Supporting Complaints Investigation for Nursing and Midwifery Regulatory Agencies

Piyawat Lertvittayakumjorn[*†], **Ivan Petej**[*], **Yang Gao**[*], **Yamuna Krishnamurthy**[*], **Anna van der Gaag**[*◇], **Robert Jago**[*], **Kostas Stathis**[*]
∗ Royal Holloway, University of London, United Kingdom
† Imperial College London, United Kingdom
◇ University of Surrey, United Kingdom
{piyawat.lertvittayakumjorn, i.petej, yang.gao}@rhul.ac.uk
yamuna.k.2018@live.rhul.ac.uk,
{anna.vandergaag, robert.jago, kostas.stathis}@rhul.ac.uk

Abstract

Health professional regulators aim to protect the health and well-being of patients and the public by setting standards for scrutinising and overseeing the training and conduct of health and care professionals. A major task of such regulators is the investigation of *complaints* against practitioners. However, processing a complaint often lasts several months and is particularly costly. Hence, we worked with international regulators from different countries (the UK, US and Australia), to develop the first decision support tool that aims to help such regulators process complaints more efficiently. Our system uses state-of-the-art machine learning and natural language processing techniques to process complaints and predict their risk level. Our tool also provides additional useful information including *explanations*, to help the regulatory staff interpret the prediction results, and similar past cases as well as non-compliance to regulations, to support the decision making.

1 Introduction

Nurses and midwives play important roles in the healthcare system as they provide highly skilled and often complex care in both hospitals and communities. To protect and prioritise the safety of the public from harmful practices, most countries have specific *health professional regulators* to set rules, monitor and shape the practice of nurses and midwives. When concerns over a nurse or midwife's practice are raised, a formal *complaint* can be submitted to the regulator, and investigations will be performed to decide further actions (e.g., warnings to the nurse/midwife in question, or even suspension of their practice). As the investigation results have significant impact on the practitioners' career and reputation, processing complaints is highly time-consuming and costly (see (NMC, 2020), p49), hence, the need for effective tools to support investigations is crucial.

In this paper, we present a decision support system to improve the *efficiency* of complaints investigation for nursing and midwifery regulators, by employing state-of-the-art machine learning and natural language processing (NLP) techniques with a human-in-the-loop. We worked closely with the UK Nursing and Midwifery Council (NMC[1]), the US Texas Board of Nursing (TBON[2]), and the Australian Health Practitioner Regulation Agency (AHPRA[3]), to understand their requirements for the system and collect data for training the machine learning models. Fig. 1 illustrates the major components and workflow of our proposed system. As new cases arrive, the system processes the corresponding complaints for each case and provides the following results: **(i)** *Risk level prediction*: each case is labelled as either *high* or *low* risk, along with a *confidence score*, which allows regulators to prioritise the new complaints. **(ii)** *Explanations* of the risk prediction results, by highlighting the most salient words in the complaint texts that led to the prediction. **(iii)** *Similar previous cases*, so that users can refer to relevant past cases to make decisions on the current case. **(iv)** *Entries in the regulation code* that a new complaint is most related to, that can help the regulators quickly link the allegations in the complaints to relevant requirements in the regulation code.

A major challenge in developing the system is *data sparsity*. Due to the sensitive nature of the healthcare data and the strict data-sharing policies of the regulators, we had access to a small amount of data (initially 1.2k complaints, later 5.7k complaints) to develop and test our system. To mitigate this problem, we use ensemble methods based

[1] https://www.nmc.org.uk/
[2] https://www.bon.texas.gov/
[3] https://www.ahpra.gov.au/

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 81–91, August 1st - August 6th, 2021.

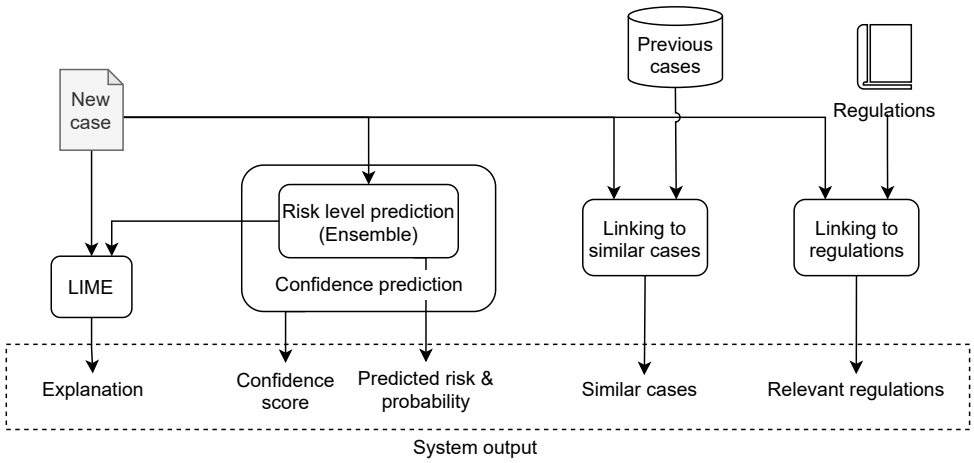

Figure 1: Workflow of the proposed system.

on both classical and neural models, including an adapted version of BERT (Devlin et al., 2019). In addition, to ensure that the predictions made by our system were *gender unbiased*, we pre-processed the text appropriately and experimented with several bias mitigation techniques. Experimental results show that the risk predictions made by the system achieved an accuracy of 0.71. An expert user evaluation, initially involving five regulatory staff at one regulator, suggests that the highlighted words and related regulation entries the system provides can not only help the regulators better understand how the predictions are made, but also allow them to provide better justifications for their decisions.

To the best of our knowledge, this is the first NLP system that supports complaints investigation for nursing and midwifery regulators.

2 Related Work

Decision Support Systems. Many NLP systems have been developed to process text data (such as records, reports, scientific papers, and social media posts) to assist in making highly critical decisions, in domains like healthcare (Bampa and Dalianis, 2020; Mascio et al., 2020; Feng et al., 2020; Proux et al., 2009), finance (Kogan et al., 2009; Wang et al., 2013), business and management (Dong and Wang, 2015; Assawinjaipetch et al., 2016; Filgueiras et al., 2019), and legislation (Rabelo et al., 2019; Soh et al., 2019; Shaffer and Mayhew, 2019). Our work proposes the first decision support system to process nursing/midwifery complaints.

Model Selection & Adaptation. *Data sparsity* is a common problem encountered by many NLP decision support systems, due to the sensitive nature of the data in certain domains and the high cost of labelling them. Hence, large neural network models do not always outperform classic feature-rich models and careful model selection is often necessary. For example, Filgueiras et al. (2019) found that, in an economic activity classification task, the SVM (Cortes and Vapnik, 1995) with TF-IDF (Salton and Buckley, 1988) representations performed better than an LSTM network (Hochreiter and Schmidhuber, 1997). On the other hand, Assawinjaipetch et al. (2016) and Mullenbach et al. (2018) showed that in complaint and clinical classification tasks, RNNs (Cho et al., 2014) or CNNs (Kim, 2014) with pre-trained word2vec embeddings (Mikolov et al., 2013) outperformed the classic machine learning models with bag-of-words representations. For each functionality in our system, we consider both classic and state-of-the-art neural network models and select the most appropriate one.

Another popular strategy to address data sparsity is to *adapt* large pre-trained models to an application domain. For example, BioBERT (Lee et al., 2019) and ClinicalBERT (Alsentzer et al., 2019) fine-tune BERT (Devlin et al., 2019) with biological and clinical trial data, to adapt BERT to their respective domains. Feng et al. (2020) performed sepsis and mortality prediction by deploying a hierarchical CNN-Transformer on top of BERT-based models. In our system, we fine-tune BERT with both nursing/midwifery complaints and other relevant data (e.g., MedSTS (Wang et al., 2020)) for downstream tasks (see §3).

Explainability is a highly desirable feature for decision support systems, especially in healthcare applications. Different types of information can be presented to users as explanations, including *attention distributions* (Mullenbach et al., 2018; Feng et al., 2020), *similar past cases* (Agirre et al., 2012; Rus et al., 2013; Cui et al., 2017; Tran et al., 2019), and *salient words* in the input text (Ribeiro et al., 2016; Lundberg and Lee, 2017). In the legal domain, to justify a verdict, relevant items in the law are often provided as explanations (Rabelo et al., 2020; Shaffer and Mayhew, 2019). Our method provides explanations in all the aforementioned forms except attention distributions, as it remains unclear whether attention distributions can be reliably used as explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

Gender Debiasing can help detect and reduce the decision support systems' biases against certain genders (Sun et al., 2019). Popular gender debiasing methods include gender swapping (Zhao et al., 2018), gender-debiased word embeddings (Bolukbasi et al., 2016; Manzini et al., 2019), adversarial training (Zhang et al., 2018), and fine-tuning (Park et al., 2018). To detect if there exists systematic biases against certain genders and reduce these biases, we test different gender debiasing methods in our system (see §5).

3 Our System

Initially, we used 1,241 real cases from one regulator to develop and test our system. Each case i consists of multiple fields, falling into three categories: the *complaint text* t_i, in which sensitive information is replaced with its corresponding entity type, e.g., all names are replaced with [PERSON]; *meta information of the case* $(c_{i1}, ..., c_{ik})$, e.g., status of the case, and who submitted the complaint; and the *investigation results*, including the *risk level* y_i of the case (high or low), and some additional assessment results $(a_{i1}, ..., a_{im})$, e.g., whether serious harm was caused to the patient or not. Table 1 presents some statistics of the dataset. Details of all fields are in the Appendix.

We understand from our collaborating regulatory agencies that the most essential functionality they need is to be able to *predict the risk level* of the case, as it allows them to prioritise the high-risk cases and better manage the workload. Hence, we formulated the problem as a binary classification task, which takes a complaint t_i and its meta-information $(c_{i1}, \cdots, c_{ik})$ as input and predicts the risk level y_i. We developed an ensemble model to predict the risk level (§3.1) and provided some additional information to further support the decision-making process of the regulator and help them interpret the prediction results (§3.2).

# High/low risk cases	766/475
# Words in each complaint	max/min/avg: 5922/5/280
# Serious harm to patient	185
# Maternity related cases	17
# Patient death	75
# Serious harm to nurse	5

Table 1: Statistics of the dataset, which has 1,241 cases received in 2019-20.

3.1 Risk Level Prediction

Due to the limited number of labelled examples, we decided to use *ensemble learning* for risk classification, exploiting the benefits of different models, both feature-rich and neural-based. In particular, we used *stacked generalisation* (Wolpert, 1992) with five base classifiers C1 – C5, detailed below.

(C1) Gradient boosting (Friedman, 2001), using the average of word2vec embeddings of words in the complaint text t_i as input. **(C2)** Adaptive boosting (AdaBoost) (Freund and Schapire, 1997) using the same input as C1. **(C3)** CNN (Kim, 2014) with t_i as input and GloVe (Pennington et al., 2014) as pre-trained word embeddings. We used the *multi-task learning* setup to train the CNN model: the model is trained to predict not only the risk levels y_i but also some additional assessment results $(a_{i1}, ..., a_{im})$. Preliminary results show that, compared to single-task learning (i.e., training the CNN for predicting only y_i), the multi-task learning setup improved the accuracy by about two percentage points. **(C4)** BERT-base (uncased), which was fine-tuned to predict the risk level. **(C5)** An ensemble which takes case meta information $(c_{i1}, ..., c_{ik})$ as input and uses three base classifiers (gradient boosting, AdaBoost, and linear SVM). Logistic regression is then used as a meta-classifier of C5.

For the main stacking model in the ensemble, we also used logistic regression, with the prediction probabilities returned by C1 – C5 as input.

3.2 Additional Information

Besides risk level predictions, our system outputs additional information to support the decision making and help users interpret the prediction results.

Confidence Scores are provided for each risk level prediction. We used a *conformal predictor* (Vovk et al., 2005) to produce the confidence scores. When the train and test data are i.i.d., the conformal predictor guarantees that the produced confidence scores are *valid*: for example, among all predictions with confidence score 0.6, the probability of the prediction being correct is 60%. We applied the conformal predictor to our ensemble model and used 40% of complaints as the *calibration set* to train the conformal predictor.

Explanations. To help regulators understand why the system labels a case as high or low risk, we used *LIME* (Ribeiro et al., 2016) to provide explanations for each prediction. LIME is a model-agnostic explanation method and does not need additional data for training. It is well suited to our system, which uses the ensemble classifier with different base models and only has access to a limited amount of data. For each case, LIME identifies the tokens that have the largest influence on the prediction probabilities and highlights these tokens as the *explanations*. Fig. 2 shows an example of the LIME explanation. If the highlighted words agree with the regulator's understanding of the key words in the text that could explain the risk prediction, then the regulator trusts the prediction results. If the regulator does not agree, then it is an indication that the prediction may not be reliable and hence the regulators need to investigate the case more carefully.

Similar Past Cases. In applications for legal decision-support, users often need to refer back to similar *past cases* to make decisions for new cases (see the *Explainability* paragraph in §2). To identify the similar past cases, we first computed the tfidf-cosine similarity scores of each of the past cases with the new case and selected the top 10 past cases with the highest similarity score. We then trained the BERT-base with 800 complaint texts (224k tokens) to create a new language model, fine-tuned the new model on two semantic similarity datasets, STSb (Cer et al., 2017) and MedSTS (Wang et al., 2020), and used the resulting model to further rank the selected past cases.

Initial results showed that the above method was very time-consuming, as, for the ranking, the fine-tuned BERT model needs to compare each sentence from the new case with each sentence from every past case. To reduce the computation time, we used *summarisation* models to generate a short summary for each case, so we could measure the similarity between cases by their summaries. We used an *extractive* summarisation model based on LSA (Ozsoy et al., 2011), which selects 1–3 representative sentences from each case to build the summary, and an *abstractive* summarisation model T5 (Raffel et al., 2020), which generates a few new sentences to summarise each case. We found that T5's summaries mostly focus on information from the first few sentences in each case. This strategy works well in summarising news articles but ignores much of the useful information in complaints. The LSA-based method, on the other hand, is not biased by the position of sentences and performs better and faster than T5, and hence we used it as the summarisation model.

Non-Compliance to Regulations. To assist regulators to check if the practice of the nurse/midwife, reported in the complaint complies with the regulations or not, our system exploits pre-trained *natural language inference* (NLI) models to detect non-compliance. Specifically, if we denote the entries in the regulation code as $R = \{r_1, r_2, \cdots, r_n\}$ and a complaint as a set of sentences $t = \{ts_1, ts_2, \cdots, ts_m\}$, then the task is to determine, for each (r_i, ts_j) pair, $i \in [1, n], j \in [1, m]$, if r_i contradicts ts_j or not. We used RoBERTa (Liu et al., 2019) fine-tuned on the MNLI dataset (Williams et al., 2018) as the NLI model. To reduce the computation time, we again used the LSA-based summarisation method to reduce the number of sentences in each complaint. The regulation entries R are from the latest NMC Code (NMC, 2015).

4 System Implementation

Backend. We used `Flask 1.0.2`, a Python based web development framework, to develop the backend of the system. We used `SQLite 3.34.0` to manage the database, `SQLAlchemy 1.2.6` for relational mapping, `Redis 3.5.3` for internal messaging and caching, `Nonconformist` for conformal prediction, and `Wtforms 2.1` to manage forms. The system receives new complaints in real time and can make predictions either in real time or batch so as to minimise the response time.

Frontend. The frontend of our web interface is implemented with `Bootstrap 4.1.3` and

Model	Accuracy	Macro F1
Majority Baseline	0.617 ± 0.032	NA
C1: Gradient Boost.	0.671 ± 0.025	0.629 ± 0.025
C2: AdaBoost	0.646 ± 0.028	0.611 ± 0.034
C3: CNNMultiTask	0.668 ± 0.029	0.623 ± 0.035
C4: BERT-base	0.680 ± 0.038	0.658 ± 0.028
C5: Meta info	0.662 ± 0.029	0.591 ± 0.056
Ensemble model	**0.708 ± 0.036**	**0.679 ± 0.032**

Table 2: Performance (mean ± standard deviation) of the risk classifiers, averaged over 10 random splits.

`Charts.js 2.5.4`. Functionalities like tool traversal, event handling, and animation are implemented using `JQuery 3.5.1`. Figure 2 shows a screenshot of a result page for a specific complaint using fictitious data. It depicts the complaint text on the left and the predicted risk as well as additional information on the right. The user can provide feedback for the predictions (accept or reject a prediction result, and provide reasons for the same). They can also provide feedback about the relevance of each similar case and regulation code, suggested by the system, to the selected case.

5 System Evaluation

Risk Level Classification results are presented in Table 2. All results were averaged over 10 runs with different random seeds, and in each run the data was randomly split into train, dev, and test sets with ratio 800:200:241. We found that all base models C1 – C5 significantly[4] outperform the majority baseline, in terms of both accuracy and macro F1, and the ensemble of the base models significantly outperforms all base models but BERT, which achieves comparable macro F1. Given the relatively small size of the data, we consider these results promising and believe that in real deployment the risk prediction performance can be further improved, as the model will have access to more labelled data.

Gender Debiasing. We aimed to answer two questions: (i) whether our risk prediction model is biased against certain genders (e.g., always associating some gender terms with the high risk class), and (ii) whether the gender biases can be reduced by using some debiasing methods. The study of ethnic biases will be conducted in the future, as most cases in our current dataset do not include any information about the ethnicity of the patients or the practitioners.

[4]Throughout this paper, p-values are computed with paired t-test and the significance level is 0.05.

Technique	Training data	Test data
Gender removing	he → ϕ	he → ϕ
Gender neutralising	he → they	he → they
Gender swapping	he → he, she	he → he

Table 3: Examples of three gender debias methods.

To measure to what extent a model is gender biased, two widely used metrics are *false positive equality difference* (FPED) and *false negative equality difference* (FNED) (Dixon et al., 2018). The lower the FPED (FNED, respectively) values, it means the gaps between the model's false positive (false negative, respectively) rates in the gender-specific and overall cases are smaller, hence suggesting lower gender bias of the model. The FPED and FNED values for our ensemble-based risk prediction model are 0.189 and 0.117, respectively (first row in Table 4). Since they are not zero, it suggests that the model does have gender biases.

To reduce the gender bias, we experimented with three methods to "clean" the data: *gender removing*, which removes all gender words from both training and test data; *gender neutralising*, which replaces each gender word with a neutral word (e.g., dad → parent) in both the training and test data; and *gender swapping*, which creates new training examples by swapping the genders (e.g., dad → mum), and train the model with both the original and the new gender-swapped data. Table 3 illustrates these gender debiasing methods. In addition to the above methods, we also tested the use of gender-debiased word embeddings (Bolukbasi et al., 2016), in base models C1 and C2, to further reduce biases. Note that, models C3 and C4 were not used as we did not debias embeddings of GloVe and BERT in C3 and C4; including them may obscure the effect of the debiased word2vec. Also, CNN and BERT are too time-consuming to train and run for ten times of the eight models.

Table 4 compares the performance of different debiasing methods. With standard word embeddings (the upper part in Table 4), all three gender debiasing methods managed to reduce gender biases, at the price of at most two percentage points loss in accuracy. However, when the gender debiasing methods are used together with the gender-debiased embeddings, the performance becomes even worse. This reminds us of existing work that questions the effectiveness of debiased embeddings (Gonen and Goldberg, 2019). Some also argue that it gets rid of more meanings beyond prejudice

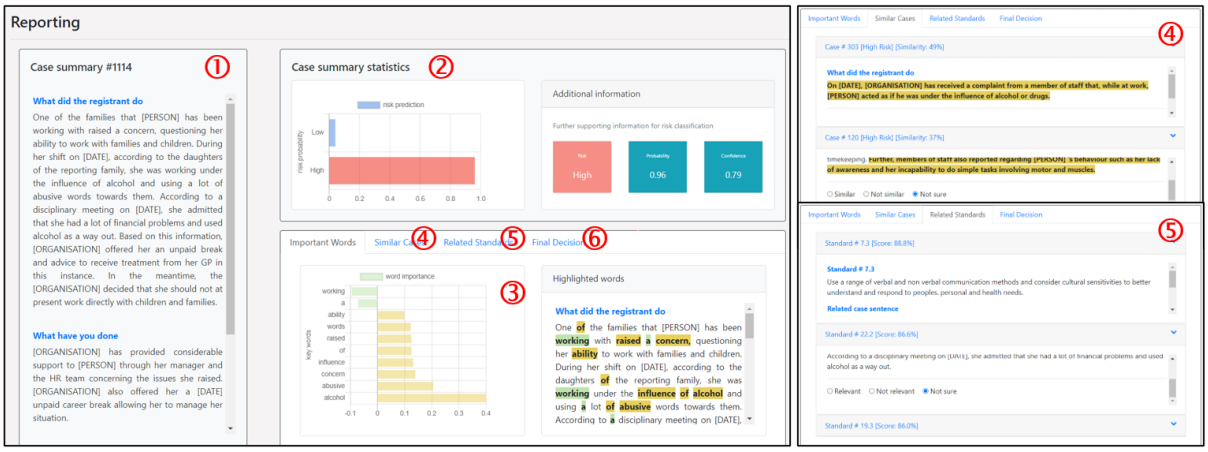

Figure 2: A screenshot of the result page for a fictitious complaint. The page consists of (1) the complaint text (2) the predicted risk level, probability, and confidence (3) word importance scores provided as the explanation by LIME (4) similar past cases (5) non-compliance to regulations (6) the final decision to be given by a case manager.

Debias Setting		Accuracy	Macro F1	FPED	FNED
O	unchanged	**0.718**	**0.688**	0.189	0.117
	remove	0.700	0.666	0.167	0.105
	neutralise	0.709	0.677	**0.129**	0.085
	swap	0.713	0.682	0.154	**0.080**
D	unchanged	0.705	0.674	0.186	0.117
	remove	0.699	0.664	0.191	0.082
	neutralise	0.707	0.675	0.190	0.101
	swap	0.708	0.676	0.186	0.117

Table 4: Performance of different gender debias methods. "O" and "D" in the leftmost column stand for original and gender-debiased embeddings, respectively.

rather than guiding the AI to act fairly (Caliskan et al., 2017). Hence, in real deployment, our system will only perform gender swapping and use the resulting data to train the ensemble model.

Human Evaluation. We invited five regulatory staff from NMC to use and evaluate our system. Each case maanager was provided with four complaints randomly sampled from our test set. They were asked to use our system to assist them in their investigation of the complaint. A questionnaire was provided to them after the test was completed, requesting their ratings (5-point Likert scores) and comments on different aspects of the system.

All participants found the usability and responsiveness of the system highly satisfactory, with average scores at 4.4 and 4.2, respectively. With respect to the quality of the risk predictions, explanations (i.e., the highlighted words), and the identified relevant regulations, participants provided moderate ratings at 2.8 for each of them. However, lower ratings (1.8) were given on the similar cases found by the system: for example, a complaint mentions that *the nurse has a strong odour of alcohol on her breath* and the experts want the system to find other cases about nurses who are inebriated or unfit to practice, but the system found cases with words like *alcohol* or *odour*, even though the words were used in very different contexts (e.g., *used alcohol as disinfectant*). We believe this is a highly challenging task as it requires not only domain knowledge but also *common sense knowledge* to capture the nuances in the complaints. We leave further investigation of this problem to future work.

As for the explanations (i.e, words highlighted by LIME), the participants reported that the highlighted words in the high-risk cases were often sensible and useful, while the words highlighted in the low-risk cases were sometimes stopwords and hence difficult to interpret. We believe the reason for this is that our models rely on the appearance of certain *keywords* (e.g., injured, died) to identify the high-risk cases, which are absent in the low-risk cases and hence the model picks up some spurious words to make the predictions. We note that, while highlighting the stopwords makes it difficult for the regulatory experts to interpret the explanations, it helps the system designers and machine learning experts better understand the problems with the system and hence allows them to improve the system accordingly. In the next version, we plan to hide stopwords highlighted by LIME from the regulatory experts to avoid confusion, but we will show them to system designers in order to help them improve the model.

6 Conclusion

In this work, we have presented the first system to support complaints investigation for nursing and midwifery regulators. The system exploits state-of-the-art text classification, summarisation, semantic similarity measurement and NLI techniques, and provides different types of information to assist the regulators, including risk level assessment, similar past cases, and non-compliance to regulations. In addition, explanations (in the form of highlighted words) are provided to improve the transparency of the system, and gender debiasing operations are performed to reduce systemic gender biases. Feedback received from domain experts confirmed the system's usefulness and potential.

We will continue our collaboration with the nursing and midwifery regulatory bodies and collect more labelled data, e.g., relevant case pairs and non-compliance to regulations; this data will help us develop domain-specific sentence similarity measurement and NLI models to further improve the performance of the system. We are considering extending the system with additional functionalities, for example, applying *active learning* (Klie et al., 2018) to allow the system learn more efficiently from human feedback and thus be constantly updated online. We also plan to perform additional experiments in control groups with domain experts to test the effectiveness of the system, e.g., by comparing the average time consumed to process a case with and without the use of our system.

Regulatory bodies in different jurisdictions face similar problems (e.g., long processing time, high cost, and an increase in the number of cases to investigate) and have similar requirements on the functionalities of the system (risk prediction, similar past cases, non-compliance to regulations). Hence, we hope this work will inspire more AI/NLP-based decision support systems across different jurisdictions, and encourage more collaborations between the NLP researchers and regulatory bodies in the legal, financial and healthcare sectors.

Acknowledgements

We thank our colleagues from the UK Nursing and Midwifery Council (NMC), the US Texas Board of Nursing (TBON), and the Australian Health Practitioner Regulation Agency (AHPRA), for their feedback to the development of the prototype. We also thank Francesco Fildani and Narinderpal Sehra from the CIM IT helpdesk at Royal Holloway, University of London, who helped us deploy the demonstration system. This project was funded by the US National Council of State Boards of Nursing (NCSBN).

Ethical Impact Statements

As our system processed highly sensitive data and its recommendations can have an impact on the person under investigation, we describe the system's potential ethical impact in different aspects below.

Data Collection. All data were collected, redacted and distributed by professionals from the regulatory agencies, strictly following all the related regulations in their respective countries.

Institutional Review. This project has been reviewed and approved by each participating institution, in line with their ethical approval process.

Expected Beneficiaries. The direct beneficiaries are the regulatory agencies, as the system improves the efficiency of their investigation and reduces the cost. The nursing/midwifery community and the patients will also benefit, as the waiting times will be reduced. Moreover, it will reduce costs which are often passed on to registrants via registration fees.

Failure Modes. Our system provides confidence scores and highlighted words to help users make sense of the predictions. Hence, even in the "failure cases" where the system provides imprecise predictions, the users can quickly identify the problems and reject the predictions (see §3). In terms of data security, our system does not edit or modify the original texts, and all texts have backup copies in secure servers; hence, the risk of data contamination or loss is minimised.

Biases. We inspected different types of potential biases and employed multiple techniques to minimise biases, as discussed in §5.

Misuse Potential. The system will be used by well-trained users from the regulatory bodies strictly inside their organisations, following all guidelines and requirements of the agencies. Hence, we believe that the potential for misuse is very low.

Potential Harm to Vulnerable Populations. Our system learns from past decisions to make new predictions. A potential risk is that, if the human decisions on the past cases have strong biases or systematic mistakes, the system may exploit those biases in its decision making. We believe the explanations produced by our system can be used to identify such systemic biases and mistakes. If users find

that certain gender-related words are highlighted, it suggests that the model heavily relies on those words to make predictions, and the regulatory staff can perform further investigations accordingly.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Panuwat Assawinjaipetch, Kiyoaki Shirai, Virach Sornlertlamvanich, and Sanparith Marukata. 2016. Recurrent neural network with word embedding for complaint classification. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 36–43, Osaka, Japan. The COLING 2016 Organizing Committee.

Maria Bampa and Hercules Dalianis. 2020. Detecting adverse drug events from Swedish electronic health records using text mining. In *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, pages 1–8, Marseille, France. European Language Resources Association.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

C. Cortes and V. Vapnik. 1995. Support vector networks. *Machine Learning*, 20:273–297.

Xiaolan Cui, Shuqin Cai, and Yuchu Qin. 2017. Similarity-based approach for accurately retrieving similar cases to intelligently handle online complaints. *Kybernetes*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Shuang Dong and Zhihong Wang. 2015. Evaluating service quality in insurance customer complaint handling throught text categorization. In *2015 International Conference on Logistics, Informatics and Service Sciences (LISS)*, pages 1–5. IEEE.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489, Online. Association for Computational Linguistics.

João Filgueiras, Luís Barbosa, Gil Rocha, Henrique Lopes Cardoso, Luís Paulo Reis, João Pedro Machado, and Ana Maria Oliveira. 2019. Complaint analysis and classification for economic and food safety. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 51–60, Hong Kong. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *COLING 2018, The 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico, August 20-26, 2018*, pages 5–9. Association for Computational Linguistics.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 86–94, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

NMC. 2015. Professional standards of practice and behaviour for nurses, midwives and nursing associates.

NMC. 2020. Nursing and midwifery council annual report and accounts 2019–2020 and strategic plan 2020-2025.

Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4):405–417.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Denys Proux, Pierre Marchal, Frédérique Segond, Ivan Kergourlay, Stéfan Darmoni, Suzanne Pereira, Quentin Gicquel, and Marie-Hélène Metzger. 2009. Natural language processing to detect risk patterns related to hospital acquired infections. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 35–41, Borovets, Bulgaria. Association for Computational Linguistics.

Juliano Rabelo, Mi-Young Kim, and Randy Goebel. 2019. Combining similarity and transformer methods for case law entailment. In *Proceedings of the*

Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019, pages 290–296. ACM.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. A summary of the coliee 2019 competition. In *New Frontiers in Artificial Intelligence*, pages 34–49, Cham. Springer International Publishing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Vasile Rus, Mihai Lintean, Rajendra Banjade, Nobal Niraula, and Dan Stefanescu. 2013. SEMILAR: The semantic similarity toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 163–168, Sofia, Bulgaria. Association for Computational Linguistics.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523.

Robert Shaffer and Stephen Mayhew. 2019. Legal linking: Citation resolution and suggestion in constitutional law. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 275–282.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 802–808, Nagoya, Japan. Asian Federation of Natural Language Processing.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Lang. Resour. Evaluation*, 54(1):57–72.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Appendix

Hyperparameters Selection

Hyperparameters of our models are selected using grid search on 250 randomly sampled cases; results are presented below. For the CNN model (base model C3 in the ensemble), we use three filter sizes (2,3 and 4) and 15 filters for each size. For the

multi-task training (base model C3), the loss function we use is $L_y + 2L_A$, where L_y and L_A are the cross-entropy losses for predicting the risk level and additional assessment results, respectively. To fine-tune BERT (base model C4), we use Adam as the optimiser with fixed learning rate 2e-5, batch size 8 and perform the training for 10 epochs.

Data Fields

Fields in the dataset are summarised in Table 5.

Category	Data Fields
Meta Information $(c_{i1}, ..., c_{ik})$	**CreateDate** (when the case was created), **CurrentStatus**(closed, in investigation, or await adjudication hearing), **Referrer** (who submitted the complaint)
Assessment Results $(a_{i1}, ..., a_{im})$	**RiskLevel** (high or low), **RiskOfRepetition** (True or False), **SeriousHarmToRegistrant** (True of False), **SeriousHarmToPatient** (True of False) **BreachOfOngoingRegulatoryIntervention** (True or False), **MaternityRelated** (True or False), **PatientDeath** (True of False), **InvestigationResults** (free text)

Table 5: Fields in the complaints dataset.