Established in 2006, Twitter is one of the most popular websites contributing to public opinion formation. It has 450 million monthly active users, who interact with each other by sharing their ideas feelings, and opinions with text and visuals. Twitter aims at giving everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers", Yet, Twitter causes exposure to unwanted tweets, especially those containing hateful expressions. Although, Twitter restricts abusive behaviors, such as harassment or expressing hate. the mechanisms it uses to detect abusive behavior are very limited: Twitter has detection algorithms and users can report tweets that may end up in down-ranking, removal of a tweet, or, suspending an account. However, it is hard to say that these mechanisms are efficient and there is a huge demand for models that detect hate tweets and filter them to offer a satisfactory user experience.

Because of this market need, the models developed here have been trained on Twitter data and provide an algorithm, which can be used to detect hate speeches.

The training data has 10197*3 values. It does not have any null value. So there is not any requirement for the deletion of data. The data has three columns: id, tweet, and label.

1- 'id' is the number of row.

2- 'tweet' contains the user, text and hashtag. There are also some special characters such as '(' '&', which require a special cleaning process)

Yet before the training model, some more steps should be taken to get better results. First of all, after removing stopwords lemmatization is used. Because it gives more reliable results than stemming[1]. Secondly, labels are dis-proportional: there is are 20 entries for 0 and only 2242 for 1. Oversampling and under-sampling are two convenient ways to solve this problem. [2] Although over-sampling gives better results, since it has already been deployed on the same data[3], I used under-sampling.

Regarding models, classification models are selected. First, the Random Forest classifier provided a 0.87 f1 score Logistic Regression as the second model gave 0.88. Lastly, Gradient Boosting Classifier resulted in a 0.84 f1 score. Yet it should be said that this model is trained with under-sampled data. If over-sampled data were used, the 1 score would be 0.94.

[1] https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/#:~:text=Stemming%20and%20lemmatization%20are%20methods,the%20word%20is%20being%20used.

[2] https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[3] https://www.kaggle.com/code/pardhasaradhireddy/hate-speech-detection-f1-score-99

As the report offers EDA analysis of Twitter data and a model detecting hate speeches in tweets, the Logistic Regression is the most effective model with a 0.88 F1 score and accuracy.