Automatic Classification of Legal Violations in Cookie Banner Texts

Marieke van Hofslot
Utrecht University
Information and Computing Sciences
m.t.vanhofslot@students.uu.nl

Almila Akdag Salah
Utrecht University
Information and Computing Sciences
a.a.akdag@uu.nl

Albert Gatt
Utrecht University
Information and Computing Sciences
a.gatt@uu.nl

Cristiana Santos
Utrecht University
Faculty of Law
c.teixeirasantos@uu.nl

Abstract

Cookie banners are designed to request consent from website visitors for their personal data. Recent research suggest that a high percentage of cookie banners violate legal regulations as defined by the General Data Protection Regulation (GDPR) and the ePrivacy Directive. In this paper, we focus on language used in these cookie banners, and whether these violations can be automatically detected, or not. We make use of a small cookie banner dataset that is annotated by five experts for legal violations and test it with state of the art classification models, namely BERT, LEGAL-BERT, BART in a zero-shot setting and BERT with LIWC embeddings. Our results show that none of the models outperform the others in all classes, but in general, BERT and LEGAL-BERT provide the highest accuracy results (70%-97%). However, they are influenced by the small size and the unbalanced distributions in the dataset.

1 Introduction

Cookie banners are a part of everyday life for EU-based users while browsing the Web. To comply with the General Data Protection Regulation (GDPR) (EU, 2018) and the ePrivacy Directive (ePD-09), website operators have to inform EU users and ask for their consent for the processing of their personal data for 'unnecessary purposes', i.e. data that is not needed for the website to function, such as user-targeted advertising (Article 29 Working Party, 2012). Accordingly, EU users have to navigate through a cookie banner and decide on whether to consent to their personal information being collected via cookies or other tracking technologies that the site embeds. A consent request needs to be unambiguous, clear, concise, and informative, and consent needs to be freely given (Articles 4(11) and 7(2) (EU, 2018)).

Research has found that 89% of cookie banners violate applicable laws (Santos et al., 2021; Soe et al., 2020; Nouwens et al., 2020). The legal study by (Santos et al., 2021) focused on processing purposes of cookie banners and confirmed that 89% of the cookie banners violated at least one legal requirement applied to the text of the stated purposes; they further detected the use of vagueness, framing, misleading wording, and technical jargon. Utz et al. (2019) noted that the text to explain the purpose of data collection was typically expressed in generic terms, and use of technical jargon was not understandable properly by the average data subject. Studies furthermore confirmed that the prevalence of "affirmative" options and positive framing could nudge users toward consenting to tracking (Hausner and Gertz, 2021; Kampanos and Shahandashti, 2021).

The *language* used in cookie banners is often formulated in a way that can confuse and impact users' privacy decisions, steering them to accept consent to tracking. Regulators, policymakers and scholars (CNIL, 2022; Gray et al., 2018; Article 29 Working Party, 2018; European Data Protection Board, 2020, 2022; Chatellier et al., 2019), confirm that certain textual strategies such as the use of motivational language and humor (European Data Protection Board, 2022; Frobrukerrådet, 2018), shame (Mathur et al., 2019), guilt (Brignull, 2010), blame (Chatellier et al., 2019), fear (Bongard-Blanchy et al., 2021) or uncertainty (European Data Protection Board, 2020) influence users' online decisions. Such textual expressions can violate the legal requirements for consent. Consent, if not obtained in compliance with the GDPR, provides invalid grounds for data processing, rendering the processing activity illegal (Article 6(1)(a) GDPR)).

There is a need to identify such textual violations and develop tools that can automatically de-

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 287 - 295

tect such textual *dark patterns* (Mathur et al., 2019) in order to provide proof of such practices (and legal evidence) to support the legal proceedings of enforcement authorities in their auditing efforts. Regulators are presently overwhelmed by the novelty and sheer scale at which such patterns are being deployed online. However, only a few studies have investigated automatic detection of legal violations in cookie banner text. Bollinger et al. (2022) used feature extraction and ensembles of decision trees for their cookie purpose classifier with which they developed a browser extension to remove cookies according to user preferences. Khandelwal et al. (2022) used a fine-tuned BERT Base-Cased model to discover and force cookie settings to disable all non-essential cookies.

These studies focus on enhancing the usability of websites for the users. In this paper, we focus on automatic detection of legal violations in cookie banner texts. We work with a dataset that is annotated by five experts for such violations, and test the performance of four state of the art deep neural network models, BERT, BERT with LIWC, LEGAL-BERT and BART in a zero-shot setting. Our aim is to understand if large, pre-trained language models can be used with little or no finetuning for auditing purposes by policymakers or consumer protection organisations. To that end, we document the shortcomings of the models to provide insight on the problems and challenges of such a classification task. Our results suggest that no model outperforms all the others in all classification tasks, suggesting a need for more data annotation in this domain, as well as signalling a potential for models which are specifically trained or fine-tuned on the task at hand.

2 Methodology

In this section we first describe the dataset, discuss the annotation and classification based on manual labels. Lastly, we describe the classification models we have used.

2.1 Dataset

In Santos et al. (2021), cookie banner texts were manually annotated according to the GDPR legal requirements and their corresponding violations. The resulting dataset consists of 407 cookie banner text segments. The texts are in English, and have an average of 3.59 sentences and 49.77 words. The most common content words (i.e. 'cookies', 'website', 'policy', etc.) are very specific to the context of cookie banners.

Annotation class	Classification labels
Consent options presence	Reject option No reject option
Framing	Negative framing Positive framing No framing
Misleading language	Deception Misleading language Prolixity Vagueness No Misleading language
Purpose	Purpose mentioned No purpose mentioned
Technical jargon	Technical jargon No technical jargon

Table 1: Annotation categories and classes

Annotation classes and classification labels. These are based on the annotation guidelines used by the five experts for the study in Santos et al. (2021), where a given annotation *class* has one or more corresponding *labels*. The original dataset annotated texts segment-wise. In contrast, the goal of the present work was to label the cookie banner as a whole, to indicate whether it contains one or more instances of language that falls under any of these labels. The labels assigned to each cookie banner are thus determined by the presence of the labels in their text segments, in the original data. Thus, some segments might belong to more than one class and label.

Due to data sparseness, some classes in the original guidelines by Santos et al. (2021) were omitted, leaving five classes in total: *Consent options presence*, *Misleading language*, *Framing*, *Purpose* and *Technical jargon* (see Table 1).

2.2 Models

In this paper, we compare the performance of the following models, as measured by their classification accuracy:

BERT (Devlin et al., 2019) is a widely-used

Transformer-based model, which serves as the basis for a variety of text classification tasks, including topic classification, and sentiment analysis. The major advantage of BERT is that it was pretrained on a large corpus, allowing it to be fine-tuned on a downstream task with a relatively small data set. We encode each cookie banner text segment into a fixed-sized vector using its BERT embedding, using this as input to a classification layer fine-tuned on the training and validation data.

BERT with LIWC features. Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) is a dictionary-based text analysis tool with linguistic, psychological and topical categories. LIWC calculates the percentage of words from the cookie banner text that fall into each category and creates a vector of all these percentages. We concatenate BERT embeddings with a LIWC vector representing all 80 categories used by LIWC. The remaining architecture is the same with BERT. For classes like *framing*, *misleading language* and *technical jargon*, we expect that LIWC will increase the performance of the model, since these features reflect the more stylistic aspects of the text.

LEGAL-BERT. LEGAL-BERT are a family of BERT models that have been pre-trained on diverse English legal text from several fields, including European legislation (EURLEX[1]), UK legislation [2], and various courts from Europe and the US[3]. Since the general LEGAL-BERT model performs better than BERT on domain-specific tasks (Chalkidis et al., 2020), we use the general LEGAL-BERT as a comparison for the BERT model. While cookie banners are not themselves legal texts, they do explain legally relevant provisions; hence, we include this model to address the utility of a domain-specific BERT model in the general legal domain.

BART in ZS-setting. Zero-shot (ZS) classification in NLP has been used to classify text on which a model is not specifically trained (Sarkar et al., 2021; Yin et al., 2019a; Ye et al., 2020). Here, we use the pre-trained BART-Large MNLI model (Lewis et al., 2019) as an out-of-the-box zero-shot text classifier, similar to (Yin et al., 2019b). To do this, we reframe the classification task as a Natural Language Inference task (NLI), where the goal is to determine whether two texts, a premise and a hypothesis, are in a relation of entailment, contradiction, or are neutral. Here, the cookie banner text is the premise and the corresponding labels are hypotheses. We use the model to estimate the probability of each label for every cookie banner text segment. The label with the highest probability is selected.

Training details and hyperparameters. For simplicity, a separate model was trained for each class. For the fine-tuned models based on BERT and BERT-LEGAL, we use a classification layer of size 768, followed by a ReLU layer, to determine the most probable label for each class. For BERT and BERT+LIWC features, we use BERT Base-cased. Since Base-Cased is not available for LEGAL-BERT, we use LEGAL-BERT Base-uncased. For the BERT-like models, the learning rate is set as 1e-6, the model is trained by using cross-entropy loss and the Adam optimizer. The training was set for 12 epochs. For reporting our results, we used a 2-fold (50/50) cross-validation setup. As our dataset is small and the class distributions are not balanced, we preferred a stratified split. Since BART is used in a zero shot-setting, cross-validation is not applicable for this model, and the results are reported accordingly. All of the models were run on a laptop with AMD Ryzen 7 5700U processor (1.80 GHz) and 16 GB DDR 4 RAM.

3 Results and Discussion

Table 2 shows the performance of these models in terms of classification accuracy, computed as a proportion of correctly labelled instances per class. We provide F1-scores for all classes in Table 3.

Accuracy performance differs for each class. Overall, we do not have a model that outperforms all the others for all classes. The best accuracy performance for each class differs.

Technical jargon: LEGAL-BERT gives the best result with 81.3%, although the difference between the models is only a few percent, BERT + LIWC's result being the lowest with 74.95%. In general F1 scores are high for the majority labels and not the minority labels, but this is especially the case

[1]Publicly available from http://eur-lex.europa.eu/

[2]Publicly available from http://www.legislation.gov.uk

[3]Cases from the European Court of Justice (ECJ), also available from EURLEX, cases from HUDOC, the repository of the European Court of Human Rights (ECHR) (http://hudoc.echr.coe.int/eng), cases from various courts across the USA, see https://case.law and US contracts from EDGAR, the database of US Securities and Exchange Commission (SECOM) (https://www.sec.gov/edgar.shtml).

Class	BERT CV	BERT+LIWC CV	LEGAL-BERT CV	BART-ZS
Consent options presence	90.7 (±0.95)	89.7 (±0.95)	85.3 (±0.55)	**91.65**
Framing	**67.4 (±0.15)**	60.7 (±1.60)	65.9 (±0.15)	58.23
Misleading language	**65.2 (±2.85)**	60.2 (±3.50)	65.1 (±0.40)	54.30
Purpose	91.9 (±0.20)	90.0 (±0.75)	**93.4 (±0.25)**	76.90
Technical jargon	79.2 (±1.65)	74.95 (±2.55)	**81.3 (±0.45)**	78.87

Table 2: Comparison of cross-validation accuracies (mean and std) with best score per class/row in bold.

Class	Label	BERT CV	BERT+LIWC CV	LEGAL-BERT CV	Test set occurr. Fold 1	Test set occurr. Fold 2	BART-ZS
Consent opt. presence	Other	0.95 (±0.01)	0.94 (±0.00)	0.92 (±0.00)	172	172	0.95
	Reject option	0.62 (±0.05)	0.61 (±0.08)	0.13 (±0.13)	32	31	0.68
Framing	No framing	0.75 (±0.01)	0.71 (±0.02)	0.76 (±0.00)	120	119	0.73
	Positive	0.58 (±0.04)	0.45 (±0.01)	0.46 (±0.01)	76	76	0.17
	Negative	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)	8	8	0.13
Misleading language	None	0.82 (±0.00)	0.78 (±0.02)	0.79 (±0.00)	134	133	0.71
	Vagueness	0.21 (±0.03)	0.17 (±0.01)	0.00 (±0.00)	34	34	0.16
	Decept. lang.	0.08 (±0.08)	0.27 (±0.09)	0.00 (±0.00)	26	25	0.04
	Prolixity	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)	10	11	0.00
Purpose	Yes	0.95 (±0.00)	0.94 (±0.00)	0.96 (±0.00)	164	164	0.87
	None	0.75 (±0.00)	0.69 (±0.03)	0.81 (±0.00)	40	39	0.00
Technical jargon	None	0.88 (±0.01)	0.85 (±0.02)	0.90 (±0.01)	166	165	0.88
	Yes	0.13 (±0.13)	0.16 (±0.04)	0.08 (±0.03)	38	38	0.09

Table 3: Cross validation F1-scores (mean and std) for all models per class label

for LEGAL-BERT with only 0.08 F1 score for the minority label.

Consent options presence: The accuracy is high for all models, but the highest score is from BART with 91.65%.

Purpose: The highest accuracy comes from LEGAL-BERT with 93.4%, where BERT is close with 91.9% and BERT-LIWC still high with 90.0%. In general, this class suffers the least from the overfitting to the majority label, and has overall higher F1 scores for both labels. BART performs the worst with 76.9%, and has the lowest F1 scores.

Misleading language and *Framing*: these labels have the lowest accuracy out of the five classes, with accuracy percentages dropping to 60% for some models. We also observe the lowest occurrences in these classes, with very low or null F1 scores. Given that these are the classes with more than two labels and rely on stylistic aspects of the text, these results are not surprising.

Misleading language: BERT and LEGAL-BERT have close scores with 65.2% and 65.1%. However LEGAL-BERT has a lower std. The Prolixity label has null F1 scores for all models.

Framing: LEGAL-BERT produces the highest accuracy score for Framing with 65.9%. The Negative Framing label has null F1 scores for all models except BART.

Model comparison: To compare the classification results of models, we used pairwise McNemar tests, see Table 4. Overall, BERT and LEGAL-BERT models achieved relatively good and similar accuracy scores across all classes. However, LEGAL-BERT's F1 scores are lower then BERT for minority classes. Comparing the two models with McNemar test we observe that they perform significantly differently for Consent options class.

Observations: When we sample instances where the models fail to classify one of the five classes correctly, we see the shortcomings of each model better (see Appendix for a list of examples, and how they are classified by each

Class	BERT / BERT+LIWC	BERT / LEGAL-BERT	BERT+LIWC / LEGAL-BERT	BERT / BART-ZS	LEGAL-BERT / BART-ZS	BERT+LIWC / BART-ZS
Consent opt. presence	.585	.000**	.011*	.716	.007*	.396
Framing	.010*	.617	.069	.011*	.028*	.533
Misleading language	.013*	1.000	.012*	.002*	.002*	.097
Purpose	.134	.238	.013*	.000**	.000**	.000**
Technical jargon	.033*	.108	.000**	1.000	.419	.208

Table 4: P-values of McNemar's test on all model combinations. $^{*}p < .05$, $^{**}p < .001$

model). In most classes, BERT and LEGAL-BERT seem to wrongly over-classify the majority label. BERT+LIWC only does this with *"Framing"* and performs well on all other classes. LEGAL-BERT fails in the class *"Framing"*, where it classifies an instance of *"No framing"* as *"Positive framing"*. Overall, BART does not perform well, but contrary to the BERT models, the incorrect classifications are not due to choosing the majority class.

Occurrence distribution: Studying the classes and their corresponding misclassifications and the F1 scores, we observe that the data distribution affects the accuracy. Classification labels that have a low amount of occurrences in the data are almost always wrongly classified, even after the application of a stratified split for training and validation (see Table 3). This means that more data should be collected and annotated for these classes. Furthermore, fine-tuning of the models during training is needed, a common solution here is adding weights to the minority classes.

Implications: The challenges of automatic classification of cookie banners are due to purposefully confusing wording, lack of classified data by experts, and the shortness of cookie banners themselves. The obtained results show that using use a state of the art classification model off the shelf or with minimal fine-tuning will not yield reliable results for auditing or helping policymakers.

4 Conclusion and Future work

In this paper, we used a cookie banner dataset previously annotated by five experts that detected legal violations. We test state of the art deep learning models such as BERT, LEGAL-BERT and BART for automatic classification of such violations in this dataset. We also combined a dictionary based approach, i.e. LIWC embeddings with BERT, and checked if this improves performance or not.

Our approach aimed to give more insight into automatic detection of legal violations of cookie banners texts by comparing frequently used models. Our results suggest that there is not one model that outperforms all the others for all classes that need to be detected. In general, BERT and LEGAL-BERT work well for all classes; however, a closer look reveals that these models are also affected by the skewed data distribution for certain classes. In contrast, BART performs worst for most of the classes, but is not affected by the small size of the data set, and by class imbalance.

We further add to the limited amount of studies on automatic detection of textual legal violations of cookie banners and laying a foundation for further research on this topic. Since the language and style of the cookie banners change rapidly, we need robust algorithms that can adapt to changes both in the legal domain and in the manner of adoption of new regulations by website operators. Hence, it is crucial to develop an efficient annotation pipeline to speed up human-in-the-loop annotation and automatic classification. Our initial tests give insight into which model performs well for which challenges, and can be used further to build such a pipeline in the future.

5 Ethical implications and limitations

In this paper, we rely on large, pretrained language models for classification, fine-tuning them on a small, manually labelled dataset.

One limitation of this approach is the limited size of the manually labelled data. While accuracy and F1 figures may suggest reasonable performance on certain classes, we cannot consider such results as final, or as indicating that the models we use are sufficiently robust to be deployed in real-world settings. Rather, the results provide a picture of what current language models can achieve in a relatively under-explored domain, and provide directions for future work. As noted in the conclud-

ing section, one important direction is to curate larger and more diverse training data for the task of cookie banner classification.

References

Article 29 Working Party. 2012. Opinion 04/2012 on cookie consent exemption (WP 194). Technical report. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp194_en.pdf.

Article 29 Working Party. 2018. Guidelines on transparency under regulation 2016/679, (wp260). Technical report.

Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. 2022. Automating cookie consent and gdpr violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association.

Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "i am definitely manipulated, even when i am aware of it. it's ridiculous!" - dark patterns from the end-user perspective. *Proceedings of ACM DIS Conference on Designing Interactive Systems*.

Harry Brignull. 2010. Dark patterns. https://www.darkpatterns.org.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Régis Chatellier, Geoffrey Delcroix, Estelle Hary, and Camille Girard-Chanudet. 2019. Shaping choices in the digital world. https://linc.cnil.fr/sites/default/files/atoms/files/cnil_ip_report_06_shaping_choices_in_the_digital_world.pdf.

CNIL. 2022. Deliberation of the restricted committee No. SAN-2021-024 of 31 December 2021 concerning FACEBOOK IRELAND LIMITED. https://www.cnil.fr/sites/default/files/atoms/files/deliberation_of_the_restricted_committee_no._san-2021-024_of_31_december_2021_concerning_facebook_ireland_limited.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

ePD-09. Directive 2009/136/ec of the european parliament and of the council of 25 november 2009 amending directive 2002/22/ec.

European Union EU. 2018. General data protection regulation.

European Data Protection Board. 2020. Guidelines 05/2020 on consent under regulation 2016/679. Technical report. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

European Data Protection Board. 2022. Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them Version 1.0 Adopted on 14 March 2022. https://edpb.europa.eu/system/files/2022-03/edpb_03-2022_guidelines_on_dark_patterns_in_social_media_platform_interfaces_en.pdf.

Frobrukerrådet. 2018. Deceived by design: How tech companies use dark patterns to discourage us from exercising our rights to privacy. https://www.forbrukerradet.no/undersokelse/no-undersokelsekategori/deceived-by-design.

Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Philip Hausner and Michael Gertz. 2021. Dark patterns in the interaction with cookie banners. *arXiv preprint arXiv:2103.14956*.

Georgios Kampanos and Siamak F. Shahandashti. 2021. Accept all: The landscape of cookie banners in greece and the uk.

Rishabh Khandelwal, Asmit Nayak, Hamza Harkous, and Kassem Fawaz. 2022. Cookieenforcer: Automated cookie notice analysis and enforcement. *arXiv preprint arXiv:2204.04221*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11k shopping websites.

Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–32.

Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Cristiana Santos, Arianna Rossi, Lorena Sanchez Chamorro, Kerstin Bongard-Blanchy, and Ruba Abu-Salma. 2021. Cookie banners, what's the purpose? analyzing cookie banner text through a legal lens. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, pages 187–194.

Rajdeep Sarkar, Atul Kr Ojha, Jay Megaro, John Mariano, Vall Herard, and John Philip McCrae. 2021. Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 102–106.

Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. 2020. Circumvention by design-dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12.

Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (un)informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, Suhang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019a. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019b. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161.

Appendix

Classification examples We provide some examples of (in)correct classifications of certain classes for all models, see Table 5. The corresponding cookie banner text segments are as follows:

1. In order to give you a better service our website uses cookies. By continuing to browse the site you are agreeing to our use of cookies. Further information. Yes, I agree.

2. This website or its third-party tools use cookies, which are necessary to its functioning and required to achieve the purposes illustrated in the cookie policy. If you want to know more or withdraw your consent to all or some of the cookies, please refer to the cookie policy. By closing this banner, scrolling this page, clicking a link or continuing to browse otherwise, you agree to the use of cookies.

3. We use cookies on this site to enhance your user experience Please read our Cookie policy for more info about our use of cookies and how you can disable them. By clicking the "I accept" button, you consent to the use of these cookies. More info I accept I do not accept.

4. This website uses cookies to enable you to place orders and to give you the best browsing experience possible. By continuing to browse you are agreeing to our use of cookies. Full details can be found here.

5. By using this site you agree to store cookies for the best site experience. More info Sure!

Banner text	Ground truth	BERT	BERT+LIWC	LEGAL-BERT	BART
1	**No framing**	No framing	No framing	Positive framing	No framing
2	**Negative framing**	No framing	No framing	No framing	Positive framing
3	**Positive framing**	No framing	No framing	No framing	Positive framing
1	**Vagueness**	No mislead. lang.	Vagueness	No mislead. lang.	Vagueness
3	**No mislead. lang.**	No mislead. lang.	No mislead. lang.	No mislead. lang.	Vagueness
4	**Deceptive lang.**	No mislead. ang.	No mislead. lang.	No mislead. lang.	Deceptive lang.
2	**Techn. jargon**	No techn. jargon	Techn. jargon	No techn. jargon	No techn. jargon
3	**No techn. jargon**	No techn. jargon	No techn. jargon	No techn. jargon	No techn. jargon
3	**Purpose ment.**	Purpose ment.	Purpose ment.	Purpose ment.	Purpose ment.
5	**No purpose ment.**	Purpose ment.	No purpose ment.	Purpose ment.	Purpose ment.
2	**No reject opt.**	No reject opt.	No reject opt.	No reject opt.	No reject opt.
3	**Reject opt.**	Reject opt.	Reject opt.	Reject opt.	No reject opt.

Table 5: Example cookie banner text segments and their corresponding classification for each model