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, gerrit.lensink, rubyhan}@berkeley.edu

Abstract

This paper presents pre-trained BERT and CNN models with the purpose of identifying and detecting violent tendencies in written communication, with respect to labeled data across toxic, hate-speech, insults, explicit threats, and other abnormal comments. This is a complex task for multiple reasons. Firstly, due to the intrinsic nature and subjectivity of a ‘violent’ comment. In addition, a ‘violent’ comment may not necessarily originate from single words such as a swear word but rather from multi-word expressions. To address these challenges, we provide comparisons of various models across different pre-processing techniques and architectures, leading to an improvement in F1 score of 5 percent while using the DistilBERT model. The results obtained suggest that using BERT models is a promising approach and could be used to better understand and identify violent patterns in speech, which may be used to improve efficacy of background checks.

1 Introduction

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 regulations. Current background checks primarily include information such as criminal and mental health history as well as 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, based on the recent Uvaldale, TX mass school shooting in May 2022.[2] In order to prevent these tragic events, firearm background checks should include a wider range of indicators that identify violent patterns in an individual that were not reflected as past criminal behavior. These additional indicators may be made possible by the amount of publicly accessible data available across the web, such as tweets and online comment sections.

2 Background

Academics and industry researchers have focused on sentiment analysis using natural language processing techniques in order to better understand language and its complex structures. These advances allow expansion of exclusively numeric models to countless amounts of new data in the form of written and spoken communication.

Many of the leading models in the sentiment analysis domain are based on linear classifiers with fairly simple feature sets based on n-grams, bag of words, and other distillations of natural language. Across a range of similar domains to violence

[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/

classification, linear architectures such as SVM and naive bayes provide publication F1 scores in the range of 70 percent (Wester et al., 2016), improving F1 scores to the low 90s as recently as 2022 (Haq and Alshehri, 2022). As noted in Wester et. al's research, including more intelligent language features even resulted in reductions to model performance. Bag of words and bi- and tri-gram models outperformed models that included part-of-speech feature sets, as well as synonym network features based on WordNet.

As BERT's popularity has risen, more recent publications have leveraged pre-trained transformers layered with neural networks to improve sentiment classification in specific domains using true language models. One relevant example of these domain-specific tunings is the HateBERT model, trained specifically for the purpose of identifying hate speech and toxicity, as well as other themes such as misogyny and homophobia (Caselli et al., 2021). Pre-trained BERT models provide surprisingly strong prediction ability across a wide range of tasks, but tailor-made solutions such as HateBERT are able to improve classification by one to three points.

While linear, feature-based models provide strong scores in current literature, we believe true language models such as BERT allow for a higher ceiling in true natural language understanding with respect to context, complex references, and more. Therefore, we aim to expand this literature by studying BERT-based models in the domain of violence detection. We compare our tuned model to other cutting edge research such as HateBERT, as well as a base BERT classifier before tuning. This paper aims to improve a BERT baseline by three to five points, and reach F1 scores approaching 70 percent, reproducing scores achieved by other language model sentiment classifiers.

3 Methods

We approach our classification improvements in two different steps: First, we focus on understanding BERT's word embeddings and applying meaningful pre-processing, as well as transformation steps to data to coerce additional distance between sentence embeddings representing violent vs. non-violent comments. Second, we fine-tune our models searching across various sets of hyperparameters.

Google/Jigsaw Toxic Comment Dataset The jigsaw toxic comment dataset contains roughly 300,000 Wikipedia comments. Each comment can be labeled with one of six toxicity types. Labels include 'toxic', 'severe toxic', 'obscene', 'threat', 'insult', and 'identity hate'. Non-toxic comments will contain a zero across each of the toxicity types, and toxic comments can be labeled with any of the six labels, with up to six per comment.

Mental Health Professionals define patterns of violent behavior not only as explicit threats and written motives, but expand indicators to emotional state, variations in font size, and other factors (Glasgow and Schouten, 2014). Because language represented across our dataset is commonly considered as possible indicators for violence, we have chosen to distill all comments containing one or more toxic labels to 'violent', and labeling all other comments as 'non-violent', simplifying the original labeling.

Violent comments represented approximately ten percent of the overall corpus, which heavily skewed models toward the majority case in early modeling.

The train and test dataset originally consisted of roughly 150,000 records each, with six features representing each of the comment types, and a plain text comment. Comments in the raw data were unaltered from their form, characters, and length, based directly on the original Wikipedia comment.

Diving deep into each of the train and test datasets, it was noticed that the test dataset had 89,186 records with a value of -1 which indicates that the comment was not used for scoring hence the label type is lost. These records were dropped, resulting in the test dataset having 63,978 records. This resulted in a 80-20 split between train and test respectively for the experiments.

Embedding-focused Transformation The raw dataset consisted of comments written using colloquial language and slangs, with varying case and symbol usages. Initial transforms stripped all symbols and case, reducing comments to lowercase, text-only strings before tokenization. These transformations were based on common techniques in language processing to reduce noise in the data, but were actually found to reduce the power of the model.

These failures provoked a deeper study of validation examples, focused on the similarity of comment-level embeddings. BERT embeddings, the primary classification input, were found to show high degrees of cosine similarity between comments that our model would like to consider polar opposites. For example, running the sentences *"i want to hurt you"* and *"can i give you a hug?"* through our embedding model yielded a cosine similarity upwards of 96 percent.

During research, style patterns were identified that were strongly correlated specifically with violent comments. For example, on average 14 percent of violent comments contain fully-capitalized tokens compared to 5 percent of non-violent comments. Additionally, punctuation such as '!' or '!!!' were more commonly used in violent text vs non-violent. This motivated a tokenization and transformation strategy that pushes our embeddings in the relative direction we would like, before the classifier is trained. One such example that led our transformation strategy is outlined in Figure 1. Sentence A represents a text with no violent intent, but has key terms like *"kill"* and *"murder"*. Compared to our initial text in sentence C, the cosine similarity is fairly high at 80 percent. However, when we retain both capitalization and punctuation, we see relative decrease in similarity between benign sentence A, and violent sentence B. The returns are not as large as we would see if the embedding layer of BERT was retrained directly for our classification task, but it is evident that we can coerce our embeddings towards more expected dissimilarities with certain tokenization.

Figure 1: Example Text and Cosine Similarity

Sentence	Text
A	To kill a mockingbird is a pretty sad book that even deals with murder
B	I am going to FIND YOU and KILL YOU!
C	I am going to find you and kill you.

Cosine Similarity

	A	B	C
A	1		
B	0.76	1	
C	0.8	0.84	1

During further exploration, patterns were identified in violent comments such as the use of upper case letters and special characters or numbers to spell out profanity. While it is easy to understand comments like these as a human, it can be quite difficult for a model to identify patterns and context. To adjust for this, transformations were applied to both the training and testing datasets. Experiments included replacing and standardizing common toxic tokens such as *"fuck"* by replacing part or all with standard tokens such as "#", or other standardized keys to represent toxic tokens. Additionally, we experimented with removing profanity altogether from the dataset. Each of these attempts reduced F1 scores, likely due to the high correlation between unique symbolization in violent text versus standard representation in non-violent text.

Any user identifying information and IP addresses were removed to anonymize the data as the model should not be relying on any features which could be identifying individuals. This transformation was necessary based on privacy concerns, and is implemented even though it is responsible for a decrease in classification power.

Final transformations were fairly simple, but our search for an optimal tokenization strategy was quite iterative.

Tuning and Architecture Exploration Our research led us to what we believe is the optimal architecture for our classification tasks (Figure 2), given constraints in time and computing resources. The majority of experiments were run on relatively small train/val corpuses with the intent of understanding how different hyperparameter combinations affect model performance. We faced major challenges in overfitting, in which the model was interestingly overconfident in predicting the minority class - 'violent' text. This led us to study different regularization techniques such as dropout, undersampling, and oversampling. Un-intuitively, reducing class imbalance resulted in increased overfitting, so overfit reduction strategies focused on dropout, early stopping, and specific location of dropout between layers. Overfitting was reduced most by mid-low levels of dropout, lower epochs, and only placing dropout between hidden layers, which provided the classifier with the ability to learn better with stable weights out of the final layer.

Figure 2: Final Model Architecture

Hyperparameter	Value
Model	DistilBert
Max Sequence Length	128
Train/Val Split	80/20
Batch Size	32
Epochs	6
Dropout	0.1
Learning Rate	1e-5
Optimizer	AdamW
Hidden Layer Nodes	50
Hidden Layers	2
Undersampling	None

Throughout experimentation, nearly every model was built on the pretrained BERT-base-cased. However, final tuning iterations proved that a simpler model such as DistilBERT was more successful in reducing overfitting. We theorize that with fewer trainable parameters in the simpler DistilBERT infrastructure, overfitting is reduced. DistilBERT also provides substantial improvements in speed, reducing train time in final iterations.

Experimentation also included adding up to 12 hidden layers with hidden sizes up to 768 nodes, but simpler architectures were most effective in reducing overfit.

Baseline In addition to models tailored for other use cases, we benchmark ourselves on a BERT implementation only including transformations related to removing IP addresses. Without fine-tuning, BERT's pooled output with one hidden layer of 150 nodes is relatively successful with an F1 score just under 65 percent (Figure 3). This initial run provided a feasible starting point to approach Caselli's HateBERT model with an F1 score in the low 70s.

4 Results and Discussion

Following over 30 iterations testing tokenization, transforms, and hyperparameter strategies, our final model performs with an F1 score of 70.1 percent (Figure 3) with the hyperparameter combinations denoted in Figure 2.

Figure 3: Performance Comparisons for Various Model Architectures

Model	Description	F1 score
BERT (baseline)	No transformations	64.9%
Interim BERT A	Unsuccessful transforms	45.1%
Interim BERT B	12 hidden layers	67.6%
BERT + CNN	Added CNN output layer	69.2%
DistilBERT (final)	All transforms included	70.1%

Our final performance is heavily weighted by relatively low precision at 57 percent and recall at 85.3 percent. Precision is heavily skewed by the high amount of false positives, related to our overfitting problem discussed in section 3. Our overclassification of 'violent' texts did not match intuition, as we faced extreme class imbalance with only 10 percent of comments represented as 'violent'.

Following research into randomly sampled cases of false positives helped us identify two trends in misclassification: first, some comments contain reference to another comment such as *"he told me: 'i'm going to kill you' and it scared me"*. Second, the context was not always perceived correctly i.e. *"I'm going to beat you"* should be perceived as non-violent in the context of a domain such as competitive sports or bets, but has violent representations elsewhere.

In the case of the first confusion, we've learned that our modeling does not understand references well. Our inability to parse who is speaking throughout the comment leads to misclassifying comments as violent, when the speaker themselves may have been referencing someone else's violent intent.

This could be improved in future research by expanding the language model to specifically identify speaker versus external reference, and weight the speaker sentiment heavier than the reference. Additionally, trying models such as SpanBERT which extract relations and label them could add intelligence to the model. Our attempt at addressing this was by completely removing text within quotations, but our relatively unintelligent method proved detrimental to our success measure. Future research focused on other NLP techniques such as part-of-speech tagging and other context-based studies should be able to reduce the impact of this phenomenon.

5 Ethical Discussion

Violence is a sensitive topic in which flawed results may cause unintended harm. Thus, we considered a few ethical concerns related to our research.

Definition of Violent Comments We define violent comments as toxic, obscene, threatening, insulting or identity hate statements. We recognize that certain commenters do use violent language between friends in jest, as part of a hyperbolic speech or in a self-deprecating manner, therefore the labeling of less severely toxic comments is largely subjected to the annotator's upbringing, experience and social circumstances (Al Kuwatly et al., 2020). This may cause a level of over- or underestimation towards certain types of violent comments. For instance, words associated with profanities, curses/swearing or insults that are present in comments such as *"bro, you gay as fuck lmao! u're going straight to hell"* will be classified as violent regardless of the tone or intent of the commenter. Hence, our methods may present biases and in turn, the model is reproducible with inherent biases, so it should be applied with caution.

Bias Additionally, we recognize that certain groups choose to use language differently, without necessarily expressing violent intent. Our models are designed to identify violent intent regardless of racial, social, or economic background. This lack of sensitivity may lead to overestimation of violence in some person groups, leading to model bias. This concern must be addressed in order to ensure models are not responsible for unfairly identifying certain groups as violent.

Privacy While more data is always welcomed as additional valuable data will presumably increase our model metrics, maintaining commenter personal data privacy is critical. We took precaution in ensuring safety, security and personal information privacy by removing IP addresses and usernames from the corpus.

6 Conclusion

Our experimentation in transformations, tokenization, and hyperparameter tuning led to a roughly five point increase in F1 score from baseline for violent intent detection based on language model architecture. We attribute the majority of this improvement towards intentional transformations targeting increasing distance between vector representations of violent and non-violent comments, pre-training.

Additionally, our research was concentrated on coercion of embeddings, and then fine-tuning a classification layer based on static BERT pre-trained vector representations. Work related to fine-tuning the embedding layer for our particular task is also expected to improve performance.

Another way of improving precision for the model would be to add more toxic data to the current dataset. Due to the class imbalance, the model was predicting a high percentage of false positives.

Finally, using the above dataset it is shown that machine learning models may be used as a means of filtering candidates applying for a gun license. Using techniques outlined in this paper may lead to a safer America, driven by more holistic background checks.

References

Al Kuwatly, H., Wich, M. and Groh, G., 2020. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. [online] ACL Anthology. Available at: https://aclanthology.org/2020.alw-1.21/

Anul Haq, M. and Alshehri, M., 2022. Insider Threat Detection Based on NLP Word Embedding and Machine Learning. [online] ResearchGate. Available at: https://www.researchgate.net/publication/357577157_Insider_Threat_Detection_Based_on_NLP_Word_Embedding_and_Machine_Learning

Caselli, T., Basile, V., Mitrović, J. and Granitzer, M., 2021. Mitrović. [online] ACL Anthology. Available at: https://aclanthology.org/2021.woah-1.3.pdf

Glasgow, K. and Schouten, R., 2014. Assessing Violence Risk in Threatening Communication. [online] ACL Anthology. Available at: https://aclanthology.org/W14-3205.pdf

Palliser Sans, R. and Rial Farràs, A., 2021. HLE-UPC at SemEval-2021 Task 5: Multi-Depth DistilBERT for Toxic Spans Detection. [online] Arxiv.org. Available at: https://arxiv.org/pdf/2104.00639.pdf

Wester, A., Ovrelid, L., Velldal, E. and Hammer, H., 2016. Threat Detection in Online Discussions. [online] ACL Anthology. Available at: https://aclanthology.org/W16-0413.pdf

Data - Jigsaw toxic comment dataset
https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview

Appendix A

For complete list of modeling trials, see: https://github.com/gerritlensink/w266_final_project#models-built