Mixed Dataset Summary

Dataset makeup

The following datasets were pre-processed, combined, and shuffled to generate the dataset:

Dataset	Number of Documents	Number of Abusive Documents	Abusive Percentage
Davidson et. al	24,783	20,620	83.2%
Conversation AI	223,549	22,468	10.0%
Impermium	6,594	1,742	26.4%
Combined	**254,926**	**44,830**	**17.6%**

The datasets can be accessed below

- Conversation AI
- Davidson et. al
- Impermium

Pre-processing

The pre-processing steps applied to the documents are listed, in order, in the table below.

- Remove embedded image links
- Remove emojis
- Remove '@' symbols marking twitter handles
- Remove '#' symbols and split following string when CamelCase was used
- Remove hyperlinks and replace with 'url' string
- Replace unicode characters with closest ASCII character
- Force all letters to lowercase
- Remove '?' and '!' symbols
- Remove punctuation
- Remove digits
- Replace character unigrams and 2-grams when 3 or more were present in an un-broken sequence (ex. 'aaa' -> 'a')
- Remove extra space characters