Classifying Hate Speech on Twitter

Steve Donahue

Hate speech, trolling, and bullying on social media have become a serious problem in the last few years. From extra-national actors carrying out cyber campaigns exacerbate political differences to teenagers viciously abusing their peers, the internet has become a hostile place. The goal of this project is to develop a model which classifies racist and sexist language, using a previously labeled data set provided in a Vidhya Analytics challenge located here.

The data provided consists of unique index values, binary labels for hate speech vs normal speech, and the text of each tweet. All tweets have been represented as originally posted, except that the usernames have been removed to keep the posters anonymous. The raw data consists of about 32,000 training tweets and another 17,000 which serve as a validation set. The format for the raw data appears as follows.

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ð□□¦ð□□¦ð□□¦
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ□¦
7	8	0	the next school year is the year for exams.ð□□¯ can't think about that ð□□ #school #exams #hate #imagine #actorslife #revolutionschool #girl

This project will clean, tokenize, and normalize the tweets using standard string processing methods and NLP procedures from the nltk module. The resulting processed tweets will be engineered into various features using the bag of words, ti-idf, word to vector, and document to vector methods. Then, SVM, Logistic Regression, Random Forest, and Light GBM classifiers will be trained on the data set and evaluated using the F1 metric.

The top performing classifiers will then be further tuned for optimal classification.

The ultimate goal of this exercise is to produce a classifier which accurately determined which social media posts are racist or sexist. With a highly accurate model, it may be integrated into social platforms as an optional filter for a user's feed or a way to voluntarily prohibit inflammatory posts from being left on an individual user's wall.

It is the goal of this project to allow social media users the ability to customize their experience by electing not to entertain hate speech where they share their personal stories, images, and experiences.