Age-rating movies classification based-on subtitle

20.04.2023

—

Baktash Ansari

Project Proposal

Overview

With the increasing demand for media content and films, and the rise of film platforms and discussions about films on social networks, access to films and series has become much easier and more accessible for children, individuals under the legal age, and those who have problems to watching inappropriate content. This also means that families and acquaintances of these individuals may want to evaluate the level of inappropriate or undesirable scenes in these films before watching them, and this can be done with subtitles and the textual content of the films.

My goal is to implement a language model that allows families and content producers to identify the age-rating of a film based on its subtitles. The applications of this language model are not only for film classification but also for comparing multiple subtitles and finding a suitable subtitle with less sensitive content for the film.

Additionally, the collected subtitles for films in different age groups can be analyzed before training the language model, and various features can be obtained for each age-rating.

Goals

1. Analyzing collected subtitles for films in different age groups and obtaining various features for each age rating.
2. Facilitating the comparison of multiple subtitles and the selection of subtitles with less sensitive content for films.
3. Build a language model that can accurately classify input subtitles based on age rating labels

Milestones

Data Collection :

Since our data is subtitles of films, it is in spoken form. The two main languages for the data are Persian and English, which are the languages of the film subtitles. However, if possible, I will use multilingual language models to see the results of the model on multiple languages. Since English subtitles for films are more standardized and I can indicate the emphasis of the text with lowercase and uppercase letters, as well as question and exclamation marks, this language is my first priority.

The data is divided into several categories based on the age restrictions of the films on IMDB, which are as follows:

- G - General Audiences: Suitable for all ages, including children.
- PG - Parental Guidance Suggested: Some material may not be suitable for children. Parents are urged to give "parental guidance." May contain some mild language, some sexual content, and some violent situations/action.
- PG-13 - Parents Strongly Cautioned: Some material may be inappropriate for children under 13. Parents are urged to be cautious. May contain some profanity, some sexual content, and/or some intense violence.
- R - Restricted: Not suitable for children under 17. Contains adult themes, adult activity, hard language, intense or persistent violence, sexually-oriented nudity, drug abuse, or other elements.

- NC-17 - No One 17 and Under Admitted: Clearly adult. Children are not admitted. Contains explicit adult content.

I can use the following websites to collect data :

Opensubtitle :

This site contains a large collection of film subtitles in various languages, which probably also provides an API for us.

IMDB :

I can use the imdb website for the age-rating of films.

I can also use other websites that provide subtitles.

Data Preprocessing :

Since the collected data is in files with the srtformat, I first need to extract all the text from the files. Then I will preprocess the data by removing noise and irrelevant information, such as stop words and punctuation. I will also tokenize the text and remove any special characters.

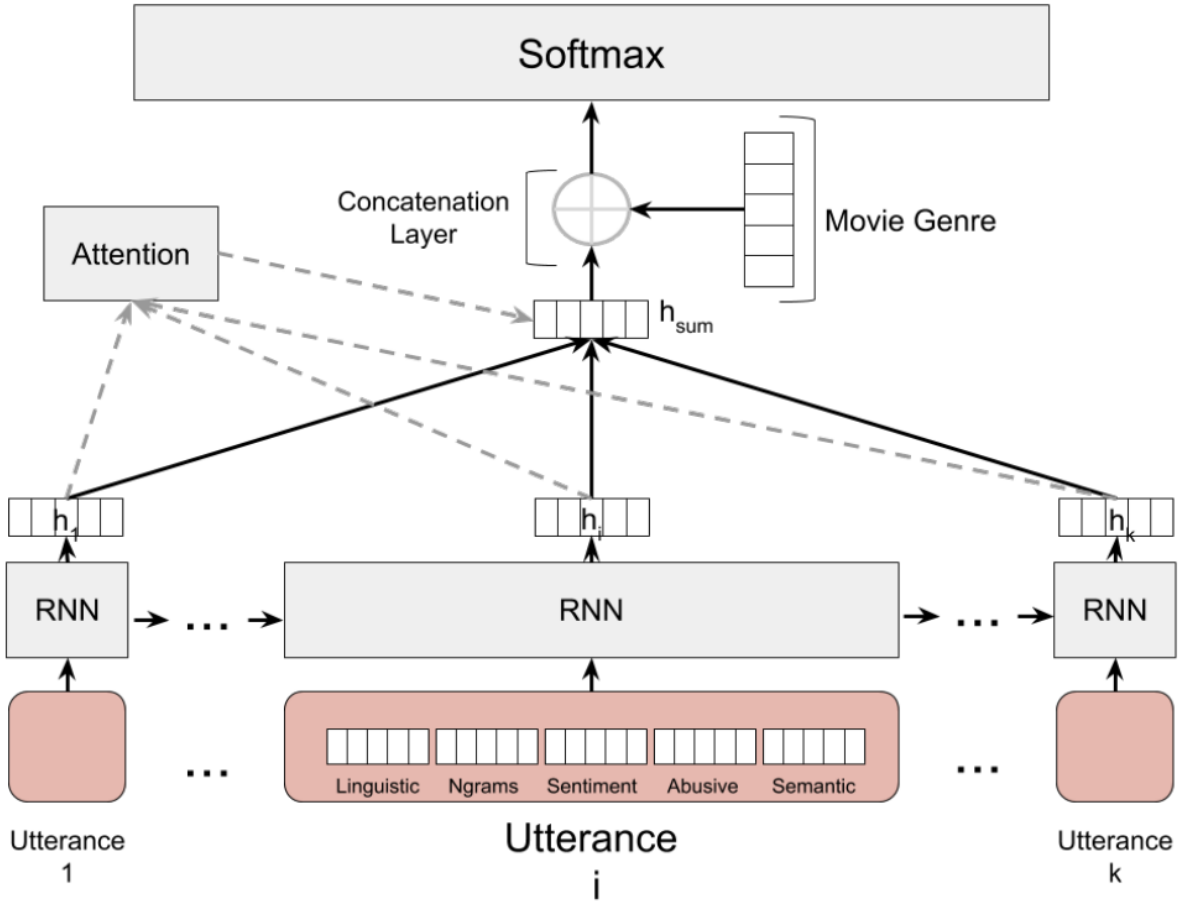
Feature Extraction :

I can extract some features from preprocess text including word frequency, n-grams, and semantic and syntatic analysis. Based on some sensitive and offensive data and words and syntatic label of each sentence, I can classify the text.

Train model :

I can use pre-trained transformer models such as BERT, RoBERTa, or DistilBERT, and fine-tune them on my subtitle dataset. These models can be fine-tuned for classification tasks such as age rating classification.

The picture below from [this article]() might be helpful :



Evaluation :

I will train and evaluate the performance of the model using appropriate metrics, such as accuracy, precision, recall, and F1-score.

Expected Outcomes :

It is expected that ultimately trained model will be able to display an appropriate age range label in the output using an input subtitle.

Challenges :

One important point about this task is that the age restrictions imposed on movies are not only based on their dialogues but also involve visual content, which creates a challenge.

Another challenge that I think may arise is the large amount of input data, as each subtitle contains a lot of content. Classifying such inputs into only five types of labels may be challenging.

Articles :

Two articles that were almost close to this idea, and I found reviewing them useful :

1. https://ojs.aaai.org/index.php/AAAI/article/view/3844/3722
2. https://aclanthology.org/Y18-1007.pdf