

Addressing Fairness in Large Language Models

Students: : Erick Pelaez Puig, Elnaz Toreihi, Valeria Giraldo, Jeslyn Chacko

Mentor: Wenbin Zhang

Instructor/Faculty: Masoud Sadjadi, Florida International University

PROBLEM

The research we have conducted this year was to address the problem of bias in large language models. We had researched many LLM metrics that measure how well the models perform certain tasks and if it has patterns of bias. These languages are usually pre-trained on the hugging face platform and tested using different data tests to test the entire scope of the metrics and compare to one another.

The metrics that I was responsible for researching, implementing, and testing were the Sentence Embedding Association Test (SEAT), DisCo, Social Group Substitution, and the Co-Occurrence Bias Score metric.

CURRENT SYSTEM

In the current system of technology, you can find many areas of bias. For example, when training with datasets, it may not be representative of the entire population and those groups would, therefore, be underrepresented in Language Models.

There are a lot of stereotypical associations that may come with gender bias. For example, some models may default to male pronouns or associate gender with a specific profession.

REQUIREMENTS

The requirements for the project was to be able to implement and understand how bias works in the different metrics that we worked with. The metrics that we learned about include:

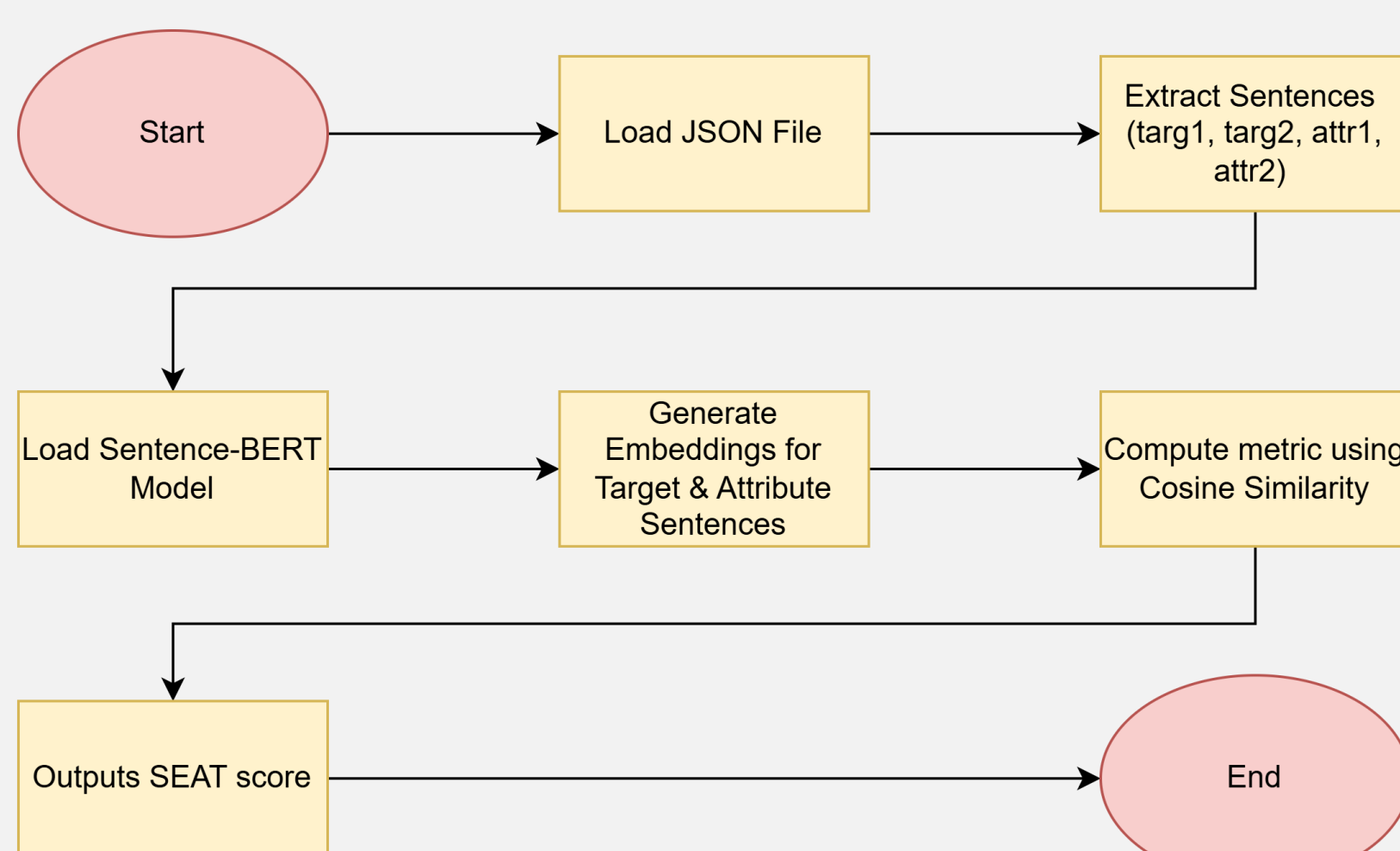
- Embedding-Based
- Probability-Based
- Generated text-based

We were to test the metrics within these categories and see how effective they are with different types of bias such as gender stereotypes.

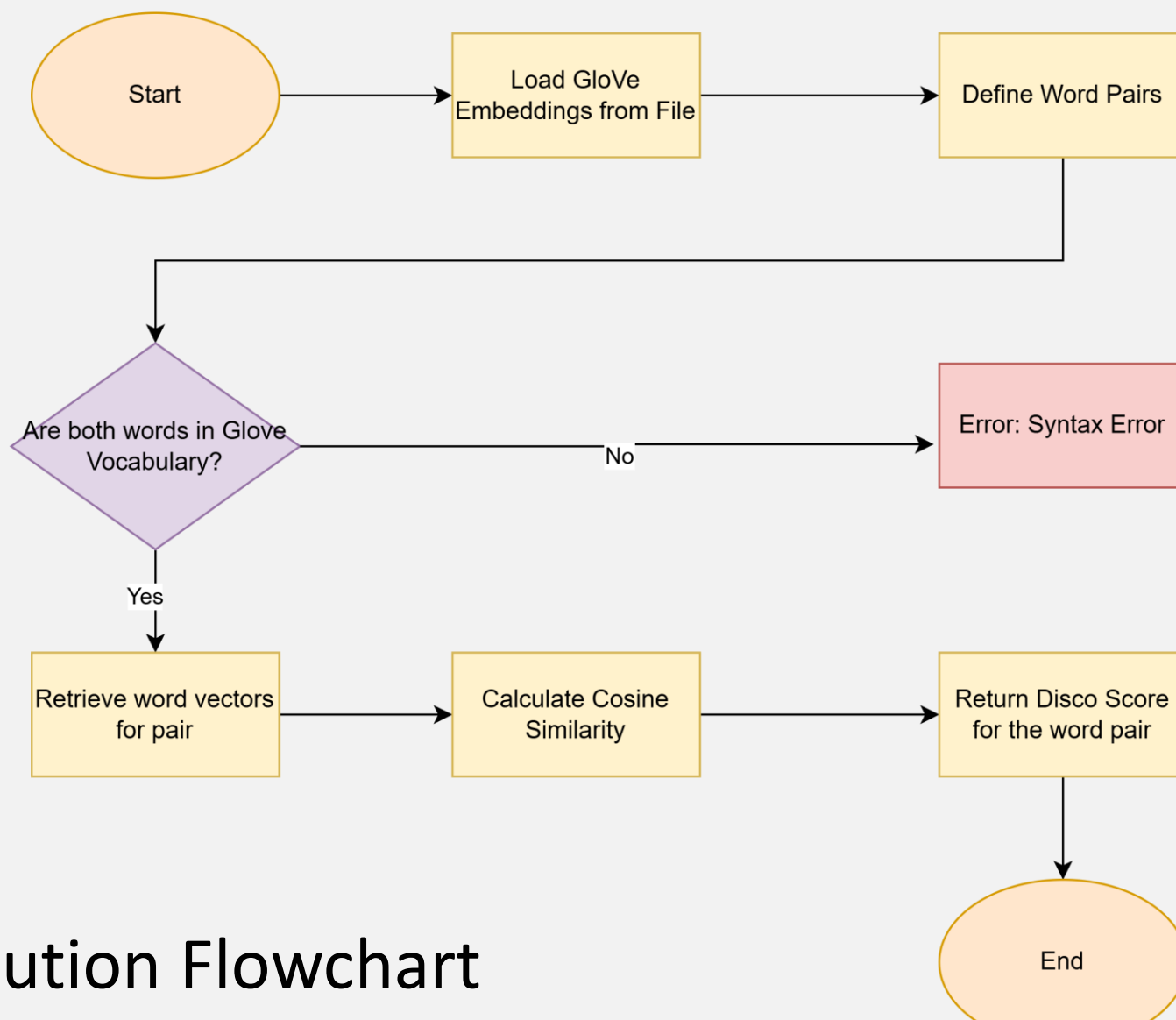
We would then compare our results to the other metrics that we are testing.

SYSTEM DESIGN

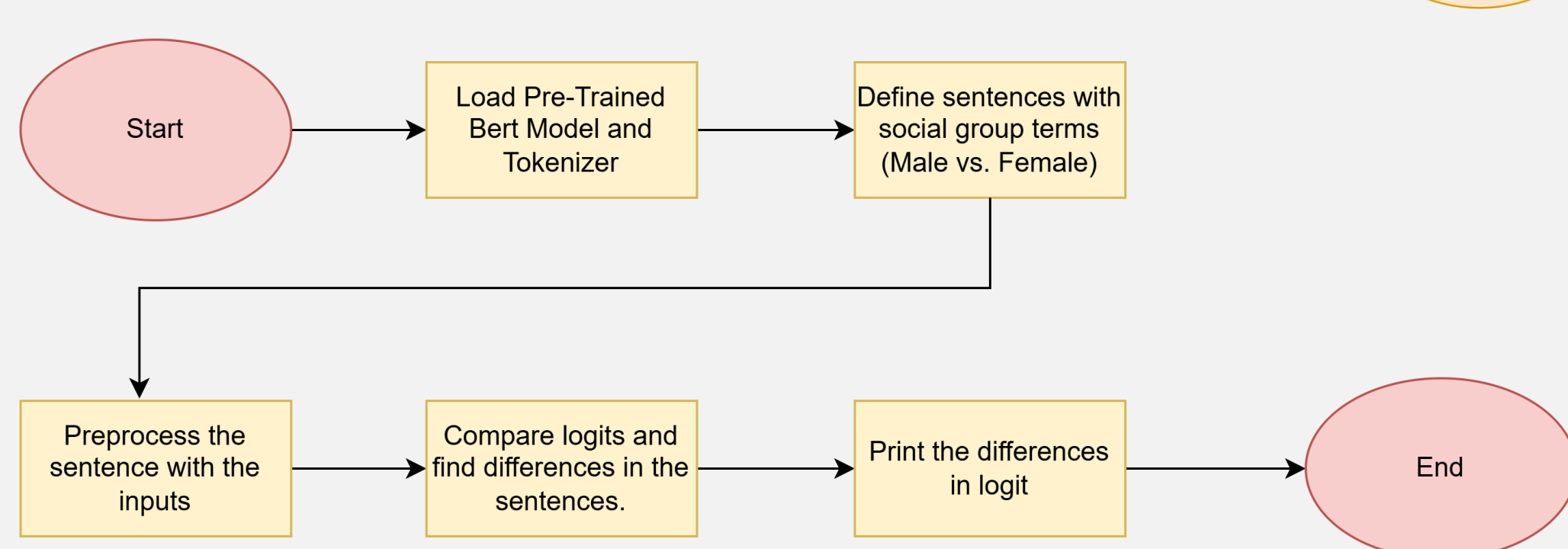
SEAT Flowchart



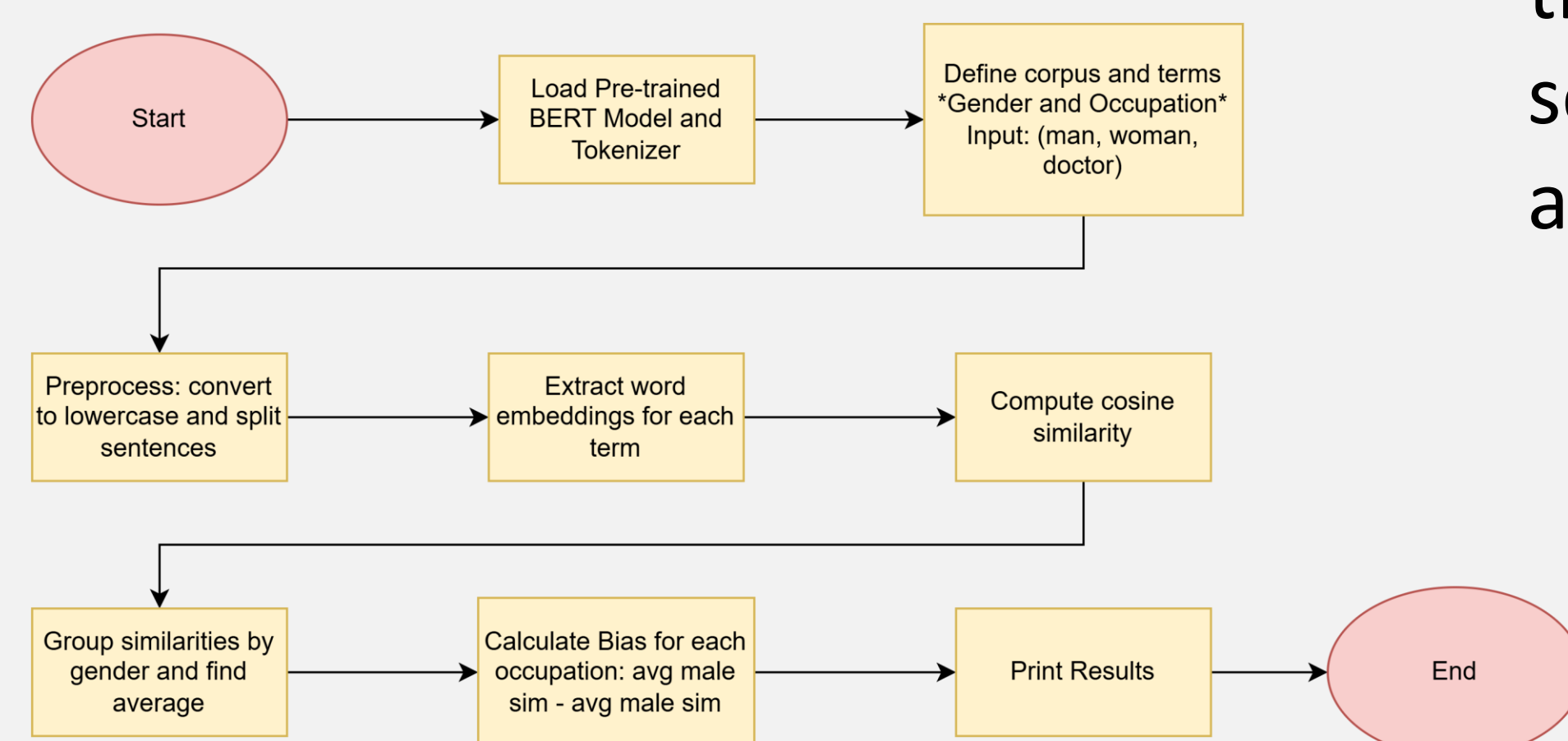
DisCo Flowchart



Social Group Substitution Flowchart



Co-Occurrence Bias Score Flowchart



OBJECT DESIGN

```
C:\Users\jesly\PycharmProjects\capstone2-project\.venv\Scripts\python.exe C:\Users\jesly\PycharmProjects\capstone2-project\.venv\src\disco.py
Loaded 400000 word vectors.
Disco score for 'man' and 'doctor': 0.8564
Disco score for 'woman' and 'doctor': 0.8570
Disco score for 'man' and 'nurse': 0.7804
Disco score for 'woman' and 'nurse': 0.8513
Disco score for 'man' and 'scientist': 0.8811
Disco score for 'woman' and 'scientist': 0.7985
Process finished with exit code 0
```

DisCO Metric showing ‘woman’ having a larger score associated with ‘nurse’ than ‘man’.

```
Logits for male version (John): tensor([[ 0.3467, -0.0741]])
Logits for female version (Jane): tensor([[ 0.3858, -0.0552]])
Difference in logits: tensor([[ -0.0391, -0.0188]])

Process finished with exit code 0
```

Social Group Substitution showing the negative differences indicating slight bias towards ‘John’ than ‘Jane’

SUMMARY

In the current system of technology, you can find many areas of bias. For example, when training with datasets, it may not be representative of the entire population and those groups would, therefore, be underrepresented in Language Models.

There are a lot of stereotypical associations that may come with gender bias. For example, some models may default to male pronouns or associate gender with a specific profession.

IMPLEMENTATION



Hugging Face



GitHub



REFERENCES

Paeaz, E., Toreihi, E., Giraldo, V., & Chacko, J. (n.d.). ERICK0726/metrictesting. GitHub. <https://github.com/Erick0726/MetricTesting>

Wang, A. (n.d.). Sent-bias/tests at master · W4NGATANG/sent-bias. GitHub. <https://github.com/W4ngatang/sent-bias/tree/master/tests>

Zhang, W., Wang, Z., & Chu, Z. (n.d.). Fairness in large language models: A taxonomic survey. <https://arxiv.org/pdf/2404.01349>

ACKNOWLEDGEMENT

The material presented in this poster is based upon the work supported by Masoud Sadjaji. We/I thank Masoud Sadjaji for his assistance and mentorship that I/we received throughout the senior design project.