The Counterfactuals at SemEval-2026 Task 9: Can Counterfactually-Inspired Preprocessing help Detect Polarization?

Teagan Johnson
Department of Computer Science, University of Colorado Boulder
teagan.johnson@colorado.edu

Abstract

This paper presents the English-language submissions of "The Counterfactuals" team for the three subtasks of Task 9 at SemEval 2026. The task aims to detect multicultural online polarization, how it is expressed, and in what contexts. The task provides a high-quality annotation dataset of posts that follows a three-level schema: polarized or not (subtask 1), polarization type classification (subtask 2), and manifestation identification (subtask 3). I construct a pointwise mutual information-based lexicon that identifies highly-correlated words with the polarized class as labeled in subtask 1. Using this lexicon, I implement a large language model data augmentation technique. I then use the preprocessed datasets to finetune a BERT model (BERTweet) for each subtask. My highest performing models placed 48th out of 60, 35th out of 36, and 17th out of 24 on subtasks 1, 2, and 3 respectively. All code is available on GitHub.[1]

Warning: This paper contains examples that may be offensive or upsetting.

1 Introduction

Online polarization refers to deepening divisions and hostility between social, political, or identity groups (Iyengar et al., 2019; Garimella and Weber, 2018). It has become a growing concern, as prior work shows that polarized discourse often precedes or co-occurs with hate speech, toxic language, and broader social conflict (Waller et al., 2021; Mathew and et al., 2020). In its extreme form, polarization can produce fragmented online communities where individuals struggle to engage in constructive dialogue, undermining social cohesion and mutual understanding (Tucker et al., 2018; Williams et al., 2015). Detecting and mitigating polarization before it escalates is therefore crucial for promoting safer and more inclusive online environments. While existing approaches have attempted large-scale detection of polarized or partisan content online, this task has commonly been framed as a coarse binary classification task (polarized or not polarized) (Conover et al., 2011; Garimella et al., 2016). More fine-grained frameworks that identify how polarization is expressed and who it targets could provide greater interpretability and practical value for researchers, platform moderators, and users. Recently, SemEval-2026 Task 9: Detecting Multilingual, Multicultural, and Multievent Online Polarization (Naseem et al., 2026a) has been proposed to encourage research related to detecting online polarized language. By participating in this task, my goal is to identify and categorize the type of polarization in online dialogue.

Spurious Correlation Example

Polarized Example
*"They constantly accuse others of violence while using their own people as **human shields**—it's pure hypocrisy."*

Non-polarized Example
*"Women are not **human shields** for men."*

Figure 1: Example of a spurious correlation: the model predicts both of these examples as polarized even though only the first example is really labeled as polarized.

I initially fine-tuned BERTweet (Nguyen et al., 2020) on the annotated training data provided by the task organizers for subtask 1 (more details on the tasks in Section 3) to classify polarization (Naseem et al., 2026a). An initial inspection of the fine-tuned model's output revealed that the presence of certain keywords seemingly disproportionately influenced the classifier's decisions. In many cases, the model appeared to rely on specific unigrams or bigrams rather than learning the deeper

[1] https://github.com/johnsont4/semeval-task9-polarization-detection/tree/main

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 387–401

contextual patterns that signal polarization. One example of this behavior can be seen in Figure 1 regarding the spurious term “human shields." This motivates my effort to reduce the model’s dependence on specific spurious correlations and encourage more robust, context-sensitive representations.

My work is guided by two research questions. **(1)** How effective is pointwise mutual information (PMI) for identifying a lexicon of spurious, polarization-inducing tokens? **(2)** Can large language models (LLMs) be used to augment the training data in ways that reduce reliance on specific unigrams and bigrams and promote learning of contextual signals?

2 Related Work

Political polarization and social media. A large body of work studies political polarization in social media, especially on Twitter (now X), finding that polarization is tightly coupled with specific topical and group-related vocabularies (Conover et al., 2011; Garimella and Weber, 2017; Barberá, 2015). This suggests that purely lexical models are especially vulnerable to learning spurious correlations between particular terms and polarized labels.

Spurious lexical correlations and identity-term bias. Similar issues have been documented in toxic language detection, where hate speech datasets over-represent certain identity terms in the positive class, leading models to conflate mentions of marginalized groups with abusive content (Waseem and Hovy, 2016; Davidson et al., 2017). Dixon et al. (Dixon et al., 2018) formalize this as *identity term bias*, showing that toxicity classifiers over-predict toxicity for benign sentences containing terms such as “gay”. My work treats highly associated terms as potential sources of unintended bias and explicitly targets them during preprocessing.

LLM-based data augmentation. Recent work leverages LLMs as generators of synthetic training data for text classification, showing improvements particularly for minority classes, though naive augmentation may introduce distributional shifts that harm robustness (Zhao et al., 2024; Chai et al., 2025). My counterfactual augmentation procedure fits within this broader paradigm.

3 Task Description

SemEval 2026 Task 9 consists of 3 subtasks covering 22 languages. I am participating in all 3 subtasks in English (Naseem et al., 2026b).

3.1 Polarization Detection Task

Subtask 1 is a binary classification task: given a post, determine whether it contains polarized content. Polarized posts display a negative attitude toward outgroups while showing blind support for ingroups, encompassing language that incites division, groupism, hatred, conflict, or intolerance. Label counts are shown in Table 1.

Polarized	Non-Polarized
1,175	2,047

Table 1: Counts for the subtask 1 dataset.

3.2 Polarization Type Classification Task

Subtask 2 is a multi-label classification task to identify which of five polarization targets are expressed: political, racial/ethnic, religious, gender/sexual, or other. Each category is an independent binary label, and a post may be associated with multiple targets simultaneously. Label counts are shown in Table 2.

Polarization Type	Count
Political	1,150
Racial/Ethnic	281
Religious	112
Gender/Sexual	72
Other	126

Table 2: Counts per label in the subtask 2 dataset.

3.3 Manifestation Identification Task

Subtask 3 is a multi-label classification task to identify which of six manifestation types are present: stereotype, vilification, dehumanization, extreme language, lack of empathy, and invalidation. As with subtask 2, each label is independent and multiple labels may apply simultaneously. Label counts are shown in Table 3.

4 Methods

Upon initial examination of the training data, I identified strong correlations between certain words (e.g. Human Shields, Zionist, Leftist) and the polarization classification. These specific words are

Polarization Manifestation	Count
Stereotype	487
Vilification	858
Dehumanization	391
Extreme Language	770
Lack of Empathy	357
Invalidation	586

Table 3: Counts per label in the subtask 3 dataset.

features that allow the model to classify based on topic-specific words rather than actual indicators of polarization. In prior work these have been referred to as *spurious correlations*. To reduce reliance on these learned shortcuts, I construct a lexicon of spurious tokens using Pointwise Mutual Information (PMI) and introduce LLM-assisted counterfactual data augmentation to explicitly break correlations between spurious tokens and labels.

4.1 Creating a Spurious Lexicon with PMI

Pointwise mutual information (PMI) has long been used to quantify word–context association and to construct lexical resources (Church and Hanks, 1990). To quantify the degree to which words are associated with polarized posts, I compute pointwise mutual information (PMI) between each token w and the polarized class label $y \in \{0, 1\}$. PMI measures how much more frequently a word co-occurs with a class than would be expected under independence:

$$\text{PMI}(w, y) = \log \frac{p(w, y)}{p(w)\, p(y)}.$$

Probabilities are estimated from frequencies in the training data where $p(w, y)$ signifies the fraction of posts containing w with label y, and $p(w)$ and $p(y)$ signify the marginal word and class probabilities. I consider words whose PMI with the polarized class exceeds 0.75 to form my *spurious lexicon*. This threshold was selected to balance lexicon size against noise: I evaluated several candidate thresholds and found that 0.75 yielded a lexicon of reasonable size without admitting too many low-frequency terms. Importantly, I don't manually curate the spurious lexicon at all. I want to evaluate the models' performance relying purely on the lexicon created by PMI. The terms included in the lexicon often encode political identities, geographic entities, or demographic descriptors that correlate with polarization in the dataset but are not in themselves evidence of polarized framing. See Table 4 for a sample of the top 10 unigrams with the highest PMI and Table 8 for the top 10 bigrams.

I construct a single global lexicon from the subtask 1 binary labels and apply it uniformly across all three subtasks. I chose this design because subtask 1 provides the cleanest signal for spurious correlations: the binary polarized/non-polarized distinction maps directly onto the PMI formulation. Computing separate label-conditional lexicons for subtasks 2 and 3 is complicated by the small per-label instance counts in those multi-label settings (e.g., only 72 Gender/Sexual instances in subtask 2), which would make per-label PMI estimates unreliable. I treat this as a deliberate simplification and acknowledge it as a limitation in the discussion.

4.2 Finetuning BERTweet Model for Classification

I used BERTweet, a RoBERTa-based model trained on ∼850M English tweets, as a baseline classifier for this task (Nguyen et al., 2020). The motivation behind using BERTweet was the similarity of its training corpus and my task data: social media text involves fragmented grammar, slang, sarcasm, and short noisy sequences that BERTweet is specifically adapted to handle. Each model consists of the BERTweet encoder with a linear classification head over the pooled `[CLS]` representation. For subtask 1 (binary), I use a single output unit with binary cross-entropy loss. For subtasks 2 and 3 (multi-label), I use one output unit per label with independent sigmoid activations and per-label binary cross-entropy loss. I fine-tuned BERTweet for all 3 subtasks, training on 3 different preprocessed datasets per subtask, resulting in 9 fully fine-tuned models in total.

4.3 LLM-Assisted Counterfactual Augmentation

To teach the model that the *label should be invariant* to changes in group-specific or topic-specific tokens, I generate counterfactual training examples using a large language model. I use the Llama-3.1-8B-Instruct language model with the following decoding parameters: temperature$= 0.8$, top-$p = 0.9$, and `max_new_tokens`$= 128$.

For subtask 1 (binary classification), my augmentation procedure generates both polarized and non–polarized counterfactuals. For each training example labeled as "Polarized" and containing a token from the spurious lexicon, I prompt the LLM

Unigram	PMI	Count in Polarized Label	Total Unigram Count
felon	1.2728	10	10
fuck	1.2728	29	29
agenda	1.1732	14	15
fascist	1.1658	26	28
kids	1.1573	12	13
let	1.1573	12	13
liberal	1.1415	21	23
genocidal	1.1352	10	11
stolen	1.1284	19	21
antisemitism	1.1207	9	10

Table 4: Top 10 unigram PMI scores associated with polarization. Full spurious unigram lexicon can be seen in Table 15.

to produce six perturbations: (1) three semantically comparable versions in which the spurious term is replaced by a *different but comparable* group such that the LLM still judges the example to be polarized, and (2) three semantically altered versions that the LLM considers non–polarizing but which retain the original spurious term. This design provides the model with balanced evidence that the presence of a spurious token is neither necessary nor sufficient for polarization, which is essential for learning invariance in the binary setting. A concrete example of my augmentation process for subtask 1 is shown in Figure 2.

For subtasks 2 and 3 (multilabel classification), however, it is substantially more difficult to generate non–polarized counterfactuals. These subtasks involve 5 and 6 labels respectively, and replacing a spurious term does not naturally yield a controlled non-polarized counterpart. As a result, my augmentation for these subtasks only generates *polarized* counterfactuals. Specifically, alternative versions of the input where the spurious term is swapped for another entity while preserving the polarization label. **This allows the model to observe variation in which entity is referenced, but without the label contrast available in subtask 1.** I return to this in the discussion.

For **subtask 1**, I ultimately cap the number of augmentations produced to 100 instances, which means the LLM produced 600 total augmentations (300 polarized, 300 non-polarized). For **subtask 2**, the LLM produced 228 total augmentations. For **subtask 3**, the LLM produced 228 total augmentations.

Malformed or placeholder outputs from the LLM (an example of which is shown in Figure 3) were identified by checking for template phrases (e.g., “first new polarized post here”) and discarded automatically. No formal human validation of label correctness was performed; I relied solely on the LLM’s own judgment to preserve labels during generation. I acknowledge this as a limitation: label noise in the augmented data may have dampened the effectiveness of the augmentation strategy, and human validation or automatic filtering based on a secondary classifier is left for future work.

4.4 Blindness Preprocessing

As a point of comparison and to help with error analysis, I run a second preprocessing step that I call “blindness" or “masked" preprocessing. Instead of replacing the identity terms found with the spurious lexicon, I replace the terms with a [BLIND] token 85% of the time (remaining 15% of the time the token is unchanged). The 85% rate was chosen to ensure that spurious tokens are masked in the majority of training examples while retaining some stochasticity, preventing the model from simply learning to rely on the absence of the token.

5 Results

For each of the fine-tuned BERT models I use the same hyperparameters which can be found in Table 17. Results tables include a **POLAR Baseline**, which is the official baseline score provided by the SemEval-2026 Task 9 organizers (Naseem et al., 2026a) for comparison. I find the optimal thresholds for each label in subtasks 2 and 3 by performing a simple grid search over thresholds in [0.05, 0.95] (step size 0.05) and choosing, for each label independently, the value that maximizes its binary F1 score. I also split the annotated training data provided by the organizers into train, validation, and test sets, using the same split across all three subtasks (Table 14). Full evaluation results on my internal splits are reported in Appendix Tables 9–13.

5.1 Binary Polarization Detection Task

I use the LLM-augmented training data to train BERTweet. The organizers provided a train set which I partition into separate train/val/test splits. The f1, precision, and recall scores on my splits are provided in Table 9 in the Appendix. The f1 score achieved after uploading my results to the SemEval Task 9 server are shown in Table 5. These f1 scores were calculated from a separate label-blind test set provided by the organizers. All models cluster around 0.77–0.78 macro-F1, indicating that BERTweet is already a strong baseline for binary polarization detection and that the augmentation provided minimal additional gain on the organizer test set.

Model	SemEval Macro-F1
POLAR Baseline	0.78
LLM-Aug BERT	0.77

Table 5: Final test set F1 scores obtained from predictions submitted to the organizer-provided evaluation server for **subtask 1**.

5.2 Polarization Type Classification Task

I extend the spurious lexicon created from subtask 1 to this subtask. The f1 scores achieved after uploading my results to the SemEval Task 9 server are shown in Table 6. The macro f1 and f1 scores per label are shown in Table 10 and Table 11, respectively. Both models match the baseline at 0.33 macro-F1, reflecting the difficulty of this task: the severe class imbalance across types (e.g., 1,150 Political vs. 72 Gender/Sexual instances) makes it challenging to generalize across all five categories.

Model	SemEval Macro-F1
POLAR Baseline	0.33
LLM-Aug BERT	0.33

Table 6: Final test set F1 scores obtained from predictions submitted to the organizer-provided evaluation server for **subtask 2**.

5.3 Manifestation Identification Task

I extend the spurious lexicon created from task 1 to this task. The f1 scores achieved after uploading my results to the SemEval Task 9 server are shown in Table 7. The macro f1 and f1 scores per label are displayed in Table 12 and Table 13 respectively. The LLM-augmented model marginally outperforms the POLAR baseline (0.43 vs. 0.41), making subtask 3 the only setting where augmentation produced a measurable gain on the organizer evaluation.

Model	SemEval Macro-F1
POLAR Baseline	0.41
LLM-Aug BERT	0.43

Table 7: Final test set F1 scores obtained from predictions submitted to the organizer-provided evaluation server for **subtask 3**.

6 Discussion

6.1 Effect of LLM-Assisted Counterfactual Augmentation

For **subtask 1**, overall macro-F1 remained almost the same as seen in Table 5. Although my augmentation strategy explicitly created contrastive polarized/non-polarized pairs designed to teach label invariance, two factors likely limited its impact. First, the augmentation volume was small: 600 generations from 100 source instances is modest relative to the 1,175 polarized training examples already available. Second, the dataset's class imbalance (1,175 polarized vs. 2,047 non-polarized; Table 1) means the model has access to substantial non-polarized signal already, reducing the marginal benefit of generated non-polarized counterfactuals.

For **subtask 2**, macro-F1 remained almost the same as baseline as seen in Table 6. Looking at the per-class F1, Table 11 shows the LLM-Aug model substantially improved performance on the *Religious* class (0.37 $\rightarrow$ 0.53) and modestly improved *Gender/Sexual* (0.16 $\rightarrow$ 0.22), while performance decreased for *Racial/Ethnic* (0.62 $\rightarrow$ 0.45). A plausible explanation is that religious and gender/sexual terms are lexically entangled with polarized framing across many contexts, so swapping them with comparable terms genuinely teaches invariance. Racial/ethnic terms, by contrast, may be semantically central enough that replacing them introduces distributional noise rather than reducing spurious reliance.

I achieve a performance boost compared to baseline for **subtask 3**. The LLM-Aug model achieved improvements in specific fine-grained categories as seen in Table 13, notably *Dehumanization* (0.42 $\rightarrow$ 0.47) and *Lack of Empathy* (0.23 $\rightarrow$ 0.24). These categories are characterized more by linguistic stance and framing than by the presence of specific

identity tokens, so reducing the model's reliance on particular entities may push it toward learning the deeper discourse cues that actually define these manifestations. This pattern is consistent with the hypothesis that entity-swapping augmentation is most effective for label categories that are structurally rather than lexically defined.

6.2 Masking vs. Augmentation: A Key Contrast

The divergent effects of masking and augmentation reveal something important about the nature of the spurious correlations in this dataset. Plain masking from the PMI-derived lexicon consistently reduced performance, particularly in fine-grained subtasks. For example, as seen in Table 11, masking dramatically harmed the *Other* category (0.24 → 0.07) and reduced *Dehumanization* (0.42 → 0.36). These results indicate that many high-PMI tokens are not purely spurious but carry genuine semantic signal for polarization; removing them deprives the model of crucial contextual information. Augmentation, by contrast, keeps the tokens present while varying the labels around them, teaching the model that the token alone is insufficient evidence — a fundamentally different inductive signal than erasure.

This contrast suggests that the right strategy for this task is not to suppress identity-related vocabulary but to diversify the label contexts in which it appears. A key limitation of the PMI lexicon in both settings is that the purely threshold-based approach introduces noise (e.g., function-word N-grams such as "let" or "is just") and inflates scores for rare terms: several top-ranked unigrams (e.g., "felon", "antisemitism") appear only 9–10 times in the polarized class (Tables 4, 15), making their PMI values statistically fragile. Future work should apply minimum frequency thresholds and linguistic filtering to better distinguish spurious cues from meaningful semantic signals.

7 Conclusion

I presented a counterfactually-inspired preprocessing pipeline for polarization detection in English social media posts. Using a PMI-derived spurious lexicon built from subtask 1 binary labels, I generated LLM-assisted counterfactual training examples and applied stochastic token masking as a complementary debiasing strategy, then fine-tuned BERTweet separately for each of the three subtasks. My results show that LLM augmentation produced modest but meaningful gains on subtask 3 (manifestation identification), particularly for fine-grained categories such as Dehumanization and Lack of Empathy, while offering little benefit for the binary and type classification subtasks. Plain masking, by contrast, consistently degraded performance, suggesting that many high-PMI tokens carry genuine semantic signal rather than acting as pure spurious shortcuts. Returning to the research questions: **(1)** PMI proves effective at surfacing a lexicon of candidate spurious tokens, though it is prone to inflating scores for rare terms and cannot distinguish genuinely spurious cues from semantically central vocabulary — making it a useful but noisy signal. **(2)** LLM augmentation can reduce reliance on specific tokens, but only partially: gains were concentrated in the most fine-grained subtask and were limited by small augmentation volume and unvalidated label quality. Key limitations include the use of a single global lexicon across subtasks, unvalidated LLM label quality, and small augmentation volume. Future work should explore label-conditional PMI lexicons, frequency-thresholded PMI estimation, and human or automatic validation of augmented examples to better realize the potential of counterfactual data augmentation for polarization detection.

8 Acknowledgements

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder (of Colorado Boulder Research Computing, 2021). I also want to extend my appreciation to Dr. Jim Martin for his support and his instruction during the CSCI-LING 5832 course at the University of Colorado, Boulder.

References

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Yaping Chai, Haoran Xie, and Joe S. Qin. 2025. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *Preprint*, arXiv:2501.18845.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and

Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 89–96.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy on social media. In *WSDM*.

Kiran Garimella and Ingmar Weber. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 528–531.

Kiran Garimella and Ingmar Weber. 2018. Political polarization on social media: a literature review. In *International Conference on Social Informatics*, pages 34–35.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. Affective polarization in the digital age. *Annual Review of Political Science*, 22:129–146.

Binny Mathew and et al. 2020. Analyzing polarization in social media: Method and application to tweets on climate change. In *WWW*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

University of Colorado Boulder Research Computing. 2021. Blanca condo cluster.

Joshua A. Tucker, Andrew Guess, Pablo Barberá, and et al. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political Polarization Workshop Report*.

Isaac Waller, Ashton Anderson, Yong-Yeol Park, and Michael Macy. 2021. Quantifying toxicity and contextualizing hate speech in online political discourse. In *ICWSM*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Hywel Williams, Jeffrey McMurray, and Tim Kurz. 2015. Network communities and political polarization online. *Political Studies*.

Huanhuan Zhao, Haihua Chen, Yunhe Feng, and Hong Jun Yoon. 2024. Improving text classification with large language model-based data augmentation. *Electronics*, 13(13):2535.

A Appendix

Bigram	PMI	Count in Polarized Label	Total Bigram Count
liberal agenda	1.2576	11	11
stolen election	1.2576	18	18
ukrainian nazis	1.1645	15	16
is just	1.1201	10	11
as human	1.1056	18	20
the radical	1.0650	14	16
the woke	1.0650	21	24
rigged election	1.0232	17	20
woke mob	1.0125	27	32
human shields	1.0013	36	43

Table 8: Top 10 bigram PMI scores associated with polarization. Full spurious bigram lexicon can be seen in Table 16.

Model	Macro F1	Precision	Recall
Base BERT	**0.78**	0.75	**0.82**
Masked BERT	0.77	**0.77**	0.77
LLM-Aug BERT	0.77	**0.77**	0.77

Table 9: Subtask 1 overall metrics.

Model	Macro F1	Precision	Recall
Base BERT	0.41	0.39	**0.76**
LLM-Aug BERT	**0.42**	**0.42**	**0.65**
Masked BERT	0.33	0.43	0.65

Table 10: Subtask 2 overall metrics.

Class	Base F1	LLM-Aug F1	Masked F1
Political	0.66	**0.67**	0.64
Racial/Ethnic	**0.62**	0.45	0.46
Religious	0.37	**0.53**	0.33
Gender/Sexual	0.16	**0.22**	0.16
Other	**0.24**	**0.24**	0.07

Table 11: Subtask 2 per-class F1 scores.

Model	Macro F1	Precision	Recall
Base BERT	**0.47**	0.38	0.70
LLM-Aug BERT	0.44	0.31	**0.82**
Masked BERT	0.46	**0.41**	0.61

Table 12: Subtask 3 overall metrics.

Class	Base F1	LLM-Aug F1	Masked F1
Stereotype	0.46	0.42	**0.50**
Vilification	**0.62**	0.54	**0.62**
Dehumanization	0.42	**0.47**	0.36
Extreme Language	**0.58**	**0.49**	0.56
Lack of Empathy	0.23	**0.24**	0.20
Invalidation	0.48	0.48	**0.50**

Table 13: Subtask 3 per-class F1 scores.

Split	Number of Examples	Percentage
Train	2,577	80%
Validation	332	10%
Test	332	10%

Table 14: Dataset split statistics for all three subtasks.

Example of LLM-Augmented Training Data

Original Input:
I love the smell of **Christian Nationalism** on New Years Eve blakeshelton

Polarized Entity-Swapping Augmentations:

- There's something nauseating about the smell of **Far Right Fundamentalism** wafting through the air on New Years Eve blakeshelton
- The stench of **Nationalist Populism** is overwhelming on this New Years Eve, a vile reminder of our country's darker side blakeshelton
- I hate the smell of **White Christian Conservatism** on New Years Eve, it's choking our democracy blakeshelton

Non-Polarized Entity-Preserving Augmentations:

- New Year's Eve gatherings often bring people together to share common values and traditions, including those related to **Christian Nationalism**.
- **Christian Nationalism** has become a significant aspect of American culture, influencing various aspects of society.
- The concept of **Christian Nationalism** is worth exploring, especially during New Year's Eve celebrations.

Figure 2: An example of my LLM augmentation procedure. Given an input containing a spurious entity (*Christian Nationalism*), the model generates (1) **polarized augmentations** by swapping the entity with semantically comparable alternatives, and (2) **non-polarized augmentations** that preserve meaning while removing inflammatory framing.

Unigram	PMI	Count in Polarized	Total Count
felon	1.2728	10	10
fuck	1.2728	29	29
agenda	1.1732	14	15
fascist	1.1658	26	28
kids	1.1573	12	13
let	1.1573	12	13
liberal	1.1415	21	23
genocidal	1.1352	10	11
stolen	1.1284	19	21
antisemitism	1.1207	9	10
full	1.1207	9	10
rigged	1.0711	20	23
racism	1.0504	12	14
hes	1.0504	18	21
hate	1.0504	18	21
shields	1.0164	36	43
nazis	1.0097	20	24
progressive	1.0097	10	12
nazi	1.0097	10	12
traitor	0.9971	19	23
mob	0.9926	28	34
musk	0.9890	23	28
woke	0.9873	32	39
feel	0.9832	9	11
literally	0.9832	9	11
radical	0.9773	22	27
everyone	0.9508	12	15
apartheid	0.9508	40	50
using	0.9508	12	15
away	0.9508	12	15
cleansing	0.9191	36	46
human	0.9143	39	50
ethnic	0.9060	38	49
being	0.8984	27	35
trumpism	0.8942	20	26
ukrainian	0.8894	23	30
enemy	0.8804	16	21
them	0.8668	40	53
talking	0.8577	9	12
shit	0.8253	11	15
fascism	0.8253	11	15
claim	0.8133	8	11
facts	0.8133	8	11
elon	0.8033	13	18
use	0.8033	13	18
always	0.7988	18	25
left	0.7947	28	39
care	0.7873	10	14
keep	0.7873	10	14
something	0.7873	10	14

Table 15: All unigrams in the corpus with a PMI greater than 0.75. This ultimately becomes the first part of my spurious lexicon.

Bigram	PMI	Count in Polarized Label	Total Count
liberal agenda	1.26	11	11
stolen election	1.26	18	18
ukrainian nazis	1.16	15	16
is just	1.12	10	11
as human	1.11	18	20
the radical	1.06	14	16
the woke	1.06	21	24
rigged election	1.02	17	20
woke mob	1.01	27	32
human shields	1.00	36	43
the deep	0.99	10	12
when the	0.99	10	12
apartheid state	0.99	34	41
is an	0.98	14	17
the election	0.97	9	11
they are	0.97	18	22
a traitor	0.97	9	11
voted for	0.97	9	11
enemy of	0.96	13	16
are the	0.96	13	16
radical left	0.95	17	21
of israel	0.94	8	10
christian nationalism	0.94	8	10
israel is	0.91	11	14
ethnic cleansing	0.90	36	46
trump is	0.90	18	23
should be	0.84	12	16
elon musk	0.84	9	12
and his	0.84	9	12
deep state	0.82	17	23
to take	0.80	8	11
an apartheid	0.77	15	21
the people	0.76	17	24
defund the	0.76	12	17

Table 16: All bigrams in the corpus with a PMI greater than 0.75. This ultimately becomes the second part of my spurious lexicon.

Setting	Value
Base model	`vinai/bertweet-base`
Max sequence length	128
Train batch size	8
Eval batch size	8
Number of epochs	3
Threshold during training metrics	0.5 (per label)
Evaluation strategy	`epoch`
Checkpoint saving	`save_strategy=no` (final model only)
Logging frequency	every 50 training steps
Metrics	macro precision, macro recall, macro F1

Table 17: Model architecture and training hyperparameters for BERTweet fine-tuning on all three subtasks.

Example of a Mistake from the LLM

Original Input:
“They" are now the **deep state**.

Polarized Entity-Swapping Augmentations:

- third new polarized post here
- second new polarized post here
- first new polarized post here

Figure 3: An example of where the LLM makes a mistake when generating additional polarized examples. The spurious token is “deep state" (a bigram), but the model only returns placeholder text instead of actual polarization examples.

```
We are studying polarization in social media posts.

- A "polarized" post clearly expresses a strong, divisive attitude or opinion,
  usually framing an "us vs them" dynamic, strongly supporting one side and/or
  attacking another.
- Here, all original posts you will see are labeled as POLARIZED (1).

You will be given a social media post that is labeled as POLARIZED (1).
It expresses a strong, divisive attitude.

Original post:
{text}

Groups or entities mentioned in the original post:
[{mentions_str}]

Your task:
- Generate 3 new posts that:
  1. Remain clearly POLARIZED (strong, divisive opinion).
  2. Keep a similar *type* of stance (e.g., critical, supportive, mocking),
     but you must change the specific groups, parties, countries, or movements
     mentioned to DIFFERENT BUT SOCIALLY COMPARABLE ones.
     For example:
     - If the original attacks "Democrats", you might attack "Republicans" or "liberals".
     - If the original talks about one country, you may switch to another country
       in a similar geopolitical context.
  3. The intensity of polarization should remain similar (don't make it neutral).
  4. Do NOT copy the original sentences; produce
        genuinely rephrased posts following the same syntactical structure.
  5. Do NOT mention exactly the same groups/entities as in the original.
  6. Do NOT add explanations; only return JSON.
```

Figure 4: Prompt used for **polarized** LLM-based data augmentation for subtask 1.

```
We are studying polarization in social media posts.

- A "polarized" post clearly expresses a strong, divisive attitude or opinion,
  usually framing an "us vs them" dynamic, strongly supporting one side and/or
  attacking another.
- Here, all original posts you will see are labeled as POLARIZED (1).

You will be given a social media post that is labeled as POLARIZED (1),
meaning it expresses a strong, divisive attitude.

Original post:
{text}

Groups or entities mentioned in the original post:
[{mentions_str}]

Your task:
- Generate 3 new posts that:
  1. Are clearly NON-POLARIZED (neutral, balanced, or mildly opinionated).
  2. Should NOT frame a strong "us vs them" conflict or use aggressive, hostile language.
  3. Stay roughly on the same topic as the original.
  4. Mention the groups found ([{mentions_str}]),
     but only in a descriptive, informational, or balanced way.
  5. Do NOT copy the original sentences;
        produce genuinely rephrased posts that may have syntactically different structure.
  6. Do NOT add explanations; only return JSON.
```

Figure 5: Prompt used for **non-polarized** LLM-based data augmentation for subtask 1.

```
"You generate counterfactual training data for polarization type classification.\n"
"Given a sentence, you rewrite it so that it preserves the same meaning and "
"polarization type, but replaces references to specific countries, political groups, "
"or identity groups with different but comparable groups.\n\n"
"STRICT OUTPUT RULES:\n"
"- Return ONLY the rewritten sentence.\n"
"- Do NOT provide any explanation or commentary.\n"
"- Do NOT number or bullet the output.\n"
"- Output MUST be a single line of text.\n"
```

Figure 6: Prompt used for LLM-based data augmentation for subtask 2.

```
"You generate counterfactual training data for polarization type classification.\n"
"Given a sentence, you rewrite it so that it preserves the same meaning and "
"polarization type, but replaces references to specific countries, political groups, "
"or identity groups with different but comparable groups.\n\n"
"STRICT OUTPUT RULES:\n"
"- Return ONLY the rewritten sentence.\n"
"- Do NOT provide any explanation or commentary.\n"
"- Do NOT number or bullet the output.\n"
"- Output MUST be a single line of text.\n"
```

Figure 7: Prompt used for LLM-based data augmentation for subtask 3.

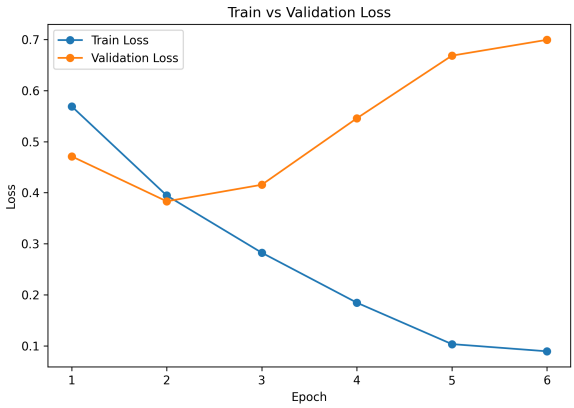

Figure 8: Loss curves for the base BERT model for subtask 1.

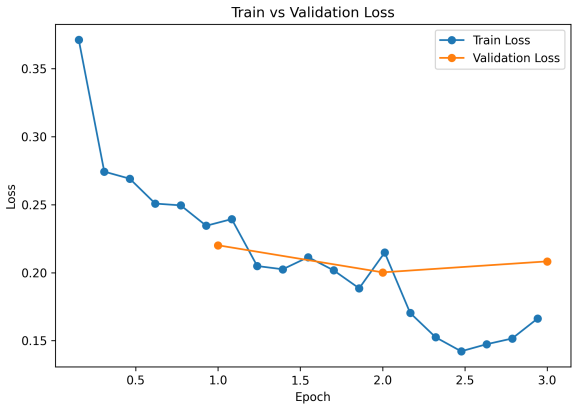

Figure 11: Loss curves for the base BERT model for subtask 2.

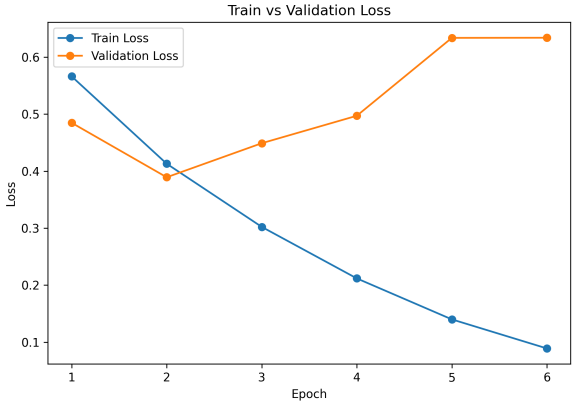

Figure 9: Loss curves for the masked BERT model for subtask 1.

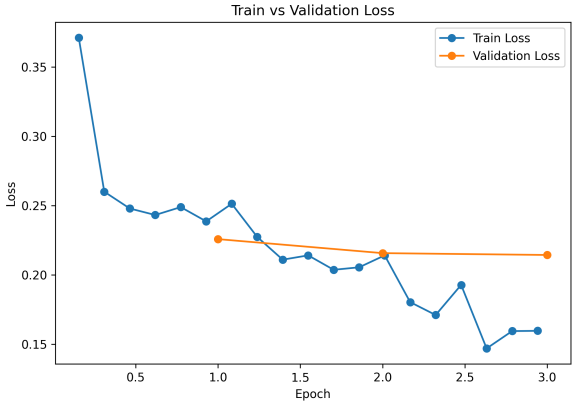

Figure 12: Loss curves for the masked BERT model for subtask 2.

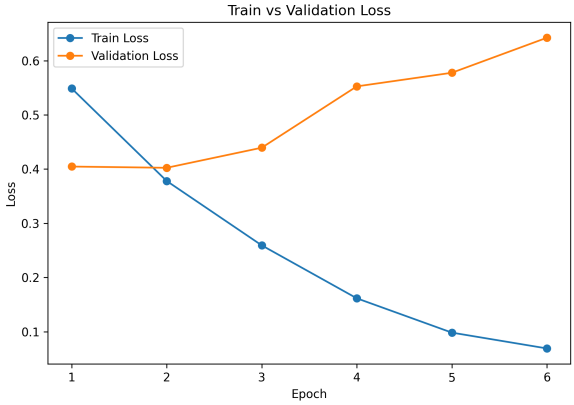

Figure 10: Loss curves for the LLM-augmented BERT model for subtask 1.

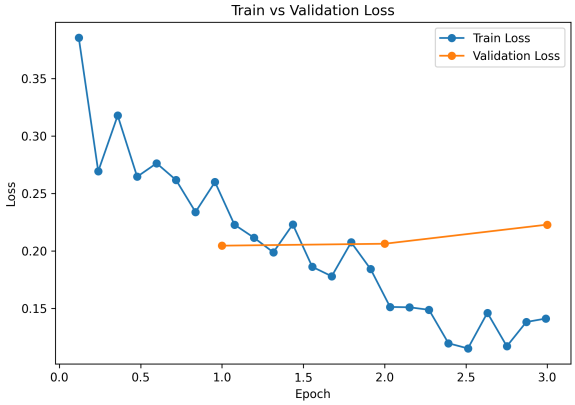

Figure 13: Loss curves for the LLM-augmented BERT model for subtask 2.

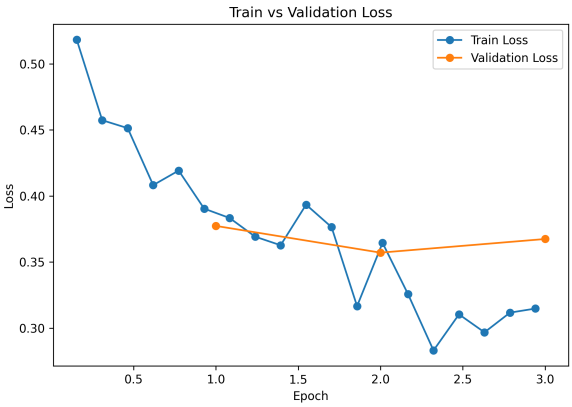

Figure 14: Loss curves for the base BERT model for subtask 3.

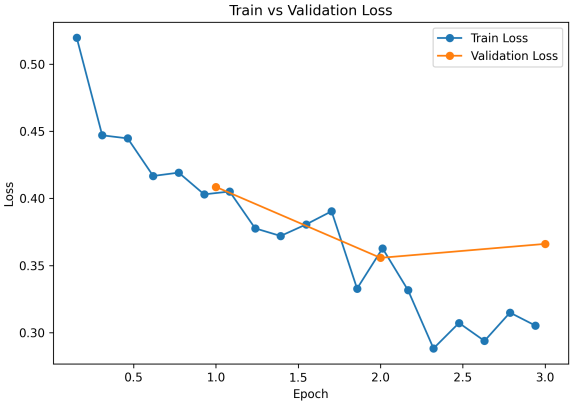

Figure 15: Loss curves for the masked BERT model for subtask 3.

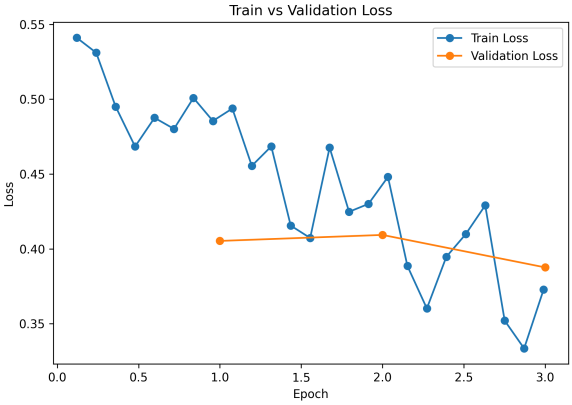

Figure 16: Loss curves for the LLM-augmented BERT model for subtask 3.