WEEK 7

Project Proposal

Group Members:

Member no.	Member name	Member age	Member country	Member speciality	Chosen topic	
1	Zyad Hussein	23	Egyptian	Robotics & Machine learning engineer	Natural Language Processing (NLP)	

Group Name:

Team Alpha

Chosen NLP Project:

Hate Speech detection using Transformers (Deep Learning).

Business understanding:

Hate speech can target a person based on his ethnicity, religion, race, colour, nationality, sex or ancestry. A person who has been targeted or faced a hate speech incidence usually face the incident in a form of verbal, written, or behavioural action. Hate speech has been lately increasing across social media platforms as well. therefore, it was needed to identify a method to tackle such issue and identify it to proceed with any for of action against the responsible party. A prominent and new method that can be implemented to tackle such issue, is the utilization of deep learning models and neural networks. A neural network model can be trained to identify a hate speech and detect it from within a complete paragraph and report it. To train such models, it is needed to acquire a large dataset with great variety of hate words, sentences, or paragraphs to ensure quality of trained model is maintained when detecting and decrease the possibility of false positives. Such model can be implemented across social media platforms to contribute in the detection of hate speech among the platform and limit the incident. It is believed that social media companies can benefit from such trained model to limit hate speech and ensure a safe experience for its users across its platforms which in return will increase the interaction of users across the platform. It will also benefit the platform in increasing the number of users as soon as the consumers recognise the efforts done by the platform to eliminate the hate speech incidents which can persuade consumers into promoting the usage of the platforms to others . Hence, increasing platforms revenues.

Business/model Requirements:

To implement the proposed method, it is required to obtain a well varied dataset to yield promising results and ensure the avoidance of overfitting. Dataset was provided to train the network. The dataset provided consists of twitter tweets. This dataset can be helpful for training a model that will work on identifying hate speech in twitter tweets as the tweet mainly consists of a word, words mixed with specials characters, a sentence or a small paragraph which is the type of data that is expected to train the model on it. The provided dataset might require data cleansing to reach training standards and it will then be divided into 70% training, 20% validation and 10% testing. The model can then be trained using the dataset and performance will be evaluated through mean average precision plots and from testing phases. The model can then be deployed on platforms or as a web application that detects hate speech within a given paragraph. EDA analysis can be performed on the given data to identify the type of data given and provide a report as well as speculations on the quality and variety of data given.