HATE SPEECH DETECTOR

Nurefşan Müsevitoğlu - 21503062
Yasin Balcancı - 21501109
Semih Teker - 21300964
Balkır Göka - 21400551
Muammer Tan - 21400967

We are planning to use a dataset used in a paper[1] which is a 100.000 English tweets labeled as abusive, hateful and normal. To use this database, we need the permission of the author, we mailed her and waiting answer. In case of not getting permission, we are planning to use an alternative dataset.

The alternative dataset contains 9,925 different posts in English which can be used to train and test our hate speech detector and to perform identification of the posts which include hate speech. The database is in the form of excel file and contains posts and their labels. 1,155 of the posts include hate speech and labeled as 1. Other 8,770 posts do not include hate speech and labeled as 0 [2].

Our aim on this project is to determine the intention of writing the post. In this case, we are going to focus on being hate speech or not. Posts are written in natural language and also they include shortcuts. Therefore, the models should understand the language used in each post and draw a conclusion from them. Additionally, since words could have different meanings depending on the context they are used, the model should understand the meaning of the posts not just by looking at each word but by analyzing it in the context. We are going to try different algorithms to train and test the data and the results are going to be compared to see which algorithm works best for this type of situation.

Techniques That Will Be Used:
Naive Bayes [3]
Logistic Regression [4]
Support Vector Machine (SVM) [5]
Random Forest [6]
Hybrid Class Semantic Classifier (HCSC)[7]
k-NN [8]

REFERENCES:

[1] Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. [Online] Available: https://arxiv.org/pdf/1802.00393.pdf Accessed: March 15, 2019.

[2] Hate_speech_dataset | Kaggle. [Online] Available: https://www.kaggle.com/pandeyakshive97/hate-speech-dataset. Accessed: March 15, 2019.

[3] Course Slide. Available: https://moodle.bilkent.edu.tr

[4] Logistic Regression | Classification. [Online] Available: https://www.youtube.com/watch?v=-la3q9d7AKQ. Accessed:March 17, 2019.

[5] An Idiot's Guide to Support Vector Machines. [Online] Available: http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf. Accessed: March 17, 2019

[6] Random Forest - short text classification. [Online] Available: https://stats.stackexchange.com/questions/343954/random-forest-short-text-classification. Accessed: March 17, 2019.

[7] A new hybrid semi-supervised algorithm for text classification with class-based semantics. [Online] Available: https://www.sciencedirect.com/science/article/pii/S0950705116301848?via%3Dihub#. Accessed: March 17, 2019.

[8] Text Classification using K Nearest Neighbors. [Online] Available: https://towardsdatascience.com/text-classification-using-k-nearest-neighbors-46fa8a77acc5 Accessed: March 17, 2019.