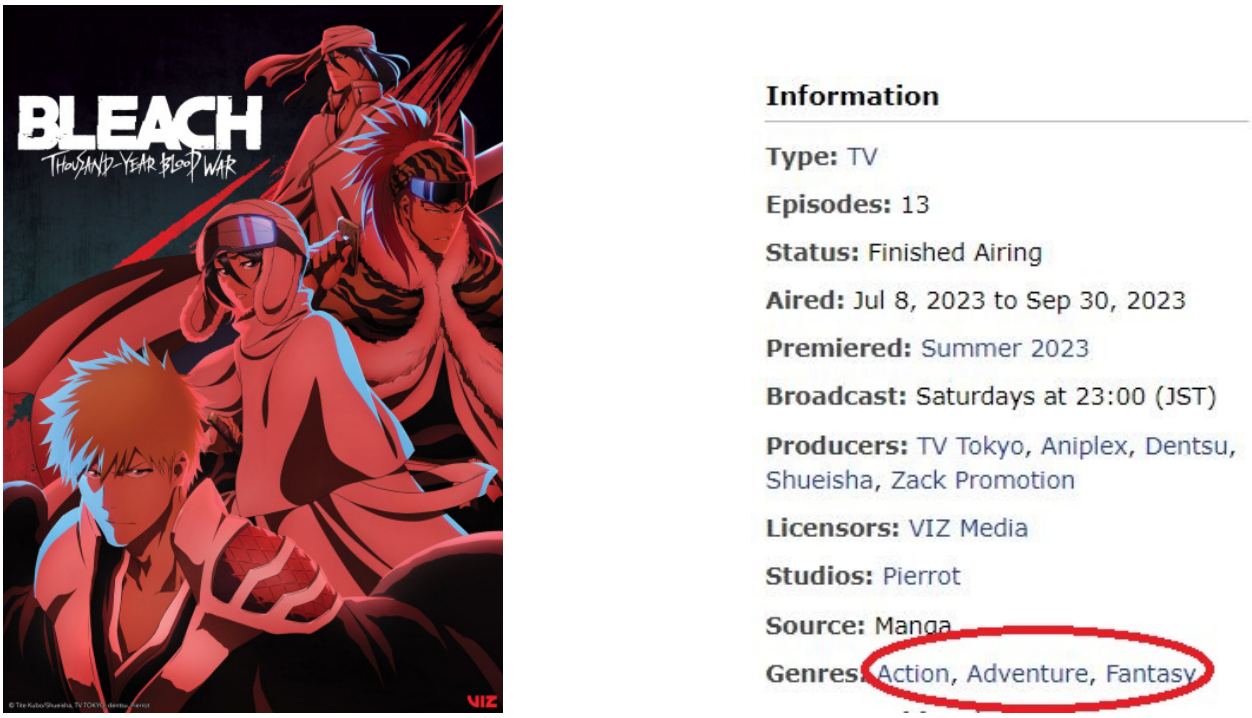
Multilabel Text Classification Challenge

Dr. Seif Eddine Bouziane

Introduction

In machine learning, multi-label classification or multi-output classification is a variant of the classification problem where multiple nonexclusive labels may be assigned to each instance. Multi-label classification is a generalization of multi-class classification, which is the single-label problem of categorizing instances into precisely one of several (greater than or equal to two) classes. In the multi-label problem the labels are nonexclusive and there is no constraint on how many of the classes the instance can be assigned to. Figure 1 below illustrates an example of an element with 3 labels (an Anime in this case) .



(a) Bleach Poster

(b) Bleach Information

Figure 1: A multi-label classification example

In this challenge, you will tackle the problem of multi-label text classification on a dataset of simple chinese tweets. The goal is to develop a machine learning model that can accurately assign one or multiple relevant labels to each tweet.

Dataset

In this challenge, you won't be using the raw dataset, as it is unlabeled and huge. Instead, you'll use a labeled subset. However, you need to perform some processing as the labels are in a separate file. You need

to connect these labels with their equivalent rows based on the tweetId to create a usable dataset consisting of 9950 tweets. The following table illustrates all the lables in the dataset.

Label	Techniques
1	Presenting Irrelevant Data
2	Straw Man
3	Whataboutism
4	Oversimplification
5	Obfuscation
6	Appeal to authority
7	Black-and-white
8	Name Calling
9	Loaded Language
10	Exaggeration or Minimisation
11	Flag-waving
12	Doubt
13	Appeal to fear or prejudice
14	Slogans
15	Thought-terminating cliché
16	Bandwagon
17	Reductio ad Hitlerum
18	Repetition
19	Neutral Political
20	Non-Political
21	Meme humor

Table 1: Propaganda Techniques

Challenge Tasks

Your task is to develop a machine learning model that can accurately classify each tweet into one or more relevant labels. You are encouraged to explore and experiment with different techniques, including but not limited to:

- Preprocessing and tokenization strategies for text
- Feature extraction and representation methods (e.g., bag-of-words, TF-IDF, word embedding, contextual embedding)
- Machine learning algorithms for multi-label classification (e.g., Binary Relevance, Classifier Chains, Adapted Algorithms)
- Deep learning architectures (e.g., CNNs, RNNs, Transformers)
- Transfer learning and fine-tuning pre-trained language models
- Techniques for handling class imbalance and noise

Evaluation

Your submissions will be evaluated using the following metrics for multilabel classification:

- Subset Accuracy

- Hamming Loss
- F1-score, Precision and Recall (micro, macro, and sample-based)
- Any other metric you see fit.

Submission Guidelines

Please submit your solution, including the following components:

- A Jupyter notebook with your code and describing your approach and all the steps from the data prepossessing to the evaluation of your model.

Resources

- A Review On Multi-Label Learning Algorithms.
- Review of Extreme Multilabel Classification.
- Comprehensive comparative study of multi-label classification methods.

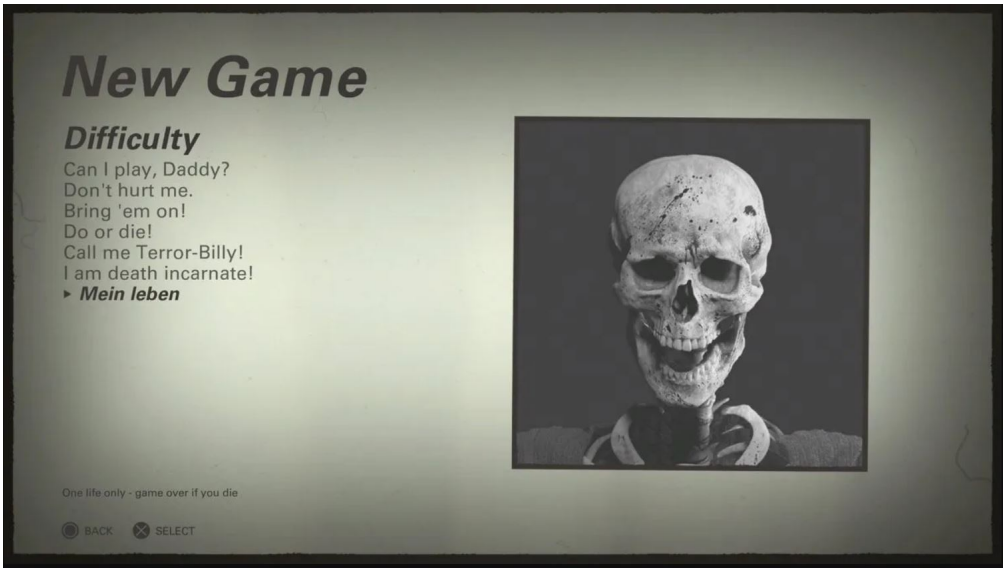


Figure 2: Challenge Difficulty

"I left everything I gathered together in one place. Now, you just have to find it." - **Gol D. Roger.**
Happy Hunting **Mr. Wick.**