Using Entities in Knowledge Graph Hierarchies to Classify Sensitive Information

Erlend Frayling, Craig Macdonald, Graham McDonald, and Iadh Ounis,
firstname.lastname@glasgow.ac.uk

Univerity of Glasgow, Glasgow G12 8QQ, UK

Abstract. Text classification has been shown to be effective for assisting human reviewers to identify sensitive information when reviewing documents to release to the public. However, automatically classifying sensitive information is difficult, since sensitivity is often due to contextual knowledge that must be inferred from the text. For example, the mention of a specific named entity is unlikely to provide enough context to automatically know if the information is sensitive. However, knowing the conceptual role of the entity, e.g. if the entity is a politician or a terrorist, can provide useful additional contextual information. Human sensitivity reviewers use their prior knowledge of such contextual information when making sensitivity judgements. However, statistical or contextualized classifiers cannot easily resolve these cases from the text alone. In this paper, we propose a feature extraction method that models entities in a hierarchical structure, based on the underlying structure of Wikipedia, to generate a more informative representation of entities and their roles. Our experiments, on a test collection containing real-world sensitivities, show that our proposed approach results in a significant improvement in sensitivity classification performance (2.2% BAC, McNemar's Test, $p < 0.05$) compared to a text based sensitivity classifier.

1 Introduction

Technology Assisted Review (TAR) [2] has been shown to improve the efficiency of government sensitivity reviewing processes through use of text classifiers to recognise sensitivities, as the classifiers can assist reviewers with predictions as to whether documents contain sensitivity or not [11]. However, training a classifier to predict sensitivities is a complex task. Sensitivity identification is not a topic-oriented task [1], and sensitivity itself can arise from factors that are implicit to the text and are not exposed in an individual textual term. Indeed, sensitivity, like the background knowledge of the concepts and entities mentioned in documents, can be latent to the text. An expert human reviewer's prior knowledge enables them to deduce latent sensitivities using their knowledge of the subject matter. On the other hand, text classifiers that are trained using the distributions of terms in the text [12], or even those trained with contextualised embeddings [5], are limited to learning from the distributions of textual features and, as such, may fail to identify latent sensitivities (even contextualised language models such as BERT [5] do not experience sensitive data).

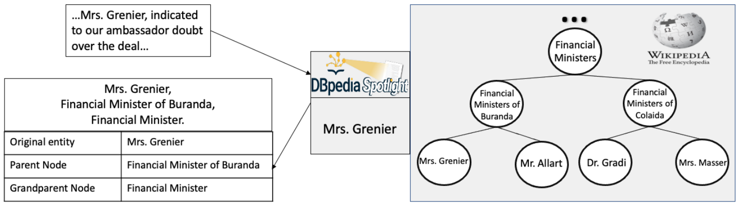
Entities such as people, places or organisations are a rich source of latent contextual information. In this work, we propose a sensitivity classification approach that aims to integrate information that is representative of what a human reviewer might possess through their prior knowledge. For example, a reviewer might know that two entities are both political leaders, and that they represent opposing political parties - a subtlety that a contextualised classifier model may not so easily pick up. Sensitivity can often be nuanced in this way. For example, in a ‘who said what about who’ situation, the specifics of ‘who’ can matter more than the ‘what’ [10] - hence, recognising that the ‘who’ are both political entities might be informative for (sensitivity) classification. To this end, we propose a novel approach to build a hierarchical relationship model of entities present in a collection of government documents, using the underlying hierarchical structure of Wikipedia. We use this structure to infer latent information about entities in documents for classification. Specifically, we attempt to identify how certain entities in documents are related by underlying hierarchical concepts; For example, that two identified politicians, though different entities by name, are both leaders of communist regimes. Experiments conducted on a collection of 1000 real government documents with actual sensitivities demonstrate that we can attain significant improvements in accuracy of sensitivity classification.

2 Related Work

Several techniques have been proposed by MacDonald et al. to improve sensitivity classification performance, including using Part of Speech (PoS) tagging and semantic word embedding features [9, 10]. To our knowledge, there has been little work concerning the central importance of entities for classifying sensitivity. In the closest work to our own, [12], the authors feature engineered an opinionated numerical score representing diplomatic risk associated with some countries mentioned in the text, from the perspective of the UK. There have been several attempts to improve models in the more general category of text classification machine learning by enhancing entity representations. E-BERT [14] is a good example, which modified the original BERT model [5] to handle entities as unique tokens and unique vector representations showed improved performance over the original model. However, it is not yet clear if the entity representation within models such as E-BERT can learn to reflect well the genericism/specialism structure that can be encapsulated in knowledge bases.

Indeed, the use of knowledge bases in classification is most prevalent in domains where substantial and specialised knowledge bases already exist, e.g. in biomedicine. One work [8] utilised a pre-existing hierarchical knowledge graph of symptoms and diseases to learn a graph convolution neural network, which improved the effectiveness of medical diagnosis. BLUEBERT [13], which follows the BERT [5] architecture, was pre-trained on abstracts from the PubMed knowledge base. This model was designed to perform the Biomedical Language Understanding Evaluation (BLUE) benchmark [13] and showed improved performance in BLUE tasks over a model pre-trained on more general datasets.

Training on specific knowledge bases for specific tasks has shown significant performance benefits versus training on general knowledge bases [13]. However,


...Mrs. Grenier, indicated to our ambassador doubt over the deal...
Mrs. Grenier,
Financial Minister of Buranda,
Financial Minister.
DBpedia Spotlight
Mrs. Grenier
Financial Ministers
Financial Ministers of Buranda
Financial Ministers of Colaida
Mrs. Grenier
Mr. Allart
Dr. Gradi
Mrs. Masser
WIKIPEDIA
The Free Encyclopedia

Original entity	Mrs. Grenier
Parent Node	Financial Minister of Buranda
Grandparent Node	Financial Minister

Fig. 1. Process of identifying entities in text & enriching with hierarchical tree entities.

in sensitivity review, we lack a publicly available knowledge base structure to use for training models. Therefore, we focus on using a general knowledge base in this work. Notably, our focus is on enhancing representations of entities in sensitive documents using knowledge base information for two reasons. Firstly, as discussed in Section 1, we hypothesise that entities are a rich source of latent information and, in some cases, sensitivity in documents. Secondly, because knowledge bases provide information about entities, this kind of information is the most easily accessible. Flisar et al. [6] applied the DBpedia semantic knowledge base, by using the DBpedia Spotlight [3] to identify DBpedia entities, and then modelling them as key concepts in short texts for classification.

DBpedia, as a semantic knowledge base that has been derived from Wikipedia encodes a plethora of semantic relationships between concepts and links to the wider semantic web. On the other hand, Kapanipathi et al. [3] used the simpler structure of Wikipedia's Category Graph (WCG), which aims to group similar Wikipedia pages in a hierarchical relationship, and hence is a self-contained knowledge structure. Instead, we propose a simpler Wikipedia-based knowledge graph, instantiated from the Wikipedia pages themselves. In the next section, we introduce our model for hierarchical modelling of entities for sensitivity review.

3 Hierarchy modelling and features

We aim to enrich the representation of documents with additional entities that can assist classifiers in identifying sensitive text, by allowing the inference of more general sensitivity rules - for instance, rather than a person's name said something, the classifier can learn that a minister in a foreign government said something, which may be more significant. To this end, we derive a knowledge base that allows to generalise from linked entities.

Figure 1 provides an overview of our approach - a sentence about "Mrs Grenier" is indicating something is being told to an ambassador in confidence (and hence may be sensitive, due to a need to preserve international relations), but a classifier that is aware that Mrs Grenier is a finance minister in "Buranda" may help that classifier to learn more generalisable classifier rules. In the following, we describe both how we build a hierarchical concept tree from Wikipedia, and also how these more general concepts are encoded into the feature representations.

3.1 Building Hierarchy Tree

The articles of Wikipedia are organised in a loose hierarchical structure, separate from that exhibited by WCG. The central principle of this alternate structure is that, for any Wikipedia article, clicking on the first linked article in the text, recursively, will, in most cases, eventually bring the user to the article for Philosophy. This forms a tree structure over the nodes (or articles), where more abstract Wikipedia articles like Science and Rational are intermediate nodes close to the root node Philosophy. More specific entities like countries and people are farther from the root.

On the other hand, while WCG has hierarchical properties, it is not fundamentally a tree structure, as each Wikipedia page can have multiple categories. From our experience in this work, the "first-link" observation creates a usable tree with the desired properties, which we call Philosophy Hierarchy Structure (PHS).

Following [6, 7], we use a Named Entity Linking tool to identify entities in documents, before generating classification features to avoid building the entire tree structure available in PHS. Indeed, in our task, we are not concerned with knowing all entities in the hierarchy, just those presently identifiable in the documents being reviewed, and the entities in their path to Philosophy. Moreover, as there are more than 6 million articles on Wikipedia, building the entire tree structure would be unnecessarily cumbersome. Therefore, we build only a local tree.

More specifically, we use the DBpedia Spotlight [3] NEL to identify all unique people, places and organisation entities. Spotlight provides a disambiguated link to the Wikipedia page for each detected entity in the document collection, which we use to retrieve the article's content. We retrieve the first link to the next (parent) article from that content. We consider this initial set of detected entities as the set of leaf nodes in our tree structure. We iteratively retrieve parent nodes for all Wikipedia articles in the initial leaf set, then for the intermediate nodes. We stop when all branches reach the Philosophy Node. In reality, the tree structure has imperfections – when creating a branch three outcomes are possible: (i) A generated branch reaches the node for Philosophy correctly, and the recursive parsing cycle is stopped; (ii) A branch of nodes forms a loop where one node in the branch points to a node further down; (iii) The branch breaks when the upper-most node cannot be parsed to obtain the next node. However, imperfect branches still contain the hierarchical information we need about entities present in the document collection and can still be used.

3.2 Feature development

Having described the production of a tree structure object, we now describe our approach to extracting features from this tree. Key to our hypothesis discussed in Section 1, we argue that certain entities in documents sharing parent nodes in the tree represents a hierarchical relationship that could be useful for classification. We identify and model these relationships for entities in a collection of documents as text features in our approach.

To generate features for a given document we find the associated set of DBpedia entities present in the text and their corresponding nodes in the tree produced in Section 3.1. For each node in the tree we identify the next N parent nodes, where N is some integer number of nodes to climb into the tree. We combine the original set of DBpedia entities for each document with the additional parent nodes to form a new extended set of entities. We expect that across a corpus of documents, parent nodes will appear in documents for which the detected

Table 1. Results from experiment on 1000 record collection. Significant improvements over the text-only baseline classifier are denoted with * (McNemar's test, $p < 0.05$).

Features	P	R	F1	BAC
↓ $N \setminus$ baseline →	0.363	0.657	0.468	0.636
0	0.369	0.661	0.474	0.641
1	0.371	0.661	0.476	0.643
2	0.370	0.657	0.473	0.641
3	0.369	0.665	0.475	0.642
4	0.373	**0.669**	0.479	0.646
5 *	**0.378**	**0.669**	**0.483**	**0.650**
6	0.374	0.665	0.479	0.646
7	0.374	**0.669**	0.480	0.647
8	0.373	**0.669**	0.479	0.646
9	0.372	0.665	0.477	0.644

DBpedia entities are different, revealing that the different entities have underlying connections represented by the parent nodes in the extended collection of entities. This extended set of entities for each document can be used as additional features in a classification task. For example, referring to Figure 1, if Mrs. Grenier retires from her position as financial minister, and a new individual (Mr Allart) takes over, the surface form name of the individual will change in newer documents. However, using the extended features would still provide the common connection of 'Financial Ministers of Buranda'. In this sense, generalisation is achieved, and a classifier may make the connection that both Mr. Allart and Mrs. Grenier share equal importance across old and new documents.

4 Experiments

We perform experiments to address two research questions, namely:
RQ1: Can a text classifier use our hierarchically enriched entity features to predict sensitivity in government documents more accurately?
RQ2: Does changing the number of added parent nodes N of the hierarchically enriched features, detailed in Section 3.2, affect classification effectiveness, and which number N is most effective in this task?

4.1 Experimental Setup

We use a collection of 1000 government documents that have been reviewed for sensitivity by experienced government reviewers. The data collection was assessed for sensitivities relating to international relations and personal information, which are common types of sensitivities defined in freedom of information settings. The collection contains 251 (25%) sensitive documents in total, across both categories of sensitivity assessed. We use a 10-fold cross-validation setup, averaging Precision (P), Recall (R), Balanced Accuracy Score (BAC) and F1 measure across folds.

We generate a hierarchical relationship tree using the process described in Section 3.1. DBpedia Spotlight detects 2226 entities in the collection, and the total number of nodes in the final tree structure is 5129. We extract several feature sets for each document. Firstly, the text of each document alone. Secondly, we extract a set of entities directly detected by DBpedia's Spotlight tool for each document (denoted $N = 0$). Further, we use this entity set for each document to feature engineer hierarchically-enhanced representations for tree depth values of $1 \leq N \leq 9$, as described in Section 3.2. We use the original entity set and all nine hierarchically enhanced sets as ten separate feature sets. Finally, we combine each of the ten sets of entities with the original text of each document to produce combined text and entity features. For classification, we apply a Multinomial Naive Bayes model. Words are represented using term frequency only, removing stopwords that occur in the Sci-Kit Learn's English stopword list.

4.2 Results

Table 1 presents the effectiveness of the Multinomial Naive Bayes classifier using different combinations of features. The table presents effectiveness in terms of Precision, Recall, F1, and Balance Accuracy for each configuration. We also test each configuration for statistical significance ($p < 0.05$) compared to the baseline that classifies documents on only their text features (denoted 'text'). Firstly, from Table 1, we note that all sets of entity features improve classifier performance when combined with the text features. The best performance increase w.r.t. the baseline occurs when classifying document text with our entity features when considering a hierarchy depth (N) of 5. This result is a 2.2% improvement in BAC score over the baseline and a 3.2% improvement in F1, which is statistically significant according to a McNemar's test ($p < 0.05$). Moreover, all experiments combining text features with entity sets outperform the baseline of BAC 0.636. Among precision and recall, we note that precision is enhanced by 4% (0.353→0.378), while recall is enhanced by 2% (0.657→0.669). Indeed, in an assistive classification task such as sensitivity review, precision is important, as false positive may cause reviewers to loose confidence in the predictions.

Thus, we answer our research questions as follows: for **RQ1**, we find that entity features making use of the PHS hierarchy can be used to identify sensitivities more accurately when used in addition to the textual features of the documents. For **RQ2**, we find that adding five levels of parent nodes to the enriched set of entities for each original entity occurring in text achieves the best performance, but all $N > 0$ outperform adding only the original entities.

5 Conclusions

In this work, we proposed a novel approach to provide a sensitivity classifier with a hierarchical representation of entities that allows a classifier to infer new generalised rules about entities and sensitivity. Moreover, we evaluated the effectiveness of our features for sensitivity classification and showed that our enhanced entity features allow a classifier to make more successful predictions about sensitivities. We showed that significant improvements can be obtained compared to a baseline text classification approach (McNemar's test, $p < 0.05$), particularly improving precision. In future work, we will apply Graph Neural Networks in conjunction with the hierarchical graph structures, which we expect to result in further classification improvements.

Acknowledgements

E. Frayling, C. Macdonald and I. Ounis acknowledge the support of Innovate UK through a Knowledge Transfer Partnership (# 12040). All authors thank SVGC Ltd. for their support.

References

1. Giacomo Berardi, Andrea Esuli, Craig Macdonald, Iadh Ounis, and Fabrizio Sebastiani. 2015. Semi-automated text classification for sensitivity identification. In Proc. of CIKM.
2. Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In Proc. of SIGIR
3. Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In Proc. of I-SEMANTICS
4. Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In Proc. of I-SEMANTICS
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Jernej Flisar and Vili Podgorelec. 2020. Improving short text classification using information from DBpedia ontology. Fundamenta Informaticae 172, 3 (2020), 261–297.
7. Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. 2014. User interests identification on Twitter using a hierarchical knowledge base. In Proc. of ESWC.
8. Bing Liu, Guido Zuccon, Wen Hua, and Weitong Chen. 2021. Diagnosis Ranking with Knowledge Graph Convolutional Networks. In Proc. of ECIR.
9. Graham McDonald, Craig Macdonald, and Iadh Ounis. 2015. Using part-of-speech n-grams for sensitive-text classification. In Proc. of ICTIR
10. Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing sensitivity classification with semantic features using word embeddings. In Proc. of ECIR.
11. Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Towards maximising openness in digital sensitivity review using reviewing time predictions. In Proc. of ECIR
12. Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. Towards a classifier for digital sensitivity review. In Proc. of ECIR.
13. Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proc. of BioNLP Workshop and Shared Task.
14. Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. E-BERT: Efficient-yet-effective entity embeddings for BERT. arXiv preprint arXiv:1911.03681 (2019).