Real-time Hate Speech Detection on Social Media with Spark

Sevak Harutyunyan

Faculty of Applied Statistics and Data Science, YSU

sevak.g.harutyunyan@gmail.com

Introduction

With the massive increase in social interactions on online social networks, there has also been an increase of hateful activities that exploit such infrastructure.

The need in software to effectively and reliably detect such hateful speech is important for analyzing public sentiment of a group of users towards another group, and for discouraging associated wrongful activities.

Models

Logistic regression is widely used to predict a binary response. It is a linear method with the loss function in the formulation given by the logistic loss:

$$L(\mathbf{w}; \mathbf{x}, y) := \log(1 + \exp(-y\mathbf{w}^T\mathbf{x})).$$

Random forest is an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.

$$f \quad \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Gradient-Boosted Trees (GBTs) are ensembles of decision trees. GBTs iteratively train decision trees in order to minimize a loss function.

Results

	precision	recall	accuracy	F1
LR	**0.73**	0.78	**0.77**	0.75
RF	0.7	0.8	0.76	0.74
GBDT	0.72	**0.81**	0.77	**0.76**

Datasets

The dataset we used is publicly available here.

count: 31962 row, 3 col

id : integer = id of a post

label : integer =
0 (no offense) #count (18217)
1 1 (offense) # count (13745)

tweet : string = the post from a user

Architecture