LING 490 Project Proposal

Hate Speech Detection System

Group: Ariane Taraki, Erik Ly, Aditi Khanna

- **Goal (Problem definition and importance)**
    - **What am I going to do?**

Our group plans to build a system to detect hate speech (sexism and racism) within text, specifically in this case, tweets. Using an already existing annotated dataset, we plan to build our system by training and testing multiple ML models and doing error analysis on the respective results.

- - **Who would benefit?**

It is understood that much of hate speech is ill-defined in the linguistics community. Despite this, we believe our system could be used as starting place for detecting and flagging racist and sexist hate speech in text, specifically within social media platforms (i.e. Twitter in this case). By doing this, one would be able to detect when a social media environment is fostering hate speech, thus providing an opportunity to combat the speech in general to help create a more inclusive community. Additionally, our error analysis will be useful for other researchers in the field for determining how different ML models work when used in the system.

- **Problem Difficulty**
    - **Why is it hard?**

We anticipate that this may be difficult because we are not very familiar with the applications of various traditional and deep learning ML models, and this is a key aspect of our project. We will have to train these models properly on the existing dataset, decide on which features to test for each model, test within the data, and analyze the results in our error analysis. When interpreting the results, we will have to come together and accurately assess not only the performance of our models, but also why these results came to be.

- - **How hard is it?**

This project is moderately difficult, as we will have to intake a lot of data (16k+ tweets), create our models and train/test them in application to this data. Despite this, since we have many resources to base our system on, this should not be too difficult to achieve. We anticipate the hardest part being our error analysis, when having to accurately assess our model results. Overall, this should be feasible for our group, but still somewhat challenging.

- **Previous Work [this bullet point is NOT mandatory now, but recommended]**
    - **What have others tried?**

This project uses data gathered in Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter from researchers Zeerak Waseem and Dirk Hovy. The dataset was created from a model that served to detect racism and sexism within tweets submitted by users from the Twitter API. Additionally, this approach to a hate speech detection system is detailed as an Old Project for this class, which we are attempting to extrapolate on (delving deeper into model comparison and analysis).

- **Approach**
    - **What approach am I going to try?**

We will first use the tweet data set from Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter by Zeerak Waseem and Dirk Hovy as the foundation for our work. We will gather these tweets by parsing the Twitter IDs in question from the Twitter API, storing them for usage alongside their annotated categorization (“sexism”, “racism” or “none”). This data will then be used for training and testing on our Machine Learning models. Features and other specifics of our ML modeling system are to be determined, however we do plan to use some similar and different features for each model to be able to conduct a more in-depth error analysis. For our error analysis, we plan to compare the resulting performance of our models, as well as why we believe the performance differences exist or do not exist.

    - **Why do I think it will work well?**

We are confident that our system will work well as we have a lot of data and existing research to check our accuracy against and glean from. Additionally, we believe that by using multiple ML models and comparing them to each other, we should be able to detect how the models perform and conclude why. As long as we are able to successfully train and test our models, we should be able to complete proper error analysis.

- **Methodology**
    - **What resources (e.g.: data, annotations, etc.) are required and how am I going to get them (unless I already have them)?**

The resources we need for this project include the dataset of 16k+ tweets that we will be using to train our model and run our script upon. This can be found at: https://github.com/zeeraktalat/hatespeech. This dataset includes tweet IDs, associated with their annotated hate speech (“sexism”, “racism”, or “none”). The tweet content themselves will be gathered from the Twitter API.

    - **What computational models do I have in mind to solve the problem?**

We are currently still in the discussion phase of which specific ML models we plan to use on our system, currently we believe we may use the traditional Naive Bayes, Decision Tree and Linear Regression models . We may decide to substitute one of these traditional models for a deep-learning model instead, for better error analysis and performance comparison based on the context of our project. A supervised learning approach makes sense for our project as well.

    - **Which aspects of your model(s) are particularly hard?**

We believe that using these models successfully, and interpreting their results will be a challenge for our group. Traditional models (Naive Bayes, Decision Tree, Linear Regression) have challenges with datasets as slight or drastic variations from the original dataset can lead to these models being unable to perform. If we decide to do more difficult deep-learning models, we will have to implement a Multilayer Perceptrons (MLP) to allow the machine to make a connection of what is hate speech from hatebase.org.

.

- **What to do if the hard steps don't work out (meaning, what is my plan B)?**

If we are unable to properly use a ML model, our backup plan would be to substitute that model out for one we understand better (i.e. substitute a deep learning model with a traditional one). Additionally another backup would be to utilize other machine learning models such as SK Learn to evaluate our system. If no ML models end up working, a last case resort would be to compare model effectiveness using resources and research on the models that already exist.

If our dataset serves to be a problem with respect to training the models, our plan B would be to create a dataset ourselves from the already existing one with a different annotation schema containing features for easier evaluation. This might be difficult, but due to the fact that we already have an existing dataset, we would just have to build upon that data using a script to detect the hate speech within the tweets, possibly annotating them with level of offensiveness (from hatebase.org) or other features gathered from additional data sources.

- **Metrics**
  - **How will I measure success?**

We plan to measure the success of our project by determining if we are able to successfully test and train our chosen models on the data, and conduct error/performance analysis on the results. We would deem those models themselves to be successful if they reach over a certain performance result threshold for precision, accuracy, and F1 scores (specific threshold to be determined).

- **Summary**
  - **Why is this project of interest to me and what will I learn by doing this project?**

As regular users of social media, we have all seen the amount of hate comments and sexist/racist speech on platforms such as Twitter. This project is interesting in that it allows us to create a system that could potentially filter out these hateful comments. In the process, we get to learn about how to build effective models using ML and compare our results to current similar state-of-the art projects. We will also learn how to work together on a project of this size, working on our own but also coming together to evaluate our combined results.

- **Split of Work**

We plan to evenly split up the work for our project three ways. We all plan to contribute towards organizing and parsing the dataset for ML usage. All three of us will be in charge of training and testing one ML model using similar processes. For our current traditional models in mind, the split would work as follows (these are not set in stone): Aditi in charge of the Naive Bayes model, Ariane in charge of the Decision Tree model, and Erik in charge of the Linear Regression model. We plan to come together and fairly discuss our results and observations from using our respective models. The resulting analysis will thus be split evenly and serve as our model error analysis. By splitting the work in this way (each being in charge of a ML model), we are able to utilize our three-person group to the best of our advantage, maximizing analysis efficiency.