School of Electronic Engineering and Computer Science

Queen Mary University of London

MSc PROJECT DEFINITION

2019-20

Project Title: An Online Multilingual Hate speech recognition system

Supervisor: Dr. Arkaitz Zubiaga

Student name: Neeraj Vashistha

Student email: ec19471@qmul.ac.uk

Student phone number: 07721176324

PROJECT AIMS:

The project aims to target hate speech in online web pages by identifying chunks of sentences/phrases and text in real-time which are related to hate speeches. Discredit, Derailing, Dominance, Sexual Harassment & Threats of Violence, Stereotype & Objectification are some of the types of Hate speeches other than cyberbullying and abuse which can be seen easily on Internet. By using existing research and combining the same with added support to other languages, and thus, creating a tool which identifies and scores a page with effective metric in real time and using the same as feedback to re-train our model.

A lot of research has been carried out in recent past in order to classify hate speeches but there is still a lot of potential for research as even today scores of people are victims of Hate speech. The existing system is designed on a set of datasets but as the language is evolving every day, the system to identify hateful content should evolve. The system should be robust, adaptive and multilingual so that it can truly improve lives.

First and a major challenge is to define Hate speech, in the recent past there have been many instances where the freedom of expression has been mixed with hate speech, and thus there is no technical definition regarding the same. Second challenge is the result of the first issue, since there are different definitions, the datasets and source capture different information thus there is no standardization for data. Third and the most important challenge is automatic detection is difficult and is not interpretable. Using machine learning, few state-of-the-art methodology have been built but these techniques are created on a certain annotated datasets which are not evolving like usage of language in current time. Further the results of these automated machine learning tools are not interpretable. One more issue with automatic hate speech detection system is that they fail to capture the societal differences, meaning and context of the text phrase, thus the current system fails to generalize well. The current system aims to solve the above issues as well as develop an evaluation strategy to detect hate speech on unlabeled data ans streaming data.

The aim of the project is to explore the current techniques and state-of-art methods and work towards solving them in an organised way.

PROJECT OBJECTIVES:

The main objective of this project is categorised into two parts, the long term objectives and short term objectives. In short term objectives, a short literature survey on the topic of Hate speech will be performed. Analysis of different methodologies, along with identification of different data set would be carried out. And quick implementation of few algorithms with error analysis shall be performed. This is done in order to gain insight and understanding of the domain knowledge, dataset and existing solutions.

The long term objective would involve in-depth analysis of different machine learning models in field NLP and other topics and to find the scope if those can be incorporated to solve the current problem. Another long term objective would be to handle streaming data and device the model into a web application to be used in real time.

Both the sub-objectives would work hand-in-hand in-order to achieve the aim, and overcome the challenges as foreseen currently and new challenges which could arise in future.

Further, these objectives are listed into the Methodologies section of how these would be intended to follow in order to achieve the project aims.

METHODOLOGY:

The project is designed in two phases. First, developing the core machine learning models and second, deploying it in an online environment where it can be evaluated and improvement/suggestion can be incorporated for the given result/ performance metric. In both the phases, below listed research/development would be carried out.

1. Identification of existing multilingual dataset containing hate speeches with proper annotation for different categories.

2. Existing research in the field of NLP and Deep Learning from text classification to sentiment analysis and to more specific state-of-the-art Hate speech recognition systems. This would be done to bring new methods to current state-of-the-art methods and understand the challenges associated with them.

3. Propose a new methodology incorporating the findings and learnings from existing.

4. Evaluation, Interpretation and error analysis of existing state-of-the-art methods with the proposed system.

5. Design of a new web application which would be used for the purpose of deployment of the above system.

6. Web application development and deployment.

7. System integration/configuration with online and real-time streaming data in web application.

PROJECT MILESTONES

The project milestones are described as follows.

1. Project Definition Document - Aims, objectives, scope, challenges and task assessment.

2. Project Requirement Specification Document - broader coverage of scope, description, product features, end users/ stakeholders, dependencies, system features, functional and non-functional requirement, internal and external requirements, product security and quality attributes.

3. Literature Review Document

4. Dataset collection, exploration and assessment

5. Data preprocessing, Feature extraction

6. Evaluation of existing algorithms

7. Model Design

8. Model Training

9. Model Validation & Testing

10. Evaluation and Error analysis of Model

11. Web application Design

12. Web application Development

13. Web application Model deployment

14. Web application Testing on streaming and non streaming data

15. Web application Model evaluation on unlabeled data

16. Alpha and Beta Bug testing

17. Final Report

REQUIRED KNOWLEDGE/ SKILLS/TOOLS/RESOURCES:

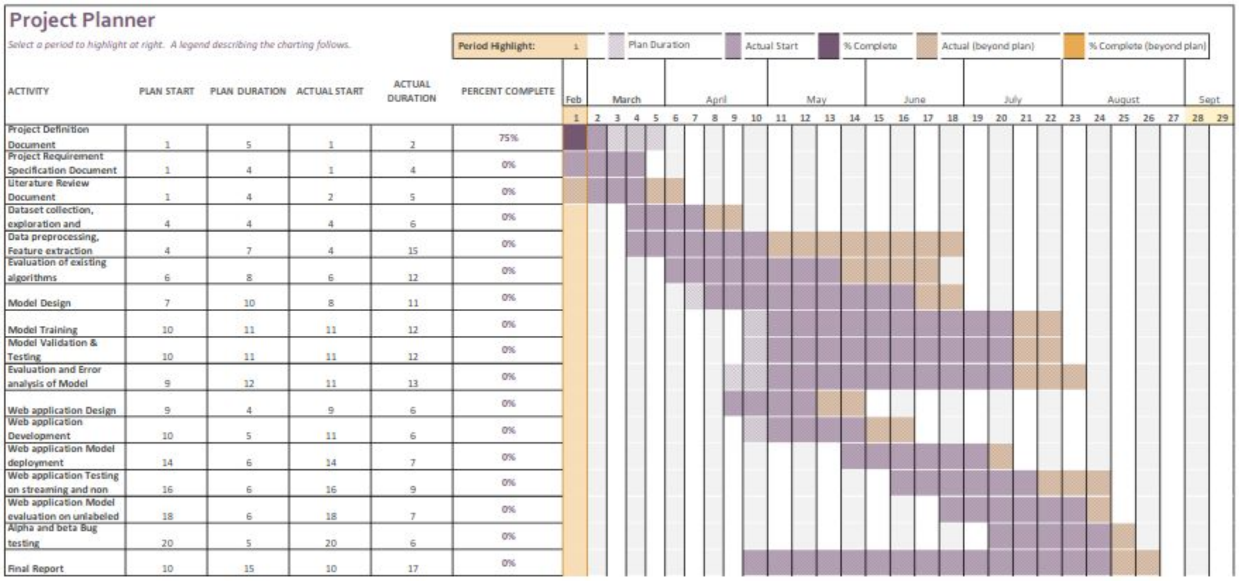
1. Tensorflow/Keras

2. Tensorflow.js

3. Web Application development - javascript

4. Deep learning libraries

TIMEPLAN



Project Planner

Select a period to highlight at right. A legend describing the charting follows.

Period Highlight: 1

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Project Definition Document	1	5	1	2	75%
Project Requirement Specification Document	1	4	1	4	0%
Literature Review Document	1	4	2	5	0%
Dataset collection, exploration and	4	4	4	6	0%
Data preprocessing, Feature extraction	4	7	4	15	0%
Evaluation of existing algorithms	6	8	6	12	0%
Model Design	7	10	8	11	0%
Model Training	10	11	11	12	0%
Model Validation & Testing	10	11	11	12	0%
Evaluation and Error analysis of Model	9	12	11	13	0%
Web application Design	9	4	9	6	0%
Web application Development	10	5	11	6	0%
Web application Model deployment	14	6	14	7	0%
Web application Testing on streaming and non	16	6	16	9	0%
Web application Model evaluation on unlabeled	18	6	18	7	0%
Alpha and beta Bug testing	20	5	20	6	0%
Final Report	10	15	10	17	0%