Draft Project Report

Title: Project: Computer Science Project (CSEMCSPCSP01)

Topic: Cyberbullying Detection in Social Media Comments

Student ID: 4252553

Student Name: Panchami Bandihalli Varadaraju

Tutor: Mugdha Kashyap

1. Abstract

Cyberbullying is a significant social problem, yet its automated detection is reputedly difficult due to the complexities of human language, context, and sarcasm. This project implements a machine learning model to classify tweets into six categories: **age, ethnicity, gender, religion, other_cyberbullying, and not_cyberbullying.** We use a pre-trained **BERT (Bidirectional Encoder Representations from Transformers)** model and fine-tune it on a dataset of 9,996 labelled tweets. This result empirically confirms that while BERT is powerful, advanced strategies are needed to capture the subtle, contextual nature of online harassment.

2. Introduction

Online social media platforms have become central to communication, but they have also enabled new platform for harassment, known as cyberbullying. This behavior can have severe and lasting psychological impacts on victims. The technical challenge lies in language itself. The line between harmless In-group joking, sarcasm, and genuine malicious intent is often blurry. A statement like You're a terrible person could be a joke between friends but for other it is a targeted attack for some people with lots of sentiment and emotions. so from this project investigates the use of a state-of-the-art natural language processing (NLP) model, BERT, to tackle this problem. The goal of this project is to build and evaluate a multi-class classifier that can not only identify cyberbullying but also distinguish between different types, providing a more granular analysis.

3. Related Work

a. **A significant body of research exists on automated cyberbullying detection**. In this project builds upon this, moving from traditional machine learning models to more advanced transformer-based architectures. Traditional Machine Learning (Baselines): Early approaches to this problem treated it as a standard text classification task. Researchers like Dadvar et al. (2013) and Nahar et al. (2014) used features like TF-IDF (Term Frequency-Inverse Document Frequency) and n-grams to represent text. These features were then fed into classifiers like Support Vector Machines (SVM), Naive Bayes, and Logistic Regression.

- Model Used: These models are very effective at identifying explicit, keyword-driven harassment
- Limitation of this method: They completely fail when context, sarcasm, or nuance is involved. They have no understanding of semantics.

b. **Early Deep Learning (LSTMs & CNNs):** With the rise of deep learning, researchers began using Recurrent Neural Networks (RNNs), specifically LSTMs (Long Short-Term Memory), and Convolutional Neural Networks (CNNs). These models were able. to learn from the sequence of words, which was an improvement.

- Model Used: Models by Pitsilis et al. (2018) showed that LSTMs could outperform traditional models by capturing some short-term contextual clues.
- Limitation of this method: Their understanding of context is still shallow (often unidirectional) and they struggle with long-range dependencies in text.

4. Technical Background

The tool used in this project is transfer learning via the BERT model.

- **BERT (Bidirectional Encoder Representations from Transformers):** BERT is a deep learning model developed by Google. Unlike older models that read text left-to-right. BERT reads the entire sequence of words at once. This bidirectional nature allows it to learn deep contextual relationships.

  For example, it can understand that the word bank means something different in river bank vs. money bank based on the words around it.

5. Method

The methodology follows a standard supervised machine learning pipeline.

a. Dataset: I have downloaded the data set from kaggel cyberbullying_tweets.csv dataset, which contains 9,996 rows. Each row has a tweet_text and its corresponding cyberbullying_type.

b. Labels: The six target classes are: age, ethnicity, gender, religion, other_cyberbullying, and not_cyberbullying. These are encoded into numerical labels 0-5 for the model.

c. Model: The Bert-base-uncased model was chosen as the base. This is a 12-layer Transformer model with 110 million parameters.

d. Data Splitting: The data was split into a training set and a validation set to test the model's performance on unseen data.

e. Preprocessing: Tokenization: The Bert Tokenizer converts raw text into a format BERT understands. This includes splitting words into sub-words (e.g., "bullying" -> "bulli" + "##ng"), adding special tokens like [CLS] (start) and [SEP] (end), and converting tokens to numerical IDs.

6. Implementation

The solution was implemented in Python using Google Colab, utilizing its free GPU access.

- Core Libraries: Pandas for data handling, Transformers (by Hugging Face) for the BERT model and tokenizer, PyTorch as the deep learning framework, and Scikit-learn for label encoding and evaluation metrics.

```
                                        608.4/608.4 kB 12.6 MB/s eta 0:00:00
Requirement already satisfied: matplotlib-venn in /usr/local/lib/python3.12/dist-packages (1.1.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.12/dist-packages (from matplotlib-venn) (3.10.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.12/dist-packages (from matplotlib-venn) (2.0.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.12/dist-packages (from matplotlib-venn) (1.16.3)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (1.3.3)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (4.60.1)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (1.4.9)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (25.0)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (3.2.5)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.12/dist-packages (from matplotlib->matplotlib-venn) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.7->matplotlib->matplotlib-venn) (1.17.0)
```

Fig: Installation of dependencies

- Data Pipeline: A custom Cyberbullying Dataset class was created in PyTorch. This class handles the tokenization of one tweet at a time. This is fed into DataLoader, which batches the data and shuffles it for effective training.

```
Loading file: cyberbullying_tweets.csv
Data Loaded: (9996, 2)
```

output actions

1 to 5 of 5 entries Filter

index	tweet_text	cyberbullying_ty
0	Every single one is a girl that would have bullied me in high school.	age
1	Weâ□□ve shown my kids a lot of #80smovies and Iâ□□m pretty sure theyâ□□re convinced i never made it through a single school lunch period without a bully flipping a lunch tray in a dorkâ□□s face or a slow clap. Sadly most of my lunches were just eaten with no applause.	age
2	The only reason i didn't get bullied for these things towards the end of high school is because i managed to become friends with the most popular girl at school. Then suddenly all the other kids started trying to get into what me and my handful of friends were into.	age
3	People who say that high school cis boys would pretend to be trans to see girls naked in locker rooms really think that coming out to your family and classmates, taking hormones, and being bullied is easier than asking a girl out on a date	age
4	I super relate to this story. I was bullied in HS, reason why I never go to our reunions nor do I like it when I ran accross my HS classmates. I never mention this to my parents but I never liked going to school because of bullies	age

Show 25 per page

Fig: Data Pipeline

- **Training Process:**

  a. Model: Bert for Sequence Classification was loaded, pre-trained, and configured for 6 output labels.

  b. Optimizer: AdamW was used, which is the standard optimizer for BERT.

  c. Scheduler: A get_linear_schedule_with_warmup was used to adjust the LR during training, which helps model stability.

Fig: Model Training

d. Epochs: The model was trained for 3 full epochs. After epochs is done the model is saved in the path '/content/bert_cyberbullying_model' in project file.

7. Risk Analysis

The risk analysis evaluated in this project by using model's implementation as evidence which are technical, and project-related risks, by using direct results from the project.

a. **TechnicalModel Overfitting:** The model learns the training data well and doesn't generalize to new, unseen tweets.
   Ex. The model is static. It was trained on a single CSV file and saved. It has no connection to live data.

b. **Project-related risks / Dataset Bias:** The cyberbullying_tweets.csv dataset is not representative of all bullying or all forms of English.
   Ex. The model is trained *only* on the data loaded from cyberbullying_tweets.csv. The labels are limited to the 6 classes encoded in cell 18. So if I give other data or different statement the model won't identify so this is the main drawback and analysis made till now.

Git link

https://github.com/PanchamiVaradaraju/Cyberbullying-detection-in-Social-Media-Comments