Status Report: Critical Information Extraction from Terms of Services Document

Aditya Ashok Dave
Univ. of Southern California
daveadit@usc.edu

Akanksha Sanjay Nogaja
Univ. of Southern California
nogaja@usc.edu

Lavina Lavakumar Agarwal
Univ. of Southern California
llagarwa@usc.edu

Shreya Venkatesh Prabhu
Univ. of Southern California
prabhus@usc.edu

Sai Sree Yoshitha Akunuri
Univ. of Southern California
akunuri@usc.edu

Abstract

The work done so far involves getting acquainted with what has already been implemented including the different preprocessing steps, models and the intricacies of the code, defining a baseline model and identifying areas of improvement to the existing work. Further we have discussed, and evaluated the challenges that we have till now, along with the anticipated challenges in future work.

1 Tasks

As a first step, we tried BERT Double, Legal BERT and Custom Legal BERT models mentioned in (L. Zheng et al., 2021) for fairness classification. These models are available in HuggingFace library https://huggingface.co/. The scores obtained are documented in table 1

Model	F1 Score
BERT Double	0.76
Legal BERT	0.753
Custom Legal BERT	0.78

Table 1: BERT Variations on ToS Dataset

In CLAUDETTE (Marco Lippi et al., 2018), the authors have identified unfair clauses using Support Vector Machines. We tried SVM for the ToS dataset by creating TF-IDF features.

ToS dataset http://claudette.eui.eu/ToS.zip was first subjected to preprocessing techniques such as converting dataset to lower case and removing accents, html tags, urls, and non alphabetical characters. TF-IDF features were created using unigrams, bigrams and trigrams for words and parts of speech tags combined. The best model was selected using leave one document out cross validation(LODOCV) technique. As per this technique, among 50 Terms of Service documents, each containing multiple clauses, 1 document was kept aside for testing purposes. Among the rest of the documents, cross validation was performed to select the best model by using 48 documents for training and 1 document for validation. The process was repeated 50 times and testing was done for each document. The test score was aggregated by computing the macro average of F1 scores obtained in each iteration of LODOCV. Test scores of important experiments can be found in table 1

Ngrams	F1 Score
(1,1)	0.8474
(1,2)	0.8714
(2,3)	0.8712
(1,3)	0.8807

Table 2: SVM with TF-IDF on ToS Dataset

TF-IDF features based on words with unigrams, bigrams and trigrams performed the best and gave the highest F1 Score of 0.88.

Next, we worked on generating word embeddings and features using BERT (Jacob Devlin et al., 2019) and RoBERTa (Yinhan Liu et al., 2019). The ToS is preprocessed using techniques similar to the SVM models, and additionally expansion of contractions was also done. The preprocessed data is then passed to the BERT Tokenizer from pretrained model "bert-base-uncased" available in HuggingFace transformers library. The generated tokens are then passed to the BERT Model, and embeddings are obtained. The BERT model has a fixed length of 512 tokens, so padding tokens were added to the tokens list. The retrieved embedding dimensions for a single clause are 512x768. A similar approach is used to create embeddings using the RoBERTa Model (pretrained roberta-base) from the HuggingFace transformers library. The retrieved embeddings are then averaged to create a sentence vector of dimension 1x768. The data is split into training and test using the 80-20 split

ratio, and fed into a non-linear SVM model with RBF kernel. The results for BERT and RoBERTa embeddings can be found in table 3.

Classifier	F1 Score
Avg. BERT Embeddings	
Non-linear SVM	0.59(macro)
Non-Linear SVM(LODOCV)	0.58(macro)
RoBERTa Embeddings	
Non-linear SVM	0.52(macro)
Non-Linear SVM(LODOCV)	0.51(macro)

Table 3: SVM with BERT & RoBERTa

Additionally, we started identifying obligatory clauses. For this, ToS dataset (9414 clauses) was first preprocessed using techniques similar to SVM along with lemmatization. A custom Word2Vec model was trained on this data, and was used along with the Google Word2Vec model to identify keywords referring to obligations. Once the keywords were identified, all the clauses that contain any of these keywords were tagged as Obligatory clauses. A total of 2747 clauses were identified as obligatory, which is around 29% of the total data. These clauses may or may not be obligatory to users. So, future work involves filtering the clauses that are obligatory to the users.

2 Risks and Challenges

One of the challenges we faced while training the SVM models was that the classes for fair and unfair were not proportionate and lead to class imbalance, which was handled using class weights.

The results obtained from SVM with simple TF-IDF features are better than the ones obtained with BERT/RoBERTa embeddings as input features. This indicates that complex techniques or models do not necessarily perform better. BERT and RoBERTa also have constraints on input token size. Any input higher than 512 tokens will have to be truncated. Moreover, the actual dimensions of the dataset obtained from BERT embeddings are 9414x512x768 which makes it computationally expensive to train neural networks on such a huge data.

As a part of identifying obligatory clauses to users, we need to map the identified obligatory clauses to entities (users/service provider). The main challenge lies in mapping the obligations to respective entities when a clause is entitled to both users and service providers.

Another major challenge is the model marking all clauses as critical information and returning the entire terms of services document to the user. The aim of the project is to reduce the content a user has to read by extracting only critical clauses.

3 Plan To Mitigate Risks and Challenges

Since BERT models have a fixed sequence length of 512 tokens, we are planning to try a few common techniques such as considering the first 512 tokens, the last 512 tokens or selecting random 512 tokens from each clause and compare the scores obtained.

Training a neural network with a huge dimensional dataset obtained from BERT embeddings is computationally expensive, hence one of the ideas we would like to employ is to count the multi-word tokens (e.g. running = ”run”, ”##ing”) as single and take the average of their embeddings. On top of that we are planning to apply PCA technique to further reduce the dimensionality for faster training. It should be noted that reducing dimensionality is a trade-off and might lead to loss of some contextual information.

For mapping obligatory clauses to entities, we plan to use dependency tree to find the subject of the clause, and then use Breadth First Search to navigate the tree and find all the tokens that are related to the subject by a parent-child relationship.

For the issue where the entire document is classified as critical information, we can compute the overall ToS unfairness score for the document as mentioned in (Guarino A. et al., 2021). Here, the authors got human experts to provide weights to various categories of unfair clauses and computed overall criticality scores for the document. A high score generally indicates that the terms are unfair and a user must seriously consider whether to proceed or not.

Apart from executing the plans for mitigating the challenges mentioned above, we would further like to finetune our preprocessing steps to take into account the special segments in the dataset, and apply various classification models. Additionally, we plan to experiment neural networks models for fairness classification on the features generated from BERT, RoBERTa and XLNet (Zhilin Yang et al., 2019). An ensemble can make better predictions and achieve better performance than any single contributing model. So we plan to ensemble our best performing models.

References

[L. Zheng et al.2021] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho 2021. *When does pretraining help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset.* Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law

[Marco Lippi et al.2018] Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, Paolo Torroni. 2018. *CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service*. Artificial Intelligence and Law.

[Jacob Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*. https://aclanthology.org/N19-1423

[Yinhan Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.*. arXiv:1907.11692 [cs.CL]

[Guarino A. et al.2021] Guarino A., Lettieri N., Malandrino D. et al. 2021. *A machine learning-based approach to identify unlawful practices in online terms of service: analysis, implementation and evaluation.* Neural Comput & Applic 33, 17569–17587

[Zhilin Yang et al.2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized autoregressive pretraining for language understanding.* . arXiv preprint arXiv:1906.08237.