# Changes in Baselines of Hatespeech Detection with pretraining and Transformers

Anonymous ACL-IJCNLP submission

## Abstract

In this project we aim to construct a new representative baseline for Deep Neural Nets on the HatefulTwitter dataset proposed by **auto˙hatespeech**. For this we implement a Transformer-based Architecture on the data and try to push the baselines proposed by the authors of the HatefulTwitter dataset. Then we will aim to provide some evaluation of explainability and performance, also with respect to industry usecases.

## 1 Introduction

In their paper **auto˙hatespeech** propose a tweet dataset with 25.000 entries and use some basic methods to establish baselines, especially interesting is that BoW approaches seem to have major problems distinguishing between hatespeech and offensive language since the distributions are so similar. This might indicate that context sensitive methods are to be used. For example Transformers/BERT-models yield such a functionality.

## 2 Review of Attention˙is˙all˙you˙need and BERT

Basicly build a foundational understanding for attention and the original BERT model. (Also important for myself since I just played around with the models in practice but never really understood the paper)

## 3 Review of auto˙hatespeech and BERT˙Transferlearning˙Hate

Compare approaches of the **auto˙hatespeech** to **BERT˙Transferlearning˙Hate** especially evaluate the feature engineering and see if there are things to expanded on.

## 4 Dataexploration

Documenting the dataexploration to motivate a lot of the feature engineering. Also gathering results on the HatefulTwitter dataset, since there are paper reviewing Hatespeech Datasets, they usually already have some facts about this perticular dataset gathered.

## 5 Feature Engineering and Parameter Finetuning

This is only fillable after I have some working code and also did some more digging on different papers and what they used as features.

## 6 Explainability and the Complexity Trade off

Test the Explainability by evaluating both models against the baselines from [**HateXplain**]

## 7 Conclusions