Goal

- Classify comments with a machine learning application and flag them as toxic and non-toxic.
- Reduce the quantity of hate speech on the internet in attempts to reduce online bullying and promote healthy, reasonable discussion in social media.

Methodology

- Preprocess dataset to automate the cleaning of noise, grammar, and spelling errors.
- Convert words and sentences into numerical representations by implementing paragraph vectors through doc2vec.
- Use the word and sentence embeddings to train a neural network that takes the embeddings classifies them as toxic or non-toxic.
- Test different network models (such as support vector machines, k-nearest-neighbors, etc...) to find the model that works best.
- Develop an end-to-end model that unites the paragraph vectors implementation and classification in one complete model/objective function such that sentence representations and classification can be learned in a unified way.

Motivations and Objectives

- Motivations
  - The Internet is burdened with unnecessary bullying.
  - Manually moderating all interaction is not feasible.
  - User reporting can be either unused or unreliable.

- Objectives
  - Automatically flag comments to prevent bullying and toxic messages.
  - Create a safe environment on the internet that promotes healthy discussion.

Research Challenges

- Convert words and sentences into vector form for analysis and classification in a neural network.
- Approximate the semantic meaning of a word dependent on its context.
- Account for sarcasm, grammatical errors, spam, non-alphanumeric characters, website URLs, and other non-helpful text.

Acknowledgement

We would like to thank Gensim for their powerful natural language processing library, and also Professors Kostas Bekris and Yongfeng Zhang for their help and insights in AI/machine learning.

References