Legal NLP Meets MiCAR:
Advancing the Analysis of Crypto White Papers

Carolina Camassa
Bank of Italy
carolina.camassa@bancaditalia.it

Abstract

In the rapidly evolving field of crypto assets, white papers are essential documents for investor guidance, and are now subject to unprecedented content requirements under the European Union's Markets in Crypto-Assets Regulation (MiCAR). Natural Language Processing (NLP) can serve as a powerful tool for both analyzing these documents and assisting in regulatory compliance. This paper delivers two contributions to the topic. First, we survey existing applications of textual analysis to unregulated crypto asset white papers, uncovering a research gap that could be bridged with interdisciplinary collaboration. We then conduct an analysis of the changes introduced by MiCAR, highlighting the opportunities and challenges of integrating NLP within the new regulatory framework. The findings set the stage for further research, with the potential to benefit regulators, crypto asset issuers, and investors.

1 Introduction

White papers are the cornerstone of information disclosure for Initial Coin Offerings (ICOs), a modern fundraising method that sells tokens to a diverse group of investors (Fisch, 2019). These fundraisings represent an alternative method for entrepreneurial finance, with similarities to IPOs, venture capital, and presale crowdfunding (Howell et al., 2020).

When a new ICO is announced, it is usually accompanied by the *white paper*, a text document that contains information for the investors and bears some similarities, at least in intent, with an IPO prospectus. However, unlike their regulated counterparts, white papers have so far operated in a largely unregulated landscape.

Thanks to this regulatory vacuum, the type of content varies widely, but it usually includes (Bourveau et al., 2022; Zhang et al., 2019) a description of the service offered and its value proposition, the composition of the project team, financial details

about the offer, and explanations of the technologies underlying the project. Due to their public accessibility and richness in data, white papers have been a frequent subject of studies aimed at extracting market insights through textual analysis (Fisch (2019); Thewissen et al. (2022); Florysiak and Schandlbauer (2022) among others).

The current ICO market is characterized by a pronounced level of *information asymmetry* between crypto issuers and investors (Bourveau et al., 2022; Chen, 2019). These circumstances, easily explained by the lack of regulation, led to investigation into the prevalence of fraud and scams among ICOs (Liebau and Schueffel, 2019; Karimov and Wójcik, 2021). It is not uncommon for white papers to be a vehicle for misinformation, sometimes resorting to imprecise claims and exaggerated language (Momtaz, 2021) to lure investors.

1.1 Introduction to the MiCA Regulation

In this context of fragmented or absent regulation of the crypto assets market, the European Union has been making an effort to unify regulation in its member states on this aspect. The vehicle for this unification is the *Regulation on Markets in Crypto Assets* (abbreviated as MiCAR). The text of the regulation was published on the Official Journal of the European Union on June 9th, 2023 (150, 2023) [1], and the norms will become fully applicable by December 30th, 2024. Its entry into force will provide clarity and protection for European citizens, but it will also require considerable effort on the part of issuers (Florysiak, 2022).

The MiCAR aims to regulate crypto asset public offerings—commonly known as ICOs—the admission of those crypto assets to trading on platforms, and additional related services to customers.

One key focus is the standardization of white paper content; the regulation mandates the inclusion of specific elements, such as details about the token of-

---

[1] OJ L 150, 9.6.2023, p. 40–205

fering, the associated risks, and team composition, with an emphasis on textual clarity and readability.

These new rules are engineered not just to to enable European citizens to make informed investment decisions, but also to assist national authorities in monitoring crypto asset offerings.

1.2 Motivation and main contributions

The goal of this study is to overview current Natural Language Processing applications to the analysis of crypto asset white papers, and identify how new applications can assist stakeholders in adapting to the upcoming MiCA regulation.

Given that much of the existing research comes from economics and finance, this work aims to bridge the domain gap, enabling future Computer Science (CS) researchers to leverage all relevant work on the topic. It also provides a starting point for researchers and practitioners in the legal NLP domain to adapt existing methods to this new topic. The organization of this work is as follows:

- A survey of NLP applications to crypto asset white papers and Initial Public Offering (IPO) prospectuses is provided in Section 2.

- Section 3 goes into detail on how the new MiCA regulation will impact the structure and content of white papers.

- Finally, Section 4 highlights the challenges

and opportunities of harnessing NLP techniques to aid all stakeholders involved in the MiCAR compliance process.

2 Related work: NLP in ICO White Papers and IPO Prospectuses

2.1 NLP and ICO white papers

A survey of the field shows that most existing studies on crypto asset white papers are published in economics and finance venues, with little overlap with CS research.

A recurring research goal is using information extracted from the content of white papers to make predictions about the likelihood of success of an ICO. In this context, the type of signal extracted can vary widely, as do the predictive models used. Some studies on the topic relies exclusively on manual text analysis (Bourveau et al., 2022; Thewissen et al., 2023; Fisch, 2019); this review focuses instead on those studies that incorporate at least one NLP method, setting dictionary-based approaches as the minimum threshold for inclusion[2].

Figure 1 shows a taxonomy of the studies grouped according to the variables of interest and the type

---

[2]Excluded work typically relies on page length (Samieifar and Baur, 2021; Bourveau et al., 2018) or readability metrics such as the Gunning-Fog index or the Flesch Reading Ease score (Zhang et al., 2019).
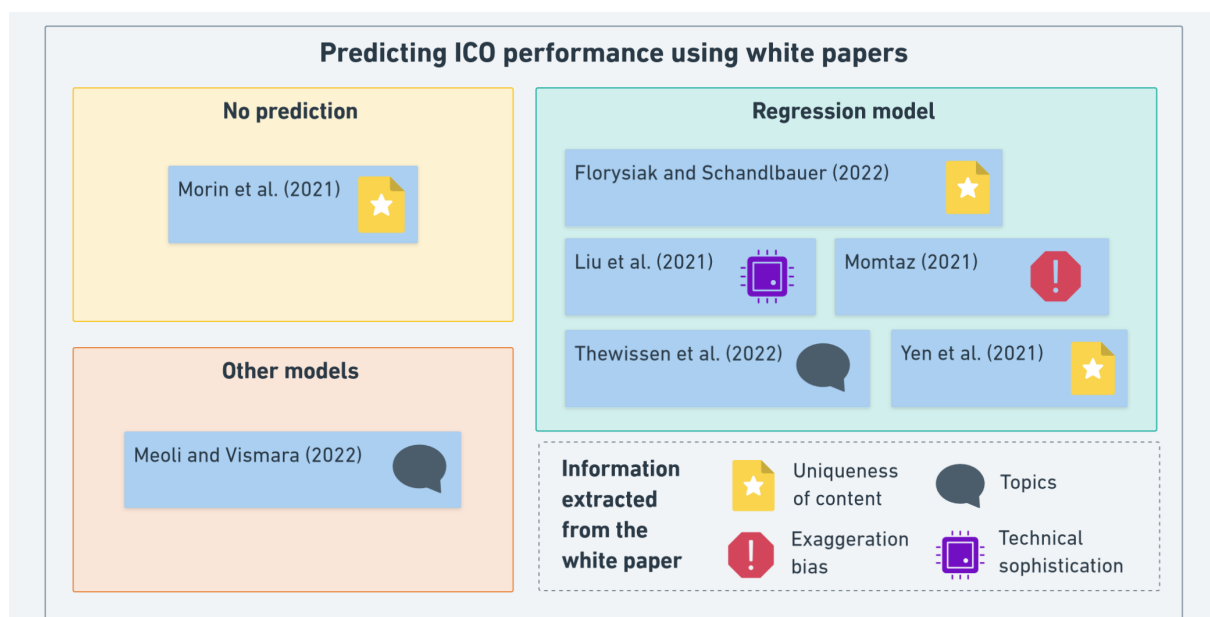

Figure 1: A shared research objective in existing studies that apply NLP to crypto asset white papers is predicting the degree of success of an ICO. The figure shows the experiments surveyed in section 2.1, grouped according to the type of prediction model used and the variables extracted from the text.

Study	Dataset size	Sentiment analysis	Topic modeling	Document similarity	NER
Florysiak and Schandlbauer (2022)	2,665			✓	✓
Liu et al. (2021)	1,629	✓	✓		
Meoli and Vismara (2022)	451	✓	✓		✓
Momtaz (2021)	495	✓			
Morin et al. (2021)	3,299			✓	
Thewissen et al. (2022)	5,210	✓	✓		✓
Yen et al. (2021)	1,064			✓	

Table 1: Overview of existing literature applying NLP techniques to crypto asset white papers, along with the dataset size and the techniques used. Most works come from the economics and finance domains.

of model, and Table 1 lists the dataset size and NLP tasks for each.

Momtaz (2021) uses a dictionary-based feature extraction method first proposed by Hu and Liu (2004) to capture the *exaggeration bias* in white papers. Momtaz argues that, in the absence of regulation, issuers of crypto asset tokens may have an incentive to positively exaggerate the quality of their venture to attract investors. The analysis confirms that there is a pervasive positive bias in the content of white papers, and that ICOs with this characteristic raise more funds in less time.

A common strategy is to use variations of the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) to extract topic information from the white papers. The output of the model can be used, for example, to quantify the amount of technical information included in the text (Liu et al., 2021). This variable shows that crypto assets with high *technology indexes* are more likely to have positive long-run performances of the associated ICOs. Thewissen et al. (2022) use SENTLDA (Bao and Datta, 2014), a sentence-level version of LDA, to extract 30 topics from an extensive collection of white papers. The topics generated by the algorithm are evaluated and manually labeled by the authors, and the topic assignment is then used as a regression variable. The results of the regression show that the topics contained in a white paper can partially explain the success of ICOs.

Given the large amount of new ICOs published each year, some research has focused on looking for similarities among published white papers. Using a term-frequency approach initially developed for IPO prospectuses by Hanley and Hoberg (2010),

Florysiak and Schandlbauer (2022) and Yen et al. (2021) assess the informational value of these documents. Here, a white paper is considered more informative if it introduces new or additional information not found in similar papers. The outcomes indicate a correlation between high informational content, or uniqueness, and success metrics such as fundraising and post-ICO market values. Separately, Morin et al. (2021) investigate text similarity across white papers using three metrics: TF-IDF, cosine similarity, and pairwise similarity. Their findings reveal that 19% of ICO white papers exhibit high similarity to previously published ones. Finally, Meoli and Vismara (2022) combine sentiment analysis, NER, topic modeling, sentiment analysis and document structure analysis to extract useful features from the text. The features are then used to train a set of binary classifiers, with different machine learning approaches, to predict whether an ICO was successful or not. The best machine learning model outperforms the benchmark, a traditional econometric forecasting model. A feature importance analysis confirms the relevance of the information extracted from the text, especially the document structure and the sentiment score.

2.2 NLP and IPO prospectuses

Textual analysis methods have also been used to extract information from the contents of IPO prospectuses. A prospectus is a document required by national financial regulatory authorities to present an investment offering—such as stocks, bonds, and mutual funds—to the public, or to obtain admission to trading on a regulated market. It contains all the details of the financial offer and must inform in-

Part	Description
A	Information about the offeror or the person seeking admission to trading
B	Information about the issuer, if different from the offeror or person seeking admission to trading
C	Information about the operator of the trading platform
D	Information about the crypto asset project
E	Information about the offer to the public of crypto assets or their admission to trading
F	Information about the crypto assets
G	Information on the rights and obligations attached to the crypto assets
H	Information on the underlying technology
I	Information on the risks

Table 2: Content requirement categories for crypto assets "other than"—one of the three types of assets regulated by MiCAR.

vestors of the risks involved with investing. These features make the documents partially similar, at least in intent, to crypto asset white papers.

Tao et al. (2018) train a custom embedding model using the WORD2VEC algorithm (Mikolov et al., 2013), to find and analyze sentences in the text that represent "forward looking statements" (FLS). These statements are meant to provide useful information about the company's future performance. Li et al. (2018) apply a keyword-based approach to the same task.

Sharpe and Decker (2022) perform sentiment analysis to look for the effect of text sentiment on the probability of an IPO being withdrawn.

Finally, Hanley and Hoberg (2010) devise a method—previously mentioned in 2.1—to separate the information content of a prospectus into *standard* and *informative* components. The components are derived from term frequency data and they are used to quantify how much of the content is also found in documents published in the same time period—or in the same industry—and how much of it is unique to a specific document.

2.3 Conclusions from the survey

While there seems to be a healthy amount of research interest in the topic, most of the existing literature covers a very limited array of NLP techniques and tasks. As an hypothesis, this might be a result of the lack of intersection between this particular domain of economics and finance research and CS research.

More advanced NLP methodologies have successfully been applied in various regulatory and legal settings. The following section contains examples of such works and argues for the transferability of these computational approaches to the analysis and regulation of future crypto asset white papers.

3 How MiCA changes the regulatory landscape

The MiCA regulation is broad and covers three categories of crypto assets: *asset-referenced tokens*, *e-money tokens*, and *crypto assets other than*[3]. The latter category also includes *utility tokens*, a specific type of crypto asset that has no financial purposes and is only intended to provide access to a good or a service supplied by its issuer. Some types of crypto assets, such as *Non-Fungible Tokens* (NFTs), are not covered by the regulation.

For each crypto asset category , the regulation contains norms regarding the content and issuance of the white paper, the conduct of business, in addition to organizational and financial requirements.

In alignment with the objective of the study, the rest of the work will focus only on the norms for white paper content, and in particular those valid for the crypto "other than" category. While the requirements vary across the three asset categories, the differences are marginal in the scope of the analysis. Another reason supporting this choice is the fact that the "other than" category has the largest set of requirements for white papers—see Appendix A.

---

[3]For brevity; the full definition in the text is "crypto assets other than asset-referenced tokens or e-money tokens".

Minimum content requirements. Each whitepaper must satisfy a comprehensive list of minimum content requirements[4] that are organized around macro-areas, listed in Table 2 in the case of "other than" crypto assets. A comparison with the other two categories is provided in Appendix A.

Table 3 shows some examples of these requirements, accompanied by related excerpts taken from existing (pre-regulation) white papers.

Other requirements In addition to minimum content requirements, other MiCAR norms impact white paper content and could have implications for future NLP applications on the topic.

- **Future value of the assets:** issuers of crypto assets are forbidden[5] from making claims about the future value of the asset in the white paper, aside from clarifying that the token might lose its value partially or in full [6]. These kinds of forward-looking statements have previously been analysed in IPO prospectuses—see Section 2.2.

- **Document language:** according to the regulation[7], *"the white paper shall be drawn up in an official language of the home Member State, or in a language customary in the sphere of international finance"*. Using a language other than English could pose some challenges that are examined in Section 4.

- **Document template:** the MiCA regulation states[8] that the white paper must adhere to the template and form established by the European Securities and Markets Authority (ESMA) and the European Banking Authority (EBA). As of October 2023, no template has been published yet, although consultations on technical requirements are in progress.

4 NLP in the MiCAR Landscape: Opportunities and Challenges

4.1 What NLP can do: exploring possibilities.

Even if there is currently a gap in NLP research applied to the topic of crypto white papers, other legal domains have benefited from collaboration with CS researchers (Zhong et al., 2020).

---

[4]As outlined in Annexes I, II, and III of the regulation.
[5]Articles *6.4, 6.5a, 19.3, 19.4a*.
[6]Does not apply to *e-money tokens*.
[7]Articles *6.9*, *19.8*, and *51.8*.
[8]Articles *6.10-6.12*

Tasks involving the interpretation and analysis of legal texts have been a popular subject for the legal NLP community. Current research mostly relies on pre-trained Transformer-based (Vaswani et al., 2017) models, with rising interest in the use of Large Language Models (LLMs) (Guha et al., 2023; Yu et al., 2023).

The rest of the section explores some avenues for applying existing NLP techniques to the MiCA regulatory framework.

Predicting the outcome of compliance checks. Given a text passage from a white paper and a MiCAR requirement, it would be beneficial to have a model capable of correctly and automatically check if the passage complies with the regulation. There have been attempts to solve this problem for different regulatory domains. For example, Zufall et al. (2020) attempt to automate the decision of whether an online post is subject to the European Union's Legal Framework against the Expression of Hatred. As part of their approach to classify hate speech from a legal standpoint, they trained GBERT (Chan et al., 2020) as a multi-class classifier on a manually annotated hate speech dataset. The model showed poor performance, suggesting that the task might be too complex for the model's capabilities. The task could be similarly challenging in the case of MiCAR, making it a good candidate for further research.

Matching passages with requirements. If automating compliance checks proves too difficult— or even undesirable—another useful application could be simply to identify which regulation norms are relevant to each text passage.

In a similar task, Ravichander et al. (2019) collects a set of annotated questions about the contents of privacy policies, in which the model must determine if the text passage contains an answer to the provided question, and classify the question-passage pair as Relevant or Irrelevant.

Abualhaija et al. (2022) use language models from the BERT family (Devlin et al., 2019), finetuned on the task of Question Answering, to assist requirements engineers in finding text passages relevant to their analysis of compliance requirements.

Question answering and information retrieval. Letting users ask questions about the content of a white paper—be it a potential investor trying to understand the project, or a regulator looking for a specific paragraph—is one of the most obvious and

potentially useful applications of NLP to this topic. Given the length and complexity of the regulation, stakeholders could also benefit from an efficient way to retrieve norms from the MiCAR text.

QA and IR systems in the legal and regulatory domain pose some specific challenges compared to traditional question answering (Abdallah et al., 2023). Among others, they usually require highly specialized datasets curated by legal experts, and there is less room for error since an imprecise or factually wrong answer could negatively impact a legal decision.

To mitigate the data problem, the authors of LEGALBENCH (Guha et al., 2023) collect over 150 annotated datasets in the legal domain, including QA tasks. Since regulated white papers will be publicly available, the research community could contribute to this effort in the future with a QA dataset inspired by the MiCA regulation.

LLMs for reasoning and summarization. As demonstrated in LEGALBENCH (Guha et al., 2023), Large Language Models can now achieve good performance on some challenging legal reasoning tasks in a zero-shot setting. While supervised approaches may still outperform these models, there is a clear advantage in using LLMs as they don't require labeled datasets. Selecting the appropriate instruction and few-shot examples for the prompt

can further improve performance (Yu et al., 2023; Trautmann et al., 2022).

Aside from reasoning, another promising application of LLMs is summarization. The emergence of training methods that use human preference data (Stiennon et al., 2020) has led to models that can produce high-quality summaries of text, including those in the legal domain (Pont et al., 2023), without domain-specific training. In the MiCAR context, both the regulation text and the white papers are long and often dense, making summarization a useful tool for regulators and investors interested in quickly analyzing their contents.

Named Entity Recognition. Some of the MiCAR requirements entail entity extraction, from the traditional kind (e.g., the list of persons involved in managing the project) to more domain-specific types of entities such as Legal Norm and Organization. Named entities are also useful for document organization and search. However, NER algorithms are particularly sensible to domain shifts: When general-purpose NER algorithms are used in texts from a narrow domain, there is often a performance degradation. To work around this limitation, Au et al. (2022) and Smădu et al. (2022) train entity recognition algorithms on annotated legal datasets in different languages. Given that crypto asset white papers are also financial

Part	Disclosure item	White paper sample
E	E5: the total number of crypto assets to be offered to the public or admitted to trading.	A total of 200 million tokens is put into circulation through a private sale and a public ICO.
	E14: Information about technical requirements that the purchaser is required to fulfil to hold the crypto assets.	To engage with the ███ protocol, individuals must first download ███ App.
H	H1: The consensus mechanism, where applicable.	The IBFT PoA is the default consensus mechanism on ███ Blockchain.
	H5: Information on the audit outcome of the technology used, if such an audit was conducted.	Two separate security assessments were conducted over a period of several months and conducted by the audit firms ███ and ███.
I	I4: A description of the risks associated with project implementation.	(...), no security measures can provide absolute protection against unauthorized access and data breaches.

Table 3: Examples of the minimum content requirements from MiCAR, accompanied by redacted excerpts taken from existing, unregulated, white papers. The first column shows the requirement category. All categories and their descriptions are listed in Table 2.

documents, NER models trained on data from the financial domain might also be an option (Salinas Alvarado et al., 2015; Francis et al., 2019).

4.2 Challenges

NLP can simplify the understanding of crypto asset white papers, aid regulators in MiCAR compliance, and reduce the effort required for issuers to generate compliant documents. It can also benefit investors by enabling them to make more informed decisions, if they are given the tools to better assimilate the contents of white papers.

However, there are also potential obstacles due to the narrow scope of application and the complexity of the subject matter.

Adapting to the legal and regulatory domain. As previously mentioned, performance degradation can occur when using general-purpose language models on narrow domains. This is not limited to entity recognition, but also impacts other tasks. One way of diminishing this effect could be making use of the models released by the legal NLP research community, such as LEGAL-BERT (Chalkidis et al., 2020) and POLBERT, a subsequent model trained on the Pile of Law (Henderson* et al., 2022) dataset.

Since white papers contain terms and language specific to the financial domain, models like FIN-BERT (Araci, 2019) and FINGPT (Yang et al., 2023) could also be appropriate. However, the world of crypto assets brings even more lexical issues with its acronyms and neologisms[9].

Domain shift between white papers and regulation. In the past, the issuers of crypto asset white papers have used very different language compared to the legal jargon found in regulatory texts. This difference might make it harder to find semantic similarities between white paper passages and the relevant regulation articles—one of the possible applications mentioned in Section 4.1.

Keymanesh et al. (2021) encounter a similar issue in developing a QA system that answers citizen queries about privacy policies, and they partially overcome it by using paraphrasing techniques for query expansion.

Language. The regulation allows using either the local language of the state the crypto asset will be issued in, or a language commonly used in finance. Although we can expect most papers to be in English, the usage of other languages might introduce difficulties, especially in combination with the aforementioned domain shift. There is ongoing research (Niklaus et al., 2023; Chalkidis et al., 2023) that aims to make available language models and NLP datasets that support multiple languages in the legal domain.

Document structure and length. In the largest study encountered (Thewissen et al., 2022), with a sample of 5210 white papers, the median document length is 30 pages, and the maximum is 167. As the number of required disclosure items rises with the introduction of MiCAR, we might see these numbers increasing.

Handling long documents in NLP presents unique challenges. Both traditional sequence models and transformer-based models struggle, with most BERT-based models having a maximum sequence length of 512 tokens. Even among recent LLMs, few exceed the limit of 4,096 tokens[10]. To get around this problem, Mamakas et al. (2022) modify LONGFORMER (Beltagy et al., 2020) and LEGAL-BERT (Chalkidis et al., 2020) to handle texts up to 8,192 tokens.

Until the regulation templates for white papers are released, it cannot be determined whether they will be required to adhere to a provided document structure. For documents that are not pre-segmented, Aumiller et al. (2021) propose a segmentation system for legal documents that uses topic modeling to split a given text into semantically coherent spans for downstream applications.

4.3 Implications for stakeholders

Regulators. NLP could help regulators by partially automating compliance checks, reducing the administrative burden on competent authorities. It can also improve the speed and accuracy of oversight with ad-hoc tools for domain experts, enabling them to respond more rapidly to market changes and potential infringements.

Issuers of crypto assets. For issuers, NLP-based tools can assist in drafting white papers that are compliant with MiCAR guidelines from the start. This can reduce the time and costs associated with legal consultations and revisions, making it easier to enter and operate within the European market.

---

[9]The Cryptopedia glossary from Gemini contains over 1000 terms unique to the crypto community.

[10]A limit which includes both input and output tokens.

Potential investors. From an investor's perspective, NLP-generated analyses of white papers can provide a more transparent and easily interpretable data source, democratizing access to investment information about the crypto asset market.

5 Conclusions

Crypto asset white papers are valuable data sources for NLP practitioners due to their public availability and rich informational content. The entry into force of the MiCAR opens new avenues for NLP applications to assist various stakeholders, including regulators, issuers, and potential investors, in navigating compliance and regulatory oversight.

Our survey of existing literature on the textual analysis of white papers revealed an active but siloed field that could benefit significantly from interdisciplinary collaboration between computer science, finance, and legal experts. Existing work in the legal NLP field could serve as a foundation for developing algorithms and models tailored to this new regulatory landscape.

Researchers have a clear opportunity to contribute to this emerging area, with the potential not only to streamline regulatory processes but also to create a more transparent and accountable crypto asset ecosystem.

Limitations

The work inevitably contains some speculative elements, given that, while the MiCAR text is public, no regulated white papers are available yet, and no regulatory workflows are yet finalized or known. The new white papers will be expected to follow a template designed by the ESMA, which has not been released yet. Therefore, the analysis of possible NLP applications, opportunities, and challenges is partially grounded in the structures and contents of existing, unregulated white papers, and it may not be entirely applicable to regulated ones.

Acknowledgments

References

OJ L 150. 2023. Regulation (eu) 2023/1114 of the european parliament and of the council of 31 may 2023 on markets in crypto-assets. *Official Journal of the European Union*, 150:40–205.

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *arXiv preprint arXiv:2304.06623*.

Sallam Abualhaija, Chetan Arora, Amin Sleimi, and Lionel C. Briand. 2022. Automated question answering for improved understanding of compliance requirements: A multi-document study. In *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 2–11.

Yang Bao and Anindya Datta. 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1371–1391.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Thomas Bourveau, Emmanuel T De George, Atif Ellahie, and Daniele Macciocchi. 2018. Initial coin offerings: Early evidence on the role of disclosure in the unregulated crypto market. *Available at SSRN*, 3193392.

Thomas Bourveau, Emmanuel T De George, Atif Ellahie, and Daniele Macciocchi. 2022. The role of disclosure and information intermediaries in an unregulated capital market: Evidence from initial coin offerings. *Journal of Accounting Research*, 60(1):129–167.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kun Chen. 2019. Information asymmetry in initial coin offerings (icos): Investigating the effects of multiple channel signals. *Electronic Commerce Research and Applications*, 36:100858.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Christian Fisch. 2019. Initial coin offerings (ICOs) to finance new ventures. *Journal of Business Venturing*, 34(1):1–22.

David Florysiak. 2022. Utility tokens, markets in crypto assets regulation (micar), and the costs of being public. *Markets in Crypto Assets Regulation (MiCAR), and the Costs of Being Public (December 7, 2022)*.

David Florysiak and Alexander Schandlbauer. 2022. Experts or charlatans? ICO analysts and white paper informativeness. *Journal of Banking & Finance*, 139(C).

Sumam Francis, Jordy Van Landeghem, and Marie-Francine Moens. 2019. Transfer learning for named entity recognition in financial and biomedical documents. *Information*, 10(8):248.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Kathleen Weiss Hanley and Gerard Hoberg. 2010. The information content of ipo prospectuses. *The Review of Financial Studies*, 23(7):2821–2864.

Peter Henderson*, Mark S. Krass*, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset.

Sabrina T Howell, Marina Niessner, and David Yermack. 2020. Initial coin offerings: Financing growth with cryptocurrency token sales. *The Review of Financial Studies*, 33(9):3925–3974.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Bedil Karimov and Piotr Wójcik. 2021. Identification of scams in initial coin offerings with machine learning. *Frontiers in Artificial Intelligence*, 4:718450.

Moniba Keymanesh, Micha Elsner, and Srinivasan Parthasarathy. 2021. Privacy policy question answering assistant: A query-guided extractive summarization approach. *arXiv preprint arXiv:2109.14638*.

Yiming Li, Baogang Wei, Yonghuai Liu, Liang Yao, Hui Chen, Jifang Yu, and Wenhao Zhu. 2018. Incorporating knowledge into neural network for text representation. *Expert Systems with Applications*, 96:103–114.

Daniel Liebau and Patrick Schueffel. 2019. Cryptocurrencies and icos: Are they scams? an empirical study. *An Empirical Study (January 23, 2019)*.

Yukun Liu, Jinfei Sheng, and Wanyi Wang. 2021. Technology and cryptocurrency valuation: Evidence from machine learning. In *Technology and Cryptocurrency Valuation: Evidence from Machine Learning: Liu, Yukun ajdnSheng, Jinfei ajd Wang, Wanyi*. SSRN.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 130–142, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Michele Meoli and Silvio Vismara. 2022. Machine-learning forecasting of successful icos. *Journal of Economics and Business*, 121:106071.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Paul P Momtaz. 2021. Entrepreneurial finance and moral hazard: evidence from token offerings. *Journal of Business Venturing*, 36(5):106001.

Andrew Morin, Marie Vasek, and Tyler Moore. 2021. Detecting text reuse in cryptocurrency whitepapers. In *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–5. IEEE.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus.

Thiago Dal Pont, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti, and Giovanni Sartor. 2023. Legal summarisation through llms: The prodigit project. *arXiv preprint arXiv:2308.04416*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Shadi Samieifar and Dirk G Baur. 2021. Read me if you can! an analysis of ico white papers. *Finance Research Letters*, 38:101427.

Jared Sharpe and Keith Decker. 2022. Prospectus language and IPO performance. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 154–162, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. Legal named entity recognition with multi-task domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Jie Tao, Amit V Deokar, and Ashutosh Deshmukh. 2018. Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. *Journal of Business Analytics*, 1(1):54–70.

James Thewissen, Prabal Shrestha, Wouter Torsin, and Anna M Pastwa. 2022. Unpacking the black box of ico white papers: A topic modeling approach. *Journal of Corporate Finance*, 75:102225.

Jennifer Thewissen, James Thewissen, Wouter Torsin, and Özgür Arslan-Ayaydin. 2023. Linguistic errors and investment decisions: the case of ico white papers. *The European Journal of Finance*, 29(7):826–868.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

Ju-Chun Yen, Tawei Wang, and Yu-Hung Chen. 2021. Different is better: how unique initial coin offering language in white papers enhances success. *Accounting & Finance*, 61(4):5309–5340.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.

Shuyu Zhang, Walter Aerts, Liping Lu, and Huifeng Pan. 2019. Readability of token whitepaper and ico first-day return. *Economics Letters*, 180:58–61.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2020. A legal approach to hate speech: Operationalizing the eu's legal framework against the expression of hatred as an nlp task.

A MiCAR Content Requirements Categories

A.1 Categories for Crypto Assets "other than"

Found in Annex I of the regulation.

- **Part A:** Information about the offeror or the person seeking admission to trading

- **Part B:** Information about the issuer, if different from the offeror or person seeking admission to trading

- **Part C:** Information about the operator of the trading platform in cases where it draws up the crypto asset white paper

- **Part D:** Information about the crypto asset project

- **Part E:** Information about the offer to the public of crypto assets or their admission to trading

- **Part F:** Information about the crypto assets

- **Part G:** Information on the rights and obligations attached to the crypto assets

- **Part H:** Information on the underlying technology

- **Part I:** Information on the risks

	Part A	Part B	Part C	Part D	Part E	Part F	Part G	Part H	Part I
Crypto "other than"	10	8	9	6	19	2	10	5	5

Table 4: Category count for each Part in the content requirements, for crypto assets "other than".

A.2 Categories for Asset-Referenced Tokens and E-money Tokens

Found in Annexes II and III of the regulation.

- **Part A:** Information about the issuer of the token

- **Part B:** Information about the token

- **Part C:** Information about the offer to the public of the token or its admission to trading

- **Part D:** Information on the rights and obligations attached to the token

- **Part E:** Information on the underlying technology

- **Part F:** Information on the risks

- **Part G:** Information on the reserve of assets

	Part A	Part B	Part C	Part D	Part E	Part F	Part G
Asset-referenced Tokens	12	5	15	17	5	6	5
E-money tokens	13	3	4	8	6	3	-

Table 5: Category count for each Part in the content requirements, for asset-referenced and e-money tokens.