Project Proposal

Kemal Şahin

Student Number: 2200765021

Burak Kurt

Student Number: 2200765010

Course Name: AIN420

Contents

1 Introduction **2**

1.1 Project Overview 2

1.2 Objectives 2

1.3 Scope 2

2 Methodology **2**

2.1 Data Collection 2

2.2 Data Preprocessing 3

2.3 Model Overview 3

2.3.1 Step 1: Offensive Content Detection 3

2.3.2 Step 2: Classification of Offensive Content 3

Abstract

This project proposes the development of a hierarchical encoder transformer model for the purpose of multi-label text classification, specifically designed to identify and categorize offensive content within textual data. Our model aims to first determine the presence of offensive content and then classify such content into four distinct categories: Sexist, Racist, Profanity, and Insult. Utilizing a dataset comprising 81,800 rows, each carefully annotated with multi-labels through the aid of Label Studio from various sources including Twitter and Kaggle, we plan to employ advanced natural language processing techniques to solve this challenge.

1 Introduction

1.1 Project Overview

In the digital age, social media platforms and online forums have become central to public discourse. While these platforms enable free expression, they also pose significant challenges in content moderation, especially concerning offensive content. Our project focuses on developing a hierarchical encoder transformer model for multi-label text classification to address this issue.

1.2 Objectives

The primary objectives of this project are as follows:

- To develop a hierarchical encoder transformer model capable of pick out between offensive and non-offensive content with high accuracy.
- To further classify identified offensive content into four distinct categories: Sexist, Racist, Profanity, and Insult.
- To utilize a preprocessed and hand-labeled dataset of 81,800 rows for model training and validation, ensuring a robust and reliable classification system.

1.3 Scope

This project is limited to the development and validation of the proposed model using the specified dataset. It focuses on Turkish-language content sourced from Twitter and Kaggle. While the model is designed to be adaptable to various types of textual data, its initial validation will be limited to the dataset at hand.

2 Methodology

2.1 Data Collection

The dataset for this project was compiled from a variety of sources, including publicly available datasets on Kaggle and a collection of tweets. The goal was to gather a diverse set of textual data that reflects a wide range of language use, including different forms of offensive content. A total of 81,800 rows of text data were collected, ensuring a rich dataset that can support the nuanced classification tasks required by our project.

2.2 Data Preprocessing

The collected data underwent a comprehensive preprocessing phase to prepare it for use in training the model. This phase included:

- **Cleaning:** Removal of URLs, special characters, and non-textual information to focus on the linguistic content.
- **Normalization:** Standardization of text to a consistent format, including converting all text to lowercase and correcting common misspellings.
- **Tokenization:** Breaking down the text into individual words or tokens to ease the analysis made by the model.
- **Labeling:** Each piece of text was meticulously annotated with relevant labels (Not offensive, Offensive, Sexist, Racist, Profanity, and Insult) using Label Studio, enabling precise multi-label classification.

2.3 Model Overview

The project uses a hierarchical encoder transformer model, designed to first pick up whether content is offensive and then classify the offensive content into one of four categories. This two-step approach allows for more throughout understanding and classification of text.

2.3.1 Step 1: Offensive Content Detection

The first step involves a binary classification model that categorizes text as either offensive or not. This model serves as a filter, ensuring that only content flagged as potentially offensive is passed on to the more detailed classification step.

2.3.2 Step 2: Classification of Offensive Content

Texts classified as offensive are then processed by a second model that categorizes them into one of four specific types of offensive content: Sexist, Racist, Profanity, and Insult.