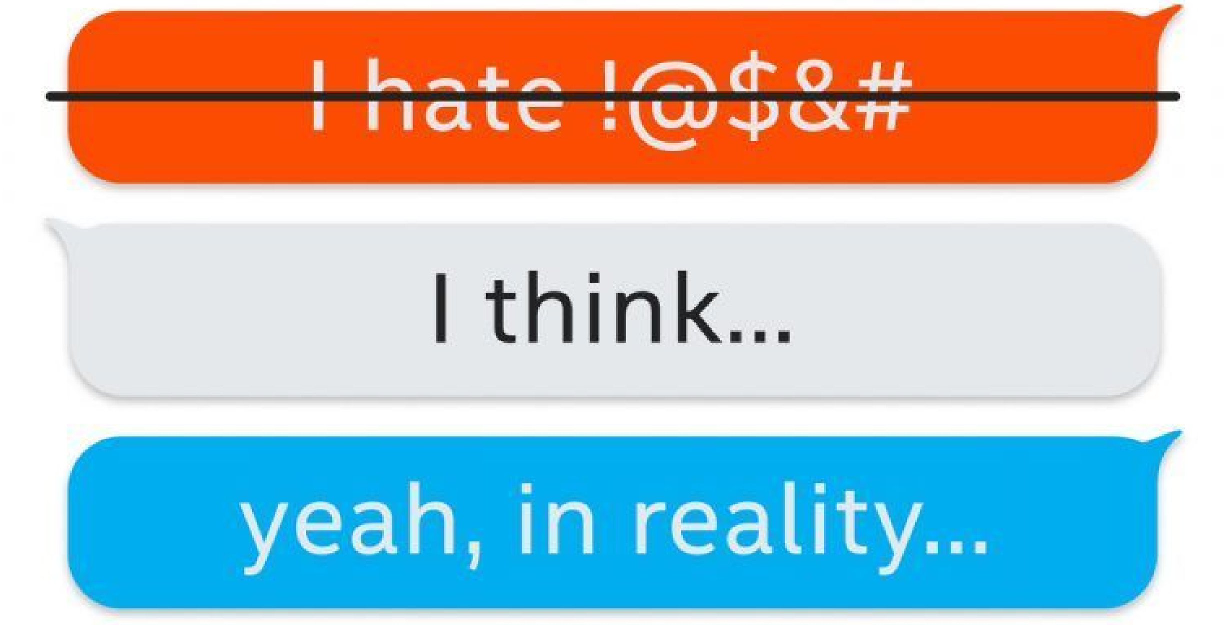
BTP-SEMINAR REPORT
17CS30018
KANISHK SINGH

GENERATIVE MODELS TO COUNTER HATE SPEECH



Introduction

The paper proposed a benchmark dataset which doesn’t ignore the conversational context and opens a lot of room for research on the task of generative hate speech intervention that generates responses to intervene during online conversations that contain hate speech.

Literature Survey Done So Far :

1.) [1909.04251] A Benchmark Dataset for Learning to Intervene in Online Hate Speech

2.) [1406.1078] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

3.) [1408.5882] Convolutional Neural Networks for Sentence Classification

4.) [1312.6114] Auto-Encoding Variational Bayes

5.) [1511.06732] Sequence Level Training with Recurrent Neural Networks

Tasks

- **Binary Hate Speech Detection (Completed)**

Two fully labeled datasets collected from Gab and REDDIT which contain conversations from different users and the indexes of the human hateful comment and the corresponding human intervention.

For Binary Hate Speech Detection, the authors of the paper have implemented four different-models mainly the Logistic Regression(LR), Support Vector Machine(SVM), Convolutional Neural Network(CNN) and Recurrent Neural Network.

I was instructed to implement the RNN model which consisted of a Bidirectional LSTM and a dense layer and the word-embeddings were initialized with Google News Word2Vec Model. The methods are evaluated on testing data randomly selected from the dataset with the ratio of 20%. The methods are evaluated using F-1 score, Accuracy and I was able to replicate the scores which were comparable as to the ones mentioned in the paper.

GITHUB LINK: Binary Hate Speech Detection and Intervention

- **Generative Hate Speech Intervention (In Progress)**

The datasets are really rich because they provide the conversational context rather than treating a sentence as an isolated instance. For the intervention task, the paper presented three models mainly Seq2Seq, Variational Auto-Encoder (VAE) and Reinforcement Learning method. I was instructed to implement the VAE model which consisted of 2 bidirectional GRU layers for the encoder followed by two independent linear layers to calculate mean and variance of the distribution of the latent variable separately.

I am midway through this task and I will complete it by Sunday 5th April positively.