Knowledge Extraction From a Small Corpus of Unstructured Safeguarding Reports*

Aleksandra Edwards[1,2], Alun Preece[1,2], and Hélène de Ribaupierre[1]

[1] School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[2] Crime and Security Research Institute,Cardiff University, Cardiff, UK

Abstract. This paper presents results on the performance of a range of analysis tools for extracting entities and sentiments from a small corpus of unstructured, safeguarding reports. We use sentiment analysis to identify strongly positive and strongly negative segments in an attempt to attribute patterns on the sentiments extracted to specific entities. We use entity extraction for identifying key entities. We evaluate tool performance against non-specialist human annotators. An initial study comparing the inter-human agreement against inter-machine agreement shows higher overall scores from human annotators than software tools. However, the degree of consensus between the human annotators for entity extraction is lower than expected which suggests a need for trained annotators. For sentiment analysis the annotators reached a higher agreement for annotating descriptive sentences compared to reflective sentences, while inter-tool agreement was similarly low for the two sentence types. The poor performance of the entity extraction and sentiment analysis approaches point to the need for domain-specific approaches for knowledge extraction on these kinds of document. However, there is currently a lack of pre-existing ontologies in the safeguarding domain. Thus, in future our focus is the development of such a domain-specific ontology.

Keywords: text mining · sentiment analysis · entity extraction.

1 Introduction

The aim of this paper is to evaluate a range of text analysis tools and approaches for extracting knowledge from a small corpus of unstructured safeguarding reports. The purpose of these reports is to describe events prior to a crime, assess agencies and practices, and to reflect on lessons learned. The reports are lengthy and complex, so extracting information across the corpus by human inspection is a time-consuming and potentially bias-prone process. Moreover, the documents are unstructured and contain a great deal of domain-specific terminology which makes them hard to analyse using automated methods. In an initial attempt at information retrieval we performed a sentiment analysis exercise hypothesising that strongly negative or positive sentences would contain key information.

* We thank David Rogers and members of the Wales Safeguarding Repository research team for their assistance. http://orca.cf.ac.uk/111010/

We also used entity extraction to identify key features in terms of individuals, organisations and locations. Further to this, we conducted a study comparing an 'inter-judgement agreement' between annotators and tools. Table 1 provides description for the text analysis tools.

Table 1. Text Analysis Tools

Tool	Sentiment Analysis	Named entity extraction
Stanford Core NLP [2]	Recursive Neural Tensor Network	CRF classifier
Google Cloud API[1]	Deep learning models	Deep learning models
Gate [1]	Generic Sentiment Analysis application	ANNIE dictionary look-up and rules
SentiStrength [3]	dictionary look-up	NA
NLTK[2]	NA	MaxEnt chunker

2 Results

Inter-human agreement versus inter-machine agreement In this pilot study, conducted by the authors of this paper, we manually annotated sentences that were used as a baseline for comparing tool performance. The study used a *description* and a *reflection* set. Both sets consisted of 100 randomly-chosen sentences from different parts of the reports. The description set consisted of sentences describing the events of the safeguarding case — often involving one or more crimes — while the reflection set consisted of findings: lessons learned and recommendations. The two sets differed in the nature of how the sentiments of the sentences can be interpreted. The highlights of the descriptive set are the events; thus, the sentiment of the sentences will be judged by the sentiment of the event. An indicative (non-verbatim) example of a descriptive sentence is: "Prison staff found the subject had hanged himself". This sentence describes a negative event, i.e., a death. The highlights of the reflection sentences are the findings. Thus, the sentiment of the sentences will be judged by the sentiment of the comment. An indicative (non-verbatim) example of a reflective sentence is: "The key finding from the review of the agencies involvement is that there was strong evidence of good inter-agency working and appropriate referrals between local services". This sentence express a positive reflection on inter-agency communication.

We measured the inter-annotator agreement and the inter-tool agreement for our sentiment analysis and entity extraction exercises (Table 2) using Fleiss' kappa. Fleiss kappa scores for the sentiment analysis showed good agreement between the annotators but a significant disagreement between the tools.

[1] Google Cloud API: https://cloud.google.com/apis/
[2] NLTK: https://www.nltk.org

The difference between the annotator scores for the two datasets suggests that humans find it easier to annotate the descriptive set rather than the reflective set while the tools did not differentiate between the two data sets. The vast majority of sentiment disagreement between the human annotators involved distinguishing between neutral vs positive/negative polarity. There was only a single instance of disagreement between positive vs negative polarity of a sentence: "The person disclosed at an appointment, that they had overdosed a month before and now felt stupid about it" (this example is paraphrased). However, the disagreement between the tools in terms of positive vs negative sentiment was considerably higher: 34% for the description and 36% for the reflections.

The Fleiss' kappa scores are low across all entity extraction categories for the software tools. Inter-human agreement for person and organisation are also low. Entities that humans disagreed on were: 'GP' (general practitioner), 'Coroner', 'Mental Health Teams', and 'Mental Health Tribunal', all of which tended to be labelled either as 'person' or 'organisation'. The low Fleiss' kappa scores show that the entity extraction task is challenging not only for software but also for non-specialist human annotators.

Table 2. Fleiss' kappa scores

	Annotators	Tools
Sentiment analysis		
Descriptions	0.6	0.1
Recommendations	0.4	0.1
Entity extraction		
Person	0.3	0.04
Organisation	0.3	0.3
Location	1.0	0.2

Sentiment Analysis Figure 1 presents the average precision, recall, and F1 between the positive, negative, and neutral sentiment categories. These results show an unsatisfactory level of performance from the tools used. Overall, the tools performed better for descriptive sentences: SentiStrength performed the best for these with around 55% for precision, recall and F1. Gate performed best for reflective sentences with F1 of 48%. The poor performance of the tools can be attributed to the fact that they are trained on datasets very different to the safeguarding domain. For example, Stanford [2] is trained on movie reviews where a phrase such as "with recommendation" has positive sentiment while the same phrase in the context of a safeguarding report might have a negative sentiment (e.g., "sentenced to life imprisonment *with recommendation* of years"). SentiStrength [3] is based on a dataset of MySpace content and uses a dictionary-based approach. It follows that sentences mentioning entities such as 'Specialist Dementia home' will match to the term dictionary 'special*' and thus have a positive sentiment.

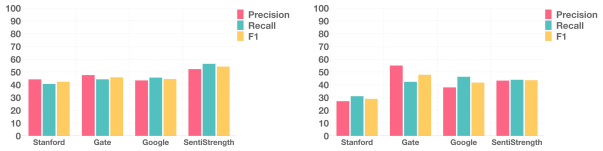

Fig. 1. Average precision, recall, F1 for sentiment analysis: description set (left), reflection set (right)

Extracting Named Entities Figure 2 shows poor performance across all categories with F1 lower than 60%. Precision and recall tend to be very unbalanced.

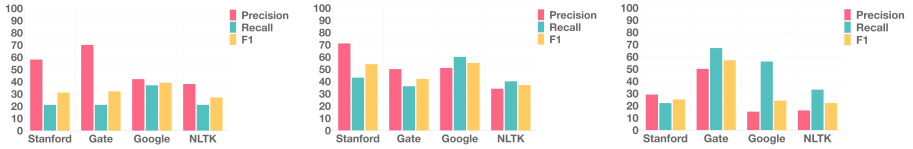

Fig. 2. Evaluation results for entity extraction: person (left), organisation (middle), location (right)

3 Conclusion and Future work

The sentiment analysis results provide no evidence that off-the-shelf sentiment analysis tools can identify key parts of the safeguarding reports. In future we plan to focus on entity extraction. The unsatisfactory results of the entity extraction tools shows the need for more domain-targeted approaches, and for knowledge extraction such as the use of an ontology. However, existing semantic web resources to the best of our knowledge lacks ontologies relating to safeguarding and crime issues. A Swoogle search (conducted 1/3/19) on terms such as 'crime', 'safeguarding', and 'mental health' found no publicly-available ontologies fitting the purpose of our domain, pointing to a need for the creation of an ontology that models safeguarding issues. In the next stage of our work, we will use word and sentence vectors to discover main themes in documents. We will then use these themes as a base for the creation of an ontology.

References

1. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting more out of biomedical documents with gate's full lifecycle open source text analytics. PLoS computational biology **9**(2), e1002854 (2013)
2. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
3. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12), 2544–2558 (2010)