Drug Recommendation and Tracking using Blockchain and Machine Learning

Gokay Saldamli[1], Amitha S. Papetla[1], Amol Sadasivan[1], Anjali Deshmukh[1],
Sreeja Madanambeti[1] and Levent Ertaul[2]
[1]San Jose State University, San Jose, CA, USA
[2]California State University, East Bay, Hayward, CA, USA

***Abstract*— Pharmaceutical drug distribution is a complex process, prescription drugs are manufactured in massive volumes and then distributed to patients. One of the significant challenges that pharmaceutical industry is encountering across the globe is counterfeit drugs. In unauthorized drug distribution, the medication can be either ineffective or fatal. We are proposing a block chain- based solution to counter drug tampering. The trackability of drugs using block-chain from manufacturer to consumer ensures transparency. The block-chain-based tracking approach ensures the transactions are stored securely using blocks inside a digital ledger to ensure security. We have proposed a solution to track the delivery of drugs at each phase of the process during the whole transaction period using a distributed ledger. Furthermore, we are incorporating recommendations of drugs to a medical professional based on reviews. In the first phase, we have setup a hype-ledger fabric-based network for recording transactional data between each of the defined participants in the architecture. Hyper-ledger Fabric framework provides channels that consist of decentralized ledgers. This framework is given permissions and each participant's access to the ledger is controlled using a smart contract. We have defined all the functionalities of each of the participants in the supply chain process in the smart contracts. In the next phase, we have applied sentiment analysis on review text using tools Text-blob and Vader. We will input these sentiments to the machine learning algorithms and evaluate the algorithms; the best performing algorithm will accomplish the recommendation of medicines to a medical practitioner.**

***Index Terms*— Drug supply chain, blockchain, Hyper ledger Fabric, healthcare, recommendation systems, machine learning**

I. INTRODUCTION

Pharmaceuticals is one of the most significant fields that contribute to well-being of our lives. Health is one of the most incredible wealth of life. Hence, to lead a healthy life, we need the help of the best quality healthcare systems. The advancement in research conducted and the world-class technologies made the manufacturing and supply of drugs most efficient. There are a variety of medications that are currently being manufactured. Every drug has its purpose and prescription for a particular ailment. Some medicines help us to build immunity, and some drugs act as armors in our fight against diseases. From multivitamins to various advanced medication play a significant role in our lives.

The current global scenario also highlights the importance of quality medication. The pandemic made this coherent that world-class health care can help us fight any deadly virus. The vaccinations and drugs developed for immunity against COVID19 prove the amount of advancement we have achieved in quality healthcare. As for any large-scale industry, pharmaceutical industries also have their setbacks. The only difference is that any fallacy in this field would lead to deadly repercussions.

One such predominant setbacks of the pharmaceutical companies is Counterfeit drugs. An industry with such significance across the globe is facing the menace of counterfeit medication. This is mainly observed in developing countries, where there is a lack of a primary healthcare system. The drugs manufactured fraudulently, with low quality, without any information regarding the source, which is not equivalent with the defined standards, and are of low quality are called Counterfeit drugs. There are millions of cases across the globe of these counterfeits, resulting in many health conditions and deaths in some situations. These counterfeits consist of legitimate ingredients, but the proportion or quantity of these components is not accurate. They may also contain toxic and unknown elements that may lead to severe health conditions for anyone consuming them instead of alleviating the pain. The fraudulent producers might also use the name of already established pharmaceutical companies to sell their fake drugs at lower prices. This will result in the effect sale of authentic and more expensive drugs and jeopardize the lives of people. Hence, the presence of these fake drugs reduces the authenticity of the health care system. These counterfeit drugs may provide momentary relief from the pain and expose the patient to deadly side effects. Thus, tracking the manufacturing of these drugs and monitoring their involvement in the supply chain has become crucial and enhances the authentication and transparency in the supply chain. Monitoring at every stage can be accomplished by the advanced technologies available.

A. Blockchain

Blockchain is one of the revolutionary technologies which Satoshi Nakamoto introduced. It is a secure distributed database that provides a vibrant digital ledger software to store the data records, transaction logs, and information related to the transactions in the form of arranged numbers of blocks. These blocks are associated with each other using cryptographic algorithms. The digital ledger in a blockchain is decentralized; thus, a copy of the ledger is shared across the network and accessed by all participants. One of the main advantages of blockchain is that they are immutable [1]. Thus once any data that goes into a blockchain cannot tamper.

Blockchain applications nowadays have gained popularity and are being used in various areas such as transportation, healthcare, finance, education, and agriculture. Blockchain can play a vital role, especially in the healthcare and pharmaceutical industries. It can help in maintaining the authenticity of the

medical records. All the data stored inside the blockchain ledger is timestamped and tamper-proof. This provides a platform for securely handling drug-related information in the pharmaceutical industry. There are three main categories for blockchain types: public, private, and consortium [2]. In this work, we would be implementing Hyperledger Fabric as our block-chain network, which is a type of private blockchain. This type of block-chain is restricted where the access is limited to an organization and not open for public. This provides the organization complete authority over managing the data.

B. Recommendation Module

We are implementing our recommendation module on the UCI drug review dataset. We are incorporating recommendations of drugs to a medical professional based on reviews along with block-chain-based tracking. We will add these recommendations tab in the application user interface so that medical professionals can get recommendations about medicines for an ailment. Doctor/Medical professional can consider patient records before approving the recommended drug. If the drug is not suitable for the user, the doctor can decline the recommendation. We have accomplished preprocessing of data, data analysis, and sentiment analysis on review text. We will input the extracted features, sentiments as features to our algorithms for recommendations. These are the sections in our recommendation system:

- Pre-processing of data
- Exploratory data analysis
- Feature extraction
- Sentiment analysis
- Machine learning algorithms
- Weighted ensemble
- Evaluation Metrics
- Final model for recommendations.

The remaining paper will consist of the following section: Section 2 explains the literature review on various blockchain based approaches for supply chain management systems. It also provides some related work on recommendation systems for drug-related information using machine learning techniques.

II. RELATED WORK AND BACKGROUND

The global open market of drugs is complicated. With the expansion of drugs in the online market, counterfeit drugs are also increased proportionally. Choi et al. [3] discusses the anomalies like counterfeit drugs, drug diversion, and drug shortage. This research proposed centralized data storage for all the license holders in a pharmaceutical supply chain. They initiated the idea of an EPICS network where the data can be inquired between trading partners. The data will not share any event in a real-time manner, and it is shortcomings of this research. On the contrary, we build a decentralized network where data will be shared across all the trading partners.

The authors in [4] has implemented a tracking system using blockchain for credit-based tracking. They have developed a tracking system using ledger technology. The main motive behind this research was to improve logistic traceability through the protected cryptographic environment. Supply chain management is elucidated from exporter to user. This network consists of four different organizations representing respectively the importer, exporter, carrier, and regulator. The study of this paper helped in understanding supply chain management between peers and the authentication process involved in every stage.

Valkanov et al. [5] delineates the centralized data storage and express the need for a decentralized system. This research explains in detail blockchain-based storage and its concepts like immutability, Connectability, Anonymity. For timely up- dates in the network, the user downloads the wallet provided by the institution, and the transparency user seeks will be granted. It emphasizes the need for synchronization in data. To increase the reliability in blockchain, the validation of records stored in the form of blocks is imperative.

Due to the vulnerability of security breaches in transaction data, the secured systems development in the supply chain process is primordial. Xie et al. [6] describes the immutability of blockchain-based transaction data storage. The cryptographic correlation in the distributed database to store the independent chains of blocks is understandable through this research paper. They have developed a tool for agriculture-based product tracking. The method of double chain storage applied in this research ensures data security. This study enlarged our idea of sensor-based blockchain tracking systems.

For recommendation of drugs based on reviews, we have reviewed the research paper presented by Garg [7]. For feature extraction, they have used a bag of words technique but mentioned its shortcoming of including many terms without contemplating. Following the feature extraction method using sentiments, they utilized TF-IDF, TF will take the frequency of word appearance into account. Inverse document frequency (IDF) is the opposite of the number of times a specific term is in the whole document; it will catch how a term is specific to a document. Furthermore, to progress in extracting features, they opted for a word2vec method that produces word embedding. Word2vec uses several deep learning methods and takes the input of corpus of text and outputs vector space. Examining this paper assisted us in extracting features concatenated with training data for developing our recommendation systems.

Bhat et.al [8] have explored item-based recommendations for newly introduced drugs. One of the most rampant challenges in recommendation systems is a cold-start problem. They have opted to collect N- best items for the user to resolve it. They have used content-based filtering and collaborative filtering techniques and proposed a hybrid approach using both methods. The hybrid method uses the advantages of content- based to fix the disadvantages of collaborative filtering. We wish to traverse through content and collaborative filtering techniques in our recommendation algorithm.

Chang et.al [9] proposed a blockchain-based framework to enhance the benefits of the supply chain process. They have achieved the removal of intermediate processes. The basic concept of this system is a ledger that shares information across the board to all the members. All the supply chain members collaborate as a result of the shared tracking information. As all the transactions happening are transparently available to all the members, the system potentially has the properties of a decentralized system.

Blockchain was initially used for monetary transactions which are independent of any centralized financial institute.

With evolution of block-chain technology we have seen applications of this in areas like insurance, patient medical records. Tanwar et.al [10] take on task of improving of limitations of healthcare systems. One of the solutions explored was to measure performance of Hyper ledger Fabric, Hyper ledger Caliper. Another proposed improvement is an Access Control Policy Algorithm to increase ease of data sharing between healthcare entities.

Kumar et.al [11] proposed a tracking method, which uses the block chain framework and implement a very authentic and lucid supply chain model where the goal was to curb the counterfeiting problem in India. They implemented a Hyper ledger-based block chain model. This network has the transactions moving from client to the peer reviewing nodes or endorsers for "signature verification", and ultimately validating the transaction. They concluded that lower the endorsements better the performance, which is achieved by reducing the latency and increasing the throughput.

III. Architecture

The architecture is divided into two subsystems one being the blockchain based tracking system and the other being the recommendation module, we have utilized various technologies to design the current system.

A. *Blockchain Hyperledger Fabric*

A web application for a drug distribution system needs to provide an interface where users can perform medicine orders, raw material supply, update the medicine data, update orders, update the records, deliver drugs, data sharing, track drug at every step of supply, drug management, customers management etc. Lot of transactional processes. For these to happen securely, we are proposing a service-oriented framework for the users which provides a ledger and smart contract functionalities as CRUD operation service APIs. A smart contract is a transparent and conflict free way to exchange the information without need of a third-party agent.

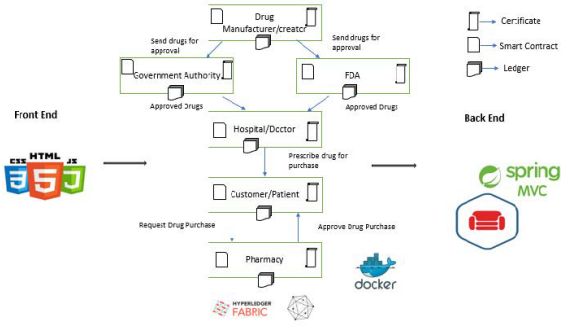


Fig. 1: Design Workflow

This is an authorization based blockchain system - only users with valid credentials can perform CRUD tasks in the system. User administrator plays a crucial role of validating and enrolling users with credentials and certificates of enrollment. Each peer node of this blockchain network has a smart contract and distributed ledger in simulated environments where they can store the data and execute the transaction as well. Due to the smart contract, these nodes can validate the transaction after the execution and write that transaction block into the ledger.

Technically the agreement between two parties is enforced by a few lines of code for smart contracts. This code consists of a defined set of regulations in which both parties agree on. This contract is automatically triggered when the specified condition meets some defined database events. There are two types of peer nodes: endorsers and committers. Endorsers peer node validates if a transaction is abiding the smart contract, if not then the transaction is denied Committer peer node validates the results of the transactions before writing it into the transaction block of the ledger. The Block-chain framework would be deployed using Hyperledger. The tools that we plan on using for deployment are Go Lang (Smart Contract).

1) *Hyperledger Fabric Features*

Hyperledger Fabric is an open-source framework for developing enterprise level block chain applications. We are utilizing the hyper ledger fabric v1.4.7 for implementing the tracking system. The latest version of the fabric delivers significant changes for both the operators and the users [12]. The updates include privacy features, improved governance of the smart contracts, new modes of operating the nodes, and support for the enterprise level applications. The following are a variety of features that are being embedded into the Hyper ledger fabric, which makes it best enterprise block chain solution:

- Assets: Smart contracts are contracts in which the contents of the buyer, seller agreement are inscribed directly into lines of code. The code and agreements are spread throughout a decentralized, distributed block-chain network. The code controls the execution of transactions, which are both trackable and irreversible.
- Chain code: Software for defining the assets and modifying the transactions of the assets.
- Ledger Features: Record of all the state transitions
- Privacy
- Security and Member ship Services
- Consensus

2) *Hyperledger Fabric's User Interface*

The web application is developed using the following technologies:

- CSS: Cascading Style Sheets (CSS) is a language of style sheets describing the look and formatting of a markup language document. Although it is most used to change the style of HTML and XHTML web pages and user interfaces, the language may be used to change the style of any XML document, including plain XML, SVG, and XUL. CSS, like HTML and JavaScript, is a foundational technology used by most websites to produce visually appealing webpages, web application user interfaces, and mobile app user interfaces.
- HTML: HTML stands for Hyper-Text Markup Language and is the standard markup language for creating web pages. HTML is a language markup that describes the semantic structure of a web page as well as presentation cues. HTML files are read by the web browsers and rendered into visible or audible web pages

- Java Script : This was used to program web page behavior which helps us to update the web page content dynamically.
- Docker: Docker is a containerization platform that is free and open source. It allows developers to package programs into containers, which are standardized executable components that combine application source code with the OS libraries and dependencies needed to run that code in any environment. Containers make it easier to deliver distributed applications, and they're becoming more popular as businesses move toward cloud-native development and hybrid multi-cloud setups. Developers can develop the containers without using Docker, but the platform makes building, deploying, and managing containers easier, simpler, and safer. Docker allows developers to use a single API to build, deploy the code, operate, and any update, and stop containers using simple commands and work-saving automation.

B. *Recommendation Module*

Recommendation systems predict the preference of the user for a given item. It does content filtering or collaborative filtering of information or a hybrid method to achieve better accuracy. It offers users various choices to purchase the best product or make the best decision. For Drug recommendations, we will deploy sentiment analysis along with the best-predicted models. First, we will analyze and pre-process data to fit the critical objective of recommending medicines. As per research [11], the best models are N-gram, Light Gradient Boosting Machine (Light GBM), Naive Bayes, decision tree. N-gram: N-gram predicts the probability of a set of co-occurring words. If N=1, it doesn't classify the emotion precisely and analyzes the words with a single corpus. N=4 classifies the emotions well; it is highly used in deep learning models. N-gram is highly useful for medicine recommendation as it helps in word similarity comparison, predictive word input, and sentiment analysis.

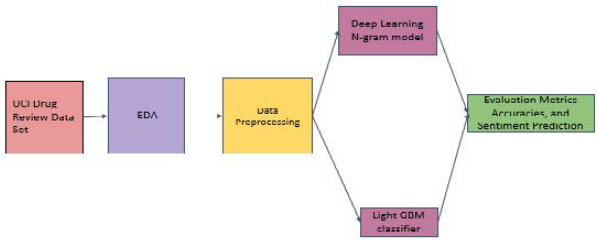


Fig. 2: Recommendation Module

Light Gradient boosting methods (GMB) like XGboost and parallel boost are very popular in machine learning. The high gradient algorithms are efficient in calculating sentiments and correlation between predicted sentiment and rating. We use models with the best accuracy and evaluation metrics are chosen to recommend drugs. The combination of recommending techniques with sentiment analysis yields a better-predicted result. Medicine recommendations can be practical for doctors to diagnose a particular disease.

The recommender engine is developed using the following technologies:

- *Python:* is the programming language for data analysis and recommendation module. Recommender System provides specific suggestions to users /medicare specialists according to their health condition and reviews. These techniques may offer decision-making capabilities. Using python pre-processing is efficient and making analysis on the sentiments of user based on the review will be coherent.
- *Numpy:* is the essential python library for Machine Learning. Numpy functions are useful for performing mathematical or scientific calculation. Mathematics is the base of machine learning, most of the mathematical tasks can be performed using NumPy.
- *Pandas:* is an open- source python library used for data manipulation and analysis. Pandas is a tool in Machine Learning used for cleaning the data and analyzing it. Pandas features includes exploring, cleaning, transforming, and visualizing from data. Matrix and vector manipulations are significant for scientific computations. Both NumPy and Pandas are essential libraries for any scientific computation, including machine learning, in python due to their intuitive syntax and matrix computation capabilities with higher performance. It provides numerous functions and different methods that speeds the process of data analysis. Pandas is a library build on the NumPy package. The two primary data structures are Series which is one dimensional and Data Frame which is two dimensional.
- *Modelling:* The recommendation module utilizes the machine learning models and algorithms from the sklearn library. The modelling is done by using a deep learning model based on N-grams and the to overcome the sentiment analysis limitation light GBM was also used. Light GBM is one of the most efficient machine learning models which employs many statistical techniques that help in boosting the gradients without making trade off with the accuracy. The modelling also y-utilizes count vectorizing and tokenizing such that the system understands the texts that are being processed.

IV. Implementation

The implementation is designed in two parts. The first part is creating a blockchain based tracking system, and the second part is the recommendation module that would utilize the data that is being tracked in the tracking system. The following pseudo code summarizes the implementation that is in progress.

A. *Blockchain Implementation*

Hyper ledger fabric is utilized to establish a tracking network between each of the participants. A web application interface was designed for each user to track the drug. Hyperledger Fabric provides predefined samples and binaries for deployment of a block-chain network. The components of the Fabric network include peer, orderer, cryptogen, configtxgen etc. The use case for the supply chain network for drug tracking was built by extending these configurations.

The configtxgen tool generates the channel and orderer artifacts such as Membership Service Providers (MSP). The MSP

contains cryptographic materials which enables the registered user to participate on a permissioned blockchain network. This tool provides the genesis block, i.e the very first block upon which additional blocks in a block-chain are added and provides permission for each peer in the network. These con- figurations are defined in the configtx.yaml file. Certificates are generated to identify and authorize different components in the network. This is done using cryptogen and the configurations of which are defined in the cryptoconfig.yaml file.

The Hyper-ledger is started using a start.sh script which then generates all of the cryptographic content and sets up the peers and orderer's as defined by the configuration file. The network generates a transaction ID every time the drug is handed off between the sender and the receiver.

1) Transaction Flow in Hyper-ledger Fabric:

We are using Java fabric SDK's, the SDK adds an abstraction layer to the wire-level proto buffer-based communication protocol that client apps utilize to communicate with a Hyperledger Fabric blockchain network. It enables Java applications to handle the Hyperledger channel and user chaincode lifecycles. The SDK also allows you to run user chaincode, query blocks and transactions on the channel, and track channel events. By implementing using Java SDK's User interface, the SDK operates on behalf of a specific user who is defined by the embedding application.

We set up the Fabric Java SDK by giving the crypto material and configuring the User context, Here users are peers who has their admin privileges, and then accessing the channel, peers, and orderer through their request ports in order to access the blockchain network that has been started. Parameters while creating a drug are name, id, type of drug, number, chemical name etc. are added as transaction request, this is the method of fabric Java SDK's. The channel instance is created and chain code is called. Chaincode/smart contract we are using 'GoLang' to create the smart contracts. In Hyperledger Fabric, a smart contract/chaincode is a program called chain-code. Go is a programming language that is used to create chain-code.

A chain-code is often used to handle business logic that has been agreed upon by network participants. A chain-code's state is only accessible to that chain-code and cannot be accessed directly by another chain-code. A chain-code in the same network can activate another chain-code to access its state if it has the required authorization. The chain-code is activated by submitting an instantiation transaction to the network by an appropriate network peer. If the government/FDA approves the transaction, the chain-code becomes operational and can accept transactions from customers via client-side applications. Any approved chain-code transactions are appended to the shared ledger. Any time after a chain-code has been instantiated, it can be upgraded by an upgrade transaction.

The orderer combines the smart contract's execution and endorsement into a block and adds it to the block-chain once all the peers, owns the smart contract have completes the transaction. This is sent out to all the network's peers. All the drugs created will be displayed. In the proposed supply-chain network architecture, the registered user initiates a transaction proposal which is performed using a specific function in a chain-code. The proposal for transaction is then submitted to the peer nodes as determined by the fabric endorsement policy. The proposal consists of various parameters such as name of the drug, holder, latitude, longitude, timestamp according to the membership service provider. The drug transaction proposal will be then executed by the defined number of peers as per the endorsement policy.

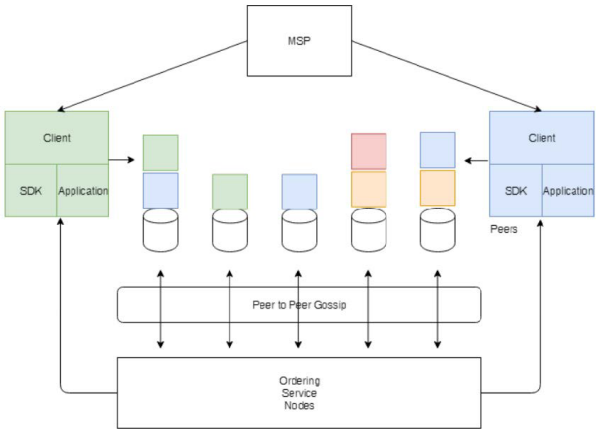


Fig. 3: Transaction flow

The endorsement policy provides detail of the number of peers that must approve a transaction on the given channel. The peers then execute the functions defined in the chain-code and must endorse the results in order to approve the proposed transaction. The output from the execution will be encrypted and recorded along with the cryptographic signatures of the peer nodes which is called as an endorsement. This endorsement consists of a transaction ID, endorser ID and endorser's signature. This transaction will then be committed into the ledger.

Once all the peer nodes complete the endorsements it is then broadcast to ordering-service in the fabric network. The response from the transaction consists of the metadata of the transaction and the endorsements. The orderer uses a consensus to calculate the order of execution of the submitted transactions in sequence for the channel. Next the ordering service collects all transactions and broadcasts them to peer node in the fabric network.

B. UI Implementation

For UI Implementation, we are using Java server pages (JSP). Creating a web application with this server side technology was a good idea. The landing page of all the APIs is DrugTrackingController.java. We have total 6 users for this web application manufacturer, fda, govt, cust, hosp, medic. Any user logs into application from the login page. After logging in using their privileges, the user is directed to its own interface page, where they can perform operations like create drug, approve/reject drug, request drug etc.

API configuration in UI: The Hyperledger application creates REST APIs to add a drug, approve a drug and fetch information of drugs. All of these are called via a go file. Some example APIs are

- Add Drug: /createDrug
- Update Drug: /createDrug
- Get list of Drugs: /readAllDrugs
- Approve: /ApproveDrug

C. *Recommendation Module*

The health care industry is leveraging the various technologies as their use cases. This leads to substantial increase in the health care data, its collection and maintenance. If valid investigations are performed on the data, then the information obtained, could be utilized to detect tumors, and many more adversities. One such significant benefits is a drug recommendation model. The drug recommendation module would be based on the various reviews given by the patients of the drugs that they were prescribed.

The data pipeline execution includes a data set, the exploratory data analysis, data pre-processing, model implementation and evaluation of results. The data workflow is described in the following sub sections.

1) *Data Set:*

The data set that the model is being trained is the UCI Machine learning drug review Data set [13].This data set gives the feedback on the drugs along with the related condition for which the drug was prescribed. The reviews are based on their benefits and the side effects that they might cause, and an ultimate remark on the drug. Overall satisfaction is depicted by rating given the drug out of 5. The data set was retrieved by Surya Kallumadi and Felix Graber by performing web crawling for their study [14]. The attributes [13] are listed below:

- urlDrugName-categorical
- Health condition-categorical
- benefitsReview-text
- sideEffectsReview-text
- commentsReview-text
- rating-numerical
- sideEffects-categorical
- effectiveness-categorical
- usefulCount is the number of users who found the review useful.

2) *EDA and Data pre-processing:*

In the data exploration section, the various attributes and the data was thoroughly investigated by the various forms of data using visualization techniques and statistical techniques. We also looked for n-grams that can best represent emotions, and the relationship with date and rating. The shape of the test data and the training data observed are as follows: Train shape: (161297, 7) Test shape: (53766, 7)

Fig. 4 illustrates the number of drugs as per the health condition. The number of drugs for the rest of the conditions all equal to 1, hence in further pre-processing these drugs would not be considered because it is not viable to recommend with only product for a given condition. Analysis is performed for a condition with at least of 2 drugs.

The sentiment analysis was performed by utilizing the techniques of Natural language Processing. This was done by stripping the review of its punctuation's and stop words. The reviews are further classified into positive and negative based on the rating attribute. Thus, converting a multi variate classification into binary classification. The ratings between 1 and 5 considered as negative, and the ratings between 6 and 10 are positive.

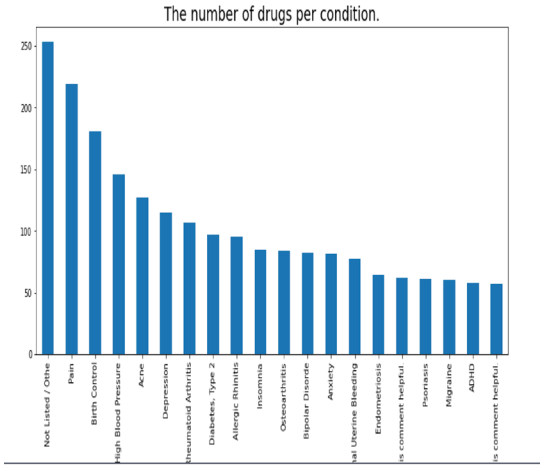


Fig. 4: Number of drugs

We have calculated the length of each review and we have plotted it against the ratings. There are ratings from 1 to 10. We opted to do this to see is there any relation between lowest or highest ratings with correspondence to length of review. We have observed all the reviews have a length more than 400 words. On an average length of reviews for all ratings is 450. The below plot shows the length of reviews against the ratings.

	uniqueID	drugName	condition	review	rating	date	usefulCount
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17
5	155963	Cialis	Benign Prostatic Hyperplasia	"2nd day on 5mg started to work with rock hard...	2	2015-11-28	43
6	165907	Levonorgestrel	Emergency Contraception	"He pulled out, but he cummed a bit in me. I t...	1	2017-03-07	5
8	74811	Keppra	Epilepsy	" I Ve had nothing but problems with the Kepp...	1	2016-08-09	11
12	191290	Pentasa	Crohn's Disease	"I had Crohn's with a resection 30 years ...	4	2013-07-06	8
...	...	...	...	...	...	...	...
53750	4705	Beliq	Weight Loss	"I have had the worst experience with this pil...	1	2015-05-02	16
53751	151266	Chantix	Smoking Cessation	"I took chantix a little over a month. It made...	1	2017-11-11	2
53754	76895	Lorcaserin	Obesity	"It didn't help me at all. I craved choco...	2	2015-08-15	8
53758	139347	Armodafinil	Narcolepsy	"This medicine kept me from sleeping the whole...	1	2014-09-30	18
53764	47656	Tapentadol	Pain	"I was prescribed Nucynta for severe neck/shou...	1	2011-11-28	20

64295 rows × 7 columns

Fig. 5: Negative Reviews

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37
7	102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is hope. I was...	10	2015-03-14	32
...	...	...	...	...	...	...	...
53760	123432	Apri	Birth Control	"I started taking Apri about 7 months ago. My ...	9	2010-08-25	18
53761	159999	Tamoxifen	Breast Cancer, Prevention	"I have taken Tamoxifen for 5 years. Side effe...	10	2014-09-13	43
53762	140714	Escitalopram	Anxiety	"I've been taking Lexapro (escitaloprgra...	9	2016-10-08	11
53763	130945	Levonorgestrel	Birth Control	"I'm married, 34 years old and I have no ...	8	2010-11-15	7
53765	113712	Arthrotec	Sciatica	"It works!!!"	9	2009-09-13	46

150768 rows × 7 columns

Fig. 6: Positive Reviews

After the above, N-grams were utilized to understand the percentage of words that were rendered in the positive or the negative sentiment. We have increased the number of grams based upon the words that are being obtained and if they are enough for classifying the sentiment. The following diagrams describe the N-grams ranging from N=1 to N=4.

When you use 1-gram, we observed that the top 5 words have the same contents, although the order of left (negative) and right (positive) are different. This means text analysis with a single corpus does not classify the emotion well. So, we will expand the corpus. Likewise, in 2-gram, the contents of the top five corpus are similar, and it is hard to classify positive and negative. This is better to classify emotions rather than previous 1-grams.From 3-gram you can see that there is a difference between positive and negative corpus.

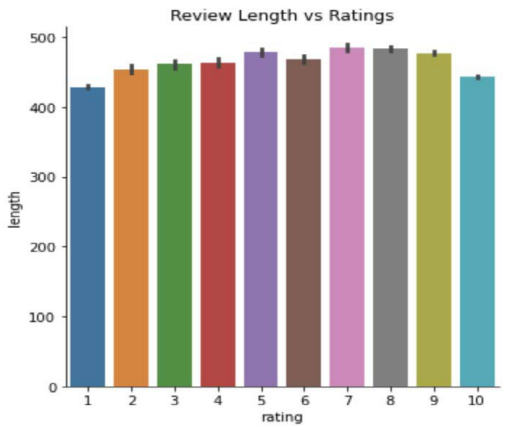


Fig. 7: Length of sequence vs rating

The 4-gram model clearly classifies the words better, and hence can be utilized in the N-gram deep learning model. Further the data is also explored based on the percentage of reviews and ratings based on the weather and the year month and day a well.

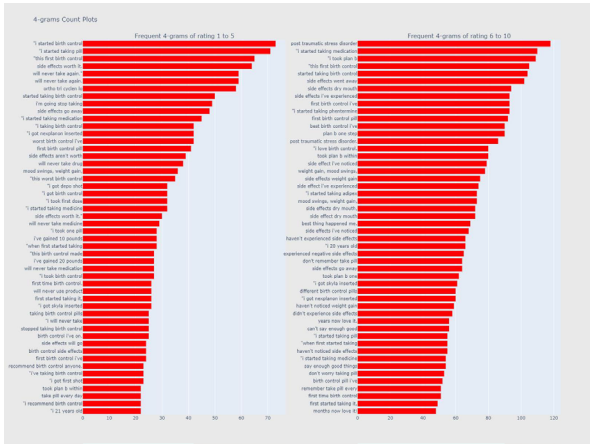


Fig. 8: Histogram

The missing values were handled. The usefulCount which is the number of users who found the review useful are also observed. It is found that in the distribution of usefulCount, the difference between minimum and maximum is 1291, which is very high with a high deviation as well. This is because people read the review irrespective of the contents being good or bad, when they look for the drugs. This makes the usefulCount very high. So when we create the model, we will normalize it by conditions, considering people's accessibility.

3) Machine Learning Model Training:

After performing the various pre-processing steps, the data is trained with the help of a few machine learning models. The proposed system [15] utilizes an NLP model called N-gram which predicts the occurrences of the various words in a given review, LightGBM and sentiment analysis.

We have implemented deep learning model using the N-gram. N-gram is exemplified as follows, say a given sentence is "This is a sentence" , if N=1 they are called unigrams, and the unigrams in the given sentence are this, is, s, sentence. if N=2, bigrams, and then bigrams are This is; is a; a sentence. If N=3; trigrams and so on. A N-gram is a connected string of N. N-Gram models are used in predicting speech and text. N-Gram models use the benefit from the elements of the Markov model, and hence are simple and scalable. Skip gram is also an interesting aspect of the N-gram modelling, where the words are not necessarily consecutive and can be skipped.

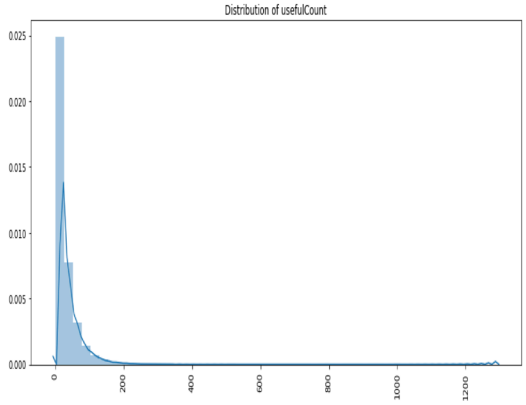


Fig. 9: Histogram

Light BGM is a machine learning model which is distributed, quick, and has extremely good performance. It is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduces memory usage. It uses two techniques one being Gradient-based One Side Sampling and the other is Exclusive Feature Bundling (EFB). The above mentioned two techniques give the Light BGM significant characteristics and also the techniques work together in such a way that they give edge to the LightBGM with respect to the other gradient boosting frame works.

The gradient-based One Side Sampling Technique for Light- GBM is explained as follows we have different data instances of variety of roles when we are computing the information gain. It is observed that the instances with larger gradients or the ones with under trained instances will contribute more to the information gain. This algorithm drops the instances with smaller gradients and keeps those instances

with gradients larger than a predefined threshold. This maintains the accuracy of the information gain obtained.

The exclusive feature bundling for LightBGM is explained as follows, in the case of high-dimensional sparse data there is a possibility of designing a nearly lossless method. This can be done by reducing the number of features, as in the sparse data the various features are mutually exclusive and these exclusive features can be bundled into a single feature called an exclusive feature bundle. This helps us in obtaining better training speeds without trading off the accuracy.

While executing the model, we created sentiment as an attribute form the rating field that is there, then split the test data into training and test data. After which count vectorization was performed. Vectorization is the process of transform the text data into a machine-readable format. The words are represented as vectors. Count Vectorizer tokenizes the text along with performing pre-processing. Tokenizing means splitting the sentence into words. It removes the punctuation marks and converts all the words to lowercase. In the process of modeling, we used deep learning model with n-gram, and additionally used a machine learning model called LightGBM to overcome the limitation of natural language processing. This improved the accuracies as well. Text Blob is used for predicting sentiments. It is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. There was k-fold cross validation to perform hyper parameter tuning. Then further there was emotional analysis performed and the drug predictions were fore-casted based on the positive and negative sentiments of the drug review data trained using the lightGBM. The positive words were calculated by taking the ratio of the number of positive words and the sum of the number of positive words and the negative words. If the ratio is less than 0.5 then the sentiment was rewarded to be positive and if greater than 0.5 then it is classified negative sentiment. Hence the final predictions are obtained by culminating the predictions from all the models. The first being from the deep learning model, the lightGBM machine learning model, and the final sentiment analysis performed with the help of an emotional dictionary, to get further accurate predictions.

V. Evaluation and Results

A. The Block-chain based drug tracking network

When we create a drug name Ibuprofen, the success transaction is processed, and a new block is created. The new block14 will possess transaction number, name of drug, Timestamp of drug creation and its approval status will be shown as pending in the transaction info. In the Fig. 10 one can see the block 14 created and the transaction info of the block.

Once the transaction is approved by FDA, the new block 15 is created and the transaction info of block 15 updates the FDA approval from pending to approved yet it shows the government approval is pending. When the Government approves the Ibuprofen drug then drug Block16 is created and its transaction info updates the status of government approval from pending to approved.

Hospital/Pharmacy can get the info of the approved drugs. Customer orders the drug for their condition and it goes to approval for hospital/pharmacy.

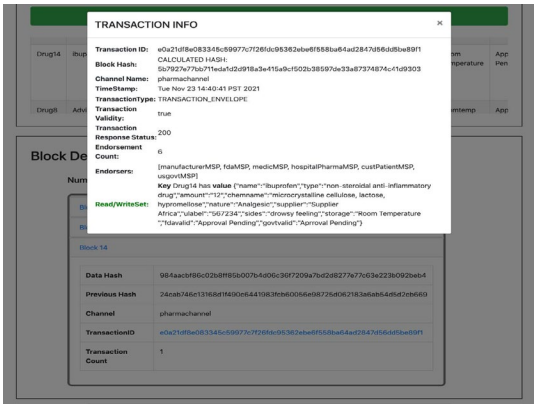


Fig. 10: New drug creation

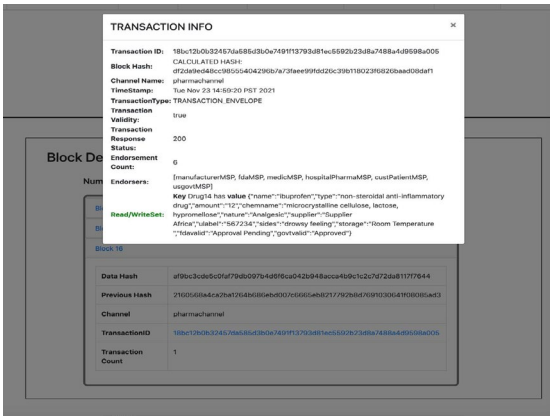


Fig. 11: Drug approved status

When the hospital/pharmacy approves/ rejects the drug ordered by the patient/customer then a new block 20 will be created. In the transaction info, it contains the information about the Endorsers (Manufacturer, FDA, USgovt, hospital/pharma, customer/patient) of drug. Once it is approved, customer/patient can view the approved status on their admin page.

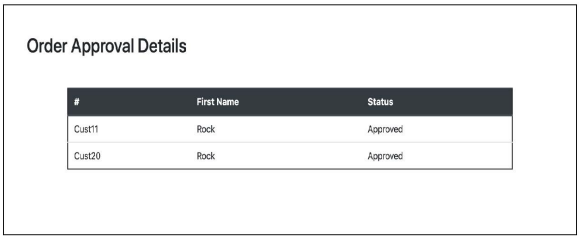


Fig. 12: Customer View of the Approved Drugs

B. Recommendation Module

After performing the initial exploratory data analysis and performing the pre-processing which includes removing the null values, the reviews which didn't have specified conditions, the conditions which had only a single drug listed under them, stripping the words of the stop words punctuations.

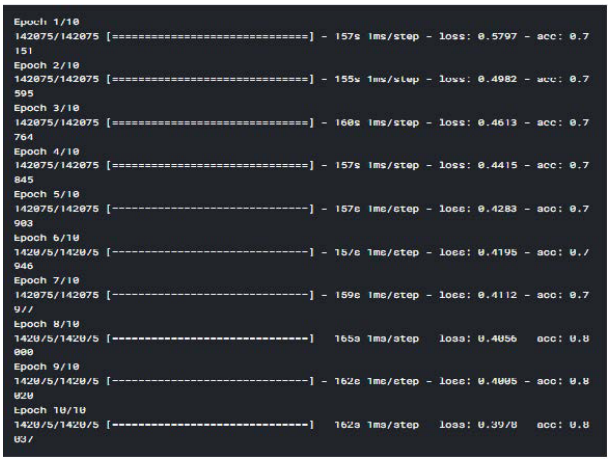
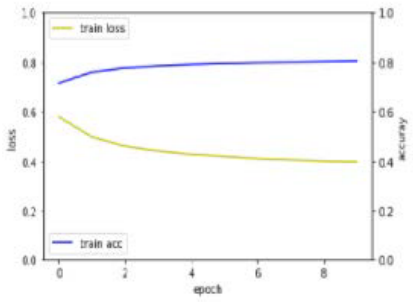


Fig. 13: Epoch

The modelling is done for the drug review data. The drug review data is evaluated based on two different models, the first being a deep learning based N-gram model, where the N was concluded to be 4 after pre-processing the word count plots with respect to the positive and negative reviews. This utilizes Tensor-flow back-end and keras models for implementing the neural network. The model summary and the accuracies and loss metrics are observed below:



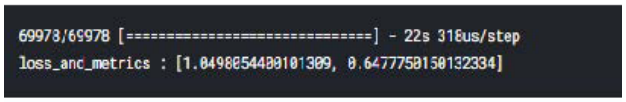


Fig. 14: Loss Curves and metrics

To further improve the accuracies observed in the N-gram based deep learning model, we tried the Light GBM machine learning model. The Light GBM is extremely fast without trading of the accuracy. This is because of the effective feature bundling and the gradient-based One Side Sampling techniques that are being utilized in the Light GBM. The final predictions culminating all the models are observed as seen in Fig. 14. The accuracies were found to be nearly 70 percent for LightGbm and the accuracies were found to be around 80 percent for the deep learning model. Thus, culminating the models and giving out the predictions with respect to the useful count was found to be most effective.

VI. Conclusion

In this paper we have introduced a block-chain-based secure supply chain management system that provides a transparent and secured way of tracking drugs from the manufacturer to the patient. We have designed a smart contract that provides various features to manage the transit of drugs at each stage of the architecture for solving the issue related to counterfeit drugs. This paper also explores the usage of Hyper-ledger Fabric to develop the back-end infrastructure that provides the building blocks of our block-chain network. Finally, we have incorporated machine learning models to predict drugs based on the health conditions and collect reviews from the patient to further improve our trained model. The result from this paper shows us that the proposed solution can provide a secure and transparent method to track drugs in a supply chain.

The future improvements include expanding the network structure to a more practical scenario. This paper considers limited number of stakeholders due to the computational requirements. In the future this platform could be scaled for a real-world scenario with the help of a larger infrastructure. The drug recommendation could be modified for providing useful analytics such as drug usages and feedback reviews. The user data could be collected to further improve the machine learning model. Future work includes a better approach on the integration of the fabric pipeline and the recommendation pipeline with a proper user interface.

References

[1] R. Chatterjee, R. Chatterjee, I. C. on Computational Intelligence, and Networks Odisha, *An Overview of the Emerging Technology: Blockchain*. 2017 3rd International Conference on Computational Intelligence and Networks (CINE), Piscataway, New Jersey :: IEEE, 2017.

[2] M. Niranjanamurthy, B. N. Nithya, and S. Jagannatha, "Analysis of blockchain technology: pros, cons and swot," *Cluster Computing*, vol. 22, no. 6, pp. 14743–14757, 2019.

[3] J. B. Choi, J. Rogers, and E. C. Jones, "The impact of a shared pharmaceutical supply chain model on counterfeit drugs, diverted drugs, and drug shortages," in *2015 Portland International Conference on Management of Engineering and Technology (PICMET)*, pp. 1879–1889, 2015.

[4] R. G.S. and M. Dakshayini, "Block-chain implementation of letter of credit based trading system in supply chain domain," in *2020 International Conference on Mainstreaming Block Chain Implementation (ICOMBI)*, pp. 1–5, 2020.

[5] V. Valkanov, M. Petrov, D. Rusev, and I. Radeva, "Modelling distributed fault-tolerant high availability storage cluster based on block-chain concepts for tracking scientific-research progress," in *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, pp. 590–595, 2020.

[6] C. Xie, Y. Sun, and H. Luo, "Secured data storage scheme based on block chain for agricultural products tracking," in *2017 3rd International Conference on Big Data Computing and Communications (BIGCOM)*, pp. 45–50, 2017.

[7] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pp. 175–181, 2021.

[8] S. Bhat and K. Aishwarya, "Item-based hybrid recommender system for newly marketed pharmaceutical drugs," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2107–2111, 2013.

[9] S. E. Chang, Y.-C. Chen, and M.-F. Lu, "Supply chain re-engineering using blockchain technology: A case of smart contract based tracking process," *Technological Forecasting and Social Change*, vol. 144, pp. 1–11, 2019.

[10] S. Tanwar, K. Parekh, and R. Evans, "Blockchain-based electronic healthcare record system for healthcare 4.0 applications," *Journal of Information Security and Applications*, vol. 50, p. 102407, 2020.

[11] A. Kumar, D. Choudhary, M. S. Raju, D. K. Chaudhary, and R. K. Sagar, "Combating counterfeit drugs: A quantitative analysis on cracking down the fake drug industry by using blockchain technology," in *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pp. 174–178, 2019.

[12] https://hyperledger-fabric.readthedocs.io/en/release-2.2/fabric model. html. [Online; accessed 15-October-2021].

[13] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+ %28Druglib.com%29. [Online; accessed 15-October-2021].

[14] F. Gra¨ßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross- data learning," in *Proceedings of the 2018 International Conference on Digital Health*, DH '18, (New York, NY, USA), p. 121–125, Association for Computing Machinery, 2018.

[15] K. Abbas, M. Afaq, T. Ahmed Khan, and W.-C. Song, "A blockchain and machine learning-based drug supply chain management and rec- ommendation system for smart pharmaceutical industry," *Electronics*, vol. 9, no. 5, p. 852, 2020.