基于BERT的网络暴力判别分类

小组成员
李若邻 2006500019
魏子华 2006500020
黎智康 2006500023

1. 题目背景介绍：

随着社交媒体在各个年龄段的使用越来越普遍，绝大多数人依靠这种必不可少的媒介进行日常交流。社交媒体的普及意味着网络欺凌可以在任何时间任何地点有效地攻击到任何人，而且是相对匿名的。互联网的普及使得这种人身攻击比传统的霸凌更难以阻止。

2020年4月15日，联合国儿童基金会针对2019冠状病毒疾病大流行期间，由于学校普遍关闭、网络线上时间增加和面对面社交活动减少而导致网络欺凌风险增加的情况发出了警告。如今，网络欺凌的统计数据令人震惊: 36.5% 的中学生感受到网络欺凌，87% 的学生观察到网络欺凌，其影响范围囊括了从学习成绩下降到抑郁到产生极端的念头。

因此，我们可以通过创建模型来自动标记潜在有害的推文来应对这种情况。

2. 数据集简介

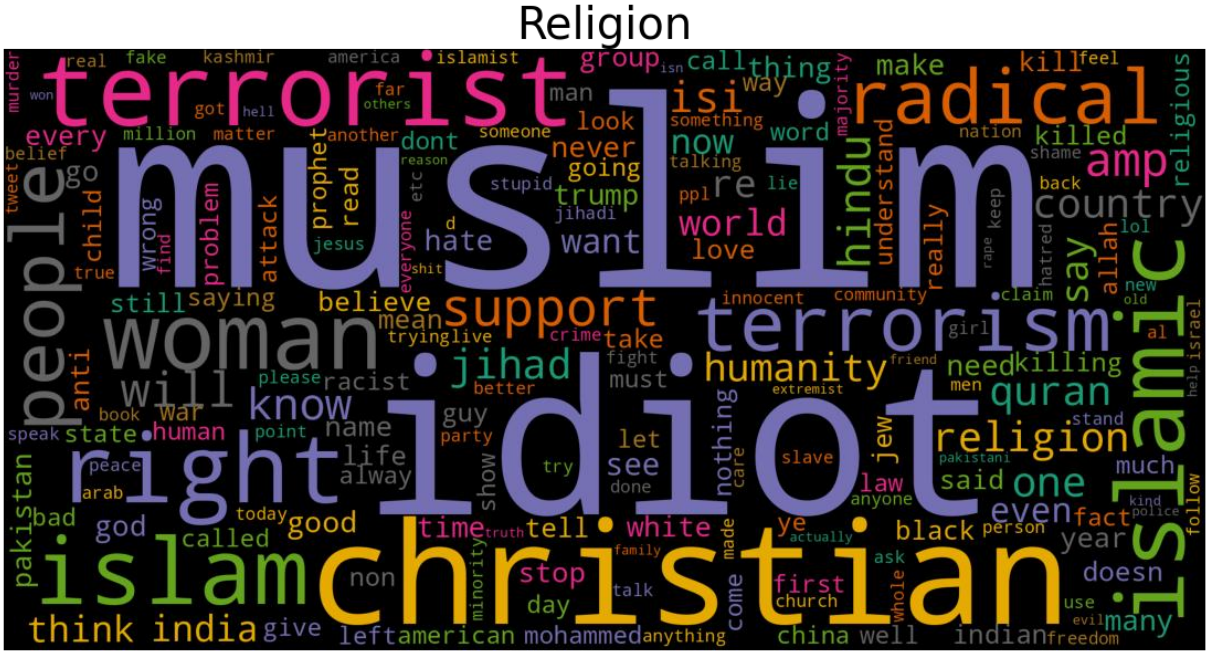
Kaggle数据集:https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

数据集内容展示:https://www.kaggle.com/code/lizakonopelko/cyberbullying-on-twitter-visualization/notebook

数据集有以下类别(组别):

tweet_text	评论的内容(英文字符)				
cyberbullying_type	Religion(宗教)	age(年龄)	gender(性别)	Ethnicity(种族)	not_cyberbulling(非网暴评论)

词云图举例:



3. 作业预期目标：

1.使用有效的数据预处理方法
2.划分测试集与训练集
3.导入bert模型，在训练集上进行预训练。
4.使用测试集预测一条推文是否与网络欺凌有关，并给出对应的网络暴力类型。
正确率应该达到80%以上

4. 使用的方法：

使用pytorch上的bert模型进行分类。BERT的全称为Bidirectional Encoder Representation from Transformers，是一种基于transformer的双向编码表征算法，其中的transformer基于多头注意力机制，同时，bert堆叠了多个transformer模型来训练。通过这种机制，bert模型能够学习到一个单词前后的上下文的关系，能够得到更好的效果。