Week 7 Report

Name: Amr Elbana
Email: amr32009363@gmail.com
Country: Egypt/ living in Germany
University: Siegen.
Specialization: NLP

Problem Description:

As mentioned, on the website of the internship, document / Text classification is one of the important applications in supervised machine learning (ML). Many of the news websites try to recommend similar news to the reader. The process of recommendation depends on the category of the news. News should be classified to groups and recommended to the users based on that. The challenge is to build a good ML system to predict the category of the online news in a high accuracy.

Business understanding

Many websites recommend news feed to the users based to the type they are always looking for. The point is that they using ads according to the number of users go to each category. Most of the profits for news websites is from the Ads so they try to catch the attention of the users and make them to stay a longer time on the websites by recommending similar news to what they are interested in.

Project lifecycle

This Schedule showing the tasks and deadline for each task to be handed.

Week	Tasks	Deadlines
Week7	- Data understanding - Business understanding	19 July 2022

Week 8 & Week 9	1. Problem description 2. Data understanding 3. Type of data used and the problem in it 4. Approaches to clean and transform your data 5. Vectorization Techniques	26 July 2022 2 AUG 2022
Week 10	1. Problem Description 2. EDA performed on the data 3. Final Recommendation	9 August 2022
Week 11	1. Problem description 2. EDA presentation for business users	`16 August 2022
Week 12	1. Select your base model 2. explore 1 model of each family. i.e. 1 model for Linear models, 1- Model for Ensemble, 1-Model for boosting. Etc. 3. Select model fits in your business requirement.	23 August 2022
Week13	1. Provide a Power point presentation. 2. Communicating Findings 3. Share the findings with stakeholders	30 August 2022

Name: NLP: Document Classification
Report date: 0/08/2022
Internship Batch: LISUM10: 30
Version: 1.0
Data intake by: Amr Elbana
Data intake reviewer: All Members
Data storage location: UCI

Tabular data details:

Total number of observations	19997
Total number of files	20
Total number of features	2
Base format of the file	.txt
Size of the data	45 Mb