Increasing Efficacy of Federal Firearm Background Checks by Identifying Patterns of Threatening Language with Natural Language Processing Techniques

Aditya Bajaj, Gerrit Lensink, Ruby Han

Masters of Information and Data Science, School of Information, University of California, Berkeley

{abajaj225,rubyhan,gerrit.lensink}@berkeley.edu

Abstract (Light)

Following recent rises in firearm violence across the United States resulting in tragic losses of lives, many Americans and policymakers are calling for increased firearm regulation. Current background checks primarily include information such as criminal and mental health history and other civil cases such as domestic violence, acquired from three separate national databases.[1] These traditional methods have proven to be ineffective in identifying individuals that should have been legally prohibited from firearm purchase or possession, as recent as the Uvaldale, TX shooting in May 2022.[2] In order to prevent these tragic events, firearm background checks should include a wider range of indicators that identify an individual's violent intent that may not be reflected as past criminal behavior. These additional indicators can be made possible by the amount of publicly accessible data available across the web, such as tweets and online comment sections. This project focuses on threat detection from user-generated text across the web by leveraging a variety of natural language processing techniques.

Data

This project leverages the 'Jigsaw Toxic Comment' dataset [3] which includes examples of comments from Wikipedia across six categories: toxic, severe toxic, obscene, threat, insult, and identity hate. We subset the training data to focus threats, insults, and identity hate, the three categories we believe are most correlated with violent behavior.

Methods

This project's aim is to classify different user-generated texts as potentially violent or non-violent. Our initial implementations will focus on common architectures employed for sentiment analysis, such as BERT, CNN, and LSTM. Following implementation of neural network architectures, we will explore other architectures such as support vector machines which are known to be successful sentiment analysis tools.

For added project complexity this project will also leverage General Adversarial Networks (GANs) with the intent of training neural networks in limited resources or imbalanced data scenarios. While we do not believe we have a limited resource topic, we would like to explore this technique as it is a common technique across both NLP and other machine learning domains.

Expected Challenges

We anticipate challenges in dealing with data that is not as clean or formatted as data sources that have been used for homework assignments. Throughout the program we have also struggled with accuracy

[1] https://giffords.org/lawcenter/gun-laws/policy-areas/background-checks/background-check-procedures/

[2] https://www.washingtonpost.com/nation/2022/05/25/uvalde-texas-school-shooting-gunman/

[3] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview

related to imbalanced data, and expect it will continue to be problematic in this new domain. Additionally, we will be exploring two complex architectures (GANs, SVMs) that have not been discussed or implemented in class.

[~406 words (body)]

Literature Review (References Only)

1. Threat Detection in Online Discussions (link)
2. Using GAN-based models to sentimental analysis on imbalanced datasets in education domain (Link)
3. Aggressive Language in an Online Hacking Forum (link)
4. HateBERT: Retraining BERT for Abusive Language Detection in English (link)
5. GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples (link)
6. Assessing Violence Risks in Threatening Communications (link – non-technical paper used for background research on 'violent communication')