Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1258–1267
Marseille, 11–16 May 2020

A Real-World Data Resource of Complex Sensitive Sentences Based on Documents from the Monsanto Trial

Jan Neerbek[1,2], Morten Eskildsen[1], Peter Dolog[3], Ira Assent[1]
[1]Department of Computer Science, Aarhus University, Aarhus, Denmark
[2]Alexandra Institute, Aarhus, Denmark
[3] Department of Computer Science, Aalborg University, Aalborg, Denmark
jan.neerbek@cs.au.dk, morten@moddi.dk, dolog@cs.aau.dk, ira@cs.au.dk

Abstract

In this work we present a corpus for the evaluation of sensitive information detection approaches that addresses the need for real world sensitive information for empirical studies. Our sentence corpus contains different notions of complex sensitive information that correspond to different aspects of concern in a current trial of the Monsanto company.

This paper describes the annotations process, where we both employ human annotators and furthermore create automatically inferred labels regarding technical, legal and informal communication within and with employees of Monsanto, drawing on a classification of documents by lawyers involved in the Monsanto court case. We release corpus of high quality sentences and parse trees with these two types of labels on sentence level.

We characterize the sensitive information via several representative sensitive information detection models, in particular both keyword-based (n-gram) approaches and recent deep learning models, namely, recurrent neural networks (LSTM) and recursive neural networks (RecNN).

Data and code are made publicly available.

Keywords: Corpus (Creation, Annotation, etc.), Statistical and Machine Learning Methods, Document Classification, Text categorisation

1. Introduction

Sensitive information detection addresses the problem of identifying (parts of) text documents that are considered sensitive in a particular application context. Sensitive information detection is of great importance in a number of applications, where unintended leak of sensitive information may incur severe negative consequences for individuals, businesses or authorities. In a study from 2017 Poneman Institute and IBM find that the average cost of a breach of sensitive information is \$3.6 million in total, for detection, escalation, notification, and after-the-fact response (Poneman Institute, 2017; Poneman Institute, 2009).

Sensitive information detection has been studied in the Natural Language Processing and Machine Learning research communities (Neerbek et al., 2018; Sánchez and Batet, 2017; Sánchez and Batet, 2016; Berardi et al., 2015; Gollins et al., 2014; Grechanik et al., 2014; Hart et al., 2011; Chow et al., 2008). In this work we focus on sensitive information detection considered as a form of text classification, where the goal is to predict whether a sentence contains sensitive information. Here we consider sensitive information to be domain specific and the definition in 4 different datasets in this work is given by labels from domain experts.

A distinguishing feature of sensitive information detection with respect to traditional text classification is that we are interested not only in the content (topics and entities) but also in context (Gollins et al., 2014). In this work we follow (Gollins et al., 2014) and focus on context within a sentence.

A limiting factor in research and evaluation of sensitive information detection methods is the lack of high quality corpora, which at least in part can be attributed to the very nature of sensitive data. Given the lack of publicly available real-world data, existing work has resorted to the creation of ad hoc evaluation data by defining particular seed keywords as sensitive (Sánchez and Batet, 2016; Grechanik et al., 2014; Chow et al., 2008), or by using two distinct data sources for sensitive and non-sensitive data (Hart et al., 2011). Evaluations using these data sources are thus limited to comparatively simple sensitive information that is captured by keyword co-occurrence alone, or may account for structural differences in data sources rather than for actual accuracy of sensitive information detection.

As an example of the types of sentences we encounter in the Monsanto corpus we provide here an example of a sentence from the $GHOST$ sensitive information type:

> But I suspect that is wishful thinking
> Are you interested in writing a column on this topic?

This sentence has been labeled sensitive by the annotators, and indeed the sentence discusses writing for Monsanto. Note that the sentence does not explicitly talk about ghost writing or even authorship of the written material. This is an example of where context influences sensitivity.

So far, a single corpus provides a language resource with real sensitive information, namely, the Enron corpus (Cormack et al., 2010; Klimt and Yang, 2004). It contains corporate documents with a variety of information content and structure, and has been used extensively to evaluate sensitive information detection (Sánchez and Batet, 2016; Hart et al., 2011; Chow et al., 2008). However, the corpus is unlabeled and more than 15 years old. Thus, the need for an up-to-date labeled corpus for state-of-the-art empirical evaluation.

We here present a new real-world sentence resource with complex sensitive information. We process, label, analyze, and characterize recently released documents that are part

of the Monsanto trial (Baum Hedlund Aristei & Goldman, 2017) as a source of great value to the research community. We provide two sets; the first contains inferred sentence level datasets based on expert labels at document level by lawyers involved in the trial ("silver" labels); the second contains labels directly annotated manually at the sentence level ("golden" labels). Following the 4 different sensitive information definitions extracted from the trial documents, we in total provide 8 classification datasets and $15,000$ labeled sentences. We release this language resource into the ELRA Catalogue of Language Resources[1] together with source code for loading of corpus and building of new models[2].

Furthermore, we characterize the complexity of the datasets in terms of the sensitive information content in the sentences. In particular, we study traditional sensitive detection methods such as n-gram or inference rule based approaches and more recent deep LSTM models and recursive neural networks. We find that all datasets present *complex* sensitive information which is not fully captured by traditional models. Complex models that consider phrase-like context capture more of the complexity of the sensitive information. Still, some sensitive information is not detected using existing methods, which provides interesting open problems for research and evaluation of future methods.

2. Related Sensitive Information Corpora and Related Work

Sensitive information corpora are scarce due to the inherently private nature of the data. This poses a challenge to research in sensitive information detection. We here review document collections used for evaluation purposes.

Open datasets such as Wikipedia has been used for detecting well-defined types of sensitive information, e.g. Personal Identifiable Information (PII); *HIV* (Health) (Sánchez and Batet, 2017) or *Catholicism* (Religion) (Sánchez and Batet, 2016). As discussed e.g. in (Neerbek et al., 2018), such forms of PII are often defined as a seed set of named entities which are comparatively easy to detect, and such resources are thus not sufficiently challenging for realistic sensitive information detection benchmarking.

WikiLeaks is used as a sensitive information source in (Hart et al., 2011), where other webpages are considered non-sensitive. That is two very different sources for content (internal secret documents versus public webpages) and successful distinction may be due to differences in data source rather than sensitivity of content. See also discussion in (Neerbek et al., 2018).

(Berardi et al., 2015) uses a corpus of 1111 historical records from UK government on *Personal Information* and *International Relations*. The corpus is not publicly available due to its sensitive nature. (McDonald et al., 2014) shows that both types of information can be modeled using features such as entity (person or country) and sentiment towards this entity, and thus does not capture aspects of sensitivity beyond such entity (sentiments).

The Enron corpus (Hart et al., 2011; Chow et al., 2008) has been partially labeled by law students as part of the TREC legal track NLP tasks (Cormack et al., 2010).

The corpus we present in this work contains real documents with complex sensitive and recent content. It complements the Enron corpus which concerns mostly finances with further complex sensitivity notions as discussed below.

3. Curation of the Monsanto Datasets

The Monsanto papers and the series of trials from which they originate are still ongoing. The trial(s) was begun in 2017 where a group sued Monsanto for claiming Roundup[3] to be safe, while Monsanto allegedly knew that Roundup could cause cancer. The Monsanto papers are internal papers from Monsanto, relevant to the trials and released due to effort by Baum, Hedlund, Aristei & Goldman law firm during the trials[4] (McHenry, 2018; Baum Hedlund Aristei & Goldman, 2017).

As part of this trial (Baum Hedlund Aristei & Goldman, 2017), law firm Baum, Hedlund, Aristei & Goldman categorizes Monsanto corporate documents into four categories (see below). No formal definition is provided, but a headline and description of the (sensitive) content in each document. Below we list the headlines and our informally derived descriptions of sensitivity notions (manually created based on content of sampled documents):

- $GHOST$, Ghostwriting, Peer-Review & Retraction. Concerns article writing and peer-reviewing by Monsanto salaried people as well as efforts in pressuring journals to retract damning studies without revealing Monsanto connection
- $TOXIC$, Surfactants, Carcinogenicity & Testing. Concerns chemical glyphosate (part of Monsanto product Roundup), in particular toxicity; declining funding for further studies; declining requested studies or declining data to regulators
- $CHEMI$, Absorption, Distribution, Metabolism & Excretion. Discussion on studies and results with regards to how animals and human react/absorb ingredients found when using Monsanto products. Discussions on starting studies deemed "risky". (Note, while $TOXIC$ is concerned with when and if Monsanto's products might cause cancer, $CHEMI$ is more concerned with the actual chemical reactions with ingredients found in Monsanto products.)
- $REGUL$, Regulatory & Government. Concerns rewards for scientists that protect Roundup business; efforts to monitor and influence regulative bodies for possible negative rulings or ratings related to Roundup/glyphosate.

We downloaded all 274 documents (emails, doc, excel, scans, etc) from lawfirm Baum, Hedlund, Aristei & Gold-

[1] http://catalogue.elra.info

[2] https://github.com/neerbek/taboo-mon

[3] Roundup is a herbicide which is developed by Monsanto

[4] https://www.baumhedlundlaw.com/toxic-tort-law/monsanto-roundup-lawsuit/

Number	21
Id	$MONGLY03934897$
Link	http://baumhedlundlaw.com/pdf/monsanto-documents/25-Invoice-Showing-Monsanto-Paid-$20000-to-Expert-Panel-Member-Dr-John-Acquavella.pdf
Link text	Invoice Showing Monsanto Paid $\$20,000$ to Expert Panel Member Dr. John Acquavella
Description	This document is an invoice dated August 31, 2015 from Monsanto to Dr. John Acquavella in the sum of $\$20,700$ for "consulting hours in August 2015 related to the glyphosate expert epidemiology panel." at $*1$.
Number	27
Id	$MONGLY02085862$
Link	http://baumhedlundlaw.com/pdf/monsanto-documents/4-Internal-Email-Further-Demonstrating-Heydens-Involvement-Drafting-Expert-Panel-Manuscript.pdf
Link text	Internal Email Further Demonstrating Heydens' Involvement in Drafting Expert Panel Manuscript
Description	This document contains an email from Dr. Heydens to Ashley Roberts regarding the introduction to the Expert Panel Manuscript. Among other features, Dr. Heydens' draft attempts to convey "that glyphosate is really expansively used." at $*1$.

Table 1: Example metadata harvested in human readable form: for each document number and id, a link to the source, a link text and a brief description are provided.

man[5] with human readable description of 120 links to documents[6]. We matched the documents with the human readable description. We resolved minor issues with matching document ids, links and descriptive texts. We use the four different types of sensitive information that Baum, Hedlund, Aristei & Goldman identified at the document level to label the documents (as discussed above).

Each document is annotated by with number, an id, a link, a link text and a description. An example is shown in Table 1. All documents are pdf documents. Some are exported from emails, word documents, and so on. Some of the documents are or contain scanned images of their text without any optical character recognition. We extracted all text encoded in the documents, but have not used OCR to transform non-text content.

Before tokenizing sentences, we removed email headers, except for the subject. We used the NLTK toolkit (Bird et al., 2009) and tokenized sentences using the Punkt sentence boundary detection approach (Kiss and Strunk, 2006), yielding $10,774$ sentences. The length distribution is shown in Table 2.

We cleaned the data further by removing very short sentences (4 words or less) and very long sentences (200 words or more). By doing so, we removed 3160 short sentences and 35 long sentences. We used label majority as the label for the dataset. We obtain a total of 7537 high quality sentences (see also Table 3).

We employ two labeling approaches to curate two sets of labels for each Monsanto datasets, to create *silver datasets* and *golden datasets*.

For the silver datasets, we assign the document label (sensitive or not with respect to each of the 4 datasets) to all sentences of that document, as provided by the lawyers at Baum, Hedlund, Aristei & Goldman. Such silver datasets thus require little human annotation effort (if we were to add more documents), as the legal experts only need to label at the document level. We thus have sensitive labels for all 7537 sentences. From documents with different labels, we uniformly at random select sentences for negative sampling for each dataset, resulting in the distribution of sentences shown in Table 4.

The silver labels are representative of application scenarios where sentence labels are not available or (too) costly to ob-

Length (characters)	Count
$[0; 4]$	1175
$[5; 19]$	1122
$[20; 74]$	2428
$[75; 124]$	2062
$[125; 299]$	3165
$[300; 499]$	573
$[500; 1000]$	195
$[1000; 3165]$	54
$[0; 3165]$	10774

Table 2: Raw sentence length distribution (in characters)

Length (words)	Count
$[5; 9]$	705
$[10; 19]$	2339
$[20; 29]$	1881
$[30; 39]$	1076
$[40; 49]$	572
$[50; 74]$	605
$[75; 99]$	203
$[100; 149]$	114
$[150; 200]$	42
$[5; 200]$	7537

Table 3: Final sentence length distribution (in words)

[5] https://www.baumhedlundlaw.com/toxic-tort-law/monsanto-roundup-lawsuit/monsanto-secret-documents/

[6] https://www.baumhedlundlaw.com/pdf/monsanto-documents/monsanto-papers-chart-1009.pdf

Dataset	Total	Train	Dev	Test
$GHOST$	6932	5900	500	532
	3466	2949	245	272
	50.00%	49.98%	49.00%	51.13%
$TOXIC$	2892	2200	340	352
	1446	1099	176	171
	50.00%	49.95%	51.76%	48.58%
$CHEMI$	2702	2100	300	302
	1351	1048	154	149
	50.00%	49.90%	51.33%	49.34%
$REGUL$	2548	1950	300	298
	1274	951	170	153
	50.00%	48.77%	56.67%	51.34%
Total	15074	12150	1440	1484
	7537	6047	745	745
	50.00%	49.77%	51.74%	50.20%

Table 4: Silver data (row 1: sentence count; row 2: sensitive sentence count; row 3: ratio of sensitive sentences)

Dataset	Total	Train	Dev	Test
$GHOST$	296	144	62	90
	77	41	14	22
	26.01%	28.47%	22.58%	24.44%
$TOXIC$	252	134	65	53
	57	26	15	16
	22.62%	19.40%	23.08%	30.19%
$CHEMI$	250	123	61	66
	32	17	5	10
	12.80%	13.82%	8.20%	15.15%
$REGUL$	275	139	69	67
	34	19	9	6
	12.36%	13.67%	13.04%	8.96%
Total	1073	540	257	276
	200	103	43	54
	18.64%	19.07%	16.73%	19.57%

Table 5: Golden data (row 1: sentence count; row 2: sensitive sentence count; row 3: ratio of sensitive sentences)

tain. In some applications, and in particular for larger documents, though, a document which contains sensitive information may also contain non-sensitive information. E.g. an email may contain greetings or best wishes which is generally not sensitive. For silver labels such documents may introduce noise. To study the impact of such noise, we also create golden labels where assignment of sensitivity is manually conducted at the sentence level. In the evaluation, we compare models constructed and tested on datasets following either labeling approach.

The golden labels are provided by 3 annotators for each sentence in a subset of about 1000 sentences. For annotation guidelines the annotators were given an introduction to the Monsanto case and the different types of sensitive information (the list introduced in the beginning of this section above), and participated in a kick-off workshop[7]. Each annotator was given the same 1073 sentences taken at random from documents labeled by the lawyers at Baum, Hedlund, Aristei & Goldman. These 1073 sentences were distributed uniformly at random across each sensitive information type. Each annotator then labels the sentence sensitive or not according to any of the sensitive information types given. We use majority of inter-annotator agreement i.e., assign sensitivity to sentences which at least 2 annotators have labeled sensitive. In our data all 3 annotators agreed on label for 65.88% of the sentences. The inter-annotator agreement can be assessed with the Fleiss Kappa (Fleiss, 1971) which takes values below or equal to 1, with 1 indicating perfect agreement and less than 0 indicating agreement by chance. Our Fleiss Kappa is 0.33 which in the rule of thumb by (Landis and Koch, 1977) can be considered a "fair agreement".

Distribution of labels in this golden annotated dataset is shown in Table 5.

[7]See also https://github.com/neerbek/taboo-mon/blob/master/doc/AnnotationDescription.txt

We observe that $GHOST$ and $TOXIC$ have sensitive ratio around 25%, where $CHEMI$ and $REGUL$ are more skewed with a sensitive ratio around 15%.

4. Empirical Characterization

We characterize the sensitivity of information in sentences in our data resource by an empirical study of existing approaches in the field. We place particular focus on comparing silver and golden labels.

4.1. Detection Models

Broadly speaking, the models for sensitive information detection can be divided into *keyword*-based and *context*-based (Neerbek et al., 2018). Keyword-based approaches assign probabilities to words (or rather, n-grams) occurring in sensitive (or non-sensitive) sentences. They differ in how they utilize these probabilities (Sánchez and Batet, 2016; Berardi et al., 2015; Grechanik et al., 2014; Hart et al., 2011; Chow et al., 2008). Context-based approaches consider the context (beyond n-grams) of a word occurrence for assigning probability of a sentence being sensitive. Dense embedding approaches can be seen as a prototypical way of encoding context for a word (e.g. (Mikolov et al., 2013; Pennington et al., 2014)). In a context-based approach, the probability of a particular word or phrase being sensitive is allowed to vary with the context (sentence, paragraph, document) in which the word appears, allowing them to detect more complex types of sensitive information that are not characterized by (co-)occurrence of keywords alone. In this evaluation we focus on sentence level sensitive information.

We quantify the complexity of our corpus by making use of these characteristic differences in keyword-based and context-based approaches, respectively. Simply put, datasets where the performance gap between the two is large, contain more complex sensitive information. We use recurrent memory cell neural networks, LSTM(Hochreiter and Schmidhuber, 1997) and recursive neural networks,

RecNN(Elman, 1990; Goller and Kuchler, 1996; Socher et al., 2013) as examples of context-based approaches. Both generate an embedding for each context and predict based on this context embedding.
Keyword-based approaches used are InfRule (Chow et al., 2008), C-san (C-sanitized) (Sánchez and Batet, 2016) and an empirical upper bound on keyword-based approaches we term *Keyword-Max*.

InfRule. One of the earliest works in the sensitive information detection domain (Chow et al., 2008) is inspired by association rule mining (Agrawal and Srikant, 1994). It considers words in a sentence as events in a probabilistic process and discovers rules which can either be simple: $w \rightarrow s$ (*word* w *implies sensitive information* s) or complex combinations using conjunction, disjunction and logical not $(w_1 \wedge w_2 \wedge \neg w_3 \wedge (w_4 \vee w_5)) \rightarrow s$. The confidence of a rule is the fraction of times it occurs and predicts correctly in the training set. We follow the setup in (Chow et al., 2008) which uses InfRule on the Enron corpus using simple rules and a constant confidence cutoff.

C-san. (Sánchez and Batet, 2016) use point-wise mutual information (PMI) between a word w and a type of sensitive information s (s can be a known sensitive word or inferred some other way) $PMI(s;w) = \log \frac{P(s \wedge w)}{P(s)P(w)}$, i.e., logarithm of the probability of the joint occurrence of word w and sensitive information s, normalized by the probability of occurrences of sensitive information s multiplied by the probability of occurrences of the word w. A sentence is considered sensitive if its PMI exceeds a sensitivity threshold. The threshold is determined using the *information content* (IC) of the sensitive information s, defined as the logarithm of the fraction of occurrences of sensitive information s: $IC(s) = -\frac{1}{\alpha}\log(P(s))$, where α is a user defined constant which reflects the cost of false negatives. A text is sensitive if for any word we have $PMI(s;w) \geq IC(s)$. The intuition behind this definition is that (for $\alpha = 1$) PMI is maximal if $PMI(s;w) = IC(s)$ and word w will predict/disclose the information s with probability 1, thus w is a good predictor. By dividing IC by $\alpha > 1$ we detect keyword-based predictors with lower than 1 probability and thus will be able to predict sensitive information even when perfect predictors do not exist.

Keyword-Max. To identify how much of the sensitive information potentially could be captured by keyword-based approaches, we include a form of (upper) empirical baseline. We allow it to set hyperparameters based on the test set, which means it is given access to additional information that in reality is not available. It is still interesting as it denotes the limit of keyword based approaches, and thereby provides a further indication of the complexity of sensitive information that cannot be captured by keyword-based approaches.

LSTM. The sequential LSTM approach builds a neural network model and for each word in a sentence applies the neural network in sequence. For a given text $t = (w_1, w_2, \ldots, w_n)$ and for each step consider a new word w_i and apply the neural network to obtain both a new memory cell state and a hidden state. Whereas the hidden state is mainly used to parse information from one step to another, the memory cell is "protected" by several gated states which allows the LSTM to carry information across longer step counts than what is generally possible using vanilla recurrent neural networks. In our previous work (Neerbek et al., 2019) we built LSTM models for sensitive information detection. Prediction is based on the state arrived at after sequentially processing every word in the sentence by adding a fully connected layer. In our evaluation we apply these models on the Monsanto datasets developed here.
Please note that the LSTM could be augmented with structural information similar to the RecNN below. In our dataset characterization, we use the LSTM as a representative of unstructured sequential deep methods, and the RecNN as a structured one. Both approaches use GloVe word embeddings (Pennington et al., 2014).

RecNN. As discussed in (Neerbek et al., 2019; Neerbek et al., 2018) the recursive neural network, RecNN, approach has been used successfully for sensitive information detection. The use of RecNN for context dependent tasks is motivated by the previous RecNN models for e.g. sentiment analysis (Socher et al., 2013) and paraphrase detection (Socher et al., 2011). In a RecNN we are given both the text t and a structure over the text S. As structure here we generate probabilistic context-free grammars (pcfg) based constituent trees (Klein and Manning, 2003), where the pcfg was trained over the Penn Treebank (Taylor et al., 2003). Let Y be the set of all nodes in the structure and all words in t, then the structure S is a mapping from each element in Y to a list of parents also in Y. The structure can be a directed acyclic graph (DAG). In this study we follow (Socher et al., 2013; Socher et al., 2011) and restrict the approach to only tree-like structures. In this case the length of the list of parents is at most 1, and the list of parents is empty for the root node in the structure. As described in Section 3, our data resource contains constituency parse trees for each sentence (text) t. We follow (Neerbek et al., 2018) where given a sentence, the root state is the last state of the evaluation of the neural network on that sentence which may carry most information about the sentence. As for the LSTM, we add a fully connected layer for predicting sensitivity. Compared to our previous work we develop transfer learning for the RecNN model between our silver and golden dataset and show improved performance of the RecNN model.

Experimental setup. Both InfRule and C-san use a cutoff of minimum confidence that a keyword must have. These cutoffs are set using the dev dataset. In contrast, Keyword-Max is allowed to set that cutoff based on the data in the test set, even though that is not available in a real application. As we observe in our study, there is a limit to the sensitive information that keyword-based approaches can successfully detect, which makes it possible for us to reliably characterize complex sensitivity in our datasets. InfRule uses default parameters on Enron data as in (Chow et al., 2008), C-san α values used in (Sánchez and Batet, 2016), namely, $\alpha \in \{1, 1.5, 2\}$. LSTM and RecNN approaches use GloVe embeddings (Pennington et al., 2014), with embedding size 100 given the relatively low number

of labeled sentences. Dropout rate 0.5 was found to work well for LSTM, while lowering dropout rate for RecNN to 0.1 yielded the best results. For LSTM we obtain the best results using AdaDelta optimizer for learning rate optimization. For RecNN the best results were found using stochastic gradient decent (SGD) with learning rate determined through line search. Please note that we are mainly interested in obtaining optimal performance for each approach such that the complexity of the datasets is accurately characterized.

4.2. Silver Labels

In the following, we characterize our data resource with the above models using silver labels for training and evaluation. For each sensitive information type we train a specific model for each of the approaches.

In Table 6, we characterize sensitive information complexity using silver sentence labels on reported accuracy scores[8]. We observe that InfRule generally finds more complex sensitive information than C-san when α is set to 1, but if this parameter is optimized, C-san captures additional sensitive information beyond InfRule results. We observe that InfRule and C-san generally perform better on $REGUL$, where differences between all models are smaller. This indicates less complex sensitive information compared to the other datasets. Additionally, we find that by giving keyword-based approaches access to test set information, in the Keyword-Max model as described above, we obtain an empirical upper limit on the less complex sensitive information as follows:

$GHOST$	$TOXIC$	$CHEMI$	$REGUL$
78.60%	73.24%	80.67%	75.00%

The context-based approaches LSTM and RecNN are capable of capturing more complex sensitive information beyond the keyword-based approaches. We observe that on silver labels LSTM has best performance on $TOXIC$ and $CHEMI$. These datasets both deal with discussions on cause and effect of chemical compounds and experimental design. Likely, this follows a more sequential buildup, presentation-wise, which the LSTM is particularly designed for capturing. Conversely, we observe that the structured approach RecNN which has access to the constituency parse tree for each text shows best performance for datasets $GHOST$ and $REGUL$. Both datasets contain many emails and are thus more conversational in nature. Accordingly, we observe that the RecNN outperforms LSTM here. This shows that complex sensitive information may show different structures in these datasets.

Overall, we conclude that all datasets contain sensitive information that can be captured by keyword-based approaches, but also more complex types that require advanced methods that exploit the context. We also note that none of the approaches achieves close to perfect accuracy, i.e., these datasets still provide potential for research on methods that can capture aspects of sensitivity that are not currently detected.

[8]More details on experiments parameters can be found in `https://github.com/neerbek/taboo-mon/blob/master/doc/ExperimentParameters.txt`

4.3. Golden Labels

We now turn to the characterization of the data with respect to the golden labels. We subdivide this study into four cases and due to space considerations we restrict our characterization experiments to our most expressive model family, the RecNN. While small differences occur, the overall conclusions remain the same.

Furthermore in our 3. case we motivate the use of transfer learning between our larger silver dataset and the smaller golden dataset as a way to characterize the level of sensitive information learnable from the silver dataset. Such characterization is based on the concept of transfer learning discussed in (Yosinski et al., 2014) for embedding based model families and thus not as such applicable to the keyword-based approaches.

In our 4. case we return to characterization using all models, including the transfer learning models and summaries the characterizations learned over the datasets.

1. Case: Silver-to-Golden In this evaluation, we build silver label based models and study how well they predict golden labels. This allows an understanding of how valuable the relatively easily obtainable document-based silver labels are when compared to human labels on sentence level. Note that in the silver dataset all the labels of the golden subset are sensitive. If the models have learned to distinguish sentences containing sensitive information from noisy, falsely labeled non-sensitive sentences then the model should predict some of the sentences correctly as non-sensitive in the golden dataset simply because the model has learned the sensitive information type. Put differently, noisy sentences which are incorrectly labeled sensitive in the silver dataset may be similar to non-sensitive sentences in the silver dataset. Consider our previous example with sensitive emails. The initial greeting may be very similar to other greetings from non-sensitive emails. Thus a model may still learn to correctly label greetings as non-sensitive even though they appear in a sensitive email. When this is the case, we say that the model has successfully learned the sensitive information type, and it is an indication of the usefulness of sensitive labels for training of sensitive information detection models.

2. Case: Golden-to-Golden Here, models trained on golden labels are evaluated against golden label test sets. This provides insight into accuracy using sentence level human labels. A major challenge with the golden dataset is its smaller size as it is based on manual effort, which may make the models prone to overfitting. We train with different types of regularization to combat overfitting.

3. Case: Silver-Transfer-to-Golden The third case outlines how transfer learning models may combine both silver and golden labels to counter both issues with noise in silver labels and issues with limited training data in golden labels. It further provides an indication about the relationship between the silver and golden labels beyond Case 1. Our study is based on transfer learning for deep neural models as discussed in (Yosinski et al., 2014) for convolutional models (CNNs). They train a layered model on one task and then *transfer* the weights to a second task that benefits if sufficiently similar. In our study, we transfer all layers

Approach	$GHOST$	$TOXIC$	$CHEMI$	$REGUL$
InfRule	57.80%	59.71%	60.33%	67.33%
C-san; $\alpha = 1$	49.60%	52.94%	54.00%	61.33%
C-san; $\alpha = 1.5$	62.40%	65.29%	67.67%	71.33%
C-san; $\alpha = 2$	72.60%	70.29%	71.33%	74.33%
LSTM	83.60%	**77.33%**	**86.67%**	82.33%
RecNN	**86.60%**	75.00%	83.67%	**87.00%**

Table 6: Characterizing complexity of silver label data using accuracy of keyword-based (top) and context-based approaches (bottom): keyword-based approaches can successfully capture the majority of sensitive content; more complex sensitive information is captured by deep learning methods; no existing method can fully recover all sensitive content

Dataset	Prec-Sen	Prec-Non-sen	Acc
$GHOST$	31.82%	61.76%	54.44%
$TOXIC$	37.50%	83.78%	69.81%
$CHEMI$	70.00%	51.79%	54.55%
$REGUL$	16.67%	73.77%	68.66%

Table 7: Dive in on performance of RecNN model; Precision and accuracy on golden label test set for models using silver labels for training.

except 1 from silver models and train the final layer using the golden training set.

4. case: Overview on Golden We provide an overall comparison of all models to characterize the golden dataset as we did with the silver dataset in Section 4.2

4.3.1. Results - 1. Case: Silver-to-Golden

In Table 7 we show precision for each class (sensitive vs non-sensitive) as well as accuracy against the golden labels. Consider a correctly predicted sensitive label as true-positive (tp), a sensitive label predicted incorrectly as non-sensitive as false-negative (fn), a correctly predicted non-sensitive label as true-negative (tn) and a non-sensitive label predicted incorrectly as sensitive as false-positive (fp), then precision sensitive is Prec-sen $= \frac{tp}{tp+fn}$, and precision non-sensitive is Prec-Non-sen $= \frac{tn}{tn+fp}$.

Due to space limitations, we here show results only for RecNN models that capture most sensitive information in our previous evaluation. The focus in this characterization is on the relationship between silver and golden labels; a final overview also on the golden labels is provided in the final characterization.

From the results in Table 7 we observe that models trained on silver labels do learn to correctly predict sensitive sentences vs non-sensitive sentence, even though all non-sensitive sentences in the golden test sets are labeled sensitive in the silver datasets. This demonstrates that datasets with silver noisy labels indeed provide useful training data for sensitive information detection models.

4.3.2. Results - 2. Case: Golden-to-Golden

In the interest of space, we only present RecNN characterization as before (results for all models are summarized in the final overview). We train models on the training data with golden labels and evaluate on the golden test sets. As the golden datasets are relatively small due to the efforts in manually labeling on sentence level, we particularly study overfitting. For this, we show performance results on training, development (validation) and test sets, separately (Table 8). As expected, the models that perform well on the training data fail to generalize well to the development and test set, i.e., experience overfitting. Concretely, the models reach almost 100% accuracy on the training set, but much lower accuracy on the development and tests sets. Model hyperparameters was found through a line search on development set[9].

Except on $TOXIC$ where we observe higher test score than just always predicting "non-sensitive", we observe that the overfitting results in worse generalization (i.e., test scores being lower than major class fraction). $TOXIC$ seems to have a high ratio of sensitive information in the test set. The data was sampled uniformly and thus the distri-

Dataset	Train	Dev	Test
$GHOST$	71.53%	77.42%	75.56%
	100.00%	79.03%	75.56%
$TOXIC$	80.60%	76.92%	69.81%
	100.00%	76.92%	71.70%
$CHEMI$	86.18%	91.80%	84.85%
	100.00%	83.61%	80.30%
$REGUL$	86.33%	86.96%	91.04%
	100.00%	82.61%	88.06%

Table 8: Accuracies on golden test set by training using golden label training set only. For each dataset, row 1 is accuracy if always predicting "non-sensitive", row 2 RecNN accuracy. Note: 100% accuracy on training set and poor test results mean overfitting due to small training sets.

Dataset	Acc	Non-sen
$GHOST$	77.78%	75.56%
$TOXIC$	71.70%	69.81%
$CHEMI$	84.85%	84.85%
$REGUL$	92.54%	91.04%

Table 9: Accuracy obtained on golden label test set using transfer learning, i.e., trained first on silver label training set, then all but one layers fixed and finetuning the final layer using the golden label training sets.

[9] https://github.com/neerbek/taboo-mon/blob/master/doc/ExperimentParameters.txt

Approach	$GHOST$	$TOXIC$	$CHEMI$	$REGUL$
InfRule	76.67%	**73.58%**	**84.85%**	92.04%
C-san; $\alpha = 1$	**77.78%**	**73.58%**	83.33%	91.04%
C-san; $\alpha = 1.5$	75.56%	69.81%	**84.85%**	91.04%
C-san; $\alpha = 2$	75.56%	69.81%	**84.85%**	91.04%
LSTM	**77.78%**	69.81%	**84.85%**	91.04%
RecNN	75.56%	71.70%	80.30%	88.06%
RecNN-tf	**77.78%**	71.70%	**84.85%**	**92.54%**

Table 10: Characterizing complexity of sensitive information on golden test sets using keyword-based approaches (top) and context-based approaches (bottom): keyword-based approaches capture more sensitive content on less noisy golden data as compared to silver data; across almost all models and datasets performance increases slightly; in particular, REGUL golden labels seem easiest to recover; transfer learning captures most sensitive content as it makes use of both silver and golden labels; no existing method can fully recover all sensitive content

bution is expected to be uniform, but for small size datasets small variance in actual numbers can lead to a biases which can contribute to the score on $TOXIC$.

The overfitting is a sign that the sensitive information types are difficult to detect and require larger samples of labeled data to detect properly. If the information types could be characterized using a simple set of keywords, then we would expect our RecNN model to be able to obtain better performance. Our results in Table 8 implies that our sensitive information types extend beyond simple keyword based definitions and in fact contain some complex information.

In the next section we address the need for additional data (using transfer learning) and show increased performance for our models when we can combine golden datasets with transferred learning from the silver datasets. This indicate a key characteristics of our sensitive information datasets, namely that they do indeed carry complex sensitive information which cannot be captured by simple keyword-based approaches alone.

4.3.3. Results - 3. Case: Silver-Transfer-to-Golden

We now turn to the combination of silver labels and golden labels using transfer learning. As discussed above, transfer learning allows making use of both silver and golden labels, thereby potentially counteracting noise and limited training data. We used the same models trained on the silver datasets as above for transfer learning with golden labels. We then trained a single layer model on top of these (fixed) representations. We fine-tuned hyper-parameters using line search and found adding data augmentation in the form of small amounts of random noise to the input embeddings worked well as regularization. We obtained the test accuracies shown in Table 9.

With transfer learning we are able to extract the most learning from the datasets, i.e., obtain the highest accuracies across the datasets. Following (Yosinski et al., 2014) we know that transfer learning performs well if the two tasks share similarities, which means that silver and golden labels are sufficiently related, and can thus be used for training and evaluation sensitive information detection models.

We have successfully transferred learning from the original models (the silver labels) to the golden labels. This furthermore implies that our document based silver labels actually provide knowledge which with relatively little effort can be utilized for sensitive information detection, even at the sentence level. Noise in the silver labels can thus be successfully ignored by the models used in our characterization.

Similar performance even in the face of noise in the silver labels furthermore implies that, all things being equal, a larger dataset with silver labels may be more valuable than a smaller golden label dataset. If available, the combination of the two labels in learning seems a promising approach indeed, both with respect to training and with respect to evaluation of approaches.

4.3.4. Results - 4. Case: Overview on Golden

We conclude the characterization by comparing all models on the golden datasets. In Table 10, we provide a complete overview over results of all the models used to characterize the golden datasets. *RecNN-tf* here denotes the transfer model discussed in the previous section.

We note that the golden dataset, as seen before, provides limited training data, which means that RecNN does not perform well. The performance of the different keyword-based methods is similar in trend as we saw in Table 6, C-san performing better than InfRule when the α parameter is chosen to match the dataset. On $REGUL$ InfRule is slightly better than C-san, both worse than RecNN-tf. The RecNN-tf model performs better than the keyword based models, except for $TOXIC$ where the small dataset sizes makes the keyword based methods better.

Overall, the transfer model RecNN-tf provides the best performance and thereby indicates how much of the sensitive information can be successfully captured by the models in our study using both silver and golden labels. It thus also provides an indication of the potential for further improvement of sensitive information detection models using this data resource.

5. Conclusion

In this work, we present new, real-world datasets based on the Monsanto documents labeled by lawyers involved in the court case. We provide labels following two different labeling approaches, *golden* and *silver*, with the data - in total 8 datasets for the sensitive information detection research community.

6. Bibliographical References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499. Morgan Kaufmann Publishers Inc.

Baum Hedlund Aristei & Goldman. (2017). Monsanto papers | secret documents. https://www.baumhedlundlaw.com/toxic-tort-law/monsanto-roundup-lawsuit/monsanto-secret-documents/. Retrieved: 2018-May-09.

Berardi, G., Esuli, A., Macdonald, C., Ounis, I., and Sebastiani, F. (2015). Semi-automated text classification for sensitivity identification. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1711–1714. Association for Computing Machinery (ACM).

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Chow, R., Philippe, G., and Staddon, J. (2008). Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 893–901. Association for Computing Machinery (ACM).

Cormack, G. V., Grossman, M. R., Hedin, B., and Oard, D. W. (2010). Overview of the trec 2010 legal track. In *Proceedings of the 19rd Text REtrieval Conference*, TREC '10.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Goller, C. and Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. In *IEEE International Conference on Neural Networks*, IEEE ICNN '96, pages 347–352. Institute of Electrical and Electronics Engineers (IEEE).

Gollins, T., McDonald, G., Macdonald, C., and Ounis, I. (2014). On using information retrieval for the selection and sensitivity review of digital public records. In *PIR'14: Privacy-Preserving IR Workshop*, SIGIR '14, pages 39–40.

Grechanik, M., McMillan, C., Dasgupta, T., Poshyvanyk, D., and Gethers, M. (2014). Redacting sensitive information in software artifacts. In *Proceedings of the 22Nd International Conference on Program Comprehension*, ICPC 2014, pages 314–325. ACM.

Hart, M., Manadhata, P., and Johnson, R. (2011). Text classification for data loss prevention. In *Proceedings of the 11th International Conference on Privacy Enhancing Technologies*, PETS '11, pages 18–37. Springer-Verlag.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 423–430. Association for Computational Linguistics (ACL).

Klimt, B. and Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, ECML '04, pages 217–226. Springer-Verlag.

Landis, R. J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

McDonald, G., Macdonald, C., Ounis, I., and Gollins, T. (2014). Towards a classifier for digital sensitivity review. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval*, ECIR 2014, pages 500–506. Springer International Publishing.

McHenry, L. B. (2018). The monsanto papers: poisoning the scientific well. *International Journal of Risk & Safety in Medicine*, 29(3-4):193–205.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS '13, pages 3111–3119. Curran Associates Inc.

Neerbek, J., Assent, I., and Dolog, P. (2018). Detecting complex sensitive information via phrase structure in recursive neural networks. In *Advances in Knowledge Discovery and Data Mining*, PAKDD '18, pages 373–385. Springer International Publishing.

Neerbek, J., Dolog, P., and Assent, I. (2019). Selective training: A strategy for fast backpropagation on sentence embeddings. In *Advances in Knowledge Discovery and Data Mining*, PAKDD '19, pages 40–53. Springer International Publishing.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543. Association for Computational Linguistics (ACL).

Poneman Institute. (2009). Fourth annual us cost of data breach study. https://www.ponemon.org/local/upload/file/2008-2009%20US%20Cost%20of%20Data%20Breach%20Report%20Final.pdf.

Poneman Institute. (2017). 2017 cost of data breach study (sponsored by ibm). https://www.ponemon.org/library/2017-cost-of-data-breach-study-united-states.

Sánchez, D. and Batet, M. (2016). C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology (JASIST)*, 67(1):148–163.

Sánchez, D. and Batet, M. (2017). Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34.

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., and

Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS '11, pages 801–809. Curran Associates Inc.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 1631–1642. Association for Computational Linguistics (ACL).

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: An overview. In A Abeillé, editor, *Treebanks*, Text, Speech and Language Technology, pages 5–22. Springer.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS '14, pages 3320–3328. MIT Press.