**Part 2 – Report**

**Intro:**

We have 50,000 samples that distribute:

Chart, pie chart

Description automatically generated

**Preprocess:**

First, we remove all non-ascii characters meaning punctuations, English letter, and digits with the assumption that these characters don’t give any value.

Second, lemmatization, and word tokenization to count misspell words and different kind of word stem as the same word

Finally, we remove stop words

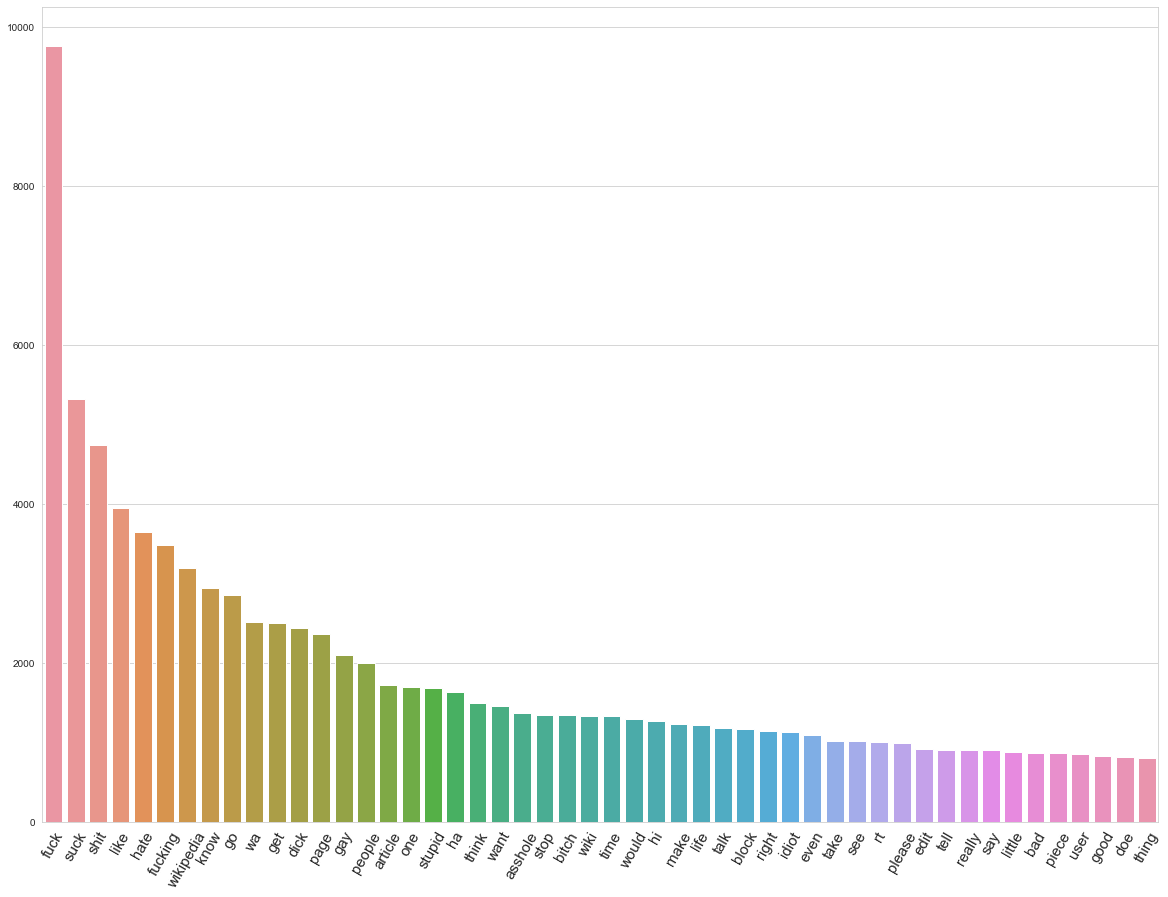
A picture containing text, computer, indoor

Description automatically generated

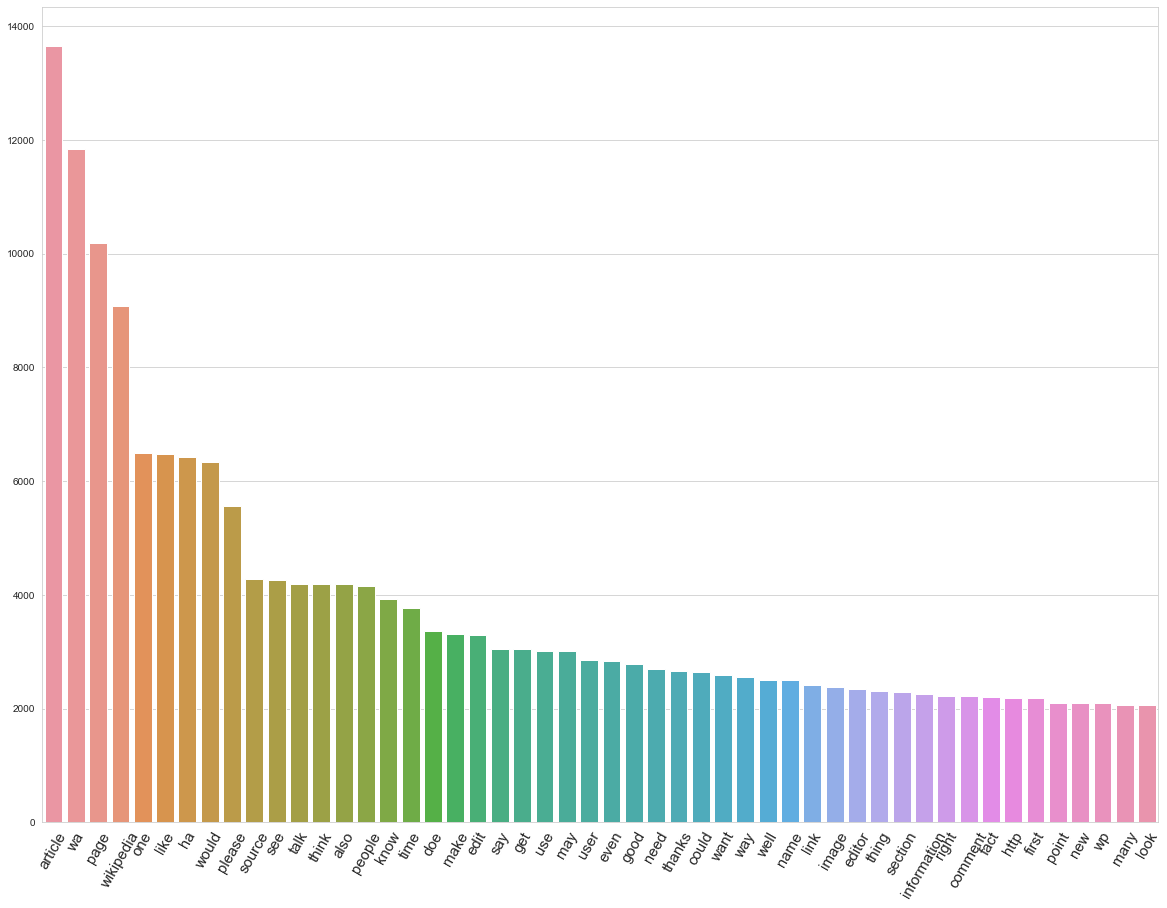
Then, we divide and shuffle the data into 20% test and 80% train.  
There was no nan values in the dataset

**Data Exploration:**

**Most frequent words in bullied class**



As we can below words like thanks, good etc. appears frequently and that make sense with the label



**Logistic Regression**

Note that we try also with bag of words after spell correction, and it doesn’t improve the results

precision recall f1-score support

0.0 0.87 0.85 0.86 6582

1.0 0.72 0.76 0.74 3418

accuracy 0.82 10000

macro avg 0.80 0.80 0.80 10000

weighted avg 0.82 0.82 0.82 10000

