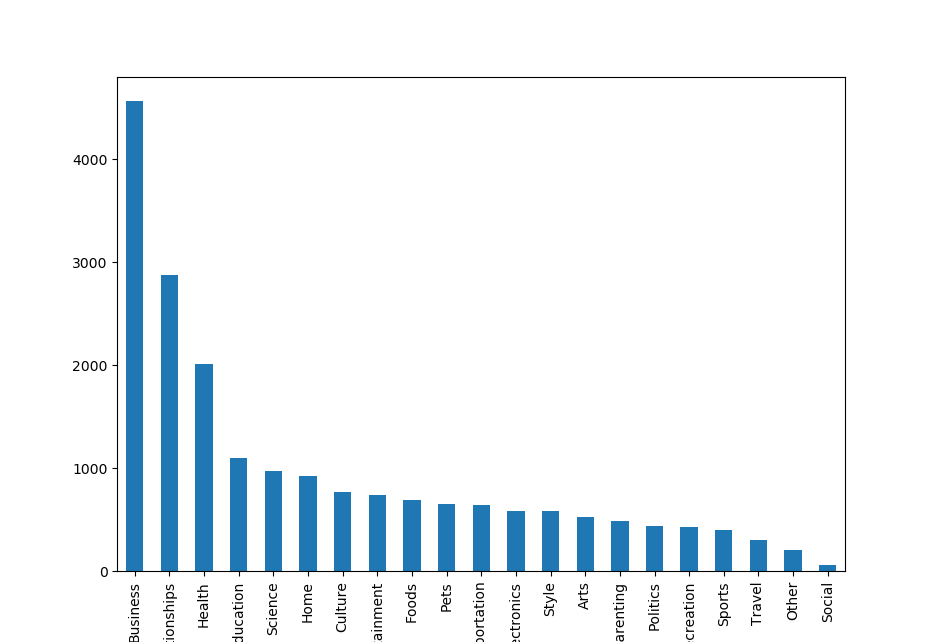
NLP Final Project

1. Statistical Report

A. 

Business 4570

Relationships 2874

Health 2011

Education 1099

Science 978

Home 926

Culture 773

Entertainment 738

Foods 695

Pets 650

Transportation 642

Electronics 590

Style 589

Arts 529

Parenting 485

Politics 442

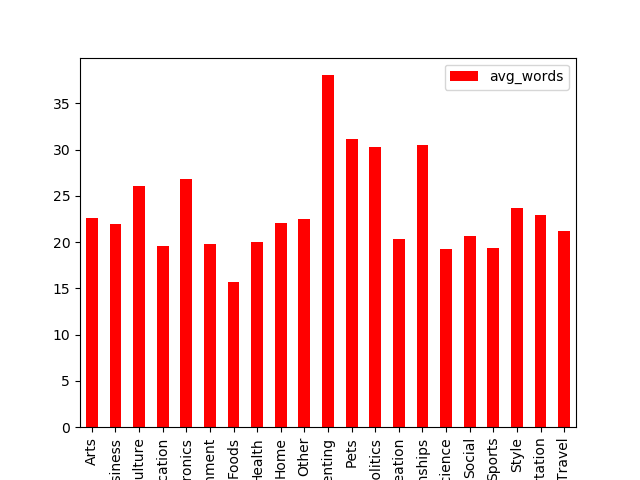
Recreation 434

Sports 400

Travel 307

Other 208

Social 60

B. Pre Cleaning

+----+----------------+-------------+

| | category | avg\_words |

|----+----------------+-------------|

| 0 | Arts | 22.6597 |

| 1 | Business | 21.9893 |

| 2 | Culture | 26.0543 |

| 3 | Education | 19.5323 |

| 4 | Electronics | 26.861 |

| 5 | Entertainment | 19.7642 |

| 6 | Foods | 15.6489 |

| 7 | Health | 19.998 |

| 8 | Home | 22.1112 |

| 9 | Other | 22.4904 |

| 10 | Parenting | 38.0392 |

| 11 | Pets | 31.1908 |

| 12 | Politics | 30.2602 |

| 13 | Recreation | 20.3041 |

| 14 | Relationships | 30.5501 |

| 15 | Science | 19.2832 |

| 16 | Social | 20.6333 |

| 17 | Sports | 19.4125 |

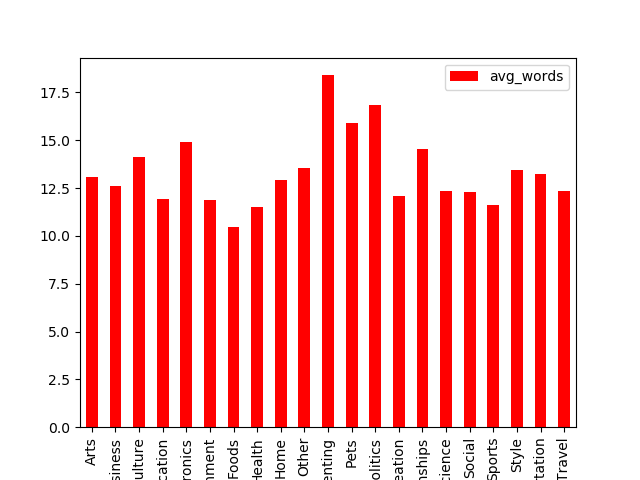
| 18 | Style | 23.6503 |

| 19 | Transportation | 22.8847 |

| 20 | Travel | 21.1857 |

+----+----------------+-------------+

Post Cleaning



+----+----------------+-------------+

| | category | avg\_words |

|----+----------------+-------------|

| 0 | Arts | 10.5066 |

| 1 | Business | 9.67155 |

| 2 | Culture | 11.6559 |

| 3 | Education | 8.7707 |

| 4 | Electronics | 12.0864 |

| 5 | Entertainment | 8.78726 |

| 6 | Foods | 7.59856 |

| 7 | Health | 8.82148 |

| 8 | Home | 10.1285 |

| 9 | Other | 10.4135 |

| 10 | Parenting | 15.8268 |

| 11 | Pets | 13.3877 |

| 12 | Politics | 13.8552 |

| 13 | Recreation | 9.06912 |

| 14 | Relationships | 11.9642 |

| 15 | Science | 9.18098 |

| 16 | Social | 9.21667 |

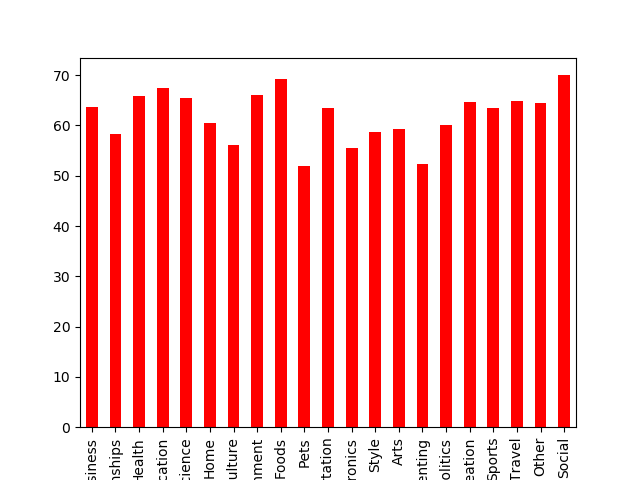
| 17 | Sports | 8.645 |

| 18 | Style | 10.8404 |

| 19 | Transportation | 10.0389 |

| 20 | Travel | 9.17915 |

+----+----------------+-------------+

C. 

Business 63.698031

Relationships 58.315936

Health 65.788165

Education 67.515924

Science 65.541922

Home 60.475162

Culture 56.144890

Entertainment 66.124661

Foods 69.208633

Pets 51.846154

Transportation 63.551402

Electronics 55.593220

Style 58.743633

Arts 59.357278

Parenting 52.371134

Politics 60.180995

Recreation 64.746544

Sports 63.500000

Travel 64.820847

Other 64.423077

Social 70.000000

D.

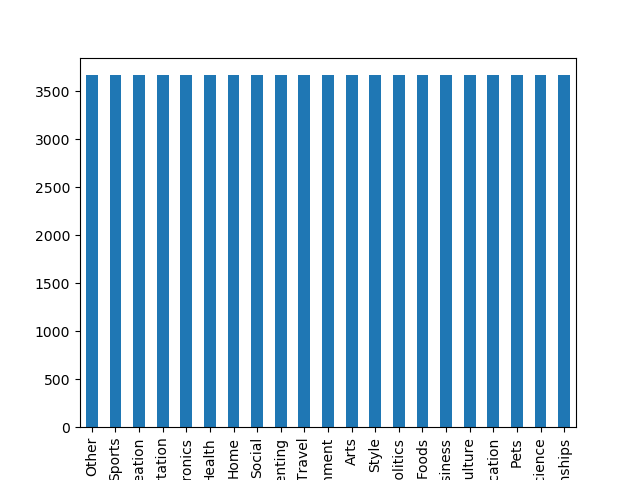
|  |  |
| --- | --- |
| Pre Cleaning  Unique Words: 49652  Total Words: 472462 | Post Cleaning  Unique Words: 24593  Total Words: 206049 |

E. Most Common Bigrams and Unigrams

(The results were interesting)

1. Pre-Processing
2. The first thing we did was to extract the Data from the HTML tags and organized it into a List of Lists with 3 Strings. Subject, Content and Category
3. We created a list of unwanted symbols such as ‘?’, ‘,’ , ‘!’,’$’… and had them removed.
4. All the words in every line were lowercased.
5. If the length of a line was shorter than 2, we had it removed.
6. We removed Stopwords from each String with NLTK.
7. We tried using the Porter Stemmer, though unfortunately it didn’t yield better results.
8. We tried removing terms which didn’t appear often, though to no gain.
9. When the cleaning was done the data was transferred into a DataFrame with 3 Columns: Subject, Content and Category
10. We created a new Combined Column which is a concatenation of Subject and Content.

J.

Having noticed in (1A) that the dataset was dominated mainly by a small number of categories, we decided to try and rebalance the data occurrences with an external library (SMOTE). This yielded better results. 

1. The additional Classifier we used was Naïve Bayes (Sklearn).

We tried tuning the hyperparameters by using an implementation of grid search.

It suggested the following inputs:

Ngram\_range=(1,2), min\_df=0.1, max\_df=0.7

|  |  |
| --- | --- |
| 1NN Accuracy  0.4302 | Naïve Bayes Accuracy  0.483 – Before Rebalancing  0.5605 – After Rebalancing |