Accepted for publication by: *Child Abuse & Neglect*

Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning

Brian E. Perron[a]*, Bryan G. Victor[b], Gregory Bushman[a], Andrew Moore[a], Joseph P. Ryan[a], Alex Jiahong Lu[ac], Emily K. Piellusch[a]

[a] Child and Adolescent Data Lab, University of Michigan School of Social Work 1080 S University Ave, Ann Arbor, MI 48109

[b] Indiana University School of Social Work, 902 West New York Street Indianapolis, Indiana 46202

[c] University of Michigan School of Information, 105 S State St, Ann Arbor, MI 48109

* Corresponding author: beperron@umich.edu

This study was supported by a grant from the Casey Family Foundation and the Michigan Department of Health and Human Services. The sponsor was not involved in the collection, analysis or interpretation of data or the writing of the manuscript.

Acknowledgements: We acknowledge the contributions of our research staff who provided considerable support for the study, including Sneha Balachander, Susan Balto, and Pnina Tranen. We would also like to thank Barbara Hiltz and Yunyu Xiao for their helpful feedback on our manuscript.

This is the author's manuscript of the article published in final edited form as:

Perron, B. E., Victor, B. G., Bushman, G., Moore, A., Ryan, J. P., Lu, A. J., & Piellusch, E. K. (2019). Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. Child Abuse & Neglect, 98, 104180. https://doi.org/10.1016/j.chiabu.2019.104180

Abstract

Background: State child welfare agencies collect, store, and manage vast amounts of data. However, they often do not have the right data, or the data is problematic or difficult to inform strategies to improve services and system processes. Considerable resources are required to read and code these text data. Data science and text mining offer potentially efficient and cost-effective strategies for maximizing the value of these data.

Objective: The current study tests the feasibility of using text mining for extracting information from unstructured text to better understand substance-related problems among families investigated for abuse or neglect.

Method: A state child welfare agency provided written summaries from investigations of child abuse and neglect. Expert human reviewers coded 2,956 investigation summaries based on whether the caseworker observed a substance-related problem. These coded documents were used to develop, train, and validate computer models that could perform the coding on an automated basis.

Results: A set of computer models achieved greater than 90% accuracy when judged against expert human reviewers. Fleiss kappa estimates among computer models and expert human reviewers exceeded .80, indicating that expert human reviewer ratings are exchangeable with the computer models.

Conclusion: These results provide compelling evidence that text mining procedures can be a cost-effective and efficient solution for extracting meaningful insights from unstructured text data. Additional research is necessary to understand how to extract the actionable insights from these under-utilized stores of data in child welfare.

Keywords: Text mining, machine learning, data science, text classification, child welfare, substance misuse

Detecting substance-related problems in investigation summaries of child abuse and neglect using text mining and machine learning

State-based child welfare agencies accumulate, manage, and store a vast amount of administrative data for regulatory reporting and coordinating services. Administrative data are considered a critical resource for empirical research, offering essential insights about child maltreatment and the public system of care for this vulnerable population. Child welfare agencies make significant investments in administrative data systems, but the overwhelming majority of investments have focused on structured data fields. *Structured data* are bits of information saved in a pre-defined format. These include dates of service, race, gender, age, type of allegations, place of residence, risk level, caseworker assignment, etc. and can be analyzed and summarized using a variety of statistical procedures. Unstructured data do not have a pre-defined format, such as case notes and written reports, so they cannot be analyzed without manual review and coding.

Even though unstructured data have inherent limitations, they serve a critical function in the provision of child welfare services. More specifically, caseworkers document essential observations about causal and risk factors of maltreatment, services processes, and service outcomes to facilitate case-level planning and decision making. Unstructured data fields also allow for documentation of phenomena that system developers did not anticipate when constructing the system. For example, at the time of writing this report, many states are facing severe problems related to opioid misuse (Vivolo-Kantor et al., 2018). Some states do not collect this information in a structured format, although caseworkers may document these problems and other problems (e.g., mental health problems and domestic violence) in case notes and reports in an unstructured text field.

Arguably, unstructured text data are one of the largest untapped data sources that are being created and managed by child welfare agencies. Qualitative methods have proven to be essential for extracting information and insights from unstructured data (e.g., Cordero, 2004; Epstein, 2011; Henry, Carnochan, & Austin, 2014; Palinkas et al., 2015), yet the nature of qualitative analysis necessarily requires every document to be carefully reviewed, coded, and analyzed. Analysis of system functioning and other regulatory monitoring requires population-level data. We cannot use qualitative methods for the analysis of system functioning and regulatory monitoring or perform qualitative analysis in (or close to) real-time like automated quantitative procedures, so they will always be retrospective descriptions of the system.

Recent advances in data science have offered new ways of working with the increased volume, types, shapes, and amounts of data, including unstructured text data. Data science is an interdisciplinary field that integrates traditional research methods with computer science and domain knowledge to manage and extract insights from various types of structured and unstructured data. This field is gaining recognition in the area of child welfare, given the emergence of for-profit companies selling predictive analytic services. However, we have limited experience in child welfare using data science as a framework for research.

One specific area of data science that has relevance to maximizing the value of unstructured text data in child welfare is *text mining,* which is the central methodological approach to our current study. Text mining is the process of extracting knowledge from unstructured text documents by converting the text into numeric data to perform various forms of quantitative analysis. Text mining uses both traditional statistical procedures (e.g., linear and logistic regression; cluster analysis) and machine learning algorithms. Machine learning algorithms are procedures that enable the computer to figure out or *learn* how a large set of data are functionally related to an outcome. The computer can use this experience to make predictions or perform a task, such as classifying documents. Taken together, text mining and machine learning algorithms offer new ways of analyzing extensive collections of unstructured text documents.

The fields of medicine, business, and law have used text mining to solve a variety of problems (Ekstrom & Lau, 2008; Fattori, Pedrazzi, & Turra, 2003; Krallinger, Erhardt, & Valencia, 2005; Netzer, Feldman, Goldenberg, & Fresko, 2012; Warrer, Hansen, Juhl-Jensen, & Aagaard, 2012; Wyner, Mochales-Palau, Moens, & Milward, 2010), but we have only a few case examples that relate specifically to child welfare. Castillo, Tremblay, and Castellanos (2014) used text mining to identify children in the foster care system who are prescribed psychotropic medication. The text mining procedures relied on the unstructured text contained in caseworker notes. Similarly, Armit and colleagues (Armit, Paauw, Aly, & Lavric, 2017) used text mining to identify cases of child abuse based on medical records. Both studies showed surprisingly high levels of accuracy. Specifically, the model to identify psychotropic medications was 90% accurate, and the model to identify child abuse was 80% accurate in comparison to expert human reviewers. We would also like to note that an information technology text (Castillo et al., 2014) and the journal *Expert Systems with Applications* (Armit et al., 2017) were the source publication of these works, neither of which are familiar to child welfare researchers. At the time of writing this report, neither study had been cited in a child welfare journal or by child welfare researchers, suggesting that the field is mostly unaware of the potential that data science in general, and text mining specifically, has to offer.

Despite the few examples available that are related to child welfare, text mining has the potential for maximizing the value of unstructured text data within existing administrative data systems. Agencies can potentially use computer models to replace or supplement resource-intensive administrative audits. Computer models can integrate with existing data systems for alerting or flagging purposes without expensive adaptations and policy changes. The current study tests the feasibility of using text mining to identify substance-related problems (SRPs) among families that have been investigated for child abuse and neglect using only the unstructured text contained in the investigation summaries written by caseworkers.

Two practical problems motivate this study. First, state child welfare agencies collect and manage vast stores of administrative data that can potentially inform efforts to improves services and system processes for vulnerable children and families. Despite the existing stores of data, system administrators, policy makers and researchers are routinely frustrated by the absence of the right data and problems with the existing data. Data quality is a problem that has consequences beyond just the child welfare system. For example, a recent report of the Administration for Children and Families highlights a federal mandate that strongly encourages data sharing among public systems of care. Data sharing can help ensure child welfare agencies and courts have timely access to the information needed for decision making (see Children's Bureau, 2018). The spirit of this mandate is on maximizing the value of existing data, but this assumes we have the right data and high quality data to share.

In the current study, our partnering agency does not collect reliable systematic data about substance-related problems among families who have been investigated for child abuse or neglect or are currently receiving services. The state uses a risk assessment tool to collect information that includes a single indicator about substance abuse for the caregiver, but recent reports have raised serious concern about the reliability and validity of these estimates. Specifically, an administrative audit revealed that 37% of the cases had improper risk assignments because of either human or software errors (Office of the Auditor General, 2018). Moreover, the risk indicators consider problems among just caregivers when, in fact, substance abuse problems within the family system can create significant obstacles to achieving successful outcomes.

Consequently, researchers and policy-makers are unable to make data-driven decisions or derive insights about socio-demographic, geographic, and temporal trends of SRPs in the current system of care. If we can reliably extract data regarding SRPs, we have the potential of extracting other data that are essential for the information needs of policymakers and researchers. This study, therefore, serves as a proof of concept for using existing data in many other ways that can maximize the value of our existing data stores to address information needs and produce practical insights.

Second, the state agency routinely performs administrative audits of case notes and other written records for quality, administrative or legal purposes. A set of coding procedures with specific keywords often define the specific procedures of the audits. Encoding the procedures into machine language allows for a computer model to perform the task, much more quickly and without fatigue. A computer model could, therefore, be applied to much larger stores of data, thereby offering both potential advantages in quality and quantity of audits. We want to note that the process for developing computer models is nearly identical to that of the administrative audit. The only difference is building and testing computer models using the manually generated codes in the administrative audit. Many agencies are already performing the bulk of the labor that is required for developing and testing the various kinds of computer models. Thus, state agencies have opportunities to make relative low-cost and low-risk investments that have potential for significant returns.

We have two specific aims for this study. Our first aim is to develop a computer model that can accurately identify whether SRPs are present or absent (hereafter defined as SRP+ and SRP-, respectively) using only the written summaries of case investigations. We conduct a series of analyses that formally assess the accuracy and help us determine whether the classifications performed by expert human reviewers (EHRs) are exchangeable with a computer model. Second, we seek to promote integration of the data science framework into child welfare research to help maximize the vast amount of unstructured text data that are otherwise amenable to analysis using only qualitative methods. Toward this end, we release our computer code in an open-source format, which documents all aspects of our data preparation, model development, and analysis. The release of the code in an open-source format is consistent with the broader philosophies in the field of data science (see Gelman & Loken, 2013).

Methods

In this section, we present a description of the model building process, organized into three parts. First, we describe specific theoretical assumptions that were central to our technical decisions regarding data preparation and model building. Second, we provide an overview of the five steps involved in constructing the models. Third, we describe some of the technical features of the models used for case identification. Our methodology is written in a manner to provide the reader with a holistic view of the feasibility study, as opposed to a detailed description of our data preparation, model specification, and performance metrics. These detailed descriptions will be the focus of future reports. However, we direct readers to our freely available, open-source code that contains all the technical details of our work: https://github.com/SSW-DataLab/cps-srp-text-mining-materials. This will allows for a step-by-step review of our procedures, in addition to the ability to test or reuse the code. Due to the highly sensitive and confidential nature of the raw data, we are unable to make these data available.

Theoretical assumptions

An essential aspect of data science is the *integration* of domain knowledge with computer science and traditional research methods. In this project, we used domain knowledge to establish a set of theoretical assumptions that guided all aspects of data preparation, model specification, and assessment of model performance. These theoretical assumptions helped our research team avoid repeated testing of the data at many decisions points.

Our primary theoretical assumption is built around the *data generating mechanisms* -- that is, the various factors and processes that influence the construction of a caseworker's narrative summary. More specifically, we carefully considered the various policies, training, and expectations of caseworkers that provide boundaries around the content and language for writing summaries. This process involved the use of qualitative research methods, including interviews, policy reviews, and observations. Understanding what kind of information could be derived from narrative summaries was our primary goal. We also considered factors that may have introduced systematic error into the data. Our confidence in the use of investigation summaries to identify SRPs was also strengthened by agency policy that specifically instructs caseworkers to record this information:

> [The] relevant facts/evidence pertaining to the allegations obtained during the investigation that resulted in the determination of whether a preponderance of evidence existed. . . . Include documentation, as appropriate, of prevalent and underlying family issues (for example, *substance abuse*, lack of parenting skills, child behavioral issues, violence in the home) and any other issues found during the investigation [italics added for emphasis]. (MDHHS, 2016, pp. 4–5).

Additionally, CPS conceptualizes various forms of drug involvement as threatened harm to the child(rent), which is an actionable form of child maltreatment. Thus, when investigations involving allegations of substance use or exposure, caseworkers are instructed to make investigation decisions on the presence or absence of evidence involving "parental substance use," "substance exposed infants," and "manufacturing, selling or distribution of substances where a child resides." (MDHHS, 2019, p. 8)

From this policy, we assumed that language concerning psychoactive substances or behaviors involving the use of psychoactive substances either indicated or negated an SRP. More specifically, terms and phrases like "getting high," "snort cocaine," and "selling drugs" indicate an SRP, whereas "no evidence of drug use" and "negative drug tests" negate an SRP.

We want to point out that the assumptions of certain terms indicating or negating an SRP are probabilistic rather than deterministic. This means that, over the long run, any given assumption will be effective at classifying documents as SRP+ or SRP- over large number of cases -- but, sometimes we will encounter errors. For example, our assumption that specific words are indicative of SRPs is complicated by words that are spelled the same but have different meanings, or *homographs*. The term “crack” could refer to either cocaine or something that is partially broken, like a window. The reader should also be aware that an accurate model does not require every assumption to be correct every time. Rather, our computer models rely on a large set of assumptions, so any single assumption that failed for a given case can be offset by other assumptions that were correct.

We also make a strong assumption that not mentioning psychoactive substances or behaviors reflected the absence of an SRP. Precisely, caseworkers may have forgotten to record this information or intentionally or unintentionally left out this information. Our models cannot detect any problem that was not explicitly documented by the case workers. From this perspective, the results we obtain are likely to underestimate rather than overestimate the true prevalence. Keep in mind this is not a limitation of the modeling approach we use, as the problem exists with any other type of administrative audit or qualitative study using unstructured text. This is a problem regarding the expectations and policies of writing case notes and reports, which has implications for how we use and make value of these data.

These theoretical assumptions were critical for preparing the data and selecting and specifying the text models. For example, when preparing the data, we used these assumptions to select data cleaning procedures that would isolate key terms from non-informative terms. Thus, constructing a dictionary involved using our theoretical knowledge as a guidepost for decision making. We selected text models that would relate our key terms with the target outcome (SRP+/-). We were not interested in exploring deeper correlational patterns that may be hidden in the natural language.

Our theoretical assumptions were also necessary for helping establish acceptable levels of model accuracy. More specifically, given the various sources of error contained in the raw and prepared data, we set a goal of achieving 80% accuracy to be considered useful for research and administrative purposes. We establish accuracy through formal comparisons with

EHRs. Our target level of accuracy was informed by prior studies that used text mining of medical and child welfare summaries (Castillo et al., 2014; Armit et al., 2017).

Model building process

Five-stages defined our model building process: 1) data identification and acquisition, 2) manual coding, 3) model selection and development, 4) performance assessment, and 5) model deployment. These stages are summarized in Figure 1 and described below.

1. Data identification and acquisition

Identifying the best available data for reliably classifying cases was our first step. The research team worked closely with state representatives (i.e., policy makers, database administrators, and field workers) to identify all possible data fields within the state administrative data system that contains data that can be used to make this inference. Following a review of the record system, state representatives concluded that summary reports from investigations of child abuse and neglect care were the best data source for the current problem. The collective experience of the state representatives, along with agency policy (MDHHS, 2016), informed our decision to use investigations as the data source.

Following IRB approval, the research team obtained all investigation summaries (hereafter referred to as *documents*) of substantiated cases of abuse and neglect from 2015 through 2017 (N = 75,843). The study team required each document to contain at least 50 words to be retained for analysis. This inclusion criterion reduced the total number of documents to 75,809. The mean number of words contained in each document was 472 (standard deviation [SD] = 285.5). All text data in this study were managed and analyzed using the statistical programming language R (version 3.4.4) (R Core Team, 2018).

2. Manual coding

After obtaining the documents, the second step was to manually code a subset of documents that indicated the presence or absence of an SRP, which is the procedure we seek to replace with a computer algorithm. *Labeling* or *tagging* data are also common ways to describe this process in data science literature. The manually coded documents serve two essential functions. First, we use some of these documents to *train* the computer. In this process, the algorithms *learn* a function for relating the text data to the classification outcome. We also use the coded documents to make systematic comparisons with the computer models.

In this study, we considered an SRP to be present if a caseworker documented the use, manufacture, and/or distribution of psychoactive substances in a manner that was considered a causal or risk factor for child abuse or neglect. In our definition, we did not limit the SRP to the caregiver or a perpetrator. Instead, our definition includes the SRP existing anywhere in the family system because of the obstacles they create in achieving important system outcomes. Thus, the coders inferred an SRP based on the particular language contained in the written report. In our definition, we regarded all forms of psychoactive substances as potentially indicative of an SRP, except cannabis, cannabis-derived substances, and infant exposure to cannabis. We recognized the significant variability in reporting practices and the ongoing changes to the state's medical marijuana policies as a significant threat to reliable coding. Moreover, many summaries mentioned cannabis but did not offer sufficient details to make an inference as to whether cannabis use was a causal or risk factor for abuse or neglect. Thus, we did not apply SRP+ codes to investigation summaries that identified cannabis as the only psychoactive substance. Excluding cannabis from our definition has important implications for the interpretation and use of the models, which is given further consideration in the discussion section of this report.

Sampling. Statistical power estimation procedures do not exist for the classification problem that we seek to solve. Thus, determining the number of documents to code manually was informed by existing text mining studies, agency resources, and ongoing qualitative reviews of our data. Ultimately, we wanted to ensure that our training data observe the full range of text patterns in our test data and future data.

The study team randomly sampled 4% of investigations from each of the 83 counties within the state, resulting in 3,094 documents. The financial resources available for the study limited the number of documents that we could manually code. After performing reliability analysis on a test set of cases (N = 91), all the documents were coded and then randomly divided into training and testing sets using a 75-25 split (training data, N = 2,217; testing data, N = 739), excluding 138 documents that were determined to be ambiguous. We selected this split to ensure our training data captures the fullest range of language patterns related to our classification outcome. More specifically, *training* data are used to teach the computer the rules for coding documents. After the computer learns to perform the task using these training data, we test the accuracy using new data – i.e., the *testing* data. These procedures are similar to factor analytic studies that use a split sample, where part of the data is used to discover a factor structure using exploratory factor analysis, and the other part is used to test the fit with confirmatory factor analysis. Doing so helps protect against over-fitting the data, which is given further discussion in the fourth step of the procedures.

Reliability analysis. Four MSW students with work experience in human services related to child welfare and substance abuse served as EHRs in this study. The EHRs were trained to review documents and identify SRP+ cases using the SRP operational definition. Inter-rater reliability was estimated using a kappa (κ) coefficient, which adjusts for chance agreement. In this study, we use the Fleiss' κ which allows for assessing reliability among more than two raters. Landis and Koch (1977) provide thresholds for interpreting kappa values, noting that values .61 - .80 indicate substantial agreement, and values > .81 indicate almost complete agreement. The reliability observed among our EHRs was as κ = .84 based on 91 documents that were independently coded by all four EHRs.

3. Model selection and specification

The next step involved selecting and specifying, also referred to as *training*, our models for the given classification problem. Many different classification models exist, so we used three criteria to guide our selection. First, any given model must have demonstrated success with text-based classification problems in both the academic literature and other professional publications. That is, we wanted a sufficient knowledge base to inform our work and promote its sustainability. Second, we wanted to test different strategies for converting text data into numeric values and classifying these numeric values on the SRP outcome. By doing so, we can build a more comprehensive understanding of how different approaches perform under different conditions. Third, the models must have corresponding open-source software options, as this is intended to promote our aim for promoting the integration of data science in child welfare research.

The process of training machine learning models is similar to traditional statistical modeling. For example, in traditional statistical modeling, we use logistic regression to relate a set of independent variables to a binary outcome. After the logistic regression model is specified, we can use the model to perform classifications with new data with some variant of a predict function in statistical software packages. Our models are constructed in the same manner as traditional usages of logistic regression. However, the significant difference is the conversion of text data to numeric variables, which are then used as variables in the model. In the following section, we describe our selection and specification of five classification models.

4. Assessment of model performance

After specifying the models with the training data, we tested the performance by classifying the remaining documents using the test data. This step allowed us to compare the model’s classification results with the codes assigned by the EHRs. We want to note that the test data were not part of the model construction. These were holdout data, so they are regarded as *new* and *unseen* data to the models, thereby providing a rigorous test of accuracy. We use three conventional estimates of accuracy (global accuracy, sensitivity, and specificity) and inter-rater agreement to assess model performance.

Global accuracy is the percentage of cases in which the model agreed with human coders. Thus, a global accuracy of 80% indicates that 80% of the SRP classification decisions of the computer matched the human rater, and 20% are different. *Sensitivity* is the identification of true positive cases. A model with 80% sensitivity correctly detects 80% of cases with an SRP (true positive), but 20% with an SRP is undetected (false negatives). *Specificity* is the identification of true negative cases. A model with 80% sensitivity correctly identifies 80% of cases without an SRP (true negative), incorrectly identifying 20% with an SRP (false positive).

We extend our assessment of accuracy to include estimates of inter-rater reliability between the EHRs and the computer models. Inter-rater reliability is an unconventional method of model assessment in machine learning studies but relevant to the current project for a couple of reasons. First, measures of accuracy assume that the gold standard for establishing the correct classification decision is the human rater. Although our study team achieved excellent levels of inter-rater reliability when manually coding the documents, we never achieved complete reliability. Therefore, we used an inter-rater reliability estimate to show the extent to which the computer models and human coders agree, without attributing error to either the computer model or human.

Our second reason for computing an inter-rater reliability estimate is to help address a principal aim of the study -- that is, to determine whether EHRs are exchangeable with computer models for this specific problem. We assess inter-rater reliability EHRs and the computer models the same way we assess inter-rater reliability among just EHRs. We regard the computer models as exchangeable if the inter-rater reliability among computer models and EHRs is roughly equivalent to the inter-rater reliability among just EHRs. We do this by exchanging ratings of a single coder with the computer model to see whether the inter-rater reliability decreases. We use the same 91 documents used to test the inter-rater reliability among the EHRs.

5. Model deployment

The final stage of this study involves selecting the model with the best performance for a given problem and classifying the remaining documents in the collection. We use the computer model to code the remaining documents in the collection (N = 73,454) and report the statewide prevalence of SRPs. To assess the stability of results, we make systematic comparisons of estimates derived from the training data and human coded data. We do not discuss the full deployment of the model within the administrative service setting, as the implementation and usage are not central to the current report. However, we retain this as a stage to help ensure the reader has a holistic view of the process.

[INSERT FIGURE 1 ABOUT HERE]

Overview of data inputs and text models

In the current report, we build a set of text mining models that rely on two unique data preparation strategies for converting text data into numeric values. Then, we test three different algorithms for establishing the functional relationship between the numeric values and the classification variable (SRP+/-).

Converting text to numeric values

In this study, we use two different strategies that convert the unstructured text data to numeric values, both of which simplify the natural human language -- a dictionary approach and a term-frequency approach. Data science literature often refers to these approaches as a *bag-*

of-words. The placement, order, and grammar of each term or combination of key terms are ignored, as though they are all held in a giant bag.

The dictionary approach involves using domain knowledge to identify terms that are likely to increase or decrease the probability of an SRP being observed by the caseworker. We grounded the dictionary approach in explicit theoretical assumptions of how case summaries are written. The development of the dictionary was iterative, including content reviews of written case summaries, reviews of agency policies and procedures, interviews with state representatives, and the domain expertise of the team members. Two different lists of positive and negative keywords comprise the keyword dictionary. Positive keywords are those that increase the likelihood of an SRP being present -- e.g., “cocaine,” “IV drug use,” “getting high,” and “drug dealing.” Negative keywords, or *negations*, decrease the likelihood of an SRP. In this case, we are trying to negate the SRP when certain substance-related language is observed. Examples of negating words include, “maintained sobriety,” “negative drug test,” and “no drug use.” The dictionary was used to identify the specific language in each document and construct counts. The different algorithms used these counts to classify each document as SRP+ or SRP-.

Our second approach to converting text to numeric values is called the *term-frequency* approach. The theoretical assumptions of the dictionary approach also guided this approach. However, the term-frequency approach has a crucial difference, insofar that it does not make an a priori specification of terms regarding their relevance or direction of association with the classification outcome. Instead, we computed the frequencies of all terms and relied on the algorithm to determine the strength and direction of association with the outcome.

The term-frequency approach starts by computing the frequencies of all the terms contained in the full collection of documents. Terms can be single terms or a contiguous set of terms referred to as *n-grams.* For example, “He tested positive for heroin” is comprised of five individual terms, referred to as *unigrams* (“He,” “tested,” “positive,” “for,” “heroin”); four *bigrams*

("He tested," "tested positive," "positive for," "for heroin"); and three *trigrams* ("He tested positive," "tested positive for," "positive for heroin").

The frequencies of n-grams are summarized in a term-frequency inverse document frequency (TF-IDF) matrix. These terms are then analyzed using a machine learning algorithm, which relates them to the target outcome (SRP +/-) (see Ignatow & Mihalcea, 2016). We created three different TF-IDF matrices using different combinations of n-grams: 1) unigrams, 2) unigrams + bigrams, and 3) unigrams + bigrams + trigrams.

Classification algorithms

In this study, we use three different classification algorithms -- a simple rule-based model, logistic regression, and random forest.

Rule-based. We specify a rule-based model (hereafter referred to as the *baseline model*) as a comparison for the machine learning algorithms. This baseline model is the only model in which the computer does not *learn* the functional relationship between the text and classification outcome. Instead, the functional relationship is computed directly based on the following scoring methodology. The dictionary is used to identify and count the number of occurrences of positive and negative terms in each document. We subtract the negative term count from the positive term count. Documents with scores > 0 are identified as SRP+, and documents with scores ≤0 are identified as SRP-. For example, assume a document with 100 total terms -- and, among those terms, five are positive, and three are negative. The final score for the document would be 5 - 3 = 2 and, therefore, SRP+.

Logistic regression. Our second model is logistic regression. Logistic regression is one of many different types of machine learning algorithms, even though the procedure for specifying the model is the same as its usage in traditional statistical research. However, the primary difference is the use of text data that have been converted to numeric values. For this model, we use the dictionary for converting text to numeric data by identifying and counting the occurrences of keywords in each document. We enter counts of keywords as independent variables (or features) in the logistic regression model. The computer then estimates the functional relationship between each variable and outcome in the same manner as traditional statistical regression modeling. After training the model, we classified documents based on a predicted probability for the model. A predicted probability > .5 indicated the document was SRP+, and < .5 indicated that the document was SRP-. We computed predicted probabilities for all documents in the test set and assigned SRP+ to cases with a predicted probability > .5, and SRP- to case with a predicted probability ≤ .5.

Random forest. Our third algorithm was a random forest, which is a standard text mining algorithm. The random forest uses the TF-IDF as input data, building and combining many different classification trees. We refer to the reader to the R packages cited in our documentation for further details on the mathematics and mechanics of the random forest algorithm. With this approach, we do not inform the model whether the keywords are positively or negatively associated with the classification outcome. Instead, the random forest algorithm *learns* this functional relationship. We specify three different random forest models using three different data inputs: 1) unigrams, 2) unigrams and bigrams, and 3) unigrams, bigrams, and trigrams. After training the model, we classified documents in the test set and compared the results with the codes assigned by the EHRs.

Results

In this section, we first report on the model accuracy, inter-rater reliability, and state-wide prevalence estimates of SRPs.

Model accuracy

Figure 2 summarizes all three measures of model accuracy. We compare the results from computer models with the codes assigned by EHRs. The EHR codes are necessarily assumed to be *correct*.

Estimates of *global accuracy* ranged from 78.3% to 93.5%. The lowest performing model was the baseline model with the simple rule-based scoring algorithm. All three random forest models ranged from 92.8% to 93.5%, which slightly outperformed the logistic regression model (90.1%). The estimates from the different machine learning models were nearly identical.

Estimates of sensitivity, or the true positive rate, ranged from 78.5% to 96.8%. With this measure, the baseline model exhibited the best performance, and the logistic regression had the lowest performance. Similar to estimates of global accuracy, the random forest models were nearly identical on measures of sensitivity, ranging from 83.6% to 85.4%.

Estimates of specificity, or the true negative rate, ranged from 70.6% to 97.3%. The best performing models were the random forest models, with estimates ranging from 96.7% to 97.3%. The performance of the logistic regression was close, with an estimate of 95.7%. The baseline model had the lowest specificity estimate (70.6%).

Inter-rater reliability

Inter-rater reliability estimates were obtained in the early stage of the research project to ensure consistency across the codes assigned by EHRs. These early-stage estimates revealed a very high degree of inter-rater reliability (κ = .84). A key question of this study was to determine whether EHRs can be replaced by a computer model for this specific classification task, without compromising reliability. We examined this question by having the computer models classify all the documents that we used to examine inter-rater reliability among the EHRs (N = 91 documents). We then used the model results as though an EHR made them. We computed four different inter-rater reliability estimates for each model. These estimates were made by replacing each EHR with the computer model. Figure 3 presents the range of estimates for each model, with a reference line showing the inter-rater reliability among just the EHRs.

Based on this analysis, the random forest models with unigrams and unigrams plus bigrams that replaced each one of the EHRs exhibited reliability estimates > .80. Moreover, a few of the permutations between subsets of the EHRs and computer models exceeded the reliability of EHRs without the computer models. The random forest using unigrams, bigrams, and trigrams exhibited a lower range of estimates compared to the other random forest models. The logistic regression model reliability estimates ranged from .77 to .83, which was considerably better than the baseline model with estimates with a range of .58 to .64.

[Insert Figure 3 about here]

Statewide prevalence estimates of substance-related problems

The next phase of the study was to deploy the models on the remaining documents in the collection (N = 73,454) to obtain prevalence estimates of SRPs. For comparison, we estimated the prevalence using all the documents coded by the EHRs (N = 2,956). An accurate model would produce roughly the same results in comparison to EHRs. As displayed in Figure 4, the models that used a machine learning algorithm (i.e., logistic regression or random forest) were very close to the overall prevalence estimate of the EHRs, whereas the deterministic model overestimated the prevalence by roughly 50%.

We disaggregated the overall prevalence and examined prevalence across the three different study years. By doing so, we observe that the prevalence estimates derived from the machine learning models (i.e., random forest and logistic regression) are nearly identical to the EHRs' estimates for the year 2015 and 2016, except for a 4% increase in prevalence. The models underestimate the increase from 2016 to 2017. This inconsistency is given further consideration in the discussion section.

[INSERT FIGURE 4 ABOUT HERE]

Discussion

Child welfare agencies collect, store and manage a vast amount of unstructured text data that is rarely used. The purpose of our study was to maximize the value of these existing data. As a proof of concept, we sought to develop a computer model that could accurately identify SRPs among families investigated for abuse and neglect. We used only the unstructured text data from investigation summaries to detect SRPs. We developed and tested a suite of computer models, and the results were compared with coding performed by EHRs. This process resulted in compelling evidence that text mining models can achieve a high degree of accuracy for the given problem, producing results that are exchangeable with EHRs. In sum, the strategies and tools that comprise the data science framework show considerable potential for helping maximize the value of unstructured text data collected and managed by child welfare agencies. In this section, we first describe the appropriate interpretations of the data and ethical use of computer models. Then, we talk in further detail about both the performance of the model, with an emphasis on potential leverage points for improvement, and possible sources of error. We highlight some important strengths and limitations of this research, along with implications for research and practice.

Model interpretation and use

The computer models developed in this study offer a cost-effective way of addressing the critical information needs of policymakers and researchers. Data derived from the models can be easily incorporated back into the central data system, allowing for the analysis of socio-demographic, geographic, and temporal trend analyses related to SRPs. However, like all uses of data, theory and evidence must guide the inferences we make from the results.

In this study, we defined SRPs existing within the family unit. Thus, we are unable to make certain inferences; inferences about an individual, such as a caregiver, parent or perpetrator, cannot be made. Additionally, our definition of SRPs focused on the use of substances that present as a causal or risk factor for child abuse or neglect, as opposed to a substance use disorder defined by the Diagnostic and Statistical Manual (DSM; American Psychiatric Association, 2013). We acknowledge that our models are likely to underestimate the *true* prevalence of SRPs. More specifically, some caseworkers may have observed an SRP but did not record this information. Our computer models cannot reliably detect SRPs that caseworkers failed to report. We assume that the absence of evidence of SRPs in a given case summary is evidence of absence, which may not hold across all cases.

In this study, we did not distinguish between different types of substances associated with a given SRP, so we cannot make any inferences about the geographic and temporal dynamics of different substances, such as opioids. In the pilot phase of the project, we coded for different types of SRPs, but we observed variability in reporting practices across caseworkers that limited our opportunity to produce reliable estimates for specific types of substances. Important information and insights can be reliably extracted from unstructured text, but this undoubtedly requires improvements to the written text rather than through coding more data or model specification.

Comments on model performance

This study examined five different text models, all of which exhibited varying degrees of performance. The varying performance naturally gives rise to the question, *Which model is the best?* It depends. In this study, our goal was to establish the feasibility of creating a model that would produce results that are exchangeable with EHRs. From this perspective, the term-frequency method for converting unstructured text to numeric data, along with the random forest algorithms produced the highest levels of global accuracy and specificity, along with the best inter-rater reliability estimates. We used three different combinations of n-grams for the models. All of these produced approximately the same results.

From these results, we consider the random forest model the *best* performing model among the comparisons. The random forest model used three different combinations of n-grams. We do not consider these differences to have any practical interpretations. Thus, we consider unigrams by themselves superior to unigrams with bigrams and trigrams because it is the easiest way to prepare the data for analysis of n-grams. Even though we identify a single model as the best performing, this by no means sets a precedent for future research. Expanding our range of insights that can be extracted from different sources of unstructured text will require significant effort. Different types of problems and data types require different methodologies, so we caution against the use of formulaic thinking.

Even though the models based on the dictionary of keywords did not perform as well as the term-frequency approach, we think these models have considerable value for other types of problems. For example, our baseline model with a simple rule-based score had the lowest global accuracy, specificity, and inter-rater reliability among all the models. However, this model was the most sensitive model, showing the highest rate of detecting true SRP+ cases in comparison to EHRs. This model could be used as part of a hybrid computer-EHR administrative audit if the goal was to identify as many true positive cases among a large sample or population of cases. EHRs can then do a more focused review of the positively identified cases. Statistically rare phenomena are well suited for this type of approach, especially if the range of professional terminology is also narrow and well defined (e.g., suicide, firearms, schizophrenia). Term-frequency models require a vast number of documents to be coded to ensure the full range of text patterns are represented in the training data. Thus, a dictionary approach would be far more efficient and likely produce superior results.

Comments on model (in)accuracy

We want to draw attention to a specific model estimate that deviated from the expected results. More specifically, the statewide prevalence estimates produced by the random forest model with unigrams exhibited almost perfect correspondence with the estimates based on EHR data for the years of 2015 and 2016. However, in 2017, the computer model deviated from the EHR estimates by roughly 4.5 percentage points.

Numerous possibilities exist to account for this deviation, including both random and systematic error. Random error refers to natural fluctuations in the data that are unpredictable and beyond the control of our procedures. From this perspective, we may be observing natural variations that would average out over time through repeated sampling. Systematic error can be caused by many factors that predictably affect inaccuracies or changes but were not directly observed or accounted for in the procedures. Unlike random error, systematic error does not average out with repeated sampling and, therefore, cannot be addressed using statistical procedures.

We discovered a policy change in 2017 that directly influenced how investigations that observed some form of cannabis use (e.g., cannabis-exposed infant) were conducted and documented. This change could have introduced systematic error into our data. More specifically, policy shifted from automatically confirming allegations of abuse when an infant tested positive for exposure to cannabis, to confirming abuse based on all facts and evidence, including the impact of substance use on parenting safety. To meet the expectations associated with the change in policy, caseworkers need to conduct more thorough assessments and provide further documentation regarding substance use. The process of conducting more thorough assessments can lead to increased detection of SRPs, which may explain the increased prevalence from 2016 to 2017. And while the policy change may explain the increase in SRPs, *why* the models did not pick up this change is a point of interest.

One possibility is that the 2017 estimates produced by the model were biased or lacked precision because we did not supply the models with a sufficient amount of training data to reflect differences in text patterns for this new policy period. To explain another way, the key features and terminology used for describing SRPs in 2015 and 2016 may have changed with this new policy, and new key features and terminology were under-represented in the 2017 data. Untangling this issue will necessarily require additional coding and analysis.

Strengths and limitations

This study makes essential contributions to child welfare research and administrative practices by demonstrating the feasibility of using text mining approaches to effectively and efficiently extract information from unstructured text data. We examined the performance of our models using a variety of metrics that help build confidence in our approach. Our methods involved freely available open source software and common text mining procedures that have demonstrated utility in a variety of fields. Open source software helps can reduce reliance on proprietary software and promote sustainability of these procedures. We also provide our computer code that was used for processing data and specifying models, which is a crucial step for ensuring the transparency of methods and promoting further integration of data science methods into child welfare research.

Although our study has notable strengths, the results of our study need to be considered in the context of the study limitations. In our work, we formally examined and reported on two different machine learning classification algorithms. We also informally tested two other algorithms called support vector machines and naive Bayes. However, these models exhibited poor performance with preliminary testing, and an additional investment of resources could not be justified to develop these models further. Future text mining initiatives should consider these as possible alternatives or points for comparisons.

We also want to acknowledge that the number of documents that were manually coded was determined almost exclusively by available financial resources. Currently, no statistical power testing is available to establish a specific number of documents to be coded. We established confidence that our amount was sufficient for the given classification problem, as the coders reported experiencing *saturation.* Saturation is a subjective judgment in qualitative research in which further data collection or coding appears unnecessary after observing what appears to the full range of text patterns for the given problem appear (Saunders et al., 2018). Coding additional documents could have further built our confidence in the precision of our estimates and allowed for additional testing of various data pre-processing steps. At the same time, the limited resources of this study can also be regarded as a strength, since integrating new methods into organizational settings, especially within the public system of care, will always be constrained by resources.

Finally, we want to recognize that our current approach to reporting our work has limitations. More specifically, we chose to report on the overall process of text mining, as opposed to providing comprehensive and detailed performance tests of different models. Consequently, we do not have the space to provide a nuanced description of the steps for processing data and specifying models. Moreover, looking more specifically at the features of the model is a necessary step for understanding what features the models used for making classification decisions. A careful analysis of the most important features can help develop our theoretical understanding of the models and the specific problems we are seeking to solve. Researchers can write future reports that mimic reporting practices of exploratory and confirmatory factor analyses.

Implications for administrative practice, research, and education

This study offers initial evidence related to the potential value of data science and text mining in child welfare research. In this section, we use these findings to suggest a range of implications for administrative practice, research and education.

Inventory of unstructured text data. First, we encourage further exploration of the existing unstructured data stored and maintained in administrative data systems. Carefully reviewing the data is necessary for determining the kinds of information we can extract from the data and the possible sources of systematic error. This work necessarily requires the use of qualitative methodologies to fully understand the range of factors that influence the generation of unstructured text among caseworkers and other personnel who contribute to the data system. From this perspective, data science initiatives in child welfare settings may be best regarded as a form of mixed methods research, even though the technical and quantitative aspects of data science dominate the literature.

Improving data quality. While the technical aspects of data preparation and model specification will have considerable influence on model performance, we want to emphasize that the most crucial leverage point for improving performance is by improving the quality of data. In the current study, we found considerable variability in the writing styles caseworkers used when writing investigation summaries. Variability in writing includes variability in the language used to describe SRPs, the overall length of summaries, the number of errors, and the use of colloquial versus technical language. To date, caseworkers have been using unstructured text data for documenting case-level details and decision making, so the variability has not been regarded as a problem. We are unaware of any studies that have examined the variability of reporting practices across case workers within this state. Despite the variability we observed, our text mining models appear to be robust to such problems.

Improving our capacity to extract useful information reliably requires improving the structure of documentation and standardizing the professional language. By doing so, we can develop models more efficiently and achieve higher levels of accuracy. These policy and procedural changes are a necessary consideration for system improvement when specific data is needed to understand and respond to emerging problems or phenomena within the field, especially if system administrators are unable (or unwilling to) integrate new fields to collect such data in a structured manner.

Cost analysis. Bringing new methodologies into child welfare research and administrative practice requires an understanding of the costs for carrying out a project and the potential returns on investment. Failing to attend to resource constraints will impede efforts to integrate and sustain new methodologies. Text mining has the potential to replace traditional administrative audits that require a team of EHRs to code a large sample of documents. As demonstrated in the current study, text mining requires the same coding procedures as a traditional administrative audit. However, the coding for traditional administrative audits is performed on an *ad hoc* basis, meaning that the coding is performed strictly for the task of generating estimates. The coded documents are rarely, if ever, used for other purposes. A text mining approach also requires a large set of documents to be coded, but these documents can be used to *train* a computer model, like we demonstrated with this study. The additional resources needed to train and validate the computer model is offset by creating an automated tool and the opportunity to vastly expand the number of documents to be analyzed. More specifically, in this study, EHRs coded a total of 3,094 documents from a sample of 75,843. This means that the we have 72,715 uncoded documents remaining in the collection. If we wanted to classify the remaining documents manually, we would need to budget between 3,500 to 7,000 hours of time for EHRs to perform this task, assuming each document takes three to six minutes to read and code. However, our trained and validated models took roughly 40 minutes to classify the remaining documents -- approximately 1,750 documents per minute.

Supplementing or changing scientific training. As pointed out by numerous data scientists, the majority of projects that are grounded in a data science framework devote nearly 80% of time and resources to preparing and managing data, which is roughly consistent with our experience (Zhang, Zhang, & Yang, 2003). However, the skills necessary for preparing and managing administrative data are not a core part of the training that child welfare scholars receive, especially those who are graduates of schools of social work. The core training curriculum in the applied social sciences needs to consider the variety of skills and knowledge needed to maximize the value of different types of data. Supplementing the existing curriculum with content areas that overlap with computer science gives doctoral students increased opportunities for selecting methodologies that can harness and maximize the value of existing data.

Conclusion

Making informed policy decisions requires the right data and high quality data. Child welfare agencies spend enormous resources collecting and managing data that often fail to meet such standards. Our study was motivated by a low-cost, low-risk opportunity to maximize the value of existing data by extracting important information from unstructured text data to help address the information needs of a state child welfare agency. Our work shows how data mining specifically and data science more broadly can be effective and efficient methods for solving certain types of data problems. We think this work provides a compelling case example for integrating innovative strategies and tools into child welfare research and practice.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). https://doi.org/10.1176/appi.books.9780890425596

Armit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*(1), 402–418. https://doi.org/10.1016/j.eswa.2017.06.035

Castillo, A., Tremblay, M. C., & Castellanos, A. (2014). Improving Case Management via Statistical Text Mining in a Foster Care Organization. In D. VanderMeer, M. Rothenberger, A. Gupta, & V. Yoon (Eds.), *Advancing the Impact of Design Science: Moving from Theory to Practice* (pp. 312–320). Miami, FL, USA: Springer International Publishing.

Children's Bureau. (2018). *Data Sharing for Courts and Child Welfare Agencies*. U.S. Department of Health and Human Services, Administration for Children & Families.

Cordero, A. E. (2004). When Family Reunification Works: Data-Mining Foster Care Records. *Families in Society; Milwaukee*, *85*(4), 571–580. http://dx.doi.org/10.1606/1044-3894.1840

Ekstrom, J. A., & Lau, G. T. (2008, May). *Exploratory text mining of ocean law to measure overlapping agency and jurisdictional authority*. 53–62. Digital Government Society of North America.

Epstein, I. (2011). Reconciling Evidence-based Practice, Evidence-informed Practice, and Practice-based Research: The Role of Clinical Data-Mining. *Social Work; Oxford*, *56*(3), 284–288.

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, *25*(4), 335–342. https://doi.org/10.1016/S0172-2190(03)00113-3

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Unpublished Manuscript*, 17.

Henry, C., Carnochan, S., & Austin, M. J. (2014). Using Qualitative Data-Mining for Practice Research in Child Welfare. *Child Welfare; Arlington*, *93*(6), 7–26.

Ignatow, G., & Mihalcea, R. (2016). *Text Mining: A Guidebook for the Social Sciences*. SAGE Publications.

Krallinger, M., Erhardt, R., & Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, *10*(6), 439–445. https://doi.org/10.1016/S1359-6446(05)03376-3

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

MDHHS. (2016). Child Protective Services Investigation Report. In *PSB 2018-004*. *Child Protective Services Manual*. Lansing, MI.

MDHHS. (2019). Complaints involving substances. In PSB 2019-001. Child Protective Services Manual. Lansing, MI

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, *31*(3), 521–543.

Office of the Auditor General. (2018). *Performance Audit Report*. Children’s Protective Services Investigations, Michigan Department of Health and Human Services. Retrieved online April 24, 2018 from https://audgen.michigan.gov/wp-content/uploads/2018/09/r431128516-0011.pdf.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., … Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, *52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

Vivolo-Kantor, A. M., Seth, P., Gladden, R. M., Mattson, C. L., Baldwin, G. T., Kite-Powell, A., & Coletta, M. A. (2018). Vital Signs: Trends in Emergency Department Visits for Suspected Opioid Overdoses — United States, July 2016–September 2017. *Morbidity and Mortality Weekly Report*, *67*(9), 279–285. https://doi.org/10.15585/mmwr.mm6709e1

Warrer, P., Hansen, E. H., Juhl-Jensen, L., & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *British Journal of Clinical Pharmacology*, *73*(5), 674–684. https://doi.org/10.1111/j.1365-2125.2011.04153.x

Wyner, A., Mochales-Palau, R., Moens, M.-F., & Milward, D. (2010). Approaches to Text Mining Arguments from Legal Cases. In E. Francesconi, S. Montemagni, W. Peters, & D. Tiscornia (Eds.), *Semantic Processing of Legal Texts* (Vol. 6036, pp. 60–79). https://doi.org/10.1007/978-3-642-12837-0_4

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381.

Figure 1. Summary of study workflow for the development of computer models to classifying unstructured text data from written investigation summaries of child abuse and neglect

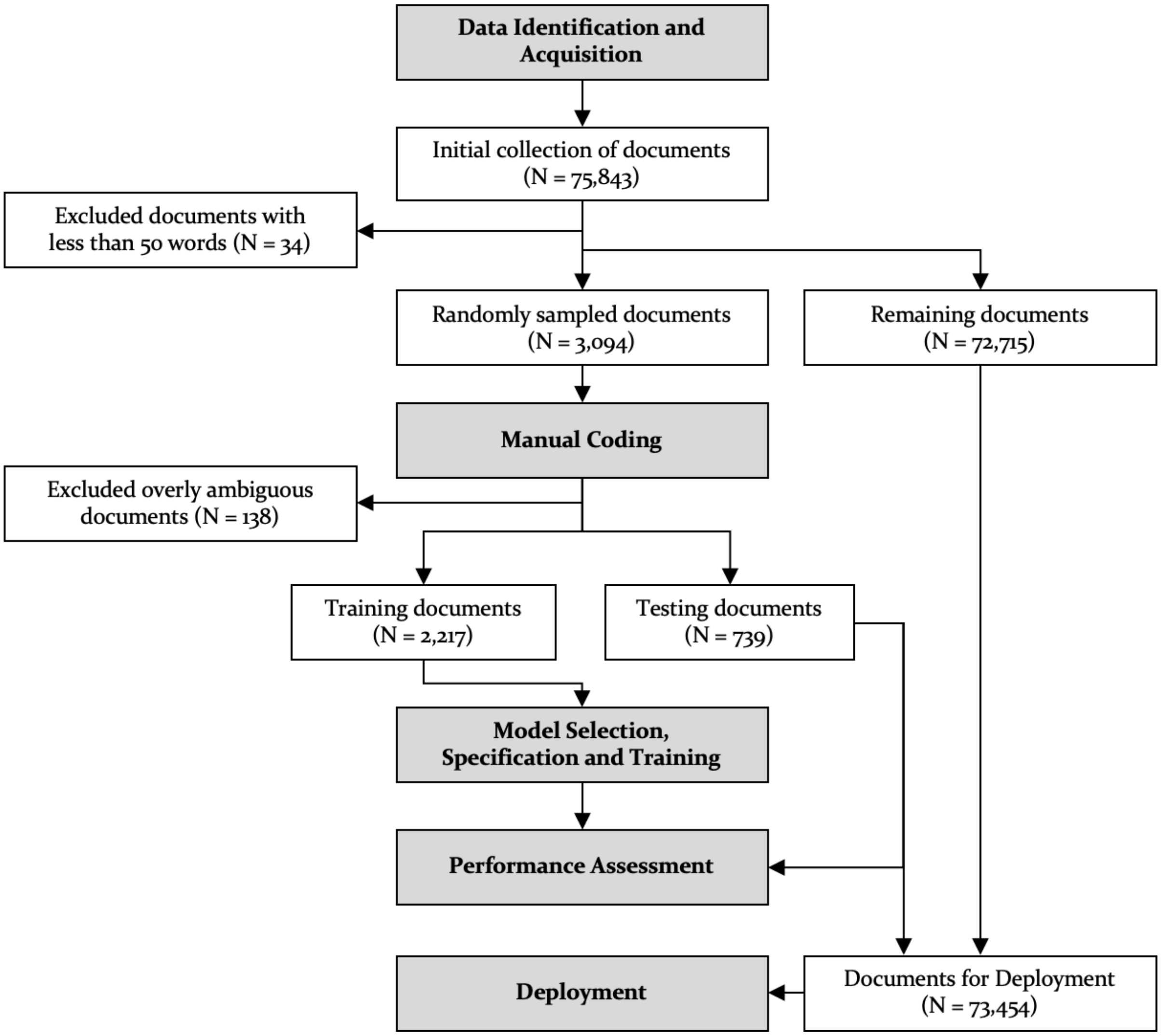
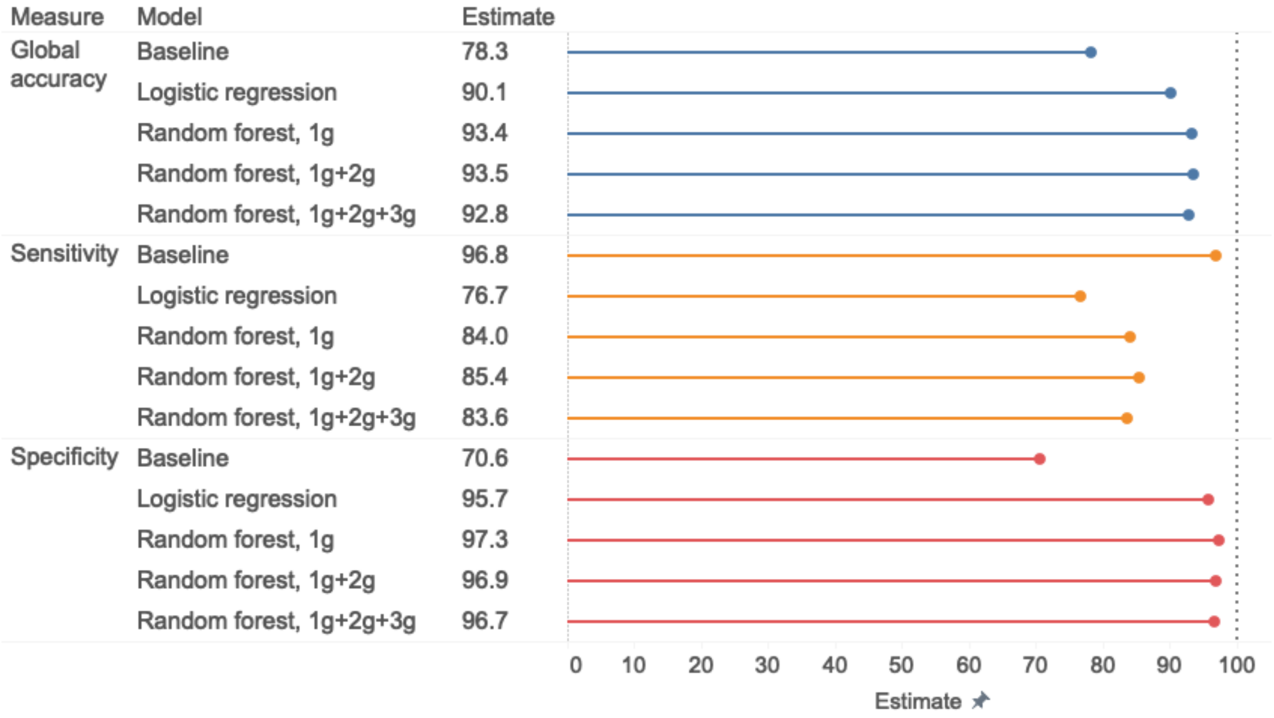

Data Identification and Acquisition
Initial collection of documents (N = 75,843)
Excluded documents with less than 50 words (N = 34)
Randomly sampled documents (N = 3,094)
Remaining documents (N = 72,715)
Manual Coding
Excluded overly ambiguous documents (N = 138)
Training documents (N = 2,217)
Testing documents (N = 739)
Model Selection, Specification and Training
Performance Assessment
Deployment
Documents for Deployment (N = 73,454)

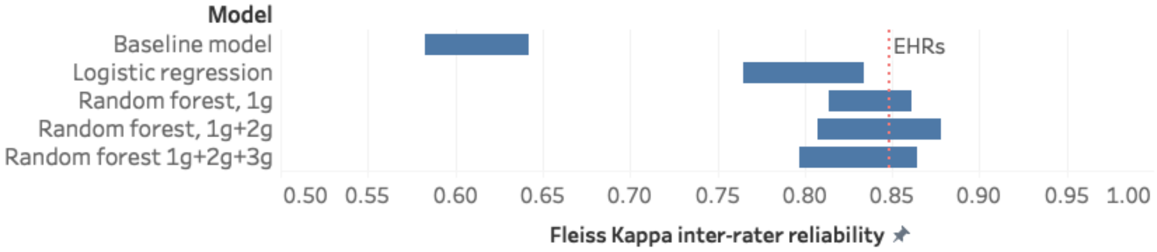
Figure 2. Accuracy of computer-based classification models when compared with expert human reviewers



Measure	Model	Estimate
Global accuracy	Baseline	78.3
	Logistic regression	90.1
	Random forest, 1g	93.4
	Random forest, 1g+2g	93.5
	Random forest, 1g+2g+3g	92.8
Sensitivity	Baseline	96.8
	Logistic regression	76.7
	Random forest, 1g	84.0
	Random forest, 1g+2g	85.4
	Random forest, 1g+2g+3g	83.6
Specificity	Baseline	70.6
	Logistic regression	95.7
	Random forest, 1g	97.3
	Random forest, 1g+2g	96.9
	Random forest, 1g+2g+3g	96.7

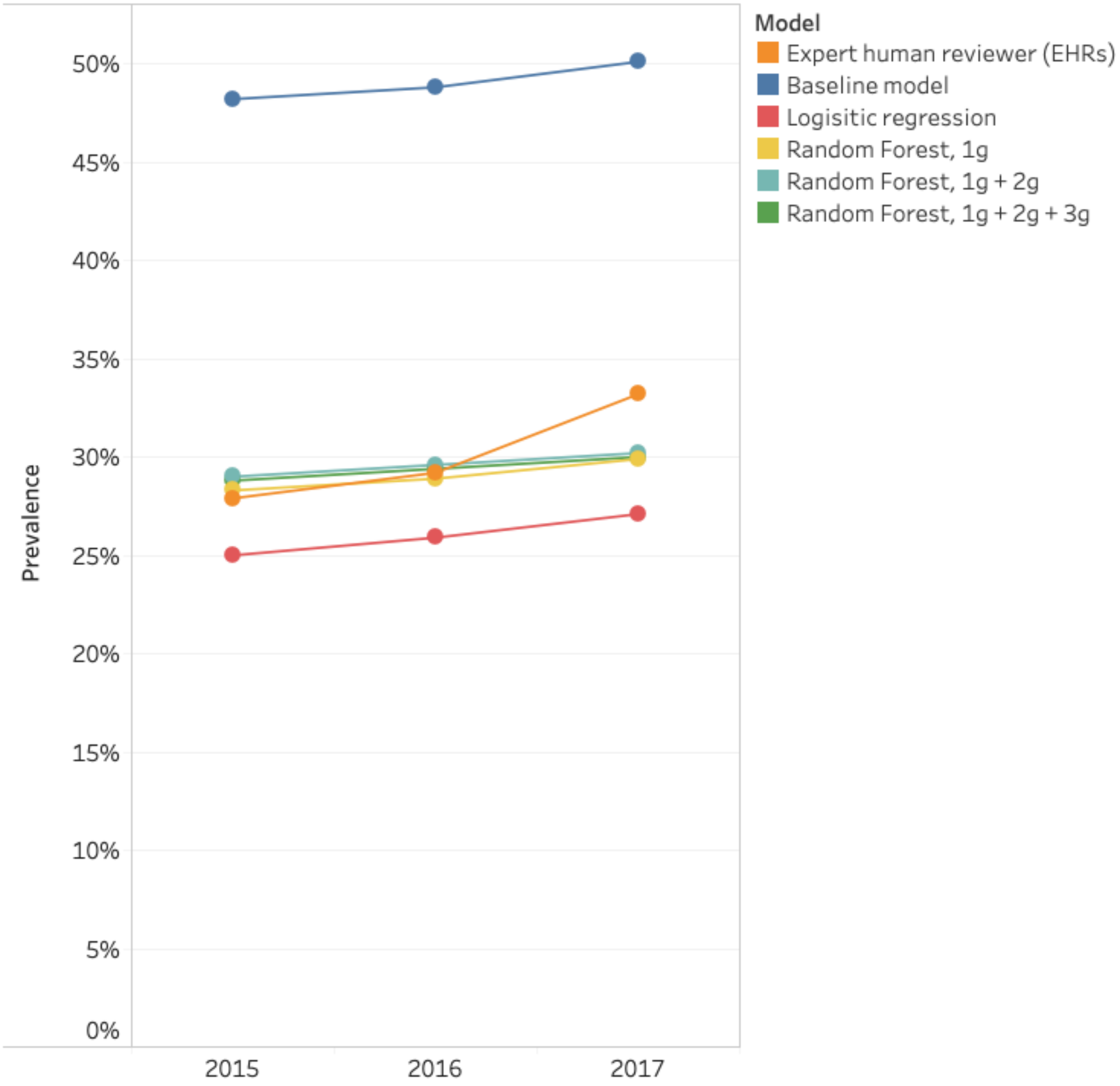
Note: Comparisons between computer models and expert human reviewers using 739 written summaries of caseworker investigations of child abuse or neglect. 1g = unigrams. 2g = bigrams, 3g = trigrams.

Figure 3. Inter-rater reliability estimates between computer models and expert human reviewers (EHRs) when classifying substance-related problems using written investigation summaries



Note: EHRs = Expert human reviewers. 1g = unigrams. 2g = bigrams, 3g = trigrams. Calculations based on four expert human reviewers and computer models, using 91 written summaries of caseworker investigations. The range of estimates was derived by replacing each EHR with the computer model and calculating the Fleiss Kappa inter-rater reliability. The inter-rater reliability among EHRs without the computer model is presented as a vertical line for reference.

Figure 4. Prevalence of substance-related problems identified based on expert human reviews and text mining models using unstructured text from caseworker investigations of child abuse or neglect, 2015-2017.



Note: Estimates for expert human reviewers (EHRs) are based on 2,956 manually reviewed and classified documents. The remaining models are model-based estimates based using 73,454 computer classified documents. 1g = unigrams. 2g = bigrams, 3g = trigrams.