HatEval - Project Scope Document

Vaibhav Bajaj
Sriven Reddy
Kripa Anne
Karthika Ramineni

Problem Statement

- **Task 1:** Classify hateful tweets (where Hate Speech against women or immigrants has been identified) as aggressive or not aggressive.

- **Task 2:** Classify hateful tweets to identify if the target harassed is a generic group of people or a specific individual.

Introduction

Hate speech is defined as the language that threatens or insults a person or a group based on race, color, gender, religion, nationality etc. With the growth of social media, a huge amount of data is being generated. Therefore, detecting and limiting Hate Speech has become extremely important.

In task 1, given some tweets containing hate speech against women or immigrants, the goal is to identify whether the tweet contains aggressive language or not as hateful tweet doesn't necessarily imply that the language used is aggressive.

And in task 2, the goal is to determine whether the target is a specific individual or a group of people.

So, in this project, the whole idea is to come up with a model for task 1 that can classify the tweets as aggressive or not aggressive and similarly, a model for task 2 that can classify the tweets as hate speech against a specific individual or a generic group.

Method:

- Data Cleaning:
    - Convert to Lowercase.
    - Removing URLs.
    - Removing mentions. (the model will be trained both ways i.e. including mentions and without removing mentions to understand and observe the learning of model.)
    - Removing Stopwords.
    - Removing numbers (or Replacing with zero.)
    - Stemming
    - Restricting Character repetitions to not more than two. (e.g. Happpppy will be converted to happy.) (Again, the model will be trained without restricting and with restricting to observe if it affects the learning.)

- Embeddings:
    - We will try to experiment with word-level embeddings and character-level embeddings and see which one performs better.

- Model:
    - In the first deliverable, we will use basic models like SVM or logistic regression for classification.
    - Later we will extend it to CNN or LSTMs and will do a comparison of all the approaches to see which one performs better.

For both the tasks mentioned above, the same method will be followed as mentioned in the three points above(Data Cleaning, Embeddings, Model) just the data will differ and hence the respective classifications.

Challenges

- Hateful tweet doesn't imply aggressive language
- Tweets are shorter in length and may not be grammatically correct
- Different forms of hatred and target

References:

- Deep Learning for Hate Speech Detection in Tweets
- Hate Speech Detection: A Solved Problem?