Project Proposal

Hate Speech Detection and Customer Review Summarization

Natural Language Processing

Team Members

[redacted]

1. Motivation:-

I think in modern society customer reviews serve as an important tool for companies decision-making. Still, such reviews may include hate speech or other vulgarity that perverts feedback and has a negative impact on the brand's image. When applied to such content, the process of filtering and extracting useful information has to be done manually, which of course, is a time-consuming task. This is what this project is set out to do, to develop a system that can

a) identify and filter hate speech in multiple languages.

b) produce a summary of genuine customer review in a precise and coherent manner.

The objective is to assist different companies in making right decisions and considering valuable and appreciable comments, leaving no consideration for toxic comments.

2.Significance:-

This work is relevant because it addresses two significant problems: content moderation and review summarization within the same intervention. In this way the system keeps filtering hate speech, and thus only non-offensive content is summarized which is suitable for businesses to preserve their good image. Thirdly, the use of the summarization enables one, or someone else, to easily get an overview of large amounts of customer reviews from which one can quickly make decisions. These tasks integrated enable result in better user experiences and at the same time ensure that businesses receive reliable feedback.

3.Objectives:-

•Detect Multilingual Hate Speech: Task: Build a model to both classify current adaptation and remove any form of abusive language in the received customer reviews from multiple languages.

•Summarize Non-Offensive Content: Provide brief extracts of non-controversial feedbacks to businesses whilst including specifics to help them act on the feedbacks.

•Enhance User Experience: In other words, by making the content free from toxic content and also giving out good summaries of the feedbacks given then the products and services offered by these firms can be enhanced.

4. Features

•Multilingual Hate Speech Detection: An improved version of the BERT model for multiple languages identifies cases of use of foul language.

•Review Summarization: T5 or BART model produces short summaries of safe text, that white masks is non-offensive articles' summarization.

•Integrated Pipeline: In polite, all the explicit, nasty, or otherwise risky reviews are removed while those with clean decent messages go through the process of being summarized for action.

5. Datasets Overview

1. OLID Hate Speech Dataset: Previously used in the detection of multiple languages with particular emphasis on obscene data. From this dataset, a model for detection of hate speech in social media posts including customer reviews will be learnt.

2. Amazon Product Reviews Dataset: This dataset consists of customers' feedbacks about the products that they use. It will then be used to scan for hate speech before being employed in creating shorter and comprehensive summaries of customers' feedback.

Action plan using Method 2 :- We will train the models of the two tasks individually – particularly the hate speech detection and the summarization subprocess – then compile them into a sequence via which the output of one subprocess is the input to the next subprocess.

Step 1: Preprocessing

•OLID Dataset: Pre-process multilingual text for hate speech identification: cleaning and tokenization.

•Amazon Reviews: Preprocessing of the review includes, erase unwanted data, word tokenization and formation of the review for summarization.

Step 2: Understanding of the model training and fine tuning

1. Hate Speech Detection:

•Model: The OLID dataset is used to fine-tune mBERT (Multilingual BERT).

•Purpose: Filter out inflammatory comments with regards to reviews before proceeding to come up with a summary.

2. Review Summarization:

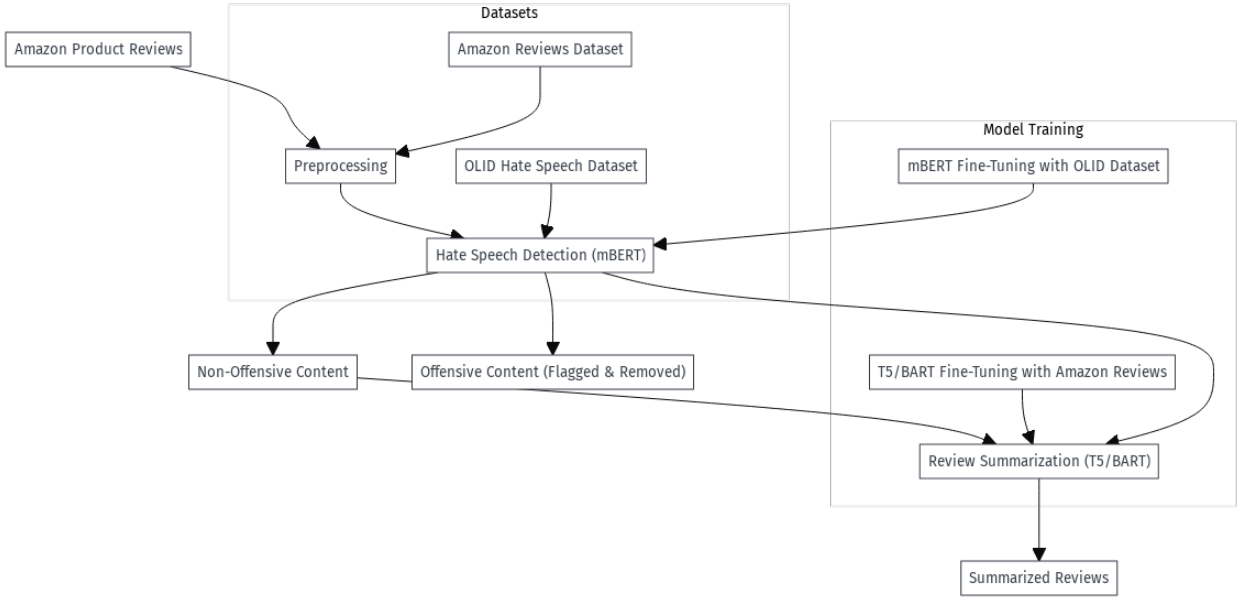
•Model: Prune T5 or BART specifically to condense non-offensive review documents.

•Purpose: Make brief conclusions based on the feedback that consumers provide.

Step 3: Model Integration

- First, reviews are pre-processed based on the hate speech detection model.
- The summarization model is then used in summing up non-offensive reviews.

6. Visualization:-



Expected Results:-

•Cleaner Summaries: To obtain reliable and actionable insights only non-offensive reviews are summarized.

•Automated Filtering: Potential mockery or offense is addressed by being filtered out therefore business get good feedback.

•Multilingual Support: The system supports multiple languages which is effective in all the world.

Conclusion:- This system coupled with hate speech detection and review summarization offers an efficient service in generating a clean summary of customer's reviews for businesses' use to enhance their decision making while fulfilling the responsibility of moderating contents with hate speech technology.

Dataset Links:- https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews

2) https://drive.google.com/file/d/1Tksi8UyzW-drFWd7maGr7MoHVa-VHQCO/view

Github Link: https://github.com/invika/Hate-Speech-Detection-and-Customer-Review-Summarization