# PPOL 6801 - Text as Data

**Final Report:** Muhammad Saad (ms4689)

**Differentially Private Topic Modelling**

*Mapping Privacy Risks & Prototyping Secure Algorithms*

## Introduction

Latent Dirichlet Allocation (LDA) is a widely used technique for uncovering latent structures in text corpora. LDA is a standard tool in natural language processing, political text analysis, and computational social science. It is frequently used to summarize documents, discover topics, and reduce high-dimensional text data into interpretable representations. However, generative text models like LDA can pose privacy risks. Since LDA is often trained on text contributions made by individuals, the parameters learned by the algorithm can include sensitive or identifiable information about individual contributors.

Recent work has confirmed that topic models may not be privacy-preserving. Evidence from relevant literature shows that text-based generative models can memorize elements of the training data, and potentially lead to leakage of sensitive information (Carlini et al., 2023). Manzonelli et al. (2024) demonstrate that similar vulnerabilities arise in LDA, which is susceptible to membership inference attacks.

Differential privacy (DP) provides a formal framework for quantifying and mitigating such privacy risks (Dwork & Roth, 2014). It is a mathematical definition of privacy that not only protects against privacy leakages in computations and algorithms, but also comes with a tuneable parameter, i.e. epsilon (ε) which can be calibrated to control the level of privacy provided. By introducing calibrated randomness, or '*noise*', into outputs, DP masks how any single individual's data influences model outputs. Several studies (Zhao, 2020; Huang and Chen, 2021; Manzonelli et al., 2024) have explored the integration of differential privacy in LDA. There approach differ along three dimensions:

1. The learning method used in LDA, including whether it is Collapsed Gibbs Sampling (Zhu et al., 2016; Huang & Chen, 2021; Huang et al., 2022; Manzonelli et al., 2024), Variational Inference (Park et al., 2018), or Spectral Approximation (DeCarolis et al., 2020).
2. Where noise is injected into the pipeline to perturb aggregate outputs from LDA, such as topic–word counts (Zhao et al., 2020; Huang & Chen, 2021), document–topic counts, or vocabulary construction (Manzonelli et al., 2024).
3. The level of privacy extended and how sensitivity is calculated within that context. Sensitivity in this case defines the maximum change that can be made by the primary unit of analysis, whether that is an individual or a document, to the aggregate LDA outputs. Some studies focus on document-level privacy (Zhu et al., 2016; Park et al., 2018; DeCarolis et al., 2020), while others focus on word-level privacy (Zhao et al., 2020; Huang & Chen, 2021).

## Project Motivation

While these approaches provide formal privacy guarantees using DP, the literature does not offer much intuition about how different design choices affect utility, interpretability, and stability of the outputs generated from LDA. Although these studies mount different privacy attacks, including membership inference and reconstruction, they do not give an interpretable overview of privacy risks that are present in the input datasets and the LDA pipeline.

This project does not have the same level of sophistication as the studies cited above. Instead, it aims to give an intuitive and interpretable understanding of how DP can be integrated within the LDA algorithm. It is motivated by four considerations: i). What are some of the privacy risks that are present in text data; ii). How do these privacy risks translate into the model outputs generated by a text analysis technique like LDA; iii). How can a simplistic differentially-private LDA pipeline be developed to afford a basic level of privacy guarantee; iv). How do different levels of noise affect the utility, specifically the interpretability of the outputs generated by LDA. Using the 20Newsgroups dataset, the analysis first demonstrates privacy risks in non-private LDA. It then implements a simplified DP algorithm by injecting Laplace noise into Collapsed Gibbs Sampling (CGS) LDA variant.

## Data

### Data Source and Unit of Observation

The analysis uses the 20Newsgroups dataset (scikit-learn developers, 2017), a publicly available corpus containing approximately 18,000 discussion posts from early internet newsgroups. The dataset comprises 20 topical categories such as politics, science, religion, and technology. Each document corresponds to a single post made by a user to a topic-specific discussion group. The posts included in the dataset are from 1993 to 1995. Although the dataset is commonly treated as anonymized, many posts contain usernames, email identifiers, or highly distinctive content. It has also been extensively used in existing literature which explores DP applications for LDA (Huang & Chen, 2021; Manzonelli et al., 2024). Text-based data made available from platforms such as Facebook or X is usually pseudonymized and several privacy protocols have been put in place. However, this is not the case of the 20Newsgroups dataset. It includes the full gambit of posts made by users, including their personally identifiable information. This is why the dataset is unique and highly suited for assessing privacy risks in topic modelling.

The unit of analysis is a post. This is a deliberate choice since it is easier to analyze a post or document-level privacy, instead of user-level or vocabulary-level privacy. The key LDA output of interest is topic-word distributions, instead of document-topic distributions and the learned vocabulary. Topic-word distributions are more intuitive to understand and are in line with the interpretability focus on the project. Because users may contribute multiple posts, the analysis also considers the user as a higher-level unit when evaluating privacy risks and sensitivity assumptions.

### Text Preprocessing & Wrangling

Text preprocessing included standard steps such as tokenization, lowercasing, removal of punctuation and stopwords, and filtering of extremely rare terms. The cleaned corpus was then converted into a document-term matrix suitable for LDA. Preprocessing choices remained the same across all models to ensure that any differences in results were attributable only to the privacy mechanisms rather than data preprocessing.

Apart from preprocessing steps and the exclusion of documents that contain no remaining tokens after cleaning, no additional data wrangling was performed. All remaining newsgroups posts were retained to resemble the original corpus. After preprocessing, the corpus consisted of cleaned documents, where each document corresponds to a single post by a user.

## Analytical Methods

**Collapsed Gibbs Sampling and Latent Dirichlet Allocation**

The project focuses on Latent Dirichlet Allocation (LDA), which is estimated using the Collapsed Gibbs Sampling (CGS) method. CGS is a commonly used procedure within the context of LDA and has also been the focus of existing DP-LDA studies (Zhao et al., 2020; Huang & Chen, 2021; Huang et al., 2022). The primary output of interest from the CGS-based LDA in this project is the topic-word distribution matrix (Φ). Φ encodes the probability of observing each word given a topic and forms the basis for interpreting the content of topics. Changes in Φ across different privacy settings are used as a proxy for model stability and interpretability, as even small changes in Φ can tweak the composition of top words within different topics.

In addition to Φ, the project also evaluates utility (within the context of DP privacy-utility tradeoff) through topic coherence measures. These quantify the consistency of words within each topic across different noise levels. Topic coherence is a simple metric for assessing topic quality and allows for comparison under different noise levels. Inspection of top-ranked words within topics was also used to complement coherence and to assess how noise affects interpretability.

**DP-LDA Mechanism**

DP is implemented by adding noise to the output statistics generated as part of the Collapsed Gibbs Sampling (CGS) for LDA. Specifically, Laplace distribution sampled noise is injected into the topic-word count matrix before computing the topic-word distribution parameter Φ as part of LDA. In standard CGS LDA, topic-word counts aggregate the number of times each word is assigned to a topic across the text data. Because these counts directly determine Φ, adding noise here provides a simplistic and interpretable mechanism for masking individual contributions on model outputs. This choice is in line with prior work on DP-LDA using CGS, which injects noise at Φ counts rather than for document-topic or vocabulary related outputs (Zhao et al., 2020; Huang & Chen, 2021).

More specifically, let $n_{kw}$ denote the actual Φ count for topic $k$ and word $w$. The DP mechanism used in the project releases a '*noisy*' version $\hat{n}_{kw} = n_{kw} + \eta_{kw}$, where $\eta_{kw}$ is noise drawn independently from a Laplace distribution with mean zero and scale parameter ($b=\Delta/\varepsilon$). Here, $\varepsilon$ denotes the privacy budget. The smaller the $\varepsilon$, the higher the noise injected into the Φ counts with the highest noise at $\varepsilon$ equals to zero. $\Delta$ is the sensitivity of the topic-word counts. Since most all the analysis is done in R, which does not have off-the-shelf and easy to use DP libraries such as OpenDP, Laplace noise is sampled using the standard parameterization *rlaplace*. This ensures that noise increases monotonically as privacy protection strengthens, or as $\varepsilon$ decreases. The use of *rlaplace* to mimic DP algorithm is another unique addition made by the project. The code and algorithm developed through it can be used by other researchers who want to implement DP using R, not Python. The noisy counts are then used to compute Φ.

### Initial Sensitivity Assumption

DP is implemented under a clearly defined sensitivity assumption before any results are examined. For the baseline, two datasets are considered adjacent if they differ by a single document. Under this document-level definition, the contribution of any one post to the Φ count matrix is limited to at most one unit. As a result, sensitivity is initially set to $\Delta = 1$. This is a simple starting point for understanding how the privacy parameter $\varepsilon$ affects Φ output.

## Results

### Initial Privacy Risk Assessment

In order to initially assess the risk of privacy leakages, the author looked at the 20 most active users with the 20Newgroups dataset (for more details, see Annex A). These were users who not only had the highest number of posts, but whose metadata also included personally identifiable information, in particular email addresses.

A quick online search by the author showed that two of the top three posters' email addresses could be traced back to actual individuals. In the case of '*sera@zuma.UUCP*', it was Ahmet Cosar (Cosar, n.d.), and for '*henry@zoo.toronto.edu*', Henry Spencer (Levithan, n.d.). This simple online research for publicly available email addresses shows the inherent privacy risk present in the 20Newsgroups data. Further analysis showed that these users consistently post in only particular groups. For Cosar it was the '*talk.politics.mideast*', and for Spencer the '*sci.space*' and '*sci.electronics*' groups. Not only that, these users almost always posted related specific issues within these groups. For example, Cosar's posts were mostly focused denying that Turkey had any role to play in the Armenian genocide and discrediting Armenians (posts such as '*Muslim women and children were massacred by the Armenians' or 'Nazi Armenian Philosophy: Race above everything and before everything'*). Therefore, these publicly identifiable authors leave a unique digital fingerprint that can be traced back to them.

Apart from their email addresses being publicly available and the unique content they generate, users like Cosar and Spencer are also outliers. Most of the users on 20Newsgroups made only a single post, whereas for these two individuals the number of posts were as high as 80 and 72 respectively. The distribution of the number of posts per user resembles a long tail. From a privacy perspective, this is important because some unique users immediately stand out by the volume of their posts. Any text analysis technique which is generative will be heavily reliant on these large-volume users. The training data will be highly skewed towards the posts made by these users and it will be easier to reconstruct their content from the output. This increases Cosar and Spencer's vulnerability to a reconstruction attack.

### Pseudonymization as Initial Privacy Preservation Technique

One common privacy-preserving technique applied to text data is pseudonymization. This involves the removal of personally identifiable information such as names, usernames, email addresses, or explicit metadata fields. In many publicly released text corpora, pseudonymization is treated as a sufficient safeguard against privacy risks. The assumption is that once identifiers are removed, individual contributors can no longer be meaningfully linked to the data.

In this project, pseudonymization of the 20Newsgroups dataset was implemented through targeted preprocessing steps. Email addresses appearing in the raw posts were detected using regular-expression-based pattern matching and removed prior to tokenization and modeling. In addition, explicit metadata fields associated with users were excluded, with only an internally assigned unique identifier retained for downstream analysis.

Despite these steps, further examination of the corpus after pseudonymization revealed privacy risks. Substantial heterogeneity existed in user participation. That is, while many users contributed only a small number of posts, a minority contributed dozens or even hundreds of messages. To assess whether such users remained distinguishable after identifier removal, the project examined stylometric patterns in the pseudonymized posts' text. Stylometry refers to the use of linguistic features, such as word usage, topical focus, and unique expression patterns, to differentiate authors based solely on their writing.

As part of the analysis, the high-dimensional document-term matrix developed after pseudonymization and preprocessing was projected onto two dimensions using Principal Component Analysis (PCA). The resulting projections showed that posts made by most contributors were restricted around one cluster, indicating these users had largely similar language patterns. However, others', such as posts made by '*user_216*', '*user_554*', and '*user_1613*' were clear outliers and had distinct linguistic patterns, as shown below in Figure 1. These outlier users with unique linguistic patterns are therefore at elevated privacy risk. The outputs by any generative text analysis technique like LDA will be particularly sensitive to these users' posts, which makes it easier to reconstruct their content and possibly re-identify them.

**Baseline LDA Estimation**

As a baseline, LDA was estimated using Collapsed Gibbs Sampling with a fixed number of topics (k = 20), mirroring the original number of newsgroups. This baseline model served two purposes. First, it provided a reference point for utility and interpretability in the absence of privacy constraints. Second, it allowed for exploration of privacy risks by examining how sensitive the learned topics were to the inclusion or removal of specific documents or users. Interestingly, all 20 topic-word distributions generated as part of the baseline LDA could easily be mapped onto the 20Newsgroups topics or groups for discussion. An illustration of 10 of these topic-word distributions is included in Figure 2 below.

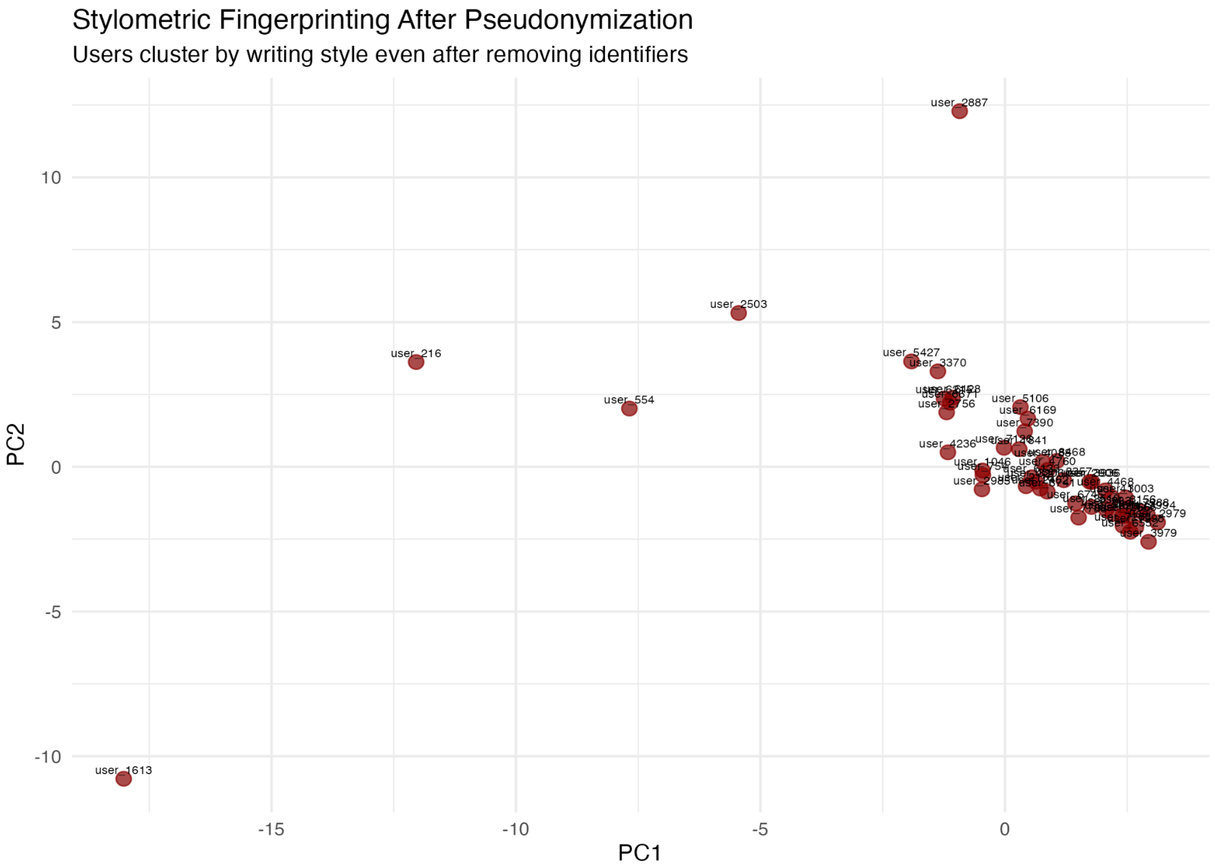


*Figure 1: Stylometric Analysis After Pseudonymization: Some users emerge as clear outliers and are at elevated privacy risks*

**Memorization Risk Analysis**

With the baseline LDA model established, the project next examined whether the learned topic-word distributions exhibit signs of memorization. Here, memorization refers to the extent to which the parameters of the model, particularly the topic-word distribution matrix Φ, are influenced by specific documents or users in the training data.

To assess memorization risk, the project evaluated the sensitivity of Φ to targeted changes in the underlying data. The focus was on posts made by Cosar (pseudonymized ID of '*user_92*') related to Turkey and the Armenian genocide in the topic that can roughly be characterized as Middle East Politics. A comparison was made for the top 10 words for Middle East Politics topic, which was Topic 3 in the baseline LDA, with the top 10 words for the Middle East Politics topic, which was Topic 7 in the LDA trained after removing posts from user_92.

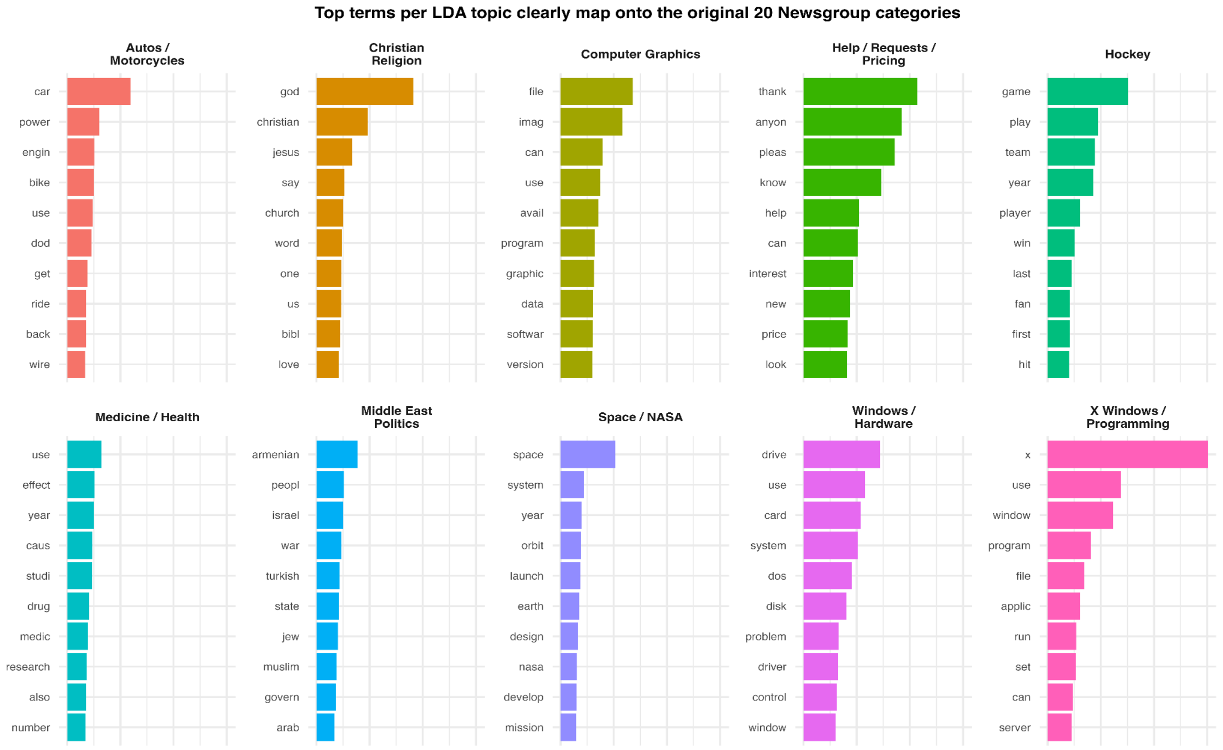


*Figure 2: Top 10 Terms for 10 LDA Groups*

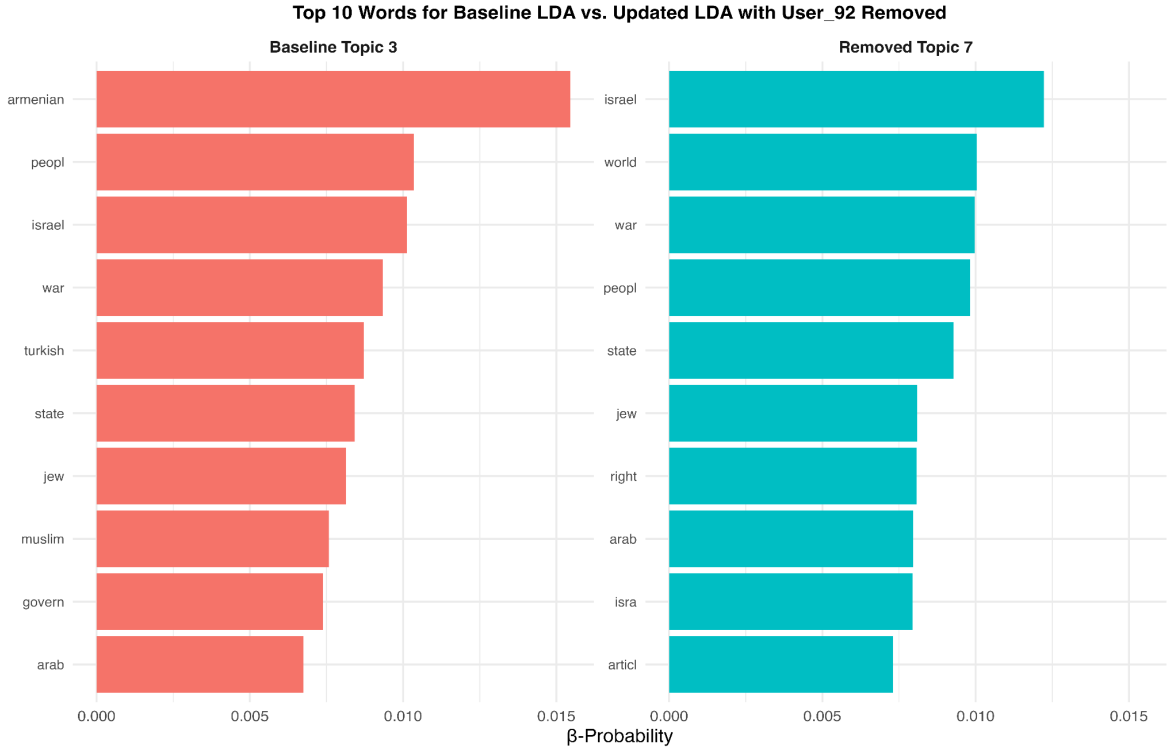


*Figure 3: Top 10 Words Comparison: Baseline LDA vs Updated LDA with User_92 Removed*

This comparison revealed that the baseline topic-word distributions were highly sensitive to the inclusion or exclusion of user_92. Whereas the top 10 words for the Middle East Topics in baseline LDA included terms such as '*armenian*' (term with highest beta-probability) and '*turkish*', these were completely missing from the top 10 words for the updated LDA, while the remaining of the top words remained largely the same. From a privacy perspective, such high sensitivity creates the potential for inference attacks, as an adversary observing model outputs can infer whether a particular user's data was included in training. These findings motivate the application of DP mechanisms, so that privacy of individuals such as Cosar can be protected.

## Differentially Private Topic Modelling

This section outlines the results of DP application to the baseline LDA model under different noise levels and sensitivity. The analysis focuses on how increasing privacy affects topic coherence, interpretability, and stability of Φ outputs.

### Prototype 1: Document-level Privacy

Under the document-level sensitivity assumption ($\Delta = 1$), topic coherence declines smoothly, i.e. monotonically, as the privacy budget $\varepsilon$ decreases. Figure 5 presents the privacy-utility curve summarizing the mean coherence drop across topics for a range of noise levels. For relatively large $\varepsilon$ (e.g., $\varepsilon \geq 5$), there is hardly any coherence loss, indicating that LDA derived topics remains largely intact even after noise injection.

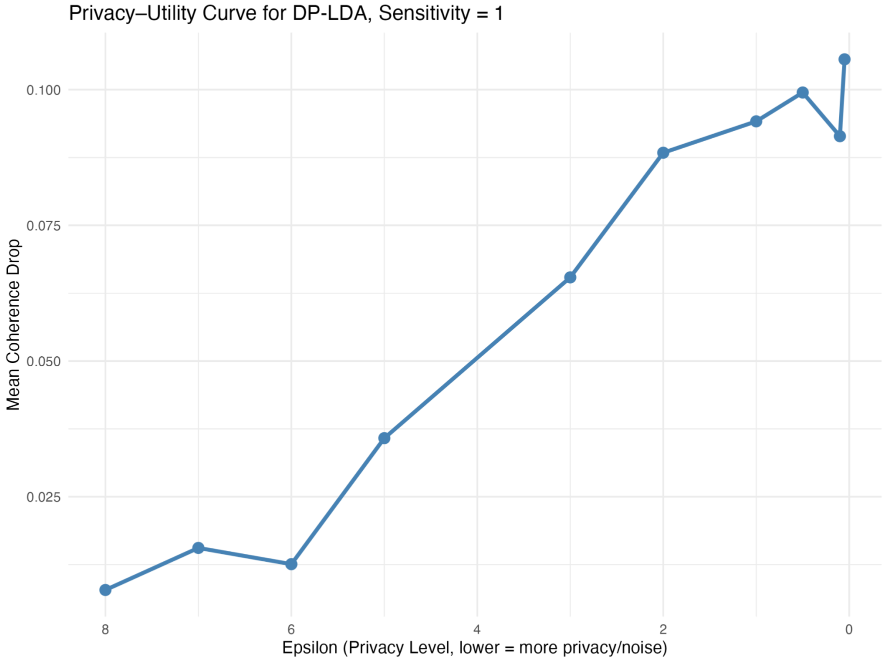


*Figure 4: Privacy-Utility Curve for Document-Level Privacy: Coherence decreases monotonically as noise level increase*

As $\varepsilon$ decreases, noise begins to noticeably affect topic composition. At $\varepsilon = 1$, coherence declines more sharply for some topics. However, core topic terms, particularly words with high beta probabilities, remain unchanged. This suggests that document-level differential privacy can provide moderate privacy protection while preserving interpretability for a range of privacy budgets. At very small ε values (e.g., ε ≤ 0.1), coherence drops substantially across most topics. Φ distributions become increasingly noisy, and the distinction between topics weakens. The resulting topics are less meaningful from an interpretability standpoint. A detailed overview of topic stability is included for document-level privacy in Annex B.

**Prototype 2: User-level Privacy**

Under the user-level privacy framework, the sensitivity was set at $\Delta = 80$. The choice of this number was based on the number of posts made by Cosar, who is the highest-volume poster in the 20Newsgroups dataset. In this case, the effect of DP noise becomes substantially more pronounced. Figure 6 shows the privacy-utility curve under this sensitivity assumption.

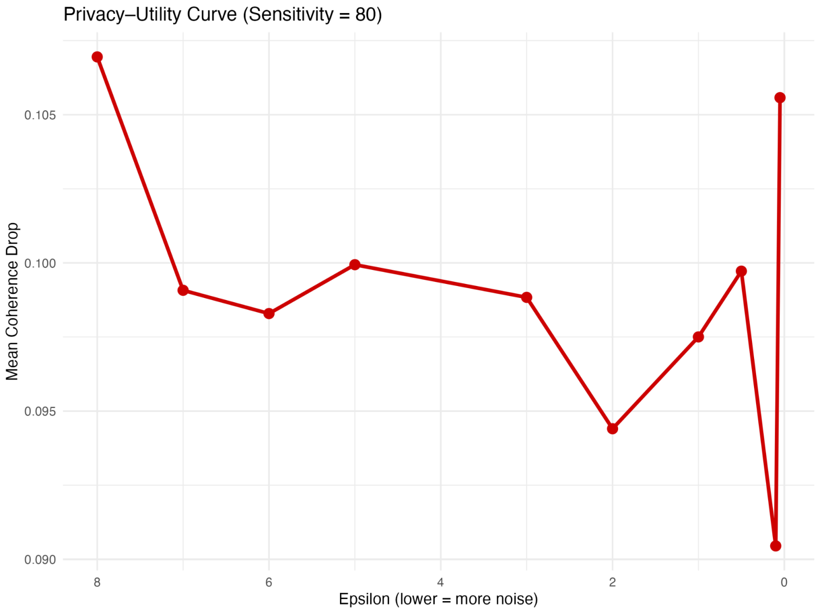


*Figure 5: Privacy-Utility Curve for User-Level Privacy: Coherence and noise do not have a monotonic relationship*

As $\varepsilon$ decreases, coherence drops rapidly, and topic–word distributions become dominated by noise. For $\varepsilon \leq 1$, many topics exhibit near-uniform word distributions. Largely in this case, coherence metrics become unstable and display non-monotonic behavior due to effects introduced by heavy noise and normalization included as part of the algorithm. A detailed overview of topic stability is included for document-level privacy in Annex C.

## Discussion

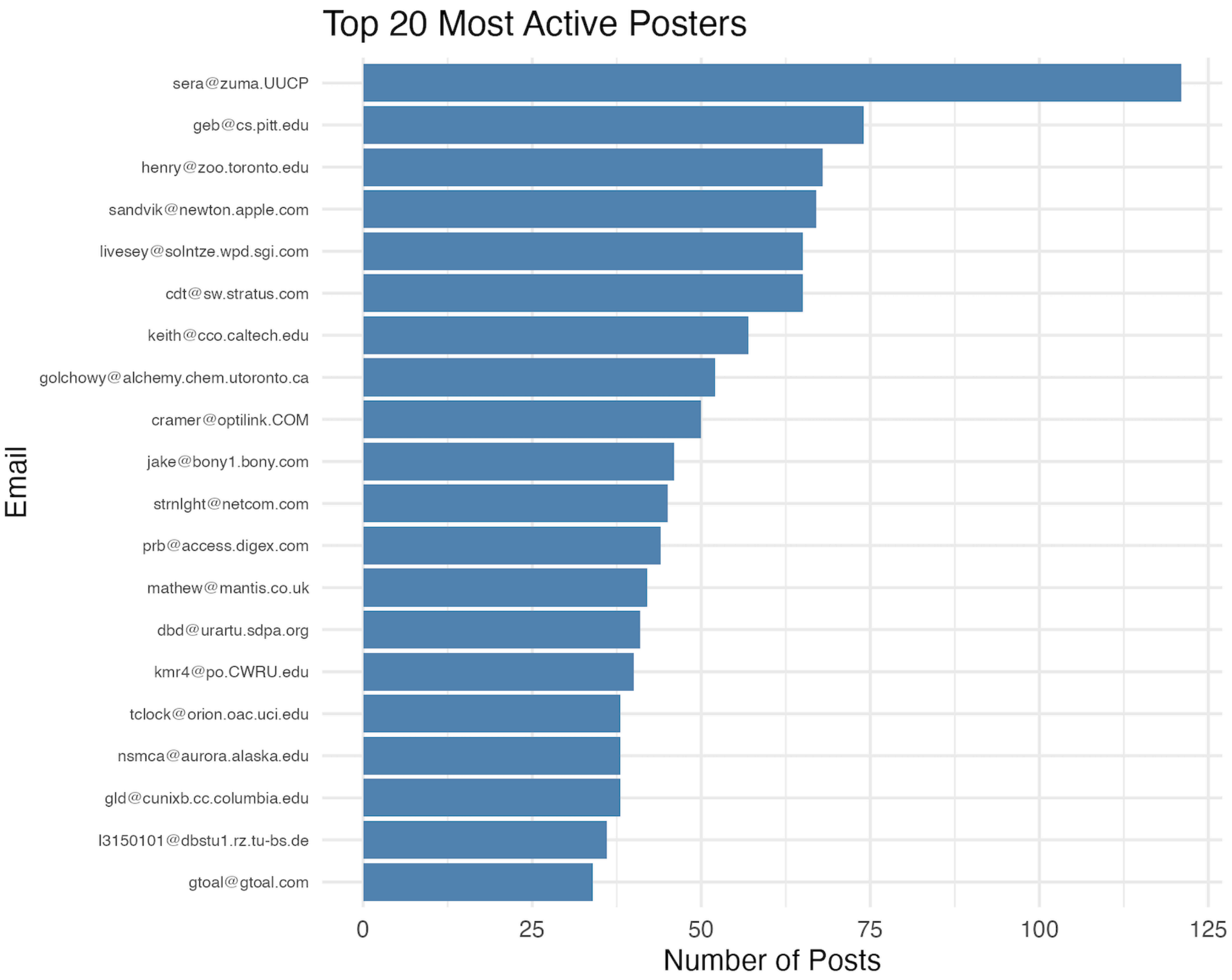
This project examined privacy risks in LDA and evaluated the extent to which DP can mitigate those risks while preserving interpretability. Using the 20Newsgroups dataset, the analysis demonstrated that standard LDA models can leak individual-level information even after personal identifiers are removed. High-volume contributors can exhibit distinct linguistic patterns that can still be picked up after pseudonymization. Stylometric analysis and targeted removal of user-specific posts further showed that topic-word distributions are highly sensitive to the inclusion or exclusion of specific individuals, which leads to memorization privacy risks.

Applying DP to the topic–word distributions revealed clear privacy-utility tradeoffs. Under document-level sensitivity assumptions ($\Delta = 1$), topic coherence degraded gradually as the noise increased. In contrast, user-level sensitivity ($\Delta = 80$) led to rapid deterioration of topic quality. The topic-word distributions become dominated by noise even at moderate noise levels. These findings underscore that the effectiveness of DP depends on how privacy is operationalized.

The author would also like to note several limitations here. DP is applied post hoc to the Φ distributions rather than integrated directly into the CGS procedure, which may limit formal or mathematically tight privacy guarantees. Sensitivity under the user-level framework is conservatively defined using a worst-case scenario of user_92 (Cosar), which may overstate individual influence. In addition, utility is evaluated primarily through topic coherence and qualitative stability.

Future work could address these limitations by implementing end-to-end DP-LDA algorithms. These could adopt adaptive sensitivity bounds, and explore alternative privacy mechanisms that better balance privacy and utility. From a policy perspective, the results should alarm researchers that use text data against treating pseudonymization as a sufficient privacy protection. Linguistic patterns and user-level participation can still expose individuals through aggregate model outputs. These risks highlight the need for stronger privacy standards and mindful use of privacy-preserving techniques in text analysis.

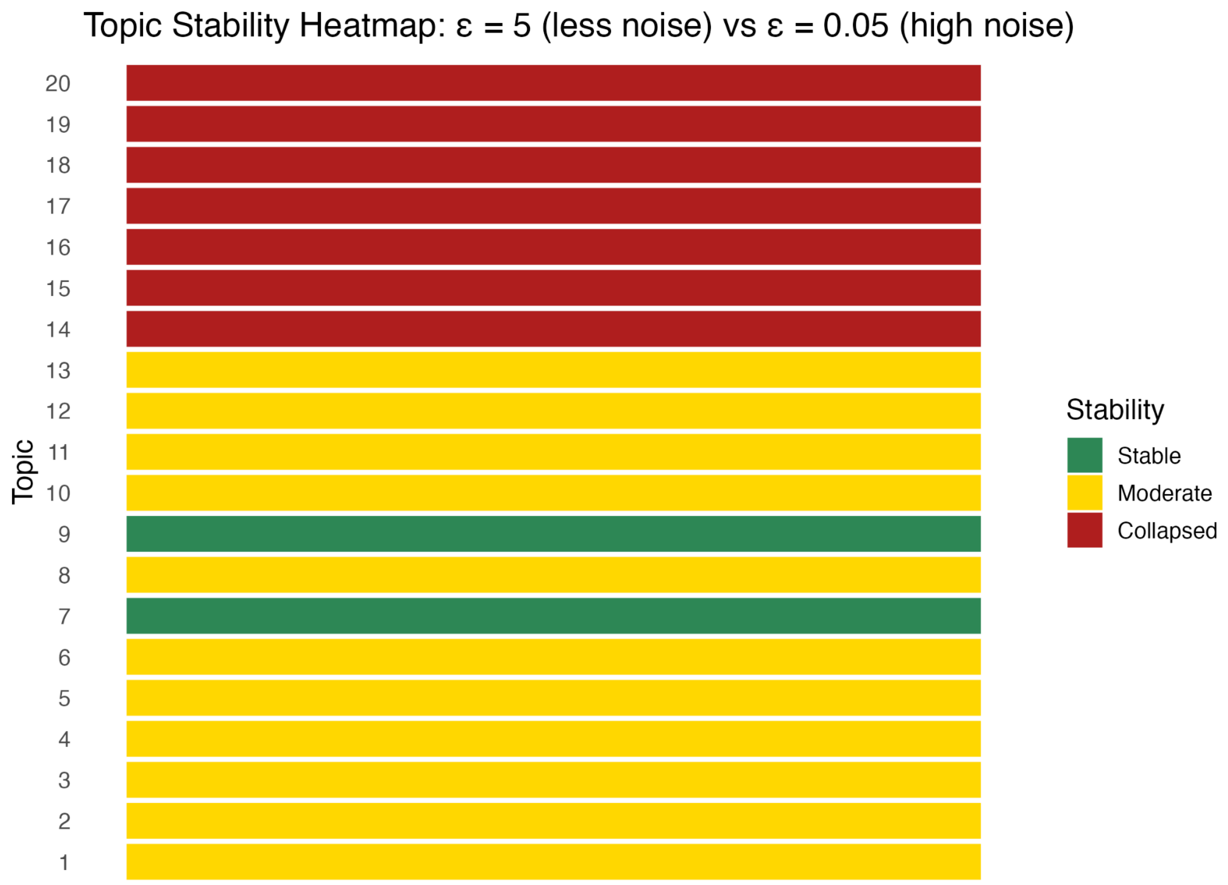
## Annex A



## Annex B

**Topic Stability Under Document-Level Privacy**

Topic stability was also assessed by comparing top-ranked words across different noise levels. Topics were qualitatively categorized as stable, moderately affected, or collapsed based on whether their top words remained largely the same after noise addition. Annex A visualizes these stability classifications. Under $\Delta = 1$, several topics remain stable even at moderate noise levels, while others show gradual degradation. Complete topic collapse is limited to the smallest $\varepsilon$ values. This reinforced the conclusion that document-level DP introduces a relatively controlled tradeoff between privacy and utility.



## Annex C

**Topic Stability Under User-Level Privacy**
Qualitative inspection confirms that topic interpretability deteriorates sharply under higher sensitivity. Core topic terms disappear from top-word lists, and topics can no longer be mapped to recognizable categories. These results highlight the difficulty of maintaining useful topic models under strong user-level privacy without alternative privacy mechanisms.

## References

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). *Quantifying memorization across neural language models.* In *Proceedings of the International Conference on Learning Representations (ICLR 2023).* https://arxiv.org/abs/2302.07800

Cosar, A. (n.d.). *Homepage*. Middle East Technical University, Department of Computer Engineering. Retrieved December 16, 2025, from https://user.ceng.metu.edu.tr/~cosar/

DeCarolis, P., Gaboardi, M., & Vadhan, S. (2020). *Rényi differential privacy mechanisms for tensor decomposition and topic modeling.* In *Proceedings of the Conference on Privacy in Machine Learning (NeurIPS Workshop).*

Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy* (Foundations and Trends in Theoretical Computer Science, Vol. 9, Nos. 3–4, pp. 211–407)

Huang, Y., & Chen, X. (2021). *Improving privacy guarantee and efficiency of latent Dirichlet allocation model training under differential privacy.* In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 159–169). Association for Computational Linguistics. https://aclanthology.org/2021.findings-emnlp.14.pdf

Huang, Y., Chen, X., & Li, Y. (2022). *Improving parameter estimation and defensive ability of LDA under Rényi differential privacy. Entropy, 24*(2), 187. https://www.sciopen.com/article/10.1007/s11390-022-2425-x

Levithan, S. (n.d.). *Regex legends*. Steven Levithan's blog. Retrieved December 16, 2025, from https://blog.stevenlevithan.com/archives/regex-legends

Manzonelli, A., Zhang, J., & Vadhan, S. (2024). *Membership inference attacks and privacy in topic modeling. Transactions on Machine Learning Research.* https://openreview.net/pdf?id=NmWp5lFL7L

Park, M., Foulds, J. R., Chaudhuri, K., & Welling, M. (2016). *Private topic modeling*. arXiv. https://arxiv.org/abs/1609.04120

scikit-learn developers. (2017). *The 20 newsgroups text dataset.* scikit-learn documentation. https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

Zhao, T., Zhang, M., & Wang, J. (2020). *Latent Dirichlet allocation model training with differential privacy. Entropy, 22*(11), 1216. https://arxiv.org/pdf/2010.04391

Zhu, T., Li, G., Zhou, W., & Phillips, P. (2016). *Differentially private data publishing and analysis: A survey. IEEE Transactions on Knowledge and Data Engineering, 29*(8), 1619–1638. https://dl.acm.org/doi/abs/10.1109/tkde.2017.2697856