Name: Faris Chaudhry
Batch: LISUM25

Hate Speech Detection

Week 7

Team Member Details

Name: Faris Chaudhry

Email: faris.chaudhry@outlook.com

Country: United Kingdom

University: Imperial College London

Specialization: NLP

Problem Description

“The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor.

Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. We will use the Twitter tweets to identify tweets containing Hate speech.”

Name: Faris Chaudhry
Batch: LISUM25

Business Understanding

Companies generally wish to censor posts that aren't in line with their policy. In this case, the company's policy is that there is to be no hate speech in user's messages (defined as discriminatory language based on factors of identity alone). It is beneficial for the company to enforce this policy to ensure the platform is accessible and advertising friendly.

YouTube faced an analogous situation. Around 2017, they pushed family-friendly content to the forefront of their platform while mitigating anything they deemed offensive or controversial. This allowed greater confidence from advertisers to market their products since their products might otherwise be shown next to unsavoury content. Therefore, the businesswise significance of doing this is: firstly, to build and maintain their user's confidence in the platform; secondly to increase revenue by increasing advertising opportunities.

Project Lifecycle

The lifecycle is in the following stages:

- Identifying and understanding the problem from a business perspective.
- Gathering data (this has been done in the data set).
- Outlining assumptions made about the data.
- Data preparation (validation and cleaning).
- Feature engineering and extraction to create parameters for the model.
- Building the model from the labelled training data.
- Deploying the model on unlabelled training data to test its efficacy.
- Repeating the last two steps to get optimal results without overfitting to the data. Trying out different model types and features to get highest precision, recall, F1-Score, et cetera.

The official deadline for this project is 30th November.