香港中文大學
The Chinese University of Hong Kong

Project Report for FTEC5510

"InsurX"-- A Novel Application for InsurTech

Chen Ming, 1155160972
Huang Yijing, 1155164917
Guo Zhongyuan, 1155162320
Ma Ziqing, 1155162325
Zhang Xincong, 1155166336
Liu Luyao, 1155169203

Contents

List of Figures

1 Abstract

Purpose – This report aims to explore the insurance business and technological in- novations for the insurance industry. Our group analyzes the strengths and weaknesses of existing insurance platforms and proposes a new InsurTech platform "InsurX" based on Optical Character Recognition (OCR), Natural Language Processing (NLP), and Blockchain technologies. The project will also help inspire further innovations in insurance and improve the insurance ecosystem in Hong Kong, mainland China or other countries and regions.
Design/methodology – The project first investigates existing InsurTech platforms in China. Also, the project delves into some of the key technologies that could transform the insurance industry. Finally, a novel InsurTech platform "InsurX" is proposed and developed.
Findings – Information asymmetry has always been a major barrier in the insurance industry. Blockchain technology allows the insurance industry to connect with the medical system more safely and effectively and provides the current insurance industry with transparency that is not available in the current paradigm. In addition, OCR is the key technology to help policyholders understand insurance policy easily. Lastly, NLP is found the essential technology to help users learn about insurance.

2 Introduction

2.1 Background

Insurance is a crucial part of the modern financial industry. It shields individuals and households from the risk of serious diseases or injuries and protects against downside risk in businesses. In the past ten years, insurance, a business that seemed to be tedious, has become one of the crown jewels thanks to the advancement in financial technologies. In particular, big data and machine learning play a pivotal role in the revolution of the insurance industry. With the help of big data and AI, insurance companies now could prevent potential frauds and provide customizable services to their clients. In addition, the incorporation of peripheral technologies such as OCR, NLP has also marked a breakthrough in the insurance industry.

InsurTech grows rapidly as a result of the technological advancements mentioned above. Many InsurTech startups have become big disrupters in

the in- industries by ingeniously combining AI, big data, blockchains, etc. to offer better services for insurance companies or insurance buyers. For instance, Galileo Platform, an HK-based InsurTech startup, re-defined insurance services with the incorporation of Blockchain technologies. Tractable, a UK InsurTech company, built an AI-based insurance assessment system that allows users to start the claims process by simply uploading images of damage to their properties. These companies have genuinely become game-changers within the whole insurance ecosystem, challenging the big giants with their innovations.

However, It's not just technologies that underpin their success. Meeting the growing demands of technologies is ultimately the recipe for their success. Decades ago, individuals or households that want to buy insurance are faced with long and tedious terminologies and standardized legal contracts. They can only access insurance protection by offline face-to-face meetings with brokers.

In the era of digitalization, people now prefer browsing or buying insurance with simple tabs and clicks on their smartphones or computers. Change of habits has dramatically changed the landscape of insurance businesses. In this report, we investigate the existing InsurTech ecosystem in China and identify key technologies to transform the traditional insurance industry. A novel InsurTech platform – "InsurX" is proposed and developed to demonstrate the power of technologies on insurance.

2.2 Insurance Market Research

There are more than 160 insurance companies in Hong Kong, which is one of the most mature insurance markets in the world. In the current environment where the three elements of "algorithm, computing power, and data" have been greatly improved, cutting-edge technologies such as big data, cloud computing, artificial intelligence, blockchain, and the Internet of Things continue to mature and begin to empower the technological transformation of the insurance industry.

Since the outbreak of COVID-19, the chain effect has accelerated the digital transformation of Hong Kong's insurance industry. Many insurance companies have rapidly implemented digitalization plans. In addition, assisting the digital transformation of traditional insurance companies has also greatly improved the entire insurance process from product design, distribution to claims settlement. For example, CoverGo shortens the design time of

insurance products through APIs, greatly improves efficiency and balances claims risks. It can be seen that the demands of the market environment and the maturity of technical means have provided the insurance industry with a new chance.

With market research, we identify four typical Chinese InsurTech companies. They have used technologies such as Machine Learning, Computer Vision, Big Data, etc. to transform the Insurance ecosystem in China. Technologies used and functions achieved are summarized in Figure 1.

Platform Name	Technology	Functions
Zhong An	Machine Learning	Accurate Pricing
		Streaming real-time data analysis
		Quickly build new problem solutions and models
	Computer vision	Authentication
		Intelligent underwriting
		Automatic text information extraction
	Cloud computing	Interactive platform and data management center
	Big Data	Risk Management
	IOT	Innovative Insurance Product Design
OneDegree	Computer vision	Accurately and quickly identify medical documents and vouchers provided by customers to improve underwriting efficiency
	BlockChain	Match the claim information and improve the credibility of underwriting claims
	Product Management System and Customizable APIs	Product development management and reduce repetitive work
Xianghubao	Computer vision	Intelligent settlement process
	Knowledge Graph	
	Remote face-to-face interviews	
	Block chain	Form non-tamperable records and ensure that the results of publicity are subject to supervision
Nuanwa	Knowledge Graph	Intelligent risk control model
	Big Data	Customer portrait, achieve all-channel customer acquisition, to ensure long-term sustainable development of business

Figure 1: Summary for the four existing Chinese InsurTech companies

2.3 Blockchain

The concept of blockchain first existed in the paper published by Satoshi Nakamoto [1] in 2008 as the concept of "proof-of-work chain". It is to add timestamps to data items and hash them in the block and spread the hash value widely. Babbit.com defines blockchain as follows, each block of the blockchain contains the hash value of the previous block, linking from genesis to the current block to form the blockchain. Each block is guaranteed to be created after the previous block in chronological order. Although various websites and literature have different definitions of blockchain, the definitions of blockchains are all similar in essence. Blockchain has the characteristics

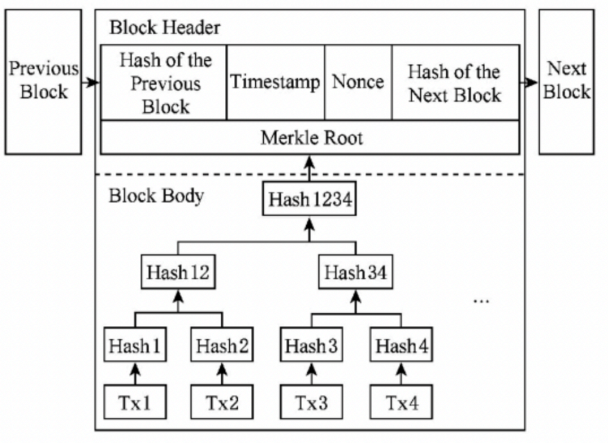


Figure 2: The structure of Blockchain

of decentralization, dis- trust, openness, immutable information, anonymity, and autonomy. The block structure of blockchain is shown in Figure 2. Each block is divided into block header and block body, involving technical elements such as chain structure, Hash algorithm, Merkle tree, and time stamp.

The basic technologies related to blockchain mainly include hash algorithm, timestamp service, Proof of Stakes, Proof of Work, P2P network technology, and asymmetric cryptographic technique at present.

2.3.1 Hash Algorithm

Hash algorithm maps input values of arbitrary length to shorter, fixed-length binary values. For example, the SHA256 algorithm [2] maps an input value of arbitrary length to a fixed-length output value of 256 bits. This binary value is called a hash. Data hashes can verify the integrity of data and the algorithm is commonly used for fast lookups and cryptographic algorithms.

Hash algorithms are widely used in blockchain, which typically does not store raw data, but rather hashes of the raw data. The node information in Merkle tree is obtained by this. Merkle tree is a hash-based tree invented by Ralph Merkle.

Merkle trees have the following characteristics: a) The data structure is generally binary tree or multi-way tree; B) Leaf nodes are hashes of data blocks; C) Non-leaf nodes are the hashes of all their child nodes.

2.3.2 Timestamp Service

As a decentralized system, blockchain differs from centralized credit systems such as banks. While the latter relies on the coercive power of a state apparatus to prevent counterfeiting, the blockchain system relies entirely on technology to solve the problem of "double payments". The system stamps each transaction with the correct time to prove that at that moment the transaction has taken place and the ownership of the money in the transaction has been transferred. The system reports an error when the previous owner of the money used it.

2.3.3 Proof of Stake and Proof of Work

Unlike Proof of Work, Proof of Stakes requires only a handful of calculations to keep a blockchain running. The former refers to the need to find a reasonable block hash, which requires a lot of computation, depending on the difficulty of the current target and the speed of the machine. When a node finds a hash value, it turns out that a lot of computation is done.

2.3.4 P2P Network Technology

P2P network technology, also known as peer-to-peer technology, is an interlinked network system without a central server, relying on user groups to exchange information. P2P network has the advantages of attack resistance and high fault tolerance because it has no centralized server. In addition, all nodes have equal status, and services are distributed on all nodes. Therefore, attacks on some nodes or networks have little impact on the whole system. The bitcoin system applies P2P technology to make each node participate in the system independently. Each node is an independent individual, and the system will not be affected by the breakdown or attack of a single node.

2.4 NLP

Natural language processing (NLP) is an artificial intelligence process to analyze human language, which is to receive natural language, exchange and translate natural language through probability-based algorithms, analyze natural language and output results. Natural language processing is in the field of computer science and computational linguistics, which is used

to study the interaction between human language and computers. Semantics refers to the relationship and mean- ing between words. Its focus is to help computers understand meaning by using the semantic structure of information such as the context of data. It is generally realized through the processes of obtaining corpus, corpus preprocessing, feature engineering, feature selection, model training, evaluation index, model online application, and so on.

At present, the difficulty of natural language processing is to eliminate ambiguity, such as ambiguity in lexical analysis, syntactic analysis, semantic analysis, and so on. The problem needs a lot of knowledge, including linguistic knowledge (such as morphology, syntax, semantics, context, etc.) and world knowledge (independent of language). Firstly, language is full of ambiguity, which is mainly reflected in the three levels of morphology, syntax, and semantics. Secondly, it is difficult to acquire, express, and apply the knowledge needed to eliminate ambiguity. It is difficult to design appropriate language processing methods and models because of the complexity of language processing.

2.4.1 GPT-2 language model

GPT-2 is an unsupervised multi-task learning language model [3]. Based on the transformer model, it uses [webtext] to crawl massive data as the training database of the language model to train a model similar to an encyclopedia. Unlike the NLP model that needs to use a large number of labeled data for training, it can solve specific problems without labeling data, GPT-2 is modeling without understanding words at all so that the model can deal with any coding language. It mainly aims at the zero-shot problem, and greatly improves the quantity, quality, and universality of training data: grab a large number of different types of web pages, and filter to regenerate high-quality training data to train a larger model. Its structure is similar to the GPT model (also known as GPT-1.0). It uses a unidirectional transformer model to make some local modifications: for example, move the normalization layer to the input position of block; Add a layer of normalization after the last self-attention block; Increase vocabulary, etc.

Nowadays, it has almost become NLP standard practice to fine-tune the pre-training model based on transformer structure. GPT-2 can be fine-tuned according to different downstream tasks.

2.4.2 Application of GPT-2

Aspects that GPT-2 can fine-tune are as follows.

1) Text generation: you can use unlabeled or tagged data to carry out adaptive pre-training on GPT-2. For example, if you want to write a composition on GPT-2, you can use a large number of unlabeled composition text for fine-tuning; If you want to use it as a question answering robot, you can supervise and train GPT-2 with labeled questions.

2) Text blank filling: if GPT-2 is to have the ability to do multiple-choice questions, it can use the labeled blank filling dataset to fine-tune, just reward the correct answer text and punish the wrong answer text. In this way, the correct text has a lower loss after training, that is, GPT-2 can choose words and fill in the blank.

3) Text classification: it can solve the problem of emotion classification or classification. GPT-2 supports zero-shot learning and does not need to change the structure of the pre-training model. It can be classified directly on unlabeled data. You can also use traditional fine-tune means to add MLP classification network after GPT-2 model and use labeled data fine-tune.

4) Text extraction: similar to text generation, the labeled news summary dataset is used to supervise and train GPT-2 to make the generated text approximate the summary content.

2.5 OCR

OCR (Optical Character Recognition) technology refers to the process of analyzing scanned image documents and processing the characters in the documents into text that can be edited on the computers. In recent years, OCR now generally refers to text detection and text recognition through deep learning [4] because of the good performance of deep learning on these two tasks. The main indicators to measure the performance of the OCR system are: rejection rate, false recognition rate, recognition speed, user-friendliness, product stability, ease of use and feasibility, etc.

The most used method of text recognition can be divided into text detection and text recognition. These two independent networks are trained to complete the text recognition task, respectively. First, the text detection network predicts the text area of the input image, and then passes the predicted text area image to the text recognition network as its input. The text recognition network then performs character recognition on the text image

transmitted by the text detection network, and finally, the text recognition result of the entire image is obtained.

In the insurance business, the length of the insurance policy is usually long, and customers usually cannot quickly capture the insured, the beneficiary, and the main insurance clauses. In InsurTech, we can upload the guarantee slip, through the OCR technology, convert the PDF file into text that can be edited on the computer. In this way, the customers can know the insurance details Clearer. The InsurTech can also save this information in the database in the future, which can facilitate the management of customer information.

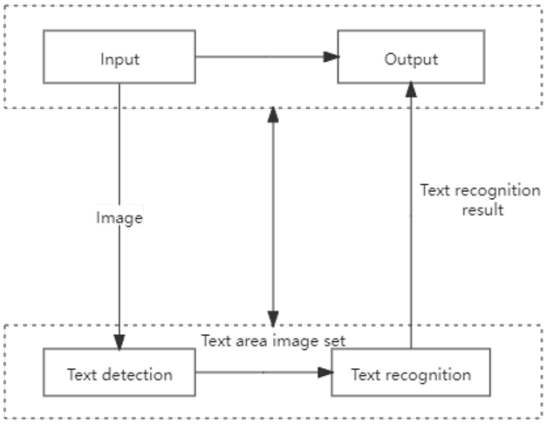


Figure 3: Algorithm of text recognition

2.5.1 Text Detection

Text detection technology is the basis of text recognition. Accurately locating the position of text can help neural network recognize the text content more accurately. Text detection is the process of detecting text areas in pictures that may contain text.

Text detection algorithms can be divided into traditional text detection algorithms and text detection algorithms based on deep learning. Traditional text detection algorithms mainly extract features manually.

2.5.2 Text Recognition

Text recognition is the process of converting the text area to be recognized into characters. Text recognition algorithms can be divided into traditional text recognition algorithms and deep learning-based text recognition algorithms. Traditional text recognition algorithms generally use template matching or SVM. When the text character set is large, the calculation will take a long time, and the algorithm accuracy will decline. The deep learning algorithms' performance is much better than the traditional method whatever in recognition accuracy, computing resource requirements, or time efficiency. Therefore, this article mainly studies text recognition algorithms based on deep learning. The text recognition algorithm mentioned in this section generally deals with horizontal or horizontal text lines. For irregular text such as multi-direction, bending, and deformation, some post-processing algorithms are needed to convert the text area into horizontal text and then recognize it.

2.6 Motivations

Insurance is an important part of the modern financial system. As an industry with over 250 years of history, insurance has established itself as a mature industry. However, some problems remained unsolved in this industry. For insurance companies, information asymmetry easily lead to adverse selection and thus causing further risk underwriting for customers. Customers are usually frustrated understanding terms and conditions on insurance policies for a lack of knowledge of insurance. Therefore, we are motivated to leverage technologies to address these problems.

2.7 Significance of the Problem

Historically, insurance companies use sophisticated mathematical equations to model actuarial risk. However, the lack of efficient methods to verify the conditions of prospective clients can easily lead to adverse selection and moral hazard.

For insurance buyers, insurance policies are not always easily understandable. The management of insurance is another problem, especially when only physical copy of insurnace contracts are available. In addition, many people can easily misinterpret insurance for lack of education on insurance.

2.8 Methodology and Tools

We first do some research on the insurance industry and investigate four existing popular insurance platforms in China. Then, we identify key technologies to solve problems of insurance. Finally, we build an application called "InsurX" with three modules called "Intelligent Consultant", "Intelligent Policy Management", and "Blockchain-based Medical Record", respectively. OCR, NLP, and Blockchain technologies are used in the three modules.

We use Python as our programming language to develop the proposed platform. A web development framework called "Dash" is used to build the front-end of our application. Technologies such as OCR, NLP, and Blockchain are implemented for the three proposed modules, respectively.

3 Intelligent Consultant

3.1 Technical Design

Traditional Q & A systems use a "knowledge map" to build the mapping of questions and answers, which is essentially retrieving pre-existing answers. It would collect answers as given, and then link the questions with the answers by knowledge map. The mapping solution cannot deal with a massive amount of information. NLP algorithm based on machine learnings can understand the text semantics and make the so-called "intelligent" response.

We hope that the insurance Q & A system understands questions adaptively. It should be able to solve problems flexibly by "understanding" the questions. It should generate answer texts with smooth sentences and flexible sentence patterns.

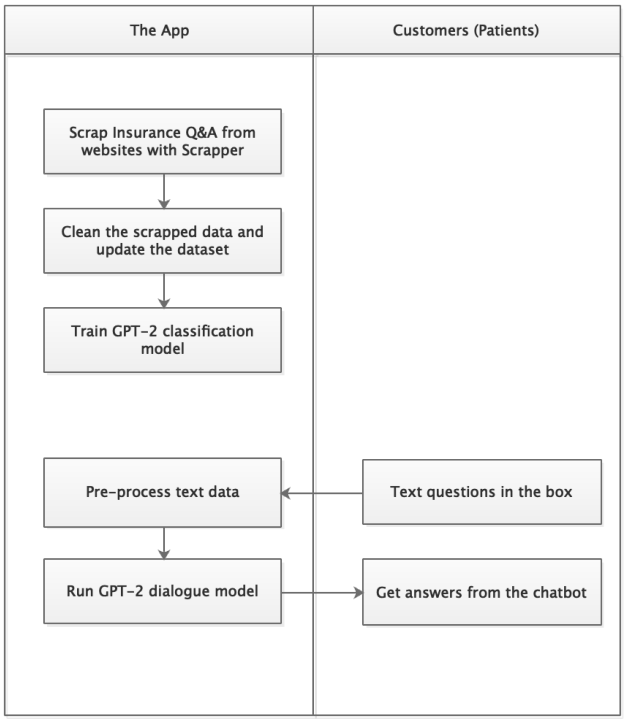


Figure 4: Flowchart of Intelligent Consultant

We fine-tune the implementation of downstream multitasking based on the GPT-2 model. Downstream tasks have two aspects. First, the expansion of the insurance corpus. Second, questions and answers. The task of corpus expansion is to collect the latest information about the insurance industry and constantly update our corpus to keep pace with the times.

3.2 Implementation

In addition to existing insurance Q & A datasets, we use the BeautifulSoup scraper to obtain insurance-related text from the web page to enrich the dataset.

For model training, we use the insurance Q & A dataset to train the GPT-2 dialogue model and use the news dataset to train the GPT-2 news classifier. The classifier is subsequently used as the filter of daily news to filter out the insurance text and continuously train the GPT-2 dialogue model.

At the webpage, the questions entered by the user are transmitted to

the back-end trained GPT-2 dialogue model, the answers are generated, and presented on the webpage.

3.2.1 Data Source and Data Processing

1. Question and answer dataset: used as a training GPT-2 dialogue model. The insurance Q & A dataset is preprocessed and sorted into TXT format, and each line contains a question and an answer.

2. Classification dataset: used to train GPT-2 classification model. Thucnews is generated based on the historical data of Sina News RSS subscription channel from 2005 to 2011. It has 740000 news documents (2.19 GB), all in UTF-8 plain text format. We re-integrate and divide the original Sina News into 14 candidate categories: finance, lottery, real estate, stock, home, education, science and technology, society, fashion, current politics, sports, constellation, game, and entertainment.

3. Insurance dataset: scraped text from insurance websites as the "insurance" category, which is subsequently merged into the classification dataset to provide data for text classification.

4. Preprocessing before data input into the model: use a tokenizer to divide each sentence into ASCII codes corresponding to words, add marks and fill zeros to the input length of the model (1024). If it is a labeled classification task, positive and negative samples of the classification dataset need to be dealt with to balance the samples.

3.2.2 Training GPT-2 Dialog Model

We use the insurance Q & A dataset to train the GPT-2 model. Before training, pre-train model parameters are loaded. The training of GPT-2 minimum model requires more than 6G GPU memory. Due to resource constraints, we use the minimum model.Hyperparameters are fine-tuned on the basis of CPM generate (GPT-2 based), batch size = 32, learning rate = 1e-5, and train 10 epochs.

3.2.3 Training GPT-2 Classification Model

We have built a news crawler classification system to explore the relevant contents of the insurance industry from the daily news, continuously update our insurance text library, and continuously train our dialogue model. Identification of insurance news is essentially a classification problem. Insurance-related news is treated as one category and un-related news as the other category. The training dataset is obtained from news datasets and insurance website crawlers. The text content labels pushed on insurance websites are labeled as "insurance". News datasets have multiple labels such as "Sports" and "Technology", which are labeled as "non-insurance"; "Insurance" and "non-insurance" are evenly selected and saved as classified datasets. We try short text classification and long text classification. Short texts use only news headlines for training, and long text adds news content as the training corpus. After the Chinese text pre-trained GPT-2 model, we add the full connection layer and return two outputs. That is, we build the binary classification model with one-hot coding. Hyperparameters are fine-tuned on the basis of CPM generate (GPT-2 based), batch size = 32, learning rate = 1e-5, and train 10 epochs. The results of the two classification models can be shown in Figure 5. The red dot is the insurance category and the blue dot is another category. The classification result of long text is better with an accuracy of 98.7%. Therefore, we suggest using the long text of news content for news screening in the insurance industry.

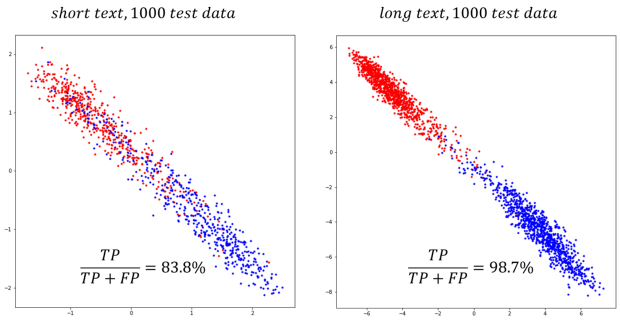


Figure 5: Model test results

3.2.4 Evaluation

The front-end display of the module is demonstrated as follows. After putting questions in the question box, texts are analyzed and answers would be displayed on the webpage.

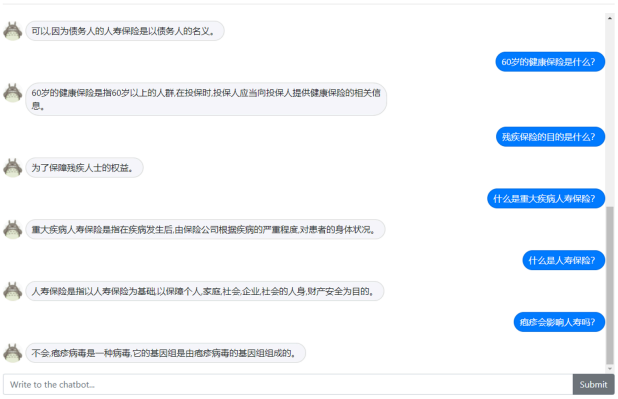


Figure 6: Intelligent Consultant front-end demonstration

4 Intelligent Policy Management

4.1 Technical Design

This functional module will be based on Baidu's EasyDL model to conduct targeted training for specific and application scenarios in the insurance industry. The model is based on Baidu's commercial model training experience, pre-trained models produced by best practices, and trained based on Baidu's self-developed EnDet entity detection model. The average accuracy of the model can reach more than 90%. In addition, the visual data management platform can be used to intelligently pre-label the uploaded pictures, and only need to check and modify to complete the labeling, and can generate virtual data in batches based on a labeled picture, quickly expand the training set, and start model training.

The model identifies specific fields for insurance industry scenarios, including 16 fields in four categories: basic insurance information, basic information of applicants, insured information, and beneficiary information. After the policy data is identified, it is structured and output in the JSON format. After the client receives the data, it builds a user-specific electronic

policy archive based on the field information and stores it locally or in the cloud to realize smart policy management.

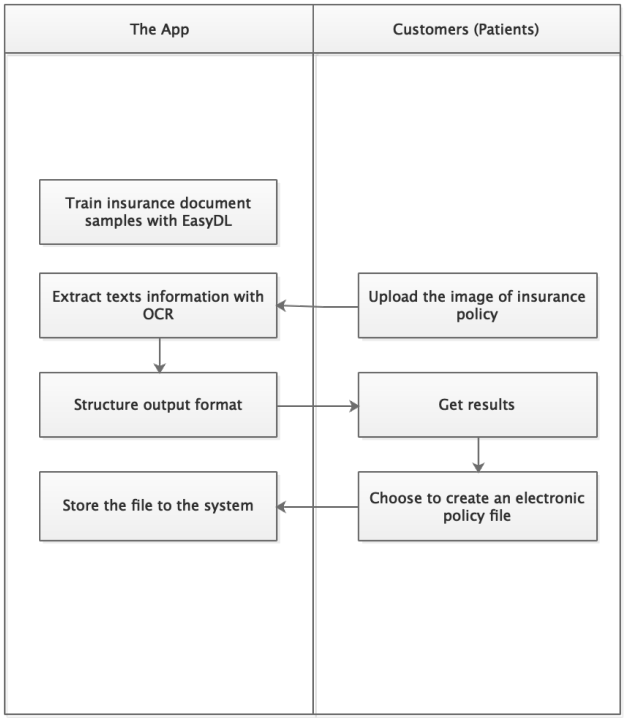


Figure 7: Flowchart of the policy recognition module

4.2 Implementation

4.2.1 Data Source and Data Processing

Since insurance policies are private property and have privacy, insurance companies and individuals will not disclose them on the Internet, and it is difficult to obtain them on the Internet. Therefore, in this module, we choose to obtain some iconic insurance policy samples from Baoxian Huize (https://www.huize.com/) and Baidu Picture (https://image.baidu.com/), Including electronic version, photo version, and handwritten version, as data samples.

4.2.2 Evaluation

Theis part will display the module functions according to all the situations that customers may encounter after entering this functional module, including OCR identification, data display, and establishment of archives.

First, after the user logs in successfully, he can enter this functional module from the "GO TO POLICY MANGER" button on the "Login successfull" page or the "POLICYMANAGER" button on the "HOMEPAGE" page (login permission is required).

After entering the policy management page, the user can drag and drop or select a policy file from the file explorer for OCR recognition and data extraction and display. The required image file is JPG, PNG, JPEG, or BMP format.

After selecting the correct policy file, this module will perform OCR identification of the policy and extract and display key field data. In this demonstration stage, two electronic versions of insurance policies, one photographic version of the policy, and one handwritten version of the policy were selected for function display to show the completeness and stability of the functions.

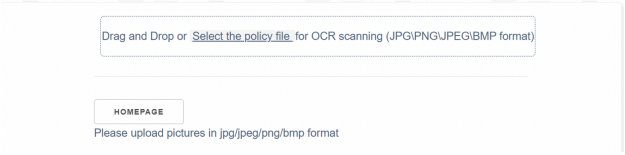
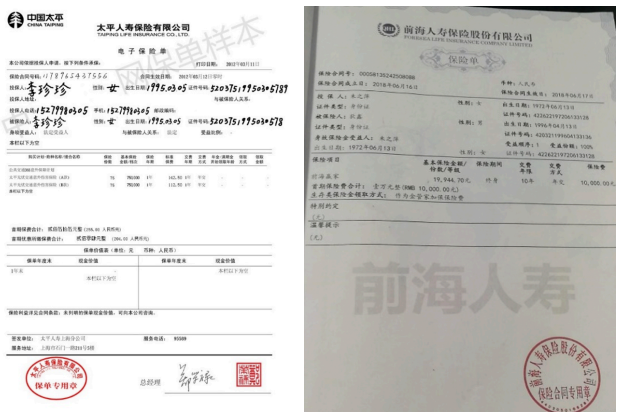


Figure 8: Upload image of insurance policy

Figure 9: Samples of electronic insurance policy

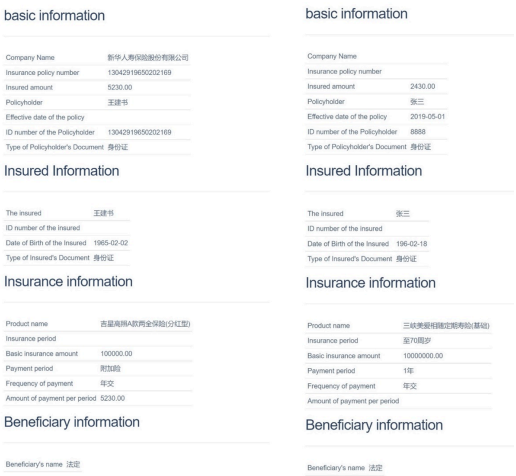


(a) A photographic insurance policy

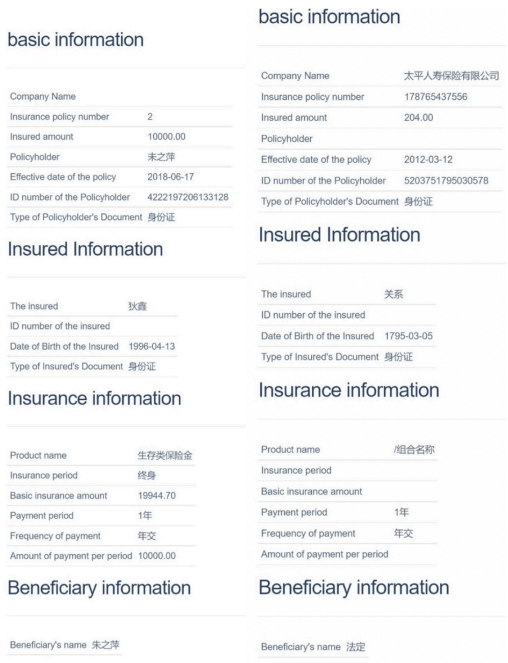
(b) A handwritten insurance policy

Figure 10: Samples of photographic and handwritten insurance policy

Upload the above four insurance policies to the system in turn, and the OCR identification and key field data extraction and display effects are shown as follows.



(a) Result for electronic samples



(b) Result for photo-graphic samples

(c) Result for hand-written samples

Figure 11: OCR text recognition results

After the policy upload and identification are completed, the user can click the "STORE ELECTRONIC INSURANCE POLICIES" button to cre-

ate a dedicated electronic policy archive, and store the key field data of the policy uploaded each time into a csv file named after the user to achieve intelligence Policy information management. In this demonstration session, the above 4 insurance policies are stored in the electronic insurance policy archive with the user name "James", and the effect is shown as follows.

username	Company Name	Insurance policy number	Insured amount	Policyholder	Effective date of the policy	ID number of the Policyholder	Type of Policyholder's Document	The insured	ID number of the insured	Date of Birth of the Insured	Type of Insured's Document	Insurance period	Basic insurance amount	Payment period	Frequency of payment	Amount of payment per period	Product name	Beneficiary's name
James	新华人寿保险股份有限公司	1.3E+16	5230	王建书		1.3E+16	身份证	王建书		1965/2/2	身份证		100000	附加险	年交	5230	吉星高照A款两全保险(分红型)	法定
James			2430	张三	2019/5/1	8888	身份证	张三		196-02-18	身份证	至70周岁	10000000	1年	年交		三峡美爱相随定期寿险(基础)	法定
James		2	10000	未之萍	2018/6/17	4.22E+15	身份证	狄鑫		#######	身份证	终身	19944.7	1年	年交	10000	生存类保险金	朱之萍
James	太平人寿保险有限公司	1.79E+11	204		2012/3/12	5.2E+15	身份证	关系		1795-03-05	身份证			1年	年交		/组合名称	法定

Figure 12: Electronic policy archives named after the user's name

5 Blockchain-based Medical Record

5.1 Technical Design

The blockchain module is integrated into the diagnosis process. A blockchain architecture usually includes the application layer, contract layer, incentive layer, consensus layer, network layer, and data layer, etc. Due to tight schedules, the project implements only the data layer. The key technology in the design is Hash 256 Algorithm, a novel hash function computed with eight 32-bit words. Diagnosis information entered by doctors would be packaged with the hash of the previous block in the blockchain to build a "transaction", after which the block would be encrypted by Hash Algorithm and added to the blockchain.

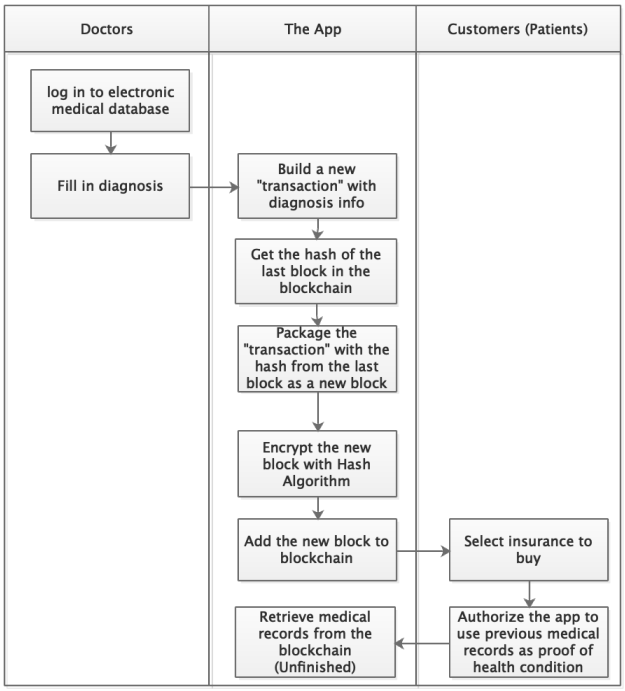


Figure 13: Flowchart of Blockchain-based medical record

5.2 Implementation

5.2.1 Data Source and Data Processing

There's no data analysis or model training in this module. However, to demonstrate the workflow, we build a simulated process called "InsurShop", which simulates the process of buying insurance online. Information for insurance companies and products is generated according to data collected from "huize.com", an online insurance broker in mainland China.

5.2.2 Evaluation

The process shows how the application integrates the blockchain technology in the "Diagnosis" and "Insurance Buying" processes. Normally, the blockchain would be held by distributed ledgers with Peer-to-Peer Networks. Due to the limited time schedule, the blockchain architecture is built without supports of distributed ledgers.

First, The doctor logs in to the website. He then fills in diagnoses for a patient.

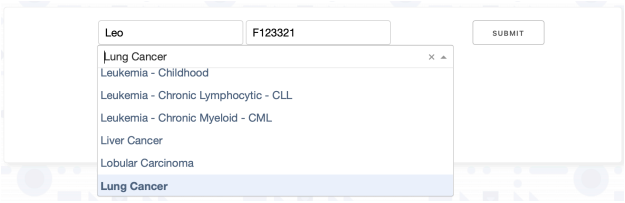


Figure 14: Doctor fills in diagnosis results

After pressing the "SUBMIT" button, a "transaction" for the medical record is created and packaged to a new block. The new block is then encrypted and added to the blockchain.

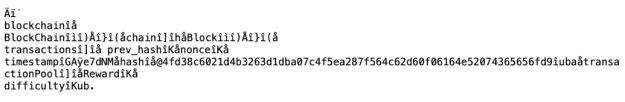


Figure 15: Open encrypted block as text

Later on, patient "Leo" logged in to the website and want to buy insurance from "INSURSHOP". He enters relevant information into the webpage.

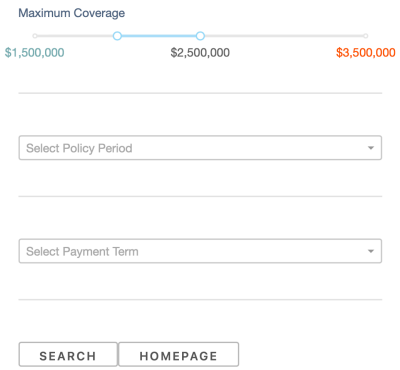


Figure 16: Customer enters relevant information at "InsurShop"

By pressing the "SUBMIT" button, the app would return matched insurances for the customer. The customer can choose to share his / her medical

records stored in the blockchain with the insurance company to verify his health conditions.

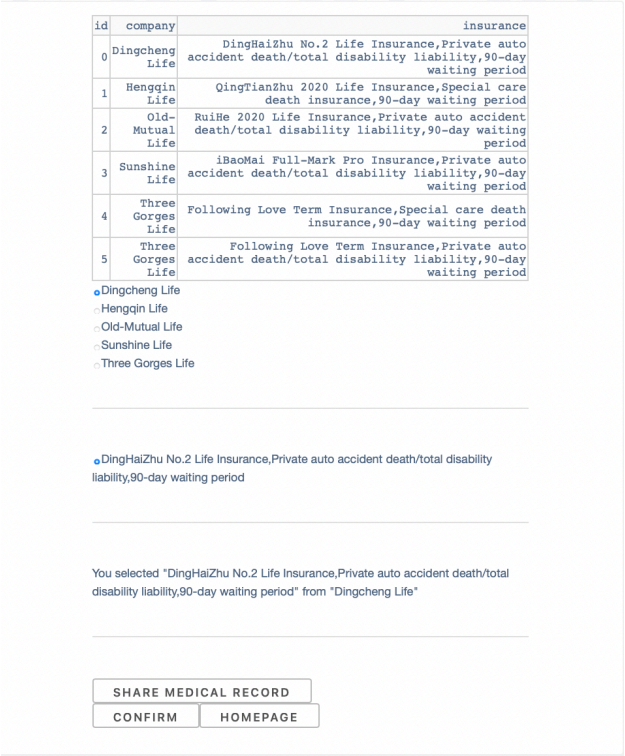


Figure 17: The app returns matched insurance for the customer

Finally, the app would forward the customer's purchasing request to the insurance company.

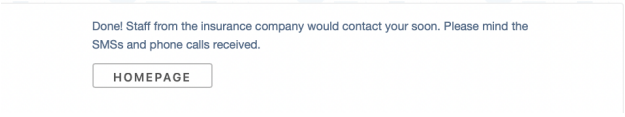


Figure 18: The app submits the purchasing request to the insurance company

6 Conclusion

The insurance industry is one of the most crucial parts of modern financial systems. In this project, we analyze the challenges faced by the insurance

industry. For insurance companies, asymmetric information and easily lead to adverse selection that could expose these companies to higher risk. The lack of efficient ways to verify the conditions of clients is the root of the risk. Also, we found that people can be frustrated by terms and conditions on insurance policies for lack of understanding of insurance. The management of insurance policies is another issue that needs to be addressed.

Looking at four major insurance platforms in mainland China, we found that technologies such as OCR, NLP, and Blockchain could transform the insurance industry on risk management and client experience improvement. Therefore, we build an InsurTech platform – "InsurX" as a prototype to demonstrate how these technologies can bring better services to insurance companies and insurance buyers. The proposed platform is developed as an application with three modules. The three modules–"Intelligent Consultant", "Intelligent Policy Management", and "Blockchain-based Medical Record", use NLP, OCR, and Blockchain technologies to implement the proposed designs. Experiments are done to verify the functionality of these modules.

"InsurX" is merely a prototype to show how technologies can transform insurance services. Some of the module, such as Blockchain-based Medical Record, has not been fully developed due to limited time. In the future, we could dig into the insurance industry and investigate other problems that need to be solved. Other technologies could also be used to improve insurance services.

References

[1] S. Nakamoto, "Re: Bitcoin p2p e-cash paper," *The Cryptography Mailing List*, 2008.

[2] G. M. Wolrich, K. S. Yap, J. D. Guilford, V. Gopal, and S. M. Gulley, "Instruction set for message scheduling of sha256 algorithm," Sept. 16 2014. US Patent 8,838,997.

[3] P. Budzianowski and I. Vulić, "Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *arXiv preprint arXiv:1907.05774*, 2019.

[4] M. Namysl and I. Konya, "Efficient, lexicon-free ocr using deep learning," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 295–301, IEEE, 2019.

Appendices

A Contribution

Chen Ming: He's the team leader who laid out the research & development plans and organizes meetings. He built the Blockchain part of this application. He designed the front-end of the application. He drafted the project proposal and the final report.
Guo Zhongyuan: He developed the NLP part of the application. He's responsible for data collection, structural design and training of dialogue model and classification model, and the APP.
Huang Yijing: He was responsible for the overall development of the intelligent policy management module, including the front-end and back-end design and implementation of functions such as OCR recognition and electronic policy filing.
Zhang Xincong: She's responsible for sourcing data needed in this project (mainly about insurance types and features on the market). Besides, She's also responsible for insurance industry existing platforms analysis and blockchain introduction.
Ma Ziqing: She made market research of existing insurance platforms. She assisted in the design of NLP module. She made the slide for the presentation. She helped write the Abstract and part of the Introduction section.
Liu Luyao: She did the background investigation and introduced algorithm of OCR technology. She analyzed some existing platforms, and compared their strength and shortcoming, and summarized the theory and procedures of OCR.

B Reflection

Chen Ming: I learned how blockchain can secure transactions by Hash Algorithm, distributed ledger, and so on. Also, I learned how to build a

simple blockchain to implement the blockchain-based medical record module. Finally, I gained insights into the insurance industry and how technologies such as blockchain can transform industry services.

Guo Zhongyuan: I learned more about NLP in this project. I made relevant thinking and implementation of the daily upgrade of APP insurance Q & A assistant. There are still areas that can be improved, such as how to serve multiple customers with limited GPUs. I will continue to explore.

Huang Yijing: I learned more about OCR from the project, and understood how to build, train and implement OCR models. At the same time, through practical applications in the project, my understanding of programming languages such as Python and HTML has also been deepened.

Zhang Xincong: As a student used to major in actuarial science, I learned more about insurerance product design when I cleared up insurance product information and realized that there was a long way to go in insurtech, because the gap between the theoretical design and reality is still wide.

Ma Ziqing: In this group project, from the early market research to the later implementation of NLP technology, I have a more comprehensive and systematic understanding of the history, current situation and development trend of the Insurtech industry. I have learned a lot and found that there still are many aspects and details can be improved by technology.

Liu Luyao: It deepens my knowledge of the insurance industry, and I learned more about the theory and algorithm of OCR in this project. I have a better understanding of the status of insurtech platforms, and the technologies used in the insurance industry. And learned the principle and process of OCR technology.