# Homework – 4

* Supreeth Bannur Sudhakara

I decided to go with Bert models for clickbait detection. I found 3 models that are trained in English language.

* 1. [https://huggingface.co/valurank/distilroberta-clickbait](https://huggingface.co/valurank/distilroberta-clickbait?text=Clicccck)
  2. <https://huggingface.co/Stremie/roberta-base-clickbait>
  3. [https://huggingface.co/elozano/bert-base-cased-clickbait-news](https://huggingface.co/Stremie/roberta-base-clickbait)

I found the dataset from hugging face for clickbait prediction. It has 32000 entries but only 300 of them were used to see the performance of the models.

<https://huggingface.co/datasets/marksverdhei/clickbait_title_classification>

The database has titles along with the label of 1/0 indicating clickbait or not!.

I did a little but of pre-processing of the data, but mostly the data was cleaned and didn’t need a lot of preprocessing.

Update: Preprocessing was dropping my accuracy of the model. My preprocessing method was removing vital information out of the data. Below are the results of the three models!

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | f1-score | Overall-Accuracy |
| 1  (distilroberta) | 0.46 | 0.63 | 0.51 |
| 1.0 | 0.31 |
| 2  (Roberta-base) | 0.62 | 0.76 | 0.75 |
| 0.97 | 0.73 |
| 3  (bert-base-cased) | 0.42 | 0.59 | 0.44 |
| 1.0 | 0.12 |

Text

Description automatically generated with low confidence

I ran the model with minimal text pre-processing, which changed the results of the best performing mode.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | f1-score | Overall-Accuracy |
| 1  (distilroberta) | 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 |
| 2  (Roberta-base) | 0.58 | 0.73 | 0.7 |
| 1.0 | 0.66 |
| 3  (bert-base-cased) | 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 |

I have checked if the predictions were correctly predicted. The results were perfectly predicted as shown in the confusion matrix shown in the image as well.

Calendar

Description automatically generated

Results and Observation:

* After doing the preprocessing of removing numbers, punctuations etc reduced the accuracy as shown in the first table or the image.
* For the preprocessed data, the Roberta-base model was found to be very high compared to the other two.
* For the minimal preprocessed data, the Roberta-base model had the least accuracy. It could be because the way the model was trained could be different. Knowing on what kind of data is used and how the data was preprocessed is necessary for understanding the working of any model and get the proper use of it.
* Distil-roberta and bert-base-cased model performed perfectly in predicting the results for the titles.
* Having different data and parameters for training the model, comparing the model’s performance aren’t exactly the best way to judge their performance.

Limitations:

* Some of the models doesn’t have enough data to understand their details such as how the model was trained, or what dataset was used to generate this model.
* Finding the right model for our problem at hand is quite a task, as some models, are fine tuned in a manner that might not work out best for us, this includes for the language of the dataset used, and the features it was trained on as well!

Future usage:

* This kind of pretrained models help in future research or even to be used in the projects as they save a lot of time that is taken for the training the data.
* As training a big data set of thousands of data takes hours and sometimes require higher compute power which might not be available to us, this helps us use better trained model without having to go through the hard part of training and fine tuning the model.