 2000

Threshold 0.003

1221 true positive 7 false positive

126 false negative 39 true negative

Threshold 0.004

1178 true positive 7 false positive

169 false negative 39 true negative

Threshold 0.005

974 true positive 6 false positive

373 false negative 40 true negative

Threshold 0.006

855 true positive 6 false positive

492 false negative 40 true negative

Set 2

Training dataset 18000

Testing dataset

Unique words 1278 of 2000

Threshold 0.003

1145 true positive 7 false positive

133 false negative 39 true negative

Threshold 0.004

1005 true positive 7 false positive

273 false negative 39 true negative

Threshold 0.005

978 true positive 6 false positive

300 false negative 40 true negative

Threshold 0.006

744 true positive 6 false positive

252 false negative 40 true negative

Set 3

Training dataset 18000

Testing dataset

Unique words 1335 of 2000

Threshold 0.003

1087 true positive 7 false positive

248 false negative 39 true negative

Threshold 0.004

945 true positive 7 false positive

390 false negative 39 true negative

Threshold 0.005

877 true positive 6 false positive

458 false negative 40 true negative

Threshold 0.006

742 true positive 6 false positive

593 false negative 40 true negative

Set 4

Training dataset 18000

Testing dataset

Unique words 1365 of 2000

Threshold 0.003

998 true positive 7 false positive

367 false negative 39 true negative

Threshold 0.004

866 true positive 7 false positive

499 false negative 39 true negative

Threshold 0.005

765 true positive 6 false positive

600 false negative 40 true negative

Threshold 0.006

612 true positive 6 false positive

753 false negative 40 true negative

Set 5

Training dataset 18000

Testing dataset

Unique words 1392 of 2000

Threshold 0.003

1348 true positive 7 false positive

44 false negative 39 true negative

Threshold 0.004

1312 true positive 7 false positive

80 false negative 39 true negative

Threshold 0.005

1276 true positive 6 false positive

116 false negative 40 true negative

Threshold 0.006

995 true positive 6 false positive

397 false negative 40 true negative

Set 6

Training dataset 18000

Testing dataset

Unique words 1376 of 2000

Threshold 0.003

1078 true positive 7 false positive

298 false negative 39 true negative

Threshold 0.004

914 true positive 7 false positive

462 false negative 39 true negative

Threshold 0.005

856 true positive 6 false positive

520 false negative 40 true negative

Threshold 0.006

812 true positive 6 false positive

564 false negative 40 true negative

Set 7

Training dataset 18000

Testing dataset

Unique words 1363 of 2000

Threshold 0.003

1305 true positive 7 false positive

58 false negative 39 true negative

Threshold 0.004

1289 true positive 7 false positive

74 false negative 39 true negative

Threshold 0.005

1168 true positive 6 false positive

195 false negative 40 true negative

Threshold 0.006

1056 true positive 6 false positive

307 false negative 40 true negative

Set 8

Training dataset 18000

Testing dataset

Unique words 1402 of 2000

Threshold 0.003

1184 true positive 7 false positive

218 false negative 39 true negative

Threshold 0.004

1178 true positive 7 false positive

169 false negative 39 true negative

Threshold 0.005

1124 true positive 6 false positive

278 false negative 40 true negative

Threshold 0.006

1095 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

1034 true positive 7 false positive

298 false negative 39 true negative

Threshold 0.004

912 true positive 7 false positive

420 false negative 39 true negative

Threshold 0.005

856 true positive 6 false positive

476 false negative 41 true negative

Threshold 0.006

764 true positive 6 false positive

568 false negative 41 true negative

Set 10

Training dataset 18000

Testing dataset

Unique words 1894 of 2000

Threshold 0.003

1657 true positive 7 false positive

237 false negative 39 true negative

Threshold 0.004

1556 true positive 7 false positive

338 false negative 39 true negative

Threshold 0.005

1489 true positive 6 false positive

405 false negative 41 true negative

Threshold 0.006

1256 true positive 6 false positive

638 false negative 41 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

17847 true positive 10 false positive

15173 false negative 36 true negative

Threshold 0.004

16879 true positive 10 false positive

16141 false negative 36 true negative

Threshold 0.005

15469 true positive 9 false positive

17551 false negative 37 true negative

Threshold 0.006

14587 true positive 9 false positive

18433 false negative 37 true negative

Results for new items corpus

Unique words: 9587

Threshold 0.003

6744 true positive 10 false positive

2843 false negative 36 true negative

Threshold 0.004

5601 true positive 10 false positive

3986 false negative 36 true negative

Threshold 0.005

5468 true positive 9 false positive

4119 false negative 37 true negative

Threshold 0.006

4956 true positive 9 false positive

4631 false negative 37 true negative

Testing new items corpus with the other two corpora

Results for Prof. Langa corpus

Unique words: 33020

Threshold 0.003

29284 true positive 8 false positive

3736 false negative 38 true negative

Threshold 0.004

28189 true positive 8 false positive

4831 false negative 38 true negative

Threshold 0.005

27923 true positive 6 false positive

5097 false negative 40 true negative

Threshold 0.006

27546 true positive 6 false positive

5474 false negative 40 true negative

Results for testing with Ukwebalana corpus

Unique words: 87033

Threshold 0.003

23459 true positive 8 false positive

63574 false negative 38 true negative

Threshold 0.004

19987 true positive 8 false positive

67046 false negative 38 true negative

Threshold 0.005

18789 true positive 6 false positive

68244 false negative 40 true negative

Threshold 0.006

17566 true positive 6 false positive

69467 false negative 40 true negative

Testing Prof. Langa corpus with the other two corpora

Results testing with new items corpus

Unique words: 9587

Threshold 0.003

8567 true positive 10 false positive

1020 false negative 36 true negative

Threshold 0.004

7456 true positive 9 false positive

2131 false negative 37 true negative

Threshold 0.005

7145 true positive 8 false positive

2442 false negative 38 true negative

Threshold 0.006

6556 true positive 8 false positive

3031 false negative 38 true negative

Results for testing with Ukwebalana corpus

Unique words: 87033

Threshold 0.003

36078 true positive 10 false positive

50954 false negative 39 true negative

Threshold 0.004

30541 true positive 9 false positive

56491 false negative 37 true negative

Threshold 0.005

24423 true positive 8 false positive

62609 false negative 36 true negative

Threshold 0.006

20910 true positive 8 false positive

66122 false negative 36 true negative