Annals of Operations Research
https://doi.org/10.1007/s10479-023-05631-9

ORIGINAL RESEARCH

AI-powered decision-making in facilitating insurance claim dispute resolution

Wen Zhang[1] · Jingwen Shi[2] · Xiaojun Wang[3] · Henry Wynn[4]

Received: 30 November 2022 / Accepted: 28 September 2023

Abstract
Leveraging Artificial Intelligence (AI) techniques to empower decision-making can promote social welfare by generating significant cost savings and promoting efficient utilization of public resources, besides revolutionizing commercial operations. This study investigates how AI can expedite dispute resolution in road traffic accident (RTA) insurance claims, benefiting all parties involved. Specifically, we devise and implement a disciplined AI-driven approach to derive the cost estimates and inform negotiation decision-making, compared to conventional practices that draw upon official guidance and lawyer experience. We build the investigation on 88 real-life RTA cases and detect an asymptotic relationship between the final judicial cost and the duration of the most severe injury, marked by a notable predicted R^2 value of 0.527. Further, we illustrate how various AI-powered toolkits can facilitate information processing and outcome prediction: (1) how regular expression (RegEx) collates precise injury information for subsequent predictive analysis; (2) how alternative natural language processing (NLP) techniques construct predictions directly from narratives. Our proposed RegEx framework enables automated information extraction that accommodates diverse report formats; different NLP methods deliver comparable plausible performance. This research unleashes AI's untapped potential for social good to reinvent legal-related decision-making processes, support litigation efforts, and aid in the optimization of legal resource consumption.

Keywords Professional service operation · Insurance claim · Civil litigation · AI · Natural language processing

Wen Zhang and Jingwen Shi have contributed equally to the manuscript.

The expressed opinions and views are solely those of the author and do not reflect the views or positions of SSGA.

Extended author information available on the last page of the article

Published online: 23 October 2023

1 Introduction

1.1 A motivational case

Litigation decision-making relies on cost–benefit analysis, weighing the terms of settlement against the expected outcome of litigation, considering transaction costs. Despite its complexity, predicting the true value of a case as accurately as possible can help participants make informed decisions and increase efficiency at the individual level, the corporate level, and the community level. We can translate this problem into estimating the "intrinsic value" of the personal injury cost in a road traffic accident (RTA) insurance claim. In such situations, both claimants and defendants must balance the resultant benefits versus the costs involved, which encompass a multitude of elements including the likely legal ramifications and the legal billings such as claims fees and attorney charges (see GOV.UK, 2021 for more cost details). Both parties can risk encountering more loss or costs, both money-wise and time-wise, if the amount of claim they initially seek or offer does not justify the fair valuation of the injury damage. From the claimant's perspective, this establishes expectations for recovery so that he would not claim an unreasonably high amount and would be better informed of whether the ultimate recovery justifies the time and expenditure spent on the matter. Consider a real-world case in which the claimant set out his initial claim to be £6000, following which the defendant responded with a counteroffer of £3400. No compromise had been reached before proceeding to court, and by the time a final order was made at £3800, the costs encountered far exceeded the incremental gain in recovery: the claimant could have obtained more by accepting the plea bargain at earlier stages. Similarly for the defendant, she might end up paying even more in settlement including costs had inaccurate assessments been made, or even worse, she could risk more serious convictions at trials. This is the case in another real-world scenario: the defendant insisted on offering £3000 while the claimant compromised from £4000 to £3750, with the final verdict being £4000.

Reasonable assessments regarding settlement prospects, enabled by automated trial outcome prediction, greatly contribute to informed decision-making and effective settlement negotiations for individuals, considering the uncertainties of legal outcomes as well as the costs in terms of time, money, effort, and trouble (Molot, 2009). The savings could be enormous when aggregated at the societal level, questioning the efficient use (Woolf, 1995) and distribution of the limited legal resources, a long-standing national demand (Jackson, 2010). According to our fieldwork, if no agreement has been reached through settlement negotiations, 2 years' awaiting a final court decision counting from the date of the incident would be an underestimate. This not only entangles both individual parties in the matter but also takes up valuable resources that the trial needs to command. This consideration is especially important for low-value RTA disputes, which do not involve intricate interactions and could have been resolved through negotiation before being brought to court.

The legal service industry has many routine and repetitive processes (Avgerinos, 2018) that can be streamlined with AI-powered solutions, improving productivity and efficiency. However, despite the growing attention and interest in professional service operations (PSO) (Lewis, 2012) such as legal and insurance industries, the implementation of AI into the functioning of PSO represents a less-explored research avenue compared to the extensive discussions on its applications in other operational management realms such as operational risk management (Araz, 2020) and procurement (Cui, 2022). For RTA insurance claim service discussed in this paper, a staggering volume of documents needs to be reviewed for each individual case to assess the cost of insurance. These needs can be catered to with the uncovering of the potential of AI in terms of automating document processing, delivering reasonable valuations, and driving informed decision-making, to fully exploit AI's pattern recognition capability with structured or unstructured data (Abrahams, 2015).

Through examining 88 cases UK Ministry of Justice (MoJ) RTA Portal-defined as the "Pre-Action Protocol for Low Value (up to £25,000) Personal Injury Claims in RTA from 31 July 2013" (Justice, 2017)-provided by a leading legal service company based in London, we found that the claimant and the defendant (or their insurance representatives) will negotiate over different issues via this online platform, among which one of the primary issues is the injury cost. The injury cost, also known as the *general cost*, is closely associated with the claimant's injury type and severity (rather than other human characteristics such as age, gender, etc.) as stated in the medical report provided by the claimant's medical expert, usually a disinterested third party like GP. According to interviews with lawyers (as summarized in the "Appendix"), most practitioners rely heavily on personal experience and conventional qualitative legal analysis when assessing the potential legal outcomes and consequences, which are subject to inaccuracy and ineffectiveness at tailoring recommendations to specific individuals and legal practitioners show great interest in how technology can help in decision-making. This makes it particularly conducive to leveraging the advances in machine learning-based tools coupled with the increasing availability and digitalization of data. Despite the promising prospects to shed new light on decision-making, proper exploitation in the context of a specific practical scenario as well as the investigation into its real-world economic rationale and societal impact is absent in existing research.

1.2 Research framework and contributions

To address this gap, we identify and quantify the relationship between the general cost and injury attributes and illustrate the automated extraction of injury information from medical reports via RegEx (Thompson, 1968), one of the most widely used information extraction techniques in the realm of natural language processing (NLP). We also examine various mining methods that process medical text directly, such as the linear method of Stochastic gradient descent (SGD) linear regression and Convolutional Neural Network (CNN) (Kim, 2014). While these methods can achieve superior predicted R^2 values of around 0.8, their lack of interpretability and transparency is a critical barrier to modelling accountability and widespread adoption. In contrast, with the infusion of domain knowledge, the RegEx-powered method that extracts precisely pertinent injury information and makes projections of the general cost, guarantees an unbeatable level of interpretability and traceability. With the rapid development of AI, growing attention has been paid to the interpretability issue (Samek, 2019) out of scientific, ethical, commercial, and regulatory considerations, especially in the context of formal decision-making such as legal reasoning where rigorous justification is necessary (Atkinson, 2020).

Concretely, the most significant novel aspect of this study is to leverage both practitioners' perspectives and state-of-the-art AI-powered techniques to establish a data-driven framework for enhanced informed decision-making in the context of civil litigation/insurance, as exhibited in Fig. 1. The study further demonstrates specifically the prediction of the general cost in RTA insurance claims. The key technical and practical contributions of this article are delineated below.

Technical Contributions comprise three areas: (i) *AI/NLP-Driven Architecture for PSO,* (ii) *Information Extraction Procedure,* and (iii) *Text Mining for Cost Prediction.* First, we contribute to the service operation literature on AI adoption (Huang, 2021; Kumar, 2018) by introducing an AI/NLP-powered structure for PSO, designed to streamline processes, harness digitalized information sources, and expedite informed decision-making. Our focus is on explainable rule-based AI solutions integrated with domain expertise, ensuring transparency in decision-making logic. We also explore the incorporation of advanced text mining techniques to fully unleash the potential of AI. Second, we put forth an interpretable methodology for automated text processing in legal settings, where there is the widespread call for explainable and interpretable AI approaches (Atkinson, 2020). This involves practical extraction of injury information from medical reports using RegEx. The process amalgamates keyword labels, numerical markers, and semantic context to precisely pinpoint injury attributes. Third, we delve into text mining techniques, including SGD linear regression and CNN, to predict costs directly from medical reports. Our comparisons highlight the superior model, achieving an R^2 value surpassing 0.8. Moreover, we demonstrate the application of learning-based data augmentation to enhance training sets with limited samples, which is a common challenge faced in industries like healthcare and legal (Perez, 2017).

Practical Contributions also consist of three aspects: (i) *Decision-making Pipeline Proposal*, (ii) *Relevance to Legal Services*, and (iii) *Vision for Social Good.* First, we advocate for an exhaustive decision-making pipeline, which introduces enhancements in text information processing methodologies. This breakthrough is pivotal for industries such as law and insurance. It furnishes them with pragmatic recommendations to capitalize on automation for improved operational efficiency. Second, by leveraging the AI-enhanced workflow, legal service entities can derive intrinsic case valuations. Utilizing domain expertise, historical datasets, and predictive analytics, they can realize better decision-making and heightened efficiency in litigation expenditures. This is paramount at individual, corporate, and societal scales. It forms the bedrock for the social good—ensuring efficient dispute resolutions, optimized legal resource utilization, and accessibility of litigation tools for all, especially the working class and those with lower incomes. Third, beyond revolutionizing industry services, the power of AI for societal good (Taddeo, 2018) extends to wider disciplines and we are moving towards a future where more informed and cost-effective decision-making is achievable by harnessing historical data and technological solutions.

The remainder of this paper is structured as follows. Section 2 reviews the theoretical background and related work and Sect. 3 describes the proposed framework in the context of the practical case. Section 4 illustrates how to extract the information automatically via RegEx, and Sect. 5 further explores the possibility of predicting from the medical report directly without manual feature selections. Section 6 demonstrates the financial benefits

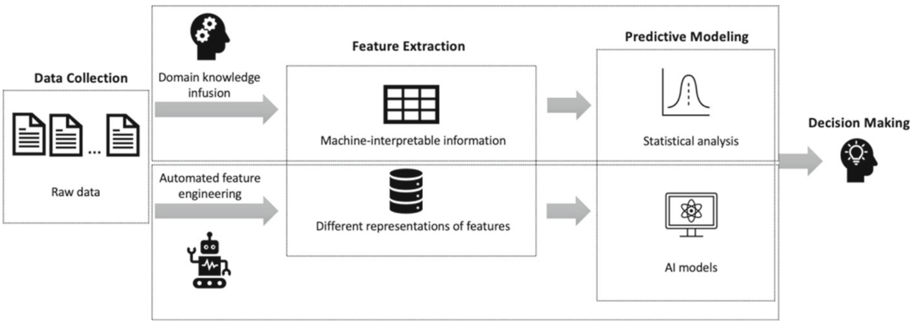

Fig. 1 Proposed AI-driven decision-making framework

of applying the proposed model, and we summarize our main findings, limitations of this research, and future research directions in Sect. 7.

2 Literature review

Information technology has the transformational power to unlock immense economic benefits that were previously inaccessible, and the professional service industry is not lagging in participating in digitalization and harnessing the potential of data. The OM community has been actively engaged in discussions on how to leverage information technology/information systems to improve the efficiency, effectiveness, and productivity of PSO (Boone, 2001; Ray, 2005). For example, how call center technologies, Electronic Medical Record (EMR), and alternative databases tap into new sources of insights constitutes compelling evidence to support such statements in a wide-ranging branch of professional service industries including legal firms (Lewis, 2012), hospitals (Dobrzykowski, 2016) and public sector operations (Karwan, 2006).

Revitalizing the landscape of information technology, AI and big data analytics present new opportunities to gain competitive advantages. Along with the accelerating pace of innovation in the development of AI technologies, the adoption of AI to enhance the broader service operation management (Huang, 2021; Kumar, 2018) is receiving growing attention in recent years. Despite the increasing interest, most studies on AI in service operations expand around conceptualizing AI-based service ecosystems or proposing research agenda (Kumar, 2018). Utilizing AI for effective decision-making and its empirical execution, however, remains less explored with a paucity of research literature and lack of established approaches. To fill in the gap, one highlight of this research is to illustrate how the AI implementation, more specifically NLP technologies, enhances service practices and quantifies PSO decision making through a case study with a UK legal service firm, as an example of a representative professional service industry that features knowledge intensity and customization (Lewis, 2012). This paper takes on a new perspective considering the practical merits of NLP in the legal realm, compared to those adopted by the computer science community.

AI in service applications can be categorized into three dimensions including mechanical AI for service standardization (i.e., AI for routine and repetitive service), thinking AI for service personalization (i.e., uncovering meaningful patterns from large data sets and making predictions) and feeling AI for service rationalization (i.e., virtual agents delivering mechanical service at the low level and advanced AI detecting emotions at the high end) (Huang, 2021). Most existing research on "Legal AI/NLP" caters more to the first two aspects, such as predicting judicial decisions by performing Support Vector Machine (SVM), CNN and FastText on court proceedings (Medvedeva, 2020; Xiao, 2018) and the settlement or dismissal of the US Federal Court cases from processed corresponding dockets (Vacek, 2019). After accounting for case heterogeneity, these NLP models bring on the leverages of mechanical AI or thinking AI to streamline legal processes and operations, unlock the benefits from data, and contribute to accelerated and informed decision-making.

Taking a slightly different pathway, this research particularly acknowledges the importance of human-interpretable explanations and leans towards deploying *Explainable-AI* solutions that ensure visibility into the decision logic by design. Different from manufactured products, service provision architecture recognizes the challenges involved with an appropriate and comprehensive combination of materials, skills, and processes to yield the "planned" or “designed” services (Goldstein, 2002). To put this into context, beyond the common efficiency and productivity considerations in OM, this research emphasizes the effectiveness or “doing the right thing” (Karwan, 2006) and aims to enhance the explainability with the infusion of domain knowledge, which is of particular importance in PSO due to its knowledge-intensive nature. The concerns about the effectiveness of information technology in OM are not unfounded in the real world: a corporate partner expressed similar concerns (Lewis, 2012) that “you cannot do the job in an efficient manner unless you distil down the knowledge that lawyers have in a form which is able to be reproduced and used by others”, highlighting the necessity of “effective” utilization of precedent knowledge.

Accordingly, rather than relying solely on “black box NLP” models such as CNN to process every single particle of legal documents, we first incorporate domain knowledge that helps narrow down the scope to illuminating the relationship between trial outcomes and injury information and then illustrate how these supporting data can be collected from medical (or clinical) records using NLP techniques in an automated manner. As mentioned above, this respects the widespread preferences for explainable and interpretable AI approaches which are particularly pronounced in the decision-making process in legal settings which calls for more rigorous reasoning and justification (Atkinson, 2020). This not only arises out of scientific, ethical, and commercial considerations but also has been backed up by regulatory requirements. For example, the EU’s General Data Protection Regulation (GDPR) (https://gdpr-info.eu/) has highlighted the importance of human-understandable interpretations of machine-derived decisions. In this regard, the second aim of this study is to address the importance of explainable AI in PSO through a practical example.

In this study, we examine medical reports to support the claimant’s insurance claim, which usually are narrative documents that describe the patient’s demographic information, accident details, treatment records, doctor’s opinions, and recommendations, etc. (see Sect. 4). These documents provide a valuable information source to investigate and exploit the associations between injury attributes and the general cost as suggested by the Judicial College Guidelines (refer to JCG later) (Lexis, 2020). To this end, we demonstrate the use of a powerful pattern-searching language, that is the regular expression (RegEx) (Thompson, 1968), which has been widely applied in clinical textual information extraction for its flexibility and convenience. To tackle the specific task here, we deploy a combination of label-based (Turchin, 2006), numeric-oriented (Murtaugh, 2015), and semantic-driven logics, while explicitly addressing the long-range context dependencies (Jagannatha, 2016) in a traceable and interpretable manner.

Complementary to the above explainable rule-based approach, we also illustrate an architecture using text mining techniques that are grounded in more advanced machine learning models, which apply feature extraction by directly transforming raw text into word representations to be fed into predictive models. We demonstrate how greater flexibility and capability are possible with two common supervised learning models CNN and SGD regression, the former selected for its proven performance in text classification and regression problems (Bitvai, 2015) and the latter for its suitability to large-scale and sparse scenarios (Smith, 2008). Contrary to the rule-based approach that relies on pre-specified dictionaries or patterns determined by domain knowledge, machine learning methods “automatically” learn features from the annotated training set after processing text through word vectorization or word embeddings, which convert the textual vocabulary into real-value vectors. Common vectorization techniques employed include bag-of-words model (BOW), which simply represents documents as multisets of words but ignores the context (Harris, 1954), and its variants such as frequency, one-hot, and TF-IDF (Term Frequency-Inverse document frequency) (Salton, 1988). Word embedding techniques such as the word2vec (Mikolov, 2013), GloVe (Pennington, 2014), and FastText (Joulin, 2016) take a step forward by more efficiently learning sparse vectors from large-scale text, better preserving semantic and syntactic relations, and thereby producing more contextualized word representations. To recap, the third objective of this paper is then to show how the NLP-driven architecture rooted in a wider range of AI modelling options and flexible implementation brings out the fuller potential of AI.

3 An integrated textual analytic framework for insurance claim dispute resolution

Leveraging both practitioners' perspectives and state-of-the-art AI-powered techniques, we establish an integrated textual analytic framework to promote informed decision-making in the context of civil litigation/insurance, as exhibited in Fig. 2. The framework investigates and exploits the informational value of the textual resources to derive quantitative insights and determine the cost of insurance claims.

Diverse perspectives and approaches can be adopted to fully utilize textual resources, as summarized in the upper and lower panels in Fig. 2. The choice of analysis method depends on the implementability and practicality that are permitted or limited by data availability, sample size, problem complexity, method efficiency, and many other considerations. In the first approach (Fig. 2), with the acquisition of solid domain knowledge that associates the objectives (such as claim cost) to some underlying quantitative items, these specific attributes are identified, extracted and translated into machine-interpretable representations that conventional statistical modelling can take as data feeds. To this end, we take advantage of text processing techniques such as regular expression, which greatly contribute to operational effectiveness by automating pattern-identification and information-extraction tasks.

In the absence of representational guidance and prior knowledge while sample size allows, text mining techniques present an alternative that performs text-driven information extraction and retrieval. In this regard, features can take a variety of forms that characterize textual information, including but not limited to text frequency, relative importance, or other learned representations that encode word semantics and reveal hidden patterns. Though less effort

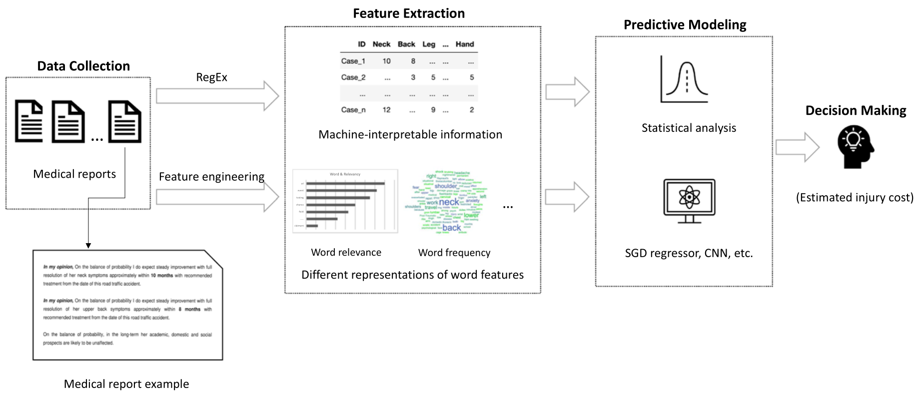

Fig. 2 Integrated textual analytic framework

is spent to build the prior domain knowledge and establish the specific purpose for model-training, this type of approach typically tends to be computationally intensive for routine use and requires expertise in model tuning.

We further illustrate the practical implementation of the framework based on real-life UK RTA insurance cases and demonstrate how this enhanced efficiency can translate into economic benefits, for defendants and claimants alike. In the first approach, we focus on elucidating the relationship between the injury severity and the general cost out of the view that the injury type-attribute carries important predictive information, according to the insights drawn from lawyers' combined experience. To extract the injury information from medical reports of various formats, we construct regular expressions solutions in a flexible and practical way to address some potential real-world challenges in automated text processing. The workflow presented here can be conveniently applied in similar scenarios while ensuring interpretability, regardless of sample size. In the second approach, we explore the feasibility and effectiveness of text mining methods that formulate cost predictions directly from textual databases, which tend to perform better with large sample sizes. To demonstrate the implementation of the second approach, we perform learning-based data augmentation to enrich and diversify the existing training set with limited samples while we believe the utility potentials can be better unleashed with larger datasets. To our best knowledge, this work presents the analytical framework in lawsuit settlement that explores textual resources in diverse ways and compares their effectiveness and appropriateness in different scenarios with real-world case applications. The relationship between the general cost and injury attribute revealed here can be conveniently incorporated and exploited to guide decision making.

4 Predict injury cost with structural injury information

In this section, we describe the data employed in this research and demonstrate the development and testing of different predictive models for estimating the general cost with the injury information.

4.1 Domain knowledge infusion

A fieldwork was conducted within a legal professional service company in the UK from 2018 to 2020 to explore applying AI and statistical predictive models in civil litigation. As an exploratory study, for the sake of ease of research, we started with a relatively straightforward business of handling low value (up to £25,000) cases of personal injury from the RTA, which requires the lawyers' professional knowledge to value the case while the legal documents are less complex compared with those high-cost cases at the same time.

In an RTA insurance claim, two types of cost need to be evaluated and negotiated between claimants and defendants: the *general cost* that is solely decided by the claimant's injury type and severity and the *special cost* which includes other costs such as income loss, repairs cost, physiotherapy cost and so on. According to our fieldwork, special cost is particularly case dependent, lacking internal characteristics for large-scale automated processing and causing little controversy if sufficient evidence (such as a receipt for repairing a car) is available, therefore we focus on the general cost only in this research (as discussed in the "Appendix"). Currently, the valuation is conducted by the lawyers/claim handlers manually, by referring to the JCG and their experience, which leads to a time-consuming process and varied results depending on the individual's subjective judgement. The growing volume of case data is not being put to good use.

For each case, there is a corresponding medical report documenting the claimant's injuries (usually multiple, such as neck and back) and this report is the primary evidence of the general cost. The lawyers interviewed mainly focus on the most serious injuries and then consider other injuries as appropriate. However, this discretion is very subjective and there is no uniform standard, which allows us to explore whether a statistical analysis could be used to find the relationship between injury and cost so that we could automatically and quickly predict the value of a new case when it comes in.

4.2 Data

This research is based on 88 low-cost RTA Portal cases that occurred between 2013 and 2019 and were provided by a UK legal service company. Figure 3 illustrates that the general cost of most cases is between £2000 and £5000, with a few cases higher than £6000, and detailed statistics are summarized in Table 1.

Since the general cost is mostly associated with the traffic accident injuries according to our interviews, we process the civil litigation proceedings manually to extract the specific injury information on each case. Summary statements of injuries are available for all cases,

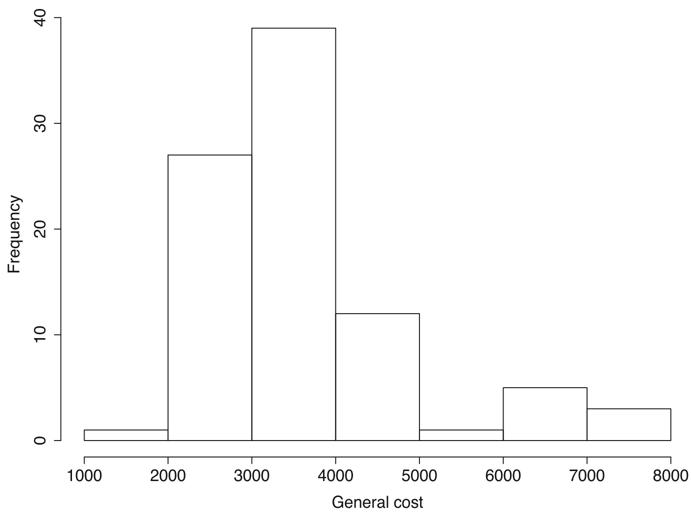

Fig. 3 General cost histogram

Table 1 Descriptive statistics summary for general cost

Descriptive statistics	Value	Descriptive Statistics	Value
Obs	88	Mean	3772
Median	3500	Mode	3000
Std. Dev	1259	Minimum	1350
Maximum	7750		

Table 2 Injury frequency

Injury type	Frequency	Injury type	Frequency
Neck	74	Back	66
Shoulder	50	Spine	19
Chest	10	Lumbar	8
Hip (s)	8	Cervical	7
Arm	4	Wrist	4

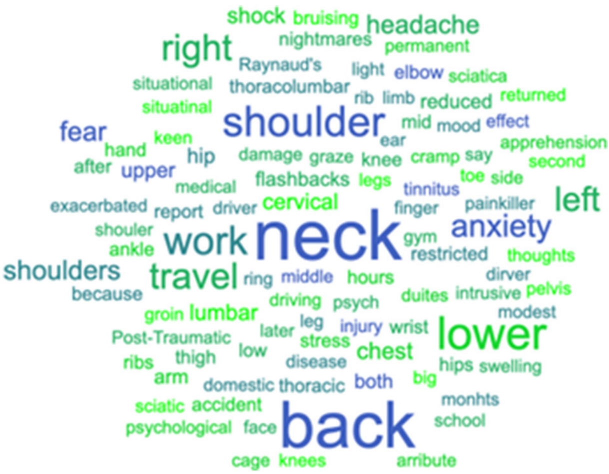

Fig. 4 Medical reports word cloud

with detailed medical reports available for 24 of them. There are 28 types of physical injuries in total and the top 10 most frequent injuries are listed in Table 2. It can be noted that the most common injuries are the neck, back, and shoulder(s) injuries. This is due to the nature of low-cost traffic accidents, which usually result in upper body injuries also called *whiplash*. The word frequency of the summary description of the symptoms is further visualized as a word cloud in Fig. 4.

4.3 Predictive modelling

When estimating a personal injury in practice, legal professionals combine the approximate range of compensation based on injury information by referring to JCG with insights from their own experience to arrive at a general cost proxy. They usually place special attention on the most severe injury and consider the rest injuries to some extent. However, this process counts heavily on individual experiences and there are no well-established or readily available analytical methods to follow.

Inspired by lawyers' beliefs on the relationship between the cost and injury, we take a step further and test this hypothesis by adopting various predictive approaches. The first model predicts the general cost based on the most severe injury, i.e., months to recover from the most severe injury, and the corresponding injury type. For the convenience of the study, we

Table 3 Linear regression of the general cost on the most severe injury with injury type indicated

Coefficient	Estimation	Std err	t-value	p-value
Intercept (a_0)	2448.54	230.42	10.626	$< 2e - 16^{***}$
Injury severity (a_1)	119.62	13.27	9.015	$5.53e - 14^{***}$
Dummy variable (a_2)	76.08	222.39	0.342	0.733
Dummy variable (a_3)	387.40	671.81	0.577	0.566

$***p < 0.001$, $**p < 0.01$, $*p < 0.05$

classify the human body into three categories of head-neck, torso, and limbs based on all possible injuries listed on JCG, and then set two dummy variables to indicate which part of the body the most serious injury belongs to. For example, the injury will be indicated as "limbs" if the most severe injury in one case is a hand injury. Further, if more than one type of injuries shares the same longest months, such as that both neck and hand require ten months to resolve, we will seek assurance from the third-ranked injury. This means, if there also exists a nine-month headache for instance, then the most severe injury will be labelled as "head-neck"; otherwise, either "head-neck" or "limbs" will be adopted randomly. This is done to consider that if the injury is severe, the area adjacent to it may also be injured. With the two dummy variables indicating the injury category, the linear regression model is

$$\mathrm{y_i} = \alpha_0 + \alpha_1 \mathrm{x_i} + \alpha_2 + \alpha_3 + \mathrm{\epsilon_i}, \quad (1)$$

where $i = 1, 2, ..., n$ ($n = 88$ in this research), y_i and x_i are the general cost and months to resolve from the most severe injury of case$i$. α_0 and α_1 represent the intercept and the coefficient of x_i and ϵ_i denotes the error term. α_2 and α_3 are dummy variables for torso and limbs respectively, i.e., when $\alpha_2 = 1$, the injury belongs to the torso or limbs when $\alpha_3 = 1$, otherwise head-neck. The regression result summarized in Table 3 reveals a statistically significant relationship between the general cost and injury severity while the cost for different types of injuries is not significantly different. One possible reason for this is that in these low-cost RTA cases, the JCG's range of compensation for each type of injury is relatively close.

Therefore, we will only consider the injury severity in the predictive model and first model will use the most severe injury to estimate the cost based on the principle of parsimony. By observation, there exists a nonlinear relationship between the injury severity and cost, and the cost tends to stabilise with the increase of the severity as demonstrated in Fig. 5. This is reasonable in practice because this suggests that a plateau in general costs should occur after a period of time, e.g., if the most serious injury takes 2 or 2.5 years to recover, it should not make a significant difference. To this end, we test and compare the performance of four common non-linear models. The first two are Quadratic and Cubic regression models within which second-degree polynomials (x^2) and third-degree polynomials (x^3) are included respectively to capture the non-linear patterns. However, since neither of them can describe the asymptotic processes (i.e., a stable cost with increased injury severity) and polynomial models are prone to overfitting due to the introduction of higher order polynomials, we also investigate two asymptotic regression models which we believe should be more suitable to this problem. Specifically, we test the logarithmic transformation and (take natural log of the input x) square root transformation (take the square root of the input x) in the linear

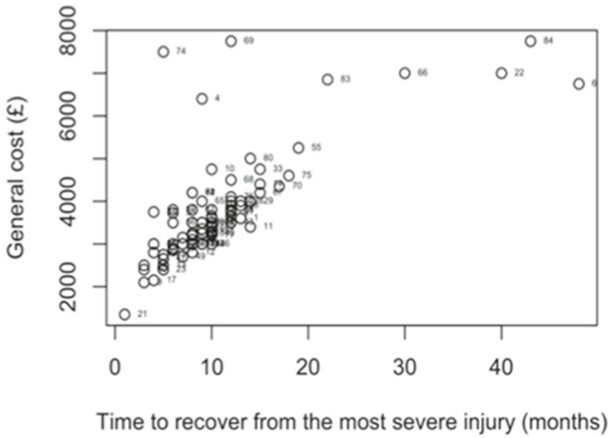

Fig. 5 Injury severity and general cost relationship

regression, both of which have slopes that asymptotically decrease to a constant. Figure 6 demonstrate the model fittings.

All models fit the data to some extent by observation and to further assess their predictive performances, we calculate and demonstrate the R^2 value, MAE (Mean Absolute Error), and RMSE (Root Mean Square Error) with Leave One Out Cross Validation (LOOCV) since LOOCV is a fairer and more reliable measurement, especially when the dataset is small, as well as the AIC and BIC values to measure model performance that account for model complexity in Table 4.

Although the Cubic model has the largest R^2 value, smallest MAE and RMSE, it tends to overfit by showing a declining trend to fit a severe case. Together with the context of the question, we recommend using the square root transformation linear regression model to

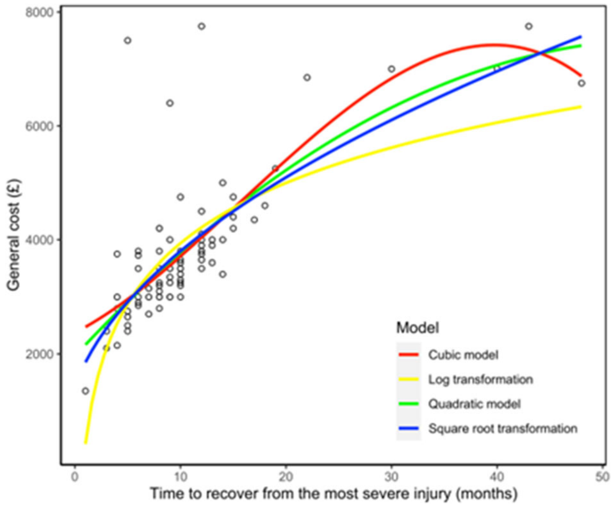

Fig. 6 Different non-linear model fittings

Table 4 Predictive performance of different nonlinear models

Model	R^2	MAE	RMSE	AIC	BIC
Quadratic model	0.521	534.964	868.193	1442.024	1451.933
Cubic model	0.530	517.315	859.074	1442.371	1454.757
Logarithmic model	0.462	625.663	918.552	1451.745	1459.177
Square root transformation	0.527	537.031	860.910	1441.249	1448.681

Table 5 Square root transformation linear regression of the general cost on the most severe injury

Coefficient	Estimation	Std err	t-value	p-value
Intercept	717.05	312.92	2.291	0.0244*
Injury severity square root	978.79	95.95	10.201	$< 2e-06$***

$* * *p < 0.001$, $** \ p < 0.01$,$*p < 0.05$

predict the general cost and the regression model is

$$y_i = \beta_0 + \beta_1 x_i^{\frac{1}{2}} + \epsilon_i, \tag{2}$$

where $i = 1, 2, ..., n$ ($n = 88$ in this research), y_i and x_i are the general cost and months to resolve from the most severe injury of case$i$, β_0 is the constant and β_1 is the coefficient of the square root ofx_i. As can be seen from Table 5, all coefficients are significant.

While adopting the most severe injury solely could achieve a good prediction performance, we notice that there are several high value cases (4, 69 and 74) with less severe injury. We further inspect them and find that the value of cases with multiple injuries can also be high. For example, the general cost in case 74 is particularly high, mainly because the claimant suffered multiple injuries from the neck, shoulder, lower back, and right hand, all of which took five months to recover from.

Hence, rather than predicting the general cost only with the most severe injury, we also examine the utility of extending the feature space by including more inputs such as the top two and three most severe injuries. One basis for these tests is that with the increased number of injury types, the general cost displays an increasing trend as shown in Fig. 7. It is worth noting that we will not predict with the full range of injuries as injuries to other areas, such as thighs and ankles, only appear in one or two cases compared to the common RTA injuries to the neck and back. If a stepwise or Lasso regression model is applied to remit overfitting by features selection (28 injury types for 88 cases), these injuries may be dropped due to insignificance. The model trained in this manner, however, is biased since it is incapable of predicting the compensation if the claimant only suffers the leg injury for instance.

Following a similar procedure for predicting the cost with single injury severity, we test the cost prediction via the top two and top three most severe injuries with both a linear regression model and non-linear models tested before. Since most predictors in nonlinear models are shown to be non-significant, we only report the linear regression results (also with LOOCV) in Table 6. The lower values for each of the evaluation metrics indicate that involving the top two and top three injuries does not contribute to improved predictive performance compared to a single injury prediction model.

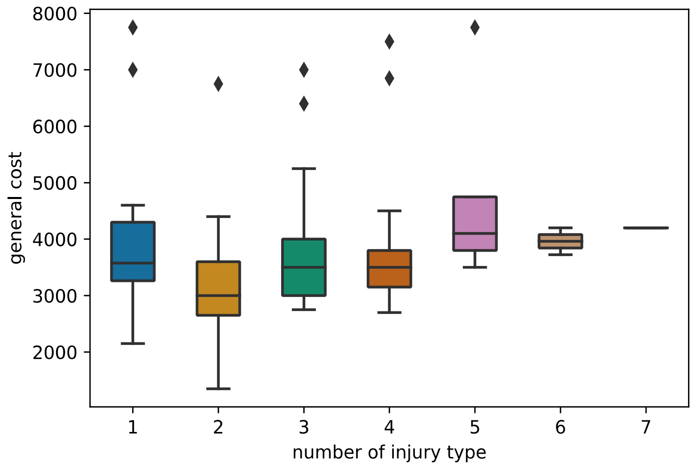

Fig. 7 Relationship between the general cost and number of injury types

Table 6 Performance of linear regression models with top two and three injuries

Model	R^2	MAE	RMSE	AIC	BIC
Linear model with top two injuries	0.479	547.728	907.434	1446.105	1456.014
Linear model with top three injuries	0.486	563.412	904.594	1444.998	1457.384

In summary, we test different prediction models in this section and prefer the square root transformation linear regression, as it both performs very well and better suits to address this research question. However, this proposal is limited by the amount of data available, and we believe that the full injury model would have provided more insight with more data. In the next section, we will demonstrate how to extract the full injury information automatically from the medical report.

4.4 Extract injury information with regular expression

While existing literature largely focuses on algorithm development in isolation from practical applications, this research seeks to not only evaluate the feasibility of RegEx in extracting and structuring medical information but also to demonstrate the usefulness of NLP structured output in a real-world litigation setting, in particular, its utilization in conjunction with decision support systems. Among various information retrieval approaches, the regular expression is chosen here over other syntactic and semantic parsers for its own advantages: a powerful metalanguage that is convenient, interpretable, transferable, customizable, and accessible. Flexibility and generalizability can be enhanced by further refinement of RegEx rules. In this section, we will explain the exploitation of RegEx with some of the most observed formats of clinic reports. For confidentiality reasons, only the structure and narratives of the medical documents are inherited while the identity and case-specific information is all made up for illustrative purposes only.

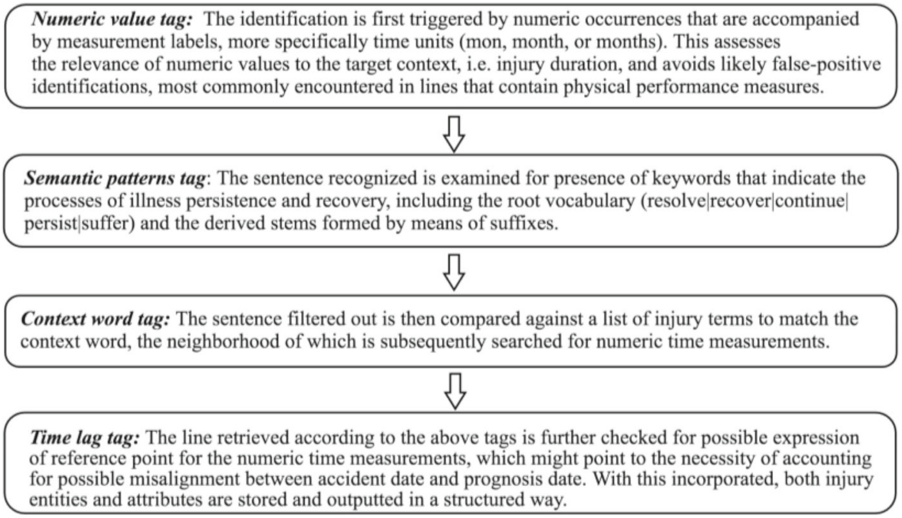

Fig. 8 Extracting entity-attribute information

Automated RegEx-based Learning of Information from Free Text in Medical Documents. To serve the specific purpose described in earlier sections, the primary objective of the medical information extraction in this research is to identify the injury category-severity pair so as to convert unstructured medical narratives into structured machine-interpretable representations. To illustrate, the piece of text *"… expect steady improvement with full resolution of her neck symptoms approximately within 11 months…"* ideally should be compressed onto a lower-dimensional fragment, consisting of the primary illness information (injury name "neck") along with its respective attributes (injury severity "11 months").

A substantial challenge here is that, due to a lack of standard and consistent lexical structure and constraints, the interaction between injury category and attribute may appear in any form, seldom corresponding perfectly with a transparent semantic interpretation. As shown in the following examples, the sentences can be composed in a variety of ways, i.e., the syntax, while the lexical and grammatical structure coalesces to communicate similar semantics, i.e., what injury takes how long to heal. To put it into predicate-argument context, "resolve" can either be the predicate in "injury-resolve-time" or the argument in "expect-resolution in time" or even the attribute of a noun as in "expect steady improvement with full resolution". These variations make identification and interpretation of the sentences based on function features ambiguous and difficult.

To cope with all this variety and uncertainty, RegEx with specific designment, implemented via Python `re` (Van Rossum, 2020), allows great flexibility and bandwidth in locating entities and capturing patterns. This section seeks to explore its potentials in extracting information from medical reports where standardized syntactic structure and format are absent. More specifically, the injury category, along with injury duration, can be abstracted using a combination of number-oriented, label-based, and semantics-driven approaches, which is performed using the following algorithm exhibited in Fig. 8.

With the RegEx module, this is achieved by exploiting a combined use of `re.search()` and `re.findall()`, regular expressions for pattern matching. Here we'll first demonstrate the working of these two RegExs with the simplest scenario, to illustrate how `re.search()` locates target information and how `re.findall()` returns matched instances, while the entire workflow will be presented in later sections. The `re.search()`, `re.search(r'.*((?<=resolve)\w*|(?<=recover)\w*|(?<=continue)\w*|(?<=prognosis)\w*|(?<=suffer)\w*)|(?<=persist)\w*).*\d+ .*',line)`, searches for the pre-defined pattern within each line, which is immediately followed by the if-statement "if match:". When successful, the search results will be held in the match object 'match', with `match.group()` producing the fully matched string. Once the sentence with "semantics" flag that indicates the expected healing time is identified by `re.search()`, `re.findall()` takes this further and derives the injury category and severity respectively via `inj_type = re.findall(r'\b('+'|'.join([injury+'s?' for injury in injury_list])+r')\b',line)` and `inj_t=re.findall(r'(\d+)mon\w*',line)`. For the former, the meta character `\b` performs a whole-words search for injury type. The OR operator, i.e., pipe character |, matches alternatives among the list of injuries, with sub-patterns enclosed by parenthesis to establish a logical group. For the latter, `\d+`finds a sequence of one or more digits preceding the timeframe keyword `\mon\w*` and picks out a list of strings corresponding to the group `(\d+)`, i.e., the count of months. Applying the operations on the Example: Word patterns to identify. C ("Appendix") gives the following ouput: `Injury Type:['back'] Injury Severity:['9'].`

Contextual Dependency. The approach described above, though easy-to-implement and straightforward, is only applicable to scenarios where the injury attribute appears as an immediate neighbor of the injury instance (see Example: Long-term contextual dependencies. A in "Appendix"). For lack of rigid structure and uniform formatting in medical reports, however, this keyword-based blocking is not as appropriate on some occasions due to its incapability to capture contextual dependencies over longer word intervals. For some, the injury type stands a short distance away from its attribute that could be located across multiple boundaries such as phrases, sentences, or even paragraphs, in which case only weak linkages exist through personal pronouns (it/they/etc.) or demonstrative pronouns (this/that/these/etc.). As shown in Example: Long-term contextual dependencies. B ("Appendix"), the specific object "neck" leads the whole paragraph as a stand-alone opening line (split from the rest by \.) and is substituted by the demonstrative pronoun "this" in subsequent references, taking another four sentences and nearly fifty words to arrive at the healing period "2 months" to be extracted.

For some, the target information is laid out in incomplete sentences or case-specific formats such as tables, the structure, and style of which cannot be fully parsed in the multiple layers of file transformations, as is the case exhibited by Example: Case-specific formats. C in "Appendix". It's technically demanding, beyond the capability of established parsers, to keep track of the report structure with the medical records scanned into PDF and the PDF then converted to text, in which processes the links between "Symptom" and "Attributable" are broken. Theoretically, the headers can be identified based on the text properties, coordinates, or relative positionings, but in practice, substantial loss of these locating information is inevitable during the transfer processes.

More complex algorithms handle these problems by explicitly utilizing contextualized word representations to capture the surrounding contextual information, which requires careful validation of the context window size parameter. This study seeks to practically address these challenges and explore real-world applications for the particular task, i.e., to extract injury entity-attribute pairs from narrative medical documents, rather than to develop rigorous pattern recognition algorithms in more complicated domains. Accordingly, to tackle this specific problem of surrounding context, we adopt a feasible and easy-to-implement

Identify context: The first occurrence of injuries *re.findall(r'\b('+'|'.join([injury+'s?' for injury in injury_list])+r')\b',line)*is recognized.

⇩

Build context: The context group is formed by recording the injury instance in a string list *inj_word* to capture the contextual dependencies across boundaries, which could be defined by phrases, sub-sentences, sentences, paragraphs.

⇩

Append context: The labelled context is to be appended at the beginning of the subsequent text fragments *line =' '.join([i_word+':',line]*until the context shifts, i.e. switching to the next injury category and repeating the '***Identify context***' step.

Fig. 9 Addressing long-term contextual dependency

approach by establishing a "content group" that associates the earlier occurrence of the "context word", i.e., injury type in this case, to the sentences that follow. The intuition behind is straightforward: instead of wishfully expecting the NLP techniques to exercise sufficient intelligence itself to appreciate the syntax, we can build the content group by identifying the aforementioned concept of interest and intentionally appending it to the following sentences. By doing this, variable context range has been implicitly accounted for to adaptively learn the long-term dependencies and combine information from multiple sentences. A simple way to achieve this is that, as illustrated in the following algorithm in Fig. 9.

This can be more selectively executed by putting an end when the timeframe keyword "month" is encountered. As seen in the Example:Output generated in "Appendix", the context label "back" has been created and assigned to the relevant text to reflect the logical structure, allowing the explicit linkage between the symptom "back" and timeframe "2 months" to be constructed.

Workflow and Other Considerations. Combining all the above considerations, the flow of the information extraction procedure can be summarized in the following chart in Fig. 10.

Post-accident reference date. On top of the above considerations, one point deserving special attention here is the reference date for the expected recovery time. The following two scenarios represent the two most common medical narrations regarding the timeline description. Example: Without Gap ("Appendix") covers the majority of cases where the expected healing time is gauged and reported with respect to the date of the accident. On some occasions as in Example: With Gap in "Appendix", the medical professional might anchor the judgment as of the examination date, resulting in a time gap between accident and diagnosis to be filled.

The "With gap" case can be flagged by identifying key adjectives indicative of relative positionings, such as "further/following/next/future/another/additional/extra/extend", or the explicit reference of "from the date of examination". Accordingly, this time gap needs to be considered when such a match occurs around the "injury-time" pattern. Applying the same logic, we specify two approaches to extracting the time gap details. The first approach searches for the "accident-prognosis" pattern by locating simultaneous appearance of "post-accident"/ "time since accident" and "prognosis"/ "examination", and capturing the time period adjacent to the keywords. The other approach tackles the time interval directly by recognizing the exact dates of the accident and the report respectively, based on which the discrepancies are assessed. Usually, both dates are accurately reported in the first few pages of documents, oftentimes on the first page. This date information can be extracted by means of `partition()` method, which returns three elements, i.e., before "match",

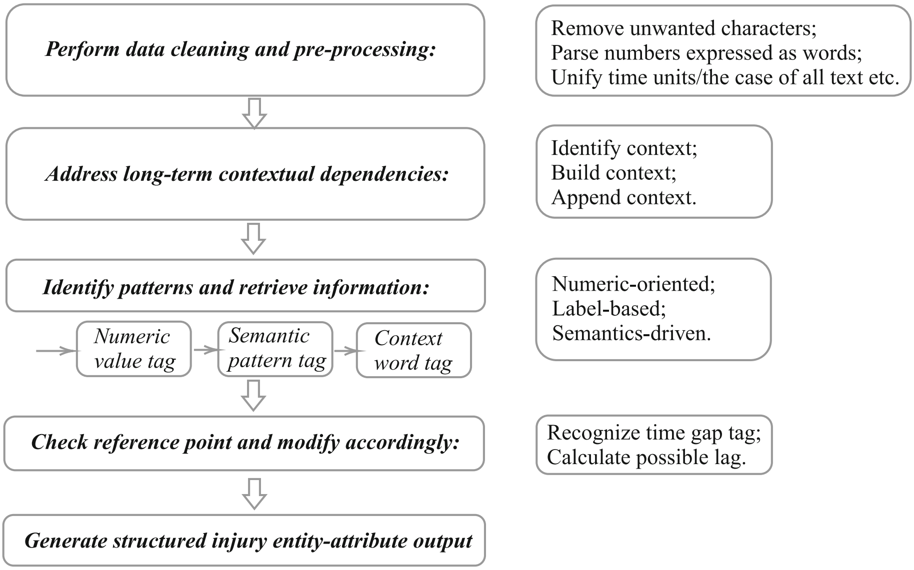

Fig. 10 Flow of the information extraction procedure

"match" and after "match". We can use `\Date of accident` as the keyword to locate the segments that contain the dates of accident and examination, and further identify the dates from the `afterkeyword` part via `datefinder()` method, based on which the time gap is calculated and injury duration is adjusted accordingly.

4.5 Predict injury cost with medical report via SGD linear regression and CNN

Instead of predicting injury cost by employing features manually selected by experts, i.e., injury type and severity caused by the corresponding accident, we also investigate the possibility of predicting the injury cost directly from the medical report via other text mining techniques such as SGD linear regression and CNN. SGD model is trained by updating its parameters iteratively using the gradient of the loss function with respect to those parameters. Stochastic gradient descent differs from regular gradient descent in that it updates the model parameters using a randomly selected subset of the training data at each iteration, rather than the entire dataset. This random sampling helps to speed up the optimization process and can help the algorithm avoid getting stuck in local minima. SGD linear regression is adopted since it has been proved to be an efficient optimizer that is suitable for large-scale and sparse machine learning problems such as NLP (Smith, 2008). Similarly, a CNN is preferred as it is reported to be one of the best performing algorithms in tackling both the text classification and regression problems (Bitvai, 2015), and the CNN used in this research is similar to the previous work using CNN for NLP task (Dereli, 2019; Kim, 2014). On the input layer, the texts will be padded as the same length and the embedding layer will set the vocabulary size and represent each word as k-dimensional real-valued representations (k depends on the word embeddings applied). The subsequent layer consists of a convolutional layer with filters and a kernel size set to the number of words processed at once. A Rectified Linear Unit (ReLU) is usually employed as the non-linear activation function at the output. Following this, a max

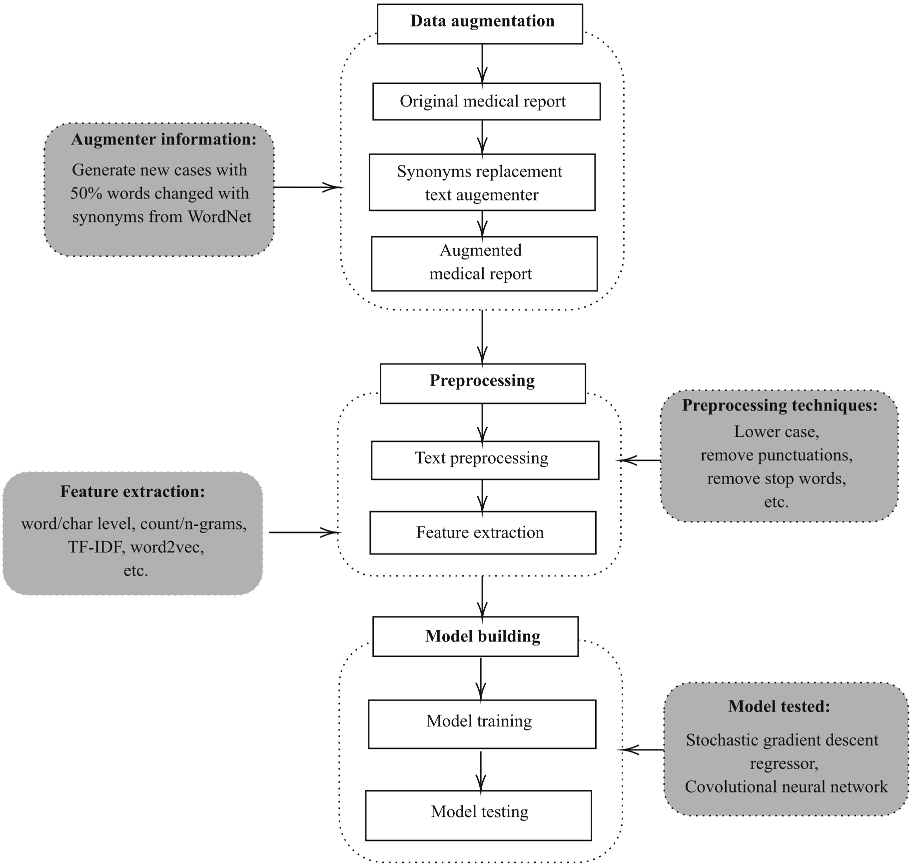

Fig. 11 Document analysis framework

pooling layer is set to merge the output from the convolutional layer, while a flatten layer is used to reduce the three-dimensional output to two dimensions, enabling concatenation. The specific structure and parameters of the CNN are explained in Sect. 5.3 and implemented with Python Keras api. Also, as only 24 full medical reports are available out of these 88 cases, which is insufficient to train and validate the model, we generate some artificial cases with the text data augmentation technique (Wei, 2019). The performance of different methods is compared.

To be specific, we explore how to make predictions in the three steps summarized in Fig. 11. First, we augment the original data set for training and testing. Then, we pre-process the raw text data and extract features with different methods to represent the text and lastly, we build various predictive models and evaluate their performance.

4.6 Data preparation

One of the main challenges in this research, like many other practical exploratory research problems, is the scarcity of data sources. In this research, although we collected 88 cases with brief injury descriptions and the corresponding general cost, only 24 of these cases have accompanying full medical reports. To better evaluate the different models with reasonably adequate samples, we adopt the text data augmenter from the `EDA` package (Wei, 2019). Explicitly, we solely utilize the word-level augmenter which substitutes words with their synonyms from WordNet (Miller, 1995) because the main purpose of augmentation for this task is to generate more medical reports without changing the meaning of the original words (at least similar), there would otherwise be little point in predicting costs from medical reports consisting of many unimportant words. In this research, we generate ten new cases for each original case by substituting 50% of the words with synonyms so that we get 264 cases in total. The 50% similarity is chosen because, on the one hand, new cases should be different from the original cases otherwise they will result in overfitting; on the other hand, too much variation will cause the meaning of the new cases to be so different from the original cases that it would not make sense to relate the general cost to the new cases. We will further discuss this issue in Sect. 7.

4.7 Pre-processing and feature extraction

After applying general text data pre-processing techniques such as lower cases, punctuation, and stop words removal, we extract different features with `TfidfVectorizer` from the `scikit-learn` package. We choose the TF-IDF vectorizer as it is an efficient representation of text that not only focuses on the frequency of words present in the corpus, but also provides the importance of the words. By employing the TF-IDF vectorizer, we can dismiss less important words and build a less complex model by reducing the input dimensions. To be specific, we test the TF-IDF models and n-grams (we use $n = 4$ and $n = 5$ to represent a reasonable term) TF-IDF models at both the word level and character level.

Top ten terms exclude those with names (for privacy reasons) obtained from these feature representation approaches are listed in Table 7. While the word-level representations make sense to some extent as they involve some terms such as "score", "explanation", it is difficult to interpret the character-level features like "a", "e". Another approach to symbolize text features is using word embeddings such as word2vec (Mikolov, 2013), GloVe (Pennington, 2014), and FastText (Joulin, 2016), which represents each unique word with a specific vector of numbers. The advantage of word embeddings is that they are dense vectors that retains

Table 7 Top ten terms identified by different vectorizer

Text vectorizer	Top ten terms
Word level TF-IDF	"Reputation", "score", "pretermit", "neglect", "lack", "escape", "billings", "explanation", "overleap", "fille"
Word n-grams TF-IDF	"Convention come face uncomfortableness", "convention look causa soreness", "rule come drive soreness", "rule look causa soreness", "convention come crusade uncomfortableness", "formula come campaign uncomfortableness", "index number chance event", "pattern appear reason irritation", "60 normal appeared cause discomfort", "medical checkup news"
Char level TF-IDF	"1", "o", "n", "r", "0", "i", "t", "a", "e",""
Char n-grams TF-IDF	"Fire", " g wom", "mr gr", "mr g", "iss", "mr wo", "g lad", "ung l", " shaw", "oley"

the semantics of different words compared to the BOW models and has been widely utilized in the sentiment analysis and text regression with different neural networks (Bitvai, 2015).

4.8 Model building and testing

With the different features described above, we train and test an SGD linear regression model (with 80% of the total data as training set and 20% as the testing set) and the result is reported in Table 8.

Models with word-level TF-IDF and character n-grams TF-IDF as features perform much better than others, achieving a predicted R^2 value of around 0.8. One possible reason why these two models perform better than others may be that the n-grams character expressions are more word-like, which seems to suggest that representation by words is more applicable to this problem. Also, with the pre-trained word embeddings of word2vec (`GoogleNews-300d-1 M.vec`), FastText (`wiki-news-300d-1 M.txt`), and GloVe (`glove.6B.100d.txt`), we test the predicting performance of a classic CNN (Kim, 2014) using the same data partitioning of the SGD linear regression model. Here, the first hidden layer is the embedding layer. Depending on the chosen pre-trained word embeddings, the embedding dimension could be 300 (word2vec and FastText) or 100 (GloVe). Then, on the convolutional layer, the configuration adapted is used with 32 filters/channels and a kernel size of 8 with a ReLU activation function. Following a max-over-time pooling layer, a flatten layer and a fully connected layer, the output layer provides a predicted value of the general cost. Table 9 demonstrates that the CNN with different word embeddings performs similarly around 0.74–0.79, which is slightly worse than the best result of predicting with SGD linear regression though the difference is not obvious. Therefore, both the SGD linear regression model and the CNN appear to predict the cost from the medical report well, although the fact that augmented data is employed in the experiment should not be ignored.

Table 8 SGD linear regression prediction results with different features

Text vectorizer	MAE	RMSE	R^2
Word level TF-IDF	359.01	533.59	0.76
Word n-grams TF-IDF	798.91	1243.39	− 0.29
Char level TF-IDF	727.92	1085.87	0.01
Char n-grams TF-IDF	263.51	393.02	0.87

Table 9 Prediction results comparison with different models

Model	MAE	RMSE	R^2
CNN with word2vec embedding	308.97	476.23	0.78
CNN with FastText embedding	340.72	465.16	0.79
CNN with GloVe embedding	389.86	511.50	0.74
SGD linear regression	263.51	393.02	0.87

4.9 Economic gains

The improved efficiency of applying the proposed method can translate into monetary values in our case study with the 88 cases. Table 10 provides a comparative analysis among the discrepancy (absolute difference) between the "intrinsic costs" (proxied by judges' final verdicts) and offers from both parties (C is short for the Claimant and D is short for the Defendant in the table) at various phrases and the difference between the "intrinsic costs" and square root transformation model implied values.

Comparing the intrinsic costs, model-implied values, and both parties' first offers (the only offers accessible to the judges) reveals that, the model-implied values better reflect the judges' perceptions than claimants' or defendants' expectations, reducing the expectation-reality discrepancies by over 40%/30% on average for claimants/defendants, respectively. Had the proposed framework been adopted across the board, both parties would immediately arrive at a consensus monetary number that represents a more sensible estimate of the intrinsic value, which leads to the efficient resolution of disputes without incurring unnecessary expenses or consuming legal resources. If only one party embraces the practice, he would be in a more advantageous and informed position with more accurate estimates and a better chance of winning. On top of that, automation facilitated by IT tools unlocks opportunities for a more cost-effective, time-saving, and productive workflow by replacing routine and repetitive manual operations and activities.

As we could reasonably expect, both parties make concessions as the negotiation progresses, as suggested by the first and final offers from both parties. If we consider the final offers as the negotiators' "bottom lines" that represent their most precise and confident valuations, these figures are by no means closer to the actual outcomes than the model-implied ones. In summary, this section demonstrates how the AI-architecture enhances the evaluation of claims, increases efficiency of service processes, and thereby leads to substantial economic gains.

Table 10 Comparative analysis among offers

Statistic	C's first offer	D's first offer	C's final offer	D's final offer	Predicted value
Cumulative difference	82,223.10	69,675	53,686	56,155	**46,105.711**
Mean	934.35	791.76	689.61	655.17	**523.93**
SD	906.70	737.97	882.68	721.49	663.03

Bold is used to emphasize the advantages of using the model

5 Discussion and conclusion

Despite the growing discussion of applying AI/data science in the financial realm and insurance especially from both the academia (Hendershott, 2021) and industry (Balasubramanian, 2018), little prior research has been conducted that meets the genuine needs of the industry and utilizes primary data. In this research, we demonstrate a practical application of AI in the insurance or legal professional service sector. We argue that this is a valuable topic since it not only brings benefits to the involved companies such as insurance or legal professional service companies but also to ordinary people and the whole society. Specifically, we examine the relationship between the injury symptoms listed in the medical report and injury cost and process the medical documents to predict the injury cost automatically via NLP techniques. We make several research contributions and identify implications for practice.

First, we demonstrate the viability of using AI effectively and efficiently to improve the decision-making process in the professional service sector in which many tasks are essentially repetitive, with a specific case of a UK legal service firm. Depending on the data availability, we propose a general framework to deal with different situations, i.e., those with a small or large amount of data. In the first situation, we build features manually with expert knowledge and conduct more traditional statistical analysis accordingly. Since the increase in injury types indicates an increasing general cost as Fig. 7 demonstrates, we further examine the prediction of the general cost with the top two and three most severe injuries as discussed in Sect. 4, whereas none of them outperforms the single injury regression model. Based on the existing data, it appears the square root transformation regression using the most severe injury can predict the cost well, while we don't rule out the possibility that a full injury model might be suitable for more general situations. The biggest advantage of this approach is its interpretability, which is of particular importance in the context of professional service where knowledge is the treasure.

Second, to showcase how automation can be achieved in this context, we explore extracting the features with RegEx. To obtain more reliable results from the second approach, many cases are required thus better suited to this era of big data. Rather than adopting the predictions of these models directly, we take advantage of them as a quick reference to help lawyers make initial judgements. We develop a rule-based NLP workflow by exploiting and leveraging a collection of keywords, rules, and logic. Though dependent on a set of pre-defined features, the rule-based frame- work well leverages domain-specific knowledge, fully reflects task-specific objectives, and achieves great interpretability and transparency. To be more specific, we combine the use of label-based, numeric-oriented, and semantic logic. Towards the last point, we address the long-range semantic dependencies by proposing an implementable and traceable procedure that explicitly deals with the semantic context.

Last, we investigate how to apply different machine learning techniques to predict the general cost from medical reports directly without manual feature selection. Using the character level n-grams ($n = 4, 5$) TF-IDF values as input features, the SGD linear regression model achieves a predictive R^2 value of 0.8 and a CNN utilizing different word embeddings performs similarly, with the highest predictive R^2 of 0.79. It is worth noting that the testing result is from augmented data (50% synonym substitution) and may suffer from overfitting problems. In machine learning and computer vision research, data augmentation is commonly used in deep learning tasks such as image recognition and object detection (Shorten, 2019) because it is a powerful technique that can improve the model performance, especially when the availability training data is limited (Perez, 2017). The scarcity of data is a common challenge faced in various sensitive industries, such as in healthcare where data access is strictly protected due to privacy concerns (Perez, 2017). Likewise, this study is encountering data limitations that are even more challenging as the data involved is not only affected by patient privacy issues but also heightened concerns regarding legal data confidentiality. At the same time, NLP is still in its early stages of applying data augmentation compared to computer vision (Shorten, 2021) and there is a lack of a researcher-proven, widely accepted approach to NLP augmentation. Therefore, we believe that augmenting NLP data in this study based on existing research (i.e., the EDA package) is a reasonable endeavour, which may provide some insights for other researchers in confidential industries, such as insurance and legal, who are also grappling with data limitations.

Despite the contributions highlighted above, there are some limitations of this study and how to address them points out several important future research avenues. As discussed above, a natural progression of this work is to analyze more predictive models if a larger data set is available. The scarcity of data makes it tricky to test the full injury model and validate the machine learning algorithms that require a much larger data set. In this research, we try to overcome the difficulty of data obtaining via NLP data augmentation, but it is possible that this augmentation will cause overfitting by introducing too many similar cases or lead to underfitting since the newly generated cases do not relate well to the cost. Studying this issue is undoubtedly an interesting topic for future research.

Also, the use of rule-based NLP procedures relies on pre-identified expressions and structures as well as pre-defined logical rules, which are not perfectly readily applicable in other situations and need to be refined with additional knowledge bases and development efforts. Future directions for research may include enhanced exploitation of existing well-curated clinical resources and domain-specific ontologies, and a simpler approach to adapting regular expressions to other tasks. For example, Murtaugh et al. (2015) presented a learning algorithm that automates the process of designing and developing regular expressions. Besides this learning approach that automatically identifies and refines patterns to be fed into the rule-based system, machine learning or deep learning techniques can also be exploited to either establish the pre-determined logic rules or provide supplementary information. So far, this study has been pursuing the rule-based or learning-based branches in a separate manner to extract pertinent information from the text and then characterize its informational value for claim costs. In a broader application context, outputs of one approach can serve as inputs of the other to combine the advantages of both. More specifically, machine learning algorithms can aid the recognition of patterns that account for complex semantic lexicon or dependence, while rule-based methods can produce explicit and interpretable features to enhance machine learning performance. This provides appealing directions for future research.

Appendix

A Interview

The following is a summary from the interviews and workshops within and outside the cooperated legal service company to clarify the legal background related to the RTA claims from 2018 to 2020. These interviews are taken as part of the research project of exploring AI and statistical predictive models in civil litigation. It is worth noting that these interviews came from different lawyers of different levels. We refer to them collectively as the lawyer for convenience.

Author: May I know what are the key decision questions you are interested in?

Lawyer: An important part of the cases we deal with daily are of low value (up to £25,000) cases of personal injury from the motor accident. Before choosing to go to court, we will negotiate with the other party on compensation for the case. Our lawyers/claim handlers run tasks on the Portal system and duplicate them on our company's case management systems.

Author: Could you explain the Portal system in detail?

Lawyer: The Portal is a tool for processing personal injury claims up to the value of £25,000. A road traffic accident (RTA) Portal was introduced in 2010 (£10,000 claims limit) and vertically extended to include motor claims up to £25,000 and horizontally extended to include Employers' Liability/Public liability (EL/PL) claims £1000 to £25,000. It is a web based claims platform where all communication is via electronic means. All information must be sent via www.claimsportal.org.uk. Both sides will negotiate and seek to settle claim and oral or paper hearing will be issued when damages cannot be agreed.

Author: We suppose that there will be costs incurred during the negotiation, is that true?

Lawyer: Yes, there will be cost, and the cost is a little complicated. Firstly, there is the fixed cost which is paid by the defendant such as the stage cost, medical report cost, etc. The detailed stage cost can be found on PART 45—FIXED COSTS. Secondly, the miscellaneous, which is usually different in different cases. Generally speaking, the longer the negotiation takes, the more parties need to pay. There are also result depended fees if parties fail to reach an agreement during the negotiation and go to court. For example, the stage three cost and issue fee will be paid by the defendant if his/her offer will be beaten.

Author: Thanks for clarification. We browsed the litigation documents and find that it basically takes at least two years from the occurrence of the accident to the final judgment of the court. Is this too long for both parties, especially considering that they are low-cost cases?

Lawyer: That's a good point. There are many reasons for the delay or even the inability to resolve through negotiation. One of the important reasons is that the two parties cannot reach an agreement on the amount of the final compensation.

Author: Does that mean that if we look up the results of a large number of historical cases and use these results to predict the compensation of a similar case, it will help us better estimate and avoid some unnecessary losses?

Lawyer: Yes, I think so. It will be very helpful if we can build some kind of systems to systematically analyze a large amount of historical data and draw certain conclusions.

Author: Great. Another thing we found is that two types of damage costs are negotiated, the general cost and special cost. What is the difference between them?

Lawyer: Basically, the general cost is entirely decided by the claimant's injury type and severity, which will be estimated by the lawyers/claim handlers with the Judicial College Guidelines. On the other hand, the special cost includes costs such income loss, repairs cost, physiotherapy cost caused by the accident and so on. The special cost is quite case-based, and we can talk about general cost first.

Author: Thanks for that. I can see that there is a corresponding compensation for each type of injury based on the severity. But what if the claimant suffered two or more injuries? Will you add up all the injury compensation?

Lawyer: I don't think we will add up the compensation for all injuries, but will focus on the most serious injury, and then consider the others as appropriate. As for the discretion, it is based on the solicitor/claim handler's experience. It will be very interesting to get an insight of the relationship between injury and compensation.

B code

Entity-attribute information retrieval

```
match = re.search(r'.*((?<=resol)\w*|(?<=recover)\w*|(?<=continue)\w*|(?<=prognosis)\w*|
    (?<=suffer)\w*)|(?<=persist)\w*).*\d+.*', line)
if match:
    print('Match: %s' % (match.group(0))
    inj_type = re.findall(r'\b('+'|'.join([injury+'s?' for injury in
        injury_list])+r')\b',line)
    print('Injury Type: %s' %(inj_type))
    if len(inj_type) == 1:
        inj_type = [i_type[:-1] if i_type[-1]=='s' else i_type for i_type in inj_type]
        inj_t = re.findall(r'(\d+) mon\w*',line)
        print('Injury Severity: %s' %(inj_t))
        if inj_t:
            Summary_Table.loc[file_num, inj_type] = inj_t[0]
```

Long-term contextual dependency

```
content=list()
inj_word=[]
for line in report['content'.split('.')]:
    if inj_word:
        for i_word in reversed(inj_word):
            line=''.join(i_word+':',line)
            match_inj = re.findall(r'\b('+'|'.join([injury+'s?' for injury in
                injury_list])+r')\b',line)
            if match_inj:
                inj_word = [i of i in list(set(match_inj))]
            if re.findall(r'\d+ mon\w*',line)
                inj_word = []
content.append(line)
```

Gap between accident and prognosis

```
keyword = 'Date of accident'
PageObj = object.getPage(0) #read the first page of the medical report
Text = PageObj.extractText()
beforekeyword, keyword, afterkeyword = Text.partition(keyword)
matches = datefinder.find_dates(afterkeyword)
detect_dates=[]
for match in matches:
    match = match.strftime("%Y-%m-%d")
    detect_dates.append(match)


d1 = datetime.strptime(detect_dates[0], "%Y-%m-%d")
d2 = datetime.strptime(detect_dates[1], "%Y-%m-%d")
abs((d2 - d1).days)
```

C Text example

Word patterns to identify

Example: Word patterns to identify

A. I believe that the frontal **headaches** will **resolve** four to five months from the date of the accident.

B. **Neck**. This was localized to the posterior aspect. …… This symptom **resolved** after 2 months.

C. In my opinion, on the balance of probability, I do expect steady improvement with full **resolution** of her **back** symptoms approximately within 9 months with recommended treatment from the date of this road traffic accident.

D. **Neck**. In my opinion, this was entirely due to the material incident. I would expect a full recovery. I would expect the **recovery** to occur over the following 5 months.

Long-term contextual dependencies

Example: Long-term contextual dependencies

A. In my opinion, on the balance of probability I do expect steady improvement with full resolution of her **neck** symptoms approximately within **11 months** with recommended treatment from the date of this road traffic accident.

B. (1) **Neck**. This was localised to the posterior aspect. The onset of this symptom was immediately after the material incident. The initial level of severity, on a scale of mild-moderate-severe, the symptoms was considered by the patient to be moderate to severe. This symptom has resolved. This symptom resolved after **2 months**.

Case-specific formats

Table 11: Example: Case-specific formats

Symptom	Attributable
Neck pain and stiffness	8 months
Soft tissue injury/superf to chest	11 months
Soft tissue injury to right knee	11 months
Superficial injury/graze to face	10 months

Time gap

Example: Time Gap

Example: With Gap

A. I have examined here approximately 2 months after this road traffic accident....... In my opinion, on the balance of probability I do expect steady improvement with full resolution of her neck symptoms approximately **within 11 months with recommended treatment from the date of this road traffic accident**.

B. ...... this symptom resolved **after 2 months**.

Example: Without Gap

A. Neck, In my opinion, this was entirely due to the material incident. I would expect a full recovery. I would expect the recovery to occur over the **following** 5 months.

B. ...... shoulder pain will resolve to the pre-accident level over the **next** 7 months.

C. ...... he will return to normal duties in 9 months **from the date of examination**.

Output

Example: Output

back: This was localised to the posterior aspect

back: the onset of this symptom was immediately after the material incident

back: The initial level of severity, on a scale of mild-moderate-severe, the symptom was considered by the patient to be moderate

back: This symptom has resolved

back: This symptom resolved after **2 months**

References

Abrahams, A. S. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 6*(24), 975–990.

Araz, O.M.-M. (2020). Role of analytics for operational risk management in the era of big data. *Decision Sciences, 51*, 1320–1346.

Atkinson, K.A.-C. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence, 2989*, 103387.

Avgerinos, E. A. (2018). Task variety in professional service work: When it helps and when it hurts. *Production and Operations Management, 27*, 1368–1389.

Balasubramanian, R. A. (2018). *Insurance 2030—the impact of AI on the future of insurance*. McKinsey & Company.

Bitvai, Z., & Cohn, T. (2015, July). Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (Volume 2: Short Papers) (pp. 180–185).

Boone, T. A. (2001). The effect of information technology on learning in professional service organizations. *Journal of Operations Management, 19*, 485–495.

Cui, R. A. (2022). AI and procurement. *Manufacturing & Service Operations Management, 24*(2), 691–706.

Dereli, N. A. (2019). Convolutional neural networks for financial text regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

Dobrzykowski, D. D. (2016). Examining pathways to safety and financial performance in hospitals: A study of lean in professional service operations. *Journal of Operations Management, 42*, 39–51.

Goldstein, S. M. (2002). The service concept: The missing link in service design research? *Journal of Operations Management, 20*, 121–134.

GOV.UK. (2021). *Payment of court fees in road traffic accident related personal injury claims under the new.* Retrieved November 2021, from https://www.gov.uk/government/publications/whiplash-reform-programme-information-and-faq/payment-of-court-fees-in-road-traffic-accident-related-personal-injury-claims-under-the-new-small

Harris, Z. S. (1954). Distributional structure. *Word, 10*, 146–162.

Hendershott, T. A. (2021). FinTech as a game changer: Overview of research frontiers. *Information Systems Research, 32*, 1–17.

Huang, M.-H.A. (2021). Engaged to a robot? The role of AI in service. *Journal of Service Research, 24*, 30–41.

Jackson, R. M. (2010). *Review of civil litigation costs*. The Stationery Office.

Jagannatha, A. N., & Yu, H. (2016, June). Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting* (Vol. 2016, p. 473). NIH Public Access.

Joulin, A. A. (2016). Fasttext. zip: Compressing text classification models. arXiv:1612.03651.

Justice. (2017). *Pre-action protocol for low value personal injury claims in road traffic accidents from 31 july 2013.* Retrieved January 10, 2021, from https://www.justice.gov.uk/courts/procedure-rules/civil/protocol/pre-action-protocol-for-low-value-personal-injury-claims-in-road-traffic-accidents-31-july-2013.

Karwan, K. R. (2006). Integrating service design principles and information technology to improve delivery and productivity in public sector operations: The case of the South Carolina DMV. *Journal of Operations Management, 24*, 347–362.

Kim, Y. (2014a). *Convolutional neural networks for sentence classification*. Association for Computational Linguistics.

Kumar, S. A. (2018). Research in operations management and information systems interface. *Production and Operations Management, 27*, 1893–1905.

Lewis, M. A. (2012). How different is professional service operations management? *Journal of Operations Management, 30*(1–2), 1–11.

Lexis. (2020). *Valuing general damages—checklist*. Retrieved May 2021, from https://www.lexisnexis.com/uk/lexispsl/personalinjury/document/393875/5N4N-02G1-F18H-M3GD-00000-00/Valuing_general_damages_checklist.

Medvedeva, M. A. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law, 28*, 237–266.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 11*, 39–41.

Molot, J. T. (2009). A market in litigation risk. *The University of Chicago Law Review, 76*, 367.

Murtaugh, M.A.-T. (2015). Regular expression-based learning to extract bodyweight values from clinical notes. *Journal of Biomedical Informatics, 54*, 186–190.

Pennington, J. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.
Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621.
Ray, G., Muhanna, W. A., & Barney, J. B. (2005). Information technology and the performance of the customer service process: A resource-based analysis. *MIS quarterly*, pp. 625–652.
Salton, G. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*, 513–523.
Samek, W. (2019b). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.
Shorten, C. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48.
Shorten, C. T. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*, 1–34.
Smith, D. A., & Eisner, J. (2008, October). Dependency parsing by belief propagation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 145–156).
Taddeo, M. (2018). How AI can be a force for good. *Science, 361*, 751–752.
Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM, 11*(6), 419–422.
Turchin, A. (2006). Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association, 13*, 691–695.
Vacek, T. (2019). Litigation analytics: Case outcomes extracted from US federal court dockets. In *Proceedings of the Natural Legal Language Processing Workshop*.
Van Rossum, G. (2020). The Python Library Reference, release 3.8.2. *Python Software Foundation, 16*.
Von Nordenflycht, A. (2010). What is a professional service firm? Toward a theory and taxonomy of knowledge-intensive firms. *Academy of Management Review, 35*(1), 155–174.
Wei, J. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv:1901.11196.
Woolf, H. (1995). *Access to Justice: Interim report to the Lord Chancellor on the civil justice system in England and Wales*. Lord Chancellor's Department.
Xiao, C. (2018). Cail2018: A large-scale legal dataset for judgment prediction. arXiv:1807.02478.

Authors and Affiliations

Wen Zhang[1] · Jingwen Shi[2] · Xiaojun Wang[3] · Henry Wynn[4]

✉ Wen Zhang
phd14wz@mail.wbs.ac.uk

Jingwen Shi
phd15js@mail.wbs.ac.uk

Xiaojun Wang
x.wang.18@bham.ac.uk

Henry Wynn
H.Wynn@lse.ac.uk

1 Business School, University of Bristol, 11-13 Tyndall's Park Road, Clifton, Bristol BS8 1PY, UK

2 State Street Global Advisors, London, UK

3 University of Birmingham, Birmingham, UK

4 London School of Economics and Political Science, London, UK