# Applying NLP to simplify Terms of Service Agreements

*Emre* Iyigün, *Ali* Moutyrek

**Abstract.** In the digital era, the complexity of Terms of Service (ToS) agreements often leads to user neglect, resulting in risks during digital interactions. We focus on making ToS agreements transparent and empowering users for informed decision-making by applying NLP. State-of-the-art methodologies in this domain include the use of large language models for summarization and simplification. Our approach incorporates text summarization, both extractive and abstractive, along with keyword extraction. We utilize models like BART and LED for abstractive summarization, assessed for fluency, accuracy, coherence, and readability through human evaluation. Our findings reveal that abstractive summarization, particularly through the fine-tuned BART model, showed overall good performance based on our evaluation technique in making ToS agreements more comprehensible for the user. It effectively transformed complex legal terms into simpler language while maintaining the essence of the agreements. This approach not only simplifies the ToS agreements but also aids in establishing a more informed user base, thereby contributing to safer and more transparent digital interactions.

## 1 Introduction

In the digital era, where interactions and transactions are predominantly online, the Terms of Service (ToS) agreements are fundamental in defining the legal framework between users and service providers. However, a critical issue has emerged: the **pervasive neglect by users** in reading these agreements. This phenomenon is not just a matter of concern but poses a substantial risk in the digital landscape.

A pivotal study that sheds light on this issue is "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services" by Jonathan A. Obar and Anne Oeldorf-Hirsch [1]. **Their research provides compelling evidence that the vast majority of users do not read ToS agreements**. By agreeing to terms without understanding them, users may unknowingly consent to privacy invasions, data sharing, and other actions that they might otherwise contest.

This alarming trend underscores the necessity of our work. By applying Natural Language Processing (NLP) techniques to simplify ToS agreements, we aim to make these agreements **more ac-**

**cessible and understandable to the user**. Our goal is to increase transparency and enhance user comprehension, thereby empowering users to make informed decisions.

Various approaches are applied to simplify of ToS agreements. We will delve into specific methods and discuss how these methods are evaluated.

## 2 Related work

Several works have addressed the challenge of simplifying ToS agreements. Here's a brief overview of these notable contributions:

- [2]: This paper by Luger et al. delves into the complexity of ToS agreements. It emphasizes the need for user consent and the challenges users face in understanding these documents. The study is a critical examination of how ToS are presented and how this impacts user comprehension and consent.
- [3]: Azose's work explores the use of advanced large language models for generating concise summaries of website ToS. This approach aims to make these often lengthy and complex documents more accessible and understandable to the user.
- [4]: Manor and Li's research focuses on summarizing contracts into plain English. This work is crucial in making legal documents, including ToS, more comprehensible to non-experts. The approach involves transforming legal jargon into simpler language, thereby enhancing transparency and understanding.
- [5]: This paper addresses the summarization of legal obligations, entitlements, and prohibitions in contracts, focusing on the specific needs of each party involved. The work by Sancheti et al. is in its approach to tailor summaries to the interests and responsibilities of different stakeholders, providing a more targeted understanding of legal documents.
- [6]: Perera and Perera's paper introduces a machine learning and NLP framework to transform complex legal documents into plain English, aiming to enhance accessibility and highlight key aspects for users.

## 3 Methodology

We delved into two different methods for simplifying ToS agreements, focusing primarily on text summarization. These methodologies were categorized into groups:

- Text summarization: **hybrid**, **abstractive**, and **extractive summarization**
- **Keyword extraction**

**Text summarization involves condensing longer texts into shorter, cohesive summaries that encapsulate the main ideas.** This process aims to enable swift comprehension of extensive documents, retain pertinent information, and enhance information retrieval efficiency [7].

Extractive summarization refers to **selecting important sentences** from a piece of text and showing them together as a summary whereas abstractive summarization refers to the task of **generating an abstract of the text** i.e., instead of picking sentences from within the text, a new summary is generated [7]. Hybrid summarization combines the extraction of key content with the generation of new, coherent summaries [8]. **Keyword extraction involves identifying important words that capture the essence of the text** [9].

In the following the used techniques are presented and explained.

- Keyword extraction
  - RAKE: Based on [9], RAKE (Rapid Automatic Keyword Extraction) algorithm starts by identifying phrases using stopwords and punctuation as delimiters, isolating potential keywords. Each word within these phrases is then scored based on its frequency and degree of co-occurrence with other words. RAKE ranks these phrases by their word scores, with higher scores indicating greater significance. The top-ranked phrases, adhering to our two-word limit, are extracted as keywords. For this, the rake-nltk Python library was used (https://pypi.org/project/rake-nltk/).
- Extractive techniques
  - RAKE was also utilized to extract sentences based on predefined obligation words. For this purpose, a list of obligation-related terms in both English and German was compiled, including words like "must", "shall", and "is required to". The text was then systematically split into individual sentences using regular expressions. Applying the RAKE algorithm to each sentence, keyphrases were extracted, and those containing the specified obligation words were identified. Then keyphrases are put together to form a summary.
  - Latent Semantic Analysis (LSA): Based on [10, 11] it works by first creating a term-document matrix to represent term frequencies across documents or text segments. It then applies Singular Value Decomposition (SVD) to this matrix, reducing its dimensionality and highlighting significant term relationships. This decomposition reveals latent concepts or themes, essentially clusters of terms that frequently co-occur, indicating the main ideas in the text. LSA identifies and extracts key sentences or passages that best represent these latent concepts. Finally, it refines the selection using cosine similarity measures in the reduced space, ensuring the summary is both comprehensive and non-redundant. For this, the LsaSummarizer from the sumy Python library was used (https://github.com/miso-belica/sumy).
  - TextRank: This algorithm that operates on the principle of graph-based ranking, similar to Google's PageRank [12]. In TextRank, the text is represented as a graph with sentences as nodes. Connections between these nodes are established based on the similarity between sentences. The

algorithm then scores each sentence based on its connections, with highly connected sentences receiving higher scores. These scores help in identifying the most important sentences in the text. The highest-scoring sentences are selected to form the summary, ensuring that they collectively represent the main themes of the original text while maintaining coherence. For this, the TextRankSummarizer from the sumy (https://github.com/miso-belica/sumy) Python library was used.

- Abstractive techniques
  - BART: Based on [13], BART (Bidirectional and Auto-Regressive Transformers) combines two key mechanisms: a bidirectional encoder (like BERT) that reads the input text and an autoregressive decoder (like GPT) that generates the summary. In abstractive summarization, BART rephrases and condenses the original text, generating new sentences that capture the core meaning while potentially introducing novel words and phrases not found in the source. For summarization, the BART-large model, specifically the variant trained on the CNN/Daily Mail dataset was utilized (https://huggingface.co/facebook/bart-large-cnn). Additionally, this model was fine-tuned using a specialized dataset comprising full-text English ToS and corresponding human-annotated summaries (https://huggingface.co/datasets/EE21/ToS-Summaries). This dataset was developed by scraping English ToS agreements along with their human-annotated summaries from tosdr.org.
  - LED: LED (Longformer Encoder-Decoder) is a specialized variant of the transformer model, adeptly designed for abstractive summarization tasks, such as the 'led-base-book-summary' which was used in this work (https://huggingface.co/pszemraj/led-base-book-summary). This model is uniquely adapted to process long documents, a challenge often encountered with standard transformer models. The key feature of LED is its sliding window attention mechanism [14]. This approach allows the model to focus on smaller, manageable segments of a long document at a time, making it efficient in handling extensive texts without losing contextual information.

  - LongT5: LongT5 is an extension of the original T5 (Text-to-Text Transfer Transformer), specifically designed to handle longer text sequences [15]. It builds upon the T5's framework, which reformulates various NLP tasks into a unified text-to-text format. LongT5's key advancement lies in its ability to process and summarize much longer documents than the standard T5 model. In this work the LongT5 that was fine-tuned on scientific literature, was used (https://huggingface.co/pszemraj/long-t5-tglobal-base-sci-simplify).
- Hybrid: In this method, first RAKE is applied as it is explained in "Extractive techniques", to generate an initial summary. Following this, the fine-tuned BART model was used to further condense and refine this summary.

In our work, we grappled with the challenge of evaluating the quality of the generated summaries. This evaluation is complex due to several factors:

- Quality Metrics Limitations: Commonly used metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) only capture overlap in terms of n-grams between the generated summary and reference summaries [7, 16]. They do not adequately measure semantic coherence, factual accuracy, or readability, which are crucial for summarization quality.
- Dependence on reference summaries: Many evaluation methods rely on comparing generated summaries with human-written reference summaries. However, the quality and representativeness of these reference summaries can vary, affecting the reliability of the evaluation.
- Challenges with factual consistency: Ensuring that the summarized content accurately reflects the facts in the original text is a significant challenge, especially for abstractive summarization models that generate new sentences. This issue is critical for maintaining the reliability and trustworthiness of the summaries [17].
- Length Bias: Some evaluation metrics can be biased towards longer or shorter summaries, which may not accurately reflect the quality of the content. This bias can lead to misleading evaluations, favoring summaries that match the length preferences of the metric rather than the actual content quality [18].

Because of the limitations mentioned above we propose a simple human-based evaluation to determine the best presented technique to simplify ToS agreements. This evaluation method aims to address the shortcomings of automated metrics by incorporating human judgment and comprehension. In the following, the evaluation process is explained.

1. Selection of ToS documents: Randomly selecting 5 to 10 ToS documents provided by [19].
2. Reading and summarization: Each selected ToS document is read thoroughly by different human reader. Subsequently, summaries are created from all the techniques mentioned.
3. Evaluation criteria and rating scale: The summaries are evaluated based on four key criteria, each rated on the scale 1 (poor), 2 (fair), 3 (good) to 4 (excellent):
   - Fluency: Assessing how smoothly the summary reads.
   - Factual Accuracy: Determining whether the summary accurately reflects the original ToS content.
   - Coherence: Evaluating if the summary is logically structured and easy to follow.
   - Readability: Judging how easy it is to understand the summary.
4. Summing individual scores: For each generated summary, the scores across all criteria were summed up to get a score for a technique. Then the total score for each technique is obtained by summing the individual scores across all generated summaries. This total score represents the overall performance of the technique in terms of the combined criteria.
5. Best technique selection: The technique with the highest total score is considered the best performer in our evaluation.

## 4 Discussion

The keyword extraction method was insufficient for simplifying ToS agreements, as it failed to represent facts. As a result, we have excluded it from further consideration.

In total, 15 ToS agreements were evaluated by three group members, each randomly selected five ToS agreements. The overall performance is shown in Figure 1.

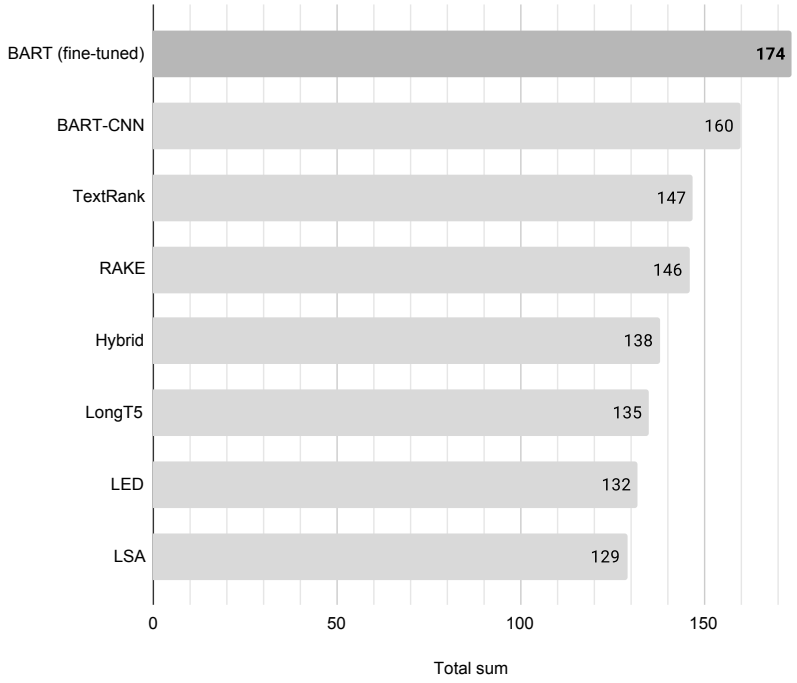


**Figure 1.** Comparison of the overall performance of different techniques

Figure 1 offers a succinct evaluation of summarization techniques, highlighting a hierarchy in performance with BART (fine-tuned) at the apex. TextRank and RAKE, while simpler, yield competitive scores, reflecting their algorithmic efficiency. LongT5 and LED, despite their lower scores, remain functional, perhaps more suited for tasks beyond the scope of this assessment. LSA's lower-end score could indicate its comparative inadequacy in dealing with complex summarization tasks. The figure indicates that fine-tuning on specific datasets enhances performance in summarizing ToS agreements.

Figure 2 presents a comparison across different summarization techniques. In this comparison, BART variations demonstrate a robust equilibrium, performing well across all measured aspects. This suggests that while BART models are fine-tuned, they maintain a strong level of performance that does not heavily favor one criterion over another. On the other hand, techniques that focus on

extraction are particularly noted for their high factual accuracy. This could be due to their method of directly utilizing source material, which inherently preserves the factual content. The bar graph visually represents these findings, allowing for a clear comparison between the different methods and their respective strengths and weaknesses in each criterion.

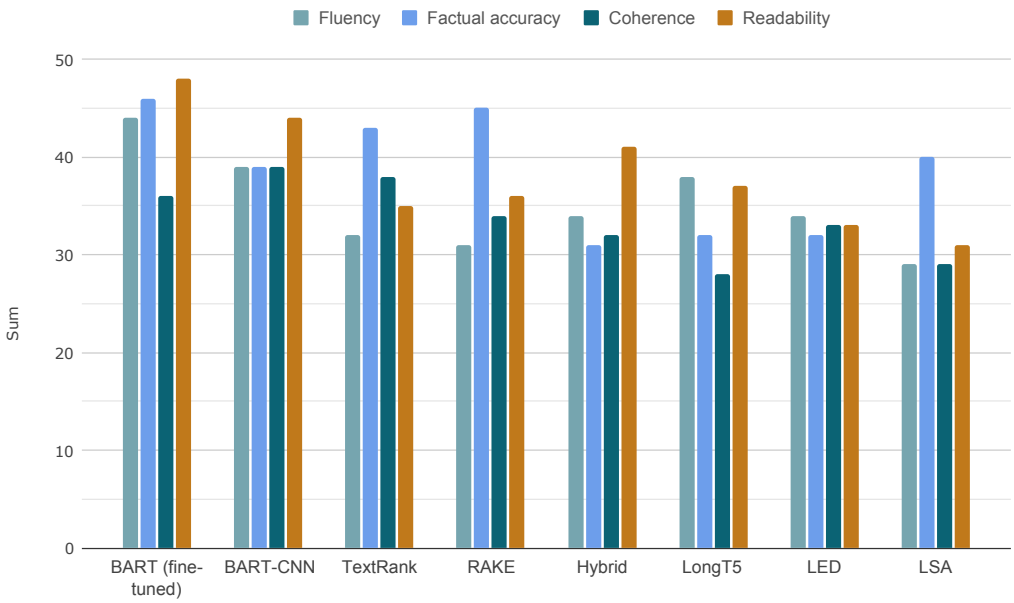


**Figure 2.** Comparison of the performance by criteria of different techniques

Figure 3 displays the mean performance of abstractive, extractive, and hybrid summarization methods across fluency, factual accuracy, coherence, and readability. Abstractive methods score high in fluency and readability but slightly lower in factual accuracy due to rewriting. Extractive summarization leads in factual accuracy due to only extracting scentences but falls behind in fluency and readability. Hybrid methods show balanced performance across all criteria, suggesting a compromise between the fluency and readability of abstractive methods and the factual accuracy of extractive methods.

It can be observed that the performance of abstractive summarization stands out as the highest among the evaluated methods.

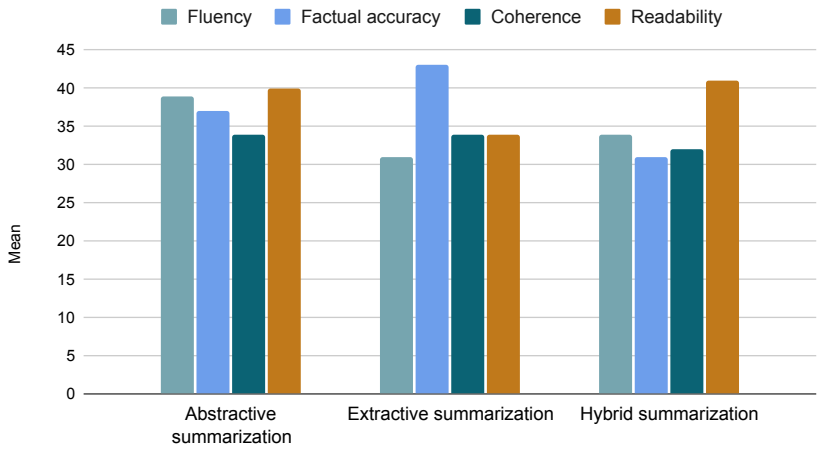


**Figure 3.** Comparison of the average performance of different methods by criteria

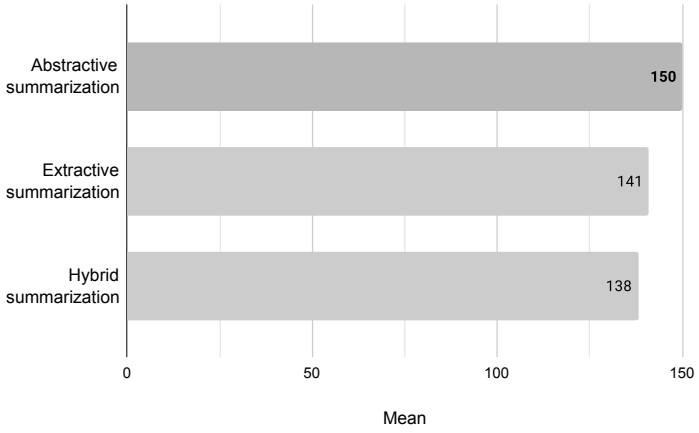


**Figure 4.** Comparison of the average performance of different methods

## 5 Conclusion

While the abstractive models have even points at all the criteria, the extractive accelarate at the factual accuracy. It is up for discussion, if all the facts in the ToS are necessary for the users.

An effective approach to improve summarization could involve a synergistic combination of abstractive and extractive methods, harnessing the strengths of each. For instance, the pointer-generator network, as discussed [20], presents an opportunity to enhance abstractive summarization.

While the proposed evaluation assigned equal weight to all criteria, a differential weighting for these

criteria could be considerd. Additionally, exploring the possibility of expert evaluation could enable the identification of strengths and weaknesses in each technique's generated summaries.

## References


[1] J.A. Obar, A. Oeldorf-Hirsch, *The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services* (Routledge, 2020), Vol. 23, pp. 128–147, https://doi.org/10.1080/1369118X.2018.1486870

[2] E. Luger, S. Moran, T. Rodden, *Consent for All: Revealing the Hidden Complexity of Terms and Conditions*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2013), CHI '13, p. 2687–2696, https://doi.org/10.1145/2470654.2481371

[3] B. Azose, *Generating website terms of service summary using a large language model* (2023), https://www.tdcommons.org/dpubs_series/6107

[4] L. Manor, J.J. Li, *Plain English Summarization of Contracts*, in *Proceedings of the Natural Legal Language Processing Workshop 2019*, edited by N. Aletras, E. Ash, L. Barrett, D. Chen, A. Meyers, D. Preotiuc-Pietro, D. Rosenberg, A. Stent (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), pp. 1–11, https://aclanthology.org/W19-2201

[5] A. Sancheti, A. Garimella, B. Srinivasan, R. Rudinger, *What to Read in a Contract? Party-Specific Summarization of Legal Obligations, Entitlements, and Prohibitions*, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by H. Bouamor, J. Pino, K. Bali (Association for Computational Linguistics, Singapore, 2023), pp. 14708–14725, https://aclanthology.org/2023.emnlp-main.909

[6] T. Perera, T. Perera, *Barrister-Processing and Summarization of Terms Conditions / Privacy Policies*, in *2021 6th International Conference for Convergence in Technology (I2CT)* (2021), pp. 1–7

[7] S. Vajjala, B.P. Majumder, A. Gupta, H. Surana, *Practical Natural Language Processing*, 1st edn. (O'Reilly Media, Inc., 2020)

[8] A. Nenkova, K. McKeown, Foundations and Trends in Information Retrieval **5**, 103 (2011)

[9] S. Rose, D. Engel, N. Cramer, W. Cowley, in *Text Mining: Applications and Theory*, edited by M.W. Berry, J. Kogan (Wiley, 2010)

[10] P. Kherwa, P. Bansal, *Latent Semantic Analysis: An Approach to Understand Semantic of Text*, in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)* (2017), pp. 870–874

[11] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Journal of the American Society for Information Science **41**, 391 (1990)

[12] R. Mihalcea, P. Tarau, *TextRank: Bringing Order into Text*, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, edited by D. Lin, D. Wu

(Association for Computational Linguistics, Barcelona, Spain, 2004), pp. 404–411, https://aclanthology.org/W04-3252

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Association for Computational Linguistics, Online, 2020), pp. 7871–7880, https://aclanthology.org/2020.acl-main.703

[14] I. Beltagy, M.E. Peters, A. Cohan, arXiv preprint arXiv:2004.05150 (2020)

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, arXiv e-prints arXiv:1910.10683 (2019), 1910.10683

[16] Y. Liu, M. Lapata, *Text Summarization with Pretrained Encoders*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019)

[17] Z. Cao, F. Wei, W. Li, S. Li, *Faithful to the Original: Fact-Aware Neural Abstractive Summarization*, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (AAAI Press, 2018), AAAI'18/IAAI'18/EAAI'18, ISBN 978-1-57735-800-8

[18] P. Sountsov, S. Sarawagi, *Length bias in Encoder Decoder Models and a Case for Global Conditioning*, in *Conference on Empirical Methods in Natural Language Processing* (2016), https://api.semanticscholar.org/CorpusID:39487

[19] P. Palka, K. Wiśniewska, R. Pałosz, A. Porębski, *Annotated terms of service of 100 online platforms* (2023)

[20] J. Deaton, *Transformers and Pointer-Generator Networks for Abstractive Summarization* (2019), https://api.semanticscholar.org/CorpusID:204778440