VLG OPEN PROJECT 2023-24

Complete Report

We all know in this modern civilization , the LLMs are used extremely to do any NLP domain related task. But the problem arises , when people use the LLM generated text in their daily lives to do their work very fast and also in very illegal activities. So to get a solution of it ,I have made a LLM generated text classification model .

Progress:

I started the project on 10th December, 2023 and successfully completed on 25th December,2023 on Kaggle competition "***LLMs-AI-generated-text-detection***". Within these days I tried lot of models to increase my accuracy on my model and Leaderboard score on Kaggle. Lets See my approach to solve this problem.

Approach:

1. I first see the dataset and plot its target column "generated" ,I saw that the dataset is highly **imbalanced** , so I used a external huge dataset of AI-generated text binary dataset ,available on Kaggle to solve the problem. I selected 45000 and 45000 rows from each of the category (Human generated and AI- generated) to balance the dataset . As a result , the dataset had 46375 rows of Human generated and 45003 AI -generated text.
2. After that I processed the text with **NLTK library** and converts into tokens with proper **padding** by using **Tokenizer**.
3. Then I used **Self-Attention Mechanism based Transformers** followed by **embedding layers** and **dense** layers as my model to feed the data and trained the model for 3 epochs.

   Results :

   - **I get a accuracy of 99% as well as validation accuracy of 99% after training the model.**
   - **I also get 98.42 % accuracy on a external dataset (drcat_v2 dataset)**
   - **I also get a leaderboard score of 0.853 on Kaggle**.

Learnings:

1. I firstly tried basic LSTMs ,GRUs on this dataset , but I got very poor results on training as well as the time taken by this model to train on this huge dataset is also very high . So I switch myself to learn transformers ,due to **its self attention and decoder-encoder attention mechanism and high speed parallelised computation**, its able to get good results on the dataset. I have also written a blog on transformers in medium. https://medium.com/@indubarnwal752/transformers-710fe70b4ed0

2. I also read many discussions on Kaggle in this competition and I come to know that in this competition **simple tf-idf and classification based models overkilled LLM based models**. I researched on it and found some points on that:

   - **I have also seen that there are many typos and grammatical errors in the train dataset. So when you used external dataset which are correct and majority in class (90K vs 1378) then the high and advanced LLMs overfits the training data , you get a high level of accuracy . But when you used this model for the hidden test dataset you get 0.85 around LB score due to the typos available in the hidden test data , in fact due to this typos the LLMs are become unable to that wrong semantics of that data.**

- **Obviously in this case Tf-idf with simple classification model overkills this LLMs because they doesn't focus on the semantics of the text , they just do the count / frequency of the word, so they get good LB as well as accuracy.**
- **But in real life scenario definitely this high LB scored models don't get good results without focusing on the semantics , definitely the LLMs traditionally overkills this simple models**.

I am grateful to VLG that they give me chance to work on this project and for clear my doubts during this project.

Name: Kousik Kumar Barnwal

Enrollment No: 22112056

Branch: Chemical Engineering (2nd Year)

Contact : 8637577602