How is unstructured data being used with the emerging technologies of Blockchain, Artificial Intelligence and Chatbots

Ba Viet Anh (Henry) Nguyen
Swinburne University of Technology

I. INTRODUCTION

In the era of big data, the volume of the data generated and collected by organizations has grown exponentially; as a result, we need to find other approaches to store, manage and analyze the data efficiently instead of traditional data processing. Traditional data processing methods have been proven inadequate in handling the complexity, volume, and variety of modern datasets. The application and development of big data will probably be limited if we stick to the traditional manner of data processing, which will definitely result in complex information sets, ununified data forms, and compromised data integrity (Li et al., 2023). In the light of these challenges with traditional data processing, the concept of unstructured data is gaining huge attention. The concept of unstructured data encompasses a wide range of information not readily categorized into predefined analytical categories. This includes written texts such as published and unpublished documents, personal diaries, field notes or transcripts of audio. Importantly, despite being called 'unstructured', these data are absolutely not unstructured; rather, they lack a predetermined organizational scheme and are instead shaped by the intentions and concerns of their creators (Boulton & Hammersley, 2006).

With the continuous development of technology, three emerging technologies which are Blockchain, Artificial Intelligence and Chatbots have shown particular promise in leveraging unstructured data.

This literature review aims to explore how unstructured data is being used with the emerging technologies of Blockchain, Artificial Intelligence and Chatbots. By mainly reviewing peer-reviewed research papers, the review will explain, analyze, critically evaluate and clarify the research in this field.

II. BLOCKCHAIN AND UNSTRUCTURED DATA

Blockchain technology, which is extremely famous for its decentralized and immutable ledger, is increasingly being utilized to handle unstructured data to expand to more applications other than its primary application (cryptocurrency application) that comes to everyone's mind first when talking about blockchain.

In the healthcare sector, unstructured data such as medical images, doctors' notes or lab reports are absolutely important for patient care. Blockchain enables secure and more efficient sharing of this sensitive information among parties. For example, a blockchain-based framework called MedShare is proposed to be able to solve the problem of sharing medical data between custodians of medical big data in an unreliable environment (Xia et al., 2017). By using blockchain, the system can probably enhance data security, integrity, and patient privacy to ensure that only authorized individuals or organizations like researchers or medical institutions are allowed to access sensitive health records. MeDShare leverages the blockchain's immutable ledger to record every transaction and data-sharing activity so that all actions taken on unstructured medical data are relatively transparent and traceable. The system uses smart contracts and access control mechanisms to monitor data usage

continuously.

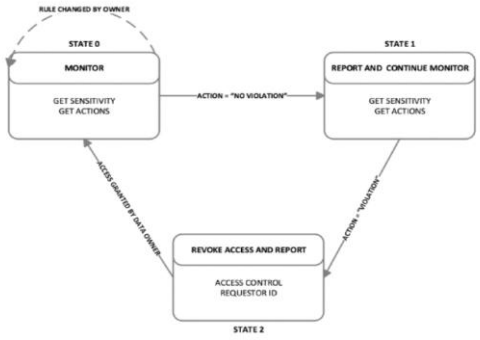

Figure 1: Xia et al., 'Smart Contracts as a Finite State Machine (FSM)', 2017, url: https://ieeexplore.ieee.org/document/7990130

For example, if a healthcare provider shares a patient's medical image, the smart contract will only permit the authorized parties. If any unauthorized access or malicious activity is detected such as attempts to alter or misuse the data, the smart control can revoke access to preserve data integrity and patient confidentiality. By integrating unstructured data with blockchain, MeDShare could enhance collaboration between hospitals, research institutions, and cloud service providers without compromising security. It allows for efficient data provenance and auditing and also organizations to track the lifecycle of medical data. This capability is essential for maintaining trust in the healthcare system and encourages the sharing of valuable medical information while safeguarding patient privacy. MedShare has demonstrated how blockchain can manage unstructured data in the healthcare sector to possibly improve healthcare outcomes. It also illustrates the potential for blockchain to revolutionize the way unstructured medical data is handled.

III. ARTIFICIAL INTELLIGENCE (AI) AND UNSTRUCTURED DATA

Nowadays, there are various types of data that do not fit the traditional databases anymore such as text, images, audio, and video. Artificial Intelligence has made significant strides by leveraging these unstructured data. This type of data comprises a large portion of information generated today, and this is also the reason that it leads AI technology to evolve to extract valuable insights from these data across various domains.

In the realm of Natural Language Processing (NLP), AI utilizes unstructured data from sources like social media, email, and articles to understand, learn and generate human language. One of the examples is the development of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019). BERT was pre-trained on vast amounts of unstructured text, including the entire Wikipedia (2,500M words) and BooksCorpus (800M words) to capture contextual relationships between words in a sentence (Devlin et al., 2019). Furthermore, BERT's bidirectional training approach allows it to understand the context of unstructured text more comprehensively.

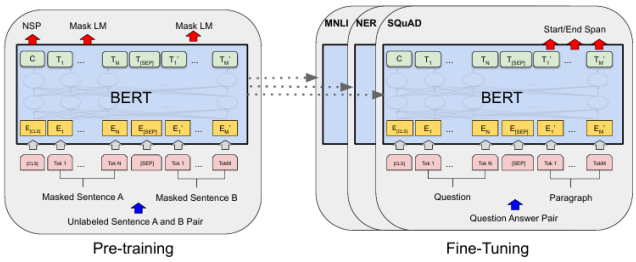

Figure 2: Delvin et al., 'Overall pre-training and fine-tuning procedures for BERT', 2019, url: https://aclanthology.org/N19-1423.pdf

BERT captures deeper linguistic patterns and relationships within the data by masking random tokens in the input and predicting them based on both left and right context (Devlin et al., 2019). This model helps model learning representations that are more contextually rich compared to traditional unidirectional language models. As a result, BERT has achieved state-of-the-art performance on several NLP tasks, including the GLUE benchmark and the SQuAD. According to the success of BERT, unstructured data (in this case: natural sentences, paragraphs, and documents from BooksCorpus and Wikipedia) has been utilized to improve the performance of AI. It demonstrates how advanced models can extract meaningful insights from complex and diverse language inputs.

In Computer Vision, AI processes unstructured image and video data to perform tasks such as object detection, image classification, and facial recognition. Krizhevsky et al. (2017) introduced ImageNet, a deep convolutional neural network (CNN) that significantly improved image

classification accuracy on the ImageNet dataset.

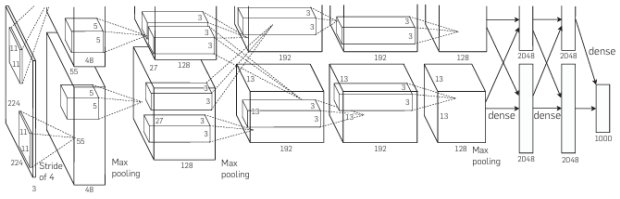

Figure 3: Krizhevsky et al., 'An illustration of the architecture of our CNN', 2017, url: https://dl.acm.org/doi/pdf/10.1145/3065386

By training on over a million high-resolution images (1.2 million), ImageNet demonstrated how deep learning could effectively handle unstructured visual data to breakthroughs in image recognition (37.5% top 1 and 17% top 5 error rates on ILSVRC-2010) (Krizhevsky et al. 2017).The key that led ImageNet to be success is that the network effectively extracted hierarchical features and information from images (unstructured data) without manual intervention by utilizing convolutional layers and techniques like ReLU activation and dropout regularization. Based on that, the model can recognize complex visual structures such as edges, textures, and shapes inherent in large, high-dimensional images (unstructured data) (Krizhevsky et al., 2017). While ImageNet sets new benchmarks, it also relies heavily on vast labelled datasets, requires considerable computational resources. This will limit its application in smaller-scale settings. Despite these challenges, ImageNet has paved the way for more advanced deep learning architectures and retained the ability to process unstructured data in computer vision tasks.

The application of AI to unstructured Audio Data has advanced speech recognition and understanding. Deep Speech is an end-to-end deep learning system for speech recognition. Trained on thousands of hours of unstructured audio data, Deep Speech has improved the accuracy of transcribing spoken language to text, even in noisy environments (Hannun et al., 2014). Deep Speech uses recurrent neural networks (RNNs) with Nesterov's Accelerated gradient methods for training to learn speech patterns from huge amounts of unstructured data to map raw audio directly to text transcriptions (Hannun et al., 2014).

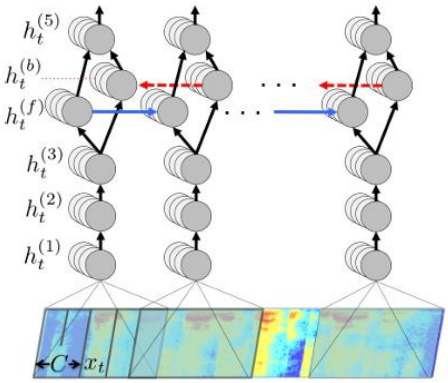

Figure 4: Hannun et al., 'Structure of RNN model and notation', 2014, url: https://arxiv.org/pdf/1412.5567

The nuances in speech, such as diverse accents, intonations, and background noises will be extracted by Deep Speech (Hannun et al., 2014). One important point that should be raised is that speech data is tightly regulated due to privacy laws but same as ImageNet, Deep Speech requires huge computational resources and data availability will limit to be scalable. However, it is assumed that learning directly from raw data (audio) without any predefined features – unstructured data, can help model in general and Deep Speech in this specific case perform better in the real-word speech recognition tasks.

However, the use of unstructured data raises some concerns. Ethical concerns arise regarding privacy and data security, especially when handling sensitive information such as personal communications or health records. Johnson et al. (2016) did raise the importance of ensuring privacy when using unstructured clinical data in AI applications. The issue of data privacy in AI is discussed a lot as the use of AI has increased significantly in recent years. Hence, in 2024, the European Union introduced Artificial Intelligence Act (AI Act) with a main focus on data privacy, which is also the first law aimed at regulating AI systems.

In summary, unstructured data is extremely essential in advancing AI technologies. By effectively processing and analyzing this type of data, AI systems can perform complex tasks across

language understanding, visual recognition, speech processing, etc. However, it is compulsory to adhere to all the rules and laws to prevent ethical issues.

IV. CHATBOTS AND UNSTRUCTURED DATA

Unstructured data has created many new challenges for traditional data processing systems because of its lack of predefined format (Shum et al., 2018). However, modern chatbots leverage unstructured data to learn, understand and generate more human-like conversations. The development of chatbots has been greatly accelerated by advances in the processing of unstructured data, especially natural language text.

Early chatbots, such as ELIZA, operated on simple pattern matching and scripted responses. While considered innovative at the time, these systems still struggle with the complexities and nuances of human language (Shum et al., 2018). The introduction of machine learning, and particularly deep learning techniques, enabled chatbots to learn from large volumes of unstructured text data from the real world and that has revolutionized chatbot capabilities.

A pivotal development was introduced by Vinyals and Le in 2015. They presented a neural conversational model using a sequence-to-sequence (Seq2Seq) framework with recurrent neural networks (RNNs).

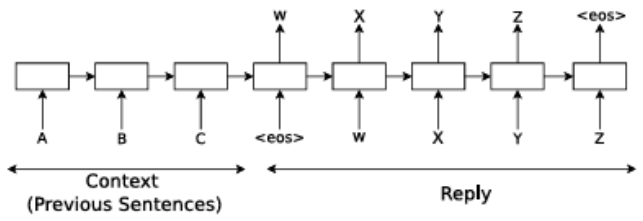

Figure 5: Vinyals & Le, 'Using the seq2seq framework for modeling conversations', 2015, url:https://arxiv.org/pdf/1506.05869

Their model was trained by large conversational datasets (unstructured data), which allows it to learn conversational patterns directly from unstructured dialogue data. Hence, it could generate appropriate responses by predicting sequences of words based on input text. This remarkable development demonstrated that chatbots have moved beyond rigid, rules-based interactions like ELIZA. The interactions look more natural thanks to this model, but this model still had limitations in maintaining coherence in longer conversations due to RNNs struggling with retaining context over time.

Building on this foundation, Serban et al. (2016) proposed a hierarchical neural network model (hierarchical recurrent encoder-decoder model (HRED)) for end-to-end dialogue systems. This model is like an upgraded version of Vinyals and Le's model when it can capture conversational context at both the utterance and dialogue levels to efficiently maintain coherence over multiple exchanges. This hierarchical structure can both manage the complexities of unstructured conversational data and allow more natural and contextually relevant interactions.

Chatbot Technology has truly broken through with the birth of transformer-based models like OpenAI's GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020). These models are trained on vast amounts of internet text data (unstructured data) to learn conversational patterns, language nuances and context in order to generate human-like dialogue. Thanks to unstructured data, these models can engage in a wide range of topics and adapt to various conversational contexts, unlike the earlier chatbot models that use rule-based systems or structured data.

Despite these advancements, there still remain some challenges in utilizing unstructured data for chatbots. One significant concern is the potential for models to generate biased or inappropriate content due to the presence of biases in training data (Sheng et al., 2019).

In summary, the use of unstructured data is fundamental for the evolution of chatbot technology. Chatbots have effectively improved in understanding and generating human-like conversations after leveraging large datasets of natural language.

V. CONCLUSION

Unstructured data has become a cornerstone in emerging technologies like Blockchain, Artificial Intelligence and Chatbots. By utilizing unstructured data such as text, images, audio, etc...., these

technologies have transcended previous limitations in data processing and analysis. Although Blockchain has not taken advantage of unstructured data much in upgrading its features, Blockchain is extremely important in storing and securing unstructured data, like in MeDShare. In contrast to Blockchain, Artificial Intelligence and Chatbots leverage vast amounts of unstructured data: for AI, it is to excel in natural language processing (NLP), computer vision, and speech recognition to develop models and applications; for Chatbots, it is to engage in more natural and contextual conversations. While significant progress has been made, challenges regarding data privacy, security, and ethical issues persist. However, with daily efforts, it is strongly believed that these limitations and challenges can be eliminated to utilize fertile unstructured data resources.

V. REFERENCE

1. Li, H., Guo, J., Yang, L., Wang, T. & QiaKai, H., 2023, 'Unstructured data processing strategy on account of blockchain technology', in Jansen, B.J., Zhou, Q. and Ye, J. (eds) Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022), Lecture Notes on Data Engineering and Communications Technologies, vol. 155, Springer, Singapore.
Link:https://link.springer.com/chapter/10.1007/978-981-19-9373-2_74

2. Boulton, D. & Hammersley, M., 2006, Analysis of unstructured data. Data collection and analysis, 2, pp.243-259.

Link:
https://books.google.com.au/books?hl=en&lr=&id=BEDTrvUH8NcC&oi=fnd&pg=PA243&dq=what+is+unstructured+data&ots=WzPEfWHCF9&sig=ZtNE__Ayyyjhw7I50sJZ-7It1g8#v=onepage&q=what%20is%20unstructured%20data&f=false

3. Xia, Q., Sifah, E. B., Asamoah, K. O., Gao, J., Du, X. & Guizani, M., 2017, 'MeDShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain', IEEE Access, vol.5, pp.14757-14767

Link: https://ieeexplore.ieee.org/document/7990130

4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., 2019, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', NAACL-HLT, vol.1, pp.4171-4186, ACL.
Link to the reference: https://aclanthology.org/N19-1423/

Link to the pdf: https://aclanthology.org/N19-1423.pdf

5. Krizhevsky, A., Sutskever, I., & Hinton, G.E., 2017, 'ImageNet Classification with Deep Convolutional Neural Networks', ACM, vol.6, pp. 84-90.
Link to reference:
https://dl.acm.org/doi/10.1145/3065386
Link to pdf:
https://dl.acm.org/doi/pdf/10.1145/3065386

6. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S. Sengupta, S., Coates, A. & Y.Ng, A., 2014, 'Deep Speech: Scaling up end-to-end speech recognition',arXiv:1412.55567, Cornell University.

Link to reference: https://arxiv.org/abs/1412.5567

Link to pdf: https://arxiv.org/pdf/1412.5567

7. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. & Mark, R.G, 2016,'MIMIC-III, a freely accessible critical care database', Scientific Data, 3, 160035.

Link to reference:
https://www.nature.com/articles/sdata201635#citeas

8. European Parliament 2019-2024, Artificial Intelligence Act, 2024, European Parliament.

Link:
https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

9. Shum, H., He, X-d., Li, D., 2018, From Eliza to Xiaolce: challenges and opportunities with social chatbots', Frontiers Inf Technol Electronic Engineering, vol.19, pp.10-26

Link:
https://link.springer.com/article/10.1631/FITEE.1700826#citeas

10. Vinyals, O. & Le, Q.V., 2015, 'A Neural Conversational Model', arXiv:1506.05869v3.

Link to pdf: https://arxiv.org/pdf/1506.05869

11. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A. & Pineau, J., 2016,'Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models', Proceedings of the AAAI Conference on Artificial Intelligence, vol.30, pp.3776-3783.

Link: https://aaai.org/papers/9883-building-end-to-end-dialogue-systems-using-generative-hierarchical-neural-network-models/

12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskerver, I., 2019, Language Models are Unsupervised Multitask Learners', OpenAI.

Link: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

13. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M, Wu, J., Winter, C., C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D., 2020, 'Language models are few-shot learners', In Proceeding of the 34th International Conference on Neural Information Processing Systems (NIPS '20) , Curran Associates Inc., NY, USA, article 159, pp.1877-1901.

Link to reference: https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Link to pdf: https://dl.acm.org/doi/pdf/10.5555/3495724.3495883

14. Sheng, E., Change, K-W., Natarajan, P., & Peng, N., 2019, 'The Woman Worked as a Babysitter: On Biases in Language Generation', pp. 3398-3403.

Link: https://www.researchgate.net/publication/336997596_The_Woman_Worked_as_a_Babysitter_On_Biases_in_Language_Generation