﻿========================

OffensEval 2020: Identifying and Categorizing Offensive Language in Social Media (SemEval 2020 - Task 12)

Turkish training data

v 1.0: December 16 2019

https://sites.google.com/site/offensevalsharedtask/

========================

1) DESCRIPTION

The file offenseval-tr-training-v1.tsv contains 31,756 annotated tweets.

The file offenseval-annotation.txt contains a short summary of the annotation guidelines.

Twitter user mentions were substituted by @USER and URLs have been substitute by URL.

Each instance contains up to 1 labels corresponding to one of the following sub-task:

- Sub-task A: Offensive language identification;

2) FORMAT

Instances are included in TSV format as follows:

ID INSTANCE SUBA

The column names in the file are the following:

id tweet subtask\_a

The labels used in the annotation are listed below.

3) TASKS AND LABELS

(A) Sub-task A: Offensive language identification

- (NOT) Not Offensive - This post does not contain offense or profanity.

- (OFF) Offensive - This post contains offensive language or a targeted (veiled or direct) offense

In our annotation, we label a post as offensive (OFF) if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.

Contact

semeval-2020-task-12-all@googlegroups.com