Systems in Language: Text Analysis of Government Reports of the Irish Industrial School System with Word Embedding

Susan Leavy
University College Dublin
Ireland
susan.leavy@ucd.ie

Mark T. Keane
University College Dublin
Ireland
mark.keane@ucd.ie

Emilie Pine
University College Dublin
Ireland
emilie.pine@ucd.ie

Abstract

Industrial Memories is a digital humanities initiative to supplement close readings of a government report with new distant readings, using text analytics techniques. The Ryan Report (2009), the official report of the Commission to Inquire into Child Abuse (CICA), details the systematic abuse of thousands of children 15 from 1936 to 1999 in residential institutions run by religious orders and funded and overseen by the Irish State. Arguably, the sheer size of the Ryan Report—over 1 million words—warrants a new approach that blends close readings to witness its findings, with distant readings that help surface system-wide findings embedded in the Report. Although CICA has been lauded internationally for 20 its work, many have critiqued the narrative form of the Ryan Report, for obfuscating key findings and providing poor systemic, statistical summaries that are crucial to evaluating the political and cultural context in which the abuse took place (Keenan, 2013, Child Sexual Abuse and the Catholic Church: Gender, Power, and Organizational Culture. Oxford University Press). In this article, we concentrate on describing the distant reading methodology we adopted, using machine learning and text-analytic methods and report on what they surfaced from the

Report. The contribution of this work is threefold: (i) it shows how text analytics can be used to surface new patterns, summaries and results that were not apparent via close reading, (ii) it demonstrates how machine learning can be used to annotate text by using word embedding to compile domain-specific semantic lexicons for feature extraction and (iii) it demonstrates how digital humanities methods can be applied to an official state inquiry with social justice impact.

Keywords: text analysis, text classification, machine learning, industrial schools, child abuse

1. Introduction

The Ryan Report (Ryan, 2009) details the findings of the Irish Government's Commission to investigate child abuse in Irish Industrial Schools, run by Catholic Religious Congregations from the 1936-1999. The Report provides an extensive catalogue of abuse carried out in these schools and had a major societal impact in Ireland with respect to public attitudes to the moral authority of the Roman Catholic Church (Donnelly and Inglis, 2010; Pilgrim, 2012). However, aspects of its narrative structure have been criticised for obscuring as much as it revealed. The anonymisation of names of the clergy for instance has been criticised for protecting the religious orders (Powell et al, 2012) and the structure of the document obscures the systematic nature of the abuse (Pine et al. 2017).

This paper reports on the use of text analytics to surface heretofore-invisible underlying patterns and enable a system-wide analysis of the contents of the report and

facilitate new kinds of reading through an interactive web-based platform[1]. It presents a distant reading methodology whereby word embedding is used to compile domain-specific semantic lexicons for feature extraction to enable machine learning classifiers to annotate excerpts of the Ryan Report according to its meaning. In the remainder of this introduction, we identify key shortcomings of the Report (see section 1.1), specify our motivation for doing the current work (section 1.2), outline the key themes in the report (section 1.3) and outline the structure of the remainder of the paper.

1.1 *Shortcomings of the Ryan Report*

The structure and narrative form of the Ryan Report organises information in a way that impedes a system-wide analysis of abuse in the Irish industrial school system. Preliminary chapters describe the historical background of the school system, the terms of the Commission of Investigation and how various selected sources were used[2]. The main body of the report is then comprised of a collection of chapters organized by school[3]. Each chapter begins with a historical overview of the school and its management. The narrative then moves to a consideration of the events involving clerics or lay staff in the school, about whom accusations of abuse were made.

Due to this segregation of information by school, the descriptions of serial abusers and their movements from school to school are distributed across many chapters. This makes it very difficult, if not impossible, for the reader to build a coherent history of a given individual who may have worked at several schools. Indeed, in the context of 432 members of religious orders spread across 66 chapters, even the

[1] https://industrialmemories.ucd.ie/
[2] See the Commission to inquire into child abuse (CICA) Report, Vol. 1.1 to Vol. 1.5 (available at: http://www.childabusecommission.ie/rpt/).
[3] See the CICA Report, Vol. 1.6 to Vol. 2.16.

most assiduous reader cannot easily connect a given individual's sequence of offenses in any coherent way. This narrative structure obscures the movement of staff between institutions, which was a common response of governmental and congregational bodies to allegations of abuse. This structure also makes institutional comparisons difficult, thus obfuscating the system-wide conditions that allowed abuse to emerge and become endemic.

Within the chapters on each school, information is further divided in sections according to individual perpetrators detailing evidence of abuse and the response of the religious orders. While this approach is consistent with the concept of individual responsibility that is fundamental to a retributive justice system (Hagan et al., 1981), such individualised narratives deflect from the complex social phenomena that contribute to the occurrence of abuse (Keenan, 2012).

1.2 Motivation for a Distant Reading of the Ryan Report

The motivation for the current work arose from the difficulties in undertaking a cross-institutional, systemic analysis. Hence, we advance a suite of techniques, using word embedding and text analytics (i.e. text classifiers) to perform distant readings of the document and annotate extracts of text based on their content. We outline a methodology for generating a set of domain-specific keywords (doing query expansion from minimal seed keywords) to compile lexicon-based features for classifiers that can be used in conjunction with other features to identify paragraphs based on their semantic content. This methodology could potentially be used to analyse the content of similarly voluminous reports resulting from other investigations (e.g. Royal

Commission Report, Australia, 2017 (covering 8,000 witness testimonies); Truth and Reconciliation Commission, Canada, 2015 (over 7,000 testimonies)).

A central motivation also concerned issues with the application of machine learning techniques in the area of digital humanities. In many digital humanities projects, although corpora are too large to conduct comprehensive close readings, they are often not large enough to employ 'big-data' methodologies such as machine learning mainly due to the cost of compiling sufficient training data (Schöch, 2013). We addressed this issue by outlining a scalable methodology that enables machine learning to be used for annotation with relatively small training-datasets.

1.3 Knowledge Categories

Annotating excerpts of the Ryan Report based on their semantic content enabled the existing narrative of the report to be deconstructed and its findings to be extracted and read in new ways. The following outlines the thematic foci of this analysis and their relevance to gaining a system-wide understanding of the dynamics of abuse at Irish industrial schools:

Witness testimonies: Extracting the accounts of individual witnesses recorded in the text, to allow us to collate and examine in detail all of the testimony embedded in the Report. Experts of testimony of witnesses in the Ryan Report are most commonly presented in the form of block quotes and preceded by a colon along with introductory text contextualising the source of quotations. Shorter in-text quotations are identified by quotation marks. The same punctuation is used also to signify extracts from historical documentary sources such as reports and letters necessitating semantic analysis of the text introducing the quitations.

9.47 A common thread running through the testimony of the complainants was that punishment was meted out indiscriminately and that this created an environment of fear. One witness, who was a resident for eight years from the early 1950s, stated:

> *it didn't really matter what you were beaten for, it was just one of those things, if they saw you there and you weren't doing something then you got beaten for it.*

(Vol. 2.9)

Transfer events: These paragraphs deal with the responses to allegations of abuse in the industrial schools where, typically, the cleric involved was transferred from one institution to another. In some cases the cleric involved was moved out of the schools system (to a parish or a Congregational House), dismissed or granted a dispensation from their vows. Those paragraphs recounting the movement of accused abusers, to enable us to view the transfer trajectories of specific individuals and to surface patterns of movement between institutions obscured by the linear narrative structure of the Report.

7.374 Br Lancelin came under suspicion of sexual involvement with boys while he was in Artane in 1944 and was transferred to Carriglea. His personal card stated:

> Suspicion had been aroused by a tendency to particular friendship with a boy in Artane.

(CICA Vol. 1.7)

Abuse events: These paragraphs detail abusive events (i.e., physical, emotional and sexual abuse) and are a crucial to understanding the scale and nature of abuse across the industrial school system. The language used to describe abusive events is complex reflecting the varied experiences of the 1,090 witnesses who gave evidence of abuse experienced at the industrial schools. Extracting such paragraphs allows us to identify, collate and examine in detail the representations of abuse in the Ryan Report.

7.124 This treatment indicates that the boy had lacerations to his arm and head, in addition to the fracture that was later diagnosed. The severity of the beating must have been obvious.

(CICA Vol. 1.7)

1.4 Outline of Paper

In the remainder of this paper, we present the techniques we used for a distant reading of the Ryan Report. We review the main collections of research relevant to our concerns in Section 2. We then describe the techniques and present the results of our research. In Section 3, we outline how we used word-embedding methods, specifically word2vec (Mikolov, 2013), to carry out feature extraction in order to classify the semantic content of excerpts of the report. Section 4 describes how these domain-specific semantic lexicons along with other features can be used in a suite of classifiers designed to automatically identify particular text items in the Report. This section also reports the results evaluating the effectiveness of these classifiers in detecting the semantic content.

2. Background

The approach we adopted in this project encompasses findings from previous studies in relation to the requirements for a digital platform to enable distant reading. It also builds upon previous approaches to using machine learning to automatically classify text.

2.1 *Digital Platforms for Humanities Research*

Widlöcher et al. (2015) outlined guidelines for platforms for humanities research demonstrating how enriching data through annotation, segmentation of documents, statistical analysis and comprehensive search functionality enables distant reading. They also emphasise the importance of retaining structural elements of documents to facilitate

close readings. This incorporation of both close and distant reading functionality within an exploratory digital interface was demonstrated in work by Hinrichs et al. (2015) and Kopaczyk (2013).

Distant reading though the extraction and exploration of relationship between entities in text is a central function of many platforms (Muralidharan and Hearst, 2013; Vuillemot et al., 2009). Jokers and Mimno (2013) emphasises distant reading using methods such as topic modelling and visualisation. In developing an approach to digitally analysing the Ryan Report we build on requirements outlined in these related digital humanities projects.

2.2 *Annotation in Humanities Research*

Analysis of the contents of the Ryan Report involved the automatic classification of paragraphs based on their content. In exploring approaches to annotation in humanities research it is important to appreciate the important role that manual annotation plays in the critical analysis of text (Jackson, 2001). Researchers gain in-depth knowledge of the corpus through the process of evaluating its meaning and annotating the text. The development of distant reading methods must therefore aim not to simply replace this interpretative stage but to enhance it. Incorporating input from domain experts into the process is key to achieving that and also ensuring the interpretability of automation so the classification process itself may be critically analysed. This is demonstrated in work by Sweetnam and Fennel (2012) who included input from experts in each stage of their annotation process.

2.3 *Automated Annotation*

There are two main approaches to automated annotation, rule-based and statistical machine learning. Chiticariu et al. (2013) outlines how the data-analytics industry primarily employs rule-based approaches to annotation and information extraction despite major developments in academia in using machine learning. This they found, is largely due to the fact that rule-based methods are interpretable, can incorporate domain knowledge easily and do not require extensive training data.

A comparable situation persists in digital humanities where despite an abundance of research developing automated methods for annotation many projects rely on manual annotation of text (Mahlow et al., 2012). This is due in large part to the domain-specificity of the language of many digital humanities corpora and the high levels of accuracy required to produce reliable analysis (Frank et al.; 2012, Hampson et al. 2013). Compiling sufficient training data to yield accurate results in this context is often costly and error prone. To address this, we explored an approach to automated classification that ensures high levels of accuracy with limited training data, while also incorporating domain knowledge and emulating the transparency and interpretability of a rule-based approach.

2.4 *Annotation Using Word Embedding and Semantic Text Classifiers*

The most relevant research on automated annotation pertains to identifying witness testimony. In the Ryan Report this information is represented as excerpts of reported speech. Our methodology therefore builds upon previous approaches to automatically identifying reported speech in text. This commonly relies on pattern-based extraction rules to detect linguistic markers such as quotation marks (Krestel et al., 2008; Pouliquen et al., 2007; Iosif et al., 2014). However, in the Ryan Report, punctuation

does signal speech, but that punctuation also signals other kinds of text so semantic information from the paragraphs has to be taken into consideration making research extracting indirect speech more relevant (Krestel et al., 2008).

Using machine learning, Schöch et al. (2016) developed an approach that involves semantic analysis using a lexicon of 81 linguistic features associated with direct speech derived from a corpus of French 18th century literature. These were used as features to train a classifier, yielding an accuracy of 84.4 percent. Weiser and Wartin (2012) developed a dictionary of verbs that introduce speech in text (reporting verbs) and used this in conjunction with pattern-based extraction rules to annotate indirect speech.

Machine learning approaches to text classification commonly use a bag-of-words approach to feature selection. However, this approach is problematic when instances to be classified are short giving rise to over-fitting (Brooks, 2013). A lexicon-based approach to feature selection can prevent address this but encounters new issues concerning the domain specificity of some corpora. Existing lexical databases such as WordNet (Miller, 1995) have been used to generate lists of synonyms from seed words to compile semantic lexicons (Argamon et al., 2007). However, they often do not recognise terms specific to particular domains such as the domain of ecclesiastical discourse used in the Ryan Report**.** Our project therefore required a methodology that used machine learning with lexicon-based features that take account of specific terms used in the Ryan Report.

In compiling domain-specific lexicons for feature extraction we called upon work by Mikolov (2013) who developed word2vec, a word-embedding algorithm. Word2vec is a set of neural network models that produces distributed representations of

words from text that reflect many aspects of their meanings. It implements the distributed semantics notion that the "meaning of a word can be determined by the company it keeps"[4]. This technique analyses word co-occurrence over large corpora representing a given word by a large vector of all the other words it is found beside it. Using these vectors one can then establish that two words are "similar" or synonymous by virtue of whether their vectors are the same or close in a multi-dimensional space. Mikolov's (2013) work provides a method for uncovering word-similarities that are tailored to the language of the Ryan Report.

This word-embedding technique was used by Chanen (2016) to identify synonyms and compile lexicons for feature extraction in order to account for the multiplicity of terms used to refer to the same semantic concept in a corpus of flight incident reports. Their method of identifying semantically related terms involved generating 20 word-2-vec ensembles, extracting terms that re-occurred over each ensemble and manually filtering antonyms and semantically dissimilar words. Given the domain-specificity of the language in the Ryan Report this approach suggests a useful way to compile lexicons that are specific to the language of the religious, industrial-school and legal worlds of the Ryan Report.

3. Distant Reading the Ryan Report: Methodology

A central aim of this project was to provide a methodology for identifying the semantic content of text in the Ryan Report and extracting given categories of information. The semantic categories identified included testimony of witnesses included in the Report (witness testimony), details of the transfer of clergy from school to school (transfer

[4] See also Latent Semantic Analysis, as a related technique, Dumais 2004; Landauer 2006; and similar methods in Turney & Pantel, 2010.

events) and descriptions of abusive (abuse events). Machine learning classification was used to annotate the text based on domain-specific semantic lexicons along with other features. In order to generate these domain-specific lexicons, word embedding was used to find terms in the Ryan Report that were semantically similar to a given set of seed words; this task can be cast as a type of query expansion or feature extraction.

These text-analytic methods for paragraph identification were extensions to our construction of a digital platform involving an exploratory web interface and database into which the significant parts of the Ryan Report were processed[5]. The core basic record in the database of this web-based system stored each paragraph from the relevant chapters in the Ryan Report. These paragraph records were then linked to other tables detailing actors in the Report (witnesses, clerics, officials), the Schools, the Congregations and time periods. Named actors were extracted using NLTK (Bird and Looper, 2002) and other information was identified using a rule-based approach. The web-interface also had a string-search facility for the paragraphs along with filters for other categories of entity (e.g., one could search on a single school or a diocese).

In the remainder of this section, we report on the other aspects of the methodology we developed to permit automated paragraph identification.

3.1 *Method*

The Corpus & Paragraph Categories

In the Ryan Report, 22 of the chapters detail events at each school. This dataset, comprising 6,839 paragraphs and 597,651 words was the corpus annotated according to its semantic content. Each paragraph is a definite unit-of-analysis in the Report, as they

[5] Digital platform was developed using the Django framework (https://djangoproject.com)

are systematically numbered and tend to focus on particular events and issues. The following are characteristic features of each semantic category:

3.2 Feature Extraction Techniques: Using Word Embedding

Using machine learning with lexicon-based features can address the issue of over-fitting of classification models when instances of text to be classified are short as is the case with paragraphs in the Ryan Report (see section 2.4). However, the language of the Ryan Report is domain-specific and general thesauri would not identify concepts such as "dispensation" as being synonymous with "dismissal". Hence, we used the word2vec algorithm (Mikolov, 2013) supplemented by synonyms generated from WordNet to find semantically related words from a set of seed-keywords building on the methodology outlined by Chanen (2016).

To compile the semantic lexicon five word-2-vec ensembles were generated from seed words. The top 30 words were extracted from each ensemble. A set of words common to each ensemble were identified and the results were then reviewed by a domain expert to validate their validity as synonyms within the context of the Ryan Report. Using this method many non-obvious synonyms were found. General synonyms were collected using the WordNet lexical database. This involved entering each seed word and compiling a list of synonyms from the results of a search in WordNet. The resulting lists were verified manually to ensure they were appropriate synonyms for the context of the Ryan Report.

Seed words were manually selected based on initial readings of the texts. In the case of paragraphs detailing transfers and direct speech, initial seed words were straight

forward to compile as terms such as ‘transfer’ and ‘said’ were commonly used in the report. However, in the case of descriptions of abuse, the language varied widely. A support vector machine-learning algorithm was used in this case to generate a classification model using 100 example paragraphs based on a bag-of-words feature set. Analysis of the support vectors highlighted words that best distinguished paragraphs describing abuse and the highest-ranking of these were used as seed words. The domain-specific semantic lexicons that resulted from this word embedding procedure are detailed in the next section.

3.3 Feature Extraction Techniques: Lexicon-Based Features

Domain-specific semantic lexicons were supplemented with other less domain-specific features. Verbs introducing reported speech, colons or quotation marks signal witness testimony in the Ryan Report. Punctuation such as commas, question marks and word contractions seemed to be used more frequently and testimony was also expressed in the first person. This information was therefore included as features to classify excerpts of direct speech in the Ryan Report (Table 1). A lexicon was also manually generated in order to filter out excerpts from written reports and letters. This lexicon included the terms: visitation, visitor, report, letter, wrote, written. ‘Visitations’ for instance is the term used for inspections of industrial schools carried out by the church. Various combinations of all features were examined to identify the optimal feature set.

Generally, the person being transferred is named in a paragraph detailing such an event. Similarly, in describing abuse, the perpetrator is commonly named. Names were therefore included as features for both of these semantic categories. In describing transfer events, the names of the institutions were often mentioned and the events

seemed often to be described in sections, which concerned abuse or named the alleged perpetrator. This information was therefore included as features for classification.

Semantic Category	Feature
Witness Testimony	Reporting Verbs: domain specific semantic lexicon Pronouns Punctuation
Transfer Events	Transfer Terms: domain specific semantic lexicon Section heading references to types of abuse Mentions of Religious actors Mentions of Institutions
Descriptions of Abuse	Abuse Terms: domain specific semantic lexicon Mentions of Religious Actors

Table 1: Feature Sets Extracted from the Ryan Report

3.4 Classifiers Used for Paragraph Identification

Separate classifiers were built for each of the paragraph categories. Using training data for each paragraph type, features were extracted based on the semantic lexicons generated using word embedding and WordNet to build feature vectors for each paragraph based on frequency counts. A random-forest classifier (Breiman, 2001) was then trained to find the relative weightings of features that predicted the content-class for given paragraphs using the Weka toolkit (Holmes et al., 1994). The random-forest algorithm was chosen because, as an ensemble learner that creates a 'forest' of decision trees by randomly sampling from the training set, it is well suited to learning from smaller datasets (Poliker, 2006).

3.5 Training Data

Sample paragraphs belonging to each paragraph category were manually selected from the Ryan Report as training data for classifiers. In order to address the issue of the cost

of compiling training data in digital humanities projects (Fran et al., 2012; Hampson et al., 2013), minimising the number of examples required was a guiding principle.

The training data consisted of 25 paragraphs detailing transfer events, 150 paragraphs containing direct speech and 100 paragraphs describing abuse. The vacience in numbers of training examples reflected the volume of instances in the report itself and the cost of compiling training data. Positive examples of each case were selected from across the report to capture the variety within the category. Negative examples were compiled through a random selection and manual verification of paragraphs.

3.6 Validation & Evaluation

Preliminary testing of the classifiers was done using 10-fold cross validation. These metrics indicated the most effective combination of features and were subsequently evaluated on a sample taken from a larger set of unseen data. The sample of unseen data was made up of 600 randomly selected paragraphs from the report. For transfer paragraphs, given the low number of training examples (25 positive and 50 negative), an interim evaluation stage was conducted by applying the classification model to a balanced set of 200 examples of unseen data from the report to further verify the optimal combination of features for classification.

4 Results & Discussion

The results showed that using word embedding to generate semantic lexicons for feature extraction is effective in yielding high accuracy where the language of a corpus is domain-specific and the volume of training data is limited. This allowed the integration of the semantic annotations in an online search tool for the report (Fig. 1). In the

following sections we outline the results of using word embedding to generate semantic lexicons and then report on effectiveness of the classifiers.

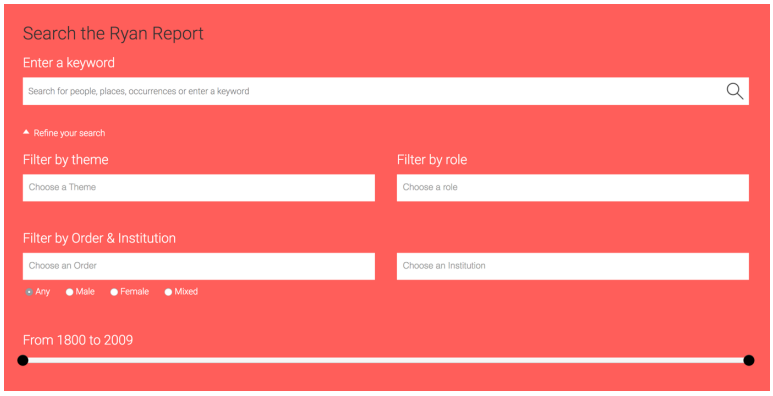

Figure 1: Search Interface for Ryan Report

4.1 *Domain-specific Semantic Lexicons: Using Word Embedding*

Semantic lexicons for each category of text were generated from an initial set of seed words derived from readings of the report. In the case of witness testimony, the seed words were the reporting verbs "said", "told" and "explained". A reading of the Report suggested some obvious key terms to describe the transfer of staff from school to school: "transfer", "dismiss" and "sack".

Seed words for abusive events were uncovered through analysis of the support vectors in a model generated by a support vector machine learning algorithm based on the words in a sample of 100 positive and negative paragraphs (details on this approach in section 3.2). This showed that terms distinguishing paragraphs describing abuse from the remainder of the report formed five semantic categories: perpetrator, abusive actions, body parts, emotions engendered in the victims and implements used in the abuse. The highest-ranking support vectors from these word types were selected as seed-words to form the semantic lexicon: abuse, beaten, raped, arms, humiliation, implement.

The word lists generated from running the word2vec algorithm on the full text of the Ryan Report are detailed in Table 2. This details the common terms among the top-30 words across 5 word-embedding ensembles generated for each seed word. After the manual verifications step, they were supplemented by general synonyms of each seed word generated from a search of the WordNet lexical database.

Text Category	Source	Feature
Witness Testimony	Seed terms	Said, told, explained
	Word embedding	Accepted, acknowledged, added, admitted, advised, agreed, alleged, angry, answered, asked, asking, asserted, assured, believed, called, claimed, commented, complained, conceded, concluded, confessed, confirmed, convinced, denied, describes, explained, explained, felt, guarantee, heard, informed, insisted, knew, described, learned, mentioned, presumed, protested, questioned, realised, recalled, recollection, recounted, relieved, remarked, remember, remembered, replied, requested, said, saw, saying, says, screams, stated, stating, suggested, surmised, tells, thinks, thought, told, warned, witnessed, reported
	WordNet	Apologise, apology, articulate, articulated, assure, assured, condone, condoned, enounced, enounce, explicate, explicated, express, expressed, narrate, narrated, pardon, pardoned, posit, posited, recite, recited, recount, recounted, said, state, stated, submit, submitted, tell, told, verbalise, verbalised
Transfer Events	Seed terms	Transfer, dismiss, sack
	Word embedding	Application, applied, apply, appointed, appointment, arrival, arrived, arriving, assigned, attended, committed, continued, converted, decision, departure, discharge, discharged, dismiss, dismissal, dismissed, dispensation, dispensed, entered, expelled, leaving, move, moved, position, posted, posting, proposal, referring, release, relieved, removal, remove, removed, replaced, request, resignation, resigned, returned, returning, sacked, sanction, seek, sending, sent, served, stayed, suspended, transferred, withdraw, withdrawal
	WordNet	Transferred, transfer, moved, remove, dismiss, dismissed, sacked
Descriptions of Abuse	Seed terms	Abuse, beaten, raped, arms, humiliation, implement
	Word embedding	Lexicons pertaining to parts of the body, abusive actions, emotion engendered in the victims and implements of abuse

Table 2: Domain-specific Semantic Lexicons

4.4Classifier Results

The results showed that the semantic lexicons generated using word embedding played a key role in producing accurate classifiers using limited training data. In classifying abuse paragraphs the words in the semantic lexicons were the sole features used. For transfer paragraphs, the semantic lexicon denoting transfer events featured in each of the combinations yielding the highest classification results. Results for classifying witness testimony were also highest when the semantic lexicon of reporting verbs were used as features. However, as was expected, features based on punctuation such as colons were also important in identifying this category of paragraph.

Classification: Witness Testimonies

In classifying paragraphs containing witness testimony, the model using a combination of all feature sets gained the highest accuracy in 10-fold cross-validation (Table 3). Most combinations of features were well-balanced between precision and recall.

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Reporting Verbs, Punctuation, Personal Pronouns	.928	.927	.927	93
Punctuation, Pronouns	.920	.920	.920	92
Punctuation, Reporting Verbs	.918	.917	.917	92
Pronouns	.914	.913	.913	91
Pronouns, Reporting verbs	.910	.910	.910	91
Punctuation	.882	.867	.865	86
Reporting Verbs	.823	.813	.812	81

Table 3: Results of 10-fold cross-validation for Witness Testimony Classification

The best performing model as indicated by the 10-fold cross validation was then was run on the remainder of the report. Based on a random sample of 600 paragraphs, an accuracy of 87 percent was achieved (Table 4).

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Reporting verbs, Writing, Punctuation, Personal Pronouns	.685	.766	.723	87

Table 4: Accuracy on sample of 600 Paragraphs for Witness Testimony Classification

Error analysis showed that false negative results were primarily due to in-text quotations of short-phrases. There were no instances of larger blocks of quotations being missed by the classifiers. The rate of false positives was relatively high primarily due to the misclassification of letters, extracts from inspection reports and diary entries. However, in many cases the source of such content can be challenging to decipher even on reading the report.

Classification: Transfer Events

When all features were included in the classifier to automatically detect paragraphs detailing the transfers of religious throughout the industrial school system, 88 percent accuracy was gained based on 10-fold cross validation (Table 5). However, when named entities were excluded as feature-sets, accuracy increased to 94%. This counter-intuitive result was verified further by applying the 3 best performing models to 200 unseen paragraphs consisting of a 50-50 balance between positive and negative examples. This showed that on a balanced set of unseen data, using all features yielded the best results (Table 6).

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Transfer terms, section heading info, mentions of school	.941	.940	.940	94
Transfer terms, section heading info	.941	.940	.940	94
Transfer terms, mentions of religious actors, section heading info, mentions of school	.882	.880	.880	88
Transfer terms, mentions of religious actors, section heading Info	.880	.880	.880	88
Transfer terms, mentions of religious actors	.865	.860	.859	86
Section heading info, mentions of school	.865	.860	.859	86
Transfer terms, mentions of religious actors, mentions of school	.842	.840	.840	84
Transfer terms, mentions of school	.825	.820	.819	82
Mentions of religious actors, section heading info	.821	.820	.820	82
Mentions of religious actors	.818	.800	.797	80
Transfer terms	.818	.800	.797	80
Mentions of religious actors, section heading info, mentions of School	.750	.740	.737	74
Mentions of religious actors, mentions of school	.744	.740	.739	74

Section heading info	.720	.720	.720	72
Mentions of school	.601	.600	.599	60

Table 5: Witness Testimony 10-Fold Cross Validation for Transfer Events

The text set of 200 sample paragraphs was comprised paragraphs that were distinctly positive and negative examples of transfer paragraphs. For this reason, a higher level of accuracy would be expected than on the rest of the report where language can often be more vague.

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Transfer Terms, Section Heading Info, Mentions of School	.937	.804	.865	89
Transfer Terms, Section Heading Info	.913	.816	.862	86
Transfer Terms, Mentions of Religious Actors, Section Heading Info, Mentions of School	.966	.832	.894	90

Table 6: Witness Testimony Accuracy on Balanced Sample of 200 Paragraphs

The final phase of evaluation for paragraphs pertaining to transfer events involved application of the best performing model, from the results of the balanced set of 200 paragraphs, to the remainder of the report and manually examining the classification of 600 randomly sampled paragraphs (Table 7). These results showed high levels of recall. However, there were quite a few false positive results leading to relatively low levels of precision.

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Transfer Terms, Section Heading Info, Mentions of School	.514	.900	.655	93

Table 7: Witness Testimony Accuracy on Random Sample of 600 Paragraphs

Error analysis showed that paragraphs that were falsely categorised as being about the transfer of clergy actually pertained to the transfer of children. However, some false positive results raised potentially new questions regarding the transfer of children throughout the industrial school system as a response of the congregations to allegations of abuse:

9.135 The boy at the centre of this allegation was transferred to another industrial school early the following year.

(CICA Vol. 1.9)

Classification: Abuse Events

The best performing model for identifying paragraphs describing abuse in 10-fold cross validation used two of the semantic categories along with the names of the alleged perpetrator (Table 8). The domain-specific semantic lexicons that were most useful included references to the emotions engendered in the victims and references to abusive actions.

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Action, emotion, mentions of religious actors	.958	.958	.958	95.7
Emotion, implement, action, mentions of religious actors	.953	.953	.953	95.3
Action, implement, mentions of religious actors	.953	.953	.953	95.2
Action, implement, emotion	.948	.948	.948	94.8
Implement, emotion, mentions of religious actors	.948	.948	.948	94.8
Emotion, implement, body, action, mentions of religious actors	.943	.943	.943	94.3
Body, action, emotion, mentions of religious actors	.939	.939	.939	93.8
Body, action, implement, mentions of religious actors	.934	.934	.934	93.4
Body, Implement, mentions of religious actors	.906	.906	.906	90.5
Body, action, implement	.904	.901	.901	90.1
Actor, body, emotion	.901	.901	.901	90.0
Emotion, implement, body, action	.884	.882	.882	88.2
Body, implement, emotion	.816	.816	.816	81.6
Body, action, mentions of religious actors	.939	.939	.939	93.8
Body, action, emotion	.881	.877	.877	87.0

Table 1: Results of 10-fold cross-validation for Abuse Events

The classification model was then run on the reminder of the report and a random sample of 600 paragraphs was manually verified yielding overall accuracies of 82 percent (Table 9).

Feature Sets	***Precision***	***Recall***	***F-Measure***	***Accuracy (%)***
Action, emotion, mentions of religious actors	.395	.816	.532	81.8

Table 9: Descriptions of Abuse Tested on Random Sample of 600 Paragraphs

While precision was reduced reflecting the complexity of the language, recall was high. Error analysis showed that false positives uncovered a similarity in the language used to describe the emotional experience of victims of abuse and some memories of young clergy when they first took up positions in the schools.

5 Conclusions

This research demonstrates how distant reading methodologies can deconstruct an official state report narrative to enable new kinds of analysis of institutional child abuse. Automatic annotation of excerpts of the report based on the meaning of the text enabled a more focussed close reading of these identified paragraphs, surfacing significant new patterns of events and language in the institutional system (Pine et al., 2017). These insights were previously obscured by the legal constraints on and narrative form of the Ryan Report, which emphasised an in-depth case-by-case study, in lieu of system-wide analysis.

The feasibility of using machine learning to annotate text for digital humanities projects can be enhanced by using word embedding for feature extraction. The cost of compiling training data and the domain specificity of the text of many projects can often be a barrier to using machine learning approaches to annotation. This research demonstrates how word embedding can be used to compile context-specific semantic lexicons as a method for extracting features for text classifiers to perform automated annotations of text. This is an innovative methodology building on an approach outlined by Chanen (2016). High accuracy was achieved using a minimal set of training examples with features based on semantic lexicons generated from the entire dataset.

There have been numerous international state investigations into the abuse of children. Wright et al. (2017) documented 40 historical child abuse enquiries to date each of which resulted in lengthy reports detailing their findings. In using automated methods to enable distant reading of the Ryan Report, this project presents an approach whereby key information may be extracted and restructured to facilitate a system-wide analysis of the findings of such investigations.

Acknowledgements

This research is part of the Industrial Memories project funded by the Irish Research Council under New Horizons 2015.

References

Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N. and Levitan, S., 2007. Stylistic text classification using functional lexical features. Journal of the Association for Information Science and Technology, 58(6), pp.802-822.

Bird, S., Klein, E., & Loper, E., 2009. Natural language processing with Python. Beijing, O'Reilly Media. Sebastopol, CA.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.
Vancouver

Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P. and Aragon, C.R., 2013, February. Statistical affect detection in collaborative chat. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 317-328). ACM.

Chanen, A., 2016, April. Deep learning for extracting word-level meaning from safety report narratives. In Integrated Communications Navigation and Surveillance (ICNS), 2016 (pp. 5D2-1). IEEE.

Chiticariu, L., Li, Y. and Reiss, F.R., 2013, October. Rule-based information extraction is dead! long live rule-based information extraction systems!. In EMNLP (No. October, pp. 827-832).

Django Software Foundation, 2016. Django (V ersion 1.9.6) [Computer Software]. Retrieved from https://djangoproject.com.

Donnelly, S. and Inglis, T., 2010. The media and the Catholic Church in Ireland: Reporting clerical child sex abuse. Journal of Contemporary Religion, 25(1), pp.1-19.

Dumais, S.T., 2004. Latent semantic analysis. Annual review of information science and technology, 38(1), pp.188-230.

Frank, A., Bögel, T., Hellwig, O. and Reiter, N., 2012. Semantic annotation for the digital humanities. Linguistic Issues in Language Technology, 7(1), pp.1-21.

Hampson, C., Munnelly, G., Bailey, E., Lawless, S. and Conlan, O., 2013, September. Improving User Control and Transparency in the Digital Humanities. In Culture and Computing (Culture Computing), 2013 International Conference on (pp. 196-197). IEEE.

Hogan, R. and Emler, N.P., 1981. Retributive justice. The justice motive in social behavior: Adapting to times of scarcity and change, pp.125-143.

Holmes, G., Donkin, A. and Witten, I. H. (1994), Weka: A machine learning workbench, in 'Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on', IEEE, pp. 357–361.

Iosif, E. and Mishra, T., 2014, April. From Speaker Identification to Affective Analysis: A Multi-Step System for Analyzing Children's Stories. In CLfL@ EACL (pp. 40-49).

Jackson, H.J., 2002. Marginalia: Readers writing in books. Yale University Press.

Jockers, M.L. and Mimno, D., 2013. Significant themes in 19th-century literature. Poetics, 41(6), pp.750-769.

Keenan, M., 2013. Child sexual abuse and the Catholic Church: Gender, power, and organizational culture. Oxford University Press.

Krestel, R., Bergler, S. and Witte, R., 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. Reporter, 1(5), p.4.

Landauer, T.K., 2006. Latent semantic analysis. John Wiley & Sons, Ltd.

Mahlow, C., Grün, C., Holupirek, A. and Scholl, M.H., 2012, September. A framework for retrieval and annotation in digital humanities using XQuery full text and update in BaseX. In Proceedings of the 2012 ACM symposium on Document engineering (pp. 195-204). ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp.39-41.

Muralidharan, A.S. and Hearst, M.A., 2014. Improving the Recognizability of Syntactic Relations Using Contextualized Examples. In ACL (2) (pp. 272-277).

Pilgrim, D., 2012. Child abuse in irish catholic settings: A non-reductionist account. Child Abuse Review, 21(6), pp.405-413.

Pine, E., Leavy, S. and Keane, M.T., 2017. Re-reading the Ryan Report: Witnessing via and Close and Distant Reading. Éire-Ireland, 52(1), pp.198-215.

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circuits and systems magazine, 6(3), pp.21-45.

Pouliquen, B., Steinberger, R. and Best, C., 2007, September. Automatic detection of quotations in multilingual news. In Proceedings of Recent Advances in Natural Language Processing (pp. 487-492).

Powell, F., Geoghegan, M., Scanlon, M. and Swirak, K., 2012. The Irish charity myth, child abuse and human rights: Contextualising the Ryan report into care institutions. British Journal of Social Work, 43(1), pp.7-23.

Ryan, S. (2009). Commission to inquire into child abuse report (Volumes I - V). Dublin: Stationery Office, Dublin. Available at: http://www.childabusecommission.ie/rpt/.

Schöch, C., 2013. Big? smart? clean? messy? Data in the humanities. Journal of Digital Humanities, 2(3), pp.2-13.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U. and Tello, J.C., 2016. Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. In Digital Humanities 2016: Conference Abstracts (pp. 346-353).

Sweetnam, M.S. and Fennell, B.A., 2011. Natural language processing and early-modern dirty data: applying IBM Languageware to the 1641 depositions. Literary and linguistic computing, 27(1), pp.39-54.

Vancouver

Turney, P.D. and Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, pp.141-188.

Vuillemot, R., Clement, T., Plaisant, C. and Kumar, A., 2009, October. What's being said near "Martha"? Exploring name entities in literary text collections. In Visual

Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on (pp. 107-114). IEEE.

Weiser, S. and Watrin, P., 2012. Extraction of unmarked quotations in newspapers. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012).

Widlocher, A., Bechet, N., Lecarpentier, J.M., Mathet, Y. and Roger, J., 2015, September. Combining Advanced Information Retrieval and Text-Mining for Digital Humanities. In Proceedings of the 2015 ACM Symposium on Document Engineering (pp. 157-166). ACM.

Wright, K., Swain, S., and Sköld, J. (2017). 'The Age of Inquiry: A global mapping of institutional abuse inquiries'. Melbourne: La Trobe University. DOI: http://doi.org/10.4225/22/591e1e3a36139