Twitter Hate Speech Detection

- Project Proposal

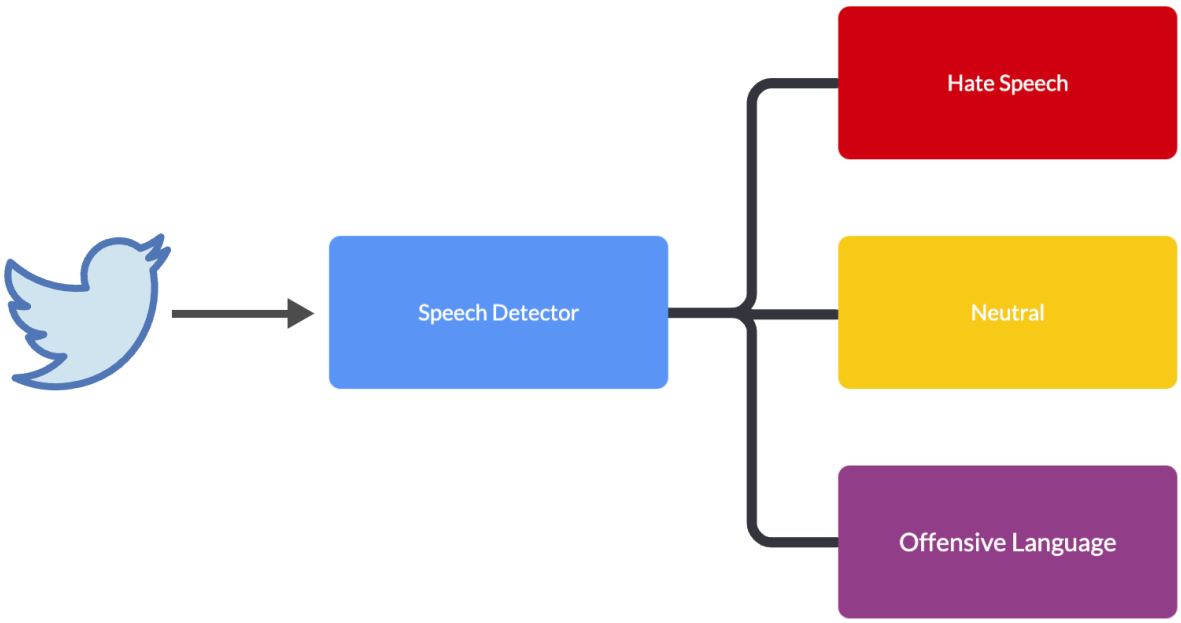
CSCE 5290: Natural Language Processing, Spring 2022

Instructor: *Dr Sayed Khushal Shah* (sayed.shah@unt.edu)

By: *Manoj Kolluri (manojkolluri@my.unt.edu)*

GitHub Link: https://github.com/ManojKolluri/NLP-Project-CSCE_5290

Idea Description



The basic idea behind this proposal is to develop machine learning and deep learning models that will be able to read a tweet and classify it to be belonging to one of 3 categories which are Hate Speech, Offensive Language and Neutral Language. At the end of the project, we would also like to integrate a User Interface which would allow the end user to be able to interact with the models and give a custom tweet as an input to these models and check the results produced by the models. Also, before integrating the User interface with the models the models would first be evaluated using certain evaluation metrics such as accuracy, F1 score, precision, sensitivity, and a confusion matrix using a testing dataset which would be separated from the main dataset prior to the training and validation processes of the models. Prior to the training process we would also be applying various text cleaning and NLP preprocessing techniques like stemming and lemmatization to remove unwanted data from the text used to train these models.

Goals & Objectives

Motivation & Significance:

Over the last 2 decades there has been a tremendous amount of growth in social media both in terms of its scale and in terms of its importance to a huge population of individuals and communities all around the world as one the major means of communication. However, the nature of these social media platforms is built in such a way that it allows any person in the world with a proper access to that platform to post or share any kind of information or express their opinion to millions of people using that platform which might sometimes be inappropriate or even repugnant. Amongst these kinds of inappropriate Hate Speech which is an abusive writing that usually expresses a certain amount of prejudice or ill intent against a particular person, or a group based on race, sex, or religion is one of the most important and dangerous aspects. Uncontrolled spread of such kind of speech can be damaging to the society in many ways therefore there is a need to automate the process of detecting and filtering out these kinds of content from the social media platforms.

The application that we are proposing to build in this project is a text classifier that is able to read a tweet from twitter which is one of the world's most famous social media platforms and try to classify the text belonging to either the category of hate speech, offensive speech or simply just neutral. We would also create a friendly user interface into which the user may be able type or enter a tweet and the models integrated within the UI classifies the tweet and generates an instant result. This would allow the users to gain a practical understanding of how these techniques work.

Objectives & Features:

The major objective of this project is to develop a certain number of machine learning and deep learning models and train them to be able to identify hate speech in twitter tweets and categorize them instantly upon giving a tweet as an input to the user interface.

Technical Goals:

1. Using different types of NLP techniques to preprocess and clean the dataset.
2. Segregating the dataset for testing and training proposes.
3. Creating Machine Learning models and evaluating them.
4. Creating a Deep Learning Model that takes tweets as inputs and evaluate its performance against the machine learning models.
5. Creating a User Interface and integrating it with the pre-trained models.

References

1. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. doi:10.1145/3041021.3054223

2. MacAvaney, S., Yao, H., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE, 14*(8), e0221152. doi:10.1371/journal.pone.0221152

3. Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ. Computer Science, 7*, e598. doi:10.7717/peerj-cs.598