Instructions for EMNLP 2023 Proceedings

Anonymous EMNLP submission

Abstract

Natural language processing (NLP) techniques have increased substantially in performance and can potentially automate some areas of legal reasoning and writing. However, increased performance has come at the cost of increased opacity of the models. XAI has been developed to combat this opacity. Nevertheless, it is unclear how such explanations should be evaluated and whether these explanations are effective because effectiveness depends on the varying purposes and end-users of the explanations. Separately, the stakeholders of data privacy regulation (i.e. organisations, data privacy regulatory authorities, and consumers of data services) can benefit particularly from the use of XAI. Hence, this capstone fills a gap in research by evaluating the effectiveness of XAI within the context of data privacy.

This capstone trains and reports the performance of machine learning classifiers that detect data practices in PlayStore apps' data privacy policies. Then, an XAI technique, LIME, is applied on these models to produce explanations of classifications of selected data practices. Finally, the explanations are evaluated by human evaluation using a survey which uniquely utilises varied types of questions to measure different values linked to explanability (effectiveness, risk, fairness and trust) from the perspective of different stakeholders of data privacy. In terms of model performance, I achieve a 5-class weighted average F1 score of 66% using SVC with Tf-IDF. In terms of human evaluation, after viewing the explanations, respondents reported a statistically significant decrease in overall trust and understandability of the models, amongst other more specific findings.

1 Motivation and significance

A question was posed in 2017: "Will there be a software service for the generation of explanations and arguments in law that will assist in structuring explanations of answers and supportive legal arguments?" (Ashley, 2017). This question seems close to getting an answer in 2023 with the release of ChatGPT (OpenAI, 2022) to the general public. Consider the following response from ChatGPT:

Indeed, GPT-4, successor of ChatGPT and released a mere 4 months after ChatGPT, scored on the 90th percentile on the Uniform Bar Exam which is significantly in excess of the passing mark to practice law in the US (Katz et al., 2023).

ChatGPT and GPT-4 are the most recent AI models that uses advanced natural language processing (NLP). NLP is a branch of AI that gives computers the ability to understand text and spoken words as how human beings would understand (IBM, 2022). While NLP techniques have substantially increased in performance in recent years, it has come at the cost of the explainability of their predictions because of models that are architecturally more complex and opaque (El Zini and Awad, 2022). This issue of explainability is exemplified by the last question I posed to ChatGPT about its view on the egg-shell skull rule:

ChatGPT does not seem to be able to explain its views like how a typical human would. Further prompting led ChatGPT to provide a list of academic papers that support its view. However, there are issues of hallucination, where ChatGPT generates answers about facts that are false or nonexistent (Alkaissi and McFarlane, 2023). This lack of explainability (and potential inaccuracies of the explanations) could significantly hinder NLP's greater adoption within the legal industry because the lawyer that uses these models ultimately bear the legal responsibility of ensuring that the analysis is legally sound. In Singapore, a legal practitioner must act with reasonable diligence and competence in the provision of services to the client[1]. A lawyer that relies on the analysis of legal tech tools but does not understand how the analysis was produced

[1] According to r5(2)(c) of the Legal Profession (Professional Conduct) Rules 2015.

could be considered lacking in diligence and competence.

Nevertheless, the intersection in skillset between data science and legal analysis is still nascent and it is unrealistic to expect all legally trained personnel to be proficient in data science to the extent required to understand the logic of machine learning models without aid. Thus, eXplainable AI (XAI) techniques and research have been rising in popularity and has most recently found its way into the the EU's proposed AI Act that places explainability requirements on AI providers (Gyevnar and Ferguson, 2023). XAI refers to methods that attempt to help the end user of a model understand how it arrives at its result (Danilevsky et al., 2020), which reduces the model's opacity. XAI can therefore bridge the gap between the lawyer and the data scientist by using XAI techniques to explain the predictions of AI models used in legal decision making.

However, there is not much consensus about what "explainable" means. While the general agreed upon goal of XAI is to "completely, accurately and clearly quantify the [model's] logic", there is no consistent use of the terms "explainability", "interpretability" and "transparency" (Danilevsky et al., 2020). Interpretability is sometimes used to describe the model's internal logic and explainability as the ability of the user to understand that logic. However, explainability is also described as techniques to explain the model's logic post-hoc without necessarily being representative of the model's true decision (Rosenfeld, 2021). Though the AI Act frequently refers to different levels of transparency, there is no explicit definition of transparency. Transparency could involve mandatory disclosure of an AI system used in decision-making which is a much broader concept than transparency used in current XAI literature that covers only the algorithmic process (Gyevnar and Ferguson, 2023). In this capstone, I use explainability, interpretability, understandability and transparency interchangeably to refer to the overall goal of XAI as stated above.

Another area of XAI is differentiating what explainability means to different users of the model who have different purposes for the explanations. For example, what could be explainable to data scientists may not be explainable to laypersons. While data scientists may find that more technical details about the model would make the model more understandable, laypersons might be more confused if too many details are provided (Rosenfeld, 2021). Therefore, assessing the effectiveness of XAI is highly dependent on the specific context and needs of the users.

In this capstone, XAI is assessed within the context of data privacy. The widespread collection and use of data by organisations in recent years has led to an increase of regulations governing data privacy, with 145 countries having enacted data protection legislation in 2021 (Gstrein and Beaulieu, 2022). With more sophisticated regulation comes increased difficulties for organisations to ensure that they are complying with these regulations, and for governments to enforce them. Given that explainability is context and user dependent, the following explains why XAI is important to the three different stakeholders in data privacy: the consumer, organisation and state.

For the consumer, it is uncontroversial that data privacy policies on websites and software are rarely read, and even if they are read, consumers are unlikely to fully understand them because of the use of extensive legalese. The ease of reading data privacy policies was measured to be roughly the same as compared to academic articles such as the Harvard Law Review. The average policy takes 17 minutes to read, and the annual reading time per consumer is more than 400 hours which is more than an hour per day[2] (Wagner, 2022). This increasing unreadability of data privacy policies poses a significant challenge to the effectiveness of data privacy regulation given that the current model of regulation depends on the consumer giving consent that is informed and freely given (Mantelero, 2014). Combined with the increasing collection and use of data by organisations, consumers have diminishing control over how their data is collected and used. Therefore, XAI could be helpful to consumers since XAI can be used to automate legal analysis of data privacy policies and explain the implications of the analysis in simple terms.

Organisations in the EU are subjected to a data subject's "right to explanation" under the General Data Protection Regulation (GDPR). Data subjects have the right to obtain an explanation of a decision reached solely through automated processing[3]. For example, a bank could use AI to predict the prob-

[2] Assuming that a consumer visits 1462 unique websites each year.

[3] This is a plausible interpretation when reading Recital 71 in conjunction with Art 22 of the GDPR.

ability of a customer defaulting on a loan. This prediction could be used to justify a decision to deny a loan to the customer. Under the GDPR, the customer has the right to not be subjected to such automated processing and obtain an explanation as to why the model made such a prediction. Therefore, using opaque AI could affect the organisation's compliance of their legal obligations.

In Singapore, as part of Personal Data Protection Commission's (PDPC) guidance on the operations management of AI models, organisations are advised to provide explanations on how AI models are used in decision making to build understanding and trust with those that use their products (Personal Data Protection Commission, 2020). In terms of communication policies, organisations are advised to develop guidelines on the type of explanations and when to provide them. Despite debate about the scope of the right to explanation and whether it is binding (Chesterman, 2021a), such legislation in addition to the PDPC's guidance for explainability supports the need for further development of XAI specifically in the context of data privacy.

One of the state's concerns when AI is used in legal decision making is the integrity of the judicial system and regulatory activities since the legal system depends as much on the justification of its decisions as it does on the decisions themselves (Chesterman, 2021b). In the context of Singapore's Personal Data Protection Act (PDPA), there is an appeal process for cases appearing before the PDPC and the higher courts can overturn the PDPC's decisions[4]. When overturning the PDPC's decisions, the higher courts do not disagree with the outcome but they disagree with the reasoning of the PDPC's decision. Assuming that AI models are sophisticated enough to write (or at least assist the writing) court judgements, if there was no explanation of the model's writing, there would be little basis for the higher courts to overturn judgements made by lower courts.

Hence, the importance of XAI applies across all three stakeholders in data privacy. Consumers can benefit from XAI by making privacy policies more readable. Organisations can use XAI to be more accountable to consumers and regulators by providing explanations for automated decisions. Transparency brought by XAI also helps the state to regulate organisations' use of AI. Therefore, I focus on NLP and XAI in the specific context of data privacy, as this context provides a realistic and practical scope for evaluating the explainability of AI models.

This capstone's objectives come broadly within the evaluation of XAI techniques. While there have been studies that evaluate XAI through human evaluation (Vilone and Longo, 2021) and non-empirical research that investigates the relevant standard for explainability across different areas of law (Hacker and Passoth, 2022), I adopt an uniquely empirical methodology to assess XAI within a data privacy context across the different use cases of stakeholders. The closest study using similar methodology is one that asked lawyers to rate visualisations generated by different XAI methods (Górski and Ramakrishna, 2021). However, this study was conducted using a dataset containing cases from the US Board of Veterans' Appeals, and did not consider evaluating how explainability can vary with different purposes of the explanations. To this end, this capstone also serves as a trial for an unique survey design that varies the purpose of explanations to assess how explainability and values related to explainability change.

2 Research questions and roadmap

I have four research questions[5]:

1. Which AI models ("classifiers")[6] perform the best on a data privacy dataset in terms of traditional performance metrics? (Section **??**)

2. Which classifiers are the most understandable to users that include laypersons and users with domain knowledge in data science and the law? (Section **??**)

3. How would users rate the explainability of a selected XAI technique? (Section **??**)

4. What are the differences in explainability if users were asked to consider the predictions of these classifiers from the perspective of a consumer, an organisation and the PDPC? (Section **??**)

To investigate these questions, I train and evaluate the performance of classifiers that classify sentences in apps' data privacy policies to a particular

[4] Per s48Q, s48R and s54 of the PDPA.

[5] All code and analysis can be found here.

[6] I refer to AI models as classifiers from hereon as the task that the AI models perform in this capstone is classifying sentences to a data practice. I elaborate more on this distinction in Section **??**.

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\^e}	ê	{\u g}	ğ
{\`i}	ì	{\l}	ł
{\.I}	İ	{\~n}	ñ
{\o}	ø	{\H o}	ő
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTeX entries.

data practice. A privacy practice is a section of an app data privacy policy that describes a certain behaviour of an app which has privacy implications[7]. XAI methods are then used to visualise why the models made such predictions. These visualisations are assessed for effectiveness by conducting a survey asking respondents to rate these visualisations according to certain metrics and values related to explainability and by using different types of questions.

Chapter **??** introduces the dataset and provides a literature review and justification for the methodology used in this capstone with regard to the NLP classifiers, XAI techniques, and human evaluation of XAI. Chapter **??** presents an exploratory data analysis (EDA) of the dataset and reports classifier performance. Chapter **??** discusses the survey results and Chapter **??** concludes the capstone.

3 Engines

To produce a PDF file, pdfLaTeX is strongly recommended (over original LaTeX plus dvips+ps2pdf or dvipdf). XeLaTeX also produces PDF files, and is especially suitable for text in non-Latin scripts.

4 Preamble

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{EMNLP2023}
```

For the final version, omit the `review` option:

```
\usepackage{EMNLP2023}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.) Please see the LaTeX source of this document for comments on other packages that may be useful. Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples. By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

5 Document Body

5.1 Footnotes

Footnotes are inserted with the `\footnote` command.[8]

5.2 Tables and figures

See Table 1 for an example of a table and its caption. **Do not override the default caption sizes.**

5.3 Hyperlinks

Users of older versions of LaTeX may encounter the following error during compilation:

```
\pdfendlink ended up in different
nesting level than \pdfstartlink.
```

This happens when pdfLaTeX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LaTeX to 2018-12-01 or later.

5.4 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (Gusfield, 1997). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. Gusfield, 1997).

[7] This will be further elaborated on when the dataset is introduced.

[8] This is a footnote.

Output	natbib command	Old ACL-style command
(Cooley and Tukey, 1965)	`\citep`	`\cite`
Cooley and Tukey, 1965	`\citealp`	no equivalent
Cooley and Tukey (1965)	`\citet`	`\newcite`
(1965)	`\citeyearpar`	`\shortcite`
Cooley and Tukey's (1965)	`\citeposs`	no equivalent
(FFT; Cooley and Tukey, 1965)	`\citep[FFT;][]`	no equivalent

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

5.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
    \bibliographystyle{acl_natbib}
    \bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
    \bibliographystyle{acl_natbib}
    \bibliography{anthology,custom}
```

Please see Section 6 for information on preparing BibTeX files.

5.6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

6 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Limitations

EMNLP 2023 requires all submissions to have a section titled "Limitations", for discussing the limitations of the paper as a complement to the discussion of strengths in the main text. This section should occur after the conclusion, but before the references. It will not count towards the page limit.

The discussion of limitations is mandatory. Papers without a limitation section will be desk-rejected without review. ARR-reviewed papers that did not include "Limitations" section in their prior submission, should submit a PDF with such a section together with their EMNLP 2023 submission.

While we are open to different types of limitations, just mentioning that a set of results have been shown for English only probably does not reflect what we expect. Mentioning that the method works mostly for languages with limited morphology, like English, is a much better alternative. In addition, limitations such as low scalability to long text, the requirement of large GPU resources, or other things that inspire crucial further investigation are welcome.

Ethics Statement

Scientific work published at EMNLP 2023 must comply with the ACL Ethics Policy. We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

Acknowledgements

This document has been adapted by Yue Zhang, Ryan Cotterell and Lea Frermann from the style files used for earlier ACL and NAACL proceedings, including those for ACL 2020 by Steven Bethard,

Ryan Cotterell and Rui Yan, ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15(2).

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Kevin D. Ashley. 2017. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

Benjamin Börschinger and Mark Johnson. 2011. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia.

Simon Chesterman. 2021a. *We, The Robots: Regulating Artificial Intelligence and the Limits of the Law*, chapter 6. Transparency. Cambridge University Press.

Simon Chesterman. 2021b. *We, The Robots: Regulating Artificial Intelligence and the Limits of the Law*, chapter 3. Opacity. Cambridge University Press.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. *CoRR*, abs/2010.00711.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(103):1–31.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Łukasz Górski and Shashishekar Ramakrishna. 2021. Explainable artificial intelligence, lawyer's perspective. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 60–68, New York, NY, USA. Association for Computing Machinery.

Oskar J. Gstrein and Anne Beaulieu. 2022. How to protect privacy in a datafied society? a presentation of multiple legal and conceptual approaches. *Philos Technol*, 35(1):3.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Balint Gyevnar and Nick Ferguson. 2023. Aligning explainable ai and the law: The european perspective. *arXiv preprint arXiv:2302.10766*.

Philipp Hacker and Jan-Hendrik Passoth. 2022. Varieties of ai explanations under the law. from the gdpr to the aia, and beyond. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 343–373. Springer.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

IBM. 2022. What is natural language processing? Last Accessed: 2023-02-11.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Alessandro Mantelero. 2014. The future of consumer data protection in the eu re-thinking the "notice and consent" paradigm in the new era of predictive analytics. *Computer Law & Security Review*, 30(6):643–660.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Last Accessed: 2023-02-11.

Personal Data Protection Commission. 2020. Model artificial intelligence governance framework second edition. Technical report, Personal Data Protection Commission. Last Accessed: 2023-2-20.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Avi Rosenfeld. 2021. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 45–50, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Isabel Wagner. 2022. Privacy policies across the ages: Content and readability of privacy policies 1996–2021. *arXiv preprint arXiv:2201.08739*.

A Example Appendix

This is a section in the appendix.