Major Project Scope Document

Ananya Arun (20171019), Vijayraj S (20171026), Sumit Bhuin (2019201034), Virat Mishra (2019201033)

Mentor: Nikhil Pinnaparaju

1. Introduction

The main objective of our project is “Structure-based hate speech detection”. Traditional methods for hate speech detection use tons of training data to mine the hateful structure but due to disproportionate use of different terms, they are prone towards learning bias against specific objects, personalities or groups. Idea is to propose a method that takes into account the grammatical structure of the sentence to predict hatefulness.

2. Plans for Implementation

2.1 Preprocessing

2.1.1 Basic preprocessing and feature extractions

For starters, in order to incorporate sentence structure in the classification process, we need to build a component that performs a part-of-speech tagging on the sentences in the dataset.

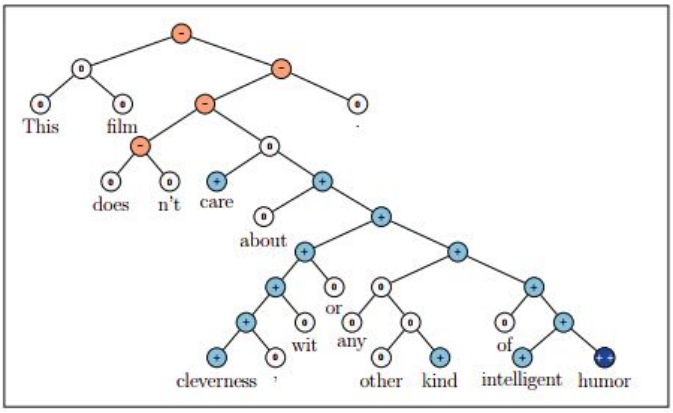
In pre-processing, we plan to remove references of proper nouns (racial, religious references, references to people etc.) and entities which are content-specific (like hashtags) to avoid bias in the classification process.

Stop words may or may not be removed, since they contribute to the sentence structure, and only final results can tell whether these would be useful in the final classification process.

2.1.2 Additional feature extraction

Sentiment Analysis

Sentiment analysis process uses NLP and text analysis to extract user’s sentiment polarity and subjective information. We plan to use Stanford NLP parser for the same, which generates a sentiment tree for each sentence in our dataset, and then annotates the sentences based on sentiment polarity and this can serve as a feature for our classifier.



A sample sentiment-tree

Linguistic analysis

It has been observed that sentences that are usually tagged as hate speech contain a lot of grammatical and spelling errors. Usually, people that express hatred against minorities are not well educated and this results in messages containing multiple misspellings or grammar errors which could be modelled into an additional feature for classification.

In our linguistic analysis, we propose to maintain an "edit distance" for every sentence. To calculate the edit distance for each word in the sentence we will consider the standard English dictionary and check for the word with smallest edit distance (That is the best match with the dictionary word). The edit distance for a sentence will be the average edit distance of all the words. (If a sentence is grammatically correct and spelt properly the edit distance will be close to 0)

Syntax analysis

Just like the above two steps, the syntax of a sentence has some relation to hate speech tagging as well. We can examine users' syntax in the sentence and check which ones promote hatred. We plan to use the Stanford NLP parser for the same. The parser tokenizes the text in the sentence to create syntax trees. We plan to experiment with using the scores of the parser as an additional feature. In case the syntax is wrong the parser will have a low score.

2.2 Classification process

In the readings, we went through as a team below, many methods were suggested to leverage the sentence structure for classification purposes. We firstly plan to start off with basic methods such as SVMs, logistic regression (with/without regularization) Naive Bayes models and CNNs, using the POS data in the form of a vector space model.

Most readings have gone with RNNs, LSTMs and modified versions of the same. We plan to implement classification using the same (since LSTM have hidden layer updates are replaced by

memory cells, making them better at finding and exposing long range dependencies in data which is imperative for sentence structures). We also plan to experiment with the additional modifications made in below readings ([1], [2]). If time permits, we may experiment with some hybrid models (like [3]), or try organising multiple well-performing models like a pseudo-random-forest-like model, where individual models get to vote on the final classification label.

2.3 Dataset selection

Most models leveraging sentence structure for classification seem to work well on datasets with single/two sentences and do not run in paragraphs. We plan to experiment with such datasets, and existing twitter datasets with the preprocessing mentioned above. For an example of the nature of datasets we plan to use, we have mentioned the links to certain datasets below.

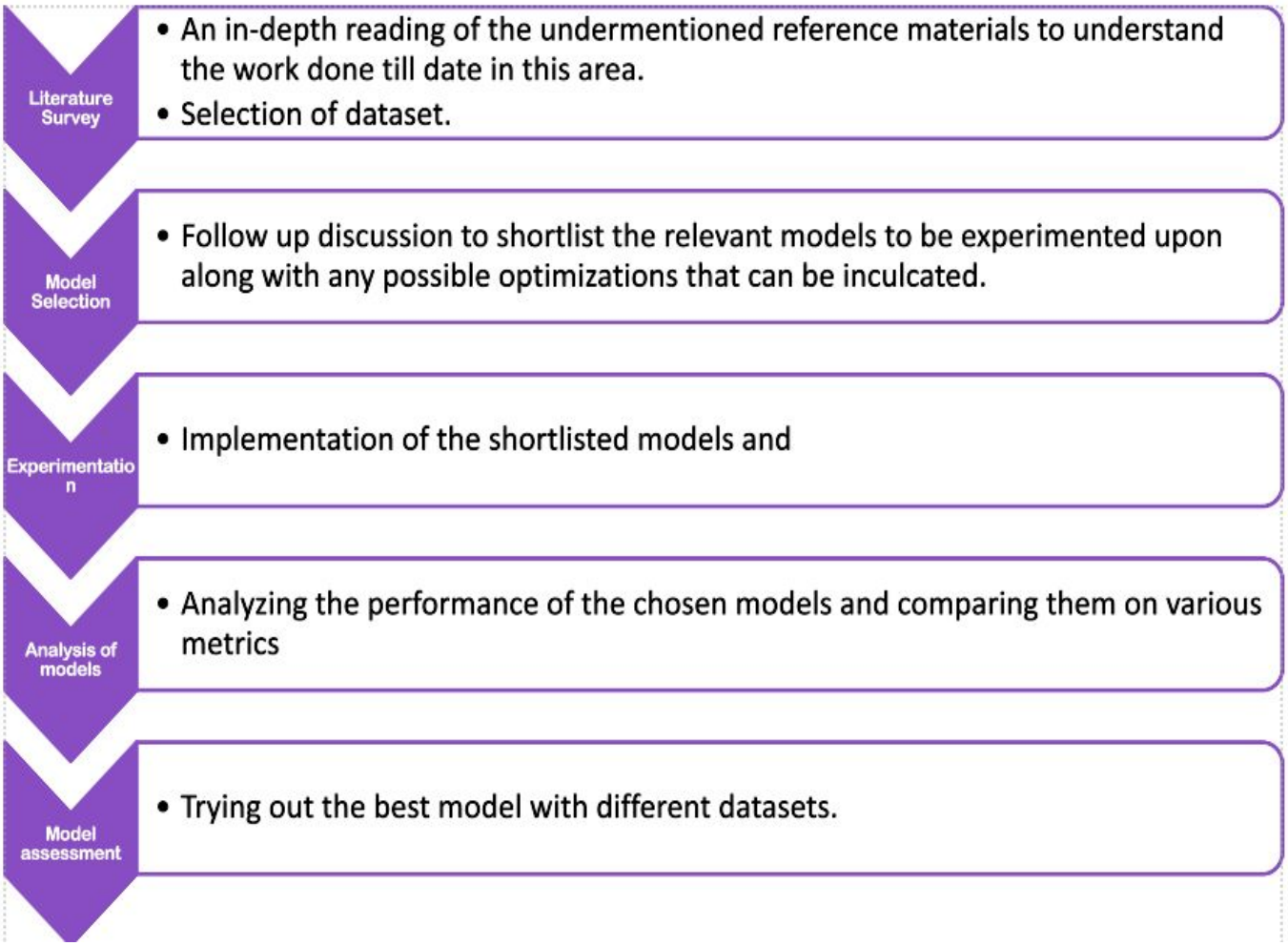
3. Relevant reading material

[1] https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/zhang.pdf
[2] http://proceedings.mlr.press/v37/zhub15.pdf
[3] http://downloads.hindawi.com/journals/cin/2019/8320316.pdf
[4] https://arxiv.org/pdf/2004.03705.pdf
[5]https://www.researchgate.net/publication/339076675_Hate_Speech_Detection_using_different_text_representations_in_online_user_comments?channel=doi&linkId=5e3c19a2458515072d83883a
[6]https://preventviolentextremism.info/sites/default/files/A%20Lexicon-Based%20Approach%20for%20Hate%20Speech%20Detection.pdf

4. Datasets

[1]https://github.com/sjtuprog/fox-news-comments/blob/master/annotated-threads/all-comments.txt
[2]https://github.com/Vicomtech/hate-speech-dataset

5. Milestones



This is the rough breakdown of our entire project, each phase expected to take roughly similar amounts of time. We will be following this workflow for our project execution.