International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

Survey on Insurance Claim analysis using Natural Language Processing and Machine Learning

Sapana Kolambe[1], Dr. Parminder Kaur[2]
[1]Research Scholar
MGM University
Chhatrapati Sambhajinagar, India
e-mail: sapanaborole07@gmail.com
[2]Associate Professor
MGM University
Chhatrapati Sambhajinagar, India
e-mail: dhingra.param@gmail.com

Abstract— In the insurance industry nowadays, data is carrying the major asset and playing a key role. There is a wealth of information available to insurance transporters nowadays. We can identify three major eras in the insurance industry's more than 700-year history. The industry follows the manual era from the 15th century to 1960, the systems era from 1960 to 2000, and the current digital era, i.e., 2001-20X0. The core insurance sector has been decided by trusting data analytics and implementing new technologies to improve and maintain existing practices and maintain capital together. This has been the highest corporate object in all three periods.AI techniques have been progressively utilized for a variety of insurance activities in recent years. In this study, we give a comprehensive general assessment of the existing research that incorporates multiple artificial intelligence (AI) methods into all essential insurance jobs. Our work provides a more comprehensive review of this research, even if there have already been a number of them published on the topic of using artificial intelligence for certain insurance jobs. We study algorithms for learning, big data, block chain, data mining, and conversational theory, and their applications in insurance policy, claim prediction, risk estimation, and other fields in order to comprehensively integrate existing work in the insurance sector using AI approaches.

Keywords-machine learning; insurance; artificial intelligence; claim; prediction; natural language processing etc.

I. INTRODUCTION

A policy issued by an insurance company provides financial protection or reimbursement to a person or an organization under the terms of an insurance contract [1]. In the insurance industry, subsidiaries are frequently employed. These affiliates may have a mutual parent or a regular parent's own shares in them, notwithstanding the fact that they are established as ordinary shares insurers. For putting life insurance theory to the test, this feature of insurance company ownership is crucial. Insurance companies safely invest the money so that it can increase and be used to pay out when a claim is made. The life and property of an individual are subject to the risk of death, disabling illness, or destruction. Financial losses could result from these risks. Transferring these risks to a third party is a wise strategy that involves insurance. Property and casualty insurance, accident and health insurance, and financial guarantors are the three most common types of insurance firms [2]. The following list of key components of the insurance industry includes claim prediction, risk prediction, and underwriting.

1.1 Claim Prediction

When you file an insurance claim, you ask your insurer's employer to pay for anything which is protected by your policy, such as an automobile accident, an incident at home, or a trip to the ER. A client may ask for an explanation if their insurance claim prediction with AI is rejected, including the reasons why and the factors that may have contributed to the rejection. A first-party claim is one that needs to be filed as soon as possible after the incident occurs and is for benefits covered by your own insurance policy. Property damage, physical injury, auto insurance coverage, crash treatment, and liability are all included in auto insurance claims.

1.2 Insurance Risk

Risk in the context of insurance refers to the likelihood that something unfavorable or unexpected may occur. Examples include fraud, the destruction or damage of useful goods, as well as personal injuries. One method for achieving this is to categories the financial load into four groups: market risk, business risk, capital sufficiency, and financial leverage.

1.3 Insurance Fraud

Insurance fraud is any action taken with the intent to defraud an insurance process. When someone files repeated insurance

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

claims for the same loss or when they claim to have lost more money than they have, it is a fraudulent insurance claim. Alternately, someone could deliberately harm the asset for which insurance is being sought. Another form of insurance fraud is when a policyholder willfully misrepresents material facts to an insurer in order to receive more favorable coverage or a reduced premium. Insurance fraud may target motor vehicles, businesses, homes, and other personal insurance claims.

1.4 Insurance Policy

An insurance policy is a legal agreement you make with underwriting to provide protection against specified risks under specific circumstances. The insurance policy outlines the issues that the company is required by law to settle and is a contract between the policyholder and the insurer. The insurer agrees, for an up-front sum known as the premium, to compensate policyholders for losses resulting from regulated perils. Medical costs, vehicle damage, lost income, and travel-related accidents are all covered by insurance. The two most widely used types of insurance are business and life insurance. There are other general insurance categories that mix different types of policies.

II. AI IN INSURANCE

Another fast-developing technology applied to insurance innovation is artificial intelligence (AI), which is employed in a wide range of back-end processes like the detection of fraudulent behavior, the development of trading algorithms, the analysis of blockchain data, and the creation of search engines for financial products and services [3]. Many disciplines are benefiting from machine learning, including robotics, computer vision, and Natural Language Processing (NLP). Many insurance applications use machine learning techniques including Logistic Regression with Penalty, Neural Network, Extra Trees Classifier, Random Forest, SVM, and GBM.

In artificial intelligence, there are three distinct approaches to learning: supervised, unsupervised, and reinforcement. The majority of insurers' researchers have utilized supervised learning to evaluate risk during the past few decades, using well-known variables in numerous permutations to produce the desired output. A precise set of objectives for unsupervised learning is encouraged for today's insurers. If there are any changes to the variables, the method will attempt to adjust them so that they better reflect the objectives.

The primary benefit of AI for the insurance sector is the easiest-to-manage data sets. Machine learning is a powerful tool for analyzing structured, semi-structured, and unstructured datasets. For the purpose of researchers and data analysts, some insurance firms provide datasets. Machine learning can improve the insurance industry's ability to forecast risk, claims, and customer behavior.

Artificial intelligence has also been used to power conversation interfaces that intelligently present clients with different types of information. These interfaces use current information, machine learning, and natural language processing. Natural language data from prior consumer interactions are supplied to chat bots, where an intelligent system processes it and trains them to respond to users' text messages quickly.

Numerous scientists are investigating cutting-edge machine learning techniques for tasks like premium leakage to expense control, debt collection, litigation, and fraud detection. The need for management solutions in the industry and the potential for machine learning approaches to be used in education are the driving forces behind this research. There is always more information available to answer questions about insurance.

III. LITERATURE REVIEW

In the insurance sector, research on risk and claim prediction has been conducted extensively utilizing AI techniques. In addition, other researchers in this field have also used chatbots, blockchain, and big data. In-depth research on insurance-related topics, including claim forecasts, predicting risks, and some AI techniques utilizing chatbots to communicate big data and data mining are detailed in this area.

Predicting health insurance claims with the help of artificial neural networks which involve both recurrent neural networks and feedforward neural networks is shown in [4]. The annual claims amounts were predicted using both models. The study's raw data were autocorrelated before being trained on the test outcomes and having the data normalized. Additionally, this study's model design included stopping criteria and considerations for model correctness. Additionally, based on the findings of this article, the artificial neural network (ANN) model outperformed human predictions and reduced the mean absolute percentage error (MAPE)by around 11.5 percent.

Proposes to use the special information acquired to create a mathematical model that will determine the dollar amounts of third-party claims for various categories of vehicles [5]. The authors recommend using a combination of linear regression, ANN, exponential smoothing, and mixed ARIMA-ANN models to forecast various types of vehicle claim amounts. In addition, the authors used root-mean-squared error (RMSE) and mean absolute percentage error (MAPE) to assess performance with a small amount of variation. According to the empirical investigation, ANN models fared better than the other models and are thought to be more accurate predictors.

In [6] researchers investigated the relative performance of XGBoost methods and logistic regression models for forecasting the probability of accident claims employing telematics data. This study demonstrates that logistic regression is a superior model due to its enhanced interpretability and predictive power.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

To anticipate the insurance claims, [2] offered four classifiers: XGBoost, J48, ANN, and Naive Bayes. The author used cutting-edge statistical techniques to handle the issue of several qualities having incomplete data. The authors claim that the J48 model followed the XGBoost model, which had the best accuracy of 92.53% and outperformed the other models.

In [7] discussed how to analyses insurance claims effectively and provided in-depth information on a variety of machine learning techniques. The performance of the machine learning models was also assessed and compared by the authors using several assessment measures. However, the study fails to provide a thorough analysis of the data and fails to give recommendations for the optimum machine learning model based on the performance criteria employed.

In [8] it explores and compares the effectiveness of various ensemble techniques, including XGBoost-based neural network models, AdaBoost, stochastic GB, random forest, and AdaBoost. The authors used XGBoost and examined the claim prediction problem's performance accuracy. According to the findings, XGBoost outperformed other ensemble learning techniques with regard to normalized Gini. However, the authors did not provide a thorough description of how the experiment was carried out or a response to the findings.

Predictive modeling for vehicle insurance claims, according to [9], used projected claim severity and frequency. Third-party property damage (TPPD) and own damage (OD) are the two categories of claims that have been taken into account. Data sets spanning the years 2001 to 2003 are used to build the predictive model [10]. The nature of insurance data, which includes its abundance of information, ambiguity, imprecision, and incompleteness, as well as the standard statistical technique [11] which is unable to manage the extreme value in the insurance data, are the key challenges in modelling motor insurance claims. The back propagation neural network (BPNN) model is suggested by their research to model the issue. The BPNN model's mechanism for resolving the problems is described in great depth.

In an insurance company, work can be done more quickly and effectively with data analytics tools, according to [12]. Four machine learning methods are shown in this study: Random Tree, ANN, REPTree, and multiple linear regressions. These methods can be used to forecast applicant risk levels. The authors also employ additional feature selection and feature extraction strategies, including principal component analysis (PCA) and correlation-based feature selection (CFS). The authors found that Multiple Linear Regression with PCA performed better than the other models, with RMSE and MAE values of 2.0659 and 1.6346, respectively. Although the authors made an effort to apply widely-appreciated supervised machine learning algorithms, they failed to thoroughly explore other widely-appreciated techniques, such as decision trees and RNNs, from which a more general conclusion or recommendation may have been drawn.

It is mentioned in [13] how much it will cost to keep working as an insurance agent. To lower total costs for the clients, the authors developed an agent-less insurance model based on an AI-driven methodology that limits the necessity for human agents in insurance companies. As a software application, this research suggested four statistical models. In order to identify a group of clients that would be most likely to purchase an item from an individual agent, the authors develop a machine-unsupervised model that can function as an agent.

The automation of the insurance claim process for the automotive industry from end to end is suggested in [14] since it is advantageous for both the customer and the company. This system receives input in the form of pictures of the damaged vehicle and outputs pertinent data, including the damaged parts and an estimation of the degree of damage (no loss, moderate or severe) to every part. This serves as a prompt to determine the repair cost estimate that will be utilized to determine the insurance claim amount. Mask R-CNN, PA Net, and an ensemble of these two, coupled with a transfer learning-based VGG16 network, are the proposed methods employed in this system to carry out various tasks of localizing and recognizing different classes of parts and defects discovered in the car.

In [15] the author has compared various machine learning algorithms to predict the claim for car insurance. The main problem handled in this work is related to an imbalanced dataset. The results obtained in the paper show the improvement of claim prediction with respect to various parameters considered for accurate calculation. It used big data [16] with 1,488,028 observations spread across 59 variables. This study shows the value of XGBoost as a model. Out of the eight models for classification used in this research, random forest and the decision tree (C50) performed the best at predicting the incidence of claims, while the Naïve Bayes model performs the poorest.

By utilizing an NLP Concept Matching technique, in [17] authors were focusing on enriching texts with semantic tags in a way that considers types, acronyms, abbreviations, and the many phasing that are used to represent concepts. The Concept Matcher effectively offers the capability to normalize the unstructured textual material into standardized tags that may be extracted and fed into various text or data mining methods. It is possible to build a relatively comprehensive knowledge model by utilizing the existing taxonomy for illnesses and treatments because the system solely deals with categorizing medical insurance claims and diseases, treatments, and related issues.

Instructions on how to construct and evaluate one's own agent applications are included in [18], along with a history of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

conversational interfaces and their evolution. In [19] a case study regarding the operation of IBM's chatbot Watson is presented after a summary of common methodologies and design decisions. Watson gained notoriety for defeating humans in the famous quiz game Jeopardy!

Today, a lot of chatbot apps have already been developed with the intention of resolving real- world issues. Because consumers tended to skim over the frequently lengthy and complex privacy notifications, one instance is PriBot, a chat assistant that can be asked questions regarding an application's privacy policy. Additionally, the chatbot takes user requests to modify their privacy options or app permissions [20].

Eight individuals were required to converse with two different chat bots in another experiment conducted by [21], one of which exhibited more robotic behavior than the other. They have obligations in this situation, such as requesting an insurance certification or ordering an insurance policy. Each of the participants naturally began to converse with one another using everyday language. When the bot didn't satisfactorily answer their questions, the users' phrases got shorter and shorter until they were only writing crucial terms. According to the survey's findings, conversational bots should ideally be designed to resemble people because consumers seem to feel more at ease speaking to people, especially when the issues at hand are important ones like their insurance plans [21].

In [22] uses TEATIME, an agent-based DM architecture, to design an artificially intelligent conversational insurance agent. TEATIME bases actions on motional states, thus when the bot is thought to be unhelpful, that feeling prompts the bot to apologies. The demonstrated example bot serves as a demonstration of idea for TEATIME and can respond to client emotions and answer inquiries about insurance, but it does not yet implement a whole business process.

Explain a text-based healthcare chatbot in [23] that serves as a companion for weight loss but also links a patient with medical staff. For gathering patient feedback, the chat interface enables non-text inputs such as scales and illustrations. The study's findings indicated that the longer a chatbot is utilized, the more automated conversations it will have, and the more highly engaged users are with it as a peer.

Three classifiers were developed to predict fraudulent claims and a percentage of the premium amount [24] based on each customer's individual characteristics. Data mining techniques are used to calculate insurance premium levels for different clients and to anticipate bogus accusations. For classification, the methods Random Forest, J48, and Naive Bayes are chosen. The results show that Random Forest outperforms the other approaches based on the synthetic dataset. Because of this, the study does not address predicting insurance claims; rather, it concentrates on fraud.

In 2017, the author [25] put out the idea of predicting if a consumer has made an insurance claim. They employ several techniques, including gradient boosting, logistic regression, Naive Bayes, random forests, least squares ridge regression, and least squares lasso regression. The data contains some NaN values as well. Mean imputation is practical since it results in a complete data set, and it may be used to replace these missing values. Convenience, even when the data are MCAR, is not a strong advantage because the resulting parameter estimations are greatly distorted by this method [26]. Moreover, if more than 10% of a subject's data is missing; mean imputation approaches produce significantly skewed estimations across all missing data circumstances [27].

In [28] built a framework and step-by-step manual for forecasting insurance claims using knowledge discovery techniques. The findings demonstrate that dimensionality reduction does not always require for this problem and that simple strategies, such as decision trees and random forests, outperform more statistically sophisticated techniques, like support vector machines, even when using tiny data sets. Unfortunately, many aspersed approaches require a significant amount of labelled data. As a result, a considerably larger dataset is required because their dataset is very sparse, resulting in very volatile outcomes.

When compared to other regression models, the Recurrent Neural Network (RNN) model excels at analyzing health insurance claims [29]. Using information gathered from medical insurance claim files, another study [30] developed LASSO, a machine learning predictive regression model, to provide the management of population health in Japan.

In order to improve their customer service, claims handling, and financing, property and casualty insurers are collecting data from a variety of sources, including agent conversations, CRM systems, telematics, home automation systems, and even online networks. More accurate coverage, pricing, and a better knowledge of the prediction all result from more precise insurance data. Datasets from the insurance industry will aid researchers and data analysts in conducting several analyses that will enhance the sector's performance as shown in table 1.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

Table 1. A succinct summary of the Insurance industry datasets available for use by researchers and data analysts

Sr. No	Dataset Name	Description
1	Insurance Reviews France	Reviews of French insurance is available in this database. User opinions on French communal health insurance. Researchers, data analysts, and sentiment analysis can all use this dataset, which is also available.
2	US Health Insurance Dataset	Insurance premiums are generated in this dataset, which includes 1338 rows of claimed information, depending upon the purchaser's year, gender, body mass index (BMI) number of children, medical conditions, and location. The dataset contains no inaccurate or missing values.
3	Health insurance data	Accidents are also covered in this dataset. Despite how challenging it is to categorize diseases, this dataset is broken down into three groups: acute, subacute, and chronic. The group's primary insureds fall into one of five groups, each having a unique set of policy modifications. Classified based on the kind of insured person and the kind of illness.
4	Health Insurance Coverage	The data in this set can be used by analysts and other academics interested in health insurance coverage. The dataset contains information about market-based healthcare plans, enrollment in Medicaid and Medicare, the percentage of uninsured individuals before and after the enactment of Obamacare, and projections of the number of people insured by an employer.

Table 2. Comparison of existing work

Sr. No	Methodology used	Dataset	Accuracy
1	J48,Neural Network, XGBoost, Naive base[44]	Kaggle: 12 variables, 30240 cases	J48-92.22% NN-91.93% XGBoost-92.53% Naïve base-83.84%
2	Naïve Bayes, Multi- Layer Perceptron, J48, Random Tree, Logistic Model Tree (LMT), Random Forest[45]	insurance2	Precision: NB-81.6% MLP-99.7% J48-99.8% RT-99.9% LMT-100% RF-100%
3	Logistic Regression, Decision Tree, Random Forest, XGBoost[15]	CAS dataset	LR- 86.52% DT-81.93% RF-85.13% XG-86.39%
4	random forests, XGBoost[49]	Turkey company Dataset	Random Forest - 0.9843 XGBoost - 0.9749
5	syntax-based approach and a semantic-based approach[62]	gold standard Dataset	recall - 90.78%, precision - 83.05%, F-Measure - 86.74%
6	Semantic Web, Modal/ Deontic Logic, and NLP [47]	Chubbs policy document	Accurancy-55%
7	Attribute extraction-CNN, DBSCAN [48]	synthetic natural language policy dataset	Precision-0.93, Recall-0.95, F1-0.94
8	NLU, LSTM, RNN, Content Filtering [46]	Cornell Movie Dialogue Dataset, Insurance Domain Dataset	Precision-0.9857, Recall-0.9857, F1-0.9857
9	BERT, k-means clustering, weighted complex network community division[63]	Cluster Dataset	Random-0.2, AlchemyAPI-0.5, Hierarchical Clustering-0.6
10	LSTM-CRF Attention model, lexicon [50]	dataset with jieba	Precision-0.9453, Recall-0.9284, F1-0.9368

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

11	Text Rank, part-of-speech[56]	social online banking domain	30.73% more accuracy than baseline
12	Word Graph Modeling Method [61]	DUC2001 and Inspec	F1-value of 0.329 as compared to 0.317 achieved by ExpandRank.

IV. KEYWORD EXTRACTION

To accurately convey the essence of a document, keywords are essential. To find keywords in both unstructured and structured texts, many Natural Language Processing methods have been devised. In contrast to structured text, it is more difficult to extract keywords from grammatically ambiguous language since it is difficult to rely on linguistic aspects in unstructured texts. A research work has been carried out to understand the different keyword methodologies used to fetch important key phrases or words from the document based on various applications. A survey based on keyword extraction methodologies are presented here.

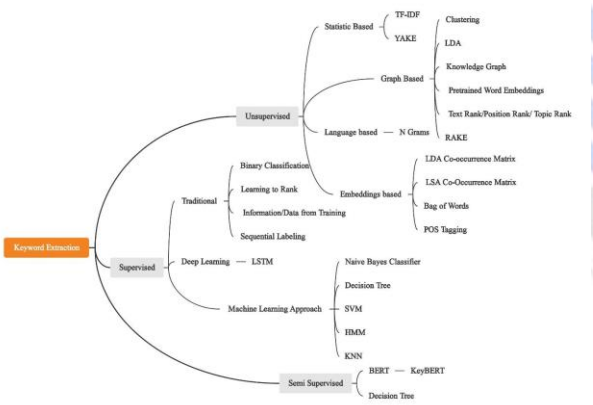

Figure 1. Keyword Extractions Techniques

In an effort to glean useful information from the unstructured data, researchers employed the hidden Markov model in addition to other natural language processing methods including POS tagging and clustering. The first text pre-processing stage in this method was completed using the methods listed below. TreeTagger [31], a probability-based decision tree, is used to perform the first stage of text labeling, known as Part of Speech tagging. An enhanced version of the ID-3 technique is used to implement a binary decision tree recursively. Using a trigram dataset, the probability was calculated by iterating over the tree. Using Penn Treebank data, this tagging method has outperformed the trigram tagger in terms of accuracy. After that, stop words were eliminated, stemming was performed on the output that had been tagged, and WordNet was used to combine comparable content. The keywords were then removed. Merging the terms with comparable content decreased the number of phrases that needed to be extracted [32].

The focus of linguistic word features is on the extraction and recognition of keywords from text documents. It includes discourse analysis, lexical analysis, and syntactical analysis. The primary resources used for lexical analysis include electronic dictionaries, tree taggers, Wordnet, n-grams, and POS patterns [33]. The syntactical analysis makes use of noun phrases and noun chunks. This strategy requires more calculation and is more accurate. The drawback of this strategy is that it necessitates subject expertise [34][35].

The C4.5 algorithm, which is an extension of the ID3 algorithm, is just one of the many methods published and used for extracting keywords from documents. It contains statistical decision tree classification, which is better suited for classes with balanced properties [36]. A keyword-generating technique named KEA is based on the TF-IDF statistics method and the Naive Bayer learning strategy. This technique is based on the term frequency and keyword location, two lexical properties. The likelihood of a word becoming a keyword increases if it appears frequently in a document. To effectively train this model, a sizable data set is required [37][38]. The KEA++ algorithm, which is an extension of the KEA keyword-generating method, chooses keywords based on three linguistic criteria. It links the synonyms of the terms and filters out those with a high node degree using a thesaurus. The vocabulary of KEA++ is governed by structures. The benefit of KEA++ is that it uses a controlled vocabulary, preventing the occurrence of nonsensical and inaccurate word extractions. The performance of KEA++ is based on the control's vocabulary [39].

A technique for extracting keywords from tweets is suggested using brown clustering, which groups words based on continuous word vectors and brown corpus. The clustering algorithm looks for clusters that have the greatest likelihood. A tweet can be between 1 and 280 characters long. As a result, the number of keywords collected must be proportional to the tweet's word count. This feature is considered in the methodology given in this research. In comparison to other existing methods, the algorithm MAUI, which combines Browns clustering and Word Vectors, holds a higher value for precision and recall [37]. Backpropagation neural networks have been suggested as a further technique for keyword extraction [40]. Journal articles made up of the corpora that were utilized to train the algorithm. Every word has a unique collection of properties embedded into it, including term frequency. This approach does not employ inverse document frequency (IDF), as doing so necessitates a system-wide study. Backpropagation Neural Network is employed through the C4.5 algorithm to adjust the system. This approach's evaluation findings show that BNN

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8461
***Article Received:** 23 July 2023 **Revised:** 15 September 2023 **Accepted:** 26 September 2023*

predicts keywords with 90.11% accuracy, 59.50% recall, and 0.717 F-Measure [41]. To extract keywords from tweets, a model based on a Recurrent Neural Network (RNN) with two hidden layers is suggested. The first layer records the keyword's information, while the second layer uses that information to extract the keyphrase. Results analysis demonstrates that this strategy outperforms the conventional state-of-the-art. [42].

Using simple statistics without training data is a rough and generally effective method. It focuses on statistics derived from non-linguistic aspects of the content, which can then be utilized to generate keywords. The text's keywords can be filtered using statistical N-gram data. The fundamental criterion is the frequency of occurrence, and it is known as TF-IDF -Term Frequency - Inverse Document Frequency [32][34]. These statistical data can subsequently be used to determine the word's support and confidence. The Apriori technique is then utilized to infer the keywords, which is often done for association rule learning and frequent itemset mining [43][37][35].

V. CONCLUSION

This study examines and presents contemporary artificial intelligence methods and cutting-edge models in the insurance industry. This study has consulted various insurance-related works, including literature on claim prediction, risk prediction, and underwriting. Based on the literature review, it is seen during the study how AI approaches have been applied in the insurance business.

XGBoost was used more frequently and performed better than any other AI technique during the survey. Numerous neural networks were also employed for a variety of insurance-related applications. Additionally, decision trees and their variants were widely used and displayed promising results in comparison to other AI techniques. However, numerous authors discussed various boosting techniques, and it appears that no clear consensus was reached on how to improve the ML algorithms. Additionally, because most researchers in this field are interested in machine learning techniques and are discouraged by the abundance of data in the field, deep learning techniques have been overlooked.

Because most researchers focus on a small number of outcome measurements for evaluating models without considering alternative performance assessment metrics, a lack of various metrics for performance assessment was also observed during the study. Almost none of the sources included in this analysis used the Explainable AI method, which visually explains why the model's performance and results may be trusted. This technique increases the model's explain ability and fairness.

While researching this problem, we also found that researchers in the insurance industry don't have access to a literature review article that would give them an overview of previous work in the field of artificial intelligence. By offering a thorough literature review of the quickly expanding literature on artificial intelligence in the insurance sector, this paper closes this research gap.

REFERENCES

[1] Kshirsagar R. and Hsu L, et.al "Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, issue. 17, pp. 15127-15136, 2021. https://doi.org/10.1609/aaai.v35i17.17776

[2] Abdelhadi S., ElBahnasy K.A., and Abdelsalam M.M, "A Proposed Model to Predict Auto Insurance Claims using Machine Learning Techniques", 2020.

[3] Mohamed H, Ruixing M, "Using Machine Learning Models to Compare Various Resampling Methods in Predicting Insurance Fraud", Journal of Theoretical and Applied Information Technology, Vol.99. No 12, pp. 2819-2833, 2021.

[4] Goundar S., Prakash S., Sadal P., & Bhardwaj A, "Health Insurance Claim Prediction Using Artificial Neural Networks", International Journal of System Dynamics Applications (IJSDA), vol. 9, no. 3, pp.40-57, 2020. http://doi.org/10.4018/IJSDA.2020070103.

[5] Selvakumar, Dipak Kumar Satpathi, P. T. V. Praveen Kumar, V. V. Haragopal, "PredictiveModelingof Insurance Claims using Machine Learning Approach for Different Ty pes of Motor Vehicles", Universal Journal of Accounting and Finance, vol. 9, no. 1, pp.1 - 14, 2021, DOI: 10.13189/ujaf.2021.090101.

[6] Pesantez-Narvaez J, Guillen M, Alcañiz M, "Predicting Motor Insurance Claims Using Telematics Data - XGBoost versus Logistic Regression", Risks. vol. 7, no.2, 2019 https://doi.org/10.3390/risks7020070.

[7] Burri RD, Burri R, Bojja RR, Buruga SR, "Insurance claim analysis using machine learning algorithms", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 6Special Issue 4, pp. 577–582, 2019, doi: 10.35940/ijitee. F1118.0486S419.

[8] Fauzan M. A., & Murfi H, "The accuracy of XGBoost for insurance claim prediction", International Journal of Advances in Soft Computing and its Applications, vol. 10, no. 2, pp. 159-171, 2018.

[9] Yunos Z. M., Ali A., Shamsyuddin S. M., Ismail N., & Sallehuddin R. S, "Predictive Modelling forMotor Insurance Claims Using Artificial Neural Networks", International Journal of Advances in Soft Computingand its Applications, vol. 8, no. 3, pp.160-172, 2016.

[10] Mnasser A., Bouani F., & Ksouri M, "Neural Networks Predictive Controller Using an Adaptive ControlRate", International Journal of System Dynamics Applications, vol. 3 no. 3, pp.127–147, 2014, doi:10.4018/ijsda.2014070106.

[11] Azar A., & Balas V, "Statistical Methods and Artificial Neural Networks Techniques in Electromyography", International Journal of System Dynamics Applications, vol. 1, no.1, pp.39–47, 2012, doi:10.4018/ijsda.2012010103.

[12] Boodhun N, Jayabalan M, "Risk prediction in life insurance industry using supervised learningalgorithms", Complex & Intelligent Systems, vol. 4, no. 2, pp.145–154, 2018 doi: 10.1007/s40747-018-0072-1/2018.

[13] K. P. Sinha, M. Sookhak and S. Wu, "Agentless Insurance Model Based on Modern Artificial Intelligence", IEEE 22ndInternational Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 2021, pp. 49-56, doi: 10.1109/IRI51335.2021.00013.

[14] R. Singh, M. P. Ayyar, T. V. Sri Pavan, S. Gosain and R. R. Shah, "Automating Car Insurance Claims Using Deep Learning Techniques," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 2019, pp. 199-207, doi: 10.1109/BigMM.2019.00-25.

[15] Sebastian Baran, Przemyslaw Rola, "Prediction of motor insurance claims occurrence as an imbalanced machine learning problem" arXiv:2204.06109v1.

[16] Hanafy, M; Ming, R., "Machine Learning Approaches for Auto Insurance Big Data" Risks 2021, 9(2) 42.

[17] Fred Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing", SIGKDD Explorations, Volume 7, Issue 1, pp. 59-66.

[18] Michael McTear, Zoraida Callejas, David Griol, "The Conversational Interface: Talking to Smart Devices", Springer, volume 6, 2016.

[19] Cahn J. "Chatbot: Architecture, design, & development', 2017.

[20] Harkous H, Fawaz K, Shin K. G, and Aberer K, "Pribots: Conversational privacy with chatbots", In Twelfth Symposium on Usable Privacy and Security, Denver, CO. USENIX Association, 2016.

[21] S'orensen I, "Expectations on chatbots among novice users during the onboarding process", 2017, http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva- 202710.

[22] Yacoubi A. and Sabouret N, "Teatime: A formal model of action tendencies in conversational agents",2018, In ICAART (2), pages 143–153.

[23] Kowatsch T., Niben M., Shih C.-H. I., R'uegger D., Volland D., et al., "Text-based healthcare chatbots supporting patient and health professional teams: Preliminary results of a randomized controlled trial on childhood obesity", In Persuasive Embodied Agents for Behavior Change (PEACH2017), 2017, ETH Zurich.

[24] Kowshalya G., & Nandhini M, "Predicting fraudulent claims in automobile insurance", In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1338-1343.

[25] Millican M., Zhang L., & Kimball D. CS 229 "Final Report: Predicting Insurance Claims in Brazil", 2017.

[26] Enders C. K. (2010),"Applied missing data analysis", Guilford press.

[27] Eekhout Iris, et al. "Missing data in a multi-item instrument were best handled by multiple imputation at the item score level." Journal of clinical epidemiology 67.3 (2014): 335- 342.

[28] Pijl T., van de Velden M., & Groenen P. "A Framework to Forecast Insurance", 2017.

[29] Yang C., Delcher C., Shenkman E and Sanjay Ranka, "Machine learning approaches for predicting high-cost high need patient expenditures in health care", BioMed Eng Online 17, No.131, 2018.

[30] Takeshima T, Keino S, Aoki R, Matsui T, and Iwasaki K., "Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data, Value in Health", Vol. 21, No. 2, 2018.

[31] http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/ DecisionTreeTagger.html. [Accessed: 07- Oct-2019].

[32] Amit Kumar Mondal and Dipak Kumar Maji, "Improved Algorithms for Keyword Extraction and Headline Generation from Unstructured Text", p. 14.

[33] M. Abulaish and T. Anwar "A Supervised Learning Approach for Automatic Keyphrase Extraction", International Journal of Innovative Computing, Information and Control, vol. 8, 2012.

[34] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", International Journal of Computer Applications, vol. 109, no. 2, pp. 18-23, 2015.

[35] Z. Zhu, M. Li, L. Chen, Z. Yang and S. Chen, "Combination of Unsupervised Keyphrase Extraction Algorithms", 2013 International Conference on Asian Language Processing, Urumqi, 2013, pp. 33-36.,2019.

[36] Turney, Peter "Learning Algorithms for Keyphrase Extraction", Inf.Retr., vol. 2, pp. 303-336, 2000.

[37] Marujo Luís & Ling, et.al (2015). Automatic Keyword Extraction on Twitter. 2. 637-643. 10.3115/v1/P15-2105.

[38] Witten, G. Paynter, E. Frank, C. Gutwin and C. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction", Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA, pp. 254-255, 1999.

[39] O. Medelyan, & I. Witten (2006). Thesaurus based automatic Keyphrase indexing. 296 - 297. 10.1145/1141753.1141819.

[40] Taemin Jo, Jee-Hyong Lee "Latent Keyphrase Extraction Using Deep Belief Networks", International Journal of Fuzzy Logic and Intelligent Systems, vol. 15, pp. 153-158, 2015.

[41] J. P. Tensuan, A. Azcarraga "Neural Network Based Keyword Extraction using Word Frequency, Position, Usage and Format Features", Research Congress 2012 De La Salle University, 2013.

[42] Wang, Yang & Gong, Yeyun & Huang, Xuanjing. (2016). Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter. 836-845. 10.18653/v1/D16-1080.

[43] S. K. Bharti and K. S. Babu "Automatic Keyword Extraction for Text Summarization: A Survey", CoRR, vol. 170403242, 2017.

[44] Shady Abdelhadi, Khaled Elbahnasy, Mohamed Abdelsalam, "A Proposed Model to Predict Auto Insurance Claims using Machine Learning Techniques", Journal of Theoretical and Applied Information Technology, Vol.98. No 22, 2020, pp. 3428-3437.

[45] Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja, Srinivasa Rao Buruga, "Insurance Claim Analysis Using Machine Learning Algorithms", International Journal of Innovative

Technology and Exploring Engineering, Volume-8, Issue-6S4, 2019, pp. 577-582.

[46] Phd Mohammad & Hussain, Omar. (2020). IntelliBot: A Dialogue-based chatbot for the insurance industry. Knowledge-Based Systems. 196. 105810. 10.1016/j.knosys.2020.105810.

[47] K. Joshi, K. Pande Joshi and S. Mittal, "A Semantic Approach for Automating Knowledge in Policies of Cyber Insurance Services," 2019 IEEE International Conference on Web Services (ICWS), 2019, pp. 33-40, doi: 10.1109/ICWS.2019.00018.

[48] K. Sane, K. Joshi and S. Mittal, "Semantically Rich Framework to Automate Cyber Insurance Services" in IEEE Transactions on Services Computing, vol., no. 01, pp. 1-1, 5555.

[49] Alohaly M., Takabi H. & Blanco E. "Automated extraction of attributes from natural language attribute-based access control (ABAC) Policies". Cybersecurity 2, 2 (2019). https://doi.org/10.1186/s42400-018-0019-2

[50] D. Çavusoğlu, O. Dayibasi and R. B. Sağlam, "Key Extraction in Table Form Documents: Insurance Policy as an Example," 2018 3rd International Conference on Computer Science and Engineering (UBMK), 2018, pp. 195-200, doi: 10.1109/UBMK.2018.8566309.

[51] T. Pu, Q. Zhang, J. Yao and Y. Zhang, "Medical Entity Extraction from Health Insurance Documents," 2020 IEEE International Conference on Knowledge Graph (ICKG), 2020, pp. 565-572, doi: 10.1109/ICBK50248.2020.00085.

[52] X. Mao, S. Huang, R. Li and L. Shen, "Automatic Keywords Extraction Based on Co- Occurrence and Semantic Relationships Between Words," in IEEE Access, vol. 8, pp. 117528-117538, 2020, doi: 10.1109/ACCESS.2020.3004628.

[53] Papagiannopoulou E, Tsoumakas G. A review of keyphrase extraction. WIREs Data Mining KnowlDiscov. 2020; 10: e1339. https://doi.org/10.1002/widm.1339

[54] Firoozeh N., Nazarenko A., Alizon F., & Daille B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), 259-291. doi:10.1017/S1351324919000457

[55] Ö. Ünlü and A. Çetin, "A Survey on Keyword and Key Phrase Extraction with Deep Learning," 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019, pp. 1-6, doi: 10.1109/ISMSIT.2019.8932811.

[56] W. Ding, P. Yu, H. Li, H. Li and X. Lu, "A new method for extracting table borders of insurance policies," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 1933-1937, doi: 10.1109/ITNEC48623.2020.9084823.

[57] R. Kedtiwerasak, E. Adsawinnawanawa, P. Jirakunkanok and R. Kongkachandra, "Thai Keyword Extraction using TextRank Algorithm," 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2019, pp. 1-6, doi: 10.1109/iSAI-NLP48611.2019.9045523.

[58] T. Weerasooriya, N. Perera and S. R. Liyanage, "A method to extract essential keywords from a tweet using NLP tools," 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 2016, pp. 29-34, doi: 10.1109/ICTER.2016.7829895.

[59] S. Lee, T. Park and M. Lee, "4W1H Keyword Extraction based Summarization Model," 2021 International Conference on Electronics, Information, and Communication (ICEIC), 2021, pp. 1-4, doi: 10.1109/ICEIC51217.2021.9369820.

[60] M. Noura, A. Gyrard, S. Heil and M. Gaedke, "Automatic Knowledge Extraction to Build Semantic Web of Things Applications," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8447-8454, Oct. 2019, doi: 10.1109/JIOT.2019.2918327.

[61] Jyothsna, K. Srinivas, B. Bhargavi, et.al, "Health Insurance Premium Prediction using XGboost Regressor," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 1645- 1652, doi: 10.1109/ICAAIC53929.2022.9793258.

[62] X. Mao, S. Huang R. Li and L. Shen, "Automatic Keywords Extraction Based on Co-Occurrence and Semantic Relationships Between Words," in IEEE Access, vol. 8, pp. 117528-117538, 2020, doi: 10.1109/ACCESS.2020.3004628.

[63] Dragoni Mauro, Villata Serena, Rizzi Williams , et.al (2016). Combining NLP Approaches for Rule Extraction from Legal Documents.

[64] Jingxia Ma, Research on Keyword Extraction Algorithm in English Text Based on Cluster Analysis, 2022, Article ID 4293102 | https://doi.org/10.1155/2022/4293102