A Case Study for Compliance as Code with Graphs and Language Models: Public release of the Regulatory Knowledge Graph

A PREPRINT

Vladimir D. Ershov*
Clausematch Research
London, United Kingdom
vladimir.ershov@clausematch.com

January 27, 2023

ABSTRACT

The paper presents a study on using language models to automate the construction of executable Knowledge Graph (KG) for compliance. The paper focuses on Abu Dhabi Global Market regulations and taxonomy, involves manual tagging a portion of the regulations, training BERT-based models, which are then applied to the rest of the corpus. Coreference resolution and syntax analysis were used to parse the relationships between the tagged entities and to form KG stored in a Neo4j database. The paper states that the use of machine learning models released by regulators to automate the interpretation of rules is a vital step towards compliance automation, demonstrates the concept querying with Cypher, and states that the produced sub-graphs combined with Graph Neural Networks (GNN) will achieve expandability in judgment automation systems. The graph is open sourced on GitHub to provide structured data for future advancements in the field.

Keywords Natural Language Processing · Language Models · Knowledge Graph · Compliance management · Banking · Rule decomposition · Applied Machine Learning · Supervised learning

1 Introduction

Regulatory compliance is a crucial aspect of many industries, and ensuring that organizations abide by relevant regulations is vital for maintaining trust and preventing potential harm. The global financial crisis of 2008 and cases such as Enron and FTX have demonstrated the severe impact on society that can occur when compliance is not enforced. However, ensuring compliance is a complex and time-consuming process that requires a thorough understanding of relevant regulations and their application to specific situations. Given the cross-organizational nature of many businesses and the constant flow of domestic and international regulatory changes, it may not be possible for most of organizations to stay up-to-date on compliance without automation. To address the challenge, this study proposes an innovative approach to compliance automation through the use of pre-trained language models and knowledge graph technology. The paper details the steps involved in the project, including data collection, machine learning (ML) models development, KG construction, and experiment outcomes. Additionally, the study outlines future steps, such as training deep learning models for relation extraction and using GNNs for decision automation.

The study draws on previous research in NLP and GNNs to support the use of language models and graph-based algorithms in compliance decision making. For example, the effectiveness of KG and GNNs in modeling compliance risks and automating decisions based on the model is demonstrated in the paper "Towards knowledge graph reasoning for supply chain risk management using graph neural networks" [1]. The paper "Large-Scale Multi-Label Text Classification on EU Legislation" [2] also demonstrates that the use of pre-trained language model even in the legal domain can often be superior to other even custom-made architectures.

*https://www.linkedin.com/in/vladimir-ershov-33559466

2 Background and related work

Different fields of compliance have been investigated for more than 30 years and this has resulted in a vast variety of papers and surveys. These include Corporate compliance [3], Medical compliance [4], and Business process compliance [5] to name a few. A recent case study illustrates that compliance remains a significant challenge due to excessive, dynamic, and complex requirements that create "impenetrable spaghetti processes" [6].

One of the most well-researched set of works composed around an abstract formal framework for regulatory compliance ([7],[8],[9],[10],[11],[12],[13]) covered by Guido Governatori establish a conceptually sound formalisation of the norms and compliance rules to describe different deontic modalities: obligations, permissions etc. The proof-of-concept prototype presented by [14] shows that once the taxonomy of entities and relations forming regulated activities and rules is defined and the formal framework is built, it is possible to use First Order Logic over it. This makes it possible to develop and apply algorithms for compliance verification over formal models. However, the process of rules transition into a formal form is manual, as is the process of developing algorithms to verify a specific type of compliance. The current paper aims to address this challenge by introducing an approach to automate the construction of formalized expressions under specified taxonomy and suggests the means to automate the construction of compliance verification algorithms as well.

The challenge faced by state-of-the-art ML systems is their lack of explainability, where predictions are made without clear explanation [15]. This is a significant issue in the field of compliance, where explainability is a crucial focus. The Explainability-by-Design Methodology [16] offers a solution by manually augmenting rules in the decision-making system to provide comprehensive explanations. It involves building a system that organizes regulatory requirements in a graph-based RDF system. However, the research also notes that the cost of manually adding structure for explainability can take a month or more for a single use-case. Therefore, a suitable system for compliance automation at scale should be able to learn and produce explanations without the need for manually labelled data after training is complete. Many studies have been focused on this by automating the rule extraction process in the form of a KG and providing rule-guided explainable decisions based on induced rules [17]. This paper presents an approach to compliance automation by extracting structure from existing regulatory rules and allowing for the general formalisation of rules to emerge, which can be used as explainable input and output for decision automation. Unlike previous works, where formal frameworks were constructed first and existing rules were then converted to a predefined form, this approach does not rely on a manually constructed algorithm.

3 Method

```
If an 'ENT' with this 'PERM' was doing this 'ACT' or 'FS' with 'PROD'
      using 'TECH' then to avoid this 'RISK' they should 'MIT'.          (1)
```

The purpose of presented approach is to explore an applicability of recent advancements in NLP [18, 19, 20] for the problem of compliance automatisation. From subject matter experts point of view, a core feature of such systems is an ability to extract obligations which in the most general case come in the form scenario (1). Given the vital nature of taxonomy for forming the rules and a current form of regulation distribution through regulatory documents suggested course of actions was based on a *Gaia knowledge extraction system* [21] :

1. To define type of entities generalised from Taxonomy and used by regulators to form rules and obligations
2. To manually tag these entities in the current regulatory documents to produce a dataset of sufficient size
3. To apply state of the art approach for Named Entity Recognition (NER) task to develop models capable of tagging these entity types in the unstructured regulatory texts
4. To manually evaluate model performance to account for concept drift as during labelling expert may change the style of tagging
5. To assess an applicability of the approach and fine tune models if needed
6. To resolve co-references in the documents and apply tagging models to form the nodes of the graph
7. To extract relations between tagged entities with syntax analysis and combine it with document hierarchy to form edges of the graph
8. To explore the outcomes and draft the future steps

3.1 Dataset

As the most general and common scenario legal experts suggested to consider scenario (1). Given that and under the assumption that entities of these types are participate in forming most of the obligations the taxonomy of entities types listed below was formed. It is listed here in way as it was described to the labelling team.

- **Permissions [PERM]** - Any permission
- **Definitions [DEF]** - Any of the terms defined in the predefined glossary
- **Risks [RISK]** - Any mention of an identified risk, liability or issue. Specifically: if it's something or a situation that regulated entities need to avoid/be aware of, if they need to have something in place to stop something, if it's something that they need to comply with or compare themselves against
- **Mitigation [MIT]** - Any mention of rule, requirement, or guidance. Specifically: if it's a rule, if it's a link to another rule, an expectation, a requirement, a process to follow, an information to include or what should be considered
- **Entities [ENT]** - Any mention of a firm, financial institution or authorised person
- **Activities [ACT]** - Any concept that pertains to an entity initiating a FS related action/activity
- **A specific FS Concept [FS]** - Any mention of a financial concept or service such as liquidity, debt, moving money, custody, financing etc.
- **A financial services product [PROD]** - Any mention of a financial product or 'vehicle' holding money
- **Any mention of tech [TECH]** - Any mention of tech i.e Digital Assets, Robo Advisory, AI, Wallets, Encryption, Crypto, DLT, APIs, Software packages, Accounting packages, Cloud, Data etc.

After approximately 6 man-months of team efforts the dataset described in the Table 1 bellow was labelled. It was done over 1880 paragraphs in the Conduct of Business Rulebook (COBS) document.

Table 1: Manually labelled Taxonomy entities

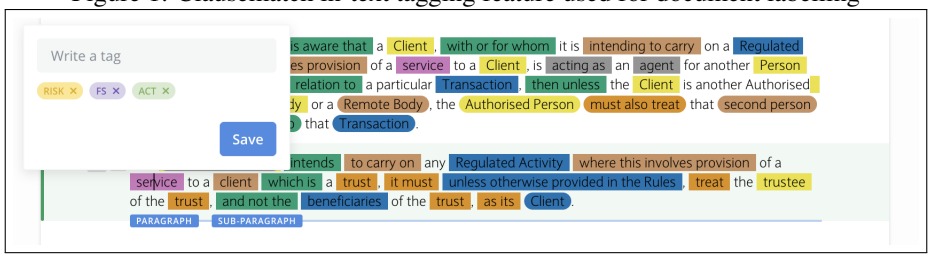
Concept	Tag	Labeled paragraphs requested	Subject matter understanding required	Entities labelled to train
Permissions	PERM	200	Easy	89
Definitions	DEF	200	Easy	2896
Risks	RISK	1000 to 10000	Hard	2170
Mitigation	MIT	1000 to 10000	Hard	3298
Entities	ENT	1000	Medium	2748
Activities	ACT	200	Easy	1444
A specific FS Concept	FS	1000	Medium	1404
A financial services product	PROD	1000	Medium	239
Any mention of technology	TECH	1000	Medium	257

3.2 Architecture

Here is the list of key components, models and libraries used to apply the advancements in NLP to the paper's task:

Labelling Initially labelling was started at Clausematch environment in the wiki style markup. Clausematch is a SAAS solution which is used for compliance and provides document storage and drafting functionality. It gives an advantage for expert to access content of the whole document and have a real-time collaboration in the browser. Disadvantage was that wiki markup requires human to type set of additional characters, which due to various inconsistencies and typos would damage up to 10% of paragraphs. Thus the in-text tagging feature shown in Figure 1 was introduced to eliminate that. *Doccano* [22] was used for a manual evaluation and labelling datasets for the fine-tuning. In comparison to Clausematch in-text tagging doccano's key advantages are an extensive set of hot keys to support the labelling work and UI with a real-time statistic on the tagging progress. Disadvantage is that each paragraph is detached from the document and an overall context of the clause is not available for the expert.

Figure 1: Clausematch in-text tagging feature used for document labelling



Applied machine learning tools For the development environment *Python 3.7.3* and *spaCy 2.2* [23] were used. The author used the *XLNet en_trf_xlnetbasecased_lg* [24] variation of pre-trained language model as a base and the NER pipe in the spaCy framework was updated during the training by applying the BILUO scheme. *Neuralcoref* [25] spaCy extension was used for coreference resolution. For the construction of graph edges presenting relations between entities in-text occurrences non-monotonic arc-eager transition-system [26] with a custom sentence segmentation [27] was used, which is available as a part of spaCy pipeline.

Infrastructure and experiment tracking AWS cloud was used as a GPU computation and storage resource provider. The experiment tracking was organised with *MLflow* [28] and S3 for the metadata storage. S3 was also used to store models, inputs and outputs for training and evaluation pipelines. These artifacts were indexed with a data version control (DVC)[2] tool.

Visualisation and graph exploration A key expectation for the applicability of new technologies is to provide an access for humans to the produced results which is particularly important in the field of governance and compliance. The Clausematch UI, as it shown in Figure 1, provides access to the entities tagged by the models. For a deep and comprehensive analysis of the graph structure, the author used *Neo4j* [29] as it supports *Cypher* [30] and the *Bloom* [31] tool. The next section will cover the results and present the visuals.

4 Results

The developed system automates the process of extracting KG from compliance documents. This includes automating the training of tagging models and running these models against text on the Clausematch platform. **The outcome of the automated data structuring in the form of a Neo4j dump we release publicly within these paper and supplementary visuals on the** *GitHub*[3]. The following sections will provide more details on the models used to produce the data and the dataset itself.

Table 2: Grouping and manual evaluation results

Tag	Dataset length	Precision	Recall	F1	Model
PERM	166	91.01	96.43	93.64	PERM
RISK	255	86.24	80.50	83.27	RISK_MIT
MIT	255	81.54	55.58	66.11	RISK_MIT
ENT	127	96.57	97.40	96.98	ENT
ACT	180	91.03	91.67	91.35	ACT_FS_PROD
FS	180	94.97	79.41	86.50	ACT_FS_PROD
PROD	180	90.32	93.33	91.80	ACT_FS_PROD

Models The process of training the models presents some challenges. It appears to not be possible to train a single XLNet-based model to capture all of the tags at once. As a result, the tags were organized into groups, and a separate model was trained and manually evaluated for each group. The results of the grouping and evaluation are presented in Table 2. Another challenge is consistency in tagging, as experts tend to reshape their perception of the meaning of taxonomy entities during the entity tagging process. For that reason, the evaluation in Table 2 was not done on the dev dataset, but instead, manual evaluation was used. The empirical perception is that up to one out of four entities may be

[2]https://dvc.org

[3]https://github.com/Vladimir-Ershov/adgm-kg1

tagged differently by the same expert after several hundred paragraphs. Since the *DEF* tag presents definitions listed in the glossary, a simplified version for tagging was used as a combination of extracted UPOS combined with lemma as a mask to match in the text.

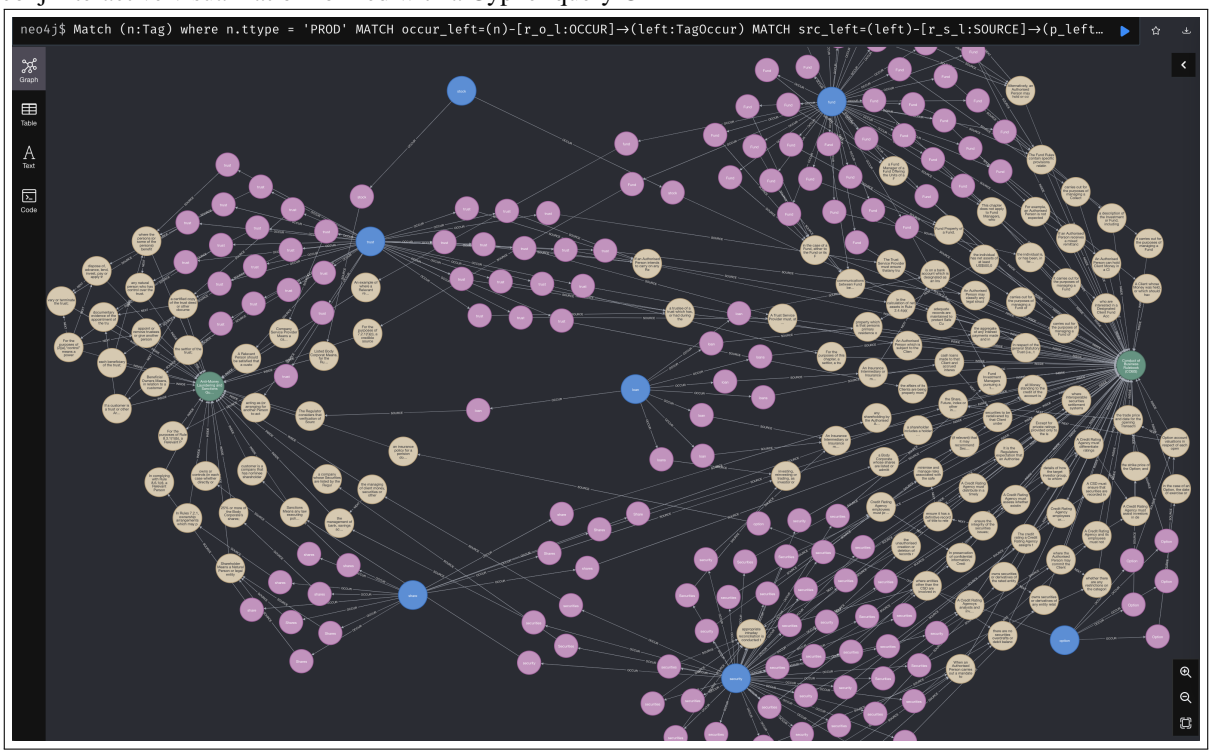
Tags The models described above were used to extract KG containing:

- Relationships: 1,209,207
- Nodes: 231,404
- "TagOccur" nodes: 173853 - which means occurrences of the tags in text
- "Tag" nodes: 35498 - for amount of entities divided into different concept. Detailed data listed in the A
- "Document" nodes: 26
- "Paragraph" nodes: 22027

Tags allow for a high-level overview of the statistics of discovered entities. Appendix B demonstrates one way to visualize it according to the detected Product in different documents. The corresponding bar plots and heatmap for the most popular tag co-occurrences can be found on *GitHub*[4].

Graph As parent and child relationships for the paragraphs in documents are part of the graph, the content can be explored using the interactive Neo4j tool, starting from the table of contents, using *Cypher* as demonstrated in Appendix D. After discussion with regulatory compliance experts, the following visual was included as **evidence of progress towards compliance automation: documents can now be analyzed on the concept level** using query C, as shown in Figure 2.

Figure 2: Two documents(green) intersected by entities identified as Product(blue) in paragraphs(tawny) by model. Neo4j interactive visualization formed with a Cypher query C



The Neo4j database allows querying the data through *REST* in the *JSON* format. This means that the regulations provided in the Regulatory Knowledge Graph, along with the results of extracted entities and relations, are available for integration into compliance automation systems as a source of ground truth. The graph structure allows for capturing

[4]https://github.com/Vladimir-Ershov/adgm-kg1

and propagating relations in a new dimension of concepts, making it possible for a downstream system to replicate an associative context expansion. As an example, Appendix F shows an application of graph algorithms to the extracted structure, capturing all the shortest paths of length 4 between entities related to *insurance* and *rule*. For visualization purposes, the Bloom tool was also used: Appendix E. Bloom allows for easy handling of graphs with 1000 or more nodes and better customization, such as conditional colour coding of the nodes.

Principal finding One of the key finding of this paper is a proposal that fine-tuned language models applied to the text for tagging and relation extraction produce interpretation in the form of layer outputs and therefore capture the meaning of a concept from the training data. A space of these embeddings introduce dimensions to automate the formalisation of statements juxtaposition from the text and makes it possible to introduce explainability in a form of a sub-graph and causal inference for an automated decision based on the KG. For that all nodes and relations in the graph derived from text have to be a result of unified interpretation. Therefore entities and relations between them have to be extracted according to the taxonomy implied by experts for compliance rule interpretation and learned by ML models. The consequence is that **the distribution of ML models containing interpretations of the rules within regulations by regulators is a vital step towards compliance automation** and will introduce a common ground where match between obligations and business internal processes is possible. Regulated companies will be able to apply released models to their internal Policies and Controls and produce KG as it would be seen by regulator. The internal KG and Regulatory KG could serve as input for GNN or other form of ML model to automate all range of compliance tasks: gap analysis, contradiction detection, compliance verification, etc. The interpretation of the rules introduced by the same ML models will allow the system to highlight factual differences between documents, rather than differences in interpretation.

Table 3: The most common relations between extracted entities as it is expected by the expert. Single cell reads as 'PERM'-[Allow]->'ACT', 'PERM'-[Authorise]->'ACT', 'PERM'-[Involving]->'ACT'

Tag	PERM	ACT	DEF	RISK	MIT	ENT	PROD	FS	TECH
PERM		Allow Authorise Involving	Involving Relating Uses	Create Increase Decreases	Must ensure Decreases	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses
ACT	Involving Relating Uses		Involving Relating Uses	Create Increase Decreases	Must ensure Decreases	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses
DEF	Involving Relating Uses	Involving Relating Uses		Create Increase Decreases	Must ensure Decreases	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses
RISK	Impact Create Increase Decreases	Impact Create Increase Decreases	Impact Create Increase Decreases		Must ensure Decreases	Impact Create Increase Decreases	Impact Create Increase Decreases	Impact Create Increase Decreases	Impact Create Increase Decreases
MIT	Must ensure Decreases	Must ensure Decreases	Must ensure Decreases	Create Increase Decreases		Must ensure Decreases	Must ensure Decreases	Must ensure Decreases	Must ensure Decreases
ENT	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Create Increase Decreases	Must ensure Decreases		Manage Controlled Owned Sell Buys	Manage Controlled Owned Sell Buys	Manage Controlled Owned Sell Buys
PROD	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Create Increase Decreases	Must ensure Decreases	Manage Controlled Owned Sell Buys		Manage Controlled Owned Sell Buys	Involving Relating Uses
FS	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Allow Authorise Cannot Involving	Create Increase Decreases	Must ensure Decreases	Involving Relating Uses	Involving Relating Uses		Involving Relating Uses
TECH	Involving Relating Uses	Involving Relating Uses	Involving Relating Uses	Create Increase Decreases	Must ensure Decreases	Manage Controlled Owned Sell Buys	Manage Controlled Owned Sell Buys	Manage Controlled Owned Sell Buys	

For the general formalization of rules to emerge, the relation extraction taxonomy must contain a unidirectional *"Is a"* relationship. This is likely to be necessary for a proper resolution of the NEL challenge. The other finding is the proposal on how to approach compliance verification. This challenge is known to have a NP-complexity [32]. The proposal is that KG has to store generalised form of the rules to enable reinforcement learning agents to navigate over sub-graphs of KG [33, 34] iterating mostly over general form of regulated rules. The key part of the solution here is the ability to make statements juxtaposition possible and to induce precedence order over implied rules. Therefore the relation taxonomy has to be extended with *"precedence"* relationship. Extracted relations between tagged entities at the current state of the graph were derived from text, but there were no classification applied to them. Table 3 presents

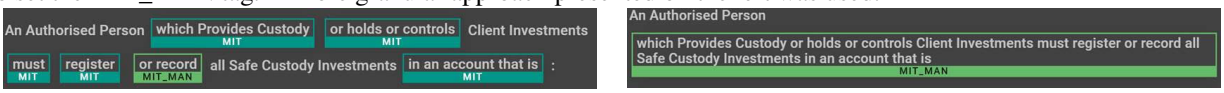
current view of the relation taxonomy formed in discussions with compliance experts as a view on connections needed to form and impose obligation. As previously stated, additional relations suggested to extend this list include: *"Is a"* and *"precedence"*.

5 Limitations and Future Work

RISK_MIT model This case study was focused on Abu Dhabi Global Market (ADGM) regulations and while the approach is expected to be applicable for other regulators, extracted taxonomy, KG extraction pipeline and models may require adjustments. Tagging models might be fine-tuned on a new taxonomy, but it likely to come with a cost of performance on ADGM taxonomy. `XLNet` baseline model might not be the most efficient to use as training on `RISK` and `MIT` tags didn't provide consistent result. The issue to highlight is that `RISK_MIT` model shows poor results: for some of the regulatory documents on a different topics the performance drops down to 43%. The reason for that is the same as with the initial model unified for all tags - the generalised concept is too vague and either model size is not capable to handle it or model weights are unlikely to converge. In order to overcome it `RISK` and `MIT` concepts have to be divided into set of sub-concepts per each and separated models should be trained. The other approach might be to train a large language models but that will likely to rise concerns regarding computational efficiency and carbon footprint [35].

Labelling and bias During tagging process experts have to decide on the interpretation for tagging boundaries as it shown at Figure 3, solving this may not eliminate the challenge and tagging might not be consistent even for a single person. The other challenge is to prepare the dataset diverse enough to reduce potential bias. The data set used for models training was compound from 1880 paragraphs of a single document. That is expected to introduce tagging artifacts which are listed in the Appendix G.

Figure 3: Two different ways to decide on the MIT tag boundaries. Subject matter expert during model evaluation have to set the MIT_MAN tag. A more granular approach presented on the left was used.



Relation extraction and GNN The future work for this study includes developing relation extraction models based on the taxonomy from Table 3. Transformer-based architectures, as presented in [36], are likely to be a reliable approach for this task. These models will be used to refine the knowledge graph and make it suitable for use as an input for a compliance automation system. GNNs [37] can be used as a baseline for such a system, but it is likely that Stepwise Reasoning Networks [38] will be more suitable for this task. The use of graph structures is motivated in addition by the NP-complexity of the regulatory compliance task [32], which is likely to limit the options for automation system design to reinforcement learning agents reasoning over sub-graphs of KG [33, 34]. This approach may result in "eventual compliance" rather than full compliance.

Graph refinement and NEL The other challenge to address is the overall refinement of the graph. Since the extraction of the entities was done on the paragraph level, coreference resolution was used to enrich the text of the paragraph based on the surrounding context. The tagged text in the normalised form of lemmas was used for NEL to generalise different occurrences of the tag under single `Tag` type in the graph. However, this unsupervised process introduced issues that could be addressed in future work.

6 Discussion and Conclusion

In conclusion, this study proposes an innovative approach to compliance automation by leveraging the capabilities of language modelling and KG technology. Through the application of ML models and manual annotation of a subset of regulations, the study demonstrates the feasibility of an executable KG that captures interpretation of regulatory taxonomy over regulations. The open-sourced KG serves as a valuable resource for future advancements in the field, and the author suggest the next step of developing automated relation extraction based on the proposed taxonomy. Additionally, the study proposes a call to action for regulators to release ML models to facilitate the capture and distribution of the meaning of regulatory taxonomy and relationships, thus enabling a more efficient and consistent interpretation of regulatory rules. Ultimately, the proposed approach aims to introduce causal inference and explainability in compliance decision-making systems through the use of the Knowledge Graphs.

7 Acknowledgments

The author of this study would like to extend his deepest gratitude to the following individuals and organizations, without whom the successful completion of this research would not have been possible.

- To Clausematch company, particularly Evgeny Likhoded and Anastasia Dokuchaeva, for their unwavering support throughout the project and commitment to solve compliance once and for all.
- To the ADGM, particularly Barry West, for providing the regulations, taxonomy, and subject matter expertise used in this study, as well as valuable feedback and insights.
- To colleagues at Clausematch, particularly Christoph Rahmede, who helped to refine the text and style of the paper.

We are honoured to have had the opportunity to work with such esteemed individuals and organizations. We are deeply grateful for their support and contributions, without which this research would not have been possible:

References

[1] Edward Elson Kosasih, Fabrizio Margaroli, Simone Gelli, Ajmal Aziz, Nick Wildgoose, and Alexandra Brintrup. Towards knowledge graph reasoning for supply chain risk management using graph neural networks. *International Journal of Production Research*, (1):1–17, 2022.

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, (2), 2019.

[3] Paul E McGreal. Corporate compliance survey. *The Business Lawyer*, pages 193–226, 2009.

[4] James A Trostle. Medical compliance as an ideology. *Social science & medicine*, 27(12):1299–1308, 1988.

[5] Michael Fellmann and Andrea Zasada. State-of-the-art of business process compliance approaches-a survey. In *EMISA Forum: Vol. 36, No. 2*. De Gruyter, 2016.

[6] Nigel Adams, Adriano Augusto, Michael Davern, and Marcello La Rosa. Why do banks find business process compliance so challenging? an australian case study. *arXiv preprint arXiv:2203.14904*, 2022.

[7] Guido Governatori, Zoran Milosevic, and Shazia Sadiq. Compliance checking between business processes and business contracts. In *2006 10th IEEE International Enterprise Distributed Object Computing Conference (EDOC'06)*, pages 221–232. IEEE, 2006.

[8] Guido Governatori and Antonino Rotolo. An algorithm for business process compliance. In *Legal Knowledge and Information Systems*, pages 186–191. IOS Press, 2008.

[9] Guido Governatori, Silvano Colombo Tosatto, and Pierre Kelsen. Towards an abstract framework for compliance. In *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops*, pages 79–88. IEEE, 2013.

[10] Guido Governatori. Business process compliance: An abstract normative framework. *IT-Information Technology*, 55(6):231–238, 2013.

[11] Guido Governatori, Mustafa Hashmi, and Moe Thandar Wynn. Modeling obligations with event-calculus. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, pages 296–310. Springer, 2014.

[12] Shazia Sadiq and Guido Governatori. Managing regulatory compliance in business processes. In *Handbook on business process management 2*, pages 265–288. Springer, 2015.

[13] Mustafa Hashmi, Guido Governatori, and Moe Thandar Wynn. Normative requirements for regulatory compliance: An abstract formal framework. *Information Systems Frontiers*, 18(3):429–455, 2016.

[14] Walid Fdhila, David Knuplesch, Stefanie Rinderle-Ma, and Manfred Reichert. Verifying compliance in process choreographies: Foundations, algorithms, and implementation. *Information Systems*, 108:101983, 2022.

[15] George Baryannis, Sahar Validi, Samir Dani, and Grigoris Antoniou. Supply chain risk management and artificial intelligence: state of the art and future research directions. *International Journal of Production Research*, 57(7): 2179–2202, 2019.

[16] Trung Dong Huynh, Niko Tsakalakis, Ayah Helal, Sophie Stalla-Bourdillon, and Luc Moreau. Explainability-by-design: A methodology to support explanations in decision-making systems. *arXiv preprint arXiv:2206.06251*, 2022.

[17] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. Jointly learning explainable rules for recommendation with knowledge graph. In *The world wide web conference*, pages 1210–1221, 2019.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 2017.

[20] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[21] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, 2020.

[22] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. URL `https://github.com/doccano/doccano`. Software available from https://github.com/doccano/doccano.

[23] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[24] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[25] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*, 2016.

[26] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1162. URL `https://aclanthology.org/D15-1162`.

[27] Joakim Nivre and Jens Nilsson. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi:10.3115/1219840.1219853. URL `https://aclanthology.org/P05-1013`.

[28] Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41(4):39–45, 2018.

[29] Ceyhun Ozgur, Jeffrey Coto, and David Booth. A comparative study of network modeling using a relational database (eg oracle, mysql, sql server) vs. neo4j. In *CONFERENCE PROCEEDINGS BY TRACK*, page 156, 2018.

[30] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1433–1445, 2018.

[31] Amy E Hodler and Mark Needham. Graph data science using neo4j. In *Massive Graph Analytics*, pages 433–457. Chapman and Hall/CRC.

[32] Silvano Colombo Tosatto, Guido Governatori, and Pierre Kelsen. Business process regulatory compliance is hard. *IEEE Transactions on Services Computing*, 8(6):958–970, 2014.

[33] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*, 2017.

[34] Zixuan Li, Xiaolong Jin, Saiping Guan, Yuanzhuo Wang, and Xueqi Cheng. Path reasoning over knowledge graph: a multi-agent and reinforcement learning based method. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 929–936. IEEE, 2018.

[35] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

[36] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*, 2019.

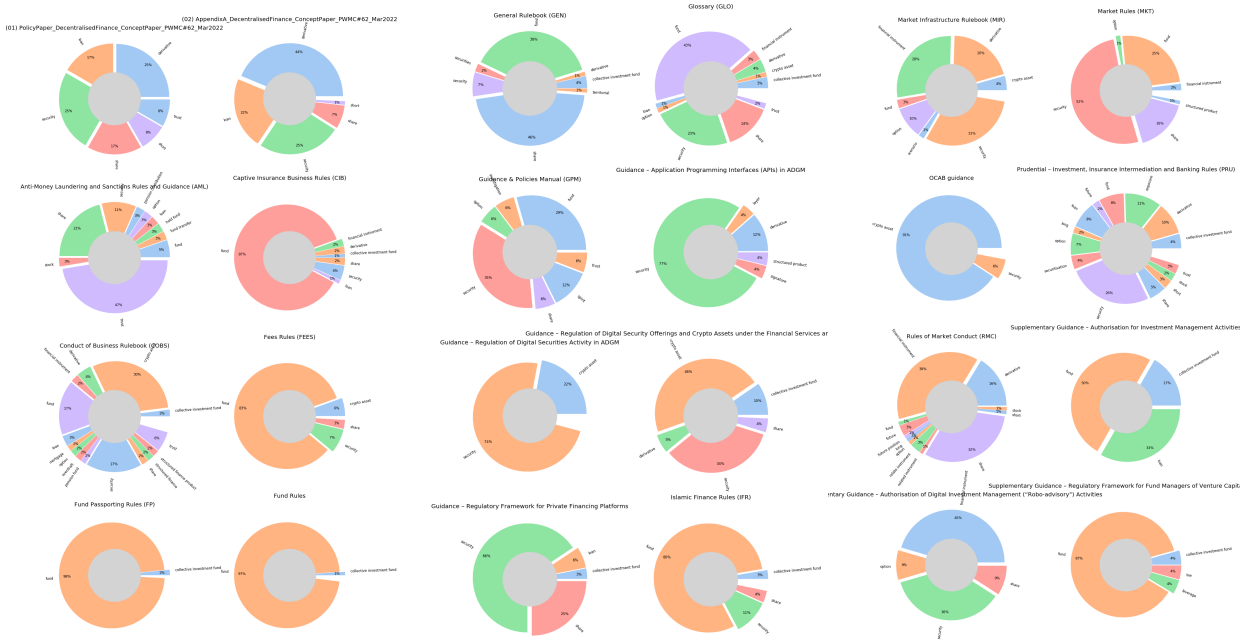
[37] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[38] Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and Kun Zhang. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. In *Proceedings of the 13th international conference on web search and data mining*, pages 474–482, 2020.

A Counts for different tags per concept

(Tag concept)	counts
MIT	20489
RISK	10737
TECH	1962
ACT	654
FS	583
ENT	526
PERM	272
DEF	202
PROD	73

B Proportions of the products mentions extracted from documents



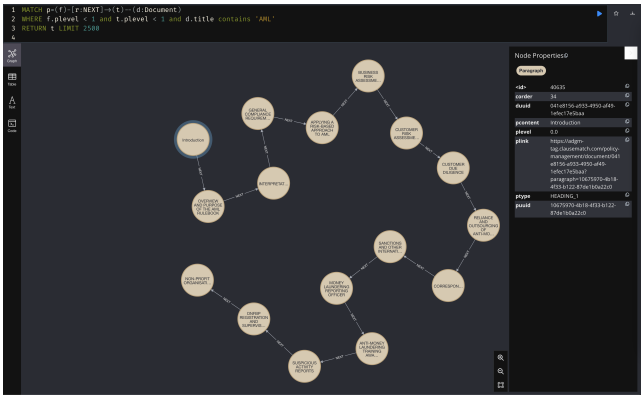
C Cypher query for intersecting two documents by entities identified as Product

```
Match (n:Tag)
where n.ttype = 'PROD'
MATCH occur_left=(n)-[r_o_l:OCCUR]->(left:TagOccur)
```

```
MATCH src_left=(left)-[r_s_l:SOURCE]->(p_left)--(d_left:Document)
where d_left.title contains '(COBS)'
MATCH occur_right=(n)-[r_o_r:OCCUR]->(right:TagOccur)
MATCH src_right=(right)-[r_s_r:SOURCE]->(p_right)--(d_right:Document)
WHERE d_right.title contains '(AML)'
RETURN occur_left, occur_right, src_right, src_left,d_left, d_right LIMIT 25000
```

D Getting table of content for the document in Neo4j. The interactive tool allows to expand each node to get internal content

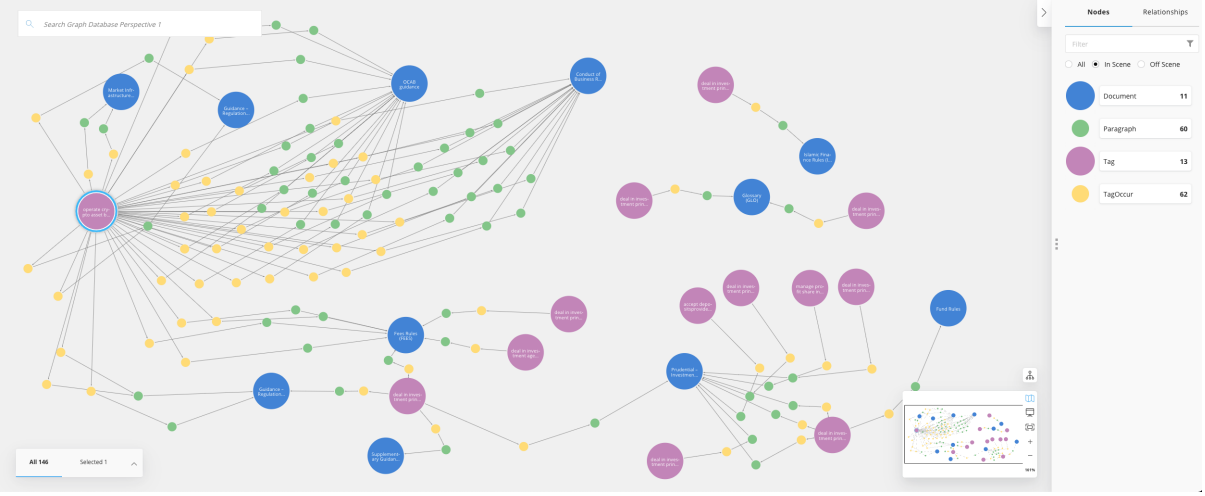
```
MATCH p=(f)-[r:NEXT]->(t)--(d:Document)
WHERE f.plevel < 1 and t.plevel < 1 and d.title contains 'AML'
```



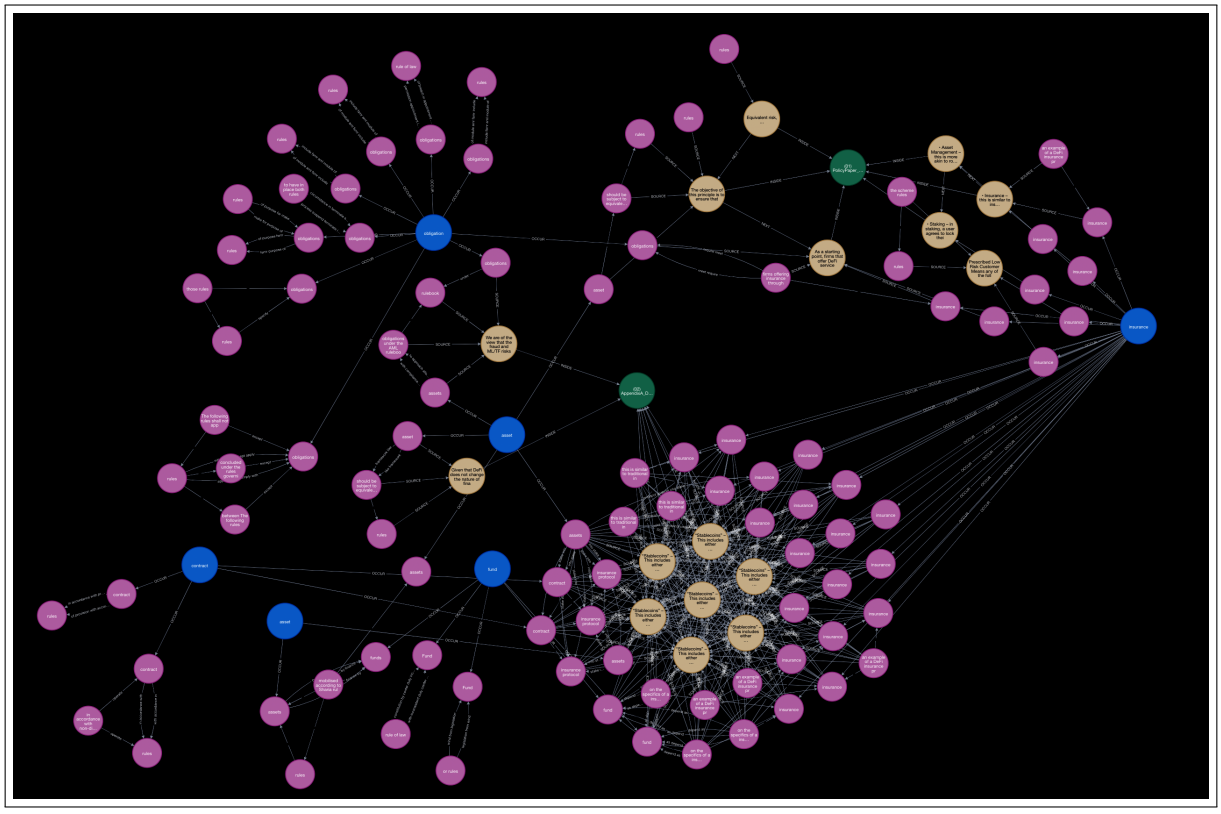
E Bloom visualization for the exploration of a usage of two permissions in the regulatory documents

```
MATCH (t:Tag) -[r1]- (to:TagOccur) -[r2]- (p:Paragraph) -[r3] - (d:Document)
WHERE t.ttype in ['PERM'] and t.lemma = 'operate crypto asset business'
RETURN t, to, p,d, r1, r2, r3
```

```
MATCH (t:Tag) -[r1]- (to:TagOccur) -[r2]- (p:Paragraph) -[r3] - (d:Document)
WHERE t.ttype in ['PERM'] and t.lemma contains 'invest' and
      t.lemma contains 'princ' and t.lemma contains 'deal'
RETURN t, to, p,d, r1, r2, r3
```



F Shortest path between insurance and rule related entities. Documents(green), concepts(blue), paragraphs(tawny). Neo4j interactive visualization formed with Cypher



```
MATCH (ent:TagOccur),(mit:TagOccur),
p = shortestPath((ent)-[*..4]-(mit))
WHERE ent.text contains 'insur' AND mit.text contains 'rule'
RETURN p LIMIT 250
```

G Snippet of tagging issues identified and cleaned manually from the graph

```
MATCH (t:Tag) -- (to:TagOccur)
where SIZE(t.lemma) <= 1
RETURN DISTINCT(to.text), count(*) as cnt
ORDER BY cnt DESC LIMIT 20
```

"(Tag text)"	counts
"unless"	75
"if the"	49
"If the"	43
"U.A.E."	41
"whether"	37
"doing"	37
"do not"	31
"whether the"	28
"U.A.E"	25
"c)"	25
"AT1"	24
"which"	21
"does not"	21
")"	18
"1"	16
"2"	16
"7"	16

```
MATCH (t:Tag) -- (to:TagOccur)
where SIZE(t.lemma) <= 0
DETACH DELETE to, t
# Deleted 1963 nodes, deleted 17115 relationships, completed after 253 ms.
```

H Tagging depends on the context and is not a direct word matching: ACT example

```
[51.58%]   =  163/ 316  times labeled for   Regulated Activity
[44.14%]   =  98/ 222  times labeled for   Shares
[55.45%]   =  61/ 110  times labeled for   Contracts of Insurance
[29.13%]   =  30/ 103  times labeled for   deposits
[54.35%]   =  25/ 46  times labeled for   Options
[63.16%]   =  24/ 38  times labeled for   Futures
[88.89%]   =  16/ 18  times labeled for   sukuk
[82.35%]   =  14/ 17  times labeled for   joint venture
[6.83%]   =  11/ 161  times labeled for   Undertaking
[90.0%]   =  9/ 10  times labeled for   Units in a Collective Investment Fund
[88.89%]   =  8/ 9  times labeled for   Service-based
[1.92%]   =  8/ 417  times labeled for   service
[100.0%]   =  8/ 8  times labeled for   credit agreement
[2.82%]   =  8/ 284  times labeled for   carrying on, in or from
[5.04%]   =  6/ 119  times labeled for   held by
[27.27%]   =  6/ 22  times labeled for   Managing Assets
[2.15%]   =  5/ 233  times labeled for   marketing
[1.0%]   =  5/ 501  times labeled for   activity
[4.76%]   =  4/ 84  times labeled for   provision of
[30.77%]   =  4/ 13  times labeled for   business activities
[16.67%]   =  3/ 18  times labeled for   holds or controls
[21.43%]   =  3/ 14  times labeled for   provision of a service to a Client
[2.14%]   =  3/ 140  times labeled for   controlled
[27.27%]   =  3/ 11  times labeled for   intends to carry on
```

```
[60.0%]   =  3/ 5  times labeled for   marketing activities
[2.68%]   =  3/ 112  times labeled for   can be conducted
[9.52%]   =  2/ 21  times labeled for   involves provision
[1.96%]   =  2/ 102  times labeled for   carrying on a Regulated Activity
[33.33%]   =  2/ 6  times labeled for   held or controlled
[100.0%]   =  2/ 2  times labeled for   develop or to undertake
[1.53%]   =  2/ 131  times labeled for   offered in
[0.49%]   =  2/ 412  times labeled for   is set up by
[11.76%]   =  2/ 17  times labeled for   provision of a service
[0.15%]   =  1/ 684  times labeled for   invest
[2.13%]   =  1/ 47  times labeled for   In the calculation
[1.18%]   =  1/ 85  times labeled for   received
[50.0%]   =  1/ 2  times labeled for   provided for the purposes
[0.26%]   =  1/ 384  times labeled for   Fund Manager
[100.0%]   =  1/ 1  times labeled for   Retail authorisation
[16.67%]   =  1/ 6  times labeled for   expertise
[0.08%]   =  1/ 1257  times labeled for   acting as
[11.11%]   =  1/ 9  times labeled for   carrying on an activity
[50.0%]   =  1/ 2  times labeled for   directly held
[25.0%]   =  1/ 4  times labeled for   must not carry on
[33.33%]   =  1/ 3  times labeled for   due and payable
[2.86%]   =  1/ 35  times labeled for   engages with
[33.33%]   =  1/ 3  times labeled for   advising or arranging
[50.0%]   =  1/ 2  times labeled for   contracts for differences
[11.11%]   =  1/ 9  times labeled for   carries on or intends to carry on
[100.0%]   =  1/ 1  times labeled for   made payable
[100.0%]   =  1/ 1  times labeled for   Large Undertaking
[100.0%]   =  1/ 1  times labeled for   giving and receiving of instructions
[3.85%]   =  1/ 26  times labeled for   demonstrate
[16.67%]   =  1/ 6  times labeled for   dedicated to
[33.33%]   =  1/ 3  times labeled for   held directly or indirectly
[100.0%]   =  1/ 1  times labeled for   promotional activities
[100.0%]   =  1/ 1  times labeled for   when it first carries on
[100.0%]   =  1/ 1  times labeled for   operated in accordance with the instructions
[4.55%]   =  1/ 22  times labeled for   contribute
[100.0%]   =  1/ 1  times labeled for   regard to its engagement
[1.85%]   =  1/ 54  times labeled for   participated
[100.0%]   =  1/ 1  times labeled for   Certificates representing certain Financial Instruments
[100.0%]   =  1/ 1  times labeled for   carried on, or held out as being carried on
[25.0%]   =  1/ 4  times labeled for   business purposes
[100.0%]   =  1/ 1  times labeled for   at the early stages of interaction
[0.47%]   =  1/ 211  times labeled for   arrangements
[100.0%]   =  1/ 1  times labeled for   inclined to act in accordance with the instructions
[100.0%]   =  1/ 1  times labeled for   Advisory and arranging
[33.33%]   =  1/ 3  times labeled for   communicating information
```