Benchmarking Hate Speech Detection in Azerbaijani with Turkish Cross-Lingual Transfer and Transformer Models

Tural Alizada[1] **and Haim Dubossarsky**[1,2,3]
[1] Queen Mary University of London
[2] Language Technology Lab, University of Cambridge
[3] The Alan Turing Institute
turalalizadeh4@gmail.com, h.dubossarsky@qmul.ac.uk

Abstract

In this paper, we investigated the task of hate-speech classification in the closely related Turkic language pair, Turkish-Azerbaijani. Transformer models can achieve strong hate-speech classification in Turkish, but their performance does not reliably transfer to closely related low-resource languages without careful evaluation. We study Turkish–Azerbaijani hate speech detection and introduce the first manually annotated Azerbaijani benchmark, comprising 1,112 YouTube comments from major news channels with severe class imbalance. We compare XLM-RoBERTa and a compact BERT-Tiny model against a TF–IDF + logistic regression baseline under monolingual training, zero-shot Turkish→Azerbaijani transfer, low-resource balanced subsampling, bilingual mixed fine-tuning, and translation-based augmentation using machine-translated Turkish data. XLM-R attains high macro-F1 in Turkish and achieves moderate zero-shot transfer to Azerbaijani, but native Azerbaijani training is fragile for the hate class. Mixed bilingual training improves robustness for both languages, whereas TF–IDF generalizes poorly to Azerbaijani.

1 Introduction

The rapid growth of user-generated content on social media has intensified the need for scalable hate speech detection systems. While transformer-based models have achieved strong performance in high-resource languages, progress in low-resource and under-resourced languages remains constrained by the scarcity of annotated data and language-specific tooling.

This work focuses on hate speech detection in Azerbaijani, a low-resource Turkic language for which no publicly available annotated dataset previously existed. We study Azerbaijani in conjunction with Turkish, a closely related Oghuz Turkic language with substantially richer resources, and investigate the extent to which cross-lingual transfer can compensate for limited native supervision. Although Turkish has publicly available resources (e.g., the Turkish Hate Speech Superset; Tonneau, 2022), Azerbaijani has lacked a manually annotated benchmark.

We introduce the first manually annotated Azerbaijani hate speech dataset, consisting of 1,112 YouTube comments collected from major Azerbaijani news channels. Using this dataset, we benchmark several modeling strategies: (i) monolingual Azerbaijani training, (ii) zero-shot Turkish→Azerbaijani transfer, (iii) bilingual mixed training under low-resource constraints, and (iv) translation-based augmentation via machine-translated Turkish data. Our experiments compare a multilingual transformer (XLM-RoBERTa), a lightweight transformer (BERT-Tiny), and a classical TF–IDF + logistic regression baseline.

Empirically, we show that while XLM-RoBERTa performs strongly on Turkish and achieves reasonable zero-shot transfer to Azerbaijani, native Azerbaijani training remains fragile under class imbalance. Mixed Turkish–Azerbaijani training improves robustness for both languages. In contrast, linear TF–IDF models that are competitive in Turkish fail on Azerbaijani, highlighting the importance of multilingual contextual representations for morphologically rich, imbalanced settings. Overall, this study provides the first empirical benchmarks for Azerbaijani hate speech detection and practical guidance on cross-lingual transfer for closely related low-resource languages. The dataset, and documentation are made publicly available at https://github.com/alizadeht/azerbaijani-hate-speech.

2 Related Work

Hate speech detection has been widely studied in NLP, with early systems relying on linear classi-

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 103–112
March 29, 2026 ©2026 Association for Computational Linguistics

fiers trained over surface features such as word and character n-grams and TF–IDF representations (Schmidt and Wiegand, 2017; Davidson et al., 2017). Despite their simplicity and interpretability, bag-of-words approaches often degrade under domain shift and struggle with phenomena that require context or compositional semantics, such as implicit abuse, sarcasm, and figurative language–challenges that are amplified in morphologically rich languages.

Transformer architectures have become the dominant paradigm for toxicity and hate speech classification due to their contextual encoding and transferability (Vaswani et al., 2017; Devlin et al., 2019). Alongside large models, compact variants have been proposed for efficiency-constrained settings (Sanh et al., 2019; Turc et al., 2019), motivating explicit comparisons between lightweight and multilingual transformers when compute and latency constraints matter.

For low-resource languages, multilingual pretraining enables cross-lingual transfer without requiring large native corpora. Models such as multilingual BERT and XLM-RoBERTa are pretrained over many languages with shared subword vocabularies and are widely used for zero-shot and few-shot transfer (Devlin et al., 2019; Conneau et al., 2020; Pires et al., 2019). However, transfer quality varies substantially across language pairs and domains, and it can be fragile for morphologically rich and underrepresented languages due to segmentation effects, domain mismatch, and culturally specific realizations of abuse (Lauscher et al., 2020; Glavaš et al., 2021). This is directly relevant for Turkish–Azerbaijani: while linguistic proximity suggests transfer potential, sociolinguistic variation and culturally grounded insults can still limit generalization.

Within Turkic languages, Turkish has comparatively stronger resources, including publicly available hate speech datasets such as the Turkish Hate Speech Superset (Tonneau, 2022), and recent work benchmarks multilingual transformers for Turkish hate speech detection (Zehir and Koç, 2023). In contrast, Azerbaijani has lacked a publicly available annotated hate speech benchmark, constraining systematic evaluation and comparisons across transfer and augmentation strategies—a gap addressed by this paper's dataset and experiments. As an example, fine-tuning GPT-2 and RoBERTa embeddings, Alizada et al. (2024) increased sentiment classification by 7-10 percent,and Zeynalov (2022) trained GPT-2 on Azerbaijani Wikipedia, but reported sparsity and token imbalance.

A complementary direction for low-resource adaptation is translation-based augmentation, where labeled data from a higher-resource language is translated to the target language to increase supervision. Prior work reports that translation or back-translation can help in some settings (Hu et al., 2020), but gains are inconsistent and depend on translation quality, the preservation of abusive pragmatics, and the handling of idioms and culturally specific expressions (Jiang et al., 2021). Because Turkish and Azerbaijani are closely related yet culturally distinct, translation-based augmentation is plausible but not guaranteed to be beneficial, motivating the controlled evaluation we include.

In summary, prior work motivates three design choices evaluated here: (i) benchmarking strong multilingual transformers against transparent linear baselines, (ii) testing zero-shot and mixed bilingual training for a related-language pair, and (iii) assessing whether machine-translated Turkish data can meaningfully supplement scarce Azerbaijani supervision.

3 Methodology

3.1 Data

We use (i) a high-resource Turkish hate speech dataset, (ii) a newly created Azerbaijani benchmark, and (iii) construct a translation-based synthetic Azerbaijani corpus for augmentation.

Turkish. We use the *Turkish Hate Speech Superset* (Tonneau, 2022), containing 41,423 labeled social media comments (13,498 hate; 27,837 non-hate), previously used for Turkish hate speech detection (e.g., Zehir and Koç, 2023).

Azerbaijani (manual). We create the first manually annotated Azerbaijani hate speech dataset of 1,112 YouTube comments collected from major Azerbaijani news channels. Labels are binary (HATE VS. NON-HATE), with substantial class imbalance (107 hate; 1,005 non-hate). Annotation was done by a single native Azerbaijani speaker with an academic background and prior experience working with Azerbaijani language data and social media text. Annotation guidelines were defined prior to labeling based on widely used definitions of hate speech in the literature. Hate speech was operationalized as content that explicitly attacks, dehumanizes, or incites hostility or violence against an individual or group based on protected characteristics such as ethnicity, nationality, religion, or gender. Content containing profanity, strong opinions, or political criticism without a clearly identifiable hateful target was labeled as non-hate.

Azerbaijani (translated from Turkish). To assess translation-based supervision, we translate the Turkish dataset into Azerbaijani with Google Translate (Google) and manually spot-check a subset to reduce obvious translation artifacts.

3.2 Models

We compare two transformer classifiers and a sparse linear baseline.

XLM-RoBERTa. A multilingual transformer pretrained on 100+ languages, including Turkish and Azerbaijani (Conneau et al., 2020).

BERT-Tiny. A compact transformer intended for efficiency-constrained settings (Turc et al., 2019).

TF–IDF + Logistic Regression. A standard bag-of-words baseline using word-level TF–IDF features (Salton and Buckley, 1988) and a logistic regression classifier.

3.3 Preprocessing and Tokenization

We apply light text normalization (Unicode normalization and punctuation standardization) and remove duplicated content/emojis when present. Emojis and non-standard symbols were removed to reduce sparsity and noise in an extremely low-resource setting and to focus the benchmark on lexical and contextual signals of hate speech rather than affective markers. While emojis can convey pragmatic or emotional cues in social media discourse, their removal was intended to improve consistency across samples and models, and to avoid overfitting to platform-specific signals.x For transformer models, we use the default tokenizers associated with each pretrained checkpoint (SentencePiece for XLM-R; WordPiece for BERT-based models) (Kudo and Richardson, 2018; Devlin et al., 2019).

3.4 Training and Evaluation

Transformer models are fine-tuned for binary classification using AdamW and class-weighted cross-entropy to mitigate label imbalance.

$$\mathcal{L} = -\sum_{i=1}^{N} \omega_{y_i} \log p(y_i \mid x_i)$$

Where ω_{y_i} compensates class imbalance.

Unless stated otherwise, we use fixed hyperparameters across experiments to support comparability: 10 epochs, batch size 64, and learning rates of 3×10^{-5} (BERT-Tiny) and 2×10^{-5} (XLM-R).

We report accuracy and macro-F1, and additionally analyze per-class precision/recall and confusion matrices to characterize minority-class behavior. Furthermore, Macro-F1 is used as the primary evaluation metric because hate speech constitutes a severe minority class in both languages, making accuracy misleading in the presence of class imbalance. Macro-F1 equally weights both classes and better reflects a model's ability to detect hate speech without being dominated by majority-class performance. We additionally report hate-class precision and recall to directly assess minority-class behavior.

3.5 Experimental Setups

We evaluate monolingual learning, cross-lingual transfer, mixed bilingual training, and translation-based augmentation. Let TR and AZ denote Turkish and Azerbaijani; "1K" denotes a 1,000-sample subset. Across all experiments, configurations were designed to explicitly test symmetry and asymmetry in cross-lingual transfer, data scarcity effects, and the trade-off between data efficiency and performance.

(1) Full-data monolingual and transfer. (i) TR→TR, (ii) TR→AZ (zero-shot transfer), (iii) AZ→AZ, and (iv) AZ→TR (zero-shot transfer). We explicitly evaluate AZ→TR transfer to assess whether low-resource datasets can export useful representations, rather than assuming one-way transfer from high-resource to low-resource languages.

(2) Low-resource controlled setting (1K). We create balanced 1,000-sample subsets per language and evaluate $\mathrm{TR}_{1K} \rightarrow \mathrm{TR}_{1K}$ and $\mathrm{AZ}_{1K} \rightarrow \mathrm{AZ}_{1K}$. For Turkish, balance is achieved by downsampling the majority class. For Azerbaijani, all available hate instances are retained and the non-hate class is downsampled accordingly, resulting in a maximally balanced subset under data constraints. This controlled setting isolates the effect of data scarcity from linguistic factors.

(3) Mixed bilingual training (1K+1K). We train on a merged 2,000-sample dataset (TR_{1K} + AZ_{1K}) and evaluate (i) on the mixed test set and (ii) separately on each language (TR_{1K} and AZ_{1K}) to assess whether bilingual fine-tuning benefits both languages. Full-dataset merging was intentionally avoided to prevent Turkish (41k samples) from dominating Azerbaijani, which would obscure low-resource effects.

(4) Translation-based augmentation. We train and evaluate on the translated Azerbaijani corpus to quantify the utility and limitations of machine-translated supervision. This setup evaluates whether synthetic Azerbaijani data can compensate for annotation scarcity, while explicitly acknowledging the risk of learning translation artefacts rather than natural Azerbaijani usage.

4 Results

In this section, performance results of XLM-RoBERTa and BERT-Tiny in five experimental settings, 10 experimental configurations are presented: (1) full-data monolingual baselines, (2) cross-lingual transfer, (3) low- resource monolingual scenarios, (4) mixed-language low- resource training, and (5) machine-translated augmentation.

The results are presented in terms of accuracy, macro- precision, macro-recall and macro-F1, with macro-F1 being highlighted because it is more resistant to class imbalance, which is a primary concern in hate speech detection where positive examples are by far under-represented. Moreover, we focus discussion on macro-F1 and hate recall; other metrics are provided for completeness.

Table 1 aggregate models performance across all experiments, and figures show confusion matrices, heat maps and comparative bar plots, indicating minority-class errors and transfer patterns.

In Table 1, Results are grouped by experimental setting (A–E). Boldface indicates the best macro-F1 score within each block. TR and AZ denote Turkish and Azerbaijani, respectively; “1K” indicates balanced 1,000-sample subsets.

4.1 High-Resource Monolingual Baselines

The monolingual full-data results provide upper-bound performance in each of the languages.

In Azerbaijani (XLM-RoBERTa) (Exp. 3), in-domain training and testing resulted in accuracy = 0.91, macro-F1 = 0.72, hate recall = 0.39, non-hate recall = 0.97. This baseline is good among a 221-sample test set, but hate recall is 48 percent lower than Turkish, a direct empirical demonstration of the data scarcity penalty in Azerbaijani. Figure 1 reveals that the misclassifications concentrate in the false-negative quadrant of hate speech, which suggests a systematic bias in favor of the majority-class predictions as opposed to random error.

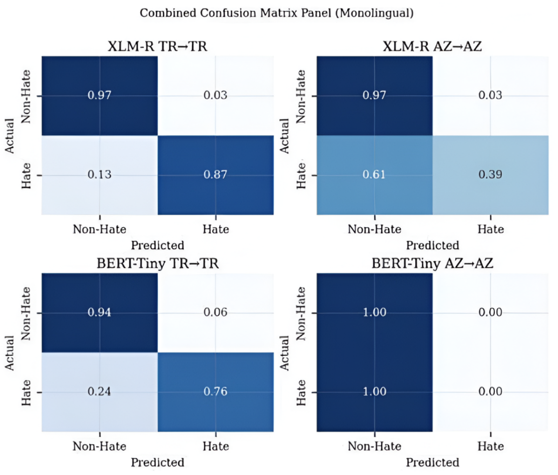

Figure 1: Confusion matrix panel

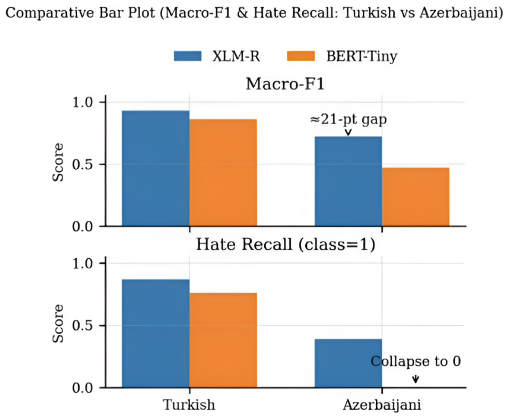

Figure 2: Comparative Bar Plot (Macro-F1 & Hate Recall: Turkish vs Azerbaijani)

The Turkish baseline (Exp.1) shows accuracy = 0.94, macro-F1 = 0.93, and balanced hate and non-hate recalls (0.87 and 0.97). The 21-point macro-F1 difference between Turkish and Azerbaijani demonstrates that high-resource languages are able to sustain more balanced decision boundaries.

BERT-Tiny reproduces this gap. On Azerbaijani it attained macro-F1 = 0.47 (Exp. 13) where hate recall collapsed to zero whereas on Turkish (Exp. 11) it attained macro-F1 = 0.86.

Figure 1 reveals the Azerbaijani-Azerbaijani confusion matrix of BERT-Tiny a near-total inability to recognize minority-class signals, replicating Davidson et al. (2017) who found shallow models reverting to majority-class predictions in the event of extreme imbalance.

Table 1: Models performance across experiments (indices in parentheses)

Setup	Model	Acc.	M-P	M-R	M-F1	Hate P
A. Full-data Monolingual Baselines						
TR → TR (1)	XLM-R	94.0	94.0	92.0	**93.0**	93.0
AZ → AZ (2)	XLM-R	91.0	79.0	68.0	72.0	64.0
TR → TR (13)	BERT-Tiny	86.0	85.0	50.0	86.0	86.0
AZ → AZ (13)	BERT-Tiny	90.0	45.0	50.0	47.0	0.0
B. Cross-Lingual Transfer						
TR → AZ (2)	XLM-R	81.0	81.0	64.0	**66.0**	64.0
AZ → TR (6)	XLM-R	68.0	66.0	55.0	46.0	65.0
TR → AZ (11)	BERT-Tiny	79.0	56.0	60.0	44.0	21.0
AZ → TR (14)	BERT-Tiny	67.0	30.0	50.0	40.0	0.0
C. Low-Resource Monolingual (1,000 samples)						
TR → TR (5)	XLM-R	81.0	80.0	75.0	**67.0**	70.0
AZ → AZ (4)	XLM-R	86.0	45.0	55.0	44.0	0.0
TR → TR (15)	BERT-Tiny	79.0	71.0	62.0	63.0	70.0
AZ → AZ (16)	BERT-Tiny	61.0	40.0	50.0	40.0	0.0
D. Mixed Low-Resource (1,000+1,000)						
TR → AZ → TR (8)	XLM-R	83.0	78.0	75.0	**76.0**	70.0
TR+AZ → TR (8)	XLM-R	95.0	95.0	95.0	95.0	95.0
TR+AZ → AZ (9)	XLM-R	96.0	79.0	89.0	84.0	96.0
TR+AZ → AZ (17)	BERT-Tiny	86.0	81.0	89.0	74.0	78.0
TR → AZ → TR (18)	BERT-Tiny	86.0	87.0	82.0	84.0	88.0
TR → AZ → AZ (19)	BERT-Tiny	91.0	95.0	55.0	53.0	100.0
E. Machine-Translated Augmentation						
TR,MT_AZ → TR,MT_AZ (10)	XLM-R	90.0	89.0	87.0	**88.0**	80.0
TR,MT_AZ → TR,MT_AZ (20)	BERT-Tiny	85.0	83.0	50.0	82.0	80.0

Overall, these patterns do not only indicate the lack of data but may reflect a focus on explicit hate: only explicit hate was recorded, and the models could not see more implicit hate, including sarcasm or coded slurs (Fortuna et al., 2020).

4.2 Cross-Lingual Transfer Performance

Evidence for Turkish → Azerbaijani transfer are found for XLM-RoBERTa (Exp. 2) that achieve macro-F1 = 0.69 with hate recall = 0.30. Transfer is also observed, although diminished, for BERT-Tiny (Exp. 12), with macro-F1 = 0.57, including hate precision = 0.21, which indicates a common false labelling of non-hate as hate. The pattern is observed also in figure 3.

Evidence for Azerbaijani → Turkish transfer are much weaker. XLM-RoBERTa (Exp. 4) shows macro-F1 = 0.44 and hate recall = 0.04, which is a 95% reduction in recall compared to Turkish monolingual training, while BERT-Tiny (Exp. 14) defaulted nearly to the non- hate predictions as well (hate recall = 0.00).

The asymmetry, which is represented graphically in figure 3, is dramatic: transfering from Turkish to Azerbaijani retains some discriminating power, whereas the opposite direction, from Azerbaijani to Turkish is disastrous. This trend reflects the one in Lauscher et al. (2020) and Glavaš et al. (2021), and indicates that the volume and variety of source data

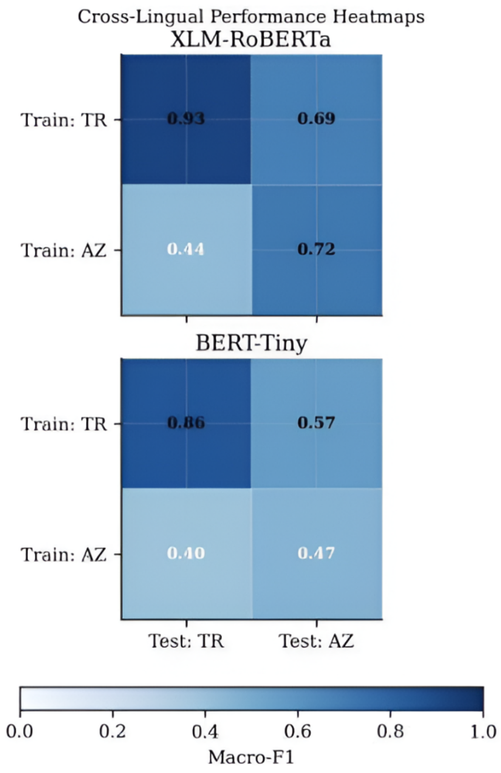

Figure 3: Cross-Lingual Performance of Macro-F1

are more significant than typological closeness in transferring to low-resource languages.

Beyond linguistic factors, domain and discourse differences likely contribute to the observed transfer asymmetry. The Turkish dataset covers a broader range of informal social media content, whereas the Azerbaijani dataset is dominated by news-related political commentary. As a result, cross-lingual transfer performance reflects not only syntactic or lexical proximity, but also mismatches in topic distribution, discourse style, and pragmatic conventions.

4.3 Restricted Low-Resource Monolingual Scenarios

Here, we examine model resilience when the size of data is restricted. To create resource-poor conditions, the two models were trained on a limited number of samples per language (1,000), and monolingually tested. Turkish-only (XLM-R, Exp. 5) had accuracy = 0.81, macro-F1 = 0.76 and a significant decrease in hate recall compared to full-data setting (0.58 vs. 0.87). This shows that there is a significant negative effect on minority class generalisation when training size is reduced by >95% even in high-resource languages.

Azerbaijani-only (XLM-R, Exp. 6) perform worse, with macro-F1 = 0.46 and 0 in hate-class recall. The model defaulted to non-hate even with class weighting, which shows a class imbalance vulnerability. This is indicative of a more destructive interaction between class imbalance and small datasets than with smaller Turkish subsets, highlighting structural disadvantages of truly low-resource languages.

The same trends were observed on the BERT-Tiny models, which had even lower macro-F1s (Exp. 15: 0.63 on Turkish, Exp. 16: 0.48 on Azerbaijani), which further validated the finding that small models exacerbate the low-data problem.

Overall, Azerbaijani results support the need of data augmentation or multilingual pretraining. In their absence, the decision surface of the model will fold to the majority class.

In several extreme low-resource Azerbaijani settings, models default to majority-class predictions despite class-weighted training. Rather than suppressing or correcting these outcomes, we report them explicitly, as they represent realistic failure modes in low-resource hate speech detection. Exposing such collapses is critical for understanding the limits of current methods and for avoiding overly optimistic conclusions based on accuracy alone.

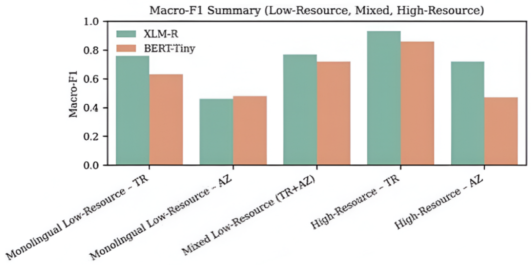

Figure 4: Macro-F1 Summary (Low Resource, Mixed, High- Resource)

4.4 Mixed-Language Low-Resource Training

To explore the idea whether the joint concatenation of small datasets of related languages can alleviate the low-resource problem, we trained both models on 1,000 Turkish + 1,000 Azerbaijani samples.

XLM-R (Exp. 7) shows macro-F1 of 0.77 on the combined-language test set, better than low-resource monolingual run. Applying the same model to Turkish-only (Exp. 8) gave 0.95 macro-F1, and to Azerbaijani-only (Exp. 9) 0.84 macro-F1. BERT-Tiny experienced comparable relative improvements (Exp. 17-19), yet were always 58 points behind XLM-R, indicating that although data mixing is beneficial to both models, the increased diversity is more useful to the larger ones.

Figure 4 summarizes Macro-F1 performance across low-resource monolingual, mixed-language, and high-resource training settings, and shows that XLM-RoBERTa consistently outperforms BERT-Tiny across all conditions, with the exception of the monolingual low-resource Azerbaijani setting.

These findings verify the fact that a little injection of related-language data enhances generalization. The enhancement is however not symmetrical as the Turkish outweighs the Azerbaijani since there is more morphological and pragmatic diversity in the Turkish hate data. This asymmetry indicates that Azerbaijani provides less unique signal, acting rather as a beneficiary of Turkish diversity than as an equal participant in transfer, which is also described in donor-receiver asymmetries in multilingual transfer by Pires et al. (2019).

4.5 Machine-Translated Augmentation

In experiment 10 and 20 we test whether translating the training set of Turkish to Azerbaijani would be able to contribute positively to a more comprehensive training data in Azerbaijani.

XLM-R (Exp. 10) reached 0.88 macro-F1, close to the result of the full Turkish monolingual baseline, and had equal precision and recall. BERT-Tiny (Exp. 20) obtained 0.82 macro-F1, which is 6 points lower.

Although the majority of lexical meaning is preserved because of the typological proximity of Turkish and Azerbaijani, syntactic and idiomatic artefacts could be confirmed by manual review of 500 samples - e.g., unnatural order of words and literal translation of culturally specific insults. This is corroborated by Jiang et al. (2021) who warn that the translation advantage plateaus when source and target languages are too close to each other, as models will tend to overfit on synthetic translation artefacts instead of learning the patterns of natural discourse.

Overall, the findings support the conclusion that Azerbaijani-only training utterly fails because of the lack of sufficient data, yet significant improvement can be provided through a transfer of Turkish → Azerbaijani, mixed low- resource training, and MT augmentation. The asymmetry of transfer (Turkish Azerbaijani vastly stronger than Azerbaijani Turkish) supports the importance of resource-rich source data in generalizing minority-class detection in morphologically rich and low-resource languages.

4.6 Classical Baseline (TF-IDF + Logistic Regression)

The sparse features on Turkish have a linear decision boundary and are competitive (macro-F1 0.85) and achieve reasonable hate recall (0.68). Nonetheless, on native Azerbaijani (macro-F1 = 0.47; hate recall = 0), performance fails as in the low-resource minority-class setting with BERT- Tiny. The results of training and testing on the translated Azerbaijani corpus provide high but overestimated scores (macro-F1 0.80) compared to native Azerbaijani, as expected due to translation smoothing artefacts as observed in our transformer experiments. Altogether, TF-IDF indicates that lexical overlap is not enough to detect Azerbaijani hate speech; multilingual pretraining in a contextual setting is needed to restore minority-class sensitivity.

Table 2: TF-IDF + LR Results

Dataset	**Acc**	**P**	**R**	**F1**	**HR**
Full AZ	0.90	0.45	0.50	**0.47**	**0.00**
Full TR	**0.88**	0.89	0.83	**0.85**	0.68
TR→AZ (trans.)	**0.84**	0.85	0.78	**0.80**	0.60

5 Discussion

This study demonstrates that linguistic relatedness may help but does not eliminate the need for target-language supervision. Turkish→Azerbaijani zero-shot transfer with XLM-R is substantially stronger than classical baselines, indicating that multilingual pretraining induces partially reusable representations across closely related languages (Conneau et al., 2020; Pires et al., 2019). However, transfer performance remains clearly below Turkish in-language results, and minority-class behavior in Azerbaijani is fragile, consistent with known limitations of zero-shot transfer for morphologically rich and underrepresented languages (Lauscher et al., 2020). This directly supports a data-centric takeaway: even between close language pairs, transfer is not a reliable substitute for curated target-language benchmarks. This conclusion aligns with recent low-resource work on polysemy emphasizing that model choice alone cannot compensate for gaps in target-language data, and calls for "Democratizing AI" by investing and curating language specific datasets (Goworek et al., 2025).

Bilingual mixed fine-tuning offers a pragmatic way to improve robustness under low-resource constraints. Training on balanced Turkish+Azerbaijani subsets improves performance compared to Azerbaijani-only training and stabilizes behavior relative to pure zero-shot transfer. The advantage of mixing high- and low-resource languages was also observed in English-Hindi (Dubossarsky and Dairkee, 2024). Additionally, this supports a view which is also suggested in cross-lingual work, that bilingual or pair-focused supervision can be more effective than relying on multilingual pretraining alone, especially when pragmatic conventions and culturally grounded insults differ despite lexical overlap (Glavaš et al., 2021). Practically, this means that modest target-language annotation, when combined with a related high-resource language, can deliver meaningful gains without requiring large Azerbaijani corpora.

The translation-based augmentation is a useful but limited substitute for native data. Translating Turkish training data into Azerbaijani increases the amount of labeled text and can improve performance over Azerbaijani-only baselines, but it also risks encoding translation artifacts and underrepresenting naturally occurring Azerbaijani discourse. This matches earlier findings that translation can help in low-resource toxicity detection, yet benefits are inconsistent and sensitive to how well pragmatic force and idiomatic expressions are preserved (Hu et al., 2020; Jiang et al., 2021). In our setting, the close relationship between Turkish and Azerbaijani makes translation plausible, but culturally specific and community-indexed hate expressions may still be poorly captured.

Our results connect to a broader point about domain specificity that extends beyond low-resource NLP (Xia et al., 2020; Toraman et al., 2022). Even in high-resource languages such as English, hate speech detection can be highly community- and context-dependent: surface-form cues may be misleading when language is used in-group (e.g., reclaimed language) or when terms shift meaning across communities. Recent work on reclaimed language shows that models and datasets that ignore this context can produce systematic errors, including elevated false positives against marginalized groups, and that reliable evaluation requires datasets and protocols tailored to the relevant communities and discourse settings (Zsisku et al., 2024). Taken together, these considerations motivate viewing dataset curation—domain coverage, community context, and label design—as a first-class component of hate speech detection research, rather than a secondary step after model selection.

6 Conclusion

This work is the first attempt to study the issue of hate speech detection in Azerbaijani with the help of the first annotated dataset of this language. The study provides an empirical baseline on Azerbaijani and Turkish by benchmarking compact and large transformer models in monolingual, cross-lingual, bilingual and machine-translated tasks.

These results indicate three important lessons. Model size is important: XLM-RoBERTa performs significantly better than BERT-Tiny in recovery of minority-class hate speech, indicating dangers of using compact models in sensitive moderation tasks. Second, cross-lingual transfer is non- symmetric: Turkish->Azerbaijani cross-lingual transfer is reasonably successful, whereas Azerbaijani->Turkish transfer fails, highlighting the importance of even modest native annotation. Third, translation is able to scale data but not nuance: whereas machine-translated corpora give high scores, they also induce artefacts that negatively affect cultural validity.

The practical implication is direct: a low-resource hate speech detector does not need to rely on brute-force scaling as much as it needs a thoughtful data design. When used with related-language data and multilingual models, small but genuine Azerbaijani annotations yield the greatest performance improvements where they are most needed: minimizing false negatives.

Furthermore, this work has threefold contributions:

1) The first publicly available Azerbaijani hate speech dataset;

2) Comparative benchmarks between transformer models in five experimental set-ups;

3) Practical suggestions on how to expand hate speech moderation to other under-represented languages.

Overall, this paper provides, through the open release of the Azerbaijani dataset, not only a basis on which future researchers can build, but also a scalable and culturally sensitive template of moderating in a low resource environment.

Ethics

All Azerbaijani comments were collected from public YouTube channels. We remove personally identifying information (e.g., usernames/links) prior analysis. Due to copyright issues, the released dataset is limited to `comment_id` and **binary hate-speech labels**, and does not include raw text or user metadata. However, it is sufficient for reproduction purposes.[1] The dataset contains hateful content; we restrict use to research and model evaluation and report error patterns to highlight potential bias and over-flagging risks.

Limitations

Our study establishes initial baselines for Azerbaijani hate speech detection, but it has several limitations.

First, the manually annotated Azerbaijani dataset is relatively small and drawn from a single domain (YouTube comments on news channels), which

[1] The dataset, necessary scripts and retrial dates will be released if the paper is accepted.

may limit generalization to other platforms, genres, and dialectal variation. Expanding it by increasing the number of instances and diversifying data sources beyond YouTube (e.g., other social media platforms or forums) would strengthen the the reliability of the results.

Another limitation of the current benchmark is that it adopts a binary label space (HATE vs. NON-HATE). While this setting is appropriate for establishing initial baselines and enabling straightforward comparison across models, it does not capture finer-grained distinctions that matter for analysis and deployment. In particular, the dataset does not differentiate hate speech types (e.g., insults vs. threats), target groups, or severity/intensity, which limit the applicability of the conclusions to deployment setup.

Third, the reliance on a single-annotator setup may introduce subjective bias, which we explicitly acknowledge. Nevertheless, the annotation criteria were applied consistently across the corpus to ensure internal coherence of labels.

Lastly, while we evaluate Turkish→Azerbaijani transfer and translation-based augmentation, we do not assess other Turkic transfer beyond Turkish. Thus we may have missed broader transfer patterns within this language group, and limit the scope of our conclusions.

Acknowledgments

H.D. is supported by the EPSRC AI Hub on Mathematical Foundations of Intelligence: An “Erlangen Programme” for AI No. EP/Y028872/1.

References

Tural Alizada, Umid Suleymanov, and Zaid Rustamov. 2024. Contextualized Word Embeddings in Azerbaijani Language. In *Proceedings of the 18th IEEE International Conference on Application of Information and Communication Technologies*, pages 1–6, Turin, Italy.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International World Wide Web Conference Companion*, pages 759–760.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-Lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

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*, pages 4171–4186.

Haim Dubossarsky and Farheen Dairkee. 2024. Strengthening the WiC: New polysemy dataset in Hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349, Torino, Italia. ELRA and ICCL.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2020. Toxic, hateful, and offensive language detection in online communication: A survey. *ACM Computing Surveys*, 53(6):122:1–122:36.

Goran Glavaš, Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Roi Reichart, and Simone Paolo Ponzetto. 2021. When is bilingual training better than multilingual? understanding Cross-Lingual Transfer in mBERT. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 6503–6517.

Google. Google translate API. https://cloud.google.com/translate (Accessed: 17 August 2025).

Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

Junjie Hu and 1 others. 2020. Back-Translation for Low-Resource Toxic Language Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1–6.

Ziang Jiang and 1 others. 2021. On Translation-Based Augmentation for Closely Related Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1245.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4483–4499.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Complex & Intelligent Systems*, 6:1–15.

Fabian Pedregosa and 1 others. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual Is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT. *arXiv*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Alexandre Tonneau. 2022. Turkish Hate Speech Superset. https://huggingface.co/datasets/atonneau/turkish_hate_speech.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

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.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Cem Zehir and Ahmet Koç. 2023. Hate speech detection in Turkish with multilingual transformers. *arXiv*.

Nijat Zeynalov. 2022. Training GPT-2 for Azerbaijani on Wikipedia: Vocabulary sparsity and token imbalance. *arXiv*.

Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, page 241–249, New York, NY, USA. Association for Computing Machinery.