International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Resolving insurance claims with Artificial Intelligence powered decision making

Balaji Dhashanamoorthi *

Master of Engineering, Control and Instrumentation, CEG, Anna University, Chennai, India.

International Journal of Science and Research Archive, 2023, 10(02), 255–271

Publication history: Received on 03 October 2023; revised on 16 November 2023; accepted on 19 November 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.10.2.0946

Abstract

This study explores how Artificial Intelligence (AI) can improve the speed and efficiency of dispute resolution in road traffic accident (RTA) insurance claims, which can benefit the society and the economy. We propose and apply a systematic AI-based method to estimate the costs and guide the negotiation process, instead of relying on official guidelines and lawyer expertise. We use 88 real-life RTA cases and find a strong correlation between the final judicial cost and the length of the most severe injury, with a high predicted R2 value of 0.527. We also demonstrate how various AI tools can help with information extraction and outcome prediction:

- How regular expression (regex) can obtain accurate injury data for further analysis;
- How different natural language processing (NLP) techniques can make predictions directly from text.

Our RegEx framework can automatically extract information from different report formats; different NLP methods can produce similar reasonable results. This research shows how AI can be used for social good to transform legal-related decision-making processes, support legal actions, and optimize legal resource use.

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

1. Introduction

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 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:

- AI/NLP-Driven Architecture for PSO,
- Information Extraction Procedure, and
- 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.

* Corresponding author: Balaji Dhashanamoorthi

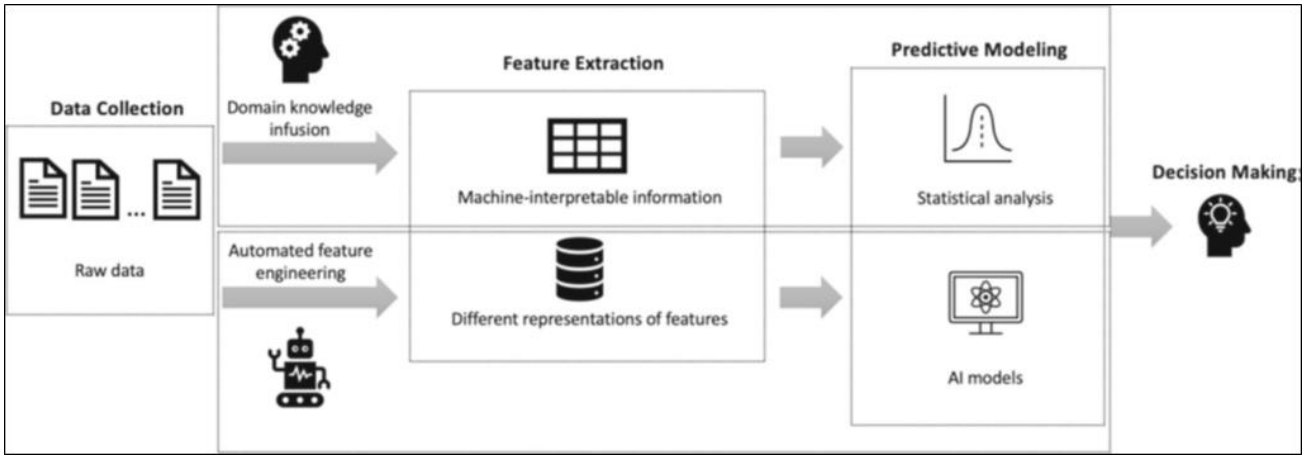

Figure 1 Proposed AI-driven decision-making framework

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 ofAI. 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 R2 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:

- Decision-making Pipeline Proposal,
- Relevance to Legal Services, and
- 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.

2. Resolving insurance claim disputes with a text-based analytic approach

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.

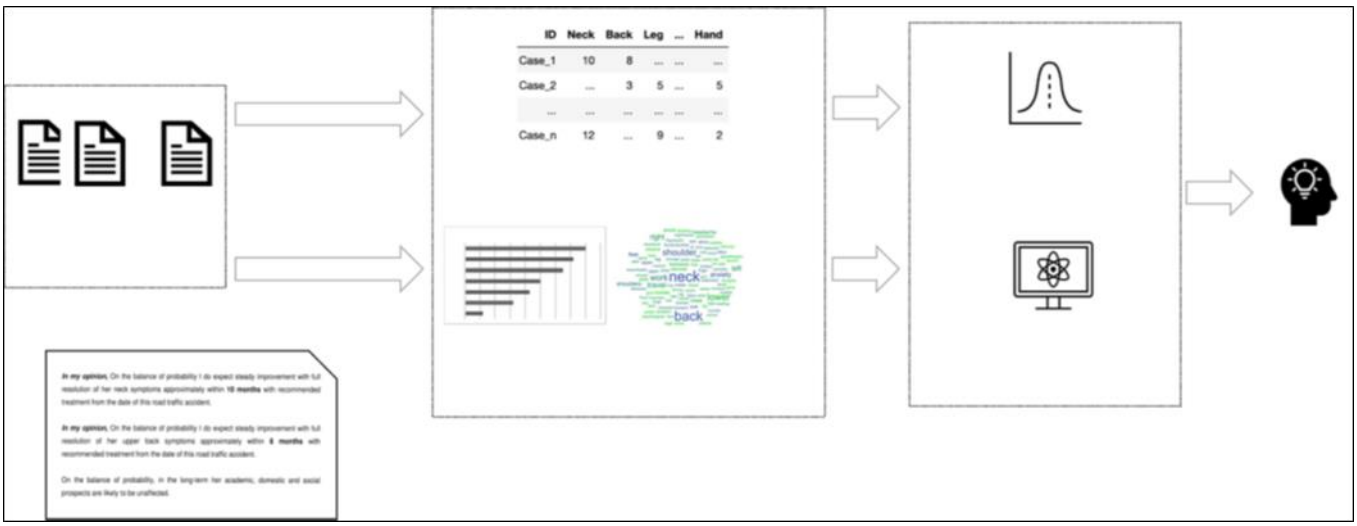

Figure 2 Integrated textual analytic framework

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 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 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.

3. 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.

3.1. 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.

3.2. 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.

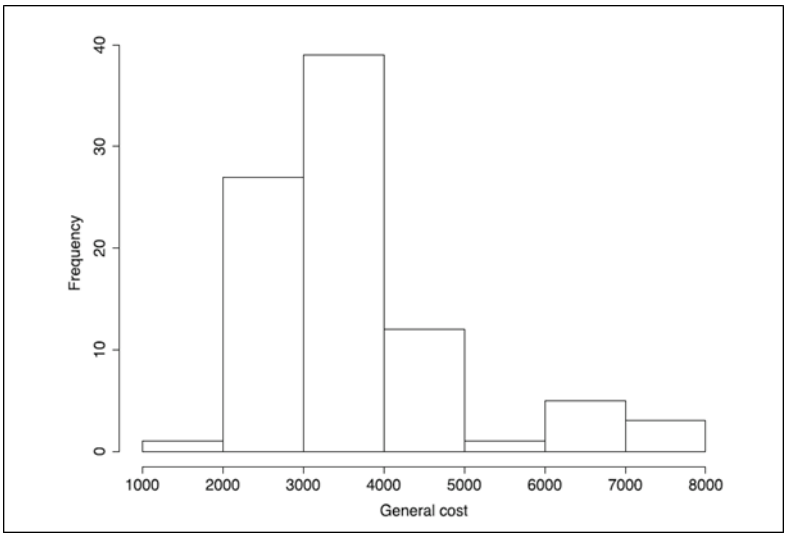

Figure 3 General cost histogram

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		

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, 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.

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

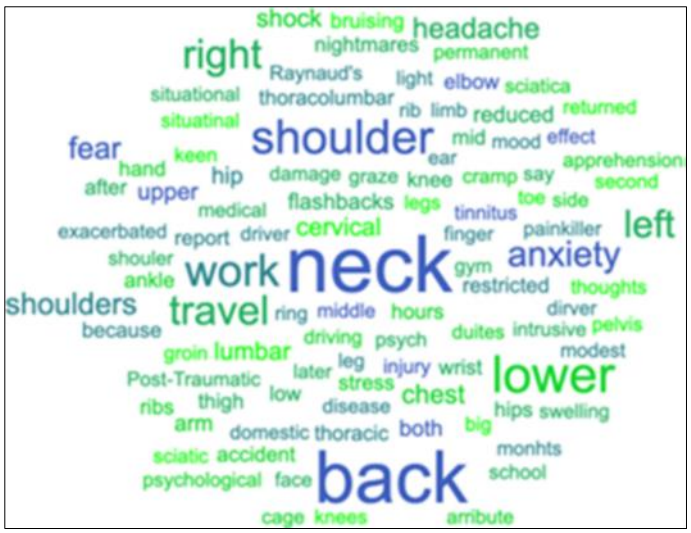

Figure 4 Medical reports word cloud

3.3. 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.

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 (a0)	2448.54	230.42	10.626	< 2
Injury severity (a1)	119.62	13.27	9.015	$e-16^{***}$ $5.53e-14^{***}$
Dummy variable (a2)	76.08	222.39	0.342	0.733
Dummy variable (a3)	387.40	671.81	0.577	0.566

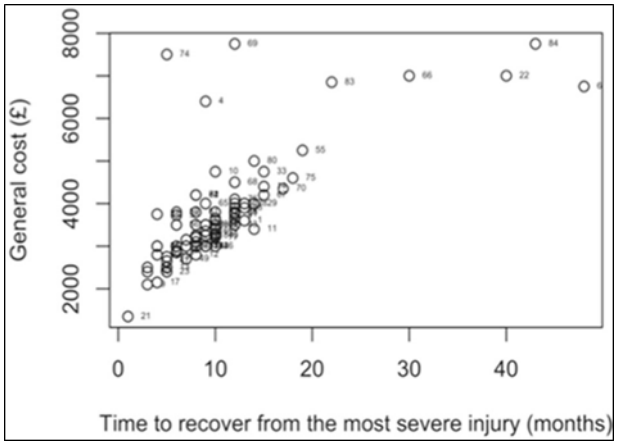

Figure 5 Injury severity and general cost relationship

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 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

$$yi \_ \alpha 0 + \alpha 1xi + \alpha 2 + \alpha 3 + \_i, \quad (1)$$

where i _ 1, 2, ..., n (n _ 88 in this research), yi and xi are the general cost and months to resolve from the most severe injury of casei . α0 and α1 represent the intercept and the coefficient of xi and _i denotes the error term. α2 and α3 are dummy variables for torso and limbs respectively, i.e., when α2 _ 1, the injury belongs to the torso or limbs when α3 _ 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 stabilize 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 (x2) and third-degree polynomials (x3) 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 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 R2 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.

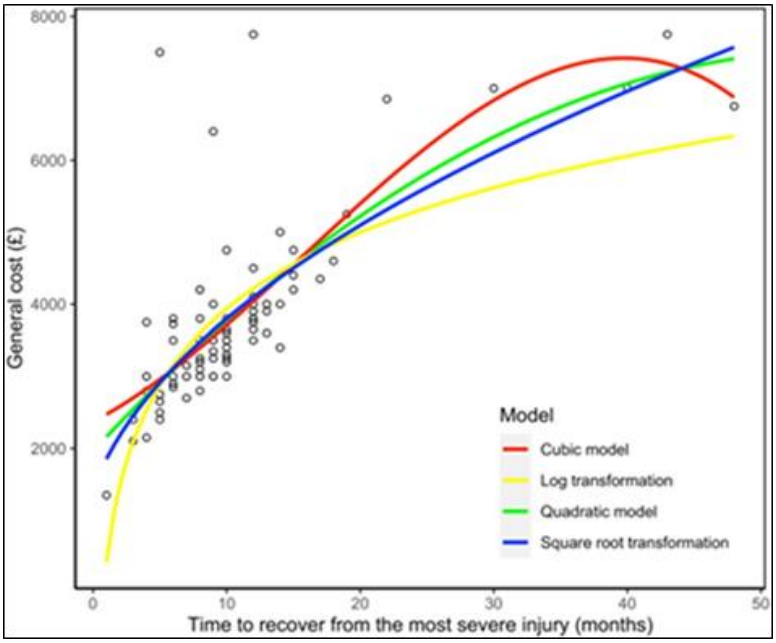

Figure 6 Different non-linear model fittings

Table 4 Predictive performance of different nonlinear models

Model	R2	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

Although the Cubic model has the largest R2 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 predict the general cost and the regression model is

$$yi \_ \beta 0 + \beta 1x1/2i + Ei, (2)$$

where i _ 1, 2, ..., n (n _ 88 in this research), yi and xi are the general cost and months to resolve from the most severe injury of case i , β0 is the constant and β1 is the coefficient of the square root ofxi . 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.

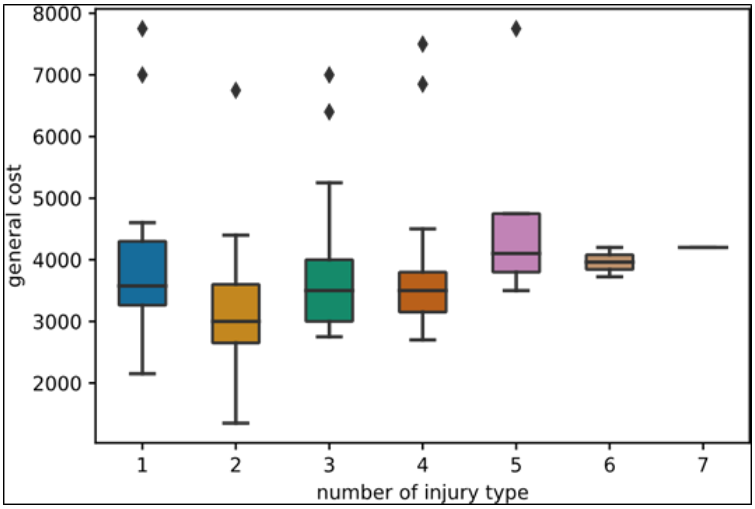

Figure 7 Relationship between the general cost and number of injury types

Table 5 Performance of linear regression models with top two and three injuries

Model	R2	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.

3.4. 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 structuringmedical 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.

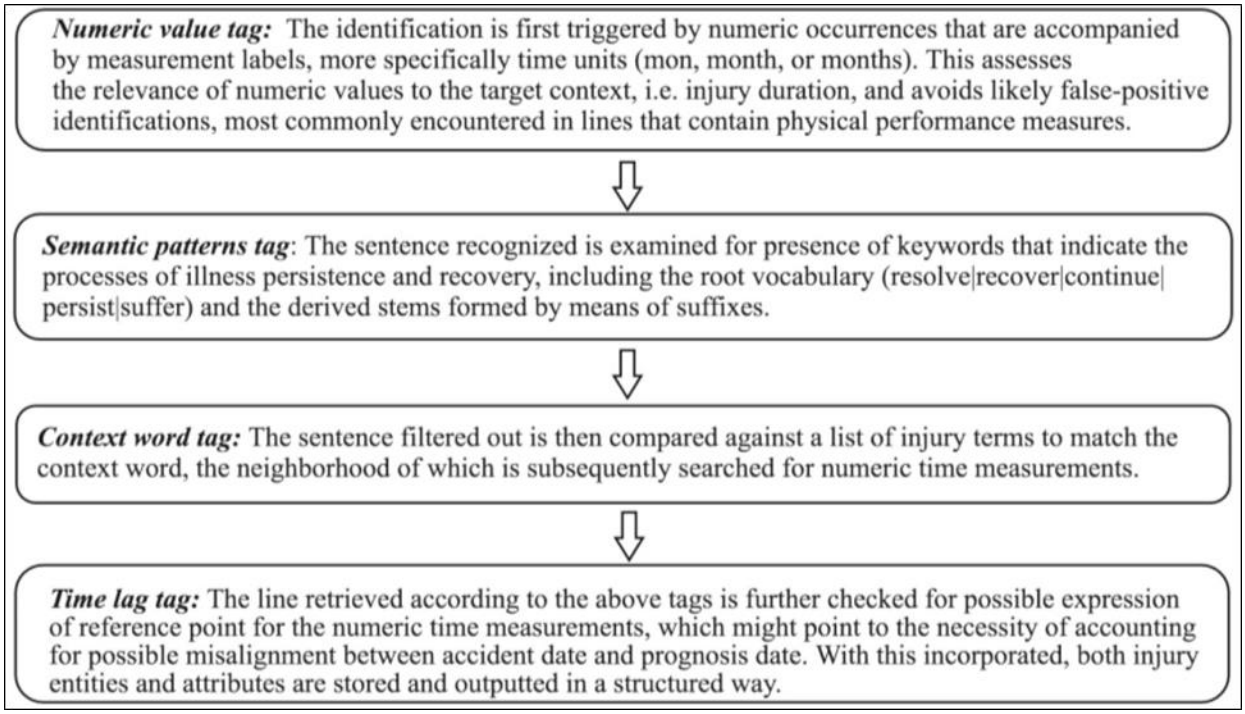

Figure 8 Extracting entity-attribute information

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 attributemay 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 Figure 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.

There.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 acrossmultiple 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

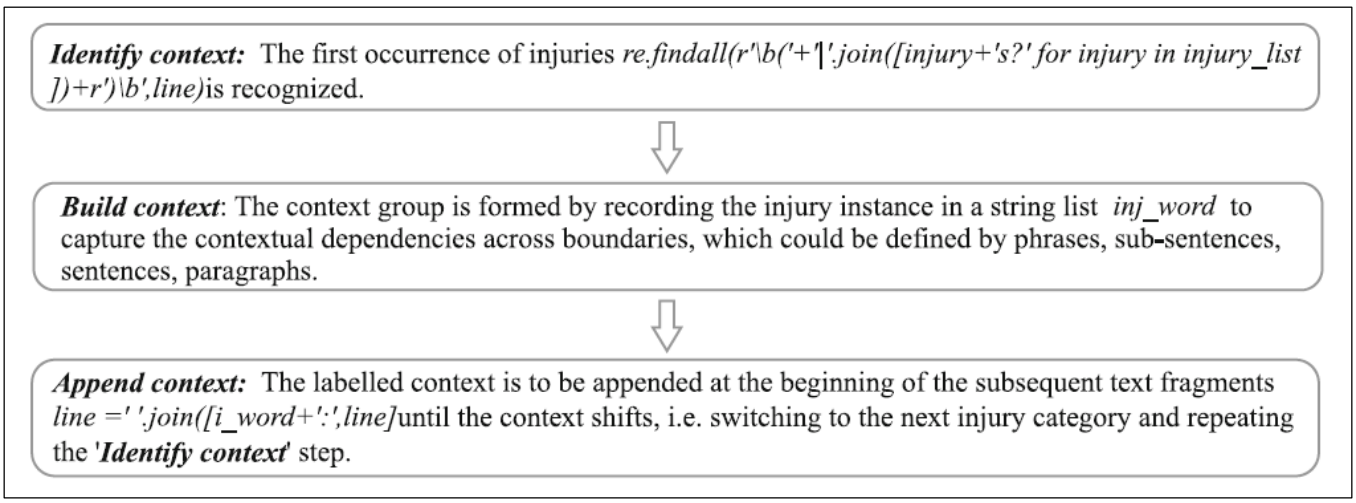

Figure 9 Addressing long-term contextual dependency

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 Figure 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 Figure 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 "postaccident"/ "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",

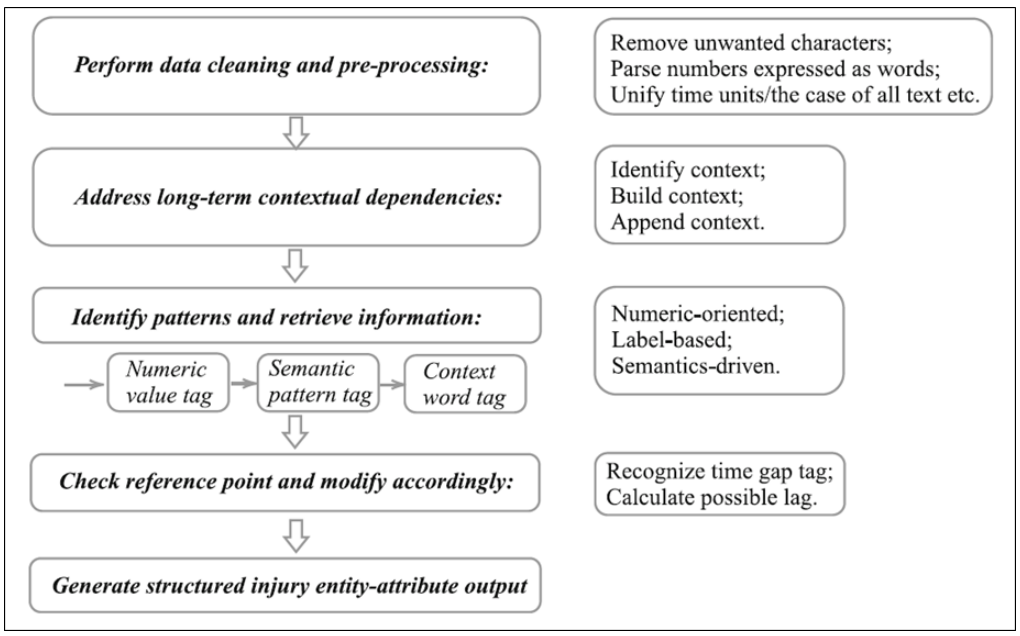

Figure 10 Flow of the information extraction procedure

"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 after keyword part via datefinder() method, based on which the time gap is calculated and injury duration is adjusted accordingly.

3.5. 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 eachword 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.

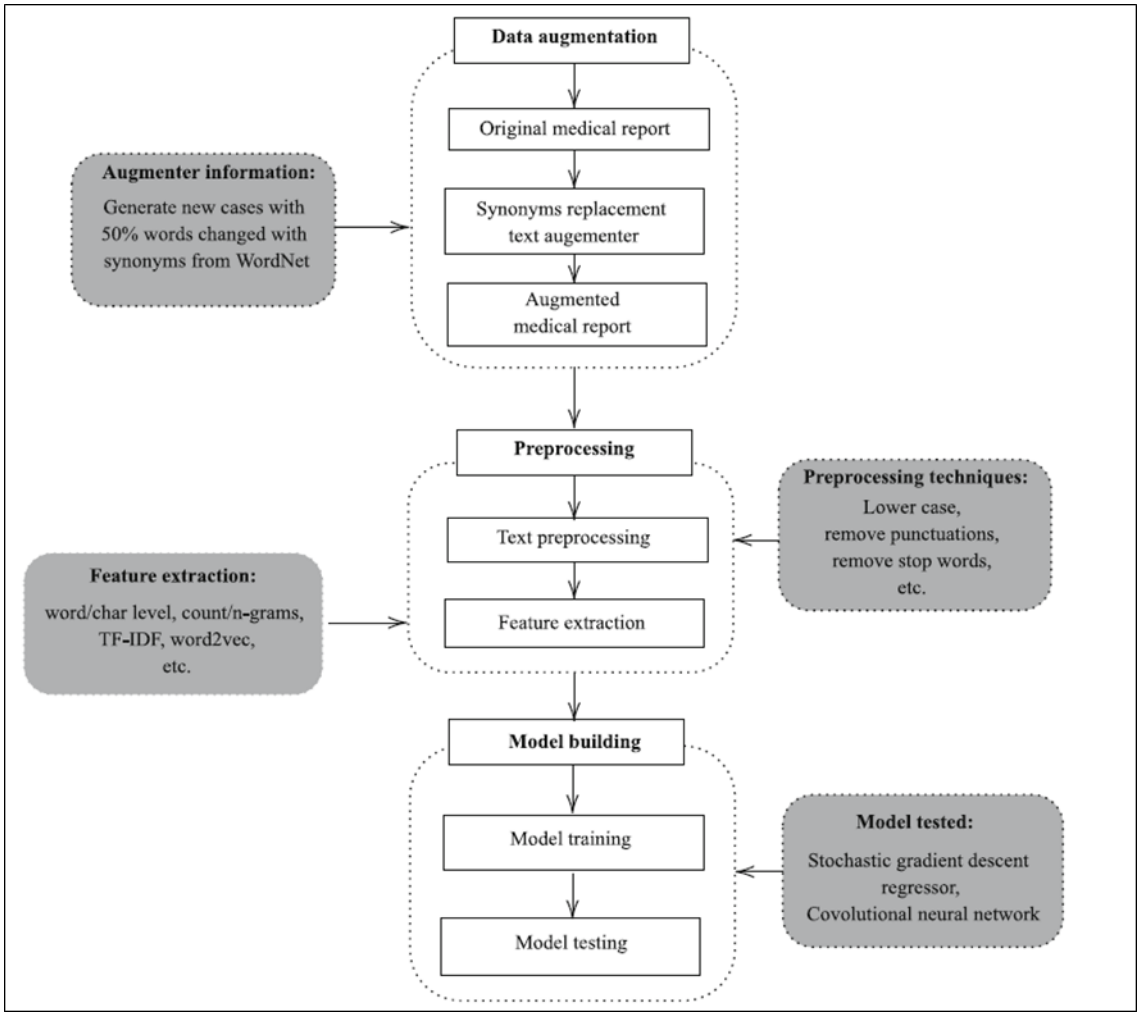

Figure 11 Document analysis framework

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 CNNare 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 Figure 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.

3.6. 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.

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

Table 6 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”

3.8. 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.

Table 7 SGD linear regression prediction results with different features

Text vectorizer	**MAE**		**RMSE**	**R2**
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 8 Prediction results comparison with different models

Model		MAE	RMSE	R2
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

Models with word-level TF-IDF and character n-grams TF-IDF as features perform much better than others, achieving a predicted R2 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 pretrained 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.

3.9. 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, themodel-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 work flow 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 9 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

4. Key results and challenges in AI

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 R2 value of 0.8 and a CNN utilizing different word embeddings performs similarly, with the highest predictive R2 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.

5. Conclusion

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.

References

[1] Abrahams, A. S. (2015). An integrated text analytic framework for product defect discovery. Production andOperations Management, 6(24), 975–990.

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

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

[4] Avgerinos, E.A. (2018). Task variety in professional servicework: When it helps and when it hurts. Production and Operations Management, 27, 1368–1389.

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

[6] Dhashanamoorthi, Balaji. Artificial Intelligence to detection fault on three phase squirrel cage induction motors subjected to stator winding fault (2023).

[7] Dhashanamoorthi, Balaji. Artificial Intelligence to detection fault on three phase squirrel cage induction motors subjected to broken bar fault (2023).

[8] Dhashanamoorthi, Balaji. Fault Detection and Identification in Three Phase Transformer using AI based FSA and PVR analysis (2023).

[9] Dhashanamoorthi, Balaji. Opportunities and Challenges of Artificial Intelligence in Banking and Financial services.

[10] Dhashanamoorthi, Balaji. Artificial Intelligence in combating cyber threats in Banking and Financial services.

[11] Dhashanamoorthi, Balaji. Efficiency Improvement On Wind Turbine Through Bump Up Stepper Motor (2022).

[12] Dhashanamoorthi, Balaji. Construction of suffix tree using key phrases for document Using down-top incremental conceptual hierarchical text clustering approach (2022).

[13] 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 JointConference on Natural Language Processing, (Volume 2: Short Papers) (pp. 180–185).

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

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

[16] 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.

[17] 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.

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

[19] 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

[20] Harris, Z. S. (1954). Distributional structure. Word, 10, 146–162.

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

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

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

[24] 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