Special Issue: Should Data Drive Private Law

Technology and Regulation

Talking at Cross Purposes?

A computational analysis of The debate on informational duties in the digital services and the digital markets acts

Fabiana Di Porto[*], Tatjana Grote[**], Gabriele Volpi[***], Riccardo Invernizzi[****]

Digital Services Act, Digital Markets Act, Big Platforms, Computational Analysis, Transparency duties

fabiana.diporto@unisalento.it

n.t.grote@lse.ac.uk

gabrielevolpi@me.com

riccardo.invernizzi03@universitadipavia.it

Since the opaqueness of algorithms used on online platforms opens the door to discriminatory and anti-competitive behaviour, increasing transparency has become a key objective of lawmakers. Leveraging the analytical power of Natural Language Processing, this paper investigates whether key terms related to transparency in digital markets were used in the same way by different stakeholders in the consultation on the EU Commission's DSA and DMA proposals. We find significant differences in the employment of terms like 'simple' or 'meaningful' in the position papers that informed the drafting of the proposals. These findings challenge the common assumption that phrases like 'precise information' are used the same way by those implementing transparency obligations and might partially explain why they frequently remain ineffective.

1. Introduction

When EU Executive Vice-President Margarethe Vestager presented the latest Commission proposals on digital platforms, the Digital Markets Act (DMA) and the Digital Services Act (DSA),[1] she compared them to the invention of the traffic light, which was created in response to the rapidly increasing importance of the car. She concluded that 'just like back then, … now we have such an increase in the online traffic that we need to make rules that put order in the chaos'.[2]

This twin-proposal suggests many new rules for digital intermediary services and online platforms.[3] With the DSA and DMA, the Commission closes a period during which stakeholders (and doctrine)[4] have been harshly discussing new ex ante rules for digital markets, both from a consumer protection and a competition law perspective.[5]

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (COM(2020)825), 15 December 2020 [hereinafter Digital Services Act, DSA]; European Commission, Proposal for a Regulation of the European Parliament and the Council on contestable and fair markets in the digital sector (Digital Markets Act) (COM(2020) 842), 15 December 2020 [hereinafter Digital Markets Act, DMA].

[*] Fabiana Di Porto is Professor of Law and technology at the University of Salento, Italy and Contract Professor of Innovation Law and Regulation at LUISS, Rome. She is Head of the Algorithmic Disclosure Regulation PRIN-funded Research Project 2019-22 (www.lawandtechnology.it, grant no 2017BAPSXF).We are thankful to Prof. Julian Nyarko from Stanford University for providing some of the data we used in the empirical part and to a number of other people for insightful discussions: Dr Roberto Marseglia, Professors Daniela Piana and Giuseppe Italiano and the anonymous referees. All mistakes remain our own.

[**] Tatjana Grote has a BA in International Relations from the Technical University of Dresden and is currently an LLM Candidate at the London School of Economics and Political Science as well as a member of the PRIN Algorithmic Disclosure Regulation project 2019-22.

[***] Gabriele Volpi has a msc Theoretical physics from the University of Pavia, Manager at DAAT consulting and Member of the PRIN Algorithmic Disclosure Regulation project 2019-22.

[****] Riccardo Invernizzi has a bsc in Mathematics - with honors - from the University of Pavia, junior analyst at DAAT consulting and Member of the PRIN Algorithmic Disclosure Regulation project 2019-22.

Received 3 Dec 2020, Accepted 5 May 2021, Published 11 Feb 2022.

[2] European Commission, Statement by Executive Vice-President Vestager on the Commission proposal on new rules for digital platforms, 15 December 2020, https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_20 _2450 (accessed 15 February 2021).

[3] There is no perfect alignment in the definition of platform services in the DSA and DMA. In the DSA, the widest concept is that of online '*intermediary service*', which covers all services within the scope of Art. 1(3), including 'online platforms' (providing hosting services) under the meaning of Art. 2(1)h DSA. In the DMA, the widest category is that of '*core online platform*'. Art. 2(2) 'online intermediation services' are one service type among the many 'core platform services' (together with e.g., cloud services, social networks, videosharing platforms). Some '*core online platforms*', then, may be designated as 'gatekeepers' (DMA, Art. 3) if they (a) have a significant impact on the internal market, (b) serve as a gateway between business and end-users, (c) enjoy an entrenched and durable position. The requisites are presumed to exist: (i) if the 'core platform service' was provided in at least 3 MS and given thresholds of average market capitalization are overcome; (ii) the core platform has more than 45 million monthly active end-users plus 10.000 business users; (iii) the thresholds in point (ii) were met in each of the last three financial years. (Art. 3, DMA).
Hence, for the sake of parallel applicability of the DSA and DMA transparency rules, not every (core) very large platform is a gatekeeper, but it is likely that every gatekeeper will also be a very large (core) online platform (see Art 3(2)b DMA).

[4] P Ibáñez Colomo, Whatever Happened to the 'More Economics-Based Approach'?, *Journal of European Competition Law & Practice* (2020) 11, 9, 473–74, (discussing the shift from the so called 'more economic approach' to the growing demand for ex ante intervention against big digital platforms in the European legal community).

[5] For challenges related to competition law, see e.g., A Ezrachi & M Stucke,

Fabiana Di Porto, Tatjana Grote, Gabriele Volpi, Riccardo Invernizzi, Talking at Cross Purposes: A computational analysis of The debate on informational duties in the digital services and the digital markets acts – Technology and Regulation, 2021, 87–106 • https://doi.org/10.26116/techreg.2021.008 • ISSN: 2666-139X

Although the two proposals differ in scope and focus,[6] both reveal that one key instrument the Commission relies upon in 'ordering' chaotic traffic in digital markets is informational duties (inclusive of both transparency and disclosure obligations).[7]

This is surprising and unsurprising at the same time. According to the standard narrative, informational duties play a central role in the realm of consumer protection[8] and serve to rebalance unequal bargaining power in trade relationships.[9] And digital markets would be no exception.[10]

On the other hand, the very utility of informational duties has been systematically questioned.[11] Overall, such duties seem to have more of a symbolic (*rectius*, political) value rather than true utility.[12] In the digital realm, many argue that extra-long disclaimers and hard-to-read terms of contract would be useless, or sometimes run counter consumers empowerment.[13] A similar argument is made for platform-to-business relations, where information duties are often considered insufficient to mitigate unequal bargaining power.[14]

This paper aims to investigate *why*, despite the long-lasting scholarly debate about their limited effectiveness, and overwhelming evidence supporting it, the DSA and DMA rely heavily on disclosure.[15] More specifically, we investigate what are the possible *sources* of ineffectiveness.

There have been many attempts to do that, the behavioral literature on disclosure being the most relevant in two regards. On one side, it has provided empirical evidence of the impact of informational arrangements[16] adopted by big digital platforms by measuring how much they affect the behavior of consumers. On the other, it has accounted for the effectiveness of disclosure duties by measuring how many consumers like or dislike them.[17] However, these studies take the legal duty as a given, an external variable. On the contrary, we contend that much can be said about their origin and the process through which this duty is formed.

Therefore, we propose to leverage the power of computational tools, among which Natural Language Processing (NLP) and Machine Learning (ML) techniques: by linguistically analyzing the debate that preceded the adoption of these duties, our empirical study suggests searching for possible sources of failure in the feedback documents to the consultation, that were input to these rules.

Our contribution innovates in several regards. First, our methodology is not effects-based, in the sense that to assess the efficacy of transparency duties, it does not look at the impact on nor the perceptions of those who receive the information, being this input context-specific. We rather analyze the *wording* that conflated the debate around the provisions establishing informational duties of

Virtual competition: the promise and perils of the algorithm-driven economy (Harvard University Press 2016), and P Marsden & R Podszun, estoring Balance to Digital Competition – Sensible Rules, Effective Enforcement, (Konrad-Adenauer-Stiftung 2020), 1-87. On consumer protection and its relation to data protection and competition law, see W Kerber, Digital markets, data, and privacy: competition law, consumer law and data protection, *Journal of Intellectual Property Law & Practice* (2016) 11(11), 856-866.

[6] Both the DMA and DSA take a resolute stance, through ex ante regulation, against the big platforms. However, the DSA aims primarily to '*ensur[e] a safe and accountable environment*' by applying asymmetric ex ante rules to online digital platforms, according to two parameters: the company's role (i. intermediary services, ii. hosting services, iii. online platforms), and size (a. large online platforms and b. very large platforms i.e., those reaching more than 45 million consumers, which will have to comply with special rules). The DSA imposes obligations on transparency, illegal content, and accountability requirements. Therefore, it addresses negative externalities and asymmetric information. On the other hand, the DMA's goal is to '*ensur[e] fair and open digital markets*' by applying asymmetric rules against large online platforms designated as '*gatekeepers*', which are addressed with a list of does and don'ts. Taken together, they can be read as an ex ante toolbox, made of a mix of competition and consumer protection rules. While the DSA amends the e-commerce directive (2000/31/EC), the DMA centers around concerns and seeks to complement EU competition rules (mostly Art 101, 102 TFEU). Finally, the DSA applies to all 'intermediary services' (Art 1), while the scope of the latter is limited to 'core platform services' offered by 'gatekeepers' as defined in Art 3 DMA.

[7] We use disclosure, transparency and informational duties interchangeably as what is relevant to the analysis is the way the terms related to the provision of information are used by the stakeholders. However, we acknowledge that there are duties owed to users and those to public authorities; and that information may well be provided for purposes of public or private disclosure, or for reasons of investigations. A taxonomy of transparency and disclosure duties is nonetheless provided for in Table 1 in the Appendix, to which reference is made in the legal analysis of Section 2.3 *below*.

[8] European Parliament resolution of 20 October 2020 with recommendations to the Commission on the Digital Services Act: Improving the functioning of the Single Market (2020/2018(INL)), 20 October 2020, 12 (no. 31, 32).

[9] See e.g., EA Posner, ProCD v. Zeidenberg and Cognitive Overload in Contractual Bargaining. *University of Chicago Law Review* , E. A. (2010) 77(4), 1181-1194.

[10] Algorithm Watch (2020), Governing Platforms – Final Recommendations, available at https://algorithmwatch.org/wp-content/uploads/2020/10/Governing-Platforms_DSA-Recommendations.pdf (accessed 17 February 2021), 1.

[11] See e.g., O Ben-Shahar & CE Schneider, Coping with the Failure of Mandated Disclosure. *Jerusalem Review of Legal Studies* (2015) *11*(1), 83–93; F Marotta-Wurgler, Even More Than You Wanted to Know About the Failures of Disclosure. *Jerusalem Review of Legal Studies* F. (2015) *11*(1), 63–74. E Zamir, & D Teichman, *Behavioral Law and Economics*. (Oxford University Press 2018), 171-177; F Di Porto, & M Maggiolino, Algorithmic Information Disclosure by Regulators and Competition Authorities. *Global Jurist*, (2019). 19(2), 11; E. Bardach & RA Kagan, *Going by the book: The problem of regulatory unreasonableness*. (Temple University Press 1982), 249-256; A Prat, The Wrong Kind of Transparency. *American Economic Review*, (2005) 95(3), 862.

[12] Di Porto & Maggiolino (n 11) 14.

[13] SK Ripken,The Dangers and Drawbacks of the Disclosure Antidote: Toward a More Substantive Approach to Securities Regulation. *Baylor Law Review* (2006) *58*(1), 160.

[14] Marsden & Podszun (n 5), 18; F Di Porto & M Zuppetta, Co-Regulating Algorithmic Disclosure for Digital Platforms, *Policy and Society* (2020) 0(0), 3-4; C Busch, Crowdsourcing, Consumer Confidence: How to Regulate Online Rating and Review Systems in the Collaborative Economy. In C Economy & A De Franceschi (Eds.), *European Contract Law and The Digital Single Market: The Implications of The Digital Revolution*, 223. (Intersentia 2016).

[15] See M Sentfleben & C Angelopoulos, The Odyssey of the Prohibition on General Monitoring Obligation on the Way to the Digital Services Act: Between Article 15 of the e-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3717022 (accessed 23 April 2021) and G Frosio (2020). Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747756, discussing transparency duties in the DSA. For an analysis of disclosure remedies in the DMA, see Ibáñez Colomo P (2021). The Draft Digital Markets Act: A Legal and Institutional Analysis, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3790276 (accessed 23 April 2021).

[16] See e.g., J Luguri & L Strahilevitz, (2021). Shining a Light on Dark Patterns. https://doi.org/10.2139/ssrn.3431205 (accessed 26/06/2021) (discussing the impact of dark patterns, including informational ones).

[17] See e.g., O Katz & E Zamir, Do People Like Mandatory Rules? The Choice Between Disclosures, Defaults, and Mandatory Rules in Supplier-Customer Relationships, JELS (2021) 18(2) 421-60 (who compare the desirability of disclosures duties, from the perspective of the consumer, as compared to mandatory rules and default rules).

the DSA and DMA. Especially, we ask whether the meaning and use of terms that were discussed and finally became parts of information duties were fully shared among the stakeholders or not. For instance, terms like 'clear' or 'unambiguous' (referred to in Art. 24 DSA and extensively discussed before its adoption) are understood the same way by online platforms using personalized ads (addressed by the duty to disclose information) and the consumers (addressee of the information piece)? If this is not, could that be a source of disclosure ineffectiveness?

To assess if this is the case, we look at the stakeholder's submissions to the Commission's public consultation over three Inception Impact Assessment documents (IAs) that were input to the DSA and DMA proposals, namely: the so-called 'New Competition Tool',[18] the 'Ex ante regulatory instrument for large online platforms'[19] (hereafter also: ex ante tools), and the (then) 'Digital Services Act'.[20]

Second, we add computational analysis to standard manual reading of submissions that is done by the Commission without the help of algorithms.[21] The total of 2.862 replies to questionnaires and feedback documents contain the comments of all stakeholders regarding the proposals put forward by the Commission in its inception IAs. They, therefore, constitute an exceptional source of knowledge about who supported and opposed these duties among them, and especially, how individuals and organizations understand and use relevant terms of transparency. While manually processing the replies might still allow identifying the need for transparency duties, there are two shortcomings of this approach. First, any manual 'analysis' of the feedback documents comes with quite substantial labor cost, something that 'distant reading' can do more efficiently.[22] Second, no human reader can quantify the extent to which the same terms are used in the same way by different stakeholders. For instance, while both a large online platform and a consumer or smaller business might speak of a need for more 'precise' information, the underlying understanding and consequent use of this term could differ. In the context of transparency obligations, this is problematic since these duties might remain ineffective if a disclosure statement is only 'readable' in the eyes of the platform drafting it, but not in the eyes of the individual consumer or the micro organization reading it.

One way to cope with such limitations is to computationally analyze the feedback submitted to the Commission through the means of a mixed supervised and unsupervised ML technique, that would *complement* standard processing by public officials in the Directorates General (DG). Specifically, we propose doing so by using Word Embedding Alignment,[23] a state-of-the-art model for translation,[24] which can be adapted to our task, i.e. monolingual translation from a language to itself to evaluate the difference in the use of the same word in different corpora.[25] As a plus, word embedding modelling is highly compatible with unsupervised learning, a feature[26] that is very useful since, as explained before, in this context we should avoid the participation of human coding during the training process as much as possible.

This way, we aim to answer two central questions: (1) Do different groups of contributors share the same understanding (measured as semantical differences between terms) and use of the central terms and issues surrounding transparency and disclosure duties in the DSA and DMA? (2) Can we identify different clusters of opinions towards key concepts and can they be a possible source of disclosure failure? Our success in finding an answer to these questions with the help of said tools will be reflected with a view to a third overarching question: (3) can computational techniques help to partially automate the collection and analysis of opinions that are inputs to a rulemaking process? If this is the case, then we should recognize their potential in supporting the creation of better information disclosure rules, as is the proclaimed goal of the DSA and DMA consultation procedure, that is disclosure rules that are less prone to failure.

The article is structured as follows. The following section outlines the informational challenges posed by digital markets and the role of transparency duties set forth in the DSA and DMA proposals in mitigating their negative effects on consumers and businesses (Section 2). We then present our computational text analysis of the consultation documents and results, showing that not only are similar opinions expressed by groups that usually belong to different clusters (i.e., medium and big organizations); but also that groups of stakeholders use central terms in different ways (Section 3). We lastly conclude by sketching how a similar procedure could help to draft smarter disclosure regulations in a larger context.

2. Informational Malpractice in the Digital Era

For many commentators, the prominent role of transparency obligations in the DSA and DMA did not come as a surprise.[27] Disclosure

[18] New Competition Tool, Inception impact assessment, Ares(2020)2877634, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12416-New-competition-tool (accessed 31 March 2021).

[19] The Ex ante regulatory instrument for large online platforms with significant network effects acting as gate-keepers in the European Union's internal market, Inception impact assessment, Ares(2020)2877647, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers

[20] The (then) Digital Services Act, Deepening the Internal Market and clarifying responsibilities for digital services, Inception impact assessment, Ares(2020)2877686, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services_en.

[21] R Senninger, Analyzing the EU Commission's Regulatory Scrutiny Board through quantitative text analysis. *Regulation & Governance,* (2020) *1;* CM Radaelli, Regulating Rule-making via Impact Assessment. *Governance* (2010). 23(1), 89–108; CA Dunlop & CM Radaelli, Impact Assessment in the European Union: Lessons from a Research Project. *European Journal of Risk Regulation* (2015) 6(1), 27–34.

[22] J. Grimmer & B.M. Stewart, Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* (2013) 21(3), 267–297.

[23] See e.g., D Alvarez-Melis & TS Jaakkola, Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890. Association for Computational Linguistics; Yehezkel Lubin, N., Goldberger, J., & Goldberg, Y. (2019). Aligning Vector-spaces with Noisy Supervised Lexicons. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 460–465.

[24] A Abdelsalam, O Bojar & S El-Beltagy, Bilingual Embeddings and Word Alignments for Translation Quality Estimation. *Proceedings of the First Conference on Machine Translation (2016): Volume 2, Shared Task Papers,* 764–771.

[25] J Nyarko & S Sanga (2020). A Statistical Test for Legal Interpretation: Theory and Applications, 25 November 2020, https://juliannyarko.com/wp-content/uploads/other/nyarko_sanga_legal_interpretation.pdf. (showing how word embedding modelling can fit very well our task).

[26] T Wada & T Iwata (2018). Unsupervised Cross-lingual Word Embedding by Multilingual Neural Language Models. arXiv:1809.02306 [cs]; A Conneau, G Lample, M Ranzato, L Denoyer & H Jégou, H. (2018). Word Translation Without Parallel Data. arXiv:1710.04087 [cs].

[27] See e.g., Global Network Initiative (2020). Thinking Through Transparency and Accountability Commitments Under The Digital Services Act, 20 July 2020, https://medium.com/global-network-initiative-collection/thinking-through-transparency-and-accountability-commitments-un-

duties of all kinds have long been conceived as a key policy instrument to tackle the manifold challenges arising from digital markets. This section will give a snapshot of these challenges focusing and explaining the role of transparency in theory and in the DSA and DMA.

2.1 Talking at Cross Purposes. The Debate on the Need to Update Informational Duties through the DSA and DMA

Consumers benefit in many ways from the impressive development of digital markets.[28] However, certain characteristics of digital markets come with new challenges and risks. Concerning consumer protection, the sale of illicit goods in online marketplaces and unfair contractual clauses are key concerns.[29] But opaque online environments, as the Crémer report rightly emphasized, may also be 'a competition policy issue'.[30]

The relationship between transparency on the one side, and competition law and consumer protection, on the other, is bidirectional. A lack of competition might force business users to accept a level of transparency they do not feel comfortable with, in absence of an alternative supplier of the online service they are consuming.[31] This is an important realization since digital markets show certain characteristics which are likely to favor highly concentrated markets.[32]

Taken together, these factors work in favor of large online platforms, which might accumulate some kind of 'gatekeeping' power and impose the level of transparency they deem appropriate on the market they dominate. Of course, they technically still underly certain transparency obligations, for instance, those included in the GDPR.[33] However, the GDPR does not cover all relevant phenomena and users.[34]

Furthermore, platforms' understanding of specific requirements like e.g., 'clear and easy' language, might effectively determine the usefulness of disclosures for consumers, the small and medium enterprises. When consumers are not able to switch to a different provider giving information in a way that better fits their needs and capacities, a lack of competition could thus result in a lack of transparency.

The other way around, there are also situations in which a lack of transparency can endanger competition due to allowing for certain anti-competitive practices. In its investigation report on competition in digital markets, the US Congress subcommittee on Antitrust, Commercial Law and Administrative Law has summarized this as follows: 'Without transparency or effective choice, dominant firms may impose terms of service with weak privacy protections that are designed to restrict consumer choice, creating a race to the bottom'.[35] Clearly, that depends on the fact that in digital markets products are mainly zero-priced, and 'privacy and quality of service can be differentiating factors'[36]; hence, granting transparency or effective choice can help ensure competition.

Such a problem may arise in case platforms manipulate the order in which offers from business customers are presented.[37] Only if the parameters used to rank products are transparent, it will be possible to know whether an online platform is distorting competition by preferencing certain offers,[38] leaving consumers in the dark about the 'trade-offs they are facing', and hence inhibiting competition in a significant manner. In particular, self-preferencing by the big tech has been long debated as a cause of competition law infringement.[39]

der-the-digital-services-act-e4dce3cee909 (accessed 22 January 2021); S Stolton(2020). Make Big Tech accountable, Austria says in Digital Services Act recommendations, Euractiv, 30 November 2020, https://www.euractiv.com/section/digital/news/make-big-tech-accountable-austria-says-in-digital-services-act-recommendations/ (accessed 22 January 2021).

[28] See Recital 1 DSA. To name just a few of these benefits: digital marketplaces facilitate cross-border trade and amplify product choices, social media allows cheap, easy, and quick communication, digital start-ups spur innovation and offer new services.

[29] Concerning contractual clauses, an empirical analysis has identified potentially unfair contractual clauses in roughly 10% of a sample of 50 online consumer contracts. M Lippi, P Pałka, G Contissa, F Lagioia, H Micklitz, G Sartor & P Torroni, CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* (2019) 27(2), 117–139.

[30] J Crémer, Y. de Montjoye & H Schweitzer, Competition policy for the digital era, European Commission Report (2019), https://data.europa.eu/doi/10.2763/407537 (accessed 14 February 2021) [hereinafter Crémer Report], 63.

[31] This problem is well-framed as follows: 'a lack of options to switch to qualitatively similar other search engines or social networks might lead users to accept also very high prices (in form of collected data) and privacy policies that do not match their specific privacy preferences'. Kerber (n5) 867.

[32] Crémer report (n 30) 2-3; M Gal & N Petit, Radical Restorative Remedies for Digital Markets. *Berkeley Technology Law Journal* (2020) 37(1), 5-6; OECD, Roundtable on Algorithms and Collusion - Executive Summary (DAF/COMP/M(2017)1/ANN3/FINAL), 26 September 2018, 5; F Scott Morton, P Bouvier, A Ezrachi, A Jullien, R Katz, G Kimmelman, D Melamed & J Morgenstern, Committee for the Study of Digital Platforms, Market Structure and Antitrust Subcommittee, Stigler Center for the Study of the Economy and the State [hereinafter Stigler report] (2019) 14. PG Picht & GT Loderer, Framing Algorithms: Competition Law and (Other) Regulatory Tools. *World Competition*, (2019) 42(3), 406.

[33] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 27 April 2016, O.J. L 119/1 [hereinafter GDPR].

[34] For instance, the GDPR is not really relevant for business users, for it covers the personal data of individuals only (Art 2(1) in connection with Art 4(1) GDPR). It does not touch on the circumstances under which data (or content) deliberately shared by an individual can be removed by a platform. Neither does it regulate how data shared by a business user of an intermediary service should be displayed and what the user ought to know about this, which is central from a competition perspective.

[35] U.S. House Committee on the Judiciary (2020). Investigation of Competition in Digital Markets. Washington, D.C.: Government Printing Office. The Subcommittee report also mentions manipulative design interfaces, so called dark patterns, nudging consumers into certain choices. Ibid, 53.

[36] Ibid, 54.

[37] Some authors argue that where consumer choices are being influenced, there is a special need for transparency duties: "A core element of such duties could be the obligation to thoroughly explain the workings of an algorithm, not on a technical level but regarding its impact on the customer, especially where it is designed to replace customer choice". Picht and Loderer (n 32) 416.

[38] Contra, L Signoret, Code of competitive conduct: a new way to supplement EU competition law in addressing abuses of market power by digital giants. *European Competition Journal*, (2020). 16(2-3), 221, at 244 (contending that where platforms gain market power by being more efficient or winning consumers based on free choice by providing better offers, this would not constitute a violation of competition law).

[39] Self-preferencing was at the heart of the Microsoft saga (see JP Jennings, Comparing the US and EU Microsoft Antitrust Prosecutions: How Level Is the Playing Field. *Erasmus Law and Economics Review*, (2006) 2, 71–86.) and was also heavily discussed by the doctrine at the time of the Google Shopping case. In fact, the Google Shopping case established that self-preferential placements are, indeed, not compatible with competition law. *Google Search (Shopping) Case C(2017) 4444*, 27 June 2017, paras 9, 10 of summary decision. See e.g., P Ackman, The Theory of Abuse in Google Search: A Positive and Normative Assessment Under EU Competition Law, in *Journal of Law, Technology & Policy*, (2) 301-372.

2.2 Legal Grounds for Updating Informational Duties

In the debate on how to react to some of these challenges, the e-Commerce Directive (ECD) has been central.[40] It is the piece of legislation the DSA updates and amends as 20 years of technological developments necessarily opened up some transparency-related lacunas.

First, platforms have quite simply become significantly larger and more important.[41] And with the reach of platforms, the amount of user-generated content has increased exponentially.[42] Hence, it is the increase in volume and magnitude of markets that justify a different approach. Second, existing rules were adopted when content moderation by automated means was not yet a widespread practice, if available at all.[43] Third, the increased relevance of recommender systems, digital nudging, personalized advertising also did not exist and was therefore not addressed by the ECD.[44]

Against the background of these developments, commentators and lawmakers have advocated in favor of significantly expanding the information duty framework of Arts 5, 6, and 10 ECD, with the aim of 'putting meaningful transparency at the heart' of new EU rules on digital services.[45]

With regards to the DMA, general shortcomings of EU competition rules when dealing with opaque online practices have been highlighted,[46] showing that law, albeit helpful, would most likely not suffice to achieve a satisfactory level of transparency.[47]

In light of these interconnected challenges for consumer protection and competition, the strong focus of the European Commission on informational duties as an easily enforceable means to increase transparency and mitigate information asymmetries seems reasonable in principle.[48]

However, over time, critics of information duties have continuously added evidence to the list of phenomena hampering the effectiveness of disclosures, which now includes e.g., information overload,[49] confirmation bias,[50] decision-making aversion,[51] the no-reading problem[52], and dislike.[53]

Despite this criticism, the Commission reports that 'many' in the consultation process have been calling for more informational duties. In the DMA, these 'many' correspond to civil society and media publishers, who 'called for an adequate degree of transparency in the market as well as the respect of consumers' autonomy and choice'.[54] In the DSA, the quest for 'algorithmic accountability and transparency audits, especially with regard to how information is prioritized and targeted' online comes from 'a wide category of stakeholders', and is particularly voiced by 'civil society and academics'.[55]

Apart from these brief notes, one cannot find more reference to the position of stakeholder groups with regards to transparency duties in the inception IAs. It is therefore relevant to see whether this synthesis duly captured the existing variegated positions. Before moving to our empirical analysis, we will briefly illustrate the actual transparency duties contained in the DSA and DMA proposals. These constitute the formalization of the debate we illustrated above, and we will use it as a blueprint for our empirical research.

2.3 The Actual Informational Duties in the DSA and the DMA

The European Commission's vision of what transparency rules might look like, as recently elucidated in the consultation on the DMA and DSA, will be briefly presented in the following. Some of these duties are new, while others are state-of-the-art for many operators. Indeed, especially those enlisted in the DSA are simply restated from the 2019 Platform-to-Business Regulation[56] and the amended Consumer Rights

[40] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on Electronic Commerce), 17 July 2000, O.J. L 178/1 [hereinafter ECD]; The ECD is considered by some as "the cornerstone of the Digital Single Market", European Parliament (n 8) 17.

[41] Given that they reach a massive number of users, illegal or otherwise problematic content and practices will now impact considerably more citizens. SB Micova & A De Streel, Digital Services Act – Deepening the Internal Market and Clarifying Responsibilities for Digital Services, Centre on Regulation in Europe Report, 2 December 2020, https://cerre.eu/publications/digital-services-act-responsibility-platforms/ (accessed 16 February 2021) [hereinafter CERRE DSA Report], 10.

[42] Alarmingly, this development has been associated with a rise in hate speech and disinformation. European Parliament, (n 8) 3.

[43] Micova & De Streel (n 41) 10.

[44] European Parliament (n 8), on page 12, mentions 'advertising, digital nudging and preferential treatment; paid advertisements or paid placement in a ranking of search results' as novel challenges to be addressed. Algorithm Watch (n 10) 1; European Parliament, (n 8) 5.

[45] Algorithm Watch (n 10) 1.

[46] The Crémer report points out several criticalities: (1) not all gatekeepers enjoy a dominant position in the sense of Art. 102 TFEU; (2) the relevant market might be substantially harder to define than in non-digital cases; (3) not every problematic practice has a demonstrable effect on the relevant market. The authors conclude that greater emphasis should be put on the theory of harm, instead. Crémer report (n 31) 3-4. Moreover, digital markets are often moving at a rapid pace, which is not necessarily a characteristic they share with competition law. Hence, there are concerns whether competition law could be applied with the necessary speed to address urgent competition needs. A de Streel, Digital Markets Act – Marking Economic Regulation of Platforms Fit for the Digital Age, Centre on Regulation in Europe Report, 24 November 2020 [hereinafter CERRE DMA report], 59; Recital 5 DMA.

[47] Information duties have also increasingly been acknowledged as competition remedies by courts, partly shifting from traditional cease and desist orders towards transparency duties see SW Waller, Access and Information Remedies in High-Tech Antitrust, *Journal of Competition Law and Economics* (2012) 8(3), 575, at 576.

[48] JC Coffee, Market Failure and the Economic Case for a Mandatory Disclosure System. *Virginia Law Review* (1984) 70(4), 717–753; SJ Grossman & JE Stiglitz, Information and Competitive Price Systems. *The American Economic Review* (1976) 66(2), 246–253; SJ Grossman & JE Stiglitz, On the Impossibility of Informationally Efficient Markets. *The American Economic Review* (1980) 70(3), 393–408; PG Mahoney, Mandatory Disclosure as a Solution to Agency Problems. *The University of Chicago Law Review* (1995) 62(3), 1047–1112.

[49] HA Simon, A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* (1955) 69(1), 99–118.

[50] A Tversky & D Kahneman, Judgment under Uncertainty: Heuristics and Biases. *Science* 1(1974) 185(4157), p. 1124–1131.

[51] O Ben-Shahar & CE Schneider, The Failure Of Mandated Disclosure. *University of Pennsylvania Law Review* (2011) 159, 727, IIdd (2015) (nt 11).

[52] For an empirical investigation of this issue, see Y Bakos, F Marotta-Wurgler & DR Trossen, Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts. *The Journal of Legal Studies*, (2014) 43(1), 1–35.

[53] Katz & Zamir (n 17).

[54] DMA, at 8 (summarizing the results of stakeholder consultations and impact assessments).

[55] DSA at 9. See also Algorithm Watch (n 10) 1; CERRE DSA report (n 41) 39; European Parliament, (n 8), 5; European Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust, COM/2020/65 final, 19.2.2020, 15.

[56] Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services OJ L 186, 11.7.2019, p. 57–79.

Directive[57].

2.3.1 DSA: Arts. 12(1), 13, 23-25, 29 and 33

As summarized in Table 1 in the Appendix, the DSA proposal includes a variety of transparency and disclosure obligations (together: informational duties) for providers of intermediary services.[58]

Art 12(1) would entail a general obligation to inform users about potential restrictions to their services contained in the terms and conditions. This information would need to be publicly available, provided in an *easily accessible format*, and written in *clear and unambiguous language.*

Whereas agreeing to the terms and conditions of a platform can be a one-time action, Art 13 DSA would oblige platforms to publish yearly reports about their content moderation practices. These reports would need to be drafted in a *clear and comprehensible language* and include certain specific information.[59]

While these obligations would apply to all providers of intermediary services, online platforms would additionally have to provide information about the out-of-court dispute settlements, content suspensions, and the use of automatic tools for content moderation (Art 23 DSA). Concerning the latter, the platform would be obliged to elucidate the '*precise* purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied'. Consequently, it seems fair to expect that the understanding of terms like 'precise' 'clear' 'unambiguous' would be crucial factors in determining the scope and form of the information provided to users.[60]

For online platforms displaying advertisements, Art 24 DSA would establish further informational duties. Advertisements and their publishers would have to be identifiable in a 'clear' and 'unambiguous manner'. Furthermore, platforms would have to share 'meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed' with the platform user. In addition to the obligations laid down in Art 24 DSA, very large online platforms within the meaning of Art 25 DSA,[61] would further need to offer application programming interfaces (APIs) to access information on the advertisements they display (Art 30(1), (2) DSA).

Apart from advertisement algorithms, rankings and recommender systems have been identified above as another platform architecture component requiring increased transparency.[62] For very large online platforms this challenge is addressed by Art 29 DSA: in their terms and conditions, very large online platforms would have to flag the use of recommender systems and explain in a 'clear, accessible, and easily comprehensible manner' how these systems work (i.e., which parameters they use and how they can be modified or influenced).[63] Again, the question of how simple, precise and understandable disclosures are understood seems central regarding the *de facto* effect of these transparency duties.

Lastly, Art 33 sets out comprehensive transparency obligations for very large online platforms.[64] These more pronounced transparency obligations for very large online platforms reflect the differentiated approach the Commission took for the design of the DSA, explicitly mentioned in Recital 39 of the proposal.[65]

2.3.2 DMA: Arts 5(g) and 6(1)g

The bottom part of Table 1 clearly shows that transparency duties in the DMA are more scarce than in the DSA and mostly relate to rankings and advertising services.[66] They are nonetheless a breakthrough in competition law, because they are ex ante policies envisaged to prevent severe hindrance to market forces from occurring. That justifies the choice to analyze them here.

The main provisions of interest are Arts 5(g) and 6(1)g DMA, especially if read in combination with Recitals 42 and 53. Art 5(g) DMA would oblige gatekeepers, with respect to their core platform services (within the meaning of Art 3(7) DMA), to 'provide advertisers and publishers ..., upon their request, with information concerning the price paid by the advertiser and publisher, as well as the amount or remuneration paid to the publisher'.[67]

Furthermore, advertisers and publishers can request, and obtain free of charge access to performance measuring tools and the information that is needed to perform their own verification to assess how satisfied they are with the advertisement product they are paying for (Art 6(1)g DMA).

While these obligations are rather specific, Art 10 DMA would open the door to add further transparency duties in the future if a market investigation pursuant to Art 17 DMA identified a need to do so for the sake of safeguarding fair competition.

[57] Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernization of Union consumer protection rules OJ L 328, 18.12.2019, p. 7–28.

[58] Above (n 7). In Table 1 (Appendix), we specify whether the norm imposes a transparency or disclosure obligation. Here we use the two as synonyms.

[59] i.e., the number of removal orders received from Member States, categorized by the type of illegal content and the average time required to remove such content; the amount of notice submitted pursuant to Art 14, any action taken thereupon, average time needed for this action, own-initiative, content moderation measures affecting availability, visibility and accessibility of information, and the number of complaints received by the internal complaint system (Art 17 DSA).

[60] For a discussion of the 'clearly, comprehensibly, and unambiguously' requirement in Art 10 e-Commerce Directive, see A Lodder & A Murray, EU Regulation of E-Commerce. (Edward Elgar Publishing 2017), 26. While case law on the matter is rather sparse, the ECJ clarified that information that can only be accessed by a number of clicks is still provided in a clear and comprehensible manner. *Bundesverband der Verbraucherzentralen und Verbraucherverbände - Verbraucherzentrale Bundesverband eV v Amazon EU Sàrl, Case C649/17*, 10 July 2019, para. 52.

[61] Per the thresholds chosen by the Commission for the designation of very large online platforms under Art 25(2) DSA and the relation with the different notion of gatekeeper in the DMA see nn 3 and 6 above.

[62] Recital 62 DSA.

[63] Moreover, the service recipient would have to be provided with an easily accessible functionality allowing her to select her preferred option for the recommender system the platform is using (Art 27(2) DSA).

[64] Not only do they have to publish reports every six months (instead of yearly), they also have to include a risk assessment (pursuant to Art 26 DSA), risk mitigation measures (pursuant to Art 27 DSA), audit reports (pursuant to Art 28(3) DSA), and audit implementation reports (pursuant to Art 28(4) DSA).

[65] For a thorough discussion of how differentiating rules better ensure the proportionality of regulatory intervention, see F Di Porto & N Rangone, Behavioural Sciences in Practice: Lessons for EU Policymakers. In A Alemanno and A Sibony (eds) Nudge and the Law, (Hart pub 2014) 20-59. With reference to transparency duties, Di Porto and Maggiolino (n 12) 12-22. See also CERRE DSA report (n 41) 11.

[66] Note that we are focusing on general informational duties, not those which only apply if there is an investigation underway (see Art 19 DMA).

[67] This is a self-enforcing obligation for gatekeepers vis-à-vis advertisers and publishers to which they provide advertising services. Gatekeepers should inform about the price paid their counterparts as well as the remuneration paid to the publisher for the publishing of an ad and for the advertising services provider by the same gatekeeper. Such transparency duty, as clarified in Recital 42, is needed for the parties to better understand the real value of the service provided.

To sum up, this section has shown that despite the many criticisms, transparency duties loom large in the DSA and DMA proposals. By analyzing in greater detail the actual disclosure duties of the two acts, we provided evidence of the way the Commission seeks to attain a high level of consumer protection and fair competition for digital services.

The analysis shows a stark contrast between what most commentators critique regarding the utility to enact more transparency duties and what the proposals purport. That suggests exploring other and new research routes to understand how these duties were implemented in the DSA and DMA proposals.

3. A Computational Analysis of The DSA and DMA Consultation Process

In this section, we ask whether informational duties are what stakeholders asked for in the consultation process and whether their actual wording in the DSA and DMA reflects the way each group uses the relevant terms. This is a relevant step, as it is important that those who implement disclosure duties (typically digital firms, be they small, medium or large) and the beneficiaries of information (individuals, but also micro-organizations) agree on the meaning of the duties (e.g., 'clear', 'accessible', or 'unambiguous language').

To do so, we leverage the power of ML and computational text analysis techniques. In the following, we present our empirical analysis of the replies and position papers submitted by stakeholders to the EU consultation process for three inception IAs. We first give a high-level description of our methodology (for a more detailed description, see Appendix),[68] before presenting our results.

3.1 Our Methodology

We collected and analyzed a total of 2,862 replies to the questionnaires and 1,862 of the respective feedback documents attached to the replies.[69] In total, we built a dataset of 3,032,418 words. To do so, we automatically downloaded all the relevant files from the Commission's website.[70] Unlike the replies (in excel), most attached submissions came in PDF format, so we first converted them into text and then constructed three large clusters.

3.1.1 Groups Identification

To identify groups of stakeholders, we relied on the Commission's categorization scheme for the organization 'size' of the feedback contributors, which groups feedback comments from (1) individuals, micro (10 employees), (2) small (50 employees), (3) medium (250 employees), and (4) large (250 or more) organizations.[71] We then aggregated the different sub-categories (3) and (4) to form three larger categories:

A. individuals and micro firms/organizations;

B. small firms/organizations; and

C. medium and big firms/organizations.

As explained in the previous paragraph, the initial clusters were based on European Commission's 'size' division. From that clustering, we aggregated medium and big firms, as suggested by: (1) the cluster size, and (2) a Kolmogorov-Smirnov test performed on the questionnaires accompanying the consultation (further explained in the Appendix).

Neither the size of companies nor the questionnaire answers we chose to perform the K-S test on were re-used for the Word Embedding Modeling (see below, A.2), hence avoiding double-dipping.

Our decision on *how* to do this aggregation was based on a qualitative and quantitative analysis of the questionnaire accompanying the feedback documents.[72]

This allowed us to find out which groups of consultation participants are the most similar and should be clustered together. Note that a Kolmogorov-Smirnov test we performed on the categorical (i.e., multiple-choice) questions in the questionnaire showed that 'medium and large' entities should be grouped together as they can be assumed to be one cluster.[73] This is per se a relevant finding, because although different in size, and despite the fact that in most economic surveys they are considered separately, medium and large entities are a cluster for the purpose of text analysis. That is justified by both qualitative and quantitative factors.

First, our algorithm assessed replies provided by firms *and* organizations together, while in economic surveys just *firms* are grouped in one cluster. It is therefore possible that the presence of organizations attenuated the distance in the use of terms.

Second, that is extremely relevant because even if medium and large entities decide through different mechanisms (e.g., taking a decision may involve only one manager in medium organizations, while requiring dozens in big ones), what we assess is the way they understand and use terms related to transparency duties. Hence, the size of

[68] The methods we used and describe hereafter largely overlap with those described in F. Di Porto et al., I see something you don't see. A computational analysis of the DSA and the DMA, appeared in (2021) Stanford Computational Antitrust, (1)6. However, there we focused our analysis on terms related to competition in digital markets and used the theoretical legal framework typical of antitrust law. In this paper, we deploy algorithms on informational duties proposed by the DSA and DMA and use theories of regulation to interpret the results of our computational analysis.

[69] Note that the replies were used partially: we only employed those drafted in English and related with disclosure terms (we manually coded these: see Appendix for further details).

[70] All the documents we used can be found under the following links. As per the DSA proposal: European Commission, Digital Services Act – deepening the internal market and clarifying responsibilities for digital services, 11 January 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-Internal-Market-and-clarifying-responsibilities-for-digital-services (accessed 28 January 2021) As per what became the DMA proposal: European Commission, Digital Services Act package – ex ante regulatory instrument of very large online platforms acting as gatekeepers, 11 January 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers; and European Commission, Single Market – new complementary tool to strengthen competition enforcement, 11 January 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12416-New-competition-tool.

[71] The Commission distinguishes the feedback also by 'types' of contributors. E.g respondents to the DSA were: the general public (66%), companies/businesses organizations (7.4%), business associations (6%), and NGOs (5.6%) authorities (2.2%), academic/research institutions (1.2%), trade unions (0.9%), and consumer/environmental organizations (0.4%) (see DSA at 8).

[72] See European Commission (n 57) for the questionnaire. A detailed description of how we analyzed the questionnaire can be found in the Appendix.

[73] This choice can not only be backed by our data, but also by some scholarly findings, e.g., R Kemp & C Lutz,Perceived barriers to entry: Are there any differences between small, medium-sized and large companies *International Journal of Entrepreneurship and Small Business,* (2006) *3(5), 538–553.* For a more detailed description as to why we cumulated medium and large entities, instead of clustering medium with small ones, see the Appendix.

organizations is not a relevant parameter, as it is semantics.

Third, by analyzing the text of organizations' opinions, as formalized in the feedback documents and replies, and later encapsulated in the DMA and DSA informational rules, we are able to capture how medium and large entities make use of terms related to transparency.

3.1.2 Word Embedding Modelling: Training the Algorithm

After having identified the most sensible way to cluster the consultation documents, we built three corpora:

- 744 documents with 35,949 unique words for corpus A (*Individuals and micro enterprises and organizations),*
- *393 documents with 32,100 unique words for corpus B (small companies/organizations),*
- *and 689 documents with 39,815 unique words for corpus C (medium and large companies/organizations)*.

We always compared two corpora, hence we analyzed three corpus pairs (A-B, B-C, A-C).

By constructing three different corpora, we were able to train a neural network on the documents of each cluster, hence having three networks that capture the intricacies of each corpus. Based on the number of times words occur next to each other, this network allowed us to calculate a vector for each word in each corpus, a so-called Word Embedding Model (more specifically, we used Gensim's CBOW word2vec model).[74] These models are remarkable in the sense that they can capture the semantic meaning of words in a set of numbers. For instance, in a well-trained model, the distance between the vector of the words 'Paris' and 'France' will be roughly the same as between 'Rome' and 'Italy'. Hence, the relative positions of vectors in the model approximately represent the meaning of certain terms. This means that while a simple algorithm would require researchers to formulate explicit rules to approximate the semantic meanings of words, ML (or the neural network, to be precise) learns the implicit rules directly from the data we feed it. This does not only increase the performance of the algorithm but also prevents an undue influence of the researchers' conscious or subconscious assumptions.[75]

3.1.3 Making sense of semantic distance

However, it needs to be noted that models trained on different corpora are not directly comparable. Since the vectors making up the models are based on the frequency of words occurring next to each other, they depend on the corpus the model was trained on. Hence, even the position of words that most definitely have the same meaning for all groups (e.g., 'and') will have very different vectors, which we would normally interpret as a semantic difference. In this case, however, the distance between the two vectors will not be the result of a different use of a word, but simply the particularities of the corpuses the model was trained on. Consequently, to make the models we trained on the different corpuses comparable, we used unsupervised vector space alignment. This allowed us to bring the vectors trained on two different corpuses together in one model space, where they would be comparable. Put differently, in the aligned model space, strongly differing vectors represent actual differences in the use of a word, instead of being a result of a different training basis.

However, we still needed to ascertain that these differences were not merely incidental, but actually of a certain significance. To do so, we employed a statistical test. This test relies on the assumption that the distance between the vectors for the same word from two different corpora can be split into three components: a semantic difference (i.e., a difference in meaning), a non-semantic difference (e.g., syntactical differences), and a random difference. We then set two assumptions: first, we assume that the semantic difference between corpora for a certain set of words (the control vocabulary) is zero. This means that we assume all stakeholder groups use words like 'and' or 'one' in the same way. Based on this, we were able to construct an empirical distribution of the non-semantic difference and the random difference, assuming that there is no semantic difference. This distribution is our second assumption.

Knowing how our vectors should look like if there was no semantic difference between the clusters, we were then able to check for each word if the distance between its vectors from two different corpora is compatible with this hypothesis of a uniform use. If it is not, we can conclude with a certain level of confidence that there is a statistically significant difference in its semantic meaning between the different corpora.

With these tools at hand, we analyzed the stakeholder submissions to the DSA and DMA consultation process. Given that the stakeholders whose opinions we analyze are to a large extent those who will either draft or receive the abundant transparency statements envisioned in the proposals,[76] their uses and view of terms related to informational duties should be of great interest both for legislators and scholars debating the factual role of informational obligations.

The questionnaires raise several points, not all of which immediately related to informational duties. For instance, the NCT questionnaire also discusses competition problems (such as agreements, self-preferencing, or collusion); while the DSA one includes questions on liability of intermediaries.

Because we are interested in the use of certain terms only, we created an initial list of 119 terms, based on the glossaries of the consultation questionnaires which explain terms that might be new to some consultation participants. However, after the first analysis, we realized that our list of terms might be too narrow for two reasons.

First, the wording of the Inception Impact Assessments (IIAs) which were discussed in the consultations differs from the final draft DSA and DMA. The change in vocabulary is especially marked in the DMA,[77] where classic concepts of competition law (such as market, dominance, efficiency gains) are mostly abandoned, and new ones are defined.[78] Since we used corpora from comments to the three IIAs

[74] T Mikolov, K Chen, G Corrado & J Dean, (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs]. R Řehůřek, R. (2019). Word2vec embeddings. https://radimrehurek.com/gensim/models/word2vec.html (accessed 22/06/2021).

[75] For instance, a researcher might assume that a word needs to be used in the same sentence at least *x times for the two to be related and design her algorithm accordingly. For our algorithm, we do not need these kinds of assumptions or rules as the algorithm learns directly from the data.*

[76] This includes the general public, authorities and consumer/environmental organizations (as addressees), and companies/businesses organizations, business associations, and trade unions (as drafters); but will exclude NGOs, and individual academics and research institutions.

[77] The difference in terminology also derives from the fact that the 'NCT' inception IA was based on Art 106 TFEU (much focused on competition), while the 'Ex-ante regulation's legal base was Art 114 TFEU (internal market). Following the consultation, the DMA proposal had its own legal base (Art 114) and terminology.

[78] As are spheres of application of the DMA in comparison with the inception IAs.

documents to run our analysis and needed it to reflect this change, we proceeded with hand-coding. Therefore, we combined words from two sources: (i) all glossaries[79] attached to previous legislation (all EU Directives and Regulations) that were recalled by the DSA and DMA proposals (for a total of 119 words); and (ii) terms related to transparency (e.g. 'disclos*', 'transparency', 'inform*' and the like) that were manually selected from the questionnaires (102 words). As a result, we ended up with a list of 194 words (102 from the DSA's questionnaire and 92 from the DMA's). (See Annex 3.1).

Furthermore, since we are interested in the specific provisions of the DSA and DMA which qualify how information should be provided (e.g., 'clear', 'accessible'), we added all those terms from the proposals' informational provisions (ten terms in total, see Annex 3.1).

Finally, stakeholders use a variety of terms to refer to the same concept. For instance, our list might include 'self-preferencing', but we would miss differences on 'self-favoring'. Our pre-defined list of terms was not able to capture this variety. Since it was also not feasible to anticipate all these variations, we chose to manually code those results that are closely related to the terms and concepts of our list *ex post*.

To perform manual coding, we relied on the legal expertise of our team, with the aid of external assistance.[80] Finally, the terms that were added manually were a total of 204, while overall the computational analysis was performed over of a total of 323 words.

3.2 Results: Different Groups, Different Uses?

We found a statistically significant difference for

1,865 word pairs between corpora A and C,

2,184 between corpora A and B and

1,113 between B and C.[81]

A detailed description of how this comparison was conducted and what 'significant' means in this context, is provided for in the Appendix (Annex 3). From all the 5,162 significant distances we found, we chose those that were relevant to our analysis, based on the selection procedure described above. This resulted in a list of 13 relevant terms

[79] Glossaries are definitions of terms usually contained in Arts. 2 of EU Directives and Regulations. Namely, we added all the glossaries from: the GDPR, the NIS Directive, the Data Governance Act proposal, the E-commerce Directive and the Platform-2-Business directive.

[80] We are thankful to Andrea Ruffo, legal scholar and teaching assistant at Luiss University of Rome for his wonderful assistance in the manual coding activities. The legal analysis was performed by Tatjana Grote and Fabiana Di Porto.

[81] It needs to be noted that many of these words are not of particular interest for us because they might identify a specific service of a certain company (e.g., the 'Gmail' email service in Google's submissions). However, some of the key buzzwords surrounding competition and transparency obligations show statistically significant differences.

Table 1: Summary of results

Term	Distance AB	Distance BC	Distance AC	Close words A	Close words B	Close words C
Consumer-centric	1.557 (0.03)**	1.625 (0.02)**	1.247 (0.16)	privacy-protecting	systems	computing
Easy	1.444 (0.04)**	1.443 (0.07)*	1.451 (0.05)*			
Easy-to-use	1.450 (0.04)**	1.427 (0.08)*	1.522 (0.02)**	deregulation		cut-off
Meaningful	0.545 (0.627)	0.670 (0.648)	1.482 (0.04)**			
Precise	1.645 (0.01)**	0.878 (0.434)	0.747 (0.497)	cartel	checklist	
Privacy-friendly			1.468 (0.04)**	misconceptions		tailor-made
Ranking	1.182 (0.15)	1.644 (0.02)**	1.452 (0.05)*		guidelines, improve, oversight	appearance, disclosing
Readable	1.051 (0.237)	1.720 (0.01)**	1.394 (0.08)*		effective, specific, clear	entities
Self-regulatory	1.340 (0.09)*	1.536 (0.04)**	0.897 (0.37)		blacklisting, sanctions, obligations	benchmarking, codes, ameliorate
Simple	1.703 (0.01)**	1.504 (0.05)*	1.158 (0.20)	formats	precise	
Understandability		1.663 (0.02)**			single-homing, practice	informs
Unregulated	1.361 (0.07)*	1.566 (0.04)**	1.822 (0.00)***	not-sufficient		mitigation
Well-informed	0.943 (0.293)	1.734 (0.01)**	1.749 (0.00)***	Confusing, explainable		Inscrutability, implementation

Note: The asterisks indicate significance at a 0.001 (***), 0.05 (**), and 0.1 (*) level, respectively.

for which we found significant differences in use and understanding.

Table 1 shows these results. The 'Distance' columns report the distance between the vectors of the same words for each corpus pair, with the respective p-value in parentheses. A grey field in the 'Distance' columns indicates that a word was not used in both of the respective corpora.

The 'Close Words' columns shine a light on some of the concepts that were closely related with the term in question in the corpora for which there was a statistically significant distance between the terms. To be precise, we computed the ten words which were most similar to the term in question[82] and then hand-coded those words which were relevant to our analysis, based on the same procedure outlined above (see the last paragraph of 3.1.2). A grey field in the 'Close words' columns means that we did not look for close words because the respective corpus was not involved in any of the significant distances or there were no meaningful close words.

Moving on to our results, we start with some terms that are of importance on a meta-level, namely those related to the overall regulatory strategy employed. Since there are different regulatory paths to ensuring transparency (e.g. by regulation or self-regulation), this is of interest as well.[83]

3.2.1 Words related to the regulatory 'meta-level'

We observe that '**self-regulatory**' is used differently by different stakeholders. Generally, we see that self-regulation seems to be a more prominent issue for medium and big companies (corpus C): while the term is only mentioned ca. 5,000 times by small companies (corpus B, with 810, 961),[84] it occurs more than 25,000 times in corpus C (which contains 1,177,120), where it is associated with the terms 'benchmarking', 'codes', and 'ameliorate'. This is reflected in Fig. 1, and could be read as a sign that self-regulation is seen as an important strategy by medium and big companies/organizations.

Differences in use also exist for the term 'unregulated'. For individuals (A) and small entities (B), an 'unregulated' digital single market does not seem like a favorable option, with 'not-sufficient' and 'precariousness' as closely related terms. (Fig. 2)

3.2.2 Words related to informational duties

With regards to *informational duties*, it is interesting to note that there is a statistically significant distance between the use of the word '**simple**' between corpus A and B (Fig. 2). While individuals and micro-businesses/organizations seem to focus on 'formats' regarding simplicity, small companies/organizations in our dataset associate the attribute '**precise**'. However, it needs to be noted that the term 'precise' also underlies some significant differences between corpora A and B, which is an important finding in light of the wording of Art. 23 DSA (Table 1).

Generally, individuals and micro-organizations (A) used the word '**simple**' roughly 20-times more often than small businesses and organizations (B).

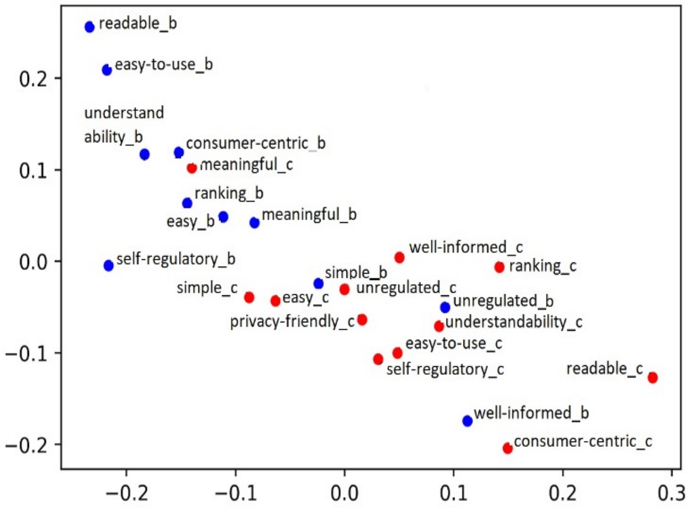


Figure 1: Aligned Vector Space Model - Corpora B & C

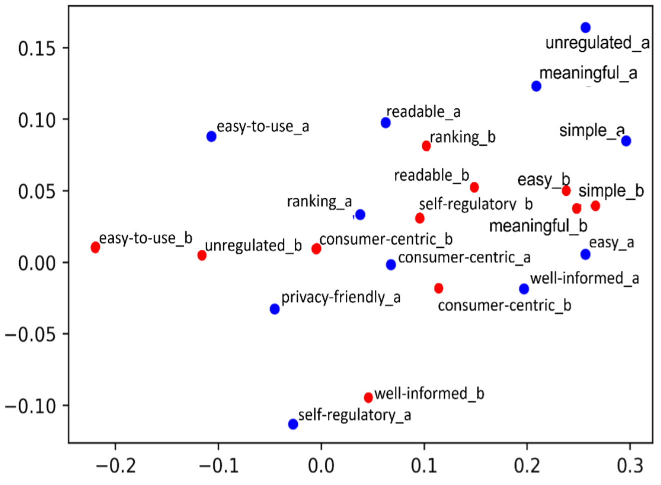


Figure 2: Aligned Vector Space Model - Corpora A & B

With regards to the obligation of advertisement system transparency laid down in Art 24 DSA, it is surprising to see that '**meaningful**' is used very differently by individuals and micro-organizations/businesses (A) than by medium and big companies (C).[85] Again, this could potentially impact the efficacy of said provision since what is deemed 'meaningful' by the drafters of the respective disclosures might be rather meaningless for their recipients.

In the comparison between corpora A and C, the term '**well-informed**' is mentioned roughly 26,000 times by individuals and micro-contributors (A; in total 1,044,337 words) compared to 18,642 mentions in corpus C (in total 1,177,120 words) and is closely related to 'explainable'. Furthermore, we find a different utilization of the terms '**easy-to-use**' and '**privacy-friendly**', respectively (see Fig. 3).

The first is interesting with a view to rules like Art 17(2) DSA, which speaks of *easy to access, user-friendly* complaint mechanisms. The latter seems to be located within slightly different contexts by different stakeholder groups: while individuals (A) heed possible 'misconceptions', medium and large companies/organizations (C) associate '**privacy-friendly**' with 'tailor-made' and 'reinforced'. Interestingly, the Commission explicitly mentions that '**privacy-friendly services**' were

[82] Our similarity measure is the cosine distance between two vectors. R Řehůřek, R. (2019). Gensim: Store and query word vectors - Similarity. https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.WordEmbeddingsKeyedVectors.similarity (accessed 30/08/2020).

[83] For a detailed discussion of regulatory strategies in disclosure regulation, see Di Porto & Zuppetta (n 14).

[84] Note that the corpus sizes indicated here refer to the overall corpus, i.e., the number of words in the documents as they were submitted. For corpus sizes indicated above we only considered the unique words for each corpus, which is why these numbers are much smaller.

[85] The use of 'Meaningful' for the corpus pair C and A might look close in Fig. 3 because the difference is not as pronounced as for some other terms, but it has p-value of 0.04, meaning that we can conclude there is a statistically significant difference.

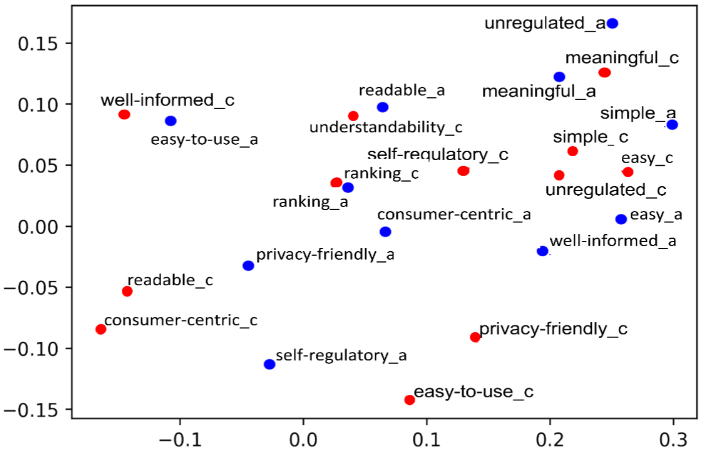


Figure 3: Aligned Vector Space Model - Corpuses A & C

one key expected outcome of the DMA in the eyes of the consultation respondents. However, what might be missing is that not all stakeholders understand the same when speaking of 'privacy-friendly'.

Comparing small companies/organizations (B) and medium/big companies/organizations (C), we find a significant distance between the vectors for the terms '**well-informed**' and '**consumer-centric**' (Fig. 1, above). The latter word is closely related to the term 'systems' in corpus B, which is unsurprising. In corpus C, we see a close association with 'computing', which is interesting since it seems to shift the focus of consumer-centric design to the processes happening behind the systems that consumers interact with.

Another intricate finding concerns the term '**ranking**', which has been central in discussions about the transparency of online platforms. This close connection between transparency and rankings is also reflected in the close words we found: small companies (B) associate rankings with '*guidelines*', medium/big companies with '*disclosing*'.

As 'ranking' is not a crucial term for transparency duties as such, this difference will not necessarily impede the effectiveness of disclosures. Nevertheless, this finding shows that there are different perceptions of some key concepts of the DSA and DMA across stakeholders.

We further find differences for the terms '**understandability**' and '**readable**'. This should be a key concern for policymakers and legal scholars when debating transparency duties: if no uniform understanding of what 'readable' transparency disclosures look like can be reached, consumers will likely have to deal with strongly differing levels of readability and understandability.

3.3 Challenges

Our algorithmic analysis of the consultation process for the DSA and DMA has shown that there are statistically significant differences between stakeholders' use and understandings of some key concepts of transparency. To the best of our knowledge, we are the first to conduct such a 'close reading' of an EU rulemaking process and discern differences in the ways a consultation relates to the rules in the context of the DSA and DMA. Our results show that NLP techniques can allow the Commission to understand not only what stakeholders say, but what they actually mean; which could substantially improve stakeholder consultations' analysis as we did here. For instance, the Commission took note of demands for more 'simple' notice-and-action procedures for content removal.[86] Yet, we discovered that the term 'simple' might not be understood in the same way across all groups of stakeholders. This could offer a first signal to the Commission that it is premature to legislate on this matter; or that a one-size-fits-all measure may not be suitable.

Linking our results back to the discussion of transparency duties and their importance for consumer protection in digital markets, our findings cast doubt on whether all stakeholders have a similar understanding and thus make similar uses of **simple**, **meaningful**, **easy-to-understand**, **readable** transparency statements. Given that the exact implementation of such duties often lies in the hands of different stakeholders, this might be one reason why transparency duties remain ineffective. For instance, our algorithm reveals that 'meaningful' is understood and used differently by the individual consumers and the medium/big platforms. This may cause Art. 24 DSA failure, as it obliges platforms to inform consumers in real-time that what is being displayed to them is an ad, in a clear and '*unambiguous* manner'. Since the literature on the failure of disclosure regulation has mostly focused on how transparency statements are perceived by consumers,[87] our focus on all stakeholders, inclusive both the recipients and drafters of disclosure statements, adds a unique, novel perspective.

Having said that, there are challenges that need to be addressed, some of which are common to the computational law scholarship,[88] others are specific to our analysis. Both offer room for improvement by future research.[89]

Concerning the analysis, in the methodology, we make two assumptions for the statistical test we perform: that words in the control vocabulary used for the vector space alignment transformation do not have a semantic difference and that the distribution of distances has the same shape also for the other words. For instance, we assume that words like 'and' or 'one' are understood in the same way by all contributors in the consultation. While this seems plausible, we cannot entirely discard the possibility of errors in the creation of the models and their alignment due to shortcomings in these assumptions. Nonetheless, our assumptions are commonly accepted in the literature.[90]

Second, our corpora are relatively small and heterogeneous since they contain documents from many different authors with potentially different styles and focuses. For instance, feedback we analyzed are in English language only, but their authors might not be native English speakers. This could introduce a bias, meaning that results may be partially driven by the particularities of our corpora. Hence, increasing the corpus size and the control vocabulary should be a top priority for future research. Another way to solve the problem would be using bootstrapping: by repeatedly and randomly changing some words in the corpora and then taking the mean value, the random term u_t^{AB} in the distribution of distances could be reduced.

Generally, it needs to be noted that our analysis focuses on the *identification* of semantically different terms. At this stage, we do not seek to provide insights into what the identified differences might be based on and how they impact the stakeholders' opinions. Therefore, it has some limitations as far as *interpretation* is concerned. Using word embedding alignment alone does not allow (yet) to show any causal relationship between differences in perceptions of transpar-

[86] DSA proposal, 8.

[87] Above (n 11).

[88] D Lim, Can Computational Antitrust Succeed? *Stanford Computational Antitrust,* https://law.stanford.edu/wp-content/uploads/2021/04/lim-computational-antitrust-project.pdf (accessed 22/06/2021), 10-13.

[89] More technical limitations are presented in the Appendix.

[90] See Nyarko and Sanga (n 25), 4.

ency and specific factors. Although we compared the most similar vectors[91] corresponding to the word pairs of interest, gaining an idea of how the meanings might differ, this still requires a certain degree of *ad hoc* interpretation. Moreover, we used ex post manual coding when selecting the results to be presented here. In the future, fully replicable, *ex ante* criteria should be used to make this selection.

Due to these limitations, our results need to be treated with caution and should be complemented by further research. Nevertheless, they constitute a first step providing interesting insights into informational duties in the DMA and DSA.

4. Concluding Remarks

This paper sets out to explore whether different stakeholders participating in the consultation process for the latest Commission proposals on new rules for digital markets (the DSA and DMA) share a similar understanding of key concepts related to one integral pillar of the new proposals: informational duties. We analyzed the replies to questionnaires and feedback documents submitted in the consultation process using the NLP technique of Word Embedding Alignment, which allowed us to identify terms that are not used in the same way by all stakeholders.

We find significant differences in the way stakeholders use words that are central in transparency duties, like 'readable', 'simple', and 'privacy-friendly'. These differences are group-specific, and hold for individuals and micro organizations; small; and medium/large organizations. If that might seem obvious at first sight, it is surprising if one considers that those participating in the consultation process on the DSA and DMA constitute a rather small epistemic community, made of legal and economic scholars, digital companies, NGOs, and IP specialists who have a high stake interest in expressing their voice and are, therefore, well-informed about the subject they discuss.

Our results should be a key concern for policymakers and legal scholars for several reasons. Differences in understanding might mean (undesirable) differences in implementation. If there is no uniform use (and understanding) of what 'readable' transparency disclosures or 'simple' complaint mechanisms look like, users will likely have to deal with strongly differing levels of readability and simplicity.

Second, this could decrease the effectiveness of transparency duties in ensuring competitive and fair markets, given that those who replied to the consultation are also those who will draft and receive the disclosures.

Third, and strictly related, different understanding and uses of words that are relevant to informational duties might also help explain why such rules fail.

The last takeaway we want to stress is that rule-makers are recommended to consider another interesting finding: that understanding and use of relevant terms of transparency (like 'simple' and 'well-informed') do not differ between medium and big organizations (corpus C), as one would expect. That is to the point to make them a sole group for the sake of text analysis. Generally, if the Commission used tools like the one applied here to complement its impact assessments and rulemaking, it could not only hear what stakeholders *say* but understand what they *mean*, which might ultimately improve the functioning of the EU's new regulatory traffic lights for digital markets.

Looking at the perspectives this paper opens, we think that our analysis, if complemented with other computational techniques, will be very useful in doctrinal studies of the future.

One scenario could be to investigate the 'rationale' of the DSA and DMA's rules. By the time the DSA and DMA will entry into force, their wording will change several times, depending on multiple interactions of the Commission, the Parliament, Council and stakeholders. Our analysis might be a first step in the direction of keeping records of textual modifications and then tracing back the statements that influenced them the most (e.g., being the most similar). Clearly, our analysis alone would not be enough and would need to be complemented with other NLP techniques. For example, text similarity techniques could be employed to map out which stakeholder opinions might have influenced the EU institutions when drafting not only its proposals but also its final rules. This might allow gaining a precise understanding of why rules were drafted in a certain way and could greatly help the interpretation of rules in light of their *telos* and their drafting history.

A second research area that our analysis could inaugurate is that of improving the drafting of disclosure statements and transparency reports, as envisaged by the two new proposals. While we considered the use and understanding of information-related terms by firms and organizations together, one could zoom in on the use of concepts by individual consumers and firms only, which will certainly differ. For instance, the phrase 'easy to use' was used differently by all three clusters. If we already find this disagreement in large, aggregated groups, the understanding of such a phrase will most likely differ between individuals. Consequently, regulators might opt for clusterized disclosures, with messages adapted to the specific informational capabilities of users' groups (as identified by our computational analysis).

That might help to overcome many of the shortcomings of current disclosure statements. While this possibility was discussed in great detail elsewhere,[92] our analysis suggests that the Commission and platforms would be well-advised to explore this possibility.

Our algorithm should be seen as the first building block of a fully-fledged tool for a more in-depth algorithmic analysis of EU rulemaking. The other building blocks might be:

- 'topic modeling',[93] which would allow rule-makers like the Commission and scholars to get an intuitive understanding of how the most important topics, that will become rules in a near future, are part of a shared view among different stakeholders or whether they emphasize different issues;
- 'document similarity'[94] could be used to cluster statements that are input to regulation before the Commission publishes a regulatory proposal. This could help to perceive certain similarities or alliances, between stakeholders, even across different groups like

[91] See n 82 above.

[92] See, e.g., F Di Porto, Algorithmic Disclosure Rules, in Artificial Intelligence and Law, (2020), https://ssrn.com/abstract=3705967 or http://dx.doi.org/10.2139/ssrn.3705967 (accessed 27 October 2021). More information on the implementation of clusterized disclosures is available at: www.lawandtechnology.it. See also: Busch, C. (2019). Implementing Personalized Law: Personalized Disclosures in Consumer Law and Data Privacy Law. *The University of Chicago Law Review* 86(2), 309–332.

[93] DM Blei, AY Ng & MI Jordan, Latent dirichlet allocation. *The Journal of Machine Learning Research*, (2003) 3, 993–1022.

[94] See, e.g., BK Triwijoyo & K Kartarina, Analysis of Document Clustering based on Cosine Similarity and K-Main Algorithms. *Journal of Information Systems and Informatics*, (2019) *1(2), 164–177.* DG annemann, Comparative Law: Study of Similarities or Differences?, in M. Reimann and R. Zimmermann (eds.), *Oxford Handbook of Comparative Law* (2d ed.) (Oxford University Press, 2019).

e.g., small companies and medium/large companies.

- Sentiment Analysis could be another means to understand if the parties to a rulemaking process agree or disagree with certain proposals or statements. In fact, we performed a first explorative sentiment analysis using a pre-trained model on those paragraphs in our documents which contain the terms of interest presented above (Table 1). While this analysis produced some interesting results,[95] a fully-developed sentiment analysis is best left for future research. Furthermore, one could cluster each statement based on the overall sentiment of a group of contributors[96] to get a better understanding of how supporters and critics of a proposal are distributed and what their main concerns and arguments are.

Overall, while we believe that discerning latent differences in the use of certain terms is a crucial capability that could significantly enhance the consultation process at the EU level, the above-mentioned additions could be combined in a fully-fledged NLP toolbox that could substantially enrich the work of both the Commission and legal scholars and provide many new insights.

Be that as it may, it is hoped that our findings will enrich the positive and normative debate about transparency rules in digital markets, inspire future research in the computational antitrust arena, and urge EU rule-makers to rethink their convictions about the use of computational tools in the consultations.

[95] For instance, we found that ‘understandability’ is seen much more favorably by small companies/organizations (B; 0.721) than by medium/big entities (C; 0.340). Similarly, we found a more positive attitude towards the terms ‘well-informed’ and ‘consumer-centric’ for individual and micro contributors (0.624) than for small companies/organizations (0.051). We also identified a negative sentiment of small companies/organizations towards the term ‘unregulated’ (-0.118). Lastly, ‘simple’ is viewed more favorably by individuals and micro contributors (A; 0.314) than by big and medium organizations/businesses (C; 0.220).

[96] See e.g., S Feng, D Wang, G Yu, C Yang & N Yang, Sentiment Clustering: A Novel Method to Explore in the Blogosphere. In Q Li, L Feng, J Pei, SX Wang, X Zhou, & QM Zhu (Eds.), *Advances in Data and Web Management.* (Springer 2009) 332–344.

Appendix

Table 1 Informational duties in the DMA and DSA

T / D duty	Digital Services Act (DSA)	Recipient of info (r) Info to be provided (i)	'How' to disclose	Core service providers (Art 2(f) DSA)	Online platforms (Art 2(h) DSA)	Very Large online platforms (Art 25)
D	Terms of service include information on content moderation and use of algorithms	(r) Users; (i) potential restrictions to their services.	'easily accessible format' written in 'clear unambiguous language'	Art 12 (Terms and conditions)		
T	Yearly reports on content moderation providing key information specified in Art 13(1) DSA	(r) Users and the general public; (i) content moderation practices	written in 'clear and comprehensible language'; need to include specific information (a. 14, 17)	Art 13 (Transparency reporting obligations for providers of intermediary services)		
D	Reasons for removing the content or disabling access	(r) Users whose content was removed or access disabled	Clear and specific statement containing the information listed in Art 15(2)		Art 15 (Statement of reasons)	
T	Additional information (with reference to Art. 13) on content suspension actions taken, use of automated means for content moderation, and out-of-court dispute settlement	(r) Users and the general public, (i) esp. about automation of content moderation and ADR	Format potentially to be specified by Commission, Art 23(4)		Art 23 (Transparency reporting obligations for providers of online platforms)	
T/D	Advertising transparency duties	(r) Users and recipients of service; (i) display that info is an ad + personalization of ad	Provided in a 'clear and unambiguous manner'		Art 24 (Online advertising transparency)	
D	Main parameters used in recommender systems must be set out in terms and conditions	(r) Users; (i) use of algorithms for recommending content	Provided in a clear, accessible, and easily comprehensible manner	/	Art 29 (Recommender Systems)	
T	Additional advertisement transparency duties to maintain in the repository and made accessible	(r) Users and the general public; (i) advertisements and their display	Repository be made publicly available through an API		Art 30 (Additional online advertising transparency)	
T	Additional information on content moderation, risk management, and auditing	(r) Users, the general public, and Digital Service Coordinator; (i) results of risk assessments and audits	-		Art 33 (Transparency reporting obligations)	
	Digital Markets Act (DMA)	Recipient of info	'How' to disclose	Gatekeepers (as defined in Art 3 DMA)		
D	Information about advertising services provided by gatekeepers for advertisers and publishers	(r) Advertisers and publishers counter-parts	-	Art 5(g) (Obligations for gatekeepers)		
D	Provide free of charge access to performance measuring tools of gatekeepers and information necessary to enable advertisers to carry our independent verification	(r) Advertisers and publishers	-	Art 6(g) (Obligations for gatekeepers susceptible of being further specified)		

Note: Informational duties (Column 1) may include either transparency duties (T) or disclosure duties (D).

Annex 1 Groups identification

To analyze the replies to questionnaires and feedback documents, we created a special scraper algorithm, which allowed us to download all the files automatically, convert them into text, and split them into three clusters. In doing this, we started by following the Commission's categorization scheme for the organization size of the feedback contributors. We then aggregated the different sub-categories into three corpora based on the typology and the dimension of the feedback contributor: Corpus A (individuals and micro organizations), B (small companies/organizations), and C (medium and large companies/organizations).

Our clustering choice is based on two considerations: First, a qualitative analysis of the questionnaires accompanying the feedback documents[97] allowed us to get an understanding of which aggregation would cluster comparable feedback contributors together. We mostly analyzed the types of feedback contributors in the sample and had a look at their replies to questions related to informational duties. Second, we conducted a quantitative analysis of the same questionnaires to ensure that our clusterization choices are solid. In particular, we sought to ensure that there is no statistically significant difference between medium and large entities in our sample since at least medium companies are often grouped with small, rather than large companies.[98] However, it needs to be noted that our feedback contributors are not only businesses but also other types of organizations. This diversity could "smooth" the differences we would have expected to find if our sample included companies only. In fact, our qualitative analysis of the questionnaires suggested that medium entities in our sample are more comparable to large businesses/organizations both in terms of entity type (whether they are from academia, civil society, private economy, etc.) and in terms of how they perceive challenges arising from digital markets (in the sense that they gave more similar answers to the pertinent multiple-choice questions in the questionnaires).[99] To test the robustness of this perception, we analyzed the answers provided for by medium and large entities to specific multiple choices questions.[100] We applied a Kolmogorov-Smirnov two-sample test[101] to understand if there is a statistically significant discrepancy between the distribution of the answers of the two groups. If that was the case, we would assume that these answers must be considered as provided by two different populations, not allowing us to treat them as a unique cluster. The results of the test are shown in Figure 1.

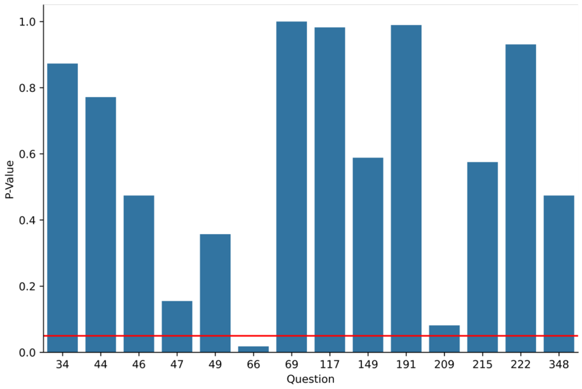


Figure 1: p-values resulting from KS-two sample test applied to the answer distributions of the considered questions. Red line highlights our significative tolerance value of p=0.05

Even using a very high tolerance **p-value** level of 0.05, only question no. 66 showed a statistically significant variation. This question alone however is mostly unrelated to our core research interest, and hence unlikely to compromise the validity of our clustering.

In total, we collected 744 documents with 35.949 words for corpus A, 393 documents with 32.100 words for corpus B, and 689 documents with 39.815 words for corpus C. We always compared two corpora, hence we analyzed three corpus pairs (A-B, B-C, A-C).

Annex 2 Training the algorithm

To discern differences in the use of certain key terms across stakeholder groups (i.e., a different semantic understanding of identical terms), we leveraged Word Embedding Models to quantify evidence of such differing understandings. This technique has already been used in various Natural Language Processing tasks, and recently also in the Computational Law literature.[102] It has been demonstrated to be very powerful and useful in providing insights into latent differences in how language is used.

The core of this technique consists in training a special neural network to convert each word contained in a corpus of texts into a vector, i.e., a set of numbers.[103] While a simple algorithm would require researchers to formulate explicit rules to somehow approximate the semantic meanings of words, ML (or the neural network, to be precise) learns the implicit rules directly from the data we feed it. This does not only increase the performance of the algorithm but also

[97] European Commission, Digital Services Act – deepening the internal market and clarifying responsibilities for digital services, 11 January 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers/public-consultation (accessed 28 January 2021).

[98] Statistically significant refers to the hypothesis of the K-S test, that the data of both groups is originating from the same population.

[99] While this could be due to the idiosyncrasy of our sample, this finding also corresponds with scholarly literature. See e.g., R Kemp & C Lutz. Perceived barriers to entry: Are there any differences between small, medium-sized and large companies. *International Journal of Entrepreneurship and Small Business*, 2006 3(5), 538–553.

[100] The questions were selected manually based on two criteria: First, we manually identified all questions relating to informational duties and competition in digital markets. In a second step, we singled out questions that had a categorical answer scale, i.e., non-text replies.

[101] L Hoboes Jr. The significance probability of the Smirnov two-sample test. *Matematica* 1958 3(5), 469-486.

[102] See e.g., Nyarko and Sanga (n 25); E Peramo, C Cheng & M Cordel, Juris2vec: Building Word Embeddings from Philippine Jurisprudence. *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 121–125; I Chalkidis & D Kampas, Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 2019 27(2), 171–198; A Mandal, K Ghosh, S Ghosh, S & S Mandal, Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law*, 2021 29(1):1-35.

[103] The Neural Network in particular is a LSTM (Long-Short Term Memory Network). See S Hochreiter & J Schmidhuber, Long Short-term Memory. *Neural Computation* 1997 9(8):1735-80. More generally, see S Lai, K Liu, S He & J Zhao, How to Generate a Good Word Embedding. *IEEE Intelligent Systems*, 2016 31(6), 5–14; Y Li & T Yang, Word Embedding for Understanding Natural Language: A Survey. In S. Srinivasan (Eds.), *Guide to Big Data Applications*. Springer International Publishing, 2018 83–104.

prevents an undue influence of the researchers' conscious or subconscious assumptions. The resulting vectors are based on the frequency of words occurring next to each other, meaning their relative positions in each phrase of the corpus and the correlation between words. The stronger two words are correlated (in their occurrence – and so in their semantic meaning)[104] in the corpus the model was trained in, the closer the corresponding vectors will be located to each other.

However, the meaning of the vectors in the model depends on their relative positions in the respective corpus; the vector of a single word alone does not give us any insights. To test if there is evidence of different semantic use of the same words between two texts, we had to assess the distance between vectors from the two different corpora corresponding to the same words. To align them, we transformed the two models geometrically.[105] This allows us to understand how a vector in one corpus relates to the vector of another corpus. After the transformation, the vectors of the two aligned corpora are comparable to each other.

For each corpus we trained a different word embedded space, and we aligned each pair of words occurring in both corpora through the means of Unsupervised Vector Space Alignment.[106]

Annex 3 Making sense of semantic distance

3.1 The Data

1. List of terms from glossaries[107]

E-commerce directive, P2B regulation, glossary of terms for DSA' questionnaire:

1. Application Programming Interface
2. Collaborative Economy Platform
3. Competent Authorities
4. Content Provider
5. Digital Service
6. Harmful Behaviours
7. Activities Online
8. Hosting Service Provider
9. Information Society Service
10. Illegal Content
11. Illegal Goods
12. Illegal Hate Speech
13. Intermediary Service
14. Intermediation Services
15. Law Enforcement Authorities
16. Notice
17. Notice Provider
18. Online Advertising
19. Online Platforms
20. Online Platform Ecosystems
21. Recommender Systems
22. Scaleup, Smart Contracts
23. Start-up
24. Trusted Flagger
25. User
26. Gatekeeper
27. Core Platform Service
28. Digital Sector
29. Online Intermediation Services
30. Online Search Engine
31. Online Social Networking Service
32. Video-Sharing Platform Service
33. Number-Independent Interpersonal Communications Service
34. Operating System
35. Cloud Computing Services
36. Software Application Stores
37. Software Application
38. Ancillary Service
39. Identification Service
40. End User
41. Business User
42. Ranking, Data
43. Personal Data
44. Non-Personal Data
45. Undertaking
46. Control
47. Recipient
48. Consumer
49. Offer Services
50. Trader
51. Intermediary Service
52. Illegal Content
53. Dissemination
54. Distance Contract
55. Online Interface
56. Digital Services Coordinator Of Establishment
57. Digital Services Coordinator Of Destination
58. Advertisement, Recommender System
59. Content Moderation
60. Terms And Conditions
61. Service Provider
62. Established Service Provider
63. Commercial Communication
64. Regulated Profession
65. Coordinated Field
66. Business User
67. Provider
68. Corporate Website User
69. Ranking
70. Mediation
71. Durable Medium

[104] This is based on the 'distributional hypothesis', which assum es that words which frequently occur together are usually also semantically related. While this approach might seem too simple to capture complex semantic meanings, the success of algorithms relying on it suggests that the claim has some merit. E Altszyler, M Sigman, S Ribeiro & DF Slezak, Comparative study of LSA vs Word2vec embeddings in small corpora: A case study in dreams database. *Consciousness and Cognition* 2017 56, 178–187.

[105] To perform this transformation, we used a "control vocabulary", containing a list of words that we can safely assume that share the same semantical meaning . The list of 1,189 words we used is, in fact, composed mainly of numbers and stop-words (like e.g., 'the'). We are thankful to Professor Julian Nyarko from Stanford University for providing us with a first list of Control keywords, to which we further added almost 2000 numerals and stop-words from the different corpuses.

[106] We used a special algorithm provided by Facebook in the library FastText. (https://github.com/facebookresearch/fastText), used in Python. P Bojanowski, E Grave, A Joulin, & T Mikolov,. Enriching Word Vectors with Subword Information, 2017. http://arxiv.org/abs/1607.04606 (accessed 22 January 2021).

[107] Terms gathered from glossaries attached to all legislation recalled by the DSA and DMA proposals plus terms taken from the glossary attached to the DSA questionnaire.

From DGA proposal:

72. Access
73. Re-Use
74. Metadata
75. Data Altruism
76. Data User
77. Data Holder
78. Data Sharing Main Establishment
79. Public Sector Body
80. Bodies Governed by Public Law
81. Public Undertaking
82. Secure Processing Environment
83. Representative

From NIS (Network and Information Systems):[108]

84. Network And Information System
85. Security Of Network And Information Systems
86. National Strategy On The Security Of Network And Information Systems
87. Operator Of Essential Services
88. Digital Service Provider
89. Incident
90. Incident Handling
91. Risk
92. Standard
93. Specification
94. Internet Exchange Point (IXP)
95. Domain Name System (DNS)
96. DNS Service Provider
97. Top-Level Domain Name Registry
98. Online Marketplace

From GDPR:

99. Processing
100. Restriction Of Processing
101. Profiling
102. Pseudonymisation
103. Filing System
104. Controller
105. Processor
106. Third Party
107. Consent
108. Personal Data Breach
109. Genetic Data
110. Biometric Data
111. Data Concerning Health
112. Enterprise
113. Group Of Undertakings
114. Binding Corporate Rules
115. Supervisory Authority
116. Supervisory Authority Concerned
117. Cross-Border Processing
118. Relevant And Reasoned Objection
119. International Organisation

II. Manually coded from the questionnaires on DSA and DMA

Manually coded from Questionnaire for the public consultation on a New Competition Tool

1. Access to data
2. adjacent/neighbouring markets
3. aftermarket
4. algorithm-based technological solutions
5. alignment of prices
6. anti-competitive
7. appropriateness
8. barriers to enter
9. binding
10. case-by-case
11. choice
12. competition
13. concentrated market
14. conditions of competition
15. copyright
16. customer lock-in
17. customer switching costs
18. data accumulation
19. data dependency
20. digital markets
21. digitisation
22. dominance-based
23. dominant
24. dual role situations
25. economies of scale
26. economies of scope
27. extreme economies of scale
28. fixed operating costs
29. gatekeeper
30. global distribution footprint
31. homogeneity of products
32. incomplete or misleading information
33. increased transparency
34. incumbency advantages
35. incumbency advantages
36. information asymmetry
37. innovation
38. inspections
39. interim measures
40. investigative powers
41. judicial review
42. lack of access to data
43. lack of competition
44. lack of transparency
45. leveraging
46. lock-in effects
47. market concentration
48. market dominance
49. market entry
50. market player
51. market power
52. market share
53. market-sharing cartels
54. monopolisation
55. multi-homing
56. multi-sided markets

[108] EU rules on the security of Network and Information Systems (NIS) are at the core of the Single Market for cybersecurity. The Commission proposes to reform these rules under a revised NIS Directive to increase the level of cyber resilience of all relevant sectors, public and private, that perform an important function for the economy and society. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX-:32016L1148&from=EN

57. network effects
58. new competition tool
59. non-binding recommendation
60. oligopolist
61. oligopolistic market structures
62. oligopoly
63. online platform
64. patents
65. penalties
66. platform
67. policy options
68. price increases
69. price leader
70. price leader-follower behavior/behaviour
71. price-fixing
72. pricing algorithms
73. procedural safeguards
74. proportionality
75. recommendations
76. regulatory barriers
77. related market
78. request of information
79. single-home
80. start-up costs
81. structural lack of competition problem
82. structural risk for competition
83. switching
84. tacit collusion
85. tailored remedies
86. tipping
87. tipping markets
88. transparency
89. two-sided markets
90. vertical integration
91. voluntary commitments
92. zero-pricing

Terms manually coded from **DSA questionnaire**

1. accountability
2. advertisement
3. algorithmic process
4. app store
5. appropriate
6. auction
7. automated detection
8. banning
9. bargaining power
10. behavioural advertising
11. blog hosting
12. bullying
13. business users
14. child sexual abuse material
15. complaint
16. conglomerate
17. conglomerate effect
18. consumer rights
19. content moderation
20. contestable
21. contextual advertising
22. control mechanism
23. counter-notice
24. coverage
25. cyber security
26. data sharing
27. dependency
28. digital identity
29. disabling
30. discrimination
31. disinformation
32. disputes
33. dissemination
34. divisive messages
35. due diligence
36. effective
37. effective measures
38. enforcement
39. ex-ante rules
40. fast-track assessment
41. flagging
42. fundamental rights
43. gender equality
44. governance
45. grooming
46. harmful
47. hate speech
48. illegal content
49. illegal medicine
50. information disclosure
51. institutional cooperation
52. internal practices
53. interoperability
54. know your customer
55. large online platform companies
56. leverage
57. liability
58. manipulation
59. market entry
60. national level
61. non-discrimination
62. non-payment
63. notice-and-action
64. notice-and-takedown
65. notifications
66. operating systems
67. oversight
68. pet trafficking
69. platforms' content policies
70. political advertising
71. price comparison
72. primary activities
73. programmatic advertising
74. proportionate
75. quality standards
76. Rating and reviews
77. Real-time bidding
78. recommendation
79. redress
80. Referral
81. reinstated content
82. removal
83. remuneration
84. reporting procedure

85. search engines
86. sector specific rules
87. self-employed
88. sharing
89. social networks
90. solidarity
91. suspension
92. tailored
93. takedowns
94. terrorist propaganda
95. trusted organisations
96. trusted researchers
97. unfair
98. unfair practices
99. unfavorable
100. user base
101. very large online platform companies
102. video sharing

Terms manually coded from the DSA and DMA proposals:

1. easily accessible
2. clear
3. unambiguous
4. specific
5. easily comprehensible
6. available
7. detailed
8. easy to access
9. user-friendly
10. precise

3.2 Statistical test

To see if there is evidence for a statistically significant semantic difference between the use of a term between the different stakeholder groups, we must perform a statistical test of their relative distance. We can model the relative distance d_t^{AB} dABt of a word t in the corpus A and B be as:

$$d_t^{AB} = \gamma_t^{AB} + \mu_t^{AB} + u_t^{AB}$$

This takes into account a semantical term γ_t^{AB}, a non-semantical term (originated from the simple different words disposition in the two corpora) and a random term . More precisely, the semantic term is defined as the difference in the usage of the same word which is driven by different understandings of the meaning of this term. Hence, this is the term we are interested in. On the other hand, the non-semantic term is defined as the term capturing all the non-semantic differences in usage, which can emanate from more frequent use of the word in different contexts, different authors, or stylistic differences. Finally, we define the random term as random differences in usage unrelated to systematic differences between the corpora. These could arise from the document-production process or the randomness of the initialization of the word-embedding algorithm's training.[109]

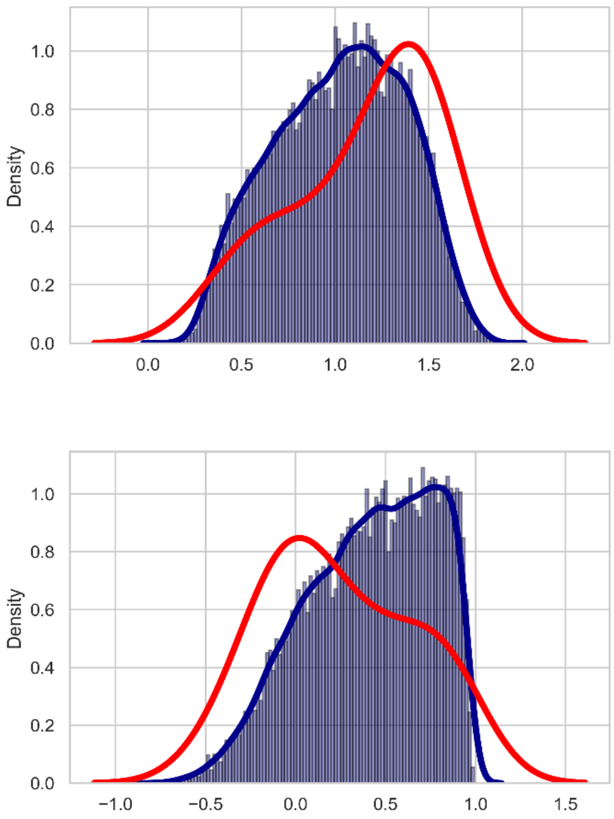
The statistical test we performed is based on two assumptions. Our first assumption is that words in the control vocabulary used for the Vector Space Alignment Transformation do not have a semantic difference, i.e., γ_t^{AB}=0. Consequently, their relative distance can give an empirical distribution of the non-semantical distance between words, composed of the only two terms $\mu_t^{AB} + u_t^{AB}$ which is our second assumption. In this manner, it is possible to construct an empirical cumulative distribution of these distances, distributed with the hypothesis of zero semantic difference.

We first built an empirical Fisher-Snedecor distribution of distances calculated with all the common words included in the Control Vocabulary. We then analyzed the distance between the vectors of a word in the two corpora, counting the number of times these values were smaller than the control words' distances in the distribution. If we accept the null hypothesis that the word we are analyzing shows no semantic difference between the different corpora, then the obtained (normalized) **p-value** tells us the probability to have a distance equal or greater than that. If this probability is small enough, we can refuse this null hypothesis with a small possibility of error. This is to say that the particular word has, indeed, a statistically significant semantic difference in the two corpora. A general acceptance value for the p-value is 0.05, which we will use as the critical threshold for our analysis.

Annex 4. Cumulative distribution of semantic differences

Figures 1 to 3 show the cumulative distribution of distances of control dictionary words (in blue) against the cumulative distribution of distances and similarities of analyzed words (in red) for each corpus pair (i.e., corpus X against corpus Y). The plot shows that the words we analyzed create a statistical distribution different from the one of the common words, as we can see from the different shapes. These differences suggest that there are significant semantic differences between the corpora.

Figure 1. Corpuses AB - Cumulative distribution of control distances (top) and similarities (bottom)



[109] Nyarko & Sanga (n 102).

Figure 2. Corpuses BC - Cumulative distribution of control distances (left) and similarities (right)

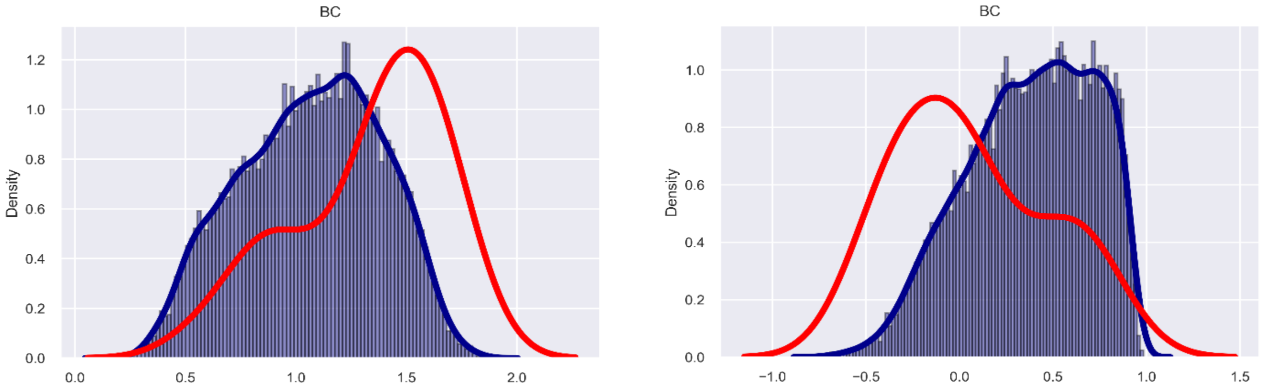
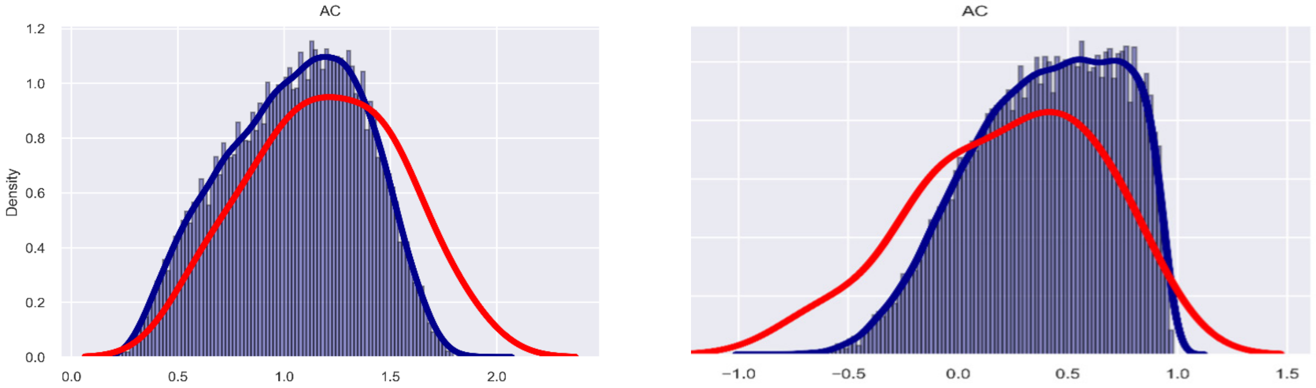


Figure 3. Corpus Pair AC - Cumulative distribution of control distances (left) and similarities (right)


AC
Density
AC