**HATE SPEECH MODERATION ON TWITTER**

PROJECT SUMMARY

# Overview - The Quick Pitch

This project is to develop a prototype solution that is based on simple sentiment analysis techniques to identify and block hate speech on Twitter, allowing an atmosphere that's non-toxic for users, as well as mitigating the skewness arising out of any violent form of communication during the 2020 US elections.

# The Problem

With the advent of social media platforms like Twitter in the past decade, individuals inclined towards racism, misogyny or homophobia have found an opportunity to reinforce their views on others who are not inclined to their beliefs. Internet culture often categorizes online hate speech as “trolling,” but the severity and viciousness of these comments has evolved into something much more sinister in recent years.

Across the United States, the inflammatory and confrontational tone of political rhetoric is creeping into public discourse and polarizing the electorate. It has created massive skew in the US elections in the past. Therefore, it's critical to block hate speech to reduce the skewness.

# The Solution

This solution has been built as a prototype to identify hate speech related to 2016 US elections, the same can be extended to block hate speeches during the 2020 US elections to avoid any skew arising out of any violent form of communication.

**Data Source:**

Twitter data mined using the following keywords, related to the US 2016 elections:

Petebuttigieg democrat impeachtrumpnow Election2020 2020election

USElection trump Obama supremacy belikepete

republicans democrats

**Steps for analysis:**

1. Static dataset for hate speech from Kaggle

2. Trained the model- using NLTK package

3. Pre-processing: tokenization, lemmatization, stop words removal

4. Algorithm- Logistic regression

5. Twitter mining for various key words

6. Applying model for classification

7. Hadoop - MapReduce for parallel operations

8. Hive – Create, load and run desired queries

9. Tableau to visualize results

# Analysis & Insights

• 7.05% of the tweets scraped were classified as hate speech

• 12.01% of the users tweeted hate-speech content

• Most common keywords in identified hate tweets

• Ability to identify the users that post the most hate tweets

• Ability to see the number of followers of these users to analyze the potential influence power or to check if these are bot accounts

References:

[1]<https://www.nyu.edu/about/news-publications/news/2019/june/hate-speech-on-twitter-predicts-frequency-of-real-life-hate-crim.html>

[2] <https://en.wikipedia.org/wiki/Online_hate_speech>

[3] <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons>