AML 2103 2, PROJECT PROPOSAL

PROJECT TITLE	Understanding and Predicting the Success of Artificial Intelligence and Machine Learning Posts

SUBMITTED BY	Mary Gomez	DATE OF SUBMISSION	11/22/2023

First Name	Last Name	Student Number
Mary	Gomez	C0891136
Rehan	Khokhar	C0896278
Andres	Correal	C0872634
Viki	Patel	C0906295
Bhavya	Vadher	C0894977
Prashant	Bhattarai	C0898956

| MOTIVATION | Artificial Intelligence (AI) and Machine Learning (ML) have become trending topics for discussion in social media, representing a significant opportunity for students in AI and ML, who could create content and capitalize on fantastic job opportunities. However, most of the time, students don't know precisely how to make sticky publications on social media.

This project aims to build a tool that can be used by anyone (especially AI and ML students) to understand and predict the success of AI and ML posts. With this, the tool's users can refine their content, increasing the chances of success. |
|---|---|
| METHOD | We will get the data from Reddit API (https://www.reddit.com/dev/api/).

To gather insights on what drives successful posts on AI and ML, we plan to create different visualizations.

- The visualizations will be built using Matplotlib and Seaborn.

Additionally, we plan to train a classification model able to predict the success of an AI and ML post.

- We will start using tree-based classification models, such as GradientBoostingClassifier or RandomForestClassifier.
- We will analyze the feature's importance. |
| INTENDED EXPERIMENTS | Determine the Most frequent words, top bigrams, and top trigrams. Those are to be grouped by ranges of score (discretizing Score, which is the measure of success)

- This will help us to understand which are the most common words, bigrams, and trigrams in successful posts.

Inspect if there is an impact on the score by the time of publishing.

Establish if there is a relation between the length of the text or any other feature with the target (Score)

Compare at least two classification models in terms of accuracy, to do this, the target "Score" will be discretized. |
| PLANNING AND MILESTONES | - Scrap the dataset from Reddit social media using Reddit API - Viki
- EDA for the dataset – Prashant
- Data Cleaning and preprocessing - Bhavya
- Visualizations - Andres
- Modeling - Rehan |

	- Reporting - Mary