Lexical Text Simplification & Detoxification

Group Number 42

PROBLEM STATEMENT

Problem Statement

Our group aims to build a system to replace complex and difficult to understand words in a text with less complex words.

Additionally, we worked towards detecting offensive text to detoxify it so that it can be viewed by a large audience without repercussions.

MOTIVATION

Motivation behind our idea

01
People with different cognitive and linguistic abilities struggle to understand the complexity of texts

02
Limits their access to information and impedes their ability to participate in daily activities

03
Address this challenge by automatically simplifying the language and structure of text while preserving the original meaning

04
Moderates the content available online and makes it less toxic and relatively safer to be out there on the web.

LITERATURE REVIEW

Literature Review

Text Detoxification using Large Pre-trained Neural Models

The authors describe a dataset of toxic text and evaluate the performance of several pre-trained language models in detecting and removing toxic language. They find that using an ensemble of models further improves performance.

The authors also propose a method for identifying and replacing specific toxic language with neutral alternatives. They show that this approach can effectively remove toxic language.

Literature Review

Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

The authors train a transformer-based language model on a large corpus to learn representations of language. They fine-tune the language model on a dataset of toxic text and corresponding civil rephrases.

The method involves identifying the toxic portions of the input text and generating alternative phrasings of those portions that are more civil.

Methodology

LEXICAL TEXT SIMPLIFICATION

LEXICAL TEXT SIMPLIFICATION

Identifying complex words

- Textstat library has a list of simple english words

- Used function 'difficult_words' from the textstat library

- Returns a word as complex if it is not present in the list of simple english words

Finding replacement words

- masked the identified complex word and utilized the MLM (Masked language modeling) to predict the appropriate words
- used the zipf frequency of the words to calculate complexity
- complex word in each sentence was replaced with one of the proposed new words
- Checked similarity between the original sentence and the modified sentence was found using cosine_similarity
- In case of multiple complex sentences in a word, all the possible permutations of the best 2 word lists of all the complex words were considered.

Methodology

DETOXIFICATION

DETOXIFICATION

- We have used the Dataset for training purposes. We have also used a pre-trained T5 model (an LLM model) for the paraphrasing section in order to retain the context of the input text.

- The code first detoxifies the input sentence using synonym replacement. For this, we detect and classify the toxic word into a category of toxicity which is then mapped to a set of appropriate synonyms.

- Then we move on to the paraphrasing, where we take in the original sentence and the detoxified sentence in order to bring back any context which was lost due to synonym replacement.

NOVELTY

Novelty

In our project, we have implemented a novel technique that combines four tasks to effectively remove toxicity from text.

- The first task involves **detecting toxic words**, followed by categorizing these words into labels.

- We then **map each label** to their **most appropriate synonyms**, achieving high accuracy in the process.

- We utilize a **paraphrasing technique** that ensures the original context of the text is retained or restored.

Results

DETOXIFICATION

Original vs detoxified text

01

Original text

black racist crap still gm website stopped talking black people long ago afraid life since 8090 crime black people kinda makes sense it oh wait thats profiling driving black know leftist marxists blm nutjobs change names fast one keep up facts speak like this wash away letting identity marxists spew racist hatred white people think not commit crimes world f would give time day clean sht first people problem.

02

Detoxified Text

A black racist deterrent has been on the internet long before black people's life since the 1980s, but black people's profiling has made it seem that this washout is not about tolerating bigotry.

Results

LEXICAL TEXT SIMPLIFICATION

Original vs simplified text

01   **Original text**
The door to his dilapidated residence was locked but the police knocked down the door with full force and captured him.

02   **Increased simplicity score**
Simplicity of original text is: 123.61
Simplicity of simplified text is : 133.58

03   **Simplified text**
The door to his run-down home was locked but the police knocked down the door with full force and arrested him.

Thanks!