# University of the Witwatersrand
# School of Computer Science and Applied Mathematics
# COMS4054A & COMS7066A:
# Natural Language Processing/Technology
# Project

## 1 Instructions

For this project you will be aiming to assist in performing systematic reviews. You will be working in groups of one, two or three members and they can be different from any of the groups you formed during the labs. A systematic review is essentially a very structured literature review and our models will have two component: 1) a binary recommendation, 2) the extraction of relevant text. The binary recommendation is a label given to a document indicating whether the model believes it is a relevant piece of literature. A label of 1 indicates that it is relevant and 0 indicates that it is not. Secondly, pieces of relevant documents should then be highlighted if it contains important information. For your project, you must perform the systematic review in an African language other than English. To achieve this you **must** perform two full fine-tuning trainings of your model. The first is a more general fine-tuning of a pretrained model and can be anything you think will be helpful for the final performance of the model. The second fine-tuning is when you then train the model to identify your topic in the chosen language. You can also choose whichever topic you like - systematic reviews are generally done in medicine, however you will already be in a low-resource domain (which medicine is) by using an African language. So being low resource twice over is not a mandatory challenge. You should take the following broad steps:

1. Choose an African language other than English and collect two datasets. One for each fine-tuning stage. Do not under-estimate this step. The decisions you make here will impact every step to come. You will be rewarded for pursuing interesting lanuages, topics and pretraining tasks. You will also have to motivate your choices here well in your write-up.

2. Decide on a pretrained model to get started - you will then fine-tune this model twice. I recommend you look at HuggingFace to help you here. Once again, how you motivate your choices here will matter a lot.

3. Clean your data and split into the usual training/validation/test split. You are expected to use these dataset splits appropriately throughout.

4. Fine-tune your pretrained model on your first, more general dataset. Obtain your results from this task and be able to interpret them in your write-up. If you model succeeds here, what does it mean for your next task?

5. Fine-tune your model on your second task and once again interpret the results.

6. Obtain and interpret the attention maps of your final model to highlight the most important pieces of your text for the topic you are identifying. You only need to show this on a subset of examples.

7. Finally, you must repeat this process for a baseline model which is not fine-tuned with the first dataset. So just jump straight from taking an initial pretrained model and fine-tune on the final topic dataset. Compare the results of training as well as the attention maps to your first model. You must be able to interpret the similarities and differences if there are any.

8. Write a report on your findings (see below).

You may use libraries to assist you in loading data, loading pre-trained models and training the neural networks. Your mark in this project will be based heavily in terms of your research methodology (how you choose your datasets, model and similar design decisions) and less on the overall metrics or the performance of the model. In many cases the African language datasets are small and result in over-fitting. You will be rewarded for noting this and similar limitations when appropriate. The metrics you present will be evaluated on how appropriate they are for the question being answered. The numerical value will merely be used to determine that your model is working as best it can. Note that you must use a train-validation-test data split to train the model, optimise hyper-parameters and evaluate its performance. Data leaks will be penalised. In terms of how large your datasets should be - your models will benefit from getting more data but this also increases computational costs. You can use only as much data as you need to complete the project. Try your best to make life easy for yourselves in how you pick topics and languages. Conversely, if you find clever ways to achieve something more difficult you will be rewarded. These factors are left deliberately vague because a large part of the assessment is seeing how you fill in these blanks yourself. Finally, ensure you show your working in the write-up. If you run an experiment which doesn't work but lends some insight, present it briefly and discuss it. We will only mark you on content which you tell us in the report and so don't waste results.

You will also be expected to write a six-page (double column) report on the topic you covered, this includes figures, tables and equations. You may use unlimited space for references and can include any other information you think is relevant in supplementary material. You can direct readers to the supplement in your 6-page report, however your markers are not required to read this if they do not want to. So make sure what you want to say is in the 6-pages. Please include your contribution statement in the appendix. The page limit is firm. You must use the IEEE format. Please see the rubric below for more details on what should be included in the write-up. This write-up should focus on your methodology, results and insight for your particular topic and should follow the format of an academic paper (provide introductions, background and so on). Creativity and insight will once again be rewarded throughout the project. Similar to your labs, you may use ChatGPT to help you write the report but as a result negative marking will be used for poor writing and formatting. Hallucinations or factually incorrect information will be heavily penalised. Your code will be checked for plagiarism and you must code the project yourselves. Finally, I will include the NeurIPS 2024 ethics statement/questionnaire as a .tex file on moodle. You must append this to your report after the references but before the supplementary material. Absence of the statement will result in the project not being marked (it takes roughly 15 minutes to fill in at the end).

## 2 Submission

**Due Date**: 22 October 2023 at 14:00.

For the submission please:

1. Submit your full code implementation.
2. Submit the report.

On the final page is a rubric which will be used to assess your write-up. I emphasise that your model accuracy or performance will only be used to determine the correctness of your approach. In other words, I care more about your thought-process, reasoning and investigation and less on your ability to obtain the best possible performance from the model. Let this guide your priorities.

# 3 Using HuggingFace Models

This section provides a brief introduction on how to fine-tune models from HuggingFace.

1. Start by installing the Transformers library from Hugging Face, which provides a user-friendly interface for working with pre-trained models. You can find the library and installation instructions on the Hugging Face Transformers GitHub page.

2. Hugging Face offers a wide variety of pre-trained models for various NLP tasks. To find the right one for your project, visit the Model Hub, where you can browse and search for models by task, architecture, or language.

3. You can easily load pre-trained models using the Transformers library. Here's a basic code snippet for loading a BERT model:

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer

model_name = "bert-base-uncased"
model = AutoModelForSequenceClassification.from_pretrained(model_name)
tokenizer = AutoTokenizer.from_pretrained(model_name)
```

4. If you need to fine-tune a pre-trained model on a specific task or dataset, HuggingFace provides resources and examples on how to do this effectively at this link.

# 4 Finding African Datasets

You are free to choose any datasets which you like as long as they are in a non-English African language and the final dataset performs the text recommendation and highlighting. Here are some suggestions on where to find African language datasets but you can also scrape the internet for other sources:

- Hugging Face has a collection of African NLP datasets available on their platform. You can search for them using keywords like "African" or "AfroNLP" on the Datasets Hub.
- This GitHub Repository contains links to various African NLP datasets for different NLP tasks.
- You can also check Masakhane's curated list of African NLP datasets at this link. Note: If you want to fine-tune a HuggingFace model with a custom dataset (a dataset outside HuggingFace), please refer to this tutorial on how to do so.

You do **not** need to use any of these datasets. I just put this here because it is good to know and is a nice place to find African language data. You may use whatever data you like as long as you can motivate it.

# 5 On the Use of ChatGPT

The use of generative models to help with writing is permitted for this submission. Using generative models for coding is not permitted. Violations of this, if detected, will be dealt with as plagiarism. All other plagiarism rules of the university stand and will be implemented as usual.

| Mode | 0% to 20% | 20% to 40% | 40% to 60% | 60% to 80% | 80% to 100% |
|---|---|---|---|---|---|
| Write-up Structure | Adequate use of language and structure. | Fair use of language and structure. | Well written and clear | Good use of language and structure. | Excellent use of language, very well written and structured |
| Method | No clear description of the architecture, data or approach to training | Some description of high-level details | Fair description on the details of the base model but little elaboration of the approach to answering the sub-topic | Base approach is described in detail and steps are well justified. Adequate discussion on the approach to answering the sub-topic | Excellent description and motivation for all parts of the method including on how the sub-topic was answered |
| Results | Results are not given or irrelevant | Some results given with inappropriate metrics | General results presented with appropriate metrics | General and sub-topic results presented with appropriate metrics | Thorough results which are appropriate for answering the sub-topic and display the general correctness of the model |
| Discussion Section | No interpretation of results or incorrect interpretation. Little knowledge of the topic displayed | Some general interpretation of results broadly | Results are interpreted which display some understanding of the mechanics of the model. Displays a basic understanding of the topic | Results are interpreted and contextualized to begin to answer the sub-topic. Clear demonstration of knowledge of the general topic | Results are interpreted and contextualized to answer the sub-topic and display insight into the working of the model or original thought on the concepts |
| Verbal Assessment (Class Discussion) | No display of knowledge on the topic. | Adequate knowledge of the topic but unable to coherently answer the question (gives random facts or irrelevant information) | Provides a concise and correct but shallow answer | The answer is clear, concise and correct. Demonstrates some deeper understanding of the topic | Demonstrates insight into the topic. Answer is clear, concise and correct. Relates the topic to other material or contextualizes the topic more broadly. |