Optimizing BERT For Classification Of Privacy Policies

Author: *Lukas Busch AUC, lukas.busch@student.auc.nl*
Supervisor: *Kasper Welbers VU, k.welbers@vu.nl*
Reader: *Giovanni Colavizza, g.colavizza@uva.nl*
Tutor*: Breanndán Ó Nualláin*
Date of Submission: 03-03-2021
Major: SCI

Summary: A growing number of studies have focussed on making privacy policies easier to read and understand. An important aspect of this field is the development of automated approaches to summarize and subtract essential information from privacy policies. For my research I will train a BERT model with the goal of improving these tasks. BERT is a state of the art model that can be applied to a variety of tasks. By using developments in the field of transfer learning I will domain-specifically train BERT on a large privacy police corpus, furthermore I will find the optimal hyper-parameters of BERT for the privacy policy domain. By including confidence scores in the classification results I aim to make my model more transparent. In doing so I hope to add to a more informed and less opaque online environment.

List of Abbreviations:

US	United States
CNN	Convolutional Neural Network
LR	Logistic Regression
SVM	Support Vector Machines

Keywords: Machine Learning, Classification, BERT, Privacy Policy, Optimization

Writing Update: For the writing update I will finish my methodology as well as set up the environment in Github

Word Count: 1943

1) Introduction:

In the modern online landscape data is a valuable resource. Therefore, it comes as no surprise that companies are collecting and selling users' data. The specifics of these data collection practices can be found in a company's privacy policy. This is an important resource internet users have to gain insight in their digital trace. However, a study from 2004 shows that only 4,5% of Americans actually read the privacy policies (Milne and Culnan, 2004). In 2008 McDonald and Cranor (2008) calculated that on average a person in the US visits nearly 1500 websites per year, and estimated that it would require almost 200 hours to read the privacy policies of all these websites. Even if one is willing to spend this time, a number of studies have shown that the average education required to comprehend a privacy policy is over high school level (Milne et al, 2004), or by other measures even at college-junior level (Srinath et al, 2020).

There have been multiple studies with the purpose of making it easier for people to read privacy policies. Most of these practices focus on summarization or classification. Either by manually analyzing them (Jensen and Pots, 2004; Bowers et al, 2017), through crowdsourcing (ToS;DR, 2012), or by using machine learning to automate the process. This last method has gained a lot of popularity in recent years. Harkous et al (2018) released the Convolutional Neural Network (CNN) based "Polisis", Liu et al (2018) reported on using Logistic Regression (LR), Support Vector Machine (SVM) and CNNs. Kumar et al (2019) explored the possibilities of FeedForward Networks and Qiu and Lie (2020) developed "Calpric", which is based on deep active learning.

In two of the above cases, the deep contextual model "BERT" (Devlin et al, 2018) has been used to measure against (Kumar et al, 2019; Qiu and Lie, 2020). In both cases the F1 scores of BERT were competitive with those of the respective models. An off the shelf BERT model even beat the FeedForward Networks from Kumar et al (2019) in four out of nine categories. This suggests a promising application of BERT for privacy policy classification. Another motivation for using BERT is the observation by Liu et al (2018) that encoding context might increase results.

In this paper I aim to train the best possible BERT model for privacy policy classification. One of the advantages of BERT is that it has been trained on 3.3 billion words and that one can use this pre-trained model for so-called "Transfer Learning": using the knowledge from a model for a specific downstream task. However, studies have shown that further pre-training BERT on a domain specific corpus can lead to increasing results in that domain (Lee et al, 2019; Beltagy et al, 2019). I plan to do this using a privacy policy corpus with over a million documents, created by Amos et al (2020). Furthermore I will look for the optimal hyper-parameters for privacy policy classification. On top of that I plan to use confidence scores of the classifications in response to Harkous et al's (2018) remark on the ambiguity of machine learning classifiers. In doing so, this research aims to add to a rapidly evolving literary field. Contributing to online transparency and the readability of privacy policies by creating better models.

2) Research Context

The research context for this proposal is divided into two parts. First (2.1) discusses an early historical overview of privacy policy related topics in the US and goes into further detail on the criticisms of natural language privacy policies. Then (2.2) focusses on machine learning applications for privacy policies as well as BERT.

2.1) Early historical overview

A privacy policy is a legal document that explains how a user's data is used by another party. In 1995 the European Parliament (1995) released a directive regarding privacy policies. The US released guidelines shortly thereafter (FTC, 1998). With the rise of the internet came more concerns about privacy. In the earlier years attempts were made by researchers to create machine readable privacy policies. In 1996 there was PICS, a project that tried to label online content (Resnick and Miller, 1996). In 2002 W3C's Platform for Privacy Preferences, or P3(P) launched a similar idea. This time with privacy as the main focus. P3P inspired other ideas, such as the european initiative of PRIME (PRIME 2004), or TAMI which was a project from MIT by Weitzner et al (2006). None of these projects gained much traction however, which makes natural language privacy policies the standard as of today.

As mentioned in the introduction these natural language policies are not without problems. There are two important criticisms that classification models aim to solve: the first is the length of documents. In 2006 Milne et al (2006) identified a trend of increased document length of website privacy policies. A recent large scale study by Amos et al (2020) confirms this. Where in 2009 the median word length of privacy policies was around 1600 words, In 2019 this doubled to 3200, with a sharper increase of length in recent years.

The second criticism is on the readability of policies: The study by Amos et al (2020) also shows an increase in difficulty of readability, they show that readability based on the FleschKincaid grade level (FKGL) (Kincaid et al, 1975) has increased from 12 to 13, equivalent to college level. Other studies show similar results (Jensen and Pots, 2004; Milne et al, 2006; Ermakova et al, 2015; Zimmeck et al, 2019; Srinath et al, 2020), where some even show results closer to a score of 14 (Srinath et al, 2020; Jensen and Pots, 2004), others, such as Ermakova et al (2015) and Zimmeck et al (2019) highlight the difference between fields and company sizes.

2.2) Machine learning

Because the machine readable initiatives of privacy policies did gain traction, researchers have adopted NLP techniques to process privacy policies. Since 2012 a nonprofit organization called "Terms of Service; Didn't Read" (ToS;DR, 2012) has used crowdsourcing to annotate privacy policies. They created an extension that is available for users to quickly acces this annotated data when on a website. This initiative led to the first few machine learning tools created for this field, as it provided valuable annotated data. In 2012 Ammar, Nadeh, Smith and Wilson (2012) retrieved 50 annotated policies from the ToS;DR website to create a classification tool.

Then in 2013 a number of researchers , including Nadeh, Smith and Wilson, launched the "Usable Privacy Policy Project" (Nadeh et al, 2013). They began this project to create more effective privacy notice and they have been responsible for multiple key developments in the field. For example, Ramanath et al (2014) used semi Hidden Markov Models to exploit the similarity of privacy policies. Or Zimmeck et al (2014), who created a browser extension called Privee, using annotated data from ToS;DR. This extension is still available, but does not function properly in many cases. Zimmeck et al (2014) concluded that there was a need for more and better structured annotated data. In 2016 Wilson et al (2016) created the OP-115 dataset, which contains 115 privacy policies and over 23K annotations.

This dataset gave rise to a number of new research, to mention a few: In 2018 Harkous et al (2018) used the OP-115 dataset in combination with 130K policies from Android apps to perform multi label classification in combination with CNNs. They released their results on a website and also created a chatbot called PriBot, which you can ask questions regarding privacy policies. In the same year Liu et al (2018) released a paper on automatic classification of privacy policies, using pre-trained word vectors from the works of Kim (2014). In the next year Kumar et al (2019) combined the OP-115 dataset with 150K unannotated privacy policies to create a classifier using FeedForward networks. They also used an off the shelf BERT model to compare their scores against. BERT gained competitive results, even though they did not perform any hyper-finetuning. Qiu and Lie (2020) created “Calpric”, a deep active learning model that has the benefit of only requiring a small amount of labeled data. Both in that aspect and in the F1 scores “Calpric” is similar to the off the shelf BERT model.

There were other initiatives that did not use the OP-115 dataset. For example, Nokbeh et al (2018) created their own annotated dataset, which they used in combination with Google prediction API to summarize policies.

Since BERT was introduced by Devlin et al (2018) it has been the focus point of much research. Liu et al (2019) with RoBERTa and Joshi et al (2019) with SpanBert (2019) have contributed to pre-training BERT. Lan et al (2020) released ALBERT, which scales better than BERT. Lee et al (2019) — BioBERT — and Beltagy et al (2019) — SciBERT — released models that create in-domain word embeddings. Thanks to the works of Zimmeck et al (2019), Amos et al (2020) and Srinath et al (2020), who have each created privacy policy datasets with over a million documents, it is now possible to create a privacy policy BERT model. On top of that there are multiple studies that discuss how to find the optimal parameters when fine-tuning BERT, such as the works of Li et al (2018), Dodge at al (2020) and Mosback et al (2020).

3) Methodology:

This research aims to train a BERT model for privacy policy classification. Using the OP-115 dataset (Wilson et al, 2016), which contains 115 English annotated privacy policies of websites, to fine-tune BERT. I will further domain specifically pre-train the model, as it has been shown that it significantly increases results (Beltagy et al, 2019; Lee et al, 2020), with the dataset created by Amos et al (2020) which contains a total of 1,071,488 English policies from websites. This research will use the same ten classification classes used by Liu et al (2018), which are in line with the annotations of the OP-115 dataset (Wilson et al, 2016). The classes are:

1) First Party Collection/Use
2) Third Party Sharing/Collection
3) User Choice/Control
4) User Access, Edit, & Deletion
5) Data Retention
6) Data Security
7) Policy Change
8) Do Not Track
9) International & Specific Audiences
10) Other

The classification of these classes will be done on text segments instead of sentences, as Liu et al (2018) showed increasing results. Taking into account the observation by Harkous et al (2018) that their work was susceptible to adversarial attacks it has to be researched if there is another way of extracting segments than the method proposed by Liu et al (2018), which uses keywords and is therefore extra susceptible to a change in vocabulary.

To optimize the hyper-parameters of my BERT model, I plan to primarily draw from the works of Mosback et al (2020), who concluded that instability of fine-tuning is due to vanishing gradient descents and not due to catastrophic forgetting, or small datasets, which was thought previously by Lee et al (2020) and Dodge et al (2020). To address this problem Mosback et al (2020) propose using small learning rates with bias correlation and to increase the numbers of iterations. To optimize time and computational efficiency I plan to use the " Hyperband" algorithm of Li et al (2018), this algorithm uses random search and is able to identify bad configurations early on. It claims be 5 to 30 times faster than the more classic Baysian approaches to hyper-parameter optimization

Similarly to Liu et al (2018), Harkous et al (2018) and Kumar et al (2019) I will validate the model by taking precision, recall and F1 scores on the OP-115 dataset using k-fold cross validation.

Citations:

Ammar, W. Wilson, S. Sadeh, N, A. Smith, N. (2012): *Automatic Categorization of Privacy Policies*: A Pilot Study. Carnegie Mellon University. Journal contribution. https://doi.org/10.1184/R1/6473072.v1

Amos, R. Acar, G. Lucherini, E. Kshirsagar, M. Narayanan, A. Mayer, J. (2020) *Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset*. arXiv 2008.09159.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: *A Pretrained Language Model for Scientific Text*. EMNLP/IJCNLP.

Bowers, J. Reaves, B. Sherman, I. Traynor, P. Butler, K. (2017). *Regulators, mount up! analysis of privacy policies for mobile money services.* In Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security (SOUPS '17). USENIX Association, USA, 97–114.

Bisong E. (2019) *Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform.* Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_7

Brown, T. Mann, B. Ryder, N. Subbiah, M. Kaplan, J. Dhariwal, P. et al (2020). *Language Models are Few-Shot Learners* Arxiv eprint: 2005.14165

Costante, Elisa & Sun, Yuanhao & Petković, Milan & Hartog, Jerry. (2012). *A machine learning solution to assess privacy policy completeness.* Proceedings of the ACM Conference on Computer and Communications Security. 91-96. 10.1145/2381966.2381979.

Devlin, J. Chang, M.-W. Lee, K. Toutanova, K. (2018) *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Dodge, J. Ilharco, G. Schwartz, R. Farhadi, A. Hajishirzi, H. Smith, N. (2020) *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*. arXiv eprint: 2002.06305

Ermakova, T. Fabian, B. Babina, E. (2015). *"Readability of Privacy Policies of Healthcare Websites".* Wirtschaftsinformatik Proceedings 2015. 73. http://aisel.aisnet.org/wi2015/73

European Union (1995) *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.* Official Journal L 281 , 23/11/1995 P. 0031 - 0050
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31995L0046

FTC (1998). *PRIVACY ONLINE: A REPORT TO CONGRESS. FEDERAL TRADE COMMISSION JUNE 1998* https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf

George R. Milne, Mary J. Culnan. (2004) *Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices,* Journal of Interactive Marketing, Volume 18, Issue 3, Pages 15-29, ISSN 1094-9968, https://doi.org/10.1002/dir.20009.

Harkous, H. Fawaz, K. Lebret, R. Schaub, F. Shin, K. G. Aberer, K. (2018) *Hamza Harkous and Kassem Fawaz and Rémi Lebret and Florian Schaub and Kang G. Shin and Karl Aberer.* arXiv eprint: 1802.02561

Jensen, C. Potts, C. (2004). Privacy policies as decision-making tools: An evaluation of online privacy notices. 471-478.

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., & Levy, O. (2019). *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. Transactions of the Association for Computational Linguistics, 8, 64-77.

Kim, K. (2014) *Convolutional neural networks for sentence classification*. arXiv preprint arXiv:1408.5882

Kincaid, J. Peter; Fishburne, Robert P. Jr.; Rogers, Richard L.; and Chissom, Brad S. (1975). *"Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel"* Institute for Simulation and Training. 56. https://stars.library.ucf.edu/istlibrary/56

Kumar, V.B. Ravichander, A. Story, P. Sadeh, N. (2019). *Quantifying the Effect of In-Domain Distributed Word Representations : A Study of Privacy Policies.*

Lan, Z. Chen, M., Goodman, S. Gimpel, K. Sharma, P. Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. ArXiv, abs/1909.11942.

Lee, J. Yoon, W. Kim, S. Kim, D. Kim, S. So, C, H. Kang, J. (2019). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*, *Bioinformatics*, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682

Lee, C., Cho, K., & Kang, W. (2020). *Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models*. ArXiv, abs/1909.11299.

Li, L. Jamieson, K. DeSalvo, G. Afshin, R. Talwalkar, A. (2018). *Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization*. arXiv eprint: 1603.06560

Liu, F. Wilson, S. Story, P. Zimmeck, S. Sadeh, N. (2018). *Towards Automatic Classification of Privacy Policy Text*. School of Computer Science Carnegie Mellon University. https://usableprivacy.org/static/files/CMU-ISR-17-118R.pdf

Liu, Y. Ott, M. Goyal, N. Du, J. Joshi, M. Chen, D. Levy, O. Lewis, M. Zettlemoyer, L. Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Facebook Ai, arXiv eprint: 2907.11692

McDonald, A, M. Cranor, L. (2008). *The Cost of Reading Privacy Policies.* I/S: A Journal of Law and Policy for the Information Society 2008 Privacy Year in Review issue http://www.is-journal.org/

Milne, George & Culnan, Mary & Greene, Henry. (2006). A Longitudinal Assessment of Online Privacy Notice Readability. Journal of Public Policy & Marketing - J PUBLIC POLICY MARKETING. 25. 238-249. 10.1509/jppm.25.2.238.

Mosback, M. Andriushchenko, M. Klakow, D. (2020) *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines*. arXiv eprint: 2006:04884

Peters, M. Neumann, M. Iyyer, M. Gardner, M. Clark, C. Lee, K. Zettlemoyer, L. (2018) *Deep contextualized word representation*. CoRR, volume: abs/1802.05365. Arxiv eprint 1802.05366.

PRIME (2004) *Privacy and Identity Management for Europe*. FP6-IST - Information Society Technologies: thematic priority under the specific programme "Integrating and strengthening the European research area" (2002-2006).
https://cordis.europa.eu/project/id/507591

Qiu, Wenjun & Lie, David. (2020). Deep Active Learning with Crowdsourcing Data for Privacy Policy Classification.

Ramanath, R. Liu, F. Sadeh, N. Smith, N. (2014). *Unsupervised Alignment of Privacy Policies using Hidden Markov Models*. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 2. 605-610. 10.3115/v1/P14-2099.

Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. ACM Trans. Internet Technol. 18, 4, Article 53 (July 2018), 18 pages. https://doi.org/10.1145/3127519

Resnick, P. Miller, J. (1996). *PICS: Internet access controls without censorship*. Commun. ACM 39, 10 (Oct. 1996), 87–93. DOI:https://doi.org/10.1145/236156.236175

Sadeh, N. Acquisti, A. Breaux, T, D. Cranor, L, F. McDonald, A, M. Reidenberg, J, R. Smith, N, A. Liu, F. Russel, C. Schauf, F. Wilson, S. (2013). T*he Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About*. Tech. report CMU-ISR-13-119, December

Srinath, M. Wilson, S .Giles, C. L. (2020) *Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies.* arXiv 2004.11131.

Stamey, J. W. Rossi, R.A. (2009). *Automatically identifying relations in privacy policies.* In Proceedings of the 27th ACM international conference on Design of communication (SIGDOC '09). Association for Computing Machinery, New York, NY, USA, 233–238. DOI:https://doi.org/10.1145/1621995.1622041

Terms of Service; Didn't Read (ToS;DR) (2012). http://tosdr.org/ index.html. Last accessed: February 2021

Weitzner, D.Abelson, H. Berners-Lee, T. Hanson, C. Hendler, J. Kagal, L. Mcguinness, D. Sussman, G. Waterman, K. (2006). *Transparent Accountable Data Mining: New Strategies for Privacy Protection.*

Zimmeck, S. Bellovin, S. M. (2014) *Privee: An Architecture for Automatically Analyzing Web Privacy Policies.* 23rd USENIX Security Symposium, USENIX Security 14. 978-1-931971-15-7, Pages 1-16

Zimmeck, S. & Story, P. Smullen, D. Ravichander, A. Wang, Z. Reidenberg, J. Russell, N. Sadeh, N.. (2019). *MAPS: Scaling Privacy Compliance Analysis to a Million Apps*. Proceedings on Privacy Enhancing Technologies. 2019. 66-86. 10.2478/popets-2019-0037.