Statement of Work

**Fake news and Hoax Detection**

[***AIDI-1002-02 - AI ALGORITHMS I***](https://durhamcollege.desire2learn.com/d2l/home/329062)

Facilitator - Mr. **Marcos Bittencourt**

***Submission By***

**Yoginder Kumar – 100800560**

Rationale Statement:

News and exchange of information is ever existed. We are living in the 21st century and we have witnessed the rise and impact of the social media & other public networks in the life of every individual. Every second, billions of posts are being shared and every statement or piece of information that is posted online influence the people, communities and societies one way or the other. In the shadow of this fact, some anti-social, fraudulent activities like fake news and hoax spread for commercial and political means often get highlighted. This creates misinformation & ruckus among citizens; causing social, financial, psychological or physical suffering. So, in order to dodge these unwanted problems in such dynamic and fast platforms, we need some sort of system that can detect the legitimacy of the content being shared; thus sensitize the information before it reaches the end user.

Proposed Solution:

This project proposes the solution in the form of a Machine Learning model that can detect the piece of information on public internet platforms and validate its legitimacy to identify whether it is genuine piece of information. Or it is a false news, or a hoax created by someone for personal gains and ill will. This will help to take immediate action to the unauthorized news and turn it down before it does harm. We are talking about the huge internet datasets and to achieve this aim, a large volume of text and other kinds of information needs to be processed and analysed by a trained NLP model. That will be able to distinguish between real news from fake news and hoaxes.

Data Requirements:

The problem statement suggests marking distinction between the Real and Fake news. We are looking to the problem by coding perspective and training the machine learning model, in that context we would need:

* **Real News Dataset**: A data set with the content and its validation of being Real news by the authorized agencies.
* **Fake News Dataset**: A data set with unreal news posted online and were flagged fake because of no genuine basis.

The ultimate dataset we need is in the form as labelled data pointing True or Fake news. If we have two datasets, we will combine them to make it one.

**Data Sources**:

1. <https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset> (Primary Dataset)
2. <https://www.kaggle.com/hassanamin/textdb3>
3. <https://www.kaggle.com/mrisdal/fake-news>

Data Philosophy:

* One assumption is that the authentication provided in the dataset is correct. In essence, statement marked True are genuine cases and False or Fake news are wrong agenda. Both are assumed to be validated by the dataset creator to be said so.
* This model needs to be installed on the machine of the user, or server-side installation for the web-based services. It can be considered as a constraint.
* This model has a limitation to predict the result, only for the specific type for what the model is trained, let us say media in this case, it would predict. Testing for completely different domain might not be as accurate. However, this can be eliminated at advanced stages.

Testing Agenda:

Development of model will be phased in Three major categories. Training, Validation and Testing. Reference to the testing process before it hits the first phase will be as follows.

**Formatted Data**

Once we format the data according to the model, now data will be ready to be used for aforesaid three major Phases. Formatted Data will be fed to the model phase wise to develop and test the model thoroughly. At the time of writing, it is planned to use 70% of the dataset for Training of the model, and 15% for Validation and Testing each, using Python with different libraries and Algorithms. Graphic reference is as below.

**Formatted Data**

Useful Links:

Data Acquisition and Understanding.

<https://github.com/ze-row/Durham-College/tree/main/AI%20Algorithms/Data%20Acquisition%20and%20Understanding>

The project approach and parameters projected are tentative and aligned with the course’s plan and suitability. It is subject to change with the course plan and/or better suitable approach with further developments, the project plan would be adjusted accordingly, and SOW will be updated to keep the users up to date.

**Edit1**: Dataset preference highlighted. Access Link shared for EDA. and few minor changes.