

M. Sakypbekova<sup>1</sup>, R. Aubakirova<sup>1\*</sup><sup>1</sup>Al-Farabi Kazakh National University, Almaty, Kazakhstan\*e-mail: [aubakirova.ramilya@gmail.com](mailto:aubakirova.ramilya@gmail.com)

## CORPUS AND BASELINE MODELS FOR KAZAKH TOXIC COMMENT DETECTION

*Abstract*

Public discourse in Kazakhstan is moving online and into the Kazakh language, yet resources for detecting toxic comments in that language remain scarce. The Kazakh corpora that exist were assembled from lists of hostile keywords within an extremism or model-safety frame, few of them are openly released, and fewer still document how they were annotated. This paper presents Kaz-ToCo, a corpus of 84,940 comments in Kazakh and in mixed Kazakh-Russian code-switched text. The comments were retrieved in May 2026 through the YouTube Data API v3 from beneath 30 Kazakhstani talk-show videos, with no keyword-based filtering at any stage of collection, and were labelled on a binary toxicity scale by an open-weight large language model deployed on local hardware. Agreement between the machine labels and the annotations of a native speaker of Kazakh, measured on a stratified sample of 100 comments, reached a Cohen kappa coefficient of 0.602, with a 95 percent confidence interval from 0.446 to 0.758. The training, validation and test splits were built at the level of the source video, which prevents comments from one video from appearing in more than one split. Two baseline classifiers are established on the corpus: term frequency and inverse document frequency features over word and character n-grams combined with logistic regression, and a fine-tuned Kazakh conversational encoder of the RoBERTa architecture. Macro-averaged F1 on the held-out test set reached 0.8086 and 0.8198 respectively. These figures fall well below the near-ceiling accuracies published on keyword-collected Kazakh corpora, and are offered as a realistic reference point. The labels, per-comment metadata, splitting protocol, random seeds and hyperparameters are released without restriction. The comment texts cannot be redistributed under the YouTube API Services Developer Policies and are available from the corresponding author upon reasonable request for non-commercial research use.

**Keywords:** Kazakh language; toxic comment detection; low-resource language processing; corpus construction; language model annotation; code-switching; text classification; encoder fine-tuning.

М. Сақыпбекова<sup>1</sup>, Р. Аубакирова<sup>1\*</sup><sup>1</sup>әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

## ҚАЗАҚ ТІЛІНДЕГІ УЫТТЫ ПІКІРЛЕРДІ АНЫҚТАУДЫҢ КОРПУСЫ МЕН БАЗАЛЫҚ МОДЕЛЬДЕРІ

*Аңдатпа*

Қазақстандағы жария коммуникация барған сайын интернетке ауысып, қазақ тіліне көшуде, алайда осы тілдегі уытты пікірлерді автоматты түрде анықтауға арналған ресурстар әлі де аз. Қолда бар қазақ тіліндегі корпустар экстремистік әрі деструктивті мазмұнды іздеп, түйінді сөздер тізімі арқылы жиналған, олардың аз бөлігі ғана ашық жарияланған, ал белгілеу рәсімі одан да сирек құжатталған. Мақалада Kaz-ToCo, яғни қазақ тілінде және аралас қазақ-орыс кодында жазылған 84 940 пікірден тұратын корпус ұсынылады. Пікірлер 2026 жылдың мамыр айында YouTube Data API v3 интерфейсі арқылы 30 қазақстандық ток-шоу бейнежазбаларының пікірлер бөлімінен, түйінді сөздер бойынша сүзгіден өткізілместен жиналды. Оларды екілік уыттылық шкаласымен жергілікті жабдыққа орнатылған ашық салмақты тілдік модель белгіледі. Стратификацияланған 100 пікірден тұратын іріктемеде машиналық белгілеу қазақ тілін ана тілі ретінде білетін маманның бағасымен үйлесті. Коэн каппа коэффициенті 0,602 құрады, 95 пайыздық сенімділік аралығы 0,446-дан 0,758-ге дейін. Оқыту, валидация және тест іріктемелері бейне деңгейінде құрылды, сондықтан бір бейненің пікірлері әртүрлі іріктемеге түспейді. Корпус үшін екі базалық классификатор құрылды: сөздік және таңбалық n-граммаларға негізделген терминдер жиілігі мен құжаттардың кері жиілігі белгілері логистикалық регрессиямен және қосымша оқытылған қазақ тіліндегі RoBERTa сәулетіндегі энкодер. Кейінге қалдырылған тест іріктемесіндегі макроорташаланған F1 өлшемі сәйкесінше 0,8086 және 0,8198 болды.

Алынған мәндер түйінді сөздер бойынша жиналған корпустарда жарияланатын шекке жуық көрсеткіштерден айтарлықтай төмен және оптимистік емес, шынайы бастапқы нүкте ретінде ұсынылады. Белгілер, пікір деңгейіндегі метадеректер, іріктемелерге бөлу хаттамасы, кездейсоқ сандар генераторының бастапқы мәндері мен гиперпараметрлер шектеусіз жарияланады. Пікірлердің мәтіндерін YouTube API Services Developer Policies ережелеріне сәйкес қайта таратуға болмайды, олар коммерциялық емес ғылыми мақсатта хат-хабар авторынан негізді сұраныс бойынша алынады.

**Түйін сөздер:** қазақ тілі; ұйғтты пікірлерді анықтау; ресурсы аз тілдерді өңдеу; корпус құру; тілдік модельмен белгілеу; код ауыстыру; мәтіндерді жіктеу; энкодерді қосымша оқыту.

М. Сақыпбекова<sup>1</sup>, Р. Аубакирова<sup>1\*</sup>

<sup>1</sup>Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

## **КОРПУС И БАЗОВЫЕ МОДЕЛИ ВЫЯВЛЕНИЯ ТОКСИЧНЫХ КАЗАХОЯЗЫЧНЫХ КОММЕНТАРИЕВ**

### *Аннотация*

Публичная коммуникация в Казахстане всё активнее уходит в интернет и переходит на казахский язык, однако ресурсов для автоматического выявления токсичных комментариев на этом языке по-прежнему мало. Имеющиеся казахоязычные корпуса собраны по спискам враждебных ключевых слов, в логике поиска экстремистского и деструктивного контента, лишь немногие из них открыты, а процедура разметки документирована ещё реже. В статье представлен Kaz-ToCo, корпус из 84 940 комментариев на казахском языке и на смешанном казахско-русском коде. Комментарии получены в мае 2026 года через интерфейс YouTube Data API v3 под 30 казахстанскими ток-шоу, без фильтрации по ключевым словам. Разметку по бинарной шкале токсичности выполнила языковая модель с открытыми весами, развёрнутая на локальном оборудовании. На стратифицированной выборке из 100 комментариев машинная разметка согласуется с разметкой носителя казахского языка на уровне 0,602 по коэффициенту каппа Коэна, при 95-процентном доверительном интервале от 0,446 до 0,758. Обучающая, валидационная и тестовая выборки построены на уровне исходного видео, поэтому комментарии одного видео не попадают в разные выборки. Для корпуса построены два базовых классификатора: признаки на основе частоты терминов и обратной частоты документов по словным и символьным n-граммам с логистической регрессией, а также дообученный казахоязычный энкодер архитектуры RoBERTa. Макроусреднённая мера F1 на отложенной тестовой выборке достигла 0,8086 и 0,8198 соответственно. Полученные значения заметно ниже околопредельных показателей, публикуемых на корпусах, собранных по ключевым словам, и предлагаются как реалистичная, а не оптимистичная точка отсчёта. Метки, покомментарийные метаданные, протокол разбиения на выборки, начальные значения генератора случайных чисел и гиперпараметры публикуются без ограничений. Сами тексты комментариев повторно распространять нельзя по правилам YouTube API Services Developer Policies, их предоставляет автор для корреспонденции по обоснованному запросу для некоммерческого научного использования.

**Ключевые слова:** казахский язык; выявление токсичных комментариев; обработка малоресурсных языков; построение корпуса; разметка языковой моделью; переключение кодов; классификация текстов; дообучение энкодера.

### **Introduction**

Kazakh-language research on harmful content has been dominated by keyword-driven collection within an extremism or model-safety frame, which has left the field without a general-purpose corpus of naturally occurring toxic comments. Kaz-ToCo fills that gap with 84,940 comments retrieved exhaustively through the YouTube Data API v3 from beneath 30 talk-show videos, with no content-based selection at the collection stage, and covering both monolingual Kazakh and Kazakh-Russian code-switching. Labelling by a locally deployed open-weight language model agreed with a native speaker at a Cohen kappa of 0.602, a moderate value that matches the cautious picture reported for toxic content in low-resource settings rather than the optimistic one reported for English-language social-science coding. Under a video-level group split, a linear model over word and character n-grams reaches a macro-averaged F1 of 0.8086 and a fine-tuned Kazakh encoder reaches 0.8198, a gap of 0.0112 that no significance test supports. The labels, the per-comment metadata, the splitting

protocol and the full training configuration are released without restriction. The comment texts, which the YouTube API Services Developer Policies do not permit to be redistributed, are available from the corresponding author upon reasonable request for non-commercial research use.

A nationwide survey of 21,000 households carried out in 2025 found that 93.4% of Kazakhstan's population aged six and older uses the internet, and that 69.5% of those users named participation in social networks among their online activities [1]. Public communication is at the same time shifting toward Kazakh. The government Commission for the Implementation of State Language Policy reported that as of June 2026 the share of Kazakh-language content in the media had reached 78%, and that roughly 60% of internet resources now operate in the state language [2]. What aggressive speech does in that space is already on record. The international Health Behaviour in School-aged Children survey found that 13.4% of Kazakhstani adolescents aged 11 to 15 had been subjected to cyberbullying [3]. A large share of the country's public conversation now happens online and in Kazakh, and official statistics already register the harm.

Automatic toxicity detection depends on labelled data in the target language, and labelled data is exactly what Kazakh lacks. Research on harmful content in Kazakh dates to the mid-2010s, but it settled early into an information-security frame. The target phenomenon gets operationalized as extremism, destructive content or cyberbullying. Corpora are assembled from lists of hostile keywords. Annotation procedures often go undocumented, and open releases remain the exception [4, 5, 6].

Three limitations follow from that framing. First, corpus construction in nearly every Kazakh-language study begins with a keyword list and automatic platform scraping [5, 6, 7]. The design produces lexical bias and topic leakage. Accuracies occasionally reported near the ceiling, up to 98% for religious-extremism detection [5], are more plausibly explained by that leakage than by the task actually being solved, a reading consistent with the broader finding that offensive-speech classifiers degrade sharply once evaluation conditions become realistic [8]. Second, annotation protocols are sparsely documented. One clear exception exists: in [4], five trained native speakers worked from written guidelines and reached a Cohen kappa of 0.82. Most of the remaining Kazakh toxicity work reports neither annotator instructions nor any agreement figure. Third, open releases are rare. Three are available. The first is a corpus of roughly three thousand religious extremist texts pulled from a social network by keyword search [9]. The second is the 10,190-text destructive-content dataset published as supplemental material with [4]. The third is a 17,827-sample profanity dataset built to align language models, which pairs every toxic input with a de-escalating model response and so serves response generation rather than comment-level classification [10]. All three were keyword-collected or safety-framed. None contains naturally occurring platform comments gathered without content-based selection, and none implements the graded, multi-facet taxonomy covering profanity, insult, threat and identity attack, nor the severity gradation whose feasibility has been demonstrated for low-resource languages [11]. The open-access practice that KazSAnDRA established for Kazakh sentiment analysis [12] has no counterpart in toxicity.

A second and methodologically stronger strand has grown up around the safety evaluation of large language models in Kazakh. The Qorgau benchmark assessed twelve models in Kazakh and Russian across more than ninety thousand responses and, in a step that matters here, validated its automated judge against human annotations of a thousand examples per language before scaling up, recording high agreement between machine and human in both languages [13]. KZ-SafetyPrompts extends the same logic to prompt collection: 5,717 natively Kazakh prompts across eleven risk categories, with a documented compilation protocol, decision rules for edge cases and deduplication procedures [14]. That a safety benchmark, rather than a toxicity corpus, offers one of the most detailed published accounts of a harmful-content annotation procedure for Kazakh says a good deal about the state of the field.

Both strands converge on a design question any new Kazakh toxicity resource has to answer: who, or what, does the labelling. Toxicity annotation is subjective. An annotator's beliefs and identity systematically shift the labels produced, and low agreement often reflects a plurality of defensible

readings rather than error [15]. Against that background, model-assisted annotation has emerged as a way to cut cost and annotator burden. The headline finding is that a zero-shot commercial chat model outperformed trained crowdworkers on both accuracy and inter-coder agreement, at roughly one-thirtieth of the cost [16]. Evidence specific to toxic content is more guarded: the same model reaches around 80% agreement with human reference labels on non-harmful content and identifies toxicity reliably at a coarse level, yet it struggles to separate hateful, offensive and toxic subcategories, and the outcome is sensitive to prompt wording [17]. Whether any of this survives the move to a low-resource language is unsettled. A direct evaluation for Marathi found a substantial quality drop and concluded that model output cannot stand in for human annotation [18], while a weak-supervision pipeline for Kazakh sequence labelling produced strong taggers from machine annotations that were confidence-filtered and backed by a small, manually verified subset [19]. One further caveat applies to the whole quantitative base: it rests on proprietary models reached through a vendor interface. Whether the findings carry over to open-weight models running on local hardware has not been tested separately.

This study was conducted to close the first of those gaps and to put evidence under the second and the third. Its aim is to build a general-purpose corpus of Kazakh-language toxic comments collected without keyword-driven selection, to document every step of the procedure, to release everything that platform policy permits to be released, and to establish reproducible baselines on the result. Three questions were put to the data. Can an open-weight language model running on local hardware label Kazakh toxicity at a level of agreement with a native speaker that makes the labels usable? Does a corpus collected without keyword filtering and evaluated under a video-level group split reproduce the near-ceiling accuracies reported elsewhere in the Kazakh literature, or noticeably lower ones? And does a Kazakh-specific pretrained encoder beat a sparse n-gram baseline by a margin that matters in practice?

## **Research methodology**

### *Data source and collection.*

The corpus is named Kaz-ToCo after Bangla-ToCo, a corpus of toxic comments in Bengali [20]. Its source material is Kazakhstani talk shows hosted on the YouTube video platform. Talk shows were chosen because the comments beneath them capture spontaneous, unmoderated reactions to whatever the panel was arguing about, rather than text elicited by a search for harmful content. This is the point at which the present collection design parts company with the keyword-built Kazakh corpora reviewed above.

Comments were retrieved in May 2026 through the YouTube Data API v3. The `commentThreads.list` method was called for each video with the `videoId` parameter, `textFormat` set to plain text and `maxResults` set to the permitted maximum of 100, and the `nextPageToken` field was followed until the API returned no further page. Top-level comments and their replies were both taken from the returned threads. Thirty videos were processed. Collection was exhaustive at the video level: every comment thread the API returned for a video was retained, with no content-based filtering at the collection stage, so the corpus's size and topical composition were set by the discussions themselves rather than by a per-class or per-topic quota. Raw collection yielded 152,244 comments.

### *Normalization and filtering.*

Comment text was normalized without altering lexical content. Line breaks and non-breaking spaces (U+00A0) were replaced with ordinary spaces. Invisible characters were removed: the zero-width space (U+200B), the word joiner (U+2060) and the byte order mark (U+FEFF). The zero-width joiner (U+200D) was kept, since removing it would break composite emoji. Runs of spaces were collapsed, and leading and trailing spaces were stripped.

Platform artefacts were not deleted but replaced with placeholder tokens, which preserves positional information: URLs became [URL], user mentions became [USER], and timecodes became [TIMECODE].

Filtering ahead of annotation proceeded in three steps. The collection was deduplicated. Language identification was then applied with the fastText classifier, and comments identified as purely Russian were dropped, while monolingual Kazakh comments and Kazakh-Russian code-switched comments were kept. Finally, entries containing no letters or digits at all, comments consisting only of emoji or punctuation for instance, were removed. These three steps reduced the collection from 152,244 raw comments to the 84,940 that make up the final corpus.

Retaining code-switched material was deliberate. Mixing Kazakh and Russian is characteristic of Kazakhstani platform discourse, and until now only the Qorgau safety benchmark had explicitly addressed a bilingual setting [13]. No dedicated resource existed for toxic code-switched comments.

#### *Data labeling.*

Comments were labelled into two classes, toxic and non-toxic, by the open-weight model cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit, served through the vLLM inference engine, version 0.21.0, and deployed locally on an ASUS DGX Spark server, whose configuration is detailed under Training procedure below. Local deployment of an open-weight model was chosen for three reasons. User comments are never sent to a third-party service. The weight version and the runtime configuration are fixed and remain available for re-running the process. And annotation cost does not scale with corpus size. Published quantitative assessments of harmful-content annotation, however, come from proprietary models [13, 16, 17], so their conclusions cannot simply be carried over to the model used here. Agreement with human annotation was therefore measured separately.

A comment was assigned to the toxic class (label 1) when it contained profanity, insults, aggression, incitement to hatred or degradation. The neutral class (label 0) covered neutral or positive text, questions and constructive criticism. Judgement was made on the meaning of the Kazakh-language utterance, taking context into account. These criteria summarize the system prompt, which was written in Russian, the working language of the bilingual environment the corpus comes from. In English translation the prompt reads:

You are an expert content moderator specializing in the Kazakh language. Classify user comments for toxicity. Rules: profanity, insults, aggression, hatred, degradation give label 1; neutral or positive text, a question, constructive criticism give label 0; judge the meaning in the Kazakh language, taking context into account; respond only with a JSON object containing a single field label, whose value is 0 or 1.

Each request carried exactly one comment. The model saw nothing else: no video title, no like counts or other metadata, no parent comment and no other comments from the thread. The word context in the instruction therefore refers to the internal context of the utterance itself. Comments whose text was empty after normalization were assigned label 0 without a model call.

Generation ran at a temperature of 0.0, capped at 64 output tokens, in structured JSON output mode. The model's reasoning mode was disabled. Response validity and typing were checked against a Pydantic schema, and requests were retried on any format violation.

#### *Validation of the machine labels.*

A hundred comments were selected at random under stratification by machine label, preserving the corpus's class proportions, and were labelled independently by a native speaker of Kazakh. Agreement between the two label sources was measured with the Cohen kappa coefficient, reported with an asymptotic 95% confidence interval and checked against a nonparametric bootstrap over 10,000 resamples. The procedure follows the validate-first-then-scale architecture used in Kazakh safety evaluation [13] and in weak supervision for Kazakh sequence labelling [19].

#### *Dataset organization.*

The corpus is a single comma-separated file with the fields video\_id, text, like\_count, is\_reply, is\_pinned, is\_favorited and label. The author\_id field, returned by the API at the collection stage, was excluded for privacy reasons.

The corpus is distributed in two parts, for a reason that is worth stating plainly. The YouTube API Services Developer Policies do not permit the redistribution of comment texts retrieved through the API, so the text column cannot be posted in a public repository. Everything else can be, and is: the

labels, the per-comment metadata, the video-level split assignment, the annotation and training code, the random seeds and the hyperparameters are released without restriction, keyed to the same row order, so that a reader who obtains the texts can rebuild the corpus exactly. The comment texts are available from the corresponding author upon reasonable request for non-commercial research use.

#### *Splitting protocol.*

Splitting was carried out at the video level, grouped by video\_id, so that comments from one video cannot land in more than one split. Comments under a single video share a topic, participants and conversational context, so a random split at the level of the individual comment would let a model lean on video-specific cues that are unavailable under deployment conditions.

To keep class balance comparable across splits, the toxic share was computed for each video, videos were binarized at the median into low-toxicity and high-toxicity strata, and a stratified split was applied at the video level. Splitting proceeded in two stages, 85/15 and then 82.4/17.6 within the larger portion, which yields roughly a 70/15/15 split by video count at a random seed of 42. The result is 20 videos in the training set and five each in validation and test.

#### *Classification models.*

Two models were evaluated. The first pairs sparse lexical features with a linear classifier. The representation concatenates two matrices of term frequency and inverse document frequency weights. The word-level matrix uses n-grams of order 1 to 2, a vocabulary capped at 100,000 features, a minimum document frequency of 2, sublinear term-frequency scaling and lowercasing. The character-level matrix uses character n-grams of order 2 to 5 computed within word boundaries, a cap of 200,000 features, a minimum document frequency of 2 and sublinear scaling. Character features were included because of Kazakh's rich morphology: a single lexeme surfaces as many word forms, and word-level features fragment across them. The resulting representation has 287,946 dimensions, 100,000 of them word-level and 187,946 character-level. The classifier is logistic regression with the liblinear solver, balanced class weights, a cap of 2,000 iterations and a fixed random seed of 42. Regularization strength was tuned by grid search over the values 0.1, 0.5, 1.0, 3.0 and 10.0 for the inverse regularization parameter, selected by macro-averaged F1 on the validation set. The value 1.0 was chosen.

The second model is the Kazakh conversational encoder kz-transformers/kaz-roberta-conversational, a model of the RoBERTa architecture of roughly 334 MB, fitted with a freshly initialized classification head over two labels. Dropout was raised to 0.25 for both the hidden and the attention layers. Input texts were tokenized with the model's own tokenizer at a maximum length of 128 tokens, padded to that length and truncated beyond it.

#### *Training procedure.*

Fine-tuning followed a two-stage freeze-then-unfreeze schedule. In the first stage the encoder was frozen for one epoch while only the classification head trained, at a learning rate of 1e-4 with no warmup. In the second stage all parameters were unfrozen and training continued for up to three epochs at a learning rate of 1e-5, with a linear scheduler and warmup over 10% of the steps.

Optimization used the AdamW algorithm with separate weight decay, 0.01 for weight matrices and 0 for biases and layer-normalization parameters. Batch size was 16 with gradient accumulation over two steps, giving an effective batch size of 32. Training ran in mixed precision with gradient clipping at a global norm of 1.0. The loss was cross-entropy with label smoothing of 0.1. Model selection used macro-averaged F1 on the validation set, with a checkpoint saved after every epoch and early stopping at a patience of two. The random seed was fixed at 42 across the random, numpy and torch generators, including the CUDA generator.

For reproducibility, the software and hardware environment is stated here. Annotation, training and inference all ran on the same machine: an ASUS DGX Spark server built on the NVIDIA GB10 Grace Blackwell superchip, which pairs a 20-core ARM64 (aarch64) CPU with an integrated Blackwell-generation GPU over 128 GB of unified memory, under Ubuntu 24.04 LTS. Data processing and the sparse baseline used numpy 2.4.6, pandas 3.0.5, scipy 1.17.1, scikit-learn 1.9.0 and joblib 1.5.3; encoder fine-tuning used torch 2.13.0 and transformers 5.16.1; the annotation model

was served through the vLLM inference engine, version 0.21.0. The exact package list is published together with the training code.

#### *Evaluation metrics.*

Models were evaluated on the held-out test set with four metrics: macro-averaged F1, micro-averaged F1, weighted F1, and the area under the receiver operating characteristic curve. Macro-averaged F1 is the primary metric and drives model selection, because it averages F1 across classes without weighting by class size and so treats the toxic and the non-toxic class equally. Micro-averaged and weighted F1 are reported for comparability with prior work that uses them, and the area under the curve captures ranking quality independently of the decision threshold. No significance test was run on the difference between the two models, a point the discussion returns to.

## **Results**

### *Characteristics of the corpus.*

Table 1 summarizes the corpus. Kaz-ToCo contains 84,940 comments from 30 videos in two classes, with a vocabulary of 136,977 distinct tokens. Comments are short: mean length is 14.38 words, the median is 10, and the range runs from 1 to 96 words. Engagement metadata is heavily skewed, with a mean like count of 5.17 against a median of 0. Replies, rather than top-level comments, account for 17.56% of the corpus.

*Table 1. Overall statistics of the Kaz-ToCo corpus*

| <i>Metric</i>                    | <i>Value</i> |
|----------------------------------|--------------|
| <i>Total comments</i>            | 84,940       |
| <i>Unique videos</i>             | 30           |
| <i>Number of classes</i>         | 2            |
| <i>Vocabulary size</i>           | 136,977      |
| <i>Mean words per comment</i>    | 14.38        |
| <i>Median words per comment</i>  | 10.0         |
| <i>Minimum words per comment</i> | 1.0          |
| <i>Maximum words per comment</i> | 96.0         |
| <i>Mean like count</i>           | 5.17         |
| <i>Median like count</i>         | 0.0          |
| <i>Reply share, %</i>            | 17.56        |

The class breakdown in Table 2 shows a corpus close to balanced: 47,614 non-toxic comments, or 56.06%, against 37,326 toxic ones, or 43.94%. The classes differ systematically along every dimension recorded. Toxic comments run longer, 16.08 words on average against 12.45 for non-toxic ones, and 12 against 9 by median. They draw fewer likes, 4.18 against 5.94 on average. And they are less often replies: the reply share among toxic comments is 14.37% against 20.06% for non-toxic ones.

*Table 2. Statistics by class, where label 0 denotes the non-toxic class and label 1 the toxic class*

| <i>Label</i> | <i>Count</i> | <i>Share, %</i> | <i>Mean words</i> | <i>Median words</i> | <i>Mean likes</i> | <i>Reply share, %</i> |
|--------------|--------------|-----------------|-------------------|---------------------|-------------------|-----------------------|
| 0            | 47,614       | 56.06           | 12.45             | 9.0                 | 5.94              | 20.06                 |
| 1            | 37,326       | 43.94           | 16.08             | 12.0                | 4.18              | 14.37                 |

The class difference in length shows up across the whole distribution, not just in the tail. Figure 1 shows that non-toxic comments cluster in a sharp mode around 50 characters, while the toxic distribution has a lower, flatter mode, is shifted to the right, and carries more mass beyond 150 characters.

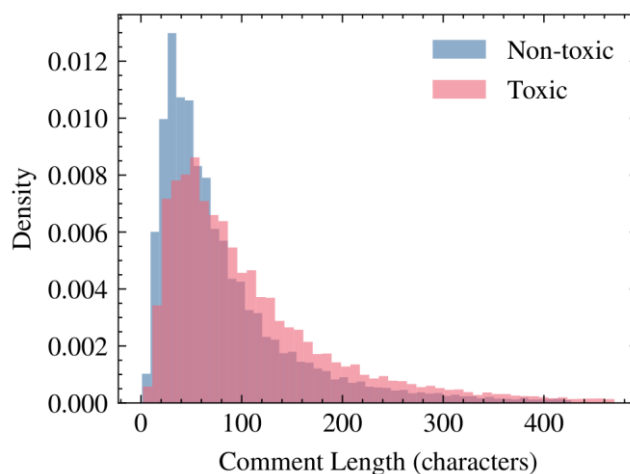


Figure 1. Density distribution of comment length in characters for the non-toxic and the toxic class

#### Composition of the splits.

Because videos vary in comment count, the per-comment proportions do not mirror the video-level split: the training set holds 62,152 comments, validation 7,780 and test 15,008. The class distribution across the three splits is shown in Figure 2. The toxic share is 43.9% in training, 44.3% in validation and 43.9% in test. Stratifying by each video's internal toxicity share keeps the balance nearly constant despite the grouping constraint, and none of the three splits is heavily imbalanced.

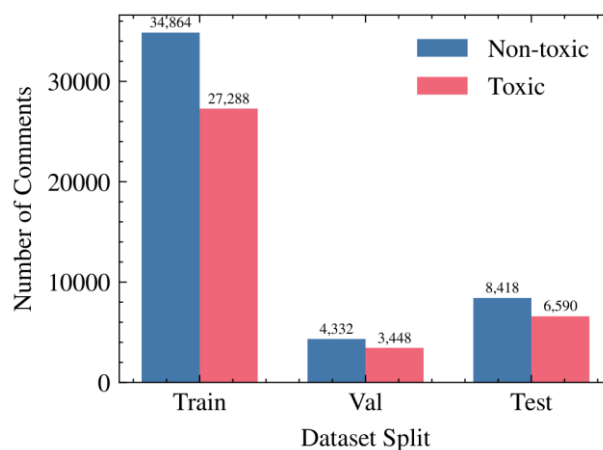


Figure 2. Class distribution across the training, validation and test splits, with counts labelled above the bars

Comment-length distributions were compared across the three splits to confirm that the video-level split had not introduced a systematic difference in text length. The distributions in Figure 3 match closely in mode, spread and right tail.

#### Agreement between machine and human labels.

On the stratified sample of 100 comments, 44 of which carried a machine toxic label, agreement between the model's labels and the native speaker's reached a Cohen kappa of 0.602, with a 95% confidence interval of 0.446 to 0.758 and an asymptotic standard error of 0.080. The nonparametric bootstrap over 10,000 resamples gives a near-identical interval of 0.443 to 0.755.

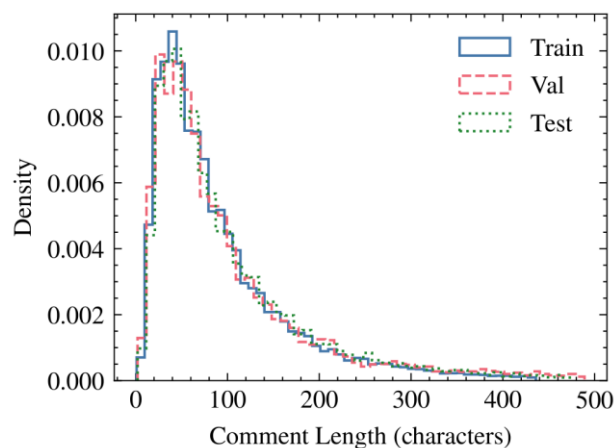


Figure 3. Density distribution of comment length in characters for the training, validation and test splits

Table 3 breaks the same 100 comments down as a contingency table. The two label sources agree on 80 comments. The 20 disagreements are asymmetric: 14 comments the human marked toxic went unflagged by the model, against 6 flagged in the opposite direction. On this sample the model assigns the toxic label less often than the native speaker does, 44 times against 52. The direction of that asymmetry bears on how the corpus's class balance should be read. A toxic share of 43.94% is high next to toxicity corpora for other languages, but the disagreement pattern indicates that the figure is not produced by the model over-assigning the toxic class. If anything, the human reference would have put it higher. The more plausible sources of the high share are the material itself, since talk-show discussions of contested social topics draw heated reactions, and the breadth of the toxic definition in the prompt, which folds aggression and degradation in with profanity and insults.

Table 3. Contingency table of human and model labels on the 100-comment validation sample

| Human label      | Model: non-toxic | Model: toxic | Total |
|------------------|------------------|--------------|-------|
| Human: non-toxic | 42               | 6            | 48    |
| Human: toxic     | 14               | 38           | 52    |
| Total            | 56               | 44           | 100   |

#### Classification quality.

Table 4 reports both baseline models on the test set. The fine-tuned Kazakh encoder scores higher on all four metrics. For both models the three F1 variants diverge by no more than 0.004, which follows from the near-uniform class distribution shown in Table 2 and indicates that neither model has collapsed into predicting the majority class.

Table 4. Performance of the two baseline models on the held-out test set

| Model                           | Macro-F1 | Micro-F1 | Weighted F1 | ROC-AUC |
|---------------------------------|----------|----------|-------------|---------|
| kaz-roberta-conversational      | 0.8198   | 0.8233   | 0.8228      | 0.9063  |
| TF-IDF with logistic regression | 0.8086   | 0.8111   | 0.8113      | 0.8982  |

#### Composition of the errors.

Tables 5 and 6 present confusion matrices over the 15,008 test comments. Relative to the linear model, the transformer produces 290 fewer false positives at the cost of 107 additional missed toxic

comments. Its edge on the aggregate metrics therefore comes from a more cautious assignment to the toxic class rather than from uniformly better separation.

Table 5. Confusion matrix on the test set for term frequency and inverse document frequency features with logistic regression

| Actual class     | Predicted non-toxic | Predicted toxic |
|------------------|---------------------|-----------------|
| Actual non-toxic | 6,939               | 1,479           |
| Actual toxic     | 1,356               | 5,234           |

Table 6. Confusion matrix on the test set for the fine-tuned kaz-roberta-conversational encoder

| Actual class     | Predicted non-toxic | Predicted toxic |
|------------------|---------------------|-----------------|
| Actual non-toxic | 7,229               | 1,189           |
| Actual toxic     | 1,463               | 5,127           |

## Discussion

*The gap between the two baselines.*

The sparse representation with a linear classifier trails the pretrained Kazakh transformer, fine-tuned through a multi-stage schedule, by 0.0112 macro-averaged F1. No significance test was run, so this difference can be called neither established nor noise. What can be said is that it is descriptively small, and that it matches the result found for destructive-content classification in Kazakh, where n-gram features with classical classifiers achieved high accuracy while a multilingual encoder combined with a recurrent layer proved competitive but not decisively better [4], a picture corroborated by comparisons from the same group [5]. Two independently built Kazakh-language corpora converging on the same qualitative outcome is worth noting, because they differ exactly along the dimension that should have mattered: the corpus in [4] was assembled by keyword search within an extremism frame, whereas Kaz-ToCo was pulled from entire videos with no content filtering at the collection stage.

*Relation to the gaps identified in the literature.*

The corpus closes the gap it was built to close. Kaz-ToCo is a general-purpose Kazakh-language toxicity resource, and it differs from the Kazakh datasets that precede it [4, 9, 10] in collection design and in scale rather than in availability. Those resources were gathered by keyword search or built for model-safety alignment, and the largest comment-level one holds 10,190 texts against 84,940 here. On availability this corpus is in one respect more constrained than they are, and the constraint is worth being explicit about: retrieval through the YouTube Data API v3 brings the comment texts under the platform's developer policies, which forbid redistributing them, so the texts are shared on request while the labels, the metadata, the splits and the code are posted openly. The open-access practice set by KazSAnDRA in sentiment analysis [12] can therefore be followed only in part. At the same time the corpus inherits a limitation from the same literature: the taxonomy is binary. The graded, multi-facet scheme covering profanity, insult, threat and identity attack, along with the severity gradation whose feasibility has been shown for low-resource languages [11], is not implemented here.

On annotation transparency, this work reports the name and the value of the agreement metric and documents the normalization, filtering and selection steps in full. It does not yet meet the standard set by KZ-SafetyPrompts, whose documented compilation protocol and edge-case rules remain among the most detailed published accounts of a harmful-content annotation procedure for Kazakh [14].

On model-assisted annotation of Kazakh toxicity specifically, this work provides the first direct evidence, and that evidence is ambiguous in exactly the way the prior literature would predict. The Cohen kappa obtained, 0.602, sits noticeably below the human-machine agreement recorded for coarse Kazakh safety judgements by the validated judge in Qorgau [13], and it does not replicate the

crowdworker-beating result reported for zero-shot annotation in English-language social-science coding [16]. It sits closer to the more cautious picture found specifically for toxic content, where coarse detection is reliable but finer distinctions degrade and the outcome is sensitive to prompt wording [17], and it fits the failure mode documented for a low-resource setting in Marathi [18]. The reading that transfer is task-dependent, and that coarse binary safety judgements survive low-resource settings better than culturally loaded distinctions, accommodates this result. A single kappa on a single corpus does not settle the dispute between the competing positions. The comparison is imprecise in a further respect: the cited figures come from proprietary models, while the annotation here used an open-weight model, and the contribution of that difference cannot be isolated.

Two caveats apply to the kappa value itself. Toxicity annotation is inherently subjective, and disagreement often reflects a plurality of defensible readings rather than error [15], so a moderate coefficient does not by itself indict the machine labels. At the same time, no calibration study exists for Kazakh-Russian bilingual annotators against which the human reference itself could be checked.

As for how realistic the evaluation is, the video-level group split removes the most obvious source of inflation in this setup. The figures obtained here, macro-averaged F1 around 0.82, fall well short of the near-ceiling accuracies reported on keyword-collected Kazakh corpora, up to 98% [5]. That is exactly what one would expect if those figures partly reflect lexical bias and topic leakage rather than genuine task difficulty [8]. The comparison spans different corpora built under different annotation schemes, so it is suggestive rather than probative.

Two gaps flagged at the outset remain open. The corpus includes Kazakh-Russian code-switched comments, which at the level of the data addresses the absence of a dedicated bilingual resource beyond the safety benchmark in [13]. But no analysis separating quality on the monolingual subset from quality on the code-switched subset was carried out, so no claim is made here about how the models handle code-switching. Nor is the encoder-choice question settled, that is, whether Kazakh-specific pretraining beats multilingual transfer on the toxicity task, since only the Kazakh-language encoder was fine-tuned, with no multilingual counterpart trained under comparable conditions.

#### *Limitations.*

The domain is narrow. All 84,940 comments come from 30 talk-show videos on a single platform. Talk-show discourse covers a particular range of topics and participants, and transfer to news comments, to other platforms, or to other Kazakh-language communities was not tested.

Labels are predominantly machine-generated. Only 100 comments carry a human label, and these were used for validation rather than as a gold-standard training subset in the style of the weak-supervision scheme in [19]. One consequence deserves to be stated outright: since the test set carries the same machine labels as the training set, the classification metrics reported above measure agreement with the annotator model's labels, not with human judgement, and a classifier that reproduced that model's systematic errors would be rewarded rather than penalized for it. A sample of 100 also keeps the confidence interval around the kappa wide. The interval from 0.446 to 0.758 spans values conventionally read as moderate agreement at the lower end and substantial at the upper. Human validation rests on a single annotator, so no human-to-human agreement estimate exists, and the reported kappa measures agreement between the model and one person.

Annotation used a single model and a single prompt formulation. The sensitivity of toxicity labels to prompt wording is documented [17] but was not measured here: neither alternative prompts nor a second annotator model for cross-checking was used. Evaluation is limited to aggregate metrics on a single test set. Breakdowns by video, by comment length and by the presence of code-switching, along with robustness checks against masked or obfuscated lexical content, were not carried out, so nothing can be said about model behaviour beyond the averaged figures. Finally, the redistribution restriction imposed by the platform's developer policies puts a step between a reader and the data. Obtaining the comment texts requires a request to the corresponding author rather than a download, which slows independent verification, and comments deleted by their authors or removed by the platform after May 2026 cannot be recovered from the API by anyone attempting to rebuild the collection from scratch.

*Prospects for further research.*

The nearest task is to refine the binary label into toxicity subtypes with a severity gradation [11], which would bring Kaz-ToCo in line with taxonomies adopted for other languages at the cost of substantially harder annotation. Next comes expanding the human-annotated subset to a size where it can serve as a gold-standard training set and running a controlled comparison of Kazakh-specific and multilingual encoders under comparable conditions. A practical extension of this work is a browser extension that flags toxic comments in real time.

## **Conclusion**

The Kazakh-language toxicity literature has not, until now, offered a general-purpose corpus of naturally occurring comments. The resources that exist were collected via keyword lists within an extremism or model-safety frame, documented unevenly, and only rarely released. This paper presents Kaz-ToCo, a corpus of 84,940 comments in Kazakh and in mixed Kazakh-Russian code-switched text, retrieved through the YouTube Data API v3 from beneath 30 Kazakhstani talk-show videos in May 2026 and labelled on a binary toxicity scale. The corpus was collected without relying on harmful keywords, normalized and filtered under a documented protocol, labelled by a locally deployed open-weight model, and checked against native-speaker annotation on a stratified sample, where the Cohen kappa came to 0.602. Splitting was performed at the video level, stratified by each video's internal share of toxic comments. Within this protocol two baselines were established. The fine-tuned kaz-roberta-conversational encoder reached a macro-averaged F1 of 0.8198 and an area under the receiver operating characteristic curve of 0.9063. The model over word and character n-grams with logistic regression reached 0.8086 and 0.8982. The gap of 0.0112 in macro-averaged F1 is small, and without a significance test it cannot be taken as establishing the transformer's advantage. It consists of 290 eliminated false positives bought at the cost of 107 additional missed detections.

The practical contribution is a resource several times larger than the few Kazakh datasets released so far, and the first collected without keyword-driven selection. The labels, the per-comment metadata, the splitting protocol, the random seeds and the hyperparameters needed to reproduce the figures reported here are published without restriction, while the comment texts, which the platform's developer policies do not permit to be redistributed, are available from the corresponding author upon reasonable request for non-commercial research use. Those figures fall noticeably short of the near-ceiling results published on keyword-collected corpora and are offered as a realistic rather than an optimistic baseline. The main limitations are the binary taxonomy, the predominantly machine-generated labels, and a human validation step carried out by a single annotator. Expanding the taxonomy, growing the human-annotated subset, comparing Kazakh-specific and multilingual encoders under comparable conditions, and testing robustness to masked text are the natural next steps. Information about the publication material. The labels, the per-comment metadata, the video-level split assignment, the annotation and training code, the random seeds and the hyperparameters that support the findings of this study are openly available at <https://github.com/cestramilya/kaz-toco>. Comment texts cannot be redistributed under the YouTube API Services Developer Policies and are available from the corresponding author upon reasonable request for non-commercial research use.

## *References*

- [1] *On the use of information and communication technologies by households in the Republic of Kazakhstan, 2025 [Electronic resource] / Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. – Astana, 2026. – URL: <https://stat.gov.kz/en/industries/business-statistics/stat-it/> (date of access: 27.08.2026).*
- [2] *Share of Kazakh-language content in media reaches 78% [Electronic resource] // KazTAG. – 2026. – URL: <https://kaztag.kz/en/news/share-of-kazakh-language-content-in-media-reaches-78> (date of access: 27.08.2026).*

- [3] Key findings from the Health Behaviour in School-aged Children survey in the Republic of Kazakhstan, HBSC 2022: Executive summary [Electronic resource] / National Center for Public Health of the Ministry of Health of the Republic of Kazakhstan. – Almaty, 2023. – URL: <https://hls.kz> (date of access: 27.08.2026).
- [4] Bolatbek M., Sagynay M., Mussiraliyeva S., Yeltay Z. Detection of offensive content in the Kazakh language using machine learning and deep learning approaches // *PeerJ Computer Science*. – 2025. – Vol. 11. – Art. e3027. – <https://doi.org/10.7717/peerj-cs.3027>
- [5] Bolatbek M., Sagynay M., Mussiraliyeva S., Zhastay Y. Machine learning methods for identifying destructive texts in the Kazakh language // *Proceedings of the 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology*. – New York: ACM, 2024. – P. 99–103. – <https://doi.org/10.1145/3701100.3701121>
- [6] Mussiraliyeva S.Z., Bolatbek M.A., Baispay G.B. Investigating long short-term memory approach for extremist messages detection in Kazakh language // *Expert Systems*. – 2025. – Vol. 42, No. 1. – Art. e13595. – <https://doi.org/10.1111/exsy.13595>
- [7] Bolatbek M., Sagynay M., Mussiraliyeva S., Baisylbayeva K., Yeltay Z. Kazakh language dataset for hate speech detection on social media text // *2024 IEEE 9th International Conference on Computational Intelligence and Applications (ICCIA)*. – Piscataway: IEEE, 2024. – P. 94–98. – <https://doi.org/10.1109/ICCIA62557.2024.10719327>
- [8] Jahan M.S., Oussalah M. A systematic review of hate speech automatic detection using natural language processing // *Neurocomputing*. – 2023. – Vol. 546. – Art. 126232. – <https://doi.org/10.1016/j.neucom.2023.126232>
- [9] Mussiraliyeva S., Bolatbek M., Omarov B. Corpus of religious extremist texts [Electronic resource]. – Mendeley Data, 2020. – VI. – <https://doi.org/10.17632/h272z7xv9w.1>
- [10] Kadyrbek N., Tuimebayev Z., Mansurova M., Viegas V. The development of small-scale language models for low-resource languages, with a focus on Kazakh and direct preference optimization // *Big Data and Cognitive Computing*. – 2025. – Vol. 9, No. 5. – Art. 137. – <https://doi.org/10.3390/bdcc9050137>
- [11] Saeed R., Afzal H., Rauf S.A., Iltaf N. Detection of offensive language and its severity for low resource language // *ACM Transactions on Asian and Low-Resource Language Information Processing*. – 2023. – Vol. 22, No. 6. – P. 1–27. – <https://doi.org/10.1145/3580476>
- [12] Yeshpanov R., Varol H.A. KazSANdra: Kazakh sentiment analysis dataset of reviews and attitudes. – *arXiv preprint arXiv:2403.19335*. – 2024. – <https://doi.org/10.48550/arXiv.2403.19335>
- [13] Goloburda M., Laiyk N., Turmakhan D. et al. Qorgau: Evaluating safety in Kazakh-Russian bilingual contexts // *Findings of the Association for Computational Linguistics: ACL 2025*. – Stroudsburg: ACL, 2025. – P. 9765–9784. – <https://doi.org/10.18653/v1/2025.findings-acl.507>
- [14] Zaghouani W., Ibrahim S.A., Muratbek A., Zhakenov O., Akhmetzhanova A. KZ-SafetyPrompts: A Kazakh safety evaluation prompt dataset for large language models. – *arXiv preprint arXiv:2605.26947*. – 2026. – <https://doi.org/10.48550/arXiv.2605.26947>
- [15] Sap M., Swayamdipta S., Vianna L., Zhou X., Choi Y., Smith N.A. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection // *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. – Stroudsburg: ACL, 2022. – P. 5884–5906. – <https://doi.org/10.18653/v1/2022.naacl-main.431>
- [16] Gilardi F., Alizadeh M., Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks // *Proceedings of the National Academy of Sciences*. – 2023. – Vol. 120, No. 30. – Art. e2305016120. – <https://doi.org/10.1073/pnas.2305016120>
- [17] Li L., Fan L., Atreja S., Hemphill L. HOT ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media // *ACM Transactions on the Web*. – 2024. – Vol. 18, No. 2. – P. 1–36. – <https://doi.org/10.1145/3643829>
- [18] Jadhav S., Shanbhag A., Thakurdesai A., Sinare R., Joshi R. On limitations of LLM as annotator for low resource languages. – *arXiv preprint arXiv:2411.17637*. – 2025. – <https://doi.org/10.48550/arXiv.2411.17637>
- [19] Aitim A. LLM-assisted weak supervision for low-resource Kazakh sequence labeling: Synthetic annotation and CRF-refined NER/POS models // *Applied Sciences*. – 2026. – Vol. 16, No. 8. – Art. 3632. – <https://doi.org/10.3390/app16083632>
- [20] Rupa S.A., Anwar M.M., Ritu N.A., Arifin K.N. Bangla-ToCo: A context-aware dataset for Bengali toxic comment detection // *Data in Brief*. – 2025. – Vol. 63. – Art. 112277. – <https://doi.org/10.1016/j.dib.2025.112277>