RTI Press

Methods Report

ISSN 2378-7813

April 2023

Improving Text Classification with Boolean Retrieval for Rare Categories: A Case Study Identifying Firearm Violence Conversations in the Crisis Text Line Database

Robert F. Chew, Kirsty J. Weitzel, Peter Baumgartner, Caroline W. Oppenheimer, Brianna D'Arcangelo, Autumn Barnes, Shirley Liu, Adam Bryant Miller, Ashley Lowe, and Anna C. Yaros

RTI INTERNATIONAL

RTI Press publication MR-0050-2304

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Chew, R. F., Weitzel, K. J., Baumgartner, P., Oppenheimer, C. W., D'Arcangelo, B., Barnes, A., Liu, S., Miller, A. B., Lowe, A., and Yaros, A. C. (2023). *Improving Text Classification with Boolean Retrieval for Rare Categories: A Case Study Identifying Firearm Violence Conversations in the Crisis Text Line Database*. RTI Press Publication No. MR-0050-2304. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2023.mr.0050.2304

This publication is part of the RTI Press Methods Report series.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

Contents

About the Authors

Robert F. Chew, MS, is a senior research data scientist and program manager in RTI International's Center for Data Science and AI.

Kirsty J. Weitzel, MS, is a research data scientist in RTI International's Center for Data Science and AI.

Peter Baumgartner, MS, is a senior research data scientist in RTI International's Center for Data Science and AI.

Caroline W. Oppenheimer, PhD, is a research public health analyst in RTI International's Mental Health, Risk & Resilience Research Program.

Brianna D'Arcangelo, BS, is a public health analyst in RTI International's Community Safety and Wellness Program.

Autumn Barnes, BA, is a public health analyst in RTI International's Substance Use, Prevention, Evaluation & Research Program.

Shirley Liu, BA, is a public health analyst in RTI International's Substance Use, Prevention, Evaluation & Research Program.

Adam Bryant Miller, PhD, is a research clinical psychologist in RTI International's Mental Health, Risk & Resilience Research Program.

Ashley Lowe, MPH, is a research public health analyst in RTI International's Transformative Research Unit for Equity.

Anna C. Yaros, PhD, is the director of RTI International's Mental Health, Risk & Resilience Research Program.

RTI Press Associate Editor

Brian Southwell

Acknowledgments

This work was supported by the Centers for Disease Control and Prevention (CDC) National Center for Injury Prevention and Control (Award number: 1R01CE003295–01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH) or CDC.

Abstract

Advancements in machine learning and natural language processing have made text classification increasingly attractive for information retrieval. However, developing text classifiers is challenging when no prior labeled data are available for a rare category of interest. Finding instances of the rare class using a uniform random sample can be inefficient and costly due to the rare category's low base rate. This work presents an approach that combines the strengths of text classification and Boolean retrieval to help learn rare concepts of interest. As a motivating example, we use the task of finding conversations that reference firearm injury or violence in the Crisis Text Line database. Identifying rare categories, like firearm injury or violence, can improve crisis lines' abilities to support people with firearm-related crises or provide appropriate resources. Our approach outperforms a set of iteratively refined Boolean queries and results in a recall of 0.91 on a test set generated from a process independent of our study. Our results suggest that text classification with Boolean retrieval initialization can be effective for finding rare categories of interest and improve on the precision of using Boolean retrieval alone.

Introduction

Recent advances in machine learning and natural language processing have made text classification[1] an increasingly popular approach for information retrieval (IR). Text classification models are trained to assign discrete units of text, such as sentences, paragraphs, or full documents, into one of several pre-defined categories by learning from prior labeled observations. Text classification for IR has widespread adoption, with applications as diverse as biomedical systematic reviews[2,3] and legal document retrieval.[4,5] An advantage of text classification over non–machine learning IR methods is that models are capable of learning how to categorize text directly from a text's vocabulary and labels, as opposed to requiring researchers to develop their own rules for how to categorize observations consistently and accurately. Additionally, text classification models can rank observations based on predicted probabilities, allowing users to view observations most likely to be relevant first.

However, despite its popularity, developing text classifiers is challenging when the categories of interest are rare. This is because the normal procedure of drawing a random sample of observations to model on will infrequently return instances of the rare category. This lack of labels from the rare class makes modeling the concept difficult or requires significant effort from the research team to find enough examples to model the rare class effectively. For example, if the underlying prevalence for a rare category is 1 percent, research teams would be required, on average, to label 50,000 observations from a random sample to find 500 examples of the rare category.

Boolean retrieval can be an attractive alternative to text classification for finding instances of rare categories. Boolean retrieval[6] is a classical, non–machine learning–based IR technique in which observations containing one or more query terms are returned to the user. These queries are defined using Boolean logic, allowing the user to construct complex rules concerning the presence or absence of terms, or combinations of terms. For example, a Boolean query for finding observations containing the biological term "cell" might contain "("cell" AND NOT "prison")" to reduce the chance of returning text related to criminal justice.

An advantage of Boolean retrieval over text classification is that it does not rely on labeled data and, therefore, can return results of rare categories without requiring numerous examples. However, fine-tuning queries can become overwhelming when optimizing for accuracy. When developing queries, there is often a tension between the precise language needed to prevent false positives and the variety of language needed to recall all relevant observations, preventing false negatives. For example, to return a comprehensive set of observations related to soft drinks, an approach designed to minimize false positives may include an exhaustive list of specific name brands or products that would rarely be used outside of the context of soft drinks (e.g., "Coca-Cola," "Pepsi"). However, these would miss many observations that contain soft drink terminology specific to certain regions of the United States (e.g., "soda" or "pop"). Adding these regional terms would help us capture more soft drink observations, reducing false negatives, but at the expense of potentially including observations that use "soda" or "pop" in a context outside of soft drinks (e.g., "baking soda" or "pop music"), increasing false positives.

In this work, we explore ways of combining the strengths of text classification and Boolean retrieval for finding rare categories in a text corpus. Specifically, we use Boolean retrieval to find candidate labeled examples of rare categories, which we then use to train text classification models. As a motivating example to illustrate this methodology, we will seek to identify all text conversations referencing firearm injury or violence in the Crisis Text Line (CTL) anonymized database from 2018 to 2021.

CTL, one of the largest crisis text services in the United States,[7] is a not-for-profit company that provides no-cost, 24–7 confidential crisis counseling via text messaging and WhatsApp. When users of CTL's service text in, they are paired with trained volunteer Crisis Counselors. The primary goal of these exchanges is to de-escalate the crisis and help the texter reach a point at which they feel calm and safe. Volunteer Crisis Counselors offer mental

Table 1. Descriptive statistics of the Crisis Text Line data in the study period (August 2018–September 2021)

Conversations (N)	2,539,460		
Messages (N)	97,915,848		
	1st Quartile	**Median**	**3rd Quartile**
Messages per conversation	21	35	50

health support at the time of crisis and, if needed or requested, referrals to community organizations and other accessible resources.

Firearm injuries are a major public health issue and a leading cause of death for individuals ages 1–44 in the United States.[8] To date, CTL has supported approximately 7 million crisis-related text conversations,[7] providing a unique opportunity to better understand the low base-rate event of firearm violence. No study to date has investigated firearm-related text messages within CTL data. Innovative approaches that accurately identify firearm-related texts within all the available CTL messages will critically advance research on firearm violence.

Methods

Data

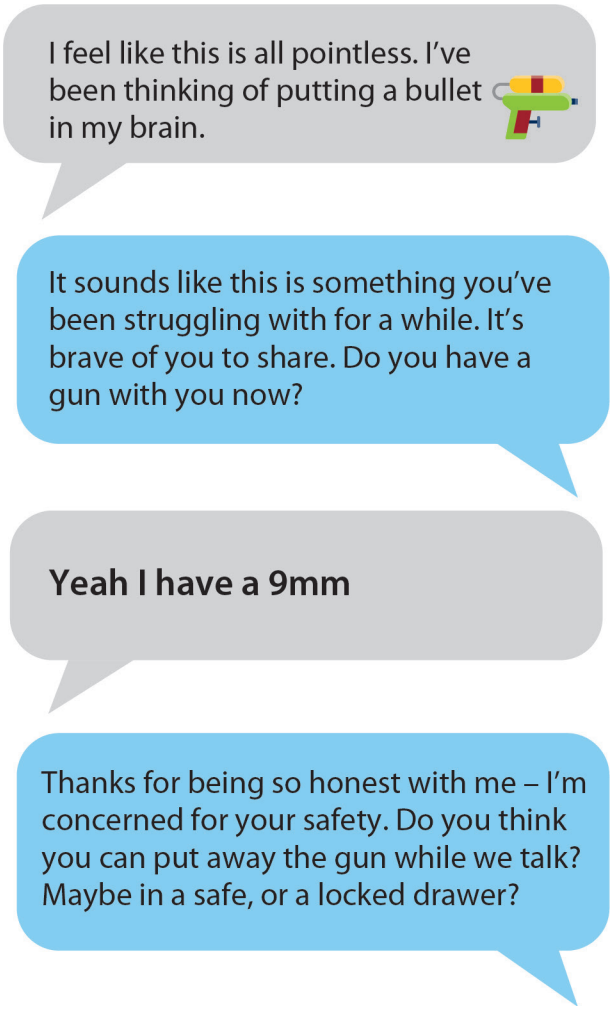
A conversation between a CTL volunteer Crisis Counselor and texter is made up of a series of text message exchanges. Figure 1 illustrates a fictional abbreviated example of a firearm-related conversation consisting of several messages exchanged between a CTL volunteer Crisis Counselor and texter. For the purposes of this study, a *message* will refer to a single text message sent by either the texter or volunteer Crisis Counselors, and a *conversation* will refer to the entire exchange of text messages related to a specific crisis instance from beginning to end.

This study uses de-identified English-language CTL message data from August 2018 to September 2021. In August 2018, CTL implemented their "Always Ask" policy, which requires volunteer Crisis Counselors to ask if the texter has had thoughts of suicide as part of each conversation. We chose August 2018 as the start date to capture only data collected while this policy was in effect. Conversations where texters did not engage after being connected with a counselor, along with conversations that led to a ban (i.e., pranks, inappropriate use) were not included in the dataset. Table 1 shows descriptive statistics of the study data.

Approach

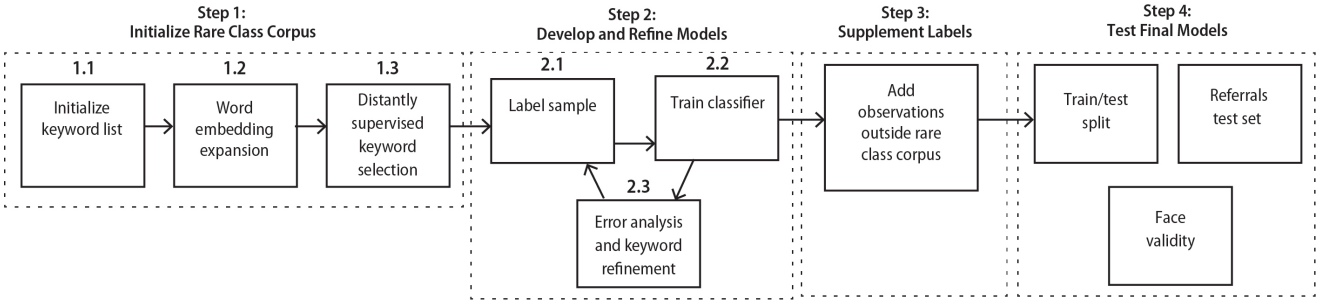
Our process of developing text classification models and Boolean queries consists of an iterative workflow (Figure 2) that allows us to refine both the models and firearm keywords over time. This workflow can be decomposed into four main steps: (1) conducting

Figure 1. Fictional abbreviated text exchange between a texter and a volunteer Crisis Counselor



 https://doi.org/10.3768/rtipress.2023.mr.0050.2304

Figure 2. Summary workflow diagram



an initial Boolean search of firearm-related keywords to return conversations with a higher likelihood of being the rare class; (2) labeling the results from the initial Boolean queries, training a text classifier, and performing error analysis (repeating as needed to obtain more training data and add more keywords); (3) supplementing labeled data with conversations from outside of the firearm corpus, to improve generalizability; and (4) validating the final text classifier numerous ways.

Step 1. Initialize the Rare Class Corpus

Typically, in machine learning, training data are assumed to come from independent and identically distributed draws from the population of interest.[9] However, taking this approach when the category of interest is rare will result in a training set with exceedingly few observations of the rare category, making modeling the concept of interest challenging. To get more observations expected to be of the rare class, we performed a Boolean search using firearm-related terms on the anonymized and de-identified CTL message text to create a firearm corpus (Step 1). Observations from this set were then sampled to create a text classification model with firearm and nonfirearm messages (Step 2.2), with a much higher proportion of the rare firearm-related messages than would be expected under a uniform random sample.

One challenge of taking this approach is that if the keywords used for Boolean retrieval do not fully capture all firearm-related conversations during our study period, the resultant text classification models may perform poorly on the types of firearm conversations that are excluded. Therefore, the main priority in developing the Boolean queries should be to capture as many potential firearm conversations as possible in the firearm corpus (maximize recall), even if it results in false positives (reduced precision). This motivates the need for several ways of expanding firearm-related keywords throughout the iterative workflow (Steps 1.1–1.3, 2.3). While we report details on our approach for creating a set of observations more likely to contain the rare class of interest, not all steps (1.1–1.3) are necessary and can be modified to take advantage of each project's unique context.

Step 1.1. Create Initial Keyword List

Our initial list comprised 23 firearm-related keywords (see Appendix) obtained from subject matter experts (e.g., "gun," "shooting"). These terms were added based on prior experience with crisis lines and knowledge of firearm terminology used in research and natural language settings (such as social media). Outside of our case study, developing an initial keyword list may be challenging if the research team is less familiar with the domain or if the concept of interest cannot be cleanly categorized by unique vocabulary alone. Even a cursory exploration of the results returned from an initial Boolean search can be informative in determining the likelihood of success for this approach.

Step 1.2. Word Embedding Expansion

To discover additional firearm-related keywords used by CTL texters, beyond the initial keyword list, we used a word embedding similarity search.[10] Word embedding models are a self-supervised learning approach that represents how a term is commonly used in text. Each embedding is a numeric vector, one for each word, that is learned directly from a text collection using an optimization routine. The optimization aims to predict a word based on the context of a small window of adjacent words. One

useful property of word embeddings is that terms with similar semantic meaning also tend to be similar numerically. Using a distance or similarity metric, we can designate how similar or dissimilar embeddings are from one another and use this information to find words used in the text conversations in a context similar to those provided in the initial keyword list. The word embeddings for this step were generated using the Word2Vec algorithm[11] trained on all CTL conversations in our study period. The word embeddings for the 23 terms in the initial set were then compared with their closest matches using the cosine similarity to see if any of the neighboring embeddings were novel firearm-related keywords. This process expanded the number of keyword terms from 23 to 37. The new terms ranged from plurals (i.e., "firearms," "rifles") to specific calibers of bullets ("9 mm").

Step 1.3. Distantly Supervised Keyword Selection

Lastly, we discovered suspected firearm-related conversations using supplementary CTL resources and searched within these to find additional firearm-related keywords. CTL volunteer Crisis Counselors fill out a survey after each conversation. If during the conversation a texter has indicated that they have a plan or access to a means of suicide, volunteer Crisis Counselors complete a free-text field in the post-conversation survey to input what means of suicide the texter indicated. We used this free-text field to identify conversations that contained one of our firearm-related terms and then manually reviewed their associated conversations to identify new candidate keywords. This approach was feasible because there were only 3 months of data available at the time of analysis (the field was added in June 2021) and because the field is only filled out for conversations with a higher suspected suicide risk (N = 814 conversations). This process expanded the keyword list from 37 to 44 terms. Additional keywords added during this process included "gsw," an abbreviation of "gunshot wound," and phrases including "shoot myself."

Although our instantiation of Step 1.3 (i.e., finding select firearm-related conversations and using them to generate keywords) was idiosyncratic to this case study, the broader approach of opportunistically finding examples of the outcome of interest and using those to develop features, or using them as labeled data, is not uncommon in the machine learning literature. In particular, distant supervision,[12] using an external data source to infer labels for an unlabeled sample, inspired our approach for this step. Although distant supervision is primarily used in the literature to generate labeled data, we modified the approach to support Boolean retrieval, allowing us to take full advantage of available CTL resources.

Step 2. Develop and Refine Text Classification Models on Rare Class Corpus

Despite having a corpus of firearm-related conversations in which each conversation contains at least one firearm keyword (Step 1), not all conversations in the corpus are about firearms, or more specifically, about firearm violence. To address this, the goal of Step 2 is to have coders label samples from this firearm corpus (Step 2.1) for two purposes: (1) to train text classification models that will allow us to classify a more specific category of firearm violence (Step 2.2), and (2) to find more relevant firearm keywords by assessing disagreements between the text classifier and labeled examples (Step 2.3). This process of expanding the firearm corpus, labeling data from the firearm corpus, and refining text classification models to better filter false positive observations is repeated over several rounds with the goal of developing a final text classifier that can then be applied to all 2.5 million conversations in our study period.

Step 2.1. Label Firearm Messages

The labeling process consisted of categorizing individual messages within a conversation based on their applicability to firearm violence. Human coders were asked to classify messages into one of three categories: "Applicable Mention of Firearm Violence" (AMF), "Nonapplicable Mention of Firearms" (NMF), or "No Mention" (NM). Our main category of interest, AMF, was used to identify all applicable mentions of firearms in which someone used, was using, or was considering using a firearm to harm or threaten any person. Messages were coded individually based on explicit statements and not on what might have been implied and required interpretation. Table 2 further describes the labeling options for the AMF code.

 https://doi.org/10.3768/rtipress.2023.mr.0050.2304

Table 2. Firearm category definitions used for labeling

Code	Meaning	Examples
Applicable Mention of Firearm Violence (AMF)	Yes, applicable	• I have a gun • He shot my mom • He made threats he was going to kill me with a gun • My thoughts were shooting my school
Nonapplicable Mention of Firearms (NMF)	Not applicable	• I would have a blast • I would take a bullet for him • Bullet journaling • Recreational firearm use (hunting, shooting range)
No Mention (NM)	No direct mention of firearms in the message	Messages that did not have a mention of firearms were coded as "No Mention," even if there were pronouns (e.g., "it," "them," "one") referring to a firearm that was previously mentioned in the conversation.

Across the iterations of this step, a team of four coders (CO, BD, AB, SL) labeled a total of 1,200 conversations, labeling each message in the conversation. Most conversations were labeled by only one coder, except for 40 conversations per round of 400 conversations that were labeled by all coders to assess inter-rater reliability. The percent agreement across coders on the sample was 93.3 percent. Gwet's AC1[13] was used as an additional metric to adjust for chance coding (AC1 = 0.929) because it does not suffer from the "high agreement, low reliability" paradox sometimes experienced with other common reliability metrics.[14,15,16]

Step 2.2. Train Firearm Text Classifier

Text classification models were created using the training data generated by the labeling process. The three-level AMF coding was transformed into a binary target by combining codes NMF and NM. The final binary classification model's target was AMF (1 = AMF, 0 = No AMF). The model was developed using spaCy,[17] a natural language processing Python package. Specifically, we use spaCy's *TextCatEnsemble* model, which combines a neural network model and a linear bag-of-words model via a stacked ensemble. The model produces a predicted probability that each message is firearm-related (AMF). A threshold of 0.5 was used to transform the probability into a binary label, with over 0.5 designated as an AMF. The performance for the final text classification model can be found in the results section.

Step 2.3. Error Analysis and Keyword Refinement

Error analysis[18] is the process of analyzing misclassified examples for the purpose of improving the model or fixing mislabeled data. For our use case, this included manually reviewing (1) cases in which the model predicted the message to be an AMF but the coder did not classify it as such; (2) cases in which the model predicted the message to be not an AMF but both the coder and Boolean search classified it as an AMF; and (3) cases in which the classifier and coder agree on the AMF designation, but the designation was missed by the Boolean search. Scenario (3) is possible because each firearm conversation contains both firearm and nonfirearm messages and both text classification and keyword matches were assigned at the message level. Most changes resulting from the error analysis were to either add keywords (e.g., "shoot [pronoun]") or change labels that were misapplied by coders. It also provided insight into common scenarios in which the Boolean retrieval, and sometimes text classifier, struggled (e.g., idioms such as "jumping the gun" or mentions of recreational gun use).

Based on insights from the error analysis, we repeated the earlier components of Step 2, adding new keywords and Boolean logic to extend the number of potential AMF conversations. We then labeled new conversations based on this extended corpus, re-trained the AMF classifier, and performed error analysis. We stopped this cycle after three rounds, using a combination of model performance and error analysis feedback as a stopping criterion. Repeating this process added 79 new keyword

 https://doi.org/10.3768/rtipress.2023.mr.0050.2304

Table 3. Volume of messages and conversations returned from Boolean retrieval

Step	Description	Terms (N)	Messages (N)	Conversations (N)
1.1	Initial Keyword List	23	82,988	52,461
1.2	Word Embedding Expansion	37	87,742	52,848
1.3	Distantly Supervised Keywords	44	105,964	62,035
2	Final Keywords	123	118,577	69,770

combinations to the Boolean query (total of 123). Table 3 depicts the growth in the number of conversations returned from Boolean retrieval after each modification to the keyword list.

Step 3. Supplement with Labeled Data Outside of the Rare Class Corpus

Although labeled data from the firearm corpus is useful for finding examples of the rare class to train on, because we only labeled messages in conversations returned from the Boolean queries, the models were only able to learn from messages and conversations from the firearm corpus. Although this process is designed to iteratively expand the firearm corpus throughout, we could systematically miss a portion of firearm conversations if our keywords were not extensive enough. Furthermore, if messages outside of the firearm corpus were substantively different than those in the firearm corpus, there are no guarantees that the model would be able to accurately predict the nonfirearm corpus conversations well.

As a final modeling step, we addressed these issues by labeling additional messages from conversations *not* included in the firearm corpus. While we could draw a uniform random sample of these conversations to label, this would return mostly nonfirearm conversations because the chance of randomly selecting a firearm conversation that does not already contain a firearm keyword is small. As an alternative, we used the current text classifier to predict on the conversations that were not part of the firearm corpus and created two groups: (1) those that were not in the firearm corpus and predicted as AMF = 1; and (2) those that were not in the firearm corpus and predicted as AMF = 0. We then drew a stratified random sample of 400 conversations across these groups to label. We did not share with the coders to which strata the conversations belonged, to avoid influencing their labeling decisions.

Conversations that were not in the firearm corpus and predicted as AMF = 1 are conversations that the text classification model believed were relevant but that did not contain a firearm keyword. Human feedback on these predicted labels is valuable for improving the text classification model because it will both expose instances where the model predictions are incorrect while also reinforcing correctly predicted observations. Conversations that were not in the firearm corpus and predicted as AMF = 0 were expected to overwhelmingly be nonfirearm related, since they were neither predicted to be AMF nor did they contain a firearm keyword. However, they were important to include in the training and test sets since nonfirearm conversations comprised most conversations in our study period, and we wanted to confirm that the final model could correctly classify them as AMF = 0.

Step 4. Test Final Model

We performed three evaluations to test our final models:

1. **Model Comparison.** Compared the performance of the final Boolean retrieval and text classification models on a random sample of the labeled data.
2. **Model Validation—Independent Test Set.** Tested the performance of the final text classification model on set of AMF conversations, generated independently from our firearm corpus approach.
3. **Model Validation—Face Validity.** Tested our assumption that the final text classification model would identify more AMF conversations than an initial Boolean search. We also compared the number of conversations predicted as AMF = 1 by the text classifier with the number returned in the expanded Boolean search.

https://doi.org/10.3768/rtipress.2023.mr.0050.2304

Model Comparison

To compare the final text classification and Boolean retrieval approaches, we assessed both on a hold-out test set of a random 20 percent of the labeled data not used for training ($N = 320$). We calculated the class-specific precision, recall, and F1 score, as well as the overall accuracy for both models.

Model Validation—Independent Test Set

For validation, we created a final test set of conversations about firearm violence generated from CTL's operations, entirely independent from our approach. Performing well on this set should give us greater assurances that our model can find conversations related to firearm injury and violence that are not dependent on the choices made in constructing the firearm corpus.

To validate our final text classification model, we used conversations containing referrals to resources related to mass shootings and gun violence as our independent test set. Resource referrals are materials shared with texters by volunteer Crisis Counselors as a means of extra support following the conversation. These referral resources provide additional information and support on a variety of topics such as suicide, gun violence, coping strategies, and domestic violence. The firearm-related referrals shared in conversations by volunteer Crisis Counselors in our study period are both from the Everytown for Gun Safety Support Fund[19]:

- *Trauma and Gun Violence*, which is described to volunteer Crisis Counselors as a referral that "provides information on combating gun violence, coping with the aftermath of a mass shooting, and has a forum for survivors."
- *Everytown Support Fund*, which is described to volunteer Crisis Counselors as a "Trauma and Gun Violence sheet [that] explains trauma after gun violence and shares coping skills."

This independent test set does not contain observations from the labeled training or validation sets used to develop the text classifiers.

Model Validation—Face Validity

Based on our mental model of how text classification should interact with the Boolean retrieval, we hypothesized that the text classification model would result in more AMF conversations than the initial keyword list, because it was built using a more extensive set of expected firearm-related messages. We also hypothesized that the text classification model would identify more AMF conversations if it performed vastly better than the final Boolean query, with the quantity returned by both converging as the gap in performance shrinks. To test this hypothesis, we estimated the number of AMF conversations using the original keywords, final keywords, and the final classification model on all conversations within the study period.

Results

Model Comparison

To compare the final text classification and Boolean retrieval approaches, we assessed both on a hold-out test set of a random 20 percent of the labeled data not used for training ($N = 320$). Our results, summarized in Table 4, show that the text classification model outperforms Boolean retrieval on most model performance metrics. Notably, we observed higher precision for AMF = 1 (0.92 vs 1.00), higher recall for AMF = 0 (0.81 vs 1.00), and higher overall accuracy (0.93 vs 0.96). The text classifier also outperformed Boolean retrieval for both classes when assessed using the F1 score (AMF = 0: 0.89 vs 0.94; AMF = 1: 0.95 vs 0.97). Boolean retrieval outperformed the text classifier in precision for AMF = 0 (0.98 vs 0.89) and recall for AMF = 1 (0.99 vs 0.95).

Model Validation

To validate the final model, we used referrals given to texters by volunteer Crisis Counselors as an independent test set. In our study period, there were 57 conversations in which a volunteer Crisis Counselor shared a gun violence-related referral, indicating the conversation was likely firearm-related. After coding, 54 of the conversations were found to be firearm related. Of the 54 conversations, the final model labeled at least one message in 49 of these conversations as an AMF, for a recall of 91 percent.

Table 4. Performance of the final Boolean retrieval and text classification model

Type	AMF Label	N	Precision	Recall	F1 Score	Accuracy
Boolean Retrieval (Keywords = 123)	0	101	**0.98**	0.81	0.89	0.93
	1	219	0.92	**0.99**	0.95	
Text Classifier	0	101	0.89	**1.00**	**0.94**	**0.96**
	1	219	**1.00**	0.95	**0.97**	

Note: AMF = Applicable Mention of Firearm Violence.

As a final test of face validity, we calculated the number of AMF conversations found using the original keywords, final keywords, and the final classification model. Table 5 summarizes the estimated number of AMF conversations in our study period using all three approaches. The final text classification model identified 2.82 percent of conversations (N = 71,839) in our study period as AMF conversations, compared with only 2.07 percent (N = 52,461) of conversations in our study period returned by the initial keywords and 2.75 percent (N = 69,770) of conversations returned by the final keyword list.

Discussion

Text classification with Boolean retrieval initialization was effective in creating a refined set of AMF conversations, a rare class that was initially expected to only appear in roughly 2 percent of all CTL conversations. These models outperformed an iterative Boolean retrieval process and were better able to distinguish between when the firearm keywords indicated an AMF and when they were used in other contexts. This was demonstrated on both an adaptively constructed set of conversations and a separate test set of conversations containing gun violence resource referrals. Our use of the gun violence resource referrals as an external test set allowed us to assess this approach on observations generated from a different mechanism than the firearm keywords, providing additional support for the method.

The test of face validity confirmed our hypothesis that the final text classifier would identify more conversations as AMF = 1 than were returned by the initial keyword list and would return a similar number of conversations to the final keyword list if their performance was similar. Although the text classifier outperformed the final Boolean retrieval overall, the iterative development process allowed both to perform well at identifying AMF. Furthermore, the iterative nature of this process demonstrated value in finding instances of the rare class, given that the final estimate of AMF conversations (N = 71,839) is well above the initial estimate using only keywords from subject matter experts (N = 52,461). Being able to identify rare categories, such as firearm violence and injury, can improve crisis lines' abilities to support people with firearm-related crises or provide appropriate resources more accurately.

This approach is designed to be flexible and allow teams to take advantage of project-specific resources. For example, we added Step 1.3 because we had access to tables with relevant information that could be used to refine the Boolean queries; while this provided more ideas for increasing the size of the firearm corpus, the approach could still be implemented

Table 5. Estimated number of AMF conversations in study period using the initial Boolean search, final Boolean search, and final text classification model

Type	AMF Conversations	
	(N)	(%)
Initial Boolean Retrieval (Keywords = 23)	52,461	2.07%
Final Boolean Retrieval (Keywords = 123)	69,770	2.75%
Final Text Classification Model	71,839	2.82%

Note: AMF = Applicable Mention of Firearm Violence.

https://doi.org/10.3768/rtipress.2023.mr.0050.2304

without this step given the multitude of other ways for developing Boolean queries (e.g., Steps 1.1 and 1.2). Additionally, while we focused on firearm injury and violence as a motivating example, our approach could be extended to any rare class in which domain knowledge can be used to identify positive cases at least partially. Although we only explored binary classification, this approach could also be expanded to multi-class classification by creating different Boolean queries for each rare class and drawing stratified samples across the rare classes to label. In our use case, each conversation in the firearm corpus often consisted of both firearm and nonfirearm messages. Since we were building text classification models at the message level, this reduced the need for labeling nonfirearm conversations at the outset because there were already both firearm and nonfirearm messages available for training from the conversations in the firearm corpus. In cases where the unit of analysis does not exhibit this hierarchy, documents outside the rare class corpus should be incorporated sooner.

Although the final text classification model improved upon the accuracy of the Boolean query, depending on the goals of the analysis, an iterative Boolean retrieval process as described above may be an attractive alternative if sufficient resources are not available to support text classification. Supervised machine learning methods require labeled data both to assess the method and to train the model. The labor required to generate these additional labels may not be worth the effort if there are competing budget or time constraints or if the intended use case has a higher tolerance for misclassification. Boolean retrieval also benefits from being fast to apply at inference time and being transparent in how it makes classification decisions. Especially when the category of interest is rare, it may be prudent to start with assessing the performance of a modified Boolean retrieval process to determine whether adding text classification would benefit the project.

Related Work

Given the ubiquity of the methods, several papers have compared the performance of Boolean retrieval and text classification. Turtle[20] compares natural language query techniques to Boolean retrieval for searching full-text legal materials and finds that the natural language systems outperform expert-crafted Boolean queries. Cohen et al.[21] also compare text classification to a search query approach and find that for most topics studied, text classification improved precision while keeping competitive recall. More recently, Westermann et al.[5] compare Boolean search to several machine learning methods, such as random forests,[22] support vector machines,[23] and a linear model using fastText[24] embeddings. They similarly find that text classification methods outperform Boolean queries, while they also acknowledge that Boolean search benefits from being highly interpretable. Although these studies do not attempt to combine text classification with Boolean retrieval, our overall findings agree with these prior works and suggest that text classification outperforms Boolean retrieval, even in the setting of rare categories.

Using domain knowledge to develop queries that are more likely to contain the class of interest is common in data programming.[25] Data programming is a weak supervision method that consists of three steps: (1) developing labeling functions to programmatically encode domain knowledge about the categories of interest; (2) combining the output of the labeling functions using a generative model to develop a best-guess estimate of the true class label; and (3) using the estimated class labels to train a supervised machine learning model. When the data modality is text, keyword-based Boolean queries such as those used in this work are often employed as labeling functions. Future work could extend our approach to incorporate ideas from data programming, such as training a generative model from the individual Boolean queries for each keyword to create a more refined rare class corpus.

A final related research area comprises methods for finding instances of rare classes to support classification. The majority of these efforts use active learning,[26] a subdomain of supervised machine learning that iteratively uses model feedback to recommend which observations to label next. Pelleg & Moore[27] propose an active learning strategy that allows the component of a mixture model to "nominate" its favorite queries to find extremely rare classes in the presence of noise. Although this method assumes a mixture model to fit the data, it

does not require a particular functional form for the mixture components. Hospedales et al.[28] have developed an active learning strategy to jointly address classification and rare class discovery, a challenging setting in which the target classes of interest are unknown a priori. They propose a generative-discriminative model pair to combine the discovery properties of generative models with the superior classification properties of discriminative models. Mullapudi et al.[29] propose an active semi-supervised method that incorporates techniques for learning under extremely imbalanced data for images[30] to label the "easy" negative examples, leaving the "hard" examples for human labelers. Lastly, closest to our work, Attenberg & Provost[31] compare using a search strategy to initialize a model with examples of the rare class to using popular active learning strategies. They also propose a hybrid model that uses search strategies (e.g., Boolean retrieval) to initialize an active learning model. They find the hybrid approach shares attractive properties of both the search and active learning strategies and outperforms both individually. Of the literature for finding instances of rare categories for classification, this work is the most like ours, in that both use search methods to find cases of the rare class to initialize a model and then perform iterations of model refinement. Our work helps clarify that this general approach can be effective even when not using active learning and provides comparisons to an iterative Boolean retrieval process.

Study Limitations

Results from our study should be interpreted within the context of several limitations. One limitation is that it is infeasible to determine exactly how many firearm conversations are in the CTL database for the study period. While our evaluation metrics suggest that the final text classification model outperforms keyword-based retrieval, the exact number of AMF conversations is still unknown and would be difficult to confirm outside of full enumeration. Another limitation is that the conversations are scrubbed of personally identifiable information by CTL before our use. Although this is effective in preventing disclosure of a texter's personal information, the algorithm may on occasion mistakenly redact language that contains firearm-related keywords. Additionally, including redactions in conversations while modeling may affect performance, although our error analysis suggests this impact is minor. The firearm-related resources shared during our study period include only referrals to organizations that focus on gun violence prevention. This scope provides less insight into model performance on other firearm-related crises, such as firearms as a means for suicide. Lastly, while we present a multi-step approach in this work, we did not have the resources to conduct an ablation study to better understand how each individual component contributed to the overall success of the method. Future work could embrace this focus, as well as replicate the approach across different datasets to better understand under which conditions it is more or less likely to be beneficial.

Conclusion

Creating text classifiers from scratch can be challenging when the category of interest is rare. In this work, we develop an approach that combines strengths from text classification and Boolean retrieval to find instances of rare categories, using the goal of finding all firearm violence conversations in the CTL database between 2018 and 2021 as a motivating case study. Identifying rare categories, like firearm violence and injury, can improve crisis lines' abilities to support people with firearm-related crises or provide appropriate resources. We find that this approach improved on a refined Boolean search alone and returned nearly 20,000 more relevant cases than an initial Boolean search using only terms provided by subject matter experts. Research teams requiring high-quality results should consider text classification with Boolean retrieval initialization for detecting rare categories of interest.

 https://doi.org/10.3768/rtipress.2023.mr.0050.2304

References

1. Dwivedi SK, Arya C. Automatic text classification in information retrieval: a survey. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16), Association for Computing Machinery, New York, NY, USA. p. 1–6. 2016. https://doi.org/10.1145/2905055.2905191
2. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8:163. https://doi.org/10.1186/s13643-019-1074-9
3. Martinez D, Karimi S, Cavedon L, Baldwin T. Facilitating biomedical systematic reviews using ranked text retrieval and classification. Proceedings of the 13th Australasian Document Computing Symposium, Hobart, Australia. 2008 Dec 8. https://www.researchgate.net/publication/216591571_Facilitating_Biomedical_Systematic_Reviews_Using_Ranked_Text_Retrieval_and_Classification
4. Moens MF. Innovative techniques for legal text retrieval. Artif Intell Law 2001;9(1):29–57. https://doi.org/10.1023/A:1011297104922
5. Westermann H, Savelka J, Walker VR., Ashley KD., Benyekhlef K. 2021. Computer-assisted creation of Boolean search rules for text classification in the legal domain. https://doi.org/10.3233/FAIA190313
6. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press; 2008. https://doi.org/10.1017/CBO9780511809071
7. Pisani AR, Gould MS, Gallo C, Ertefaie A, Kelberman C, Harrington D Individuals who text crisis text line: key characteristics and opportunities for suicide prevention. Suicide Life Threat Behav 2022;52(3):567–82. https://doi.org/10.1111/sltb.12872
8. Centers for Disease Control and Prevention. WISQARS (Web-based Injury Statistics Query and Reporting System). 2021 [cited 2022 Oct 19]. Available from: https://www.cdc.gov/injury/wisqars/index.html
9. Mehryar M, Rostamizadeh A, Talwalker A. Foundations of machine learning. Cambridge (MA): MIT Press; 2012.
10. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers), Association for Computational Linguistics, Berlin, Germany. p. 367–77. 2016. https://doi.org/10.18653/v1/P16-1035
11. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems 26. Curran Associates, Inc.; 2013 [cited 2020 Apr 24]. p. 3111–9. Available from: https://proceedings.neurips.cc//paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
12. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore. p. 1003–11. 2009 [cited 2022 Oct 18]. Available from: https://aclanthology.org/P09-1113 https://doi.org/10.3115/1690219.1690287
13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(1):29–48. https://doi.org/10.1348/000711006X126600
14. Falotico R, Quatto P. Fleiss' kappa statistic without paradoxes. Qual Quant 2015;49(2):463–70. https://doi.org/10.1007/s11135-014-0003-1
15. Feng C. Mistakes and how to avoid mistakes in using intercoder reliability indices. Methodology 2015;11(1):1–10. https://doi.org/10.1027/1614-2241/a000086
16. Zhao X, Liu JS, Deng K. Assumptions behind intercoder reliability indices. Ann Int Commun Assoc 2013;36(1):419–80. https://doi.org/10.1080/23808985.2013.11679142
17. Montani I, Honnibal M, Honnibal M, Van Landeghem S, Boyd A, Peters H, explosion/spaCy: v3.0.0: transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more. 2021. https://doi.org/10.5281/zenodo.4486473

18. Ng A. Machine learning yearning. 2022 [cited 2022 Oct 7]. Available from: https://www.mlyearning.org/

19. Everytown support fund. 2022 [cited 2022 Oct 18]. Available from: https://everytownsupportfund.org

20. Turtle H. Natural language vs. Boolean query evaluation: a comparison of retrieval performance. SIGIR '94. London: Springer; 1994. pp. 212–20. https://doi.org/10.1007/978-1-4471-2099-5_22

21. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13(2):206–19. https://doi.org/10.1197/jamia.M1929

22. Breiman L. Random forests. Mach Learn 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324

23. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97. https://doi.org/10.1007/BF00994018

24. Bojanowski P, Grave E, Joulin A, Mikolov T.. Enriching word vectors with subword information. Trans Assoc Computational Linguistics 2017;5:135–46. https://doi.org/10.1162/tacl_a_00051

25. Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in neural information processing systems 29. Curran Associates, Inc.; 2016 [cited 2020 Apr 26]. p. 3567–75. Available from: https://proceedings.neurips.cc//paper/6523-data-programming-creating-large-training-sets-quickly.pdf

26. Settles B. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 2012;6(1):1–114. https://doi.org/10.1007/978-3-031-01560-1

27. Pelleg D, Moore A. Active learning for anomaly and rare-category detection. In Saul L, Weiss Y, Bottou L, editors. Advances in neural information processing systems 17. Cambridge (MA): MIT Press; 2004 [cited 2022 Oct 6]. Available from: https://proceedings.neurips.cc/paper/2004/hash/8c59fd6fbe0e9793ec2b27971221cace-Abstract.html

28. Hospedales TM, Gong S, Xiang T. Finding rare classes: active learning with generative and discriminative models. IEEE Trans Knowl Data Eng 2013;25(2):374–86. https://doi.org/10.1109/TKDE.2011.231

29. Mullapudi RT, Poms F, Mark WR, Ramanan D, Fatahalian K. Learning rare category classifiers on a tight labeling budget. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada. 2021. p. 8403–12. https://doi.org/10.1109/ICCV48922.2021.00831

30. Mullapudi RT, Poms F, Mark WR, Ramanan D, Fatahalian K. Background splitting: finding rare classes in a sea of background. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA. 2021. p. 8039–48. https://doi.org/10.1109/CVPR46437.2021.00795

31. Attenberg J, Provost F. Why label when you can search? alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10), Association for Computing Machinery, New York, NY, USA. 2010. p. 423–32. https://doi.org/10.1145/1835804.1835859

Appendix

Initial Keyword List:

1. firearm
2. shotgun
3. gun
4. pistol
5. rifle
6. strapped
7. blaster
8. glock
9. revolver
10. gat/gatt
11. shooter
12. leng
13. rod
14. banger
15. packing heat
16. packing a heater
17. burner
18. bomb
19. hammer
20. bullet
21. fire stick
22. cannon
23. roscoe

 https://doi.org/10.3768/rtipress.2023.mr.0050.2304

www.rti.org/rtipress **RTI Press publication MR-0050-2304**