ASSIGNMENT 2

PREDICTING FREEZING OF GAIT FROM FALL REPORTS

Version 1.0

1 PROBLEM DESCRIPTION

Participants
Participants in this study were adults previously diagnosed with Parkinson's disease by a movement disorders specialist. All participants provided written informed consent prior to participation according to procedures approved by the Institutional Review Board of Emory University.

Study design
One-year prospective observational study. All participants were assessed with a detailed cognitive and motor battery of indicators of potential fall risk during a single study visit. After enrollment, study participants were prospectively tracked for incident falls for a nominal one-year period. During the observation period, they were queried about the presence and circumstances of any falls at monthly intervals using mail, email, phone, or text, at the discretion of the participant. Details of subsequent falls were recorded by participants and verified by study staff. Approximately 1/3 or participants enrolled went on to fall during the observation period.

Data set and study question
The dataset includes a variety of variables, ranging from demographic information like gender, race, and age, to clinical scales such as the *MDS-UPDRS Part III* and *Hoehn Yahr Stage*. One of the key variables is **fog_q_class**, which indicates whether a patient experiences **freezing of gait (FOG)** at baseline. Another crucial variable is **fall_description**, which provides a third-person narrative of the circumstances surrounding each fall event experienced by the participant.

Your task is to build a Natural Language Processing (NLP) model that uses the text data from **fall_description** to predict the **fog_q_class** variable. In essence, you will be investigating whether the descriptive accounts of patients' falls can serve as a predictive marker for experiencing freezing of gait.

2 TASKS

Overview
This is a binary classification problem, and your model should aim to classify each patient as either experiencing **FOG (fog_q_class = 1) or not (fog_q_class = 0)** based on the textual descriptions of their

falls. Feel free to include additional factors, such as gender, in your model. Then you will have to design and execute a thorough NLP-driven study to:
(i) implement an automatic classifier
(ii) identify text features that are indicative of the classes

Minimum set of specific tasks

- Implement an automatic classifier
- Cross-validate on the training set
- Tune hyperparameters
- Apply some sort of ensemble classification
- Compare at least 5 classifiers + a Naive Bayes baseline
- Engineer at least 4 features + n-grams
- Identify the best classifier & feature set combination
  - Evaluate the performance of classifiers based on overall micro-averaged F1 score. However, report all of the following in all evaluations.
    - Accuracy, micro-averaged F1 score, and macro-averaged F1 score.
- For the best classifier:
  - Perform training set size *vs.* performance graph. Is the learning improving with data? Can we estimate how much annotated data we need?
  - Perform an *ablation study* (re-run experiments with one feature set removed at a time)

3 DATA

[SEE CANVAS PAGE]

4 SUBMISSION

You will submit a 4-page report (excluding references). The report should be representative of a short paper submission to a journal. The report should contain the following sections: **Abstract**, **Introduction**, **Methods**, **Results**, **Discussion**, and **Conclusion**. Figures and tables will count towards this 4-page limit.

You can read NLP papers from the following journals for reference: **Journal of the American Medical Informatics Association** (https://academic.oup.com/jamia), **Bioinformatics (**https://academic.oup.com/bioinformatics**)**, **JAMA Network Open** (https://jamanetwork.com/journals/jamanetworkopen), **Journal of Biomedical Informatics** (https://www.journals.elsevier.com/journal-of-biomedical-informatics).

You will be following the ACL template.
Overleaf: https://www.overleaf.com/latex/templates/acl-2023-proceedings-template/qjdgcrdwcnwp
LaTeX: https://github.com/acl-org/acl-style-files

Word: https://github.com/acl-org/acl-style-files/tree/master/word

In your report, provide a link to your GitHub repository.

Due date: October 24.