CS314b - Machine Learning | Dr. Kusum Kumari Bharti

Fighting Hate with Machine Learning
A Literature Review

Introduction

Hate has been a political tool in recent years where majoritarianism has taken over the leaders and inclusive development remains an eye-wash. The growth of online social media services in the past decade like Facebook, Twitter, Instagram, WhatsApp etc. has provided certain elements yet another platform to spew venom in the minds of the people. Online Hate is deemed to be more personal and creates a long lasting impact given the scale and the manner through which it is spread. The targeted individual or community is also affected in a wide number of ways including trolling, rape threats, misogyny and even death threats resulting in mental torture and in extreme cases, even physical abuse.

Methodology

It hence becomes imperative for social media platforms and young engineers to use the technology at their disposal to eradicate targeted hate from Social Media Platforms.A number of platforms use manual flagging of content to determine whether the content violates their ill-framed policies. A lot of social media users experience trauma on a daily basis being subject to racial and communal hate on these platforms that are apparently within the content guidelines. It is important that there we realise that there is a very fine demarcation between censorship and fighting hate. Criticism of opinions is not hate but attacking someone for their opinion is. A tool must hence be developed to help a user ‘set their boundaries’ and create a custom classifier for each user. This way, we would not censor the ‘hateful’ content but will hide it from the concerned user so as to create a safe space for the her/him. At the same time, we would report the flagged content to the platform so that it may take action as required by law. This way, we are making social media platforms a safe space for everyone without censoring content on the platform.

Literature

We have reviewed two papers so far on the subject of our project.

1. Implementation Of Naive Bayes Classifier Algorithm On Social Media (Twitter) To The Teaching Of Indonesian Hate Speech (2017 International Conference on Sustainable Information Engineering and Technology (SIET))
2. Hate Speech on Twitter - A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection

While both of these papers are within the domain of our machine learning study, we observe that the first paper uses a simpler 'Naive Bayes Classifier' approach to machine learning rather than the second one that uses a 'Unigram' approach that requires a lot of annotation to be done before we start using the datasets for classification.

Course of Action

Since our methodology aims to create a custom model suited to our users, we should go with the Naive Bayes Model as it is easy to calculate and store for reuse.

We plan on building a classic Naive Bayes Classifier for each of our users to start with. As the user uses our tool, he will flag the offensive content that will shape his classifier more to his suiting. This way each classifier will be unique and will cater to the 'personal boundaries' of each user

Additionally, we will be working with Twitter since it is not possible to search for posts with the Facebook Graph API which is ironically very rigid as opposed to the content policy guidelines of the social media platform.

We will be using the Twitter API to focus on making Twitter a safe-space for the users.

Team

Ajwad Shaikh (2017333) | Manjot Singh (2017330) | Kanishk Goyal (2017322)