Title: *Evaluating Criminal Justice Reform During COVID-19: The Need for a Novel Sentiment Analysis Package*

Authors: Divya Ramjee[1]*, Louisa H. Smith[2]*, Anhvinh Doanvo[3], Marie-Laure Charpignon[4], Alyssa McNulty[5], Angel N. Desai[6], Maimuna S. Majumder[7]

**These authors contributed equally to this work.*

[1] Department of Justice, Law and Criminology; School of Public Affairs, American University; 4400 Massachusetts Ave NW, Washington, DC 20016 [ORCID: 0000-0001-6159-3952]
[2] Department of Epidemiology; Harvard T.H. Chan School of Public Health, Harvard University; 677 Huntington Avenue, Boston, MA 02115 [ORCID: 0000-0001-9029-4644]
[3] [ORCID: 0000-0003-0924-2287]
[4] Institute for Data, Systems, and Society; Massachusetts Institute of Technology; 77 Massachusetts Avenue, Cambridge, MA 02139 [ORCID: 0000-0002-5786-2627]
[5] Department of Epidemiology and Biostatistics; School of Public Health, Texas A&M University; 212 Adriance Lab Rd., College Station, TX 77843 [ORCID: 0000-0002-4281-8236]
[6] Division of Infectious Disease; University of California Davis Health; 4150 V Street, Suite 310, Sacramento, CA 95817 [ORCID: 0000-0001-8962-9427]
[7] Computational Health Informatics Program; Boston Children's Hospital and Harvard Medical School; 300 Longwood Avenue, Boston, MA, 02115 [ORCID: 0000-0002-3986-4303]

Corresponding Authors:

Divya Ramjee
Department of Justice, Law and Criminology
School of Public Affairs
American University
4400 Massachusetts Ave NW
Washington, DC 20016
dr1208a@american.edu

Maimuna Majumder
Computational Health Informatics Program
Boston Children's Hospital and Harvard Medical School
300 Longwood Avenue
Landmark 5506, Mail Stop BCH3187
Boston, MA 02115
maimuna.majumder@childrens.harvard.edu

Abstract:

Existing natural language processing lexicons that underlie current sentiment analysis (SA) algorithms may not perform adequately in certain academic disciplines depending on contextual complexities. The health and safety of incarcerated persons and correctional personnel have been prominent in news media discourse during the COVID-19 pandemic, potentially highlighting the need for a novel SA lexicon and algorithm that is tailored for the examination of public health policy in the context of the criminal justice system. We utilized a text corpus consisting of news articles at the intersection of COVID-19 and criminal justice to analyze the performance of existing lexicons collected across state-level outlets between January and May 2020. Our results demonstrated that sentence sentiment scores provided by three popular SA packages differ considerably from manually-curated ratings. This dissimilarity was especially pronounced when the text was more polarized, whether negatively or positively. A randomly selected set of 1,000 manually scored sentences, and the corresponding binary document term matrices, were used to train two new sentiment prediction algorithms (i.e., linear regression and random forest regression) to verify the performance of the manually-curated ratings. By better accounting for the unique context in which incarceration-related terminologies are used in news media, both of our proposed models outperformed all existing SA packages considered for comparison. Our findings suggest that there is a need to develop a novel lexicon, and potentially an accompanying algorithm, for analysis of text related to public health within the criminal justice system, as well as criminal justice more broadly.

Keywords: COVID-19, criminal justice, sentiment analysis, text analysis, public health, health, safety, public policy, lexicon, algorithm, NLP

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has cast light on the organizational and structural issues within the United States criminal justice system that adversely impact the health of incarcerated people, as well as correctional workers and staff. Incarceration and detention facilities are disproportionately affected by infectious disease outbreaks,[1] and COVID-19 prompted the U.S. Department of Justice to consider prisoner release and home confinement as mitigation options to control transmission in March 2020.[2] While this policy only applied to facilities under the control of the U.S. Bureau of Prisons, states have made varying decisions regarding prisoner release, perhaps in part due to public opinion and activist movements.[3] News media outlets in particular have served not only to highlight existing public opinion, but also to help shape public perceptions based on their coverage.[4]

To understand public support for and against release of incarcerated individuals, we used existing natural language processing (NLP) lexicons and related algorithms to assess sentiment in news media coverage towards prisoner release and criminal justice reform over the course of the pandemic. NLP tools and techniques provide rapid means for analyzing large amounts of text and are increasingly used in social science and policy contexts.[5] Sentiment analysis (SA) is an NLP subfield that pairs sentiment lexicons, i.e., dictionaries of words and phrases with rated sentiment polarity, with specific algorithms that account for important syntactical and contextual features.[5] Common practical applications of SA span a wide range of fields including economics, marketing, politics, and public health.[5]

To our knowledge, an SA lexicon specific to the field of criminal justice does not exist. This field is unique in that much of the related vocabulary is inherently negative, though the intentions and motivations of the discourse may be positive.[3,4] Thus, we hypothesized that due to dual use[6] (i.e., using a system developed for a purpose separate from the one for which it was designed), existing SA packages (i.e., lexicon-algorithm pairs) would be insufficient for accurately gauging sentiment in news media coverage related to public health crises within the criminal justice system, particularly during the COVID-19 pandemic. To test our hypothesis, we manually rated sentiment scores on a text corpus of news media articles related to COVID-19 and incarceration. Our manually-curated scores were then compared to ratings from existing SA packages for each sentence of the selected sample. Building on a training set consisting of our manual ratings as the reference outcome, we derived two novel algorithms (i.e., a linear regression model and a random forest regression model) to improve on currently available SA tools that are not tailored to text at the intersection of public health and the criminal justice system.

Results

Our experiment and analyses considered the following existing SA packages that are most frequently used in the NLP literature: SocialSent[7], VADER[8], and Stanford CoreNLP[9].

Sentiment Scoring and Lexical Analysis

Overall, sentences that were manually scored to have neutral sentiment were consistently rated as neutral by the above listed SA packages. Figure 1 shows three such sentences (sentences 2-4), with scale-standardized scores (see Methods) that deviated least from our ratings, all neutral in sentiment (i.e., either fully neutral in sentiment or with equal amounts of positive and negative sentiment). However, sentences with more extreme positive or negative sentiment

polarity, as ascertained by our manual ratings, were more often scored differently across SA packages (sentences 1, 5-7 in Figure 1). This division appeared driven by words related to criminal justice and public safety (e.g., "innocent", "violent", "defense", "threat", "vulnerable", "safety", "care", "negative").

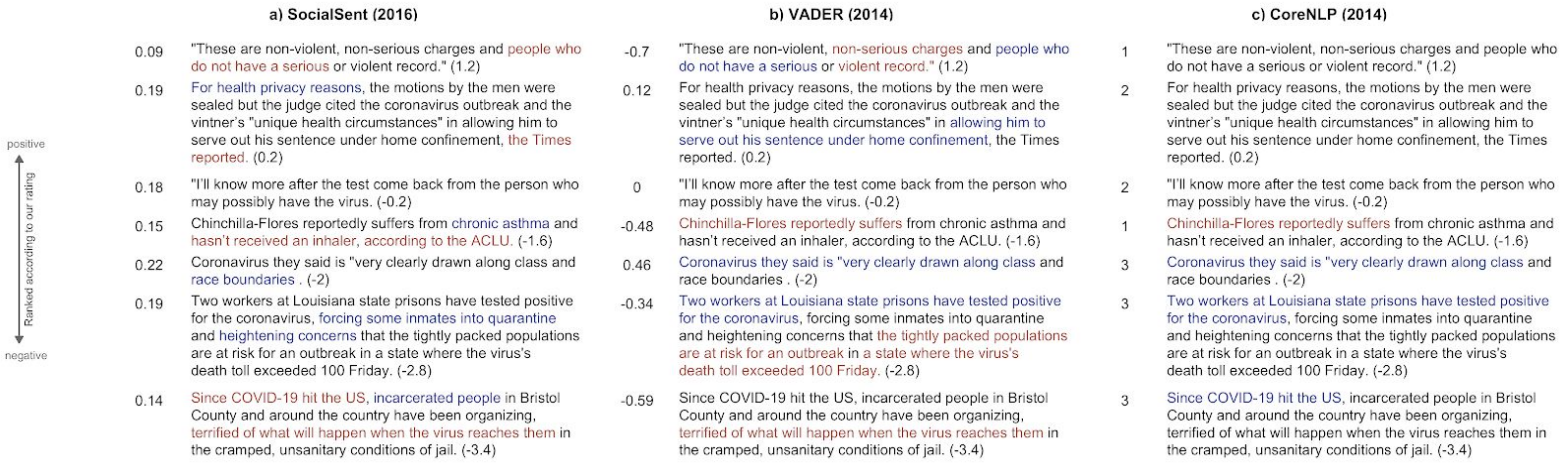


Figure 1. Standardized sentence sentiment scores that deviated least from manually-curated scores, exemplified by three distinct SA packages. SocialSent was considered because it is specifically attuned to social science contexts. VADER was also included for comparison because it is one of the most widely used SA packages. Finally, CoreNLP was used due to its accuracy in sentiment scoring by a recent systematic review of SA in public health. The sentences are arranged from top to bottom in order of most positive sentiment score to most negative, as determined by manually curators. The left, middle, and right panels correspond with the SocialSent, VADER, and Stanford CoreNLP SA packages, respectively. To the left of the sentences are the sentiment values assigned by each SA package on their standardized scale; our ratings follow each sentence in parentheses. Colors indicate relative sentiment associated with the selected portion of the sentence, with red and blue indicating negative and positive sentiment, respectively, as determined by running each algorithm on separate phrases within the sentences.

To investigate, we used the 68 words (i.e., 3.6% of the overall corpus vocabulary) that appeared most often across all sentences (i.e., in at least 10) and compared the average sentence sentiment score for each word. This was conducted using our manually-curated sentence sentiment scores against the popular SA packages of SocialSent, VADER, and Core NLP. Results present similar patterns across SA packages (Figure 2). Our manually-curated SA scores generally associated criminal justice and public safety terminologies (e.g., "safety", "attorney", "community") with positively-scored sentences while the three existing SA packages yielded more neutral or negative sentiment ratings than our scores (e.g., an average score of 0.55 (95% CI -0.06, 1.17) for "community" compared to 0.06 (-0.36, 0.48); -0.02 (-0.62, 0.58); and -0.10 (-0.62, 0.41) from SocialSent, VADER, and Core NLP, respectively) (full results in Table S1 of the Supplement). However, certain criminal justice-related terminology, including "detention", "facility", "sentence", and "justice", appeared in sentences we rated slightly more negatively in sentiment, on average, compared to the existing SA packages (e.g., an average score of -0.18 (-0.49, 0.12) for "facility" compared to 0.10 (-0.19, 0.40); -0.08 (-0.40, 0.24); and -0.04 (-0.32, 0.24) from SocialSent, VADER, and Core NLP, respectively), though this was less consistent

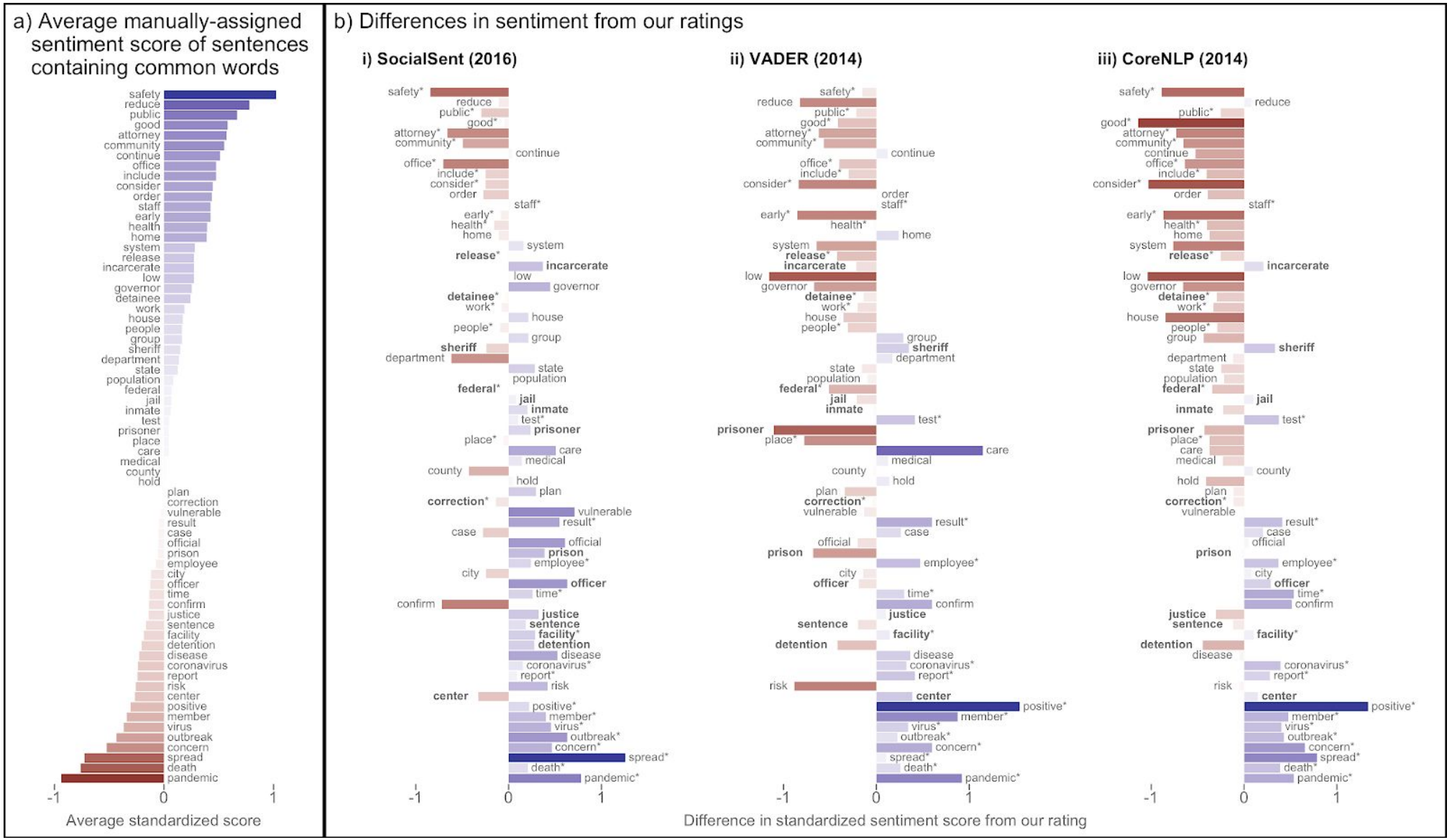


Figure 2. Average standardized sentiment scores for sentences containing each of the 68 most frequently appearing words. The words are arranged from top to bottom in order of most positive (blue) average sentiment to most negative (red), according to our human-curated scores (Panel A). Colors in Panel B indicate the relative score from each of three SA packages compared to our ratings, with red (blue) indicating that the algorithm-rated sentences with a given word more negatively (positively) than our team did. The color gradient indicates the intensity of the average standardized sentiment score for that term across sentences (i.e., light colors refer to terms rated more neutral, while darker colors refer to terms rated more negatively or more positively). The length of each bar depicts the average difference between a given package's scores and ours. Words with direction of sentiment that differ from ours in each package are marked with an asterisk, and words relating to the criminal justice system are in bold.

across packages. Criminal justice-related words associated with neutrally-scored (or equally positive and negative) sentences as determined by both existing SA packages and our manual curation included "jail", "inmate", "prisoner", "medical", and "test". For terminology specific to public health and the pandemic (e.g., "disease", "positive", "virus", "outbreak", "spread", "pandemic"), our manually-curated scores were primarily associated with negatively-scored sentences, with the exception of the word "health", while the three existing SA packages rated these as more positive in sentiment compared to our scores (e.g., an average score of -0.31 (-0.46, -0.15) for "positive" compared to -0.09 (-0.44, 0.27); 1.24 (1.03, 1.45); and 1.03 (0.65, 1.41) from SocialSent, VADER, and Core NLP, respectively).

Proof of Concept Machine Learning Algorithms

To validate the proof of concept derived from our manually-curated sentiment analysis, we developed two machine learning (ML) algorithms – a linear regression model and a random forest regression model – using our sentiment ratings. After standardization of sentiment scores for the three existing SA packages and our two ML models, we trained and tested all algorithms on our text corpus. As is evidenced by the lowest mean absolute difference between our manually-curated scores and predicted sentiment scores (Table 1), both of our models strongly outperformed all three tested SA packages – signifying an important initial step in the development of a new SA package.

Table 1. Comparison of Model Fit Between Existing SA Packages and Our Model

SA Model	**Mean Absolute Difference in Standardized Score Prediction** ***(standard error)***
SocialSent	1.04 *(0.02)*
Stanford CoreNLP	1.03 *(0.02)*
VADER	0.95 *(0.03)*
Trained Linear Regression (binary DTM)	0.82 *(0.03)*
Trained Random Forest Regression (binary DTM)	0.76 *(0.04)*

DTM = Document Term Matrix

Discussion

Our results suggest existing SA packages may be unable to accurately gauge sentiment in the text of news articles at the intersection of public health and criminal justice, especially in the context of the COVID-19 pandemic. VADER, one of the most widely used SA packages, scored many of the most frequently used words (Figure 1 and 2) as negative, despite our identification

of the words as being neutral or positive within their respective sentence contexts. SocialSent performed better, rating these words more positively than VADER. The fact that the SocialSent SA lexicon is specifically tuned to social science contexts might explain this difference. Overall, existing SA packages performed similarly to each other, with an average error of roughly 1.0 for standardized score predictions (Table 1). However, our models' performance demonstrates the limited utility of these packages – not only for analyzing texts that include both public health and criminal justice content, but also for texts related to criminal justice more broadly.

As suggested by our ML algorithms' outperformance of existing SA packages, words used in texts related to public health within the criminal justice system are contextually unique. Our results reinforce the importance of human curation as an initial step towards building a training dataset that serves the development of a new SA lexicon and algorithm, specific to this intersectional subject. These results demonstrate the importance of a new sentiment rating protocol (i.e., lexicon-algorithm pair) with texts specific to criminal justice. Based on our analyses, we plan to expand upon our findings and aim to develop a novel SA lexicon and algorithm (i.e., package) tailored to texts related to public health crises within the criminal justice system, and potentially for the field of criminal justice overall.

The pandemic instigated the U.S. Department of Justice, and specifically the U.S. Bureau of Prisons, to publicly acknowledge health-related shortcomings in the U.S. prison system and address reforming early release and home confinement measures. Public attention and news media coverage has concurrently increased, with particular attention to criminal justice reform initiatives and systemic racial inequities.[10] This preliminary study illustrates that existing SA packages are inadequate for accurate assessment of sentiment in texts regarding such current events. We hope to use our evolving NLP work on various corpora to gauge the scope of public health and reform measures for incarcerated persons, public support for or against criminal justice reform related to public health, and important factors mediating reform policy decisions in response to the pandemic.

Methodology

A recent systematic review[5] of SA in public health identified support vector machines and naïve Bayes classifiers as the most accurate algorithms (~70-80% accuracy) in the field, leading us to consider SA packages Stanford CoreNLP[9] and VADER[8] for our study. We also included SocialSent[7], which uses a novel algorithm to derive content-specific sentiment lexicons for texts related to social science.

Sentiment Scoring

MediaCloud[11], a searchable platform for articles from news outlets around the world, was used to collect articles related to COVID-19 and criminal justice from January 1, 2020 through May 25, 2020 at the state-level in the U.S. (see Supplemental for search query criteria). We subsequently scraped the full text of each available article. To avoid introducing event-specific coverage in our corpora of texts, we selected May 25th – the date of George Floyd's death – as our end date. This particular event spurred an increase in news media coverage pertaining to criminal justice reform across the United States, specifically related to excessive use of force by law enforcement.[12] Additionally, some stories about his death also discussed the topic of COVID-19 transmission during protests. Thus we limited our scope to only news articles published before George Floyd's death, since the coverage of this event could affect our results.

We then used simple random sampling to select 1,000 sentences from our text corpus of 126,552 unique sentences for manually-curated sentiment rating in two phases. Additionally, we validated that the word frequency in this subset and the overall dataset were comparable. The first 500 sentences were scored (negative, neutral, or positive) independently by five members of the research team (DR, AD, AM, MC, TC), which were used as a learning phase for the development of a standardized sentence scoring approach. All curators subsequently convened to reconcile rating discrepancies and ensure that all individuals agreed on how to score each sentence for our experiment. The second 500 sentences were then used for our experimental results and scored by the same five members of the research team on an integer scale from -4 (most negative) to 4 (most positive). This set of sentiment ratings was further averaged across curators to compute the final score for each sentence.

The second set of 500 sentences was additionally scored by SocialSent, VADER, and Stanford CoreNLP. All sentiment scores were then standardized (i.e., mean = 0 and standard deviation = 1 within scores from a given algorithm), and scores were compared between SA packages for each sentence. After lemmatization and removal of stop words, we then summarized sentiment related to the 68 words (3.3%) that appeared in at least 10 sentences by calculating the mean score across those sentences. We additionally assessed a selection of sentences to determine which sentiment scores from existing SA packages either deviated from or were consistent with our ratings (Figure 1). We further isolated most frequently appearing words and compared our sentence sentiment scoring with those from SocialSent, VADER, and Stanford CoreNLP (Figure 2).

Machine Learning Algorithms

We developed a proof of concept using our manually-curated scores for the first 500 sampled sentences. We compared the performance of our algorithms against the performance of SocialSent, VADER, and Stanford CoreNLP, using scores on the second set of 500 sampled sentences. All scoring systems were standardized (i.e., to have mean = 0 and standard deviation = 1) for comparison. We used binary document term matrices (DTMs) from our text corpus (i.e., a value is 1 if a word appears in a sentence; otherwise, the value is 0). 10-fold cross-validation was used to train and test a linear regression model and a random forest regression model on DTMs to predict sentiment scores. We then compared the scores predicted from our models to the scores predicted from each of the SA packages (standardized to the same training data). We computed the mean absolute difference between these predicted scores in the test sets and the manually-curated scores, considered as the reference (i.e., ground truth) scores (Table 1).

Acknowledgments

Thank you to Tori L. Cowger MPH (referenced as “TC” in our article) for assistance with sentiment scoring. Thank you also to Shagun Gupta MASc, for assistance with data analysis on a previous version of this manuscript. This work was supported in part by grant T32HD040128 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH).

References

1. S.A. Kinner *et al*. (2020). "Prisons and custodial settings are part of a comprehensive response to COVID-19." *Lancet Public Health*, 5(4): e188-e189.
2. U.S. Department of Justice, Bureau of Prisons, Federal Bureau of Prisons COVID-19 Action Plan, https://www.bop.gov/resources/news/20200313_covid-19.jsp.
3. J.V. Roberts & M.J. Hough. Understanding Public Attitudes to Criminal Justice. McGraw-Hill: Berkshire (2005).
4. J. T. Pickett. (2019). "Public Opinion and Criminal Justice Policy: Theory and Research." *Annual Review of Criminology,* 2(1): 405-428.
5. A. Zunic *et al*. (2020). "Sentiment Analysis in Health and Well-Being: Systematic Review." *JMIR Medical Informatics*, 8(1): e16023.
6. D. Hovy & S.L. Spruit. (2016). "The social impact of natural language processing." *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 2: 591–598.
7. W.L. Hamilton, K. Clark, J. Leskovec, & D. Jurafsky. (2016). "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora." *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2016, 595–605.
8. C. J. Hutto & E. Gilbert. (2015). "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, ICWSM 2014.*
9. C. Manning *et al*. (2014) "The Stanford CoreNLP natural language processing toolkit." *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2014*, 55–60.
10. C. Flanders & S. Galoob. (2020). "Progressive Prosecution in a Pandemic." *The Journal of Criminal Law and Criminology*, 110(4): 685-706.
11. MediaCloud [Internet]. Cambridge (MA). [cited 2020 Aug 16]. Available from: https://mediacloud.org/
12. K.J. Mullinix, T. Bolsen, & R.J. Norris. (2020). "The feedback effects of controversial police use of force." *Political Behavior*, 1-18.

Supplemental

Data Set

The following search query was used to collect articles for this study: title:((coronavirus OR COVID* OR SARS) AND (jail OR prison* OR incarc* OR detain* OR offend* OR inmate* OR "correctional facility" OR "detention center") NOT Weinstein* NOT Exotic* NOT Tekashi* NOT 6ix9ine* NOT Tiger* NOT Avenatti* NOT "R".

Once aggregated, duplicate titles were removed if they had (1) the same date, (2) same title regardless of punctuation and capital letters, or (3) potentially different news outlets but were from the same state. Only state-level news outlets were included in this study, resulting in 2200 distinct news outlets for inclusion, with an average of 7.3 articles per outlet. Additionally, sentences that shared beginnings and/or ends of articles (often advertising statements) were manually extracted and removed from the body text across all news articles to keep core content only (117 template strings were removed; see 01_download_articles.Rmd code for more details).

For our manual curation methodology, we scored sentence sentiment on an integer scale from -4 (most negative) to 4 (most positive), using 0 for neutral. Our ratings were based on the understanding of the context of the sentences and the emotions evoked.

Robustness Check

For further general comparison of our proposed approach with existing SA packages, we additionally examined the Afinn[1] lexicon, Liu and Hu opinion lexicon[2], and SentiWordNet[3] (results available in our code and data repository). The Afinn lexicon consists of a repository of English terms that were manually rated by Finn Årup Nielsen for valence (i.e., sentiment polarity), using an integer scoring system ranging between -5 (most negative) and +5 (most positive). The lexicon was developed based on the vocabulary used in micro-blogs and social media posts. While Afinn is better suited for analysis of sentiment in short text, we included it as a comparator for its recency, simplicity of use through R and Python wrappers, and its sizable lexicon (3300+ terms).

Additionally, the WordNet database consists of 100,000+ words that occur in varying contexts, and we included two other lexicons derived from the WordNet database for comparison: Liu and Hu opinion lexicon and SentiWordNet. The Liu and Hu opinion lexicon uses synonymy and antonymy relations iteratively, contains about 6800 terms, and was developed in 2004. SentiWordNet, developed in 2006, is a lexical resource for opinion mining that assigns each synset in the WordNet database three sentiment scores: positivity, negativity, objectivity. It was primarily designed for mining the expression of opinions in online forums and product reviews as well as longer pieces of written content, such as news articles. Additionally, it has the advantage of mapping each term with one or several scores, depending on the number of senses.

Our manually-curated models outperformed these SA packages in addition to the three main SA packages.

Table S1. Average score across sentences containing the 68 most-common words in our data.

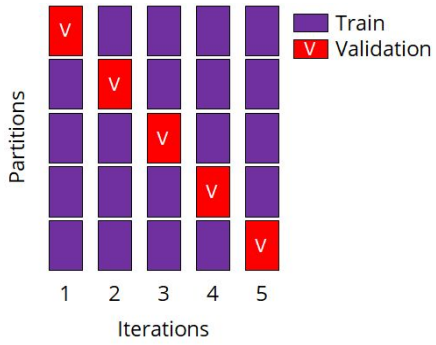
	Average standardized score (95% confidence interval) across sentences containing word			
word	*SocialSent*	*CoreNLP*	*Vader*	*Manual curation*
attorney	-0.09 (-0.63, 0.45)	-0.16 (-0.71, 0.40)	-0.05 (-0.56, 0.45)	0.57 (0.06, 1.08)
care	0.55 (-0.08, 1.17)	-0.33 (-0.78, 0.13)	1.19 (0.61, 1.77)	0.04 (-0.89, 0.97)
case	-0.33 (-0.65, 0.00)	0.15 (-0.19, 0.50)	0.22 (-0.12, 0.55)	-0.05 (-0.28, 0.19)
center	-0.59 (-1.04, -0.14)	-0.11 (-0.54, 0.31)	0.13 (-0.45, 0.70)	-0.26 (-0.67, 0.15)
city	-0.36 (-0.89, 0.17)	-0.04 (-0.50, 0.42)	-0.26 (-0.75, 0.24)	-0.12 (-0.62, 0.38)
community	0.06 (-0.36, 0.48)	-0.10 (-0.62, 0.41)	-0.02 (-0.62, 0.58)	0.55 (-0.06, 1.17)
concern	-0.06 (-0.46, 0.34)	0.14 (-0.44, 0.71)	0.08 (-0.48, 0.64)	-0.52 (-1.19, 0.15)
confirm	-0.85 (-1.43, -0.28)	0.38 (-0.16, 0.91)	0.46 (-0.04, 0.97)	-0.14 (-0.52, 0.25)
consider	0.20 (-0.34, 0.74)	-0.59 (-0.95, -0.22)	-0.39 (-1.01, 0.23)	0.45 (0.06, 0.84)
continue	0.55 (0.03, 1.07)	-0.01 (-0.52, 0.49)	0.63 (0.15, 1.11)	0.51 (-0.17, 1.19)
coronavirus	-0.09 (-0.28, 0.11)	0.15 (-0.06, 0.37)	0.08 (-0.14, 0.30)	-0.24 (-0.43, -0.05)
correction	-0.14 (-0.48, 0.20)	-0.12 (-0.42, 0.19)	-0.04 (-0.37, 0.29)	0.00 (-0.28, 0.28)
county	-0.41 (-0.67, -0.15)	0.11 (-0.19, 0.41)	-0.02 (-0.32, 0.29)	0.02 (-0.26, 0.30)
death	-0.56 (-0.85, -0.27)	-0.37 (-0.66, -0.09)	-0.51 (-0.82, -0.19)	-0.76 (-1.07, -0.45)
department	-0.48 (-0.80, -0.16)	0.02 (-0.36, 0.40)	0.31 (-0.06, 0.68)	0.14 (-0.14, 0.41)
detainee	0.21 (-0.10, 0.53)	-0.05 (-0.70, 0.59)	0.10 (-0.65, 0.86)	0.24 (-0.45, 0.93)
detention	0.08 (-0.30, 0.45)	-0.65 (-0.94, -0.35)	-0.62 (-1.08, -0.16)	-0.20 (-0.81, 0.40)
disease	0.30 (-0.33, 0.94)	-0.27 (-0.92, 0.37)	0.14 (-0.48, 0.75)	-0.23 (-0.93, 0.48)
early	0.34 (-0.04, 0.72)	-0.44 (-1.02, 0.13)	-0.43 (-1.11, 0.26)	0.43 (-0.17, 1.02)
employee	0.16 (-0.28, 0.61)	0.30 (-0.41, 1.00)	0.40 (-0.24, 1.04)	-0.07 (-0.60, 0.45)
facility	0.10 (-0.19, 0.40)	-0.08 (-0.40, 0.24)	-0.04 (-0.32, 0.24)	-0.18 (-0.49, 0.12)
federal	0.06 (-0.51, 0.62)	-0.27 (-0.76, 0.21)	-0.44 (-1.02, 0.14)	0.07 (-0.62, 0.76)
good	0.54 (0.05, 1.03)	-0.56 (-0.96, -0.16)	0.17 (-0.69, 1.02)	0.58 (-0.04, 1.20)
governor	0.70 (-0.06, 1.46)	-0.40 (-1.03, 0.23)	-0.41 (-1.12, 0.29)	0.26 (-0.41, 0.92)
group	0.37 (-0.13, 0.88)	-0.27 (-0.92, 0.37)	0.46 (-0.20, 1.11)	0.16 (-0.69, 1.01)

health	0.24 (-0.08, 0.55)	-0.01 (-0.34, 0.33)	0.37 (0.03, 0.70)	0.39 (0.02, 0.76)
hold	0.05 (-0.45, 0.56)	-0.40 (-0.87, 0.07)	0.15 (-0.58, 0.87)	0.01 (-0.67, 0.69)
home	0.28 (-0.12, 0.69)	0.02 (-0.61, 0.65)	0.63 (0.03, 1.24)	0.39 (-0.19, 0.97)
house	0.38 (-0.12, 0.89)	-0.67 (-1.02, -0.33)	-0.18 (-0.76, 0.39)	0.17 (-0.49, 0.83)
incarcerate	0.64 (-0.04, 1.32)	0.48 (-0.33, 1.29)	0.06 (-0.67, 0.78)	0.27 (-1.04, 1.59)
include	0.23 (-0.16, 0.62)	0.07 (-0.48, 0.63)	0.18 (-0.50, 0.86)	0.48 (0.04, 0.91)
inmate	0.26 (0.09, 0.44)	-0.16 (-0.36, 0.03)	0.03 (-0.16, 0.22)	0.06 (-0.10, 0.22)
jail	0.15 (-0.09, 0.38)	0.16 (-0.08, 0.41)	-0.15 (-0.40, 0.10)	0.07 (-0.19, 0.32)
justice	0.18 (-0.37, 0.74)	-0.44 (-0.92, 0.03)	-0.04 (-0.64, 0.56)	-0.14 (-0.72, 0.44)
low	0.29 (0.07, 0.51)	-0.77 (-1.08, -0.45)	-0.88 (-1.52, -0.24)	0.27 (-0.20, 0.74)
medical	0.16 (-0.38, 0.70)	-0.21 (-0.66, 0.24)	0.14 (-0.35, 0.63)	0.02 (-0.34, 0.38)
member	0.06 (-0.48, 0.59)	0.14 (-0.55, 0.82)	0.53 (-0.03, 1.10)	-0.34 (-0.77, 0.09)
office	-0.23 (-0.62, 0.16)	-0.16 (-0.81, 0.49)	0.08 (-0.41, 0.57)	0.48 (0.07, 0.89)
officer	0.50 (-0.01, 1.02)	0.16 (-0.54, 0.86)	-0.31 (-0.94, 0.31)	-0.12 (-0.78, 0.53)
official	0.56 (0.21, 0.90)	0.00 (-0.49, 0.48)	-0.25 (-0.73, 0.22)	-0.05 (-0.48, 0.38)
order	0.17 (-0.26, 0.59)	0.05 (-0.54, 0.64)	0.46 (-0.11, 1.02)	0.44 (0.04, 0.83)
outbreak	0.19 (-0.06, 0.45)	0.00 (-0.49, 0.48)	-0.21 (-0.70, 0.28)	-0.44 (-1.01, 0.14)
pandemic	-0.16 (-0.83, 0.52)	-0.40 (-0.87, 0.07)	-0.02 (-0.79, 0.76)	-0.94 (-1.86, -0.02)
people	0.07 (-0.28, 0.42)	-0.12 (-0.45, 0.20)	-0.14 (-0.44, 0.15)	0.16 (-0.25, 0.58)
place	-0.01 (-0.84, 0.82)	-0.33 (-0.78, 0.13)	-0.74 (-1.40, -0.07)	0.04 (-0.60, 0.69)
plan	0.29 (-0.20, 0.79)	-0.11 (-0.63, 0.41)	-0.34 (-0.89, 0.21)	0.00 (-0.50, 0.50)
population	0.09 (-0.27, 0.45)	-0.13 (-0.57, 0.31)	-0.01 (-0.33, 0.31)	0.08 (-0.27, 0.44)
positive	-0.09 (-0.44, 0.27)	1.03 (0.65, 1.41)	1.24 (1.03, 1.45)	-0.31 (-0.46, -0.15)
prison	0.34 (0.14, 0.53)	-0.08 (-0.30, 0.14)	-0.73 (-0.90, -0.57)	-0.05 (-0.27, 0.16)
prisoner	0.28 (-0.08, 0.64)	-0.39 (-0.72, -0.05)	-1.06 (-1.38, -0.75)	0.04 (-0.33, 0.41)
public	0.37 (-0.01, 0.76)	0.42 (-0.13, 0.96)	0.45 (0.04, 0.86)	0.67 (0.17, 1.17)
reduce	0.68 (0.11, 1.24)	0.86 (0.13, 1.59)	-0.04 (-0.73, 0.65)	0.78 (0.30, 1.26)
release	0.26 (0.08, 0.43)	0.02 (-0.22, 0.25)	-0.15 (-0.38, 0.08)	0.27 (0.06, 0.49)
report	-0.15 (-0.56, 0.26)	0.04 (-0.39, 0.47)	0.17 (-0.22, 0.57)	-0.24 (-0.57, 0.09)

result	0.50 (0.14, 0.87)	0.37 (-0.43, 1.16)	0.56 (-0.18, 1.29)	-0.04 (-0.71, 0.62)
risk	0.16 (-0.17, 0.50)	-0.31 (-0.92, 0.30)	-1.14 (-1.53, -0.74)	-0.26 (-0.97, 0.46)
safety	0.18 (-0.34, 0.70)	0.14 (-0.50, 0.77)	0.87 (0.40, 1.34)	1.02 (0.38, 1.66)
sentence	0.02 (-0.42, 0.46)	-0.28 (-0.73, 0.16)	-0.36 (-0.84, 0.13)	-0.16 (-0.62, 0.30)
sheriff	-0.09 (-0.70, 0.52)	0.48 (-0.35, 1.32)	0.50 (-0.31, 1.31)	0.15 (-0.38, 0.67)
spread	0.53 (0.04, 1.02)	0.06 (-0.57, 0.69)	-0.62 (-1.36, 0.11)	-0.73 (-1.58, 0.12)
staff	0.45 (0.19, 0.71)	0.44 (0.02, 0.86)	0.44 (0.02, 0.86)	0.43 (0.02, 0.83)
state	0.41 (0.12, 0.69)	-0.12 (-0.43, 0.19)	-0.03 (-0.34, 0.28)	0.13 (-0.18, 0.43)
system	0.44 (0.14, 0.74)	-0.48 (-1.01, 0.05)	-0.36 (-0.96, 0.24)	0.28 (-0.18, 0.75)
test	0.15 (-0.07, 0.37)	0.42 (0.13, 0.72)	0.46 (0.21, 0.72)	0.05 (-0.13, 0.23)
time	0.12 (-0.26, 0.51)	0.41 (-0.15, 0.96)	0.17 (-0.42, 0.75)	-0.13 (-0.78, 0.51)
virus	0.09 (-0.26, 0.44)	0.04 (-0.35, 0.43)	-0.02 (-0.40, 0.35)	-0.37 (-0.84, 0.11)
vulnerable	0.68 (0.15, 1.20)	-0.05 (-0.60, 0.49)	-0.17 (-1.00, 0.66)	-0.03 (-0.75, 0.68)
work	0.11 (-0.28, 0.51)	-0.14 (-0.59, 0.30)	-0.02 (-0.55, 0.52)	0.19 (-0.32, 0.70)

Cross-Validation Technique

The machine learning models we trained were tested via a technique known as cross-validation, which enabled us to obtain a confidence interval for the predictive accuracy of our model (see figure below). In cross-validation, we randomly divided the dataset into *k* partitions (e.g., "five-fold cross-validation" yields five partitions). We then trained the model on *k - 1* partitions, reserving the last partition as the validation dataset from which we made predictions and collected accuracy metrics. We repeated this train-test process *k* times so that every partition of the data serves as a test dataset once. The average of our accuracy metrics suggests how well our model tends to perform, while the standard deviation of these metrics indicates how these metrics might vary due to randomness in unseen data. This is because different parts of the data are used to train and test the model in each iteration.



Code and Data Availability

All analyses were run using R version 4.0 and Python version 3.8. All scripts used for analyses, as well as text data from news media and human-curated scores on the 1000 sampled sentences, are located on the following Github repository:
https://github.com/COVID19-DVRN/crim-sentiment

Supplemental References

1. F. Å. Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." arXiv preprint arXiv:1103.2903 (2011).
2. M. Hu. and B. Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004.
3. S. Baccianella, A. Esuli, and F. Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." Lrec. Vol. 10. No. 2010. 2010.