



(51) International Patent Classification:

G06F 16/35 (2025.01) G06F 40/289 (2020.01)
G06F 40/30 (2020.01)

(21) International Application Number:

PCT/IN2024/051334

(22) International Filing Date:

20 July 2024 (20.07.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

202321049117 20 July 2023 (20.07.2023) IN

(71) Applicant: **JIO PLATFORMS LIMITED** [IN/IN]; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN).

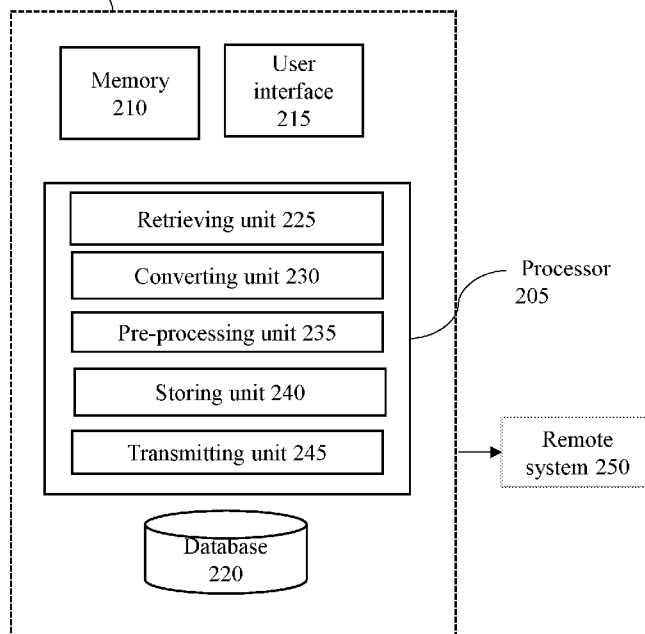
(72) Inventors: **BHATNAGAR, Aayush**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN).

MURARKA, Ankit; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **KOLARIYA, Jugal Kishore**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **KUMAR, Gaurav**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 380006 (IN). **SAHU, Kishan**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **VERMA, Rahul**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **MEENA, Sunil**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **GURBANI, Gourav**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **CHAUDHARY, Sanjana**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMED-

(54) Title: SYSTEM AND METHOD FOR PRE-PROCESSING DOCUMENTS

System 120

FIG. 2



(57) Abstract: The present disclosure relates to a system (120) and a method (500) for pre-processing the documents is provided. The method (500) includes the step of retrieving one or more documents to be pre-processed from a database (220). The method (500) further includes the step of converting the one or more documents into a text file. The method (500) further includes the step of pre-processing the text file to a format compatible for training a model.

ABAD 380006 (IN). **GANVEER, Chandra Kumar**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **DE, Supriya**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **DEBASHISH, Kumar**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN). **MEHUL, Tilala**; OFFICE-101, SAFFRON, NR. CENTRE POINT, PANCHWATI 5 RASTA, AMBAWADI, GUJARAT, AHMEDABAD 380006 (IN).

(74) **Agent: JAPHET, Chinthan**; K Law (Krishnamurthy and Co.), 4th Floor, Prestige Takt, No 23, Kasturba Road Cross, Bangalore, Bangalore 560 001 (IN).

(81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*
 — *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

SYSTEM AND METHOD FOR PRE-PROCESSING DOCUMENTS

FIELD OF THE INVENTION

[0001] The present disclosure relates to the field of wireless communication networks, and more specifically relates to a method and a system for pre-processing documents.

BACKGROUND OF THE INVENTION

[0002] Artificial intelligence (AI) models have shown remarkable capabilities in various applications, including natural language processing, text generation, and information retrieval. Training such models typically involves feeding large amounts of textual data to algorithms that learn patterns and generate contextually appropriate responses.

[0003] Word documents are a widely used format for storing textual information, containing valuable knowledge and insights across various domains. However, directly training AI models on Word documents poses several challenges. These challenges arise due to the presence of structural elements such as headers, footers, tables, images, and formatting tags within the documents. Such elements can interfere with the learning process and hinder the model's ability to generate accurate and meaningful responses.

[0004] Conventionally, training data is pre-processed to remove noise and irrelevant elements, ensuring that the model can focus on the relevant text content. However, existing pre-processing techniques are often generic and not specifically tailored for Word documents. Consequently, they may not adequately handle the complex structure and formatting elements present in Word documents.

[0005] There is, therefore, a need for an improved method and system that can pre-process documents in a manner that retains the raw data while effectively extracting and normalizing the text content. This pre-processing should remove formatting tags, irrelevant structural elements, and maintain the integrity of the textual information.

Additionally, the pre-processing should enhance tokenization, handle special considerations, and provide a suitable framework for ML training.

SUMMARY OF THE INVENTION

[0006] One or more embodiments of the present invention provides a method and a system for pre-processing documents.

[0007] In one aspect of the present invention, the method for pre-processing the documents is disclosed. The method includes the step of retrieving from a database one or more documents to be pre-processed. The method further includes the step of converting the one or more documents into a text file. The method further includes the step of pre-processing the text file to a format compatible for training a model.

[0008] In an embodiment, the one or more documents include at least one of Method of Procedure (MOP) documents, technical specifications documents and product releases documents.

[0009] In an embodiment, the step of converting the one or more documents into the text file includes the step of eliminating one or more non-relevant elements from the one or more documents including at least one of images, tables, graphs, page numbers, and structural components.

[0010] In an embodiment, the step of pre-processing the text file to a format compatible for training the model includes the steps of extracting text content from the text file and normalizing the text content extracted utilizing at least one normalizing technique. The at least one normalizing technique includes lower casing the extracted text content and removing irrelevant formatting elements from the extracted text content. The irrelevant formatting elements includes at least one of sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines.

[0011] In an embodiment, the pre-processed text file is stored in a file system of the database for training the model. In an embodiment, the method further includes the step of storing the pre-processed text file in the database. In an embodiment, the

method further comprises the step of transmitting the pre-processed text file to a remote system to train the model therein.

[0012] In another aspect of the present invention, the system for pre-processing the documents is disclosed. The system includes a retrieving unit configured to retrieve one or more documents to be pre-processed from a database. The system further includes a converting unit configured to convert the one or more documents into a text file. Further, the system includes the pre-processing unit configured to pre-process the text file to a format compatible for training the model.

[0013] In yet another aspect of the invention, a non-transitory computer-readable medium having stored thereon computer-readable instructions that when executed by a processor. The processor is configured to retrieve one or more documents to be pre-processed from a database. The processor is configured to convert the one or more documents into a text file. The processor is configured to pre-process the text file to a format compatible for training the model.

[0014] Other features and aspects of this invention will be apparent from the following description and the accompanying drawings. The features and advantages described in this summary and in the following detailed description are not all-inclusive, and particularly, many additional features and advantages will be apparent to one of ordinary skill in the relevant art, in view of the drawings, specification, and claims hereof. Moreover, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes and may not have been selected to delineate or circumscribe the inventive subject matter, resort to the claims being necessary to determine such inventive subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] The accompanying drawings, which are incorporated herein, and constitute a part of this disclosure, illustrate exemplary embodiments of the disclosed methods and systems in which like reference numerals refer to the same parts throughout the different drawings. Components in the drawings are not necessarily to scale, emphasis

instead being placed upon clearly illustrating the principles of the present disclosure. Some drawings may indicate the components using block diagrams and may not represent the internal circuitry of each component. It will be appreciated by those skilled in the art that disclosure of such drawings includes disclosure of electrical components, electronic components or circuitry commonly used to implement such components.

[0016] **FIG. 1** is an exemplary block diagram of a communication system for pre-processing documents, according to one or more embodiments of the present disclosure;

[0017] **FIG. 2** is an exemplary block diagram of a system for pre-processing the documents, according to one or more embodiments of the present disclosure;

[0018] **FIG. 3** is an exemplary block diagram of an architecture of the system of the **FIG. 2**, according to one or more embodiments of the present disclosure;

[0019] **FIG. 4** is a signal flow diagram for pre-processing the documents, according to one or more embodiments of the present disclosure; and

[0020] **FIG. 5** is a flow chart illustrating a method for pre-processing the documents, according to one or more embodiments of the present disclosure.

[0021] The foregoing shall be more apparent from the following detailed description of the invention.

DETAILED DESCRIPTION OF THE INVENTION

[0022] Some embodiments of the present disclosure, illustrating all its features, will now be discussed in detail. It must also be noted that as used herein and in the appended claims, the singular forms "a", "an" and "the" include plural references unless the context clearly dictates otherwise.

[0023] Various modifications to the embodiment will be readily apparent to those skilled in the art and the generic principles herein may be applied to other embodiments. However, one of ordinary skill in the art will readily recognize that the present disclosure including the definitions listed here below are not intended to be

limited to the embodiments illustrated but is to be accorded the widest scope consistent with the principles and features described herein.

[0024] A person of ordinary skill in the art will readily ascertain that the illustrated steps detailed in the figures and here below are set out to explain the exemplary embodiments shown, and it should be anticipated that ongoing technological development will change the manner in which particular functions are performed. These examples are presented herein for purposes of illustration, and not limitation. Further, the boundaries of the functional building blocks have been arbitrarily defined herein for the convenience of the description. Alternative boundaries can be defined so long as the specified functions and relationships thereof are appropriately performed. Alternatives (including equivalents, extensions, variations, deviations, etc., of those described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. Such alternatives fall within the scope and spirit of the disclosed embodiments.

[0025] Referring to **FIG. 1**, **FIG. 1** illustrates an exemplary block diagram of a communication system **120** for pre-processing documents, according to one or more embodiments of the present disclosure. The communication system **120** includes a network **105**, a User Equipment (UE) **110**, a server **115**, and a system **120**. The UE **110** aids a user to interact with the system **120** for pre-processing the documents. In an embodiment, the UE **110** is one of, but not limited to, any electrical, electronic, electro-mechanical or an equipment and a combination of one or more of the above devices such as smartphones, virtual reality (VR) devices, augmented reality (AR) devices, laptop, a general-purpose computer, desktop, personal digital assistant, tablet computer, mainframe computer, or any other computing device.

[0026] For the purpose of description and explanation, the description will be explained with respect to the UE **110**, or to be more specific will be explained with respect to a first UE **110a**, a second UE **110b**, and a third UE **110c**, and should nowhere be construed as limiting the scope of the present disclosure. Each of the first UE **110a**, the second UE **110b**, and the third UE **110c** is configured to connect to the server **115**

via the network **105**. In alternate embodiments, the UE **110** may include a plurality of UEs as per the requirement. For ease of reference, each of the first UE **110a**, the second UE **110b**, and the third UE **110c**, will hereinafter be collectively and individually referred to as the “User Equipment (UE) **110**”.

[0027] The network **105** may include, by way of example but not limitation, at least a portion of one or more networks **105** having one or more nodes that transmit, receive, forward, generate, buffer, store, route, switch, process, or a combination thereof, etc. one or more messages, packets, signals, waves, voltage or current levels, some combination thereof, or so forth. The network **105** may also include, by way of example but not limitation, one or more of a wireless network, a wired network, an internet, an intranet, a public network, a private network, a packet-switched network, a circuit-switched network, an ad hoc network, an infrastructure network, a Public-Switched Telephone Network (PSTN), a cable network, a cellular network, a satellite network, a fiber optic network, or some combination thereof.

[0028] The network **105** includes, by way of example but not limitation, one or more of a wireless network, a wired network, an internet, an intranet, a public network, a private network, a packet-switched network, a circuit-switched network, an ad hoc network, an infrastructure network, a Public-Switched Telephone Network (PSTN), a cable network, a cellular network, a satellite network, a fiber optic network, or some combination thereof. The network **105** may include, but is not limited to, a Third Generation (3G), a Fourth Generation (4G), a Fifth Generation (5G), a Sixth Generation (6G), a New Radio (NR), a Narrow Band Internet of Things (NB-IoT), an Open Radio Access Network (O-RAN), and the like.

[0029] The communication system **120** includes the server **115** accessible via the network **105**. The server **115** may include by way of example but not limitation, one or more of a standalone server, a server blade, a server rack, a bank of servers, a server farm, hardware supporting a part of a cloud service or system **120**, a home server, hardware running a virtualized server, one or more processors executing code to function as the server **115**, one or more machines performing server-side functionality

as described herein, at least a portion of any of the above, some combination thereof. In an embodiment, the entity may include, but is not limited to, a vendor, a network operator, a company, an organization, a university, a lab facility, a business enterprise side, a defense facility side, or any other facility that provides service.

[0030] The communication system **120** further includes the system **120** communicably coupled to the server **115** and the UE **110** via the network **105**. The system **120** is adapted to be embedded within the server **115** or is embedded as the individual entity. However, for the purpose of description, the system **120** is illustrated as remotely coupled with the server **115**, without deviating from the scope of the present disclosure.

[0031] Operational and construction features of the system **120** will be explained in detail with respect to the following figures.

[0032] **FIG. 2** illustrates an exemplary block diagram of the system **120** for pre-processing the documents, according to one or more embodiments of the present disclosure.

[0033] As per the illustrated embodiment, the system **120** includes one or more processor **205**, a memory **210**, a user interface **215** and a database **220**. For the purpose of description and explanation, the description will be explained with respect to one processor **205** and should nowhere be construed as limiting the scope of the present disclosure. In alternate embodiments, the system **120** may include more than one processor **205** as per the requirement of the network **105**. The one or more processors **205**, hereinafter referred to as the processor **205** may be implemented as one or more microprocessors, microcomputers, microcontrollers, digital signal processors, central processing units, state machines, logic circuitries, single board computers, and/or any devices that manipulate signals based on operational instructions.

[0034] As per the illustrated embodiment, the processor **205** is configured to fetch and execute computer-readable instructions stored in the memory **210**. The memory **210** may be configured to store one or more computer-readable instructions or routines

in a non-transitory computer-readable storage medium. The memory **210** may include any non-transitory storage device including, for example, volatile memory such as RAM, or non-volatile memory such as disk memory, EPROMs, FLASH memory, unalterable memory, and the like.

[0035] In an embodiment, the user interface **215** includes a variety of interfaces, for example, interfaces for data input and output devices, referred to as Input/Output (I/O) devices, storage devices, and the like. The user interface **215** facilitates communication of the system **120**. In one embodiment, the user interface **215** provides a communication pathway for one or more components of the system **120**.

[0036] In an embodiment, the database **220** is one of, but not limited to, a centralized database, a cloud-based database, a commercial database, an open-source database, a distributed database, an end-user database, a graphical database, a No-Structured Query Language (NoSQL) database, an object-oriented database, a personal database, an in-memory database, a document-based database, a time series database, a wide column database, a key value database, a search database, a cache databases, and so forth. The foregoing examples of database **220** types are non-limiting and may not be mutually exclusive e.g., the database **220** can be both commercial and cloud-based, or both relational and open-source, etc.

[0037] In order for the system **120** to pre-process the documents, the processor **205** includes one or more modules/units. In one embodiment, the one or more modules includes, but not limited to, a retrieving unit **225**, a converting unit **230**, a pre-processing unit **235**, a storing unit **240**, and a transmitting unit **245**. The system **120** is further communicably coupled to a remote system **250**.

[0038] The retrieving unit **225**, the converting unit **230**, the pre-processing unit **235**, the storing unit **240**, the transmitting unit **245**, and the remote system **250**, in an embodiment, may be implemented as a combination of hardware and programming (for example, programmable instructions) to implement one or more functionalities of the processor **205**. In the examples described herein, such combinations of hardware and programming may be implemented in several different ways. For example, the

programming for the processor **205** may be processor-executable instructions stored on a non-transitory machine-readable storage medium and the hardware for processor **205** may comprise a processing resource (for example, one or more processors), to execute such instructions. In the present examples, the memory **210** may store instructions that, when executed by the processing resource, implement the processor **205**. In such examples, the system **120** may comprise the memory **210** storing the instructions and the processing resource to execute the instructions, or the memory **210** may be separate but accessible to the system **120** and the processing resource. In other examples, the processor **205** may be implemented by electronic circuitry.

[0039] In order to initiate the pre-processing of the documents, the retrieving unit **225** of the processor **205** is configured to retrieve one or more documents to be pre-processed from the database **220**. In an embodiment, the one or more documents include at least one of Method of Procedure (MOP) documents, technical specifications documents (TSD) and product release documents. In an embodiment, the MOP refers to detailed steps and procedures for various planning, deploying, managing, and maintaining the network **105**. In another embodiment, the MOP documents facilitate, for example, but not limited to, efficient implementation of technologies, optimized performance, reliability, and security throughout the network **105**. The MOP documents include, but are not limited to, site survey and planning MOP, base station installation MOP, network slicing MOP, service deployment MOP, security and compliance MOP.

[0040] In one embodiment, the TSDs refers to detailed guidelines specifying the requirements, standards, and technical parameters essential for designing, deploying, operating, and maintaining the network **105** and services. The TSDs includes, but not limited to, radio access technology standards, network architecture, performance requirements, security standards, interoperability and compatibility, regulatory compliance, testing and validation procedures, documentation and reporting.

[0041] In one embodiment, the product release documents in the network **105** refers to broad documents. The product release documents specify the details and

specifications of recently launched products related to the technologies. The product release documents serve as essential guidelines for such as, but not limited to, manufacturers, vendors, and stakeholders involved in the development, deployment, and commercialization of products and solutions. In one embodiment, the products and solutions include, but are not limited to, modems and routers, small cells, Internet of things (IoT) devices, network slicing solutions, edge computing solutions. In one embodiment, the product release documents include, but not limited to, product overview, technical specifications, features and capabilities, network compatibility, performance metric, security feature, deployment guidelines, regulatory compliance.

[0042] On receipt of the one or more documents, the converting unit **230** is configured to convert the one or more documents into a text file. In an embodiment, the converting unit **230** is configured to eliminate one or more non-relevant elements from the one or more documents. The one or more non-relevant elements include at least one of, but not limited to, images, tables, graphs, page numbers, and structural components as provided in the one or more documents.

[0043] The pre-processing unit **235** is configured to pre-process the text file to a format compatible for training a model. In one embodiment, the model includes at least one of, but not limited to, Machine Learning (ML), deep learning, natural language processing, reinforcement learning models, and generative models. The model utilizes a variety of ML techniques, such as supervised learning, unsupervised learning, and reinforcement learning.

[0044] In one embodiment, the supervised learning is a type of machine learning algorithm, which is trained on a labeled dataset. The supervised learning refers to each training example paired with an output label. The supervised learning algorithm learns to map inputs to a correct output. In one embodiment, the unsupervised learning is a type of machine learning algorithm, which is trained on data without any labels. The unsupervised learning algorithm tries to learn the underlying structure or distribution in the data in order to discover patterns or groupings. In one embodiment, the

reinforcement learning is a type of machine learning where an agent learns to make decisions by performing actions in an environment to maximize cumulative reward. The agent receives feedback in the form of rewards or penalties based on the actions it takes, and it learns a path that maps states of the environment to the best actions.

[0045] In order to pre-process the text file, the pre-processing unit **235** is configured to extract text content from the text file. In one embodiment, the text content is at least one of, but not limited to, words and characters as provided in the text file.

[0046] Thereafter, the pre-processing unit **235** normalizes the extracted text content utilizing at least one normalizing technique. In an embodiment, the at least one normalizing technique includes lower casing the extracted text content and removing irrelevant formatting elements from the extracted text content. In an embodiment, for lower casing of the extracted text content, pre-processing unit standardizes and structures the extracted text content. On doing so, natural language processing tasks such as text classification, sentiment analysis and language modeling, is performed effectively and accurately on the extracted text content.

[0047] Further, in an embodiment, the irrelevant formatting elements include at least one of, but not limited to, sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines. Accordingly, the pre-processing unit **235** is configured to remove the irrelevant formatting elements from the extracted text content.

[0048] Subsequent to pre-processing the text file to the compatible format, the storing unit **240** is configured to store the pre-processed text file in a file system **330** of the database **220**. More specifically, the pre-processed text file is stored in the file system **330** of the database **220** for model training.

[0049] In one embodiment, the file system **330** refers to structured way of storing and organizing extracted context on the storage medium, such as, but not limited to, the hard drive, SSD (Solid State Drive), and any other storage device. The transmitting unit **245** is configured to transmit the pre-processed text file to the remote system **250**

to train the model. The remote system **250** typically refers to the computing infrastructure that leverages the capabilities of the network connectivity to support tasks such as, but not limited to model training, data processing, and remote access to computational resources. In an alternate embodiment, the pre-processed text file is accessed from the storing unit **240** to aid in training of the model at a local system, such as, but not limited to, the system **120**.

[0050] **FIG. 3** is an exemplary architecture **300** of the system **120** for pre-processing the documents, according to one or more embodiments of the present invention. It is to be further noted that the documents is interchangeably referred to as “the documents” and “the one or more documents”, without limiting and/or deviating from the scope of the present invention.

[0051] The exemplary embodiment as illustrated in the **FIG.3** includes an integrated system **305**, a Load Balancer **310**, an application server **315**, a pre-processing module **320**, a machine learning (ML) training module **325**, and a file system **330**.

[0052] The integrated system **305** of the architecture **300** is configured to input the one or more documents for pre-processing to the application server **315** via the load balancer **310**. In an embodiment, the one or more documents includes different types of documents, such as but not limited to, MOP documents, technical and functional specification documents, product release documents. In one embodiment, the MOP documents in the network **105** typically refer to detailed instructions that outline specific steps and processes for performing tasks such as but not limited to, deployment, configuration, maintenance, or troubleshooting of network elements or/and services. The integrated system **305** documents are crucial for ensuring consistency, efficiency, and safety in the operations and management of network **105**.

[0053] Upon receiving the input for pre-processing the documents from the integrated system **305**, the load balancer **310** distributes client requests among one or more of the application servers **315**. The load balancer **310** facilitates scalability, high availability, and efficient resource use. In one embodiment, the load balancer **310** directs document processing requests to the application server **315**.

[0054] In one embodiment, the application server **315** is configured to pre-process the one or more documents. The application server **315** executes techniques to convert the one or more documents into plain text files ensuring raw data is preserved throughout the conversion process. The application server **315** facilitates features, such as, but not limited to, increased processing capacity, fault tolerance, and load balancing capabilities. The application server **315** generates a clean and formatted text file from the one or more documents while stripping away unnecessary formatting. In this regard, in one embodiment, the application server **315** includes but not limited to, a pre-processing module **320**, a ML training module **325**, a file system **330**, to increase functionality and performance of the application server **315**.

[0055] The pre-processing module **320** of the application server **315**, is configured to execute the pre-processing steps on the one or more documents in the application server **315**. The pre-processing module **320** performs several operations such as but not limited to, removing elements, such as images, tables, graphs, page numbers, footnotes. Further, the pre-processing module **320** removes other structural components, which are not relevant for the ML training module **325** for the conversion of the one or more documents to text file format. In one embodiment, the pre-processing module **320** includes, but not limited to, the text extraction, normalization, language processing, cleaning hypertext markup language (HTML) tags, spell checking, and metadata removal segmentation.

[0056] In an exemplary embodiment, pre-processing additional text algorithms are integrated within the pre-processing module **320**. The pre-processing module **320** facilitates objective to the effectiveness and depth of pre-processing applied to the text file. In one embodiment, the pre-processing module **320** employs lowercase conversion for uniformity and eliminates irrelevant formatting elements such as but not limited to, sequence numbering, tabs, extra whitespace, headings, headers, footers, special characters, blank lines, and redundant lines.

[0057] Further, in one embodiment, the architecture **300** further includes the file system **330**. The file system **330** is designed to store the pre-processed the one or more

documents as received from the pre-processing module **320**. In one embodiment, the file system **330** holds essential extracted text content, optimized for ML training purposes. In one embodiment, the file system **330** comprising stored outputs can then be accessed by the ML training module **325** for further use.

[0058] In one embodiment, the ML training module **325** is utilized in different network **105** optimizations, predictive maintenance, traffic management, and security enhancement. The ML training module **325** is implemented within a clustered environment, specifically integrated into the application server **315** of the cluster.

[0059] **FIG 4** is a signal flow diagram for pre-processing of the one or more documents. For the purpose of description, the signal flow diagram is described with the embodiments as illustrated in **FIG. 2** and should nowhere be construed as limiting the scope of the present disclosure.

[0060] At step **405**, the retrieving unit **225** is configured to retrieve the one or more documents from the database **220** to be pre-processed. In one embodiment, the one or more documents include at least one of, but not limited to, the Method of Procedure (MOP) documents, the technical specifications documents, and the product release documents.

[0061] At step **410**, upon retrieving the one or more documents from the database **220** by the retrieving unit **225**, the retrieved one or more documents are transmitted to the converting unit **230** for conversion process. In the conversion process, the one or more documents are converted into the text file by using the converting unit **230**.

[0062] At step **415**, upon converting the one more documents into the text file by the converting unit **230**, the converted text file is transmitted to the pre-processing unit **235** to pre-process. The pre-processing unit **235** is configured to pre-process the text file to the format compatible for training the model. In an embodiment, the pre-processing unit **235** is configured to extract the text content from the text file. The extracted text content is normalized utilizing at least one normalizing technique. In an embodiment, the normalizing technique includes lower casing the extracted text

content and removing irrelevant formatting elements from the extracted text content. In an embodiment, the irrelevant formatting elements includes at least one of, but not limited to sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines.

[0063] At step **420**, upon pre-processing the text file to the format compatible for training the model by the pre-processing unit **235**, the storing unit **240** is configured to store the pre-processed text file in the database **220**.

[0064] At step **425**, upon storing the pre-processed text file in the database, the transmitting unit **245** is configured to transmit the pre-processed text file to the remote system **250** to train the model. In one embodiment, the remote systems **250** by providing faster, more reliable, and responsive connectivity. The remote system **250** provides numerous advantages which includes, but is not limited to, high speed, low latency, reliability, scalability, and enhanced security. In an embodiment, the remote system **250** benefits enable innovative applications across industries such as healthcare, transportation, manufacturing, and smart cities, transforming how remote operations and communications are managed and optimized. In an alternate embodiment, the pre-processed text file is accessed from the storing unit **240** to aid in training of the model at a local system, such as, but not limited to, the system **120**.

[0065] Referring to **FIG. 5**, **FIG. 5** illustrates a flow diagram of the method **500** for pre-processing the one or more documents, according to one or more embodiments of the present disclosure. The method **500** is adapted for pre-processing the one or more documents, **102**. For the purpose of description, the method **500** is described with the embodiments as illustrated in **FIG. 2** and should nowhere be construed as limiting the scope of the present disclosure.

[0066] At step **505**, the method **500** includes the step of retrieving from the database **220** one or more documents to be pre-processed. The one or more documents include at least one of, Method of Procedure (MOP) documents, technical specifications documents, and product releases documents. In an embodiment, the MOP refers to detailed steps and procedures for various planning, deploying, managing, and

maintaining the network **105**. In an embodiment, MOPs facilitate but not limited to, efficient implementation of technologies, optimizing performance, reliability, and security throughout the network **105**. The MOP includes but not limited to, site survey and planning MOP, base station installation MOP, network slicing MOP. service deployment MOP, security and compliance MOP.

[0067] At step **510**, the method **500** includes the step of the converting the one or more documents into a text file. The conversion process includes the step of eliminating by one or more non-relevant elements from the one or more documents including at least one of, images, tables, graphs, page numbers, and structural components.

[0068] At step **515**, the method **500** includes the step of the pre-processing the text file to the format compatible for training the model by the pre-processing unit **235**. In one embodiment, the extracted text content is normalized utilizing at least one normalizing technique. In an embodiment, the at least one normalizing technique includes lower casing the extracted text content and removing irrelevant formatting elements from the extracted text content. In an embodiment, the irrelevant formatting elements include at least one of sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines. In an embodiment, the pre-processed text file is stored in a file system of the database for model training.

[0069] The present invention discloses a non-transitory computer-readable medium having stored thereon computer-readable instructions. The computer-readable instructions are executed by the processor **205**. The processor **205** is configured to retrieve one or more documents to be pre-processed from the database **220**. The processor **205** is configured to convert the one or more documents into a text file. The processor **205** is configured to pre-process the converted text file to a format compatible for training a model.

[0070] A person of ordinary skill in the art will readily ascertain that the illustrated embodiments and steps in description and drawings (FIG.1-5) are set out to explain

the exemplary embodiments shown, and it should be anticipated that ongoing technological development will change the manner in which particular functions are performed. These examples are presented herein for purposes of illustration, and not limitation. Further, the boundaries of the functional building blocks have been arbitrarily defined herein for the convenience of the description. Alternative boundaries can be defined so long as the specified functions and relationships thereof are appropriately performed. Alternatives (including equivalents, extensions, variations, deviations, etc., of those described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. Such alternatives fall within the scope and spirit of the disclosed embodiments.

[0071] The present disclosure incorporates technical advancement for pre-processing the converted text file to the format compatible for training the model. By eliminating the one or more non-relevant elements from the one or more documents including at least one of, images, tables, graphs, page numbers, and structural components. The training model involves optimizing machine learning procedures repetitively to minimize prediction errors using labeled data. The process builds the model capable of making accurate classifications based on patterns identified in the training model.

[0072] The present invention provides various advantages including ML training module **325** for various the one or more documents, removal of irrelevant elements, enhanced tokenization, improved accuracy and efficiency, efficient information retrieval and knowledge extraction. The invention enables ML training module **325** of diverse the one or more document types, including MOP documents, technical specifications, and product releases. This broad applicability enhances the effectiveness and versatility of the system **120**. The pre-processing removes formatting tags, headers, footers, tables, and other structural components that can hinder the ML training module **325** by focusing solely on the text content the system **120** ensures accurate analysis and modeling.

[0073] In this regard, the pre-processing techniques optimize tokenization, enabling the AI model to understand the text structure at the granular level to improve the generation of accurate and contextually appropriate responses. Improved an accuracy to lowercasing, punctuation handling, and special considerations enhance the uniformity and quality of the pre-processed data. The system **120** efficiently extracts valuable insights, leading to more accurate ML training module **325** and improved system **120** effectiveness. The application server **315** facilitates efficient information retrieval and knowledge extraction by pre-processing the one or more documents. The application server **315** enables the system **120** to extract relevant information and generate contextually appropriate responses, enhancing overall system **120** performance.

[0074] The present invention offers multiple advantages over the prior art and the above listed are a few examples to emphasize on some of the advantageous features. The listed advantages are to be read in a non-limiting manner.

REFERENCE NUMERALS

- [0075] Communication system – 100
- [0076] Network – 105
- [0077] User Equipment – 110
- [0078] Server – 115
- [0079] System – 120
- [0080] Processor -205
- [0081] Memory – 210
- [0082] User Interface– 215
- [0083] Database- 220
- [0084] Retrieving unit – 225
- [0085] Converting unit – 230
- [0086] Pre-processing unit – 235
- [0087] Storing unit – 240
- [0088] Transmitting unit – 245
- [0089] Remote system - 250
- [0090] Integrated system – 305
- [0091] Load Balancer - 310
- [0092] Application server – 315
- [0093] Pre-processing module – 320
- [0094] Machine Learning training module – 325
- [0095] File system – 330

CLAIMS:**We Claim:**

1. A method (500) for pre-processing documents, the method comprising the steps of:
 - retrieving (505), by one or more processors (205), from a database (220), one or more documents to be pre-processed;
 - converting (510), by the one or more processors (205), the one or more documents into a text file; and
 - pre-processing (515), by the one or more processors (205), the text file to a format compatible for training a model.
2. The method (500) as claimed in claim 1, wherein the one or more documents include at least one of, Method of Procedure (MOP) documents, technical specifications documents, and product releases documents.
3. The method (500) as claimed in claim 1, wherein the step of, converting (510), the one or more documents into the text file, includes the step of:
 - eliminating, by the one or more processors (205), one or more non-relevant elements from the one or more documents including at least one of, images, tables, graphs, page numbers, and structural components.
4. The method (500) as claimed in claim 1, wherein the step of, pre-processing (515), the text file to a format compatible for training a model, includes the steps of:
 - extracting, by the one or more processors (205), text content from the text file; and

normalizing, by the one or more processors (205), the text content extracted utilizing at least one normalizing technique, wherein the at least one normalizing technique includes:

lower casing the extracted text content; and

removing irrelevant formatting elements from the extracted text content, wherein the irrelevant formatting elements includes at least one of, sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines.

5. The method 500 as claimed in claim 1, wherein the pre-processed text file is stored in a file system 330 of the database (220) for model training.
6. The method 500 as claimed in claim 1, wherein the method further comprises storing, by the one or more processors (205), the pre-processed text file in the database (220).
7. The method (500) as claimed in claim 1, wherein the method further comprises the step of:

transmitting, by the one or more processors (205), the pre-processed text file to a remote system (250) to the train the model therein.
8. A system (120) for pre-processing documents, the system (120) comprising:

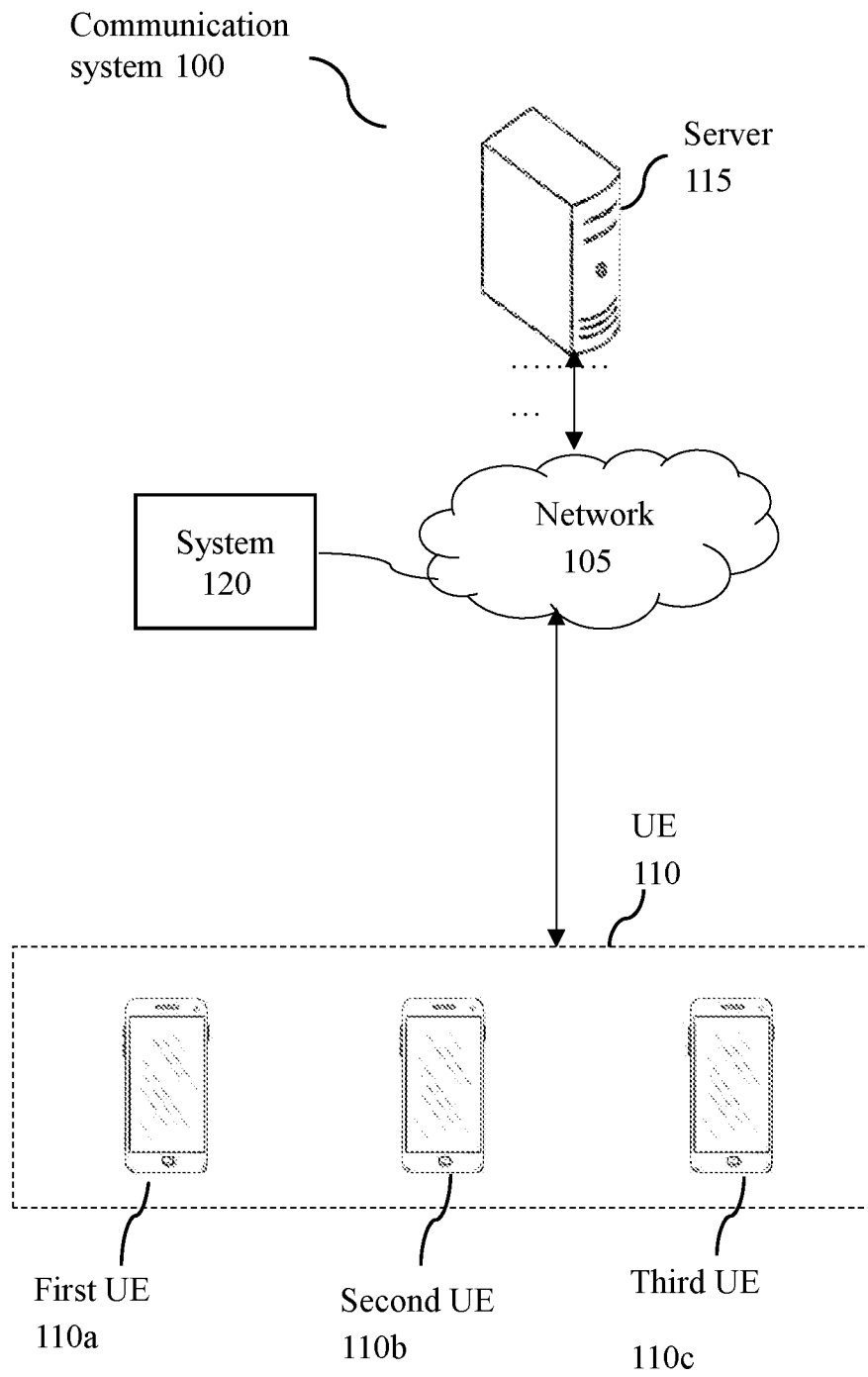
a retrieving unit (225), configured to, retrieve, from the database (220), one or more documents to be pre-processed;

a converting unit (230), configured to, convert, the one or more documents into a text file;

a pre-processing unit (235), configured to, pre-process, the text file to a format compatible for training a model.

9. The system (120) as claimed in claim 8, wherein the one or more documents include at least one of, Method of Procedure (MOP) documents, technical specifications documents, and product releases documents.
10. The system (120) as claimed in claim 8, wherein the converting unit (230), converts, the one or more documents into a text file, by:
 - eliminating, one or more non-relevant elements from the one or more documents including at least one of, images, tables, graphs, page numbers, and structural components.
11. The system (120) as claimed in claim 8, wherein the pre-processing unit (235), pre-processes, the text file to a format compatible for training a model, by:
 - extracting, text content from the text file;
 - normalizing, the text content extracted utilizing at least one normalizing technique, wherein the at least one normalizing technique includes:
 - lower casing the extracted text content; and
 - removing irrelevant formatting elements from the extracted text content, wherein the irrelevant formatting elements includes at least one of, sequence numbering, tabs, extra whitespaces, headings, tags, headers, footers, special characters, blank lines, and redundant lines.
12. The system (120) as claimed in claim 8, wherein the pre-processed text file is stored in the file system 330 of the database (220) for model training.
13. The system 120 as claimed in claim 8, wherein a storing unit (240) is configured to, store, the pre-processed text file in the database (220).

14. The system (120) as claimed in claim 8, wherein a transmitting unit (245) is configured to transmit, the pre-processed text file to the remote system (250) to train the model therein.
15. A non-transitory computer-readable medium having stored thereon computer-readable instructions that, when executed by the processor (205), causes the processor (205) to:
 - retrieve, from the database (220), one or more documents to be pre-processed;
 - convert, the one or more documents into a text file; and
 - pre-process, the text file to a format compatible for training a model.

**FIG. 1**

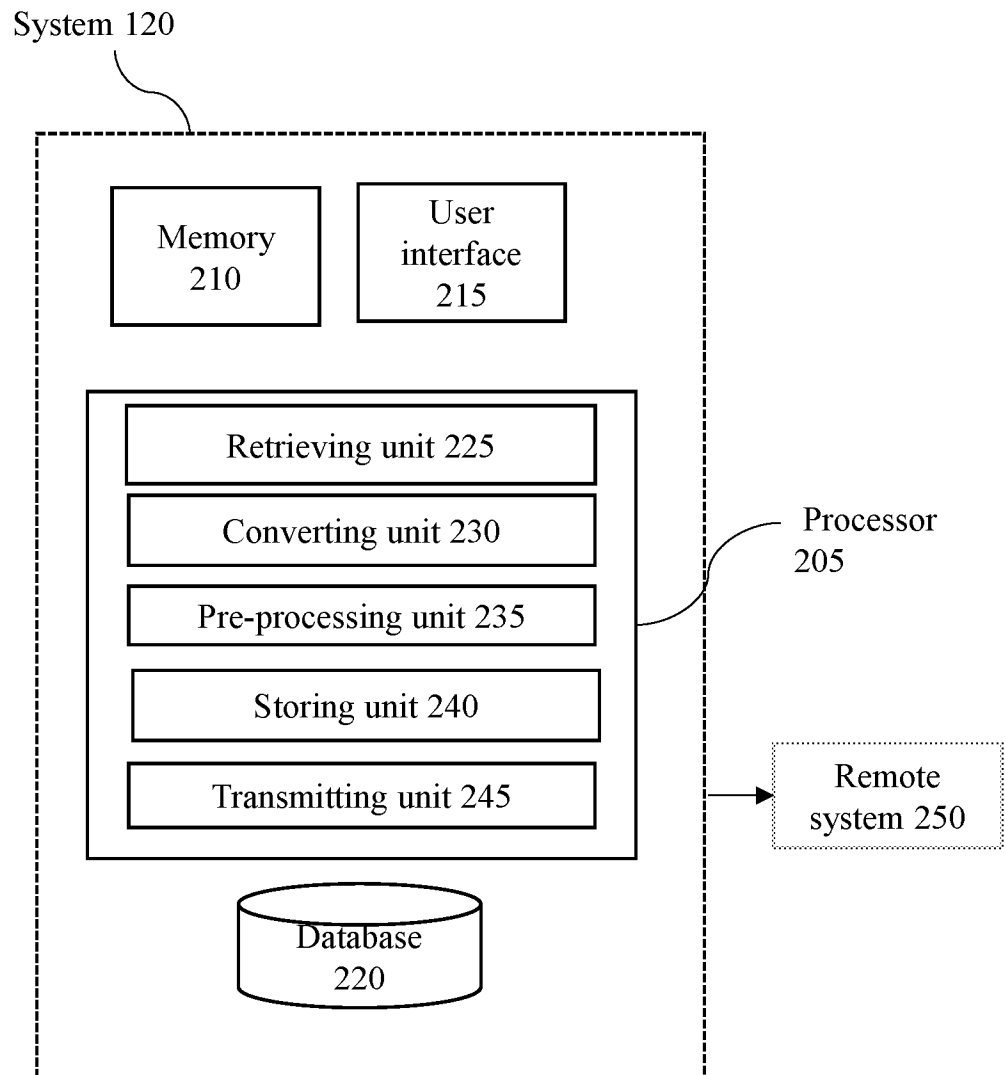


FIG. 2

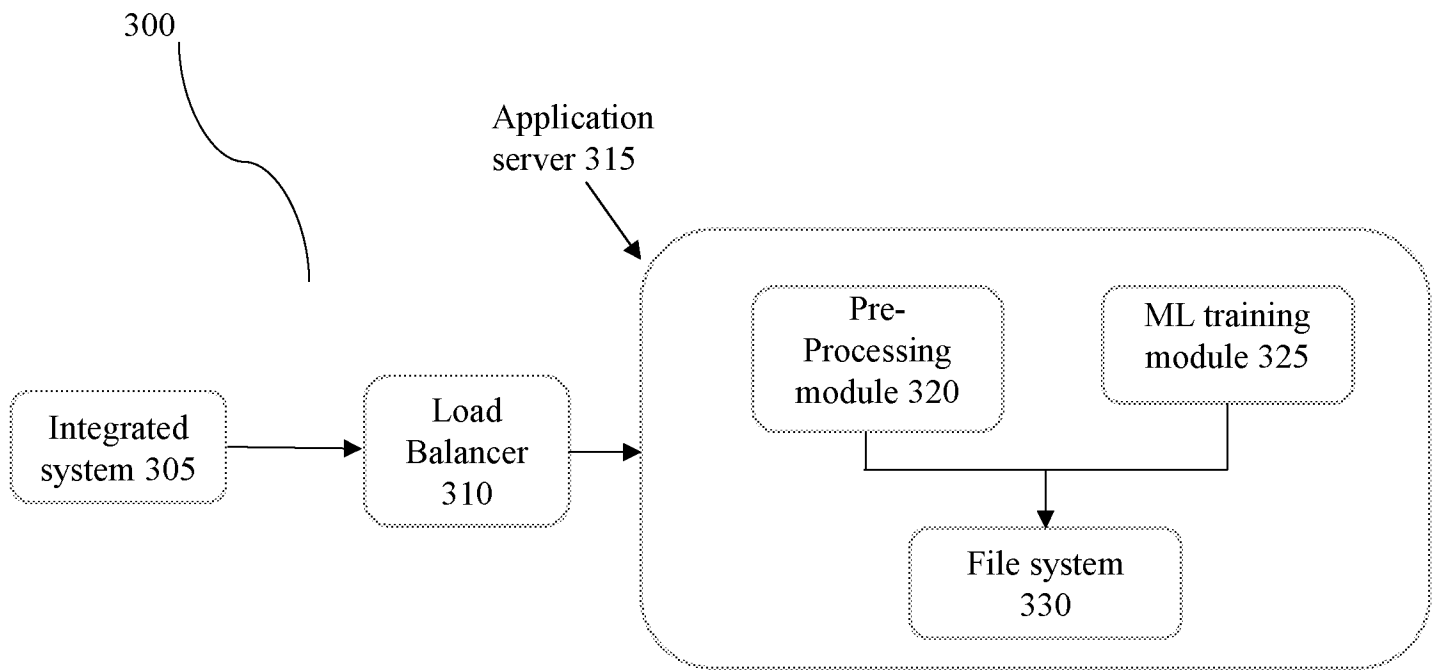


FIG. 3

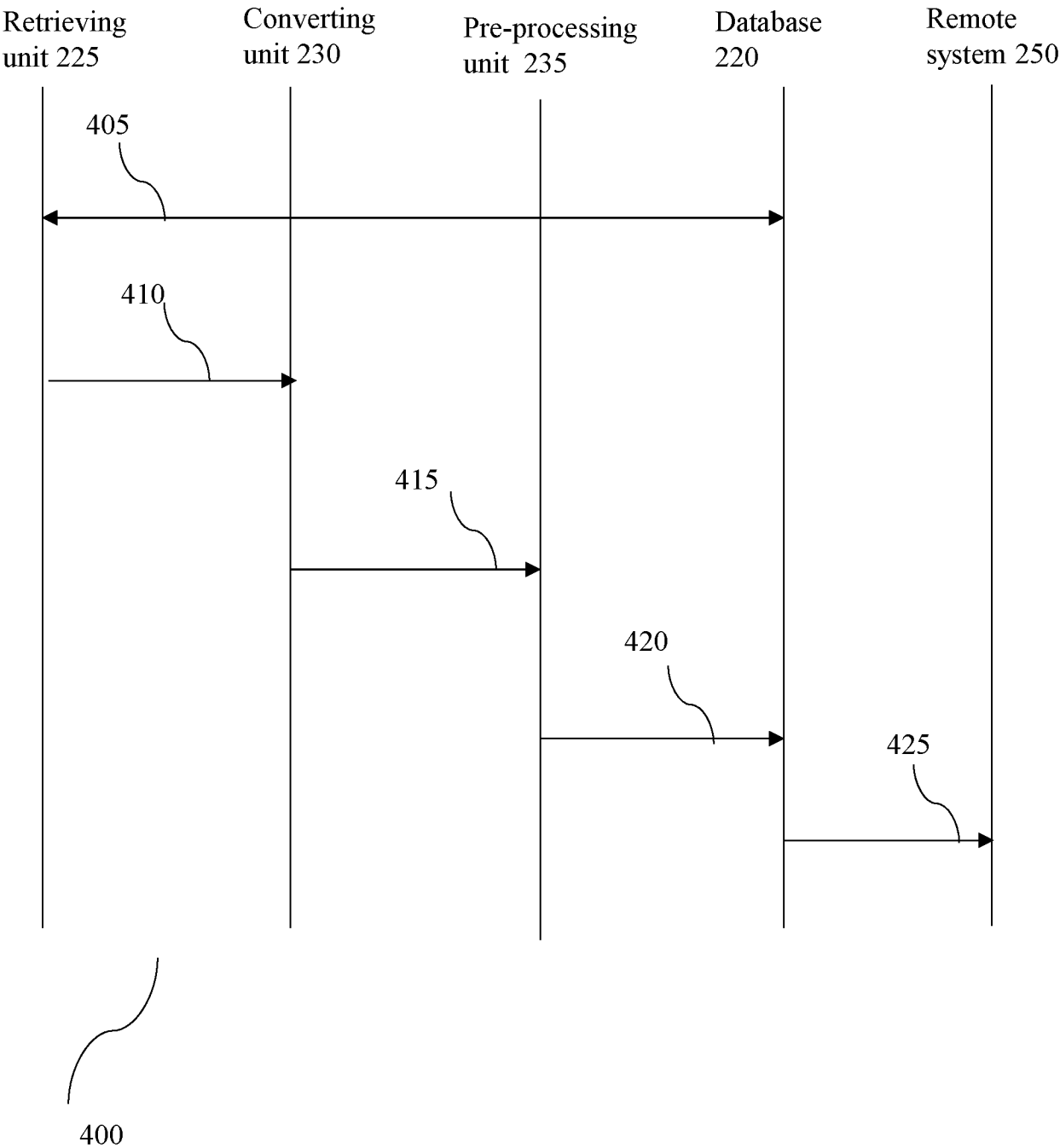
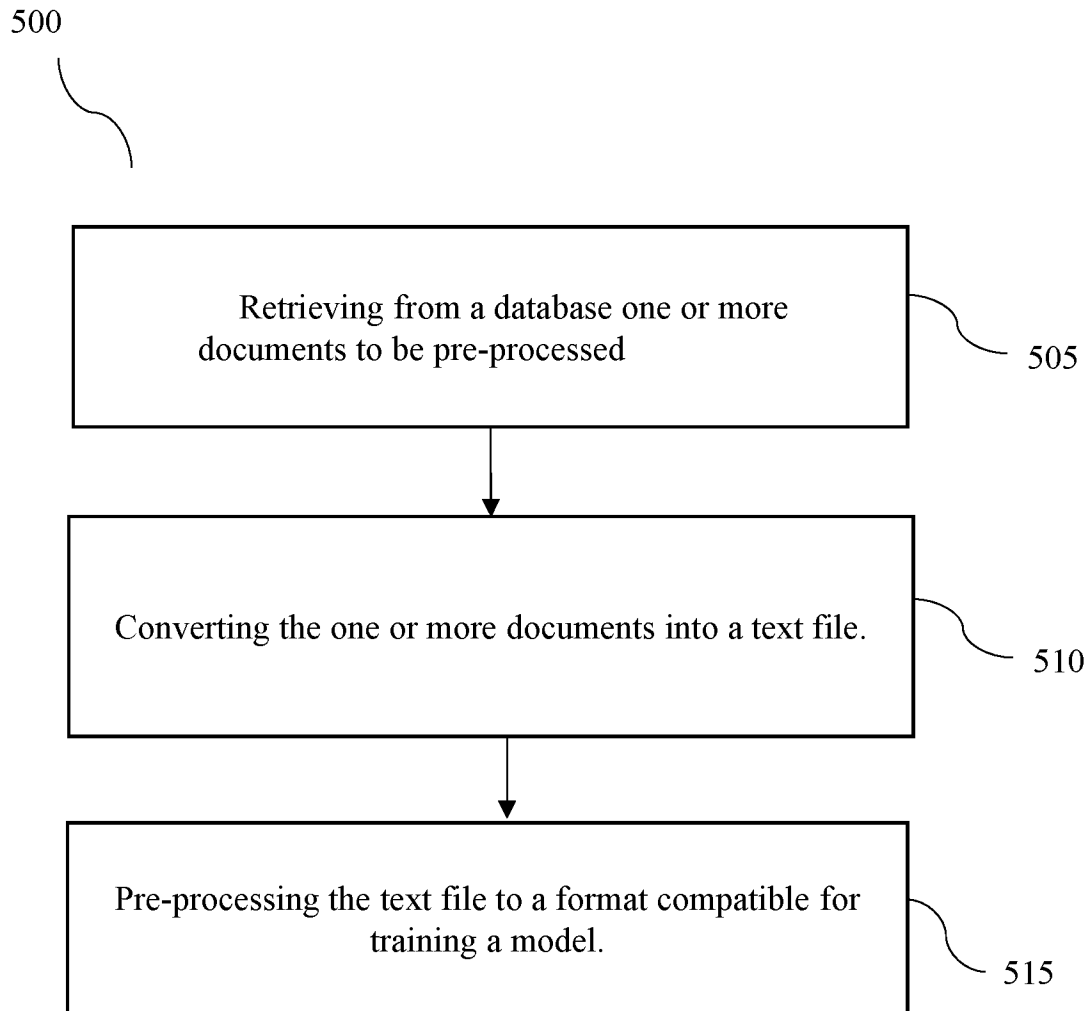


FIG. 4

**FIG. 5**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IN2024/051334

A. CLASSIFICATION OF SUBJECT MATTER

G06F16/35, G06F40/30, G06F40/289 Version=2024.01

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic database consulted during the international search (name of database and, where practicable, search terms used)

PatSeer, IPO Internal Database

Keywords: preprocess, extract, learning, model, database.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 115934942 A (HUADI COMPUTER GROUP CO LTD) 7th April, 2023 (07/04/2023) (English Translation Copy from Google Patents), Claims 1 and 6; Abstract.	1-15
Y	Kasper Welbers et al: ("Text analysis in R", 11 (4). pp. 245-265; Publication Date: 02/11/2017; DOI: https://doi.org/10.1080/19312458.2017.1387238), Page 2: 2 Data.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"I" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

05-12-2024

Date of mailing of the international search report

05-12-2024

Name and mailing address of the ISA/

Indian Patent Office

Plot No.32, Sector 14, Dwarka, New Delhi-110075

Facsimile No.

Authorized officer

Tamal Saha

Telephone No. +91-1125300200