Luring as a Proxy: Evaluating Corpus Transferability for Cybergrooming Detection

Shiying Fan[*] [†] **and Mareike Bassenge**[*] [†] **and Martin Steinebach**
Fraunhofer SIT | ATHENE, Darmstadt, Germany
{shiying.fan, mareike.bassenge, martin.steinebach}@sit.fraunhofer.de

Abstract

As the use of digital devices and social media grows among younger users, cybergrooming has emerged as a critical social concern for protecting vulnerable minors online. However, research on automated cybergrooming detection remains limited due to data scarcity. Building on previous studies that conceptualize cybergrooming as a form of luring communication, this paper investigates the potential transferability of corpora from luring or manipulative contexts for cybergrooming detection.

1 Introduction

Cybergrooming involves the sexual exploitation of minors by adults in digital environments (Wachs et al., 2012). It is an escalating social issue, causing severe psychological and potential physical harm to vulnerable minors (Schittenhelm et al., 2025). Automated cybergrooming detection systems (Inches and Crestani, 2012) can help protect young users on digital platforms, but their development is constrained by data scarcity and legal restrictions, limiting access to sufficient suitable training data for building accurate detection models (An et al., 2025a).

Prior research has noted that cybergrooming often involves the use of luring communication strategies (LCT) to manipulate minors (Olson et al., 2007). This raises a key question: *Can a model trained on publicly accessible luring communication corpora generalize across domains to detect cybergrooming?* While transfer learning has proven effective for tasks such as toxicity and hate speech detection (de Paula et al., 2023), sentiment classification (Myagmar et al., 2019), abusive language detection (Bose, 2023), or cyberbullying detection (Teng and Varathan, 2023), its potential for cybergrooming detection remains underexplored.

Addressing this research gap, this paper investigates the cross-domain adaptation potential of luring corpora for cybergrooming detection. We employ a two-stage evaluation: (1) analyzing semantic, topical, emotional, toxicity, and stylistic feature alignment across luring, grooming, and normal contexts, and (2) assessing model generalization when fine-tuned on each luring corpus for cybergrooming detection. To our knowledge, this is the first study to systematically evaluate the potential of luring communication data for cybergrooming detection, addressing the challenge of limited data availability in this low-resource NLP task.

2 Related Work

Transfer learning involves reusing generally (e.g., BERT and GPT) or task-specifically pretrained models for related tasks or domains (Pan and Yang, 2009) and has shown success in multiple NLP tasks (Vu et al., 2020). It is particularly valuable in low-resource settings, where data collection and annotation are costly, and training models from scratch is expensive (Weiss et al., 2016; Rogers et al., 2020). In cybergrooming research, transfer learning is typically applied through domain adaptation of pretrained language models to cybergrooming data (Vogt et al., 2021; Agarwal et al., 2022). Given the sensitivity, scarcity, and legal constraints of grooming data, publicly available luring communication corpora offer a privacy-preserving alternative for model development.

Beyond corpus availability, effective cross-domain transfer depends on the alignment between source and target data (Ruder, 2019; Gururangan et al., 2020). Corpus alignment can be assessed through multiple means such as semantic or topical similarity, as well as similar stylistic or linguistic features. Accordingly, transferability evaluation offers insights into the suitability of source data and the feasibility of cross-domain transfer (Eaton et al.,

[*]Equal contribution.
[†]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 371–385

2008; Seah et al., 2012; Jelenić et al., 2023). Domain adaptation focuses on transferring knowledge between a source domain (general luring context) and a target domain (cybergrooming context) under a shared label space and task definition (Pan and Yang, 2009). Its effectiveness can be empirically evaluated through zero-shot transfer performance using metrics such as accuracy, precision, and recall when target labels are available. Thus, this study aims to assess the transferability of luring corpora for cybergrooming detection by examining both cross-domain feature alignment and the models' predictive performance across domains.

3 Corpora

The luring-context corpora in this study include texts from **advertisements** (C_{Ads}) (Jaykin, 2023), **marketing** (C_{Mark}) (Isotonic, 2023), **phishing** (C_{Phish}) (Alam, 2024), **threatening** language (C_{TEL}) (Gales et al., 2022), and **persuasive** language in different contexts (C_{DP}) (Jin et al., 2024), (C_{CMV}) (Tan et al., 2016), (C_{P4G}) (Wang et al., 2019). Additional corpora include **flirty** messages (C_{FlirtM}) (Voituret, 2025) and dialogues (C_{FlirtD}) (Friedman, 2024), **manipulative** (C_{MM}) (Yuxin Wang, 2024) and **deceptive** texts (C_{Mafia}) (de Ruiter and Kachergis, 2018), and **social engineering** (C_{SoEng}) (Davinroy et al., 2024). Cybergrooming conversations (C_{groom}) are collected from Perverted Justice[1]. We also include two corpora representing non-luring normal contexts for comparison: Daily Dialogues (C_{DD}) (Li et al., 2017) and Conversation Chronicles (C_{CC}) (Jang et al., 2023). Among these, C_{DP}, C_{MM}, C_{SoEng}, and C_{CC} are synthetically generated. For detailed corpus statistics, see subsection A.1 and for text samples from the collected luring communication corpora, see subsection A.2.

4 Methodology

As stated in section 1, we propose an evaluation framework with two components: **feature alignment analysis** and **model generalization assessment**. Feature alignment analysis measures statistical distributional similarities between corpora in luring contexts and a cybergrooming corpus. Model generalization assessment measures how effectively models trained on luring-context data transfer to cybergrooming detection.

4.1 Feature Alignment Analysis

We assess feature alignment across five dimensions: semantic, topic, emotional, toxicity, and stylistic alignment. Semantic alignment is evaluated using SBERT (Reimers and Gurevych, 2019) through cosine similarity, maximum mean discrepancy, and PCA subspace angle distance between corpora. Topic alignment is measured with Turftopic (Kardos et al., 2025), extracting the top 10 topics and encoding them with BOW and SBERT embeddings, followed by the same similarity metrics. Emotional, toxicity, and stylistic alignments are assessed using a fine-tuned RoBERTa model[2] on GoEmotions (Demszky et al., 2020), Detoxify (Hanu and Unitary team, 2020), and LFTK (Lee and Lee, 2023), respectively. These features are then analyzed using feature-wise methods (Kolmogorov-Smirnov test, Mann-Whitney U rank-biserial correlation, and ANOVA) and multivariate methods (MANOVA, Canonical Correlation Analysis, and PCA with Spearman's correlation) to capture both individual and overall feature alignment. The implementation of these methods is detailed in subsection A.3. To aggregate results across metrics for better comparison and identify most aligned corpus with C_{groom}, we employ a ranking-based scoring scheme. For each metric, corpora paired with C_{groom} are ranked and assigned scores of +2 (highest), +1 (second), 0 (third), and -1 (others). We then sum the scores into a final alignment score indicating relative corpus similarity to C_{groom}. The complete results for each metric are shown in subsection A.3.

4.2 Model Generalization Assessment

To evaluate model generalization, we adopt BERT (Devlin et al., 2019) as the base model and create balanced training sets by pairing luring-context data with an equal number of normal-context data (see subsection A.4). To mitigate bias from text length variations, we truncate all inputs to 128 tokens. We fine-tune separate BERT models on each balanced dataset and empirically evaluate them through zero-shot transfer performance by using the luring label as a proxy. The test set consists of the 500 longest conversations from C_{groom} and 500 longest conversations from C_{DD} and C_{CC}. Since conversations are longer than the classifier maximum input token size, we segment each test conversation into 128-token sliding windows (75%

[1] http://www.perverted-justice.com/

[2] Hugging Face model - roberta-base-go-emotions

overlap), with a maximum of 25 windows per conversation for a fine-grained analysis. Window-level predictions are then aggregated using three strategies: *Majority Voting*, *Max Pooling*, and *Mean Pooling*. Model performance is measured using accuracy, macro F1, and MSE. Models achieving higher accuracy and macro F1-scores, along with low MSE, are considered more transferable[3].

5 Results and Discussion

This section presents the evaluation results of both feature alignment and model performance.

5.1 Feature Alignment Analysis

The analysis of feature alignment addresses three questions: (1) Do grooming conversations align with monologic or interactive discourse? (2) Which corpora align most closely with grooming data, and how do grooming interactions differ from normal daily communication? (3) Which alignment indicators most effectively distinguish luring and grooming contexts from normal contexts?

As shown in Table 1, although C_{groom} involves interactive grooming communication, it aligns more closely with monologic corpora, as the highest alignment scores for most features are observed within these corpora. In terms of specific features, C_{groom} exhibits the strongest *semantic and topic alignment* with flirty messages (C_{FlirtM}) rather than flirty dialogues (C_{FlirtD}), followed by mental manipulation dialogues (C_{MM}). While persuasive, marketing, deceptive, and social engineering corpora present to be less semantically similar to C_{groom}, daily dialogues (C_{DD}) demonstrates relatively close semantic alignment to C_{groom}, indicating that grooming discourse often adopts everyday conversational content. In contrast, *toxicity alignment* reveals the strongest similarity between C_{groom} and threatening language (C_{TEL}), despite their low semantic and topical overlap. *Emotional alignment* is highest with daily dialogues (C_{DD}), but with moderate similarity to threatening (C_{TEL}) and persuasive (C_{P4G}) communication. *Stylistically*, C_{groom} aligns most closely with deceptive discourse (C_{Mafia}) rather than with flirty, advertising, or other persuasive and manipulative corpora.

In general, cybergrooming discourse represents characteristics of flirty, manipulative and everyday communication, with threat-related toxicity and deceptive stylistic features. It is semantically and emotionally closer to daily conversations than to most luring-context corpora, making content- and emotion-based features less effective for differentiating cybergrooming interactions. Thus, risk-related signals are primarily conveyed through toxicity and stylistic features rather than through content or emotion. While C_{FlirtM} is the most transferable corpus across all observed features, when focusing solely on toxicity and stylistic features, we conclude that C_{TEL}, C_{Mafia}, and C_{MM} are the most suitable corpora for cybergrooming detection.

5.2 Model Generalization Assessment

The analysis of model performance addresses three questions: (1) Which discourse type in the luring corpora generalizes better for cybergrooming detection? (2) Which corpora contribute most to the cross-domain predictive performance of BERT? (3) How do the model performance results align with the findings from subsection 5.1?

Table 2 shows that no clear conclusion can be drawn regarding whether monologic or interactive corpora generalize better for cybergrooming detection. The results vary across aggregation strategies and corpus characteristics. Models trained on threatening language (C_{TEL}) and persuasive discourse (C_{P4G}) achieve the highest accuracy, macro F1-scores, and the lowest MSE values, with C_{TEL} performing best under max pooling and C_{P4G} under majority voting. Corpora featuring deceptive language (C_{Mafia}) and mental manipulation (C_{MM}) also transfer effectively, with $BERT_{Mafia}$ performing moderately under max pooling and $BERT_{MM}$ under mean pooling 1/3. For comparison, models trained on persuasive communication from C_{DP} and social engineering discourse (C_{SoEng}) exhibit the weakest performance across all metrics[4].

As shown in subsection 5.1, thematic similarity alone is insufficient to distinguish cybergrooming interactions. Model evaluation further reveals that BERT fine-tuned on C_{FlirtM} consistently performs poorly across all strategies, despite its strong semantic and topic alignment with grooming data. C_{TEL} in contrast, shows only moderate overall alignment but the highest toxicity alignment, achieves the best transfer performance under two strategies. Similarly, C_{Mafia}, which shares the

[3] We also calculated transferability metrics based on the feature layer of each model with inconclusive results; see subsubsection A.4.5.

[4] These models indicate label inversion in domain transfer which requires further evaluation out of scope for this study.

Luring Discourse	Corpus Pair	Semantic	Topic	Toxicity	Emotion	Stylistic	Total
Monologic	C_{groom} vs. C_{Ads}	-9	-1	-4	-6	-6	-26
	C_{groom} vs. C_{Mark}	-9	-6	-5	-5	-2	-27
	C_{groom} vs. C_{Phish}	-9	-6	-5	-6	-6	-32
	C_{groom} vs. C_{Mafia}	-8	-1	-4	-6	**9**	-10
	C_{groom} vs. C_{TEL}	-9	-5	**7**	0	-3	-10
	C_{groom} vs. C_{FlirtM}	**13**	**1**	-3	-3	-6	**2**
Interactive	C_{groom} vs. C_{FlirtD}	0	-3	-3	-2	-3	-11
	C_{groom} vs. C_{CMV}	-9	-4	-6	-6	-2	-27
	C_{groom} vs. C_{DP}	-9	-6	-6	-6	-6	-33
	C_{groom} vs. C_{P4G}	-9	-5	-3	0	-5	-22
	C_{groom} vs. C_{MM}	2	-1	2	-2	-5	-4
	C_{groom} vs. C_{SoEng}	-8	-6	-6	-5	-4	-29
	C_{groom} vs. C_{DD}	0	-2	-6	**3**	-4	-9
	C_{groom} vs. C_{CC}	-8	-3	-6	-4	-5	-26

Table 1: Feature alignment ranking results. **BOLD** indicates the strongest alignment between paired corpora.

Luring Discourse	Model	Majority Voting			Max Pooling			Mean Pooling 1/3			Mean Pooling 2/3		
		Acc.	Macro F1	MSE	Acc.	Macro F1	MSE	Acc.	Macro F1	MSE	Acc.	Macro F1	MSE
Monologic	$BERT_{Ads}$	0.52	0.37	0.48	0.78	0.77	0.22	0.56	0.46	0.44	0.50	0.34	0.49
	$BERT_{FlirtM}$	0.50	0.33	0.50	0.50	0.33	0.50	0.50	0.33	0.50	0.50	0.33	0.50
	$BERT_{Mark}$	0.51	0.37	0.49	0.59	0.57	0.40	0.52	0.42	0.48	0.51	0.34	0.49
	$BERT_{Phish}$	0.50	0.33	0.50	0.49	0.42	0.50	0.50	0.33	0.50	0.50	0.33	0.50
	$BERT_{Mafia}$	0.77	0.75	0.24	**0.98**	**0.98**	**0.02**	0.85	0.84	0.15	0.69	0.65	0.31
	$BERT_{TEL}$	0.81	0.80	0.19	**0.99**	**0.98**	**0.02**	**0.89**	**0.89**	**0.10**	0.69	0.67	0.30
Interactive	$BERT_{CMV}$	0.58	0.51	0.42	0.6	0.54	0.39	0.66	0.64	0.34	0.53	0.39	0.47
	$BERT_{FlirtD}$	0.64	0.60	0.36	0.65	0.61	0.35	0.71	0.69	0.29	0.58	0.49	0.42
	$BERT_{MM}$	0.82	0.82	0.18	0.75	0.73	0.25	0.88	0.88	0.12	0.74	0.72	0.26
	$BERT_{DP}$	*0.39*	*0.28*	*0.61*	*0.02*	*0.02*	*0.98*	*0.28*	*0.22*	*0.72*	*0.47*	*0.32*	*0.53*
	$BERT_{P4G}$	**0.94**	**0.94**	**0.06**	0.56	0.45	0.44	0.88	0.88	0.12	**0.91**	**0.91**	**0.08**
	$BERT_{SoEng}$	0.43	0.29	0.58	0.06	0.06	0.94	0.34	0.26	0.66	*0.47*	*0.32*	*0.53*

Table 2: Model performance for generalization assessment. **BOLD** indicates the best performance and *ITALIC* indicates the worst performance for each aggregation strategy. 1/3 and 2/3 are thresholds for assigning the label.

same overall alignment ranking as C_{TEL} and exhibits strong stylistic similarity with grooming data, achieves moderate but stable transfer performance. These findings suggest that toxicity and stylistic cues contribute more to model effectiveness than semantic similarity.

Conversely, C_{P4G}, despite weak overall alignment with grooming data, demonstrates strong transferability to the given task. While C_{P4G} presents good transfer performance, other persuasive corpora do not significantly contribute to cybergrooming detection, as concluded in subsection 5.1. This suggests that only certain persuasive strategies are reflected in cybergrooming interactions. Additionally, C_{MM}, which aligns moderately with C_{groom}, also shows stable, moderately strong performance in cybergrooming detection.

To summarize, models fine-tuned on advertising, phishing, flirty, social engineering, or persuasive contexts consistently underperform across all aggregation strategies. Consistent with corpus transferability findings, the model evaluation results indicate that the most transferable models are $BERT_{TEL}$, $BERT_{Mafia}$, and $BERT_{MM}$. Furthermore, $BERT_{P4G}$, despite showing weaker feature alignment with C_{groom}, also demonstrates effective transferability.

6 Conclusion and Future Work

This paper examined the transfer potential of corpora from luring contexts in cybergrooming detection. To this end, we assessed the feature alignment of various corpora with cybergrooming data and fine-tuned BERT models on each corpus for cross-domain predictive evaluation. Our results indicate that thematic similarity is only weakly predictive of transfer effectiveness. Instead, corpora with salient toxicity and distinctive stylistic features performed better. Emotional features also contributed moderately to transferability. The most effective corpora

for transfer learning were those representing threatening language (C_{TEL}), persuasive (C_{P4G}) and manipulative (C_{MM}) interactions, and deceptive communication (C_{Mafia}).

Among these corpora, C_{MM} shows the strongest thematic alignment with C_{DD} (normal-context data), while C_{P4G} exhibits greater emotional and stylistic similarity, and C_{Mafia} aligns in toxicity. These complementary patterns demonstrate that different luring characteristics contribute in varied ways to cybergrooming detection, pointing to a complex synthesis of benign and luring characteristics. This observation further suggests that cybergrooming is inherently dynamic across stages, thereby motivating a more fine-grained analysis of staged conversations to identify stage-specific characteristics and their correspondence to luring strategies.

Building on these findings, future work will expand the scope of the corpora by incorporating additional conversational data that capture a similar yet broader range of luring-relevant characteristics. We will move beyond the current feature alignment approaches by incorporating deeper linguistic analyses. The evaluation of model generalization will also shift from a primary focus on predictive performance to a more comprehensive framework for assessing model generalization and transferability beyond BERT-based architecture.

Data scarcity remains a bottleneck in cybergrooming research. While synthetic dataset generation through data augmentation and few-shot LLM prompting can expand available resources, such approaches remain dependent on existing data quality and may underperform relative to original data (Schmidhuber and Kruschwitz, 2024). Although recent advances in generative LLMs enable zero-shot data generation, this raises ethical concerns due to the creation of harmful content. Given the demonstrated relevance of the synthetic C_{MM} corpus for cybergrooming detection, a more ethically grounded alternative is to generate "normal" conversational data that implicitly captures harmful behavioral patterns, thereby supporting dataset expansion while mitigating ethical risks.

Limitations

In this study, we analyzed corpora at the conversation level by representing each conversation with both participants as a single textual unit. While this approach allows us to capture overall interaction pattern, it limits the observation of fine-grained behaviors at the message level, across interactional stages, or within one-sided exchanges. Future research could address this limitation by considering more fine-grained representations, such as message-level modeling, windowed contextual segments, or one-sided conversational views, to gain more specific insights into luring strategies and cybergrooming behavior. In this context, deeper linguistic analyses can further support both global and localized examination of distinctive behavioral signals, enabling characterization of cybergrooming dynamics.

Moreover, the sliding window approach and aggregation methods for determining model transferability can also be further optimized to analyze the transforming nature of conversations. Our approach only included the first 25 windows of a conversation (and aggregated the results) which is sufficient for most samples but fails to capture later development in longer conversations with thousands of messages. Therefore, as the direct domain transfer focuses the earlier stages of cybergrooming (and is more applicable to early detection of cybergrooming (Vogt et al., 2021; An et al., 2025b)), insights on later stages are lacking and could provide further valuable insights.

Additionally, it is important to mention that the cybergrooming corpus used in this study is outdated and monolingual. This limits the extent to which our findings can be generalized to modern communication or other languages.

Finally, we acknowledge that the current evaluation framework does not fully capture the generalization of a model, as our evaluation of transferability is limited to feature alignment analysis and model cross-domain predictive performance. However, the framework provides valuable insights into whether luring communication corpora can be effectively used for transfer. Building on this foundation, future work will focus on developing a more comprehensive evaluation framework for assessing cybergrooming detection models pretrained on corpora that reflect the characteristics identified in this study.

Ethical Considerations

The corpora used in this study are publicly available, and all sources from which they were obtained are properly cited and referenced in the paper. Where applicable, the corpora have been

anonymized to ensure that no personally identifiable information is included.

We acknowledge that the corpora used in this study contain potentially harmful content reflecting luring and manipulative communication patterns associated with cybergrooming. While such data is essential for cybergrooming research, they also raise ethical concerns due to their potential misuse, such as for training predatory bots or generating deceptive content. To mitigate these risks, we emphasize the importance of careful ethical review and adherence to responsible NLP guidelines in any downstream use or deployment context.

Furthermore, we want to emphasize the importance of performance metrics in such sensitive contexts. To make sure that the least of predatory behaviors remain undetected, a high recall score that minimizes False Negatives (FN) is preferred. This in turn commonly raises the number of False Positives (FP). For sensitive contexts, especially involving minors, full autonomous moderation remains impossible as models will always require human-in-the-loop to ensure reliable and ethical handling.

Acknowledgements

This research is supported by ATHENE under the ChARM project, which focuses on the technical protection of minors, including CSAM and cybergrooming detection.

References

Nancy Agarwal, Tuğçe Ünlü, Mudasir Ahmad Wani, and Patrick Bours. 2022. Predatory conversation detection using transfer learning approach. In *Machine Learning, Optimization, and Data Science*, pages 488–499, Cham. Springer International Publishing.

Naser Abdullah Alam. 2024. Phishing email dataset.

Heajun An, Marcos Silva, Qi Zhang, Arav Singh, Minqian Liu, Xinyi Zhang, Sarvech Qadir, Sang Won Lee, Lifu Huang, Pamela Wisnieswski, and Jin-Hee Cho. 2025a. Toward integrated solutions: A systematic interdisciplinary review of cybergrooming research. *CoRR*, abs/2503.05727.

JinMyeong An, Sangwon Ryu, Heejin Do, Yunsu Kim, Jungseul Ok, and Gary Lee. 2025b. Revisiting early detection of sexual predators via turn-level optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4713–4724, Albuquerque, New Mexico. Association for Computational Linguistics.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. 2019. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pages 2309–2313. IEEE.

Matthew Barnett. regex: Alternative regular expression module for python. `https://pypi.org/project/regex/`. Python package version 2025.11.3.

Tulika Bose. 2023. *Transfer learning for abusive language detection*. Theses, Université de Lorraine.

Michael Davinroy, James Cook, Kirill Trapeznikov, Nathan Schurr, Laura Cassani, and Lin Ai. 2024. Seconvo.

Angel Felipe Magnossao de Paula, Paolo Rosso, and Damiano Spina. 2023. Mitigating negative transfer with task awareness for sexism, hate speech, and toxic language detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Bob de Ruiter and George Kachergis. 2018. The mafiascum dataset: A large text corpus for deception detection. *CoRR*, abs/1811.07851.

Myle Demszky and 1 others. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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.

Eric Eaton, Marie Desjardins, and Terran Lane. 2008. Modeling transfer relationships between learning tasks for improved inductive transfer. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 317–332. Springer.

Myles Friedman. 2024. Flirty conversations dataset. Last accessed on 2025-12.

Tammy Gales, Andrea Nini, and Ellen Symonds. 2022. The threatening english language (tel) corpus.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alex Smola. 2012. A kernel two-sample test. In *Journal of Machine Learning Research*, volume 13, pages 723–773.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and Pierre Gérard-Marchant and Kevin Sheppard and Tyler Reddy and Warren Weckesser and Hameer Abbasi and Christoph Gohlke and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Long-Kai Huang, Ying Wei, Yu Rong, Qiang Yang, and Junzhou Huang. 2021. Frustratingly easy transferability estimation. In *International Conference on Machine Learning*.

Giacomo Inches and Fabio Crestani. 2012. Overview of the international sexual predator identification competition at pan-2012.

Isotonic. 2023. Marketing email samples. Last accessed on 2025-12.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606, Singapore. Association for Computational Linguistics.

Patadiya Jaykin. 2023. Advertisement copy dataset. Last accessed on 2025-12.

Fran Jelenić, Josip Jukić, Nina Drobac, and Jan Snajder. 2023. On dataset transferability in active learning for transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2282–2295, Toronto, Canada. Association for Computational Linguistics.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Márton Kardos, Kenneth C Enevoldsen, Jan Kostkan, Ross Deans Kristensen-McLachlan, and Roberta Rocca. 2025. Turftopic: Topic modelling with contextual representations from sentence transformers. *Journal of Open Source Software*, 10(111):8183.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Julien Plu, Philippe Schmid, Sylvain Gugger, Mariama Drame, Lewis Tunstall, Thomas Wolf, and Alexander M. Rush. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Wes McKinney. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56.

B. Myagmar, J. Li, and S. Kimura. 2019. Transferable high-level representations of BERT for cross-domain sentiment classification. In *Proceedings of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, Athens. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Loreen N Olson, Joy L Daggs, Barbara L Ellevold, and Teddy KK Rogers. 2007. Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. 2022. Transferability estimation using bhattacharyya class separability. In *2022 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 9162–9172.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Catherine Schittenhelm, Maxime Kops, Maeve Moosburner, Saskia M Fischer, and Sebastian Wachs. 2025. Cybergrooming victimization among young people: A systematic review of prevalence rates, risk factors, and outcomes. *Adolescent Research Review*, 10(2):169–200.

Maximilian Schmidhuber and Udo Kruschwitz. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Chun-Wei Seah, Yew-Soon Ong, and Ivor W Tsang. 2012. Combating negative transfer from predictive distribution differences. *IEEE transactions on cybernetics*, 43(4):1153–1165.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. 2022. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *Computer Vision – ECCV 2022*, pages 286–302, Cham. Springer Nature Switzerland.

Moein Sorkhei, Christos Matsoukas, Johan Fredin Haslum, Emir Konuk, and Kevin Smith. 2025. k-NN as a simple and effective estimator of transferability. *Transactions on Machine Learning Research*.

Baochen Sun and Kate Saenko. 2015. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, volume 4, pages 24–1.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Teoh Hwai Teng and Kasturi Dewi Varathan. 2023. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, 11:55533–55560.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. 2019. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1405.

Matthias Vogt, Ulf Leser, and Alan Akbik. 2021. Early detection of sexual predators in chats. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4985–4999.

Tom Voituret. 2025. Flirty or not ia extended dataset. Last accessed on 2025-12.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Sebastian Wachs, Karsten D Wolf, and Ching-Ching Pan. 2012. Cybergrooming: Risk factors, coping strategies and associations with cyberbullying. *Psicothema*, pages 628–633.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. 2021. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*.

Saeed Hassanpour Soroush Vosoughi Yuxin Wang, Ivory Yang. 2024. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3764.

A Appendix

A.1 Corpus Statistics

This section presents the statistics of the corpora that were used for this study (Table 3). References to each corpus can be found in section 3. Corpus samples went through multiple preprocessing steps and selection, including concatenating all messages of conversation corpora into one string for analysis, removal of structural markers, normalizing special characters. Demojizing is employed for some analysis steps like BERT emotion. As shown in Table 4, the length of samples from different corpora varies. Since we focus on longer conversations, we select only those samples that meet predefined length thresholds for each corpus, as indicated in "Selection" column of Table 3.

Context	abbr.	Corpus	#Samples	#Final	#Final	Selection	Discourse
Luring	C_{Ads}	Advertisements	1,100	696	10,025	len(str)>30	Monologic
	C_{FlirtM}	Flirty Messages	2,523	1034		len(str)>10	
	C_{Mark}	Marketing Emails	487	487		all	
	C_{Phish}	Phishing Emails	42,845	619		30<len(str)<32	
	C_{Mafia}	Mafiascum	2,235	2235		all	
	C_{TEL}	Threatening Language	715	239		len(str)>30	
	C_{CMV}	Change My View	3,051	3051	9,975	all	Interactive
	C_{FlirtD}	Flirty Dialog	347	317		random	
	C_{MM}	Mental Manipulation	2,640	2640		all	
	C_{DP}	Daily Persuasion	77,987	6950		len(str)>43	
	C_{P4G}	Persuasion for Good	1,017	1017		all	
	C_{SoEng}	Social Engineering	715	715		all	
Grooming	C_{groom}	Cybergrooming	585	585	585	all	
Normal	C_{DD}	Daily Dialog	12,376	9995	20,000	len(str)>30	
	C_{CC}	Conversation Chronicles	200,000	9995		len(str)>1200	

Table 3: Corpora Overview

Context	Corpus	#Samples	Min	Max	Mean(std)	Median	Discourse
Luring	C_{Ads}	1,100	8	43	31(3.78)	32	Monologic
	C_{FlirtM}	2,523	1	101	10.47(7.58)	9	
	C_{Mark}	487	60	390	170.28(45.64)	166	
	C_{Phish}	42,845	1	107,710	121.35(594.75)	55	
	C_{Mafia}	2,235	201	205,374	16,114.91(18,666.04)	9,893	
	C_{TEL}	715	4	4,611	175.39(329.00)	89	
	C_{CMV}	3,051	974	120,051	9,009.72(9,628.47)	5,939	Interactive
	C_{FlirtD}	347	8	105	61.22(22.24)	60	
	C_{MM}	2,640	6	1,156	95.97(84.51)	71	
	C_{DP}	77,987	10	83	34.45(6.86)	34	
	C_{P4G}	1,017	35	1,241	345.88(141.37)	324	
	C_{SoEng}	715	271	1,444	612.90(237.07)	521	
Grooming	C_{groom}	585	162	70,548	8,826.35(10,542.92)	4,976	
Normal	C_{DD}	12,376	5	850	107.39(74.86)	89	
	C_{CC}	200,000	602	2,465	1,058.9(126.23)	1050	

Table 4: Corpora Statistics of string length

A.2 Text Samples from Corpora

Table 5 presents short text samples from each of the luring corpora, providing a brief illustration of their content and style.

Corpus	Sample
C_{Ads}	"Discover our Ruffleneck sweater - a perfect blend of style and warmth! Stay cozy in a bold oversized collar that elevates your winter wardrobe."
C_{FlirtM}	"I can't wait to pick you up from the couch and take you to the kitchen counter and pleasure you and then pick you back up and pin you up against the kitchen wall and continue."
C_{Mark}	"Dear Valued Customer, Do you find yourself struggling to maintain proper posture throughout the day? Look no further than the BackJoy Portable Seat Cushion! This innovative cushion is designed to improve your posture while you sit, no matter where you are."
C_{Phish}	"girls young ahh man site makes hard warning us govt want see teen needs fucked teenteenteen voted 1 teen site net horniest youngest teen sluts world many categories teens include daddy girls babysitters cheerleaders get free access"
C_{Mafia}	"What information? If anything we'll be more receptive to any possible changes on our wagon and in addition, when we move our vote, you'll know it's a serious vote. Saki's replace out was not alignment indicative and to implicate it as such is completely bullshit, Sakura."
C_{TEL}	"i went home and loaded in my truck and spend the past 3 hours looking for this douche with the expressed intent of crushing him in that little piece of shit under cover gray impala hooking up my tow chains and dragging his stupid ass down to creek hills and just drowning him in the falls."
C_{CMV}	"How are the things we create unnatural? It is only through natural occurrences that we have this ability, why is it that we would give the credit of these things solely to man, as opposed to nature? We are not separate from nature, thus, how can any of our actions or creations be unnatural?"
C_{FlirtD}	"Only when I'm talking to someone as captivating as you.", "Flattery will get you everywhere.", "Is that an invitation?", "Maybe it is, maybe it isn't. ;)", "Mystery girl, I like it. What's your favorite thing to do late at night?", "Hmm, probably stargazing. What about you?"
C_{MM}	"I can almost feel your doubt about what I'm saying, Mary.", "I can't give up looking for her, Mrs. Redi, no matter what you're hinting at.", "I have no intention whatsoever of hinting. Your sister, Mary, is a murderess. She killed Irving August – stabbed him out of fright when he discovered where she was hiding."
C_{DP}	"Hey Kevin, I understand that you might think academic awards are pointless, but they actually play a significant role in promoting research with real-world applications."
C_{P4G}	"Good morning. How are you doing today?", "Hi. I am doing good. How about you?", "I'm doing pretty good for a Tuesday morning.", "Haha. Same here, but it really feels like a Monday.", "Ugh yes it does!"
C_{SoEng}	"Hi Nancy, I hope this message finds you well. My name is Cynthia, and I am a journalist with 'Tech Forward Daily'. I've been following the remarkable developments at your AI startup, and I'm truly impressed by how it's shaping the industry. I would love to conduct an interview with you."

Table 5: Short Text Samples from Luring Corpora

A.3 Feature Alignment Analysis

This section presents detailed alignment analysis results for semantic, topical, toxicity, emotional and stylistic features. As stated in section 4, semantic similarity between corpora was evaluated by embedding all texts using SBERT (Reimers and Gurevych, 2019), and then computed cosine similarity (implemented with `sklearn`), maximum mean discrepancy (MMD) (Gretton et al., 2012), and principal component analysis (PCA) (using `sklearn`) with subspace angle distance (Sun and Saenko, 2015) between the resulting embedding matrices of corpus pairs.

Context	Cosine Similarity			MMD			PCA subspace angle degree			Rank Score
	first	last	middle	first	last	middle	first	last	middle	
C_{groom} vs. C_{Ads}	0.09	0.09	0.08	0.24	0.18	0.25	76.13	72.47	74.07	-9
C_{groom} vs. C_{FlirtM}	**0.18**	**0.18**	**0.17**	**0.17**	**0.11**	**0.17**	**64.79**	**62.17**	**60.76**	13
C_{groom} vs. C_{Mark}	0.06	0.07	0.05	0.23	0.16	0.24	78.24	73.75	76.59	-9
C_{groom} vs. C_{Phish}	0.08	0.08	0.06	0.30	0.24	0.31	75.09	73.81	73.78	-9
C_{groom} vs. C_{Mafia}	0.11	0.09	0.14	0.25	0.20	0.30	75.61	74.20	73.61	-8
C_{groom} vs. C_{TEL}	0.09	0.10	0.09	0.21	0.14	0.22	74.51	69.84	73.42	-9
C_{groom} vs. C_{CMV}	0.01	0.02	0.06	0.23	0.16	0.23	76.13	76.16	73.95	-9
C_{groom} vs. C_{FlirtD}	**0.20**	**0.20**	**0.19**	0.21	0.15	0.23	69.21	64.73	67.31	0
C_{groom} vs. C_{MM}	0.12	0.15	0.13	**0.19**	**0.11**	**0.18**	70.52	57.00	63.88	2
C_{groom} vs. C_{DP}	0.03	0.02	0.03	0.24	0.18	0.24	73.90	73.84	73.65	-9
C_{groom} vs. C_{P4G}	0.10	0.07	0.05	0.46	0.35	0.44	70.76	72.67	75.61	-9
C_{groom} vs. C_{SoEng}	0.13	0.08	0.08	0.36	0.23	0.36	77.54	72.81	73.60	-8
C_{groom} vs. C_{DD}	0.11	0.12	0.09	**0.17**	**0.11**	**0.19**	66.90	**61.96**	65.79	0
C_{groom} vs. C_{CC}	0.11	0.10	0.14	0.20	0.17	0.24	67.70	67.37	66.49	-8

Table 6: Semantic alignment analysis. **BOLD** indicates the best alignment for each metric.

For assessing topic alignment, we applied Turftopic (Kardos et al., 2025) to each corpus and extracted the top 10 topics per instance. The resulting topic words are encoded using Bag-of-Words (BOW) (using `CountVectorizer` from `sklearn`) and SBERT embeddings, following the same set of similarity metrics for semantic alignment.

Context	Cosine Similarity		MMD		PCA Subspace Angle Degree		Rank Score
	BOW	SBERT	BOW	SBERT	BOW	SBERT	
C_{groom} vs. C_{Ads}	0.05	0.22	0.35	0.87	81.83	84.24	-1
C_{groom} vs. C_{FlirtM}	**0.30**	0.24	**0.24**	0.31	**69.48**	93.04	1
C_{groom} vs. C_{Mark}	0.13	0.22	0.27	0.32	82.18	90.80	-6
C_{groom} vs. C_{Phish}	0.13	0.21	0.29	0.47	81.12	88.66	-6
C_{groom} vs. C_{Mafia}	0.07	0.13	1.0	1.15	89.5	90.27	-1
C_{groom} vs. C_{TEL}	0.22	0.25	0.27	0.31	77.55	**81.10**	-5
C_{groom} vs. C_{CMV}	0.12	0.19	**0.25**	0.35	83.32	89.34	-4
C_{groom} vs. C_{FlirtD}	0.24	**0.30**	0.29	0.30	75.16	85.49	-3
C_{groom} vs. C_{MM}	**0.31**	**0.28**	**0.25**	0.30	**68.15**	90.19	-1
C_{groom} vs. C_{DP}	0.16	0.22	0.27	0.33	82.62	90.48	-6
C_{groom} vs. C_{P4G}	0.16	0.21	0.36	0.68	79.2	90.96	-5
C_{groom} vs. C_{SoEng}	0.22	0.24	0.31	0.47	78.68	90.60	-6
C_{groom} vs. C_{DD}	**0.30**	0.25	**0.24**	**0.27**	74.27	92.27	-2
C_{groom} vs. C_{CC}	0.18	0.24	**0.25**	0.32	78.69	96.20	-3

Table 7: Topic alignment analysis. **BOLD** indicates the best alignment for each metric.

For the remaining alignment analyses, we annotate each corpus using a fine-tuned RoBERTa model[5] on GoEmotions (Demszky et al., 2020), Detoxify (Hanu and Unitary team, 2020), and LFTK (Lee and Lee, 2023) to obtain emotional, toxicity, and stylistic features. Because each annotation produces multiple numerical features, we evaluate alignment from two complementary perspectives: feature-wise analysis and multivariate analysis.

For feature-wise analysis, we used the Kolmogorov-Smirnov (KS) test, Mann-Whitney U (MWU) rank-biserial correlation, and ANOVA. Statistical analyses were performed using `scipy` library (v1.16.2),

[5] https://huggingface.co/SamLowe/roberta-base-go_emotions

with `stats.ks_2samp` for KS test, `stats.mannwhitneyu` for MWU, and `stats.kstest` for ANOVA. The KS statistic captures full distributional divergence by measuring the maximum difference between cumulative distributions. The effect size derived from MWU test quantifies differences in central tendency, particularly median shifts. ANOVA assesses differences in feature means and variances. Together, these tests provide a comprehensive view of alignment at the individual-feature level.

For multivariate analysis, we apply MANOVA, Canonical Correlation Analysis (CCA), and PCA with Spearman's correlation. MANOVA was implemented with `statsmodel` library (v0.14.6) using `multivariate.manova`, CCA and PCA were implemented with `sklearn`, and Spearman's correlation with `scipy.stats.spearmanr`. MANOVA, using Wilks' Lambda, evaluates multivariate mean differences across all features. CCA measures linear correspondence between feature sets. Spearman's rank correlation of PCA demonstrate structural similarity in global variance patterns. These multivariate approaches collectively reveal how corpora align in terms of their overall feature structures. For KS, MWU, and ANOVA, smaller values indicate stronger alignment, whereas for MANOVA, PCA+Speaman, and CCA, larger values indicate stronger alignment.

Context	**Feature-wise**			Multivariate			**Rank Score**
	KS	**MWU**	**ANOVA**	**MANOVA**	**PCA+Spearman**	**CCA**	
C_{groom} vs. C_{Ads}	0.96	0.99	667.55	0.24	**0.83**	0.1	-4
C_{groom} vs. C_{FlirtM}	0.61	0.57	**87.87**	0.58	0.66	0.14	-3
C_{groom} vs. C_{Mark}	0.96	0.99	468.24	0.27	0.7	0.19	-5
C_{groom} vs. C_{Phish}	0.81	0.91	393.67	0.44	0.77	0.09	-5
C_{groom} vs. C_{Mafia}	**0.43**	**0.56**	394.69	0.53	0.6	0.1	-4
C_{groom} vs. C_{TEL}	**0.33**	**0.17**	**52.02**	0.63	0.56	0.2	7
C_{groom} vs. C_{CMV}	0.77	0.89	1695.12	0.29	0.71	0.14	-6
C_{groom} vs. C_{FlirtD}	0.91	0.97	294.91	0.34	0.37	0.34	-3
C_{groom} vs. C_{MM}	**0.42**	**0.49**	114.98	**0.89**	0.68	0.15	2
C_{groom} vs. C_{DP}	0.99	1.0	6832.21	0.15	0.37	0.12	-6
C_{groom} vs. C_{P4G}	0.95	0.99	961.48	0.21	**0.9**	0.13	-3
C_{groom} vs. C_{SoEng}	1.0	1.0	704.38	0.22	0.68	0.11	-6
C_{groom} vs. C_{DD}	0.89	0.96	4589.55	0.46	0.75	0.11	-6
C_{groom} vs. C_{CC}	0.97	1.0	9555.66	0.16	0.64	0.11	-6

Table 8: Toxicity alignment analysis based on 6 Detoxify features. **BOLD** indicates the best alignment for each metric.

Context	**Feature-wise**			Multivariate			**Rank Score**
	KS	**MWU**	**ANOVA**	**MANOVA**	**PCA+Spearman**	**CCA**	
C_{groom} vs. C_{Ads}	0.68	0.48	589.0	0.04	0.25	0.35	-6
C_{groom} vs. C_{FlirtM}	0.51	0.49	235.0	0.17	0.14	0.42	-3
C_{groom} vs. C_{Mark}	0.65	0.33	421.60	0.06	0.18	0.49	-5
C_{groom} vs. C_{Phish}	0.79	0.77	762.27	0.04	0.17	0.49	-6
C_{groom} vs. C_{Mafia}	0.51	0.18	711.77	0.10	0.25	0.37	-6
C_{groom} vs. C_{TEL}	0.49	0.36	145.90	0.14	0.30	**0.57**	0
C_{groom} vs. C_{CMV}	0.53	0.34	846.94	0.12	0.34	0.43	-6
C_{groom} vs. C_{FlirtD}	0.54	0.24	**199.71**	0.12	0.24	**0.51**	-2
C_{groom} vs. C_{MM}	0.42	0.25	552.45	0.17	**0.57**	0.36	-2
C_{groom} vs. C_{DP}	0.56	0.30	2146.60	0.09	0.27	0.35	-6
C_{groom} vs. C_{P4G}	**0.40**	**0.03**	363.10	0.13	0.44	0.49	0
C_{groom} vs. C_{SoEng}	0.66	0.15	562.07	0.05	0.34	0.37	-5
C_{groom} vs. C_{DD}	**0.36**	0.25	1561.06	0.19	**0.68**	0.37	3
C_{groom} vs. C_{CC}	0.43	0.10	2325.81	0.11	0.24	0.39	-4

Table 9: Emotion alignment analysis based on 28 GoEmotions features. **BOLD** indicates the best alignment for each metric.

Context	Feature-wise			Multivariate			Rank Score
	KS	**MWU**	**ANOVA**	**MANOVA**	**PCA+Spearman**	**CCA**	
C_{groom} vs. C_{Ads}	0.75	0.42	2714.39	0.0	0.12	0.94	-6
C_{groom} vs. C_{FlirtM}	0.69	0.42	1943.38	0.01	0.34	0.95	-6
C_{groom} vs. C_{Mark}	0.69	0.18	1425.07	0.01	0.44	0.99	-2
C_{groom} vs. C_{Phish}	0.79	0.47	2362.72	0.0	0.16	0.93	-6
C_{groom} vs. C_{Mafia}	**0.36**	**0.08**	734.89	0.04	0.53	0.96	9
C_{groom} vs. C_{TEL}	0.60	0.16	**412.12**	0.02	0.42	0.89	-3
C_{groom} vs. C_{CMV}	0.48	0.27	2615.43	0.01	0.53	0.96	-2
C_{groom} vs. C_{FlirtD}	0.72	0.41	944.23	0.01	0.35	0.99	-3
C_{groom} vs. C_{MM}	0.64	0.34	1930.55	0.03	0.40	0.96	-5
C_{groom} vs. C_{DP}	0.74	0.25	7467.03	0.01	0.21	0.95	-6
C_{groom} vs. C_{P4G}	0.58	0.15	1037.15	0.02	0.41	0.96	-5
C_{groom} vs. C_{SoEng}	0.59	0.13	1469.73	0.01	0.39	0.95	-4
C_{groom} vs. C_{DD}	0.66	0.36	4412.40	0.04	0.29	0.94	-4
C_{groom} vs. C_{CC}	0.56	0.22	4408.34	0.02	0.34	0.96	-5

Table 10: Stylistic alignment analysis base on 220 LFTK features. **BOLD** indicates the best alignment for each metric.

A.4 Model Generalization Assessment

In this section, we describe the fine-tuning process in detail. We preprocess texts by removing HTML tags, URLs, emails, user mentions, and numbers. Then we use BERT (Devlin et al., 2019) for semantic text representation. For each text input, we use 128 tokens because text samples tend to be shorter in chat conversations. We then fine-tune a pre-trained BertForSequenceClassification (Wolf et al., 2020) model for a binary classification task on a balanced dataset. The train and validation set are shuffled and stratified, and the test size is set to 0.33. Model performance evaluation is based on accuracy, precision, recall, and F1-score.

A.4.1 Software and Versions

A list of used python packages for fine-tuning the BERT models.

- transformers 4.57.1 (Wolf et al., 2020)
- numpy 2.2.4 (Harris et al., 2020)
- scikit-learn 1.7.2 (Pedregosa et al., 2011)
- regex 2025.11.3 (Barnett)
- torch 2.9.1 (Paszke et al., 2019)
- datasets 4.4.1 (Lhoest et al., 2021)
- pandas 2.3.3 (McKinney, 2010)

A.4.2 Model Architecture

This part shows the example architecture of a model with a learning rate of 3e-5. The training process for each model follows the same hyperparameters. The only individually adjusted parameter is learning rate depending on dataset size and type of text. The specific learning rate is selected to create a well-performing model after 3 epochs without overfitting by making sure performance on the first epoch is around chance. In the validation step the metric for the best performing model is the F1-score with the evaluation metrics accuracy, precision, recall, and F1-score calculated to compare the different models' performance.

```
model_name = 'bert-base-uncased'
tokenizer = BertTokenizerFast.from_pretrained(
    model_name)

def tokenize_function(examples):
```

```
    return tokenizer(
        examples["text"],
        truncation=True,
        padding="max_length",
        max_length=128)

config = BertConfig.from_pretrained(
    model_name,
    num_labels=2,
    hidden_dropout_prob=0.3,
    attention_probs_dropout_prob=0.3,
    classifier_dropout=0.3)

model = BertForSequenceClassification.from_pretrained(
    model_name,
    config=config)

training_args = TrainingArguments(
    eval_strategy = 'epoch',
    save_strategy = 'epoch',
    learning_rate = 3e-5,
    per_device_train_batch_size = 32,
    per_device_eval_batch_size = 32,
    num_train_epochs = 3,
    weight_decay = 0.01,
    warmup_ratio = 0.1,
    load_best_model_at_end = True,
    metric_for_best_model = 'f1',
    greater_is_better = True,
    save_total_limit = 2,
    report_to = 'none')
```

A.4.3 Learning Rates

2e-5 $BERT_{TEL}$, $BERT_{FlirtD}$, $BERT_{Mafia}$

1e-5 $BERT_{SoEng}$

5e-6 $BERT_{Mark}$, $BERT_{FlirtM}$

3e-6 $BERT_{Ads}$, $BERT_{DP}$, $BERT_{MM}$, $BERT_{P4G}$

2e-6 $BERT_{Phish}$

1e-6 $BERT_{CMV}$

A.4.4 Model Performance in-domain

Table 11 shows the performance of each model in-domain, including number of training and validation samples as described in subsection 4.2. Evaluation metrics are accuracy, precision, recall and F1-score. Each separate model has been trained to satisfactory performance. Because of the probabilistic nature of model training, results when reproducing this approach may vary. The training process for each model depends on the number of samples and varies between a few minutes and hours.

A.4.5 Feature-based

Table 12 shows feature-based evaluation of the different models. Using the penultimate layer of each model (removing the classification head), we calculate transferability metrics with the extracted features and true labels. Input, like during training, is truncated to 128 tokens. The transferability metrics are: LogME (Logarithm of Maximum Evidence) (You et al., 2021), NCE (Negative Conditional Entropy) (Tran et al., 2019), H-Score (Bao et al., 2019), k-NN (Sorkhei et al., 2025), TransRate (Huang et al., 2021), GBC (Gaussian Bhattacharyya Coefficient) (Pándy et al., 2022), SFDA (Source-Free Domain Adaptation) (Shao et al., 2022).

LogME approximates log maximum evidence via Laplace method on feature-label posterior. H-Score evaluates inter-class variance relative to feature covariance for class separability. k-NN assesses transferability by accuracy of nearest-neighbor classification on source-extracted target features. NCE measures uncertainty in predicting target labels from source features via optimal transport. TransRate

Model	#Train	#Val	Accuracy	Precision	Recall	F1
$BERT_{Ads}$	932	460	0.98	0.96	1.00	0.98
$BERT_{DP}$	9313	4587	0.91	0.85	1.00	0.92
$BERT_{FlirtM}$	1385	683	0.97	1.00	0.94	0.97
$BERT_{Mark}$	652	322	0.96	0.93	1.00	0.96
$BERT_{Phish}$	829	409	0.99	0.99	0.99	0.99
$BERT_{TEL}$	320	158	0.91	1.00	0.81	0.89
$BERT_{CMV}$	4088	2014	0.95	0.91	0.99	0.95
$BERT_{FlirtD}$	424	210	0.90	0.84	1.00	0.91
$BERT_{MM}$	3537	1743	0.92	0.87	0.99	0.92
$BERT_{Mafia}$	2994	1476	0.98	1.00	0.96	0.98
$BERT_{P4G}$	1362	672	0.92	0.88	0.96	0.92
$BERT_{SoEng}$	958	472	0.92	0.86	1.00	0.92

Table 11: In-Domain Model Performance on 0.33 validation size

gauges transfer quality through feature distribution alignment or prediction rates. GBC models feature distributions as Gaussians to score task compatibility. SFDA proxies adaptation potential without source data access, focusing on target-only metrics.

Higher scores in all metrics indicate better transferability. The scores from the feature-based assessment of transferability vary only slightly and are therefore difficult to properly evaluate for the best fitting model. The scores are very close to each other since they all use the same base model for fine-tuning which makes using these metrics less effective in determining the best model. We also show $BERT_{base}$ (bert-base-uncased) as a baseline to compare against, showing that fine-tuning on specialized corpora slightly improves transferability.

Model	LogMe	NCE	H-Score	k-NN	TransRate	GBC	SFDA	Avg Score
$BERT_{Ads}$	-0.0121	0.9955	0.8708	0.9970	0.9970	**1.0000**	0.0101	0.6940
$BERT_{DP}$	-0.0114	0.9941	**1.8755**	0.9950	0.9900	0.9980	**0.0317**	**0.8389**
$BERT_{FlirtM}$	-0.0.121	**1.0000**	0.8588	0.9970	0.9970	0.9970	0.0109	0.6927
$BERT_{Mark}$	-0.0124	**1.0000**	0.8286	**0.9980**	0.9970	**1.0000**	0.0081	0.6873
$BERT_{Phish}$	-0.0125	**1.0000**	0.8515	0.9990	0.9970	0.9980	0.0094	0.6918
$BERT_{TEL}$	-0.0119	0.9972	**0.9031**	**0.9990**	0.9970	0.9970	0.0104	**0.6991**
$BERT_{CMV}$	-0.0122	**1.0000**	0.7964	0.9990	**0.9980**	**1.0000**	0.0091	0.6843
$BERT_{FlirtD}$	**-0.0105**	**1.0000**	0.8480	0.9990	0.9970	0.9990	**0.0210**	0.6934
$BERT_{MM}$	-0.0116	0.9961	0.8753	**1.0000**	**0.9990**	0.9990	**0.0281**	0.6979
$BERT_{Mafia}$	**-0.0090**	0.9972	**2.0166**	1.0000	0.9970	0.9990	**0.1430**	**0.8777**
$BERT_{P4G}$	-0.0126	**1.0000**	0.8008	0.9980	0.9970	**1.0000**	0.0096	0.6847
$BERT_{SoEng}$	**-0.0112**	**1.0000**	**1.0377**	0.9970	0.9970	**1.0000**	0.0193	**0.7199**
$BERT_{base}$	-0.0132	0.9951	0.7769	0.9970	0.9970	1.0000	0.0079	0.6801

Table 12: Transferability Metrics, feature-based, n = 1000, **BOLD** - best performance