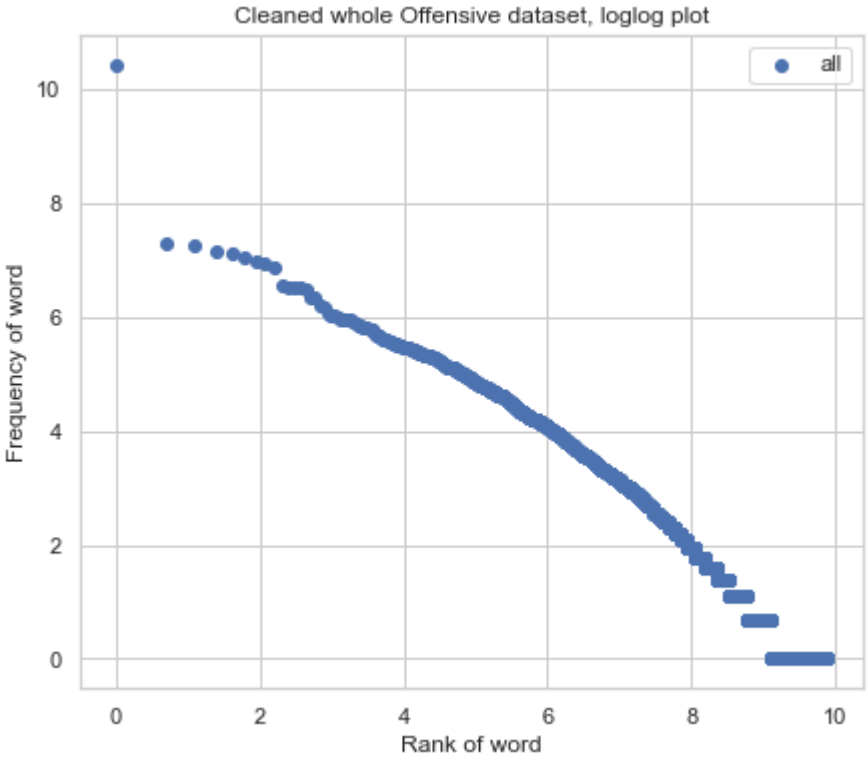
Statistics for whole Offensive dataset:

Top 20 in freq_words
 {'user': 34024, 'liberals': 1485, 'gun': 1427, 'control': 1286, 'antifa': 124
4, 'like': 1172, 'maga': 1067, 'conservatives': 1029, 'people': 968, 'get': 71
3, 'one': 688, 'amp': 682, 'trump': 682, 'know': 668, 'would': 581, 'think': 57
5, 'good': 491, 'right': 476, 'us': 440, 'want': 423}

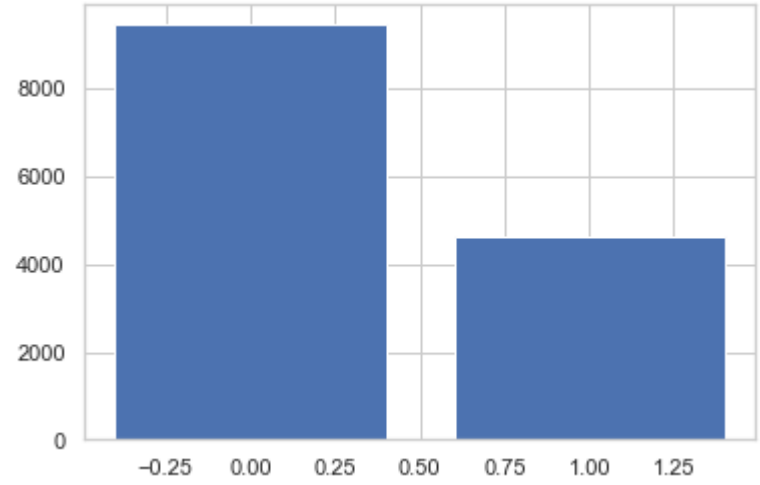
Top 20 least common words in freq_words
 [('countrymen', 1), ('breitbarters', 1), ('progressing', 1), ('reads', 1), ('u
nto', 1), ('optic', 1), ('vetsresistsquadron', 1), ('scout', 1), ('lifetimes',
1), ('obummer', 1), ('plethora', 1), ('peruse', 1), ('lighting', 1), ('torche
s', 1), ('checkmarks', 1), ('verifiedhate', 1), ('lmfaoooooo', 1), ('shitbiscui
t', 1), ('tempe', 1), ('licensereporters', 1)]

The number of characters: 1813028
The number of tokens is: 402395
The number of sentences is: 21458
The average number of tokens per sentence is: 14
The number of unique tokens are: 19557
The tokens ratio is: 0.049
The number of total tokens after removing stopwords are: 176178
Zipf's law, LogLog Plot



Distribution of true labels:

Plot of labels distribution:



	Label no.	Count	Percentage
0	0	9458	67.09
1	1	4639	32.91