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

Background and Objectives

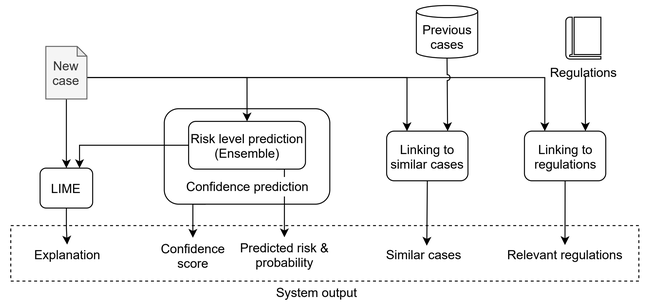
Nurses and midwives play important roles in the healthcare system as they provide highly skilled and often complex care in both hospitals and communities. To protect and prioritise the safety of the public from harmful practices, most countries have specific *health professional regulators* to set rules, 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. Processing complaints is highly time-consuming and costly 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.

Implementation

The most essential functionality regulatory agencies 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. To make this prediction, we formulated the problem as a binary classification task and developed an ensemble model. Our system also provided some additional information to further support the decision-making process of the regulator and help them interpret the prediction results:

- **Confidence scores** to assess the certainty of each prediction
- **LIME** to provide **explanations** for each prediction
- Reference to **similar past cases**
- **Natural Language Inference** (NLI) models to detect **non-compliance**

System Design



Evaluation

Risk Level Classification results are presented in Figure 3. We found that all base models C1 – C5 significantly 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.

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 in both the training and test data; and *gender swapping*, which creates new training examples by swapping the genders, and train the model with both the original and the new gender-swapped data. Figure 4 illustrates these gender debiasing methods using the *false positive equality difference* (FPED) and *false negative equality difference* (FNED) metrics.

User Interface

Our system is primarily implemented with `Flask 1.0.2`, `Bootstrap 4.1.3` and `Charts.js 2.5.4`. We invited five regulatory staff from NMC to use and evaluate our 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.

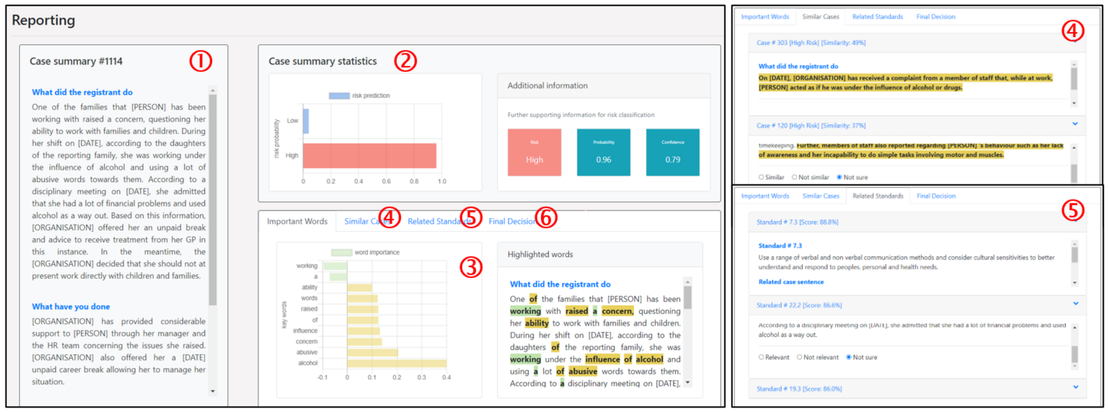


Fig. 2: A screenshot of the result page for a fictitious complaint. The page consists of (1) the complaint text (2) the predicted risk level, probability, and confidence (3) word importance scores provided as the explanation by LIME (4) similar past cases (5) non-compliance to regulations (6) the final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

Results

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**

Fig. 3: Performance (mean ± standard deviation) of the risk classifiers, averaged over 10 random splits.

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

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

Conclusion and Future Work

We 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 (with highlighted words as explanations), similar past cases, and non-compliance to regulations. Also, gender debiasing operations are performed to reduce systemic gender biases. Feedback received from domain experts confirmed the system's usefulness and potential. We hope this work will inspire more AI/NLP-based decision support systems across different jurisdictions, and encourage further collaborations between NLP researchers and regulatory bodies