This archive contains Turkish offensive language corpus as used in

OffensEval 2020 shared task

<https://sites.google.com/site/offensevalsharedtask/home>. Please see

the competition web site for more information.

The files under offenseval-tr-training-v1 are the training set

distrubuted during OffensEval 2020. The directory

offenseval-tr-testset-v1/ contains the test set without labels

(offenseval-tr-testset-v1.tsv) and the gold standard labels

(offenseval-tr-labela-v1.tsv) as \*comma-separated\* file.

There is not quoting or special escape characters in this version of

the TSV files. The files contain one instance per line, where fields

are separated by tab. Newlines in the original tweets were replaced

with three space characters. Reading these files with most csv/tsv

libraries with default options will be inaccurate. Either disable

quoting behavior, or simply read the data line-by-line and split at

tab characters manually.

The data and annotations are described in the following paper. Please

cite it if you use this resource.

@inproceedings{coltekin2020,

title={A Corpus of Turkish Offensive Language on Social Media},

author={\c{C}\"{o}ltekin, \c{C}a\u{g}r{\i}},

year={2020},

inproceedings={Proceedings of the 12th International Conference on Language Resources and Evaluation},

organization={ELRA}

}

The annotations are licensed under Creative Commons Attribution license (CC-BY).