Mulualem Asmare: MSDS 692 Project proposal 2

1. **Name, Contact info (e.g. email/phone):** Name: Mulualem Asmare Email: masmare@regis.edu Phone:3039016770
2. **Title of the project:** Developing model for prediction of text generated by AI vs human.
3. **High level description of the project: what question or problem are you addressing?** Currently, we live in an era dominated by generative artificial intelligence (AI). Despite its useful application in various sectors, it poses a potential threat to our survival due to its potential for generating and proliferating misinformation. In recent years, AI-generated misinformation has been on the rise. For instance, according to The Washington Post, published on December 17, 2023, by Pranshu Berma, AI-generated fake news has increased by a staggering 1000 percent since May of 2023. The increase in AI-generated fake news poses a threat to our security and stability. Recognizing the potential threat of AI-generated fake news, this project aims to develop a classification model capable of detecting whether an AI or a human generates a given text. Additionally, this project seeks to identify common patterns inherent in AI-generated and human-generated texts.
   References:
   Berma, P. (2023, December 17). *How ai fake news is creating a 'misinformation superspreader' - the* ... The rise of AI fake news is creating a 'misinformation superspreader.' https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/
4. What types of data science task is it?
   - EDA
   - Data visualization
   - Data transformation
   - Classification Model development
5. **Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?** I plan to use a combined dataset from three sources. One dataset is obtained from the Hugging Face Transformer ecosystem using their API in JSON format. This dataset is raw data and requires feature engineering and cleaning. Another dataset is obtained from Kaggle in text format, titled 'Human or Machine Generated Text?'. This dataset needs cleaning and feature engineering to combine it with other datasets. The third dataset is obtained from Kaggle and is named 'AI-and-Human Text,' containing 300,000 rows and 3 columns. This dataset is fairly clean; however, I will perform feature engineering to combine it with the other dataset. Due to the expected volume of data, I plan to use as much data as possible for my machine learning, considering the limitations of my computer.
6. planned to use the data from Kaggle. The dataset I intend to use has about 500K unique AI and human generated essays, totaling 1.11 GB. The data contains two columns. The first column, named 'text' contains the essays generated by AI and humans. The second column named 'generated' contains labels: 1 for AI- generated and 0 for human-generated texts.
7. **How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?** To analyze

the data, I will use Jupiter Notebook Python 3 for interacting and analyzing data. I will employ data exploration techniques to observe word count and character count for AI- and human-generated text. This will provide insights into the number of words and characters they use. Additionally, I will identify the distribution of classes in the target variable to observe whether there is a class imbalance. I will clean the data by removing special characters, numbers, and stop words. I will also transform the data into lowercase. Subsequently, I will identify the most and least frequent words and perform n-gram analysis on text generated by AI and humans to observe patterns in word usage. For visual representation, I will use word clouds to compare the words used by AI and humans and identify patterns. Following that, I will tokenize the text to break it into tokens (individual words) and perform lemmatization to change the words into their basic root form, reducing dimensionality for better machine learning model learning. Next, I will use TF-IDF vectorization to convert the text into numerical vectors for machine-learning tasks. I am planning to use logistic regression, random forest, and naive Bayes models. I will utilize accuracy metrics to test the accuracy of each model and choose the best prediction model.

8. **Describe any anticipated difficulties and problems. Discuss how you may overcome the problems**. I may encounter several anticipated difficulties and problems during the project, such as system and library compatibility issues, error codes, code issues, and various unexpected challenges. I plan to resolve these issues by utilizing resources such as Assignment I did in my past class, Google, reading articles, watching related videos, and visiting GitHub repositories, Stack Overflow, Kaggle, and other relevant platforms.
9. **Suggest a timeline for the project. This should be a weekly breakdown of what you plan on doing each week.**
   Week1: Project proposal
   Week2: Project proposal and Data collection
   Week3: Data Cleaning and Exploration
   Week4: Data preparation
   Week5: Data preparation
   Week6: Model development
   Week7: Model evaluation and Testing
   Week8: Finish up project, Conclude the observation and finding.
10. Create GitHub repository for your Practicum project. Add this proposal, begin a ReadMe document, and begin adding your data to your repository. Add a link to your GitHub repository to this document.