**!! Tabellen moeten nog aangepast worden !!**

Results for both datasets with a Gaussian Naive Bayes (without the ‘Other’ class):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| South Park  Naive Bayes: Gaussian  Without Other | Baseline | Unigram | Bigram | Trigram | Improvement  of best model |
| Characters: 3 | 0.39 | 0.40 | **0.42** | **0.42** | 0.03 |
| Characters: 5 | **0.34** | 0.25 | 0.32 | 0.29 | x |
| Characters: 7 | **0.32** | 0.20 | 0.28 | 0.25 | x |
| Characters: 10 | **0.29** | 0.19 | 0.27 | 0.25 | x |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Game of Thrones  Naive Bayes: Gaussian  Without Other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0.46 | 0.49 | **0.51** | 0.45 | 0.05 |
| Characters: 5 | 0.30 | 0.33 | **0.35** | 0.26 | 0.05 |
| Characters: 7 | 0.24 | 0.25 | **0.26** | 0.22 | 0.02 |
| Characters: 10 | 0.21 | 0.20 | **0.23** | 0.16 | 0.02 |

For Naive Bayes it was not possible to put in the whole dataset and all the features, this results in a memory error. So we experimented with smaller parts of the dataset and/or less features, different combinations showing only very little differences in results, the best results we obtained when using the whole dataset and 20000 features. The results when including the ‘Other’ class were not able to beat the baseline scores, it simply always chose the Other class which is also the baseline

**Naive Bayes**

**Whole dataset**

**20000 features**

SVM

[:10000]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| South Park  Svm:  Without other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0,4 | 0,53 | **0,67** | 0,61 | 0,27 |
| Characters: 5 | 0,33 | **0,71** | 0,67 | 0,61 | 0,38 |
| Characters: 7 | 0,31 | **0,71** | 0,67 | 0,61 | 0,4 |
| Characters: 10 | 0,29 | **0,71** | 0,67 | 0,61 | 0,42 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Game of Thrones  Svm:  Without other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0,45 | 0,52 | **0,6** | 0,58 | 0,15 |
| Characters: 5 | 0,3 | **0,66** | 0,6 | 0,58 | 0,36 |
| Characters: 7 | 0,24 | **0,66** | 0,6 | 0,58 | 0,42 |
| Characters: 10 | 0,2 | **0,66** | 0,6 | 0,58 | 0,46 |

After experimenting on a smaller part of the dataset, our findings are that the results when including the ‘Other’ class were not able to beat the baseline scores.

KNN

[:10000]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| South Park  Knn:  Without other | Baseline | Unigram | Bigram | Trigram | Improvements of best model |
| Characters: 3 | 0,4 | 0,59 (K=19) | 0,61 (K=1) | **0,62 (K=1)** | 0,22 |
| Characters: 5 | 0,33 | 0,59 (K=19) | 0,61 (K=1) | **0,62 (K=1)** | 0,29 |
| Characters: 7 | 0,31 | 0,59 (K=19) | 0,61 (K=1) | **0,62 (K=1)** | 0,31 |
| Characters: 10 | 0,29 | 0,59 (K=19) | 0,61 (K=1) | **0,62 (K=1)** | 0,33 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| South Park  Knn:  With other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0,65 | **0,66 (K=37)** | **0,66 (K=19)** | 0,65 (K=21) | 0,01 |
| Characters: 5 | 0,58 | **0,79 (K=17)** | 0,78 (K=15) | 0,78 (K=3) | 0,21 |
| Characters: 7 | 0,55 | **0,79 (K=17)** | 0,78 (K=15) | 0,78 (K=3) | 0,24 |
| Characters: 10 | 0,52 | **0,79 (K=17)** | 0,78 (K=15) | 0,78 (K=3) | 0,27 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Game of Thrones  Knn:  Without other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0,45 | **0,62 (K=61)** | 0,49 (K=25) | 0,6 (K=1) | 0,17 |
| Characters: 5 | 0,3 | **0,62 (K=61)** | 0,49 (K=25) | 0,6 (K=1) | 0,32 |
| Characters: 7 | 0,24 | **0,62 (K=61)** | 0,49 (K=25) | 0,6 (K=1) | 0,38 |
| Characters: 10 | 0,2 | **0,62 (K=61)** | 0,49 (K=25) | 0,6 (K=1) | 0,42 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Game of Thrones  Knn:  With other | Baseline | Unigram | Bigram | Trigram | Improvement of best model |
| Characters: 3 | 0,84 | **0,85 (K=9)** | **0,85 (K=7)** | **0,85 (K=7)** | 0,01 |
| Characters: 5 | 0,75 | **0,89 (K=9)** | **0,89 (K=5)** | **0,89 (K=7)** | 0,14 |
| Characters: 7 | 0,69 | **0,89 (K=9)** | **0,89 (K=5)** | **0,89 (K=7)** | 0,2 |
| Characters: 10 | 0,63 | **0,89 (K=9)** | **0,89 (K=5)** | **0,89 (K=7)** | 0,26 |

For KNN it was not possible to put in the whole dataset and all the features, this results in a memory error. So we experimented with smaller parts of the dataset and/or less features, different combinations showing only very little differences in results, the best results we obtained when using the 20000 rows of the dataset and 20000 features. We also experimented with different values for K, the value of K which leads to the highest results can be seen in the table. The results when including the ‘Other’ class were not able to beat the baseline scores, it simply always chose the Other class which is also the baseline.

**KNN**

**Dataset [:20000]**

**Features 20000**

1-grams:

Baseline: 0.43733333333333335

best score for South Park with 3 characters, without class "Other", sample size of 3000 and ngram=1: 0.4653333333333333

k = 39

Baseline: 0.36933333333333335

best score for South Park with 5 characters, without class "Other", sample size of 3000 and ngram=1: 0.38533333333333336

k = 23

Baseline: 0.344

best score for South Park with 7 characters, without class "Other", sample size of 3000 and ngram=1: 0.3586666666666667

k = 63

Baseline: 0.344

best score for South Park with 10 characters, without class "Other", sample size of 3000 and ngram=1: 0.356

k = 61

Baseline: 0.47733333333333333

best score for Game of Thrones with 3 characters, without class "Other", sample size of 3000 and ngram=1: 0.45866666666666667

k = 61

Baseline: 0.33466666666666667

best score for Game of Thrones with 5 characters, without class "Other", sample size of 3000 and ngram=1: 0.3413333333333333

k = 59

Baseline: 0.26666666666666666

best score for Game of Thrones with 7 characters, without class "Other", sample size of 3000 and ngram=1: 0.24

k = 55

Baseline: 0.22533333333333333

best score for Game of Thrones with 10 characters, without class "Other", sample size of 3000 and ngram=1: 0.19333333333333333

k = 55

2-grams:

Baseline: 0.43733333333333335

best score for South Park with 3 characters, without class "Other", sample size of 3000 and ngram=2: 0.42

k = 3

Baseline: 0.36933333333333335

best score for South Park with 5 characters, without class "Other", sample size of 3000 and ngram=2: 0.2986666666666667

k = 11

Baseline: 0.344

best score for South Park with 7 characters, without class "Other", sample size of 3000 and ngram=2: 0.3293333333333333

k = 1

Baseline: 0.344

best score for South Park with 10 characters, without class "Other", sample size of 3000 and ngram=2: 0.26

k = 5

Baseline: 0.47733333333333333

best score for Game of Thrones with 3 characters, without class "Other", sample size of 3000 and ngram=2: 0.45066666666666666

k = 61

Baseline: 0.33466666666666667

best score for Game of Thrones with 5 characters, without class "Other", sample size of 3000 and ngram=2: 0.2853333333333333

k = 69

Baseline: 0.26666666666666666

best score for Game of Thrones with 7 characters, without class "Other", sample size of 3000 and ngram=2: 0.22266666666666668

k = 27

Baseline: 0.22533333333333333

best score for Game of Thrones with 10 characters, without class "Other", sample size of 3000 and ngram=2: 0.18133333333333335

k = 63

3-grams:

Baseline: 0.43733333333333335

best score for South Park with 3 characters, without class "Other", sample size of 3000 and ngram=3: 0.37066666666666664

k = 5

Baseline: 0.36933333333333335

best score for South Park with 5 characters, without class "Other", sample size of 3000 and ngram=3: 0.37466666666666665

k = 1

Baseline: 0.344

best score for South Park with 7 characters, without class "Other", sample size of 3000 and ngram=3: 0.3506666666666667

k = 1

Baseline: 0.344

best score for South Park with 10 characters, without class "Other", sample size of 3000 and ngram=3: 0.3333333333333333

k = 21

Baseline: 0.47733333333333333

best score for Game of Thrones with 3 characters, without class "Other", sample size of 3000 and ngram=3: 0.432

k = 15

Baseline: 0.33466666666666667

best score for Game of Thrones with 5 characters, without class "Other", sample size of 3000 and ngram=3: 0.33466666666666667

k = 1

Baseline: 0.26666666666666666

best score for Game of Thrones with 7 characters, without class "Other", sample size of 3000 and ngram=3: 0.23733333333333334

k = 57

Baseline: 0.22533333333333333

best score for Game of Thrones with 10 characters, without class "Other", sample size of 3000 and ngram=3: 0.20133333333333334

k = 67

**Naive Bayes**

**Whole dataset**

**20000 features**

Baseline: 0.3896400065157192

South Park 3 1 0.39990226421241243

Baseline: 0.3896400065157192

South Park 3 2 0.428408535592116

Baseline: 0.3896400065157192

South Park 3 3 0.4179833849161101

Baseline: 0.4563894523326572

Game Of Thrones 3 1 0.49391480730223125

Baseline: 0.4563894523326572

Game Of Thrones 3 2 0.5091277890466531

Baseline: 0.4563894523326572

Game Of Thrones 3 3 0.4432048681541582

Baseline: 0.3366189576019444

South Park 5 1 0.24615176883607887

Baseline: 0.3366189576019444

South Park 5 2 0.3166351606805293

Baseline: 0.3366189576019444

South Park 5 3 0.29003510667026733

Baseline: 0.2953156822810591

Game Of Thrones 5 1 0.3340122199592668

Baseline: 0.2953156822810591

Game Of Thrones 5 2 0.35030549898167007

Baseline: 0.2953156822810591

Game Of Thrones 5 3 0.2572980312287848

Baseline: 0.3151153943697692

South Park 7 1 0.19781891960436215

Baseline: 0.3151153943697692

South Park 7 2 0.2789753994420492

Baseline: 0.3151153943697692

South Park 7 3 0.2509510524980979

Baseline: 0.23753351206434317

Game Of Thrones 7 1 0.24718498659517427

Baseline: 0.23753351206434317

Game Of Thrones 7 2 0.26327077747989275

Baseline: 0.23753351206434317

Game Of Thrones 7 3 0.22144772117962466

Baseline: 0.2865566037735849

South Park 10 1 0.19304245283018867

Baseline: 0.2865566037735849

South Park 10 2 0.2659198113207547

Baseline: 0.2865566037735849

South Park 10 3 0.2455188679245283

Baseline: 0.20737741790373368

Game Of Thrones 10 1 0.19928025191183085

Baseline: 0.20737741790373368

Game Of Thrones 10 2 0.22582096266306792

Baseline: 0.20737741790373368

Game Of Thrones 10 3 0.1632928475033738