BOISHOMMO: Benchmarking Class Imbalance in Bangla Multi-Label Hate Speech Detection

Showrov Azam[1*], Sifat Khan[1], Rashed Hossain[1], Nadim Mahmud[1], and Md Abdullah Al Kafi[1]

[1] Department of Computer Science and Engineering, Daffodil International University, Dhaka-1216, Bangladesh

{azam15-5843*, khan15-5869, hossain15-5679, nirob22205101549, kafi.cse}@diu.edu.bd

Abstract. Class imbalance is an endemic and problematic issue in the Natural Language Processing (NLP) field, especially when Natural Language Processing is applied to low-resource languages (LRLs), in which annotated corpora often do not reflect the biased distributions of real-world hate speech. In an attempt to reduce this shortcoming, we introduce BOISHOMMO, a multi-labeled Bangali dataset specifically made available to enable the benchmarking of the agglutinative language-based class imbalance phenomena in a language with a population of over 250m; spoken by over 250 million people. The social-media comments contained in BOISHOMMO are annotated by native speakers on 2,499 comments, in ten categories that overlap, as follows: Race, Behaviour, Physical, Class, Religion, Disability, Ethnicity, Gender, Sexual Orientation, and Political. The final annotation process was done by majority voting and stabilized by both the Cohen and Fleiss Kappa statistics to guarantee that there was inter-rater consistency. In order to demonstrate the usefulness and complexity of the dataset, we performed the baseline classification experiments with the use of Logistic Regression and Linear Support Vector Classifiers (Linear SVC). The best performance achieved a Macro F1 -score of only 0.2101, which proved the existence of strong performance differences among majority categories (e.g., Behaviour) and minority categories (e.g., Disability). These numerical results support the idea that BOISHOMMO represents a challenging standard of testing the algorithms that are sensitive to class imbalance. It is expected that the publishing of this dataset will contribute to future studies in the fairness-conscious content moderation and multi-label classification of Indic languages.

Keywords: Class Imbalance Analysis, Multi-label classification, Social media annotation, Low-resource NLP, Morphologically rich language, Inter-annotator agreement.

1 Introduction

High-quality annotated data is of growing importance to the safety-critical moderation tasks of natural language processing (NLP). Nevertheless, current corpora can tend to

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_36

lose the context of low-resource languages and complicated, intersectional abuse, resultant in an overall disposition towards high-resource languages and binary labelling. Multi-label and natively written materials are severely underrepresented in Bangla which is spoken by more than 250 million people in comparison with inequalities in real life. Such sparseness constrains the methodological innovations in social computing. Also, there should be a solution to class imbalance: models that are trained on artificially balanced data are more likely to overfit majority classes and underfit harm to minority, recreating system-level injustice.

We fill this gap with a standardized, multi-label dataset of Bangali hate speech, BOISHOMMO, which will be helpful in analyzing the issue of class imbalance. The data is a collection of 2,499 comments on Facebook news pages of open access (e.g., Prothom Alo, Jugantor) and it includes a naturalistic discourse on social and political issues. Unlike the previously transliterated sources, BOISHOMMO uses the native Bangla script as the means of linguistic enrichment of the language. The corpus realizes ten overlapping dimensions of hate: Race, Behaviour, Physical, Class, Religion, Disability, Ethnicity, Gender, Sexual Orientation and Political annotated to allow error analysis in fine detail. Notably, the label assignment represents credible long-tail constraints, and it is also that preserves the natural skew which would otherwise be obscured by artificial balancing.

To guarantee reliability, data were annotated by three native speakers based on strict guidelines and final labels were done by majority voting and confirmed through Cohen and Fleiss Kappa. In addition to curation, BOISHOMMO also presents a difficult challenge to predictive modelling. The classical machine learning models were used as a baseline, and the results showed a low level of Macro F1-score (approximately 0.21) on test data. This performance discrepancy directly measures the performance challenge of the task, which validates the fact that standard classifiers have a challenge with the realistic imbalance between datasets. It therefore makes the resource a critical probe to devise data-level solutions (e.g. resampling) and algorithm-level interventions (e.g. cost-sensitive learning). Following a rigorous code of ethics, BOISHOMMO will provide a proven tool for developing context-sensitive hate speech detectors in the low-resource contexts.

2 Literature Review

A significant amount of literature has been developed in automated hate speech detection, though not all languages, label designs, or scripts have gotten equal attention, leaving low-resource and morphologically rich languages such as Bangali relatively underserved. Systematic reviews and foundational surveys note that first-mover surveys and research tended to focus on English and binary labelling, which obscures errors in minority classes and restricts generalization to intersectional or nuanced harms [5], [18]. This issue of garbage in, garbage out of abusive language datasets, in which weak curation results in biased models, has been repeatedly pointed out, leading to more explicit definitions, stronger annotation guidelines, and explicit reporting of class distributions and sampling frames to allow robust, reproducible research [18], [6], [19],

[1]. In this emerging topography, recent Bangla work encompasses native script as well as transliterated corpora, audio and written forms, and multi-label schemes, and is a move towards realism in both the construction and the assessment of data [10], [6], [8].

Seminal baselines showed that hate speech can be differentiated by character n-grams and lexical features but also revealed that high accuracy could be misleading about failure modes based on subtlety, context and class imbalance [12], [3], [19], [1]. Subsequent analysis reveals that the definition of labels, sampling, and guidance of annotators can have a significant effect on datasets, and the importance of clear documentation and multi-dimensional analysis beyond aggregate accuracy is justified [18], [6], [19], [13]. Systematic blind spots (e.g. templated paraphrases, identity terms) not visible to headline metrics can also be further exposed by functional test suites and behavioural testing, which highlights the need to conduct specific evaluations in addition to standard splits [16], [15]. In this connection, explicitly imbalanced curated corpora are essential in the development and testing of mitigation strategies including reweighting, resampling, focal losses, and threshold calibration that enhance minority-label detection [6], [2], [4], [7], [11]. Multi-label tasks should always be analyzed using per-label measures and calibration analysis, as micro-average scores may mask poor recall of rare, yet socially relevant categories [6], [16], [4], [7].

The resources in Bangla have increased on three complementary fronts, audio offensiveness, transliterated multi-label text, and native-script multi-label hate categories. BAAD takes care of automatic speech recognition of offensive Bangali speech and also offers thousands of audio samples of offensive and near-offensive speech and forms a basis of speech-based moderation pipelines based on dialectal and acoustic variability [9]. This audio emphasis supplements text corpora and emphasizes multimodality as an operational need of real-world safety systems, in which abusive content occurs in speech, text, and code-mixed forms across platforms. BANTH, in transliterated Bangla, presents a large multi-label dataset that is representative of YouTube discourse, where users often combine scripts and languages; by further pre-training encoders on transliterated corpora and using translation-based prompting, BanTH reports high gains and notes domain adaptation as a determinant of performance [8]. Such transliteration-mindful approaches are related to larger multilingual NLP trends, and they demonstrate that zero-shot and few-shot detection can be improved by mapping under-resourced forms into high-resource representation spaces, either through further pre-training or translation [6], [13]. BOISHOMMO adds to the native-script side with a multi-label corpus obtained using public Facebook news pages, annotated in parallel categories (e.g., Religion, Ethnicity, Gender, Behaviour) and explicitly reporting an imbalanced distribution; large inter-annotator agreement on explicit abuse exists alongside underreporting of subtle harms [10]. This explicit asymmetrical and multi-label architecture render BOISHOMMO an appropriate testbed to test data-level and algorithm-level remedies in realistic settings and is useful in studies of cost-sensitive learning, class-sensitive augmentation, and per-label thresholding [10], [2], [7], [11].

Classical linear baselines have been replaced with transformer-based models with language- and domain-adaptive pre-training through modeling. The modern results are dominated by BERT-style models and multilingual encoders (e.g., XLM-R), particularly when further trained on in-domain or transliterated corpora [6], [13]. The key

aspect of multi-label hate speech is that label dependencies can be captured: classifier chains, conditional decoding, and label graph attention can be used to achieve higher minority label recall on co-occurrence patterns (e.g., Religion with Ethnicity) [13], [14]. In the case of transliterated Bangla, domain-specific tokenization and further pre-training provides state-of-the-art performance, which supports the claim that subword segmentation and vocabulary coverage should capture both social-media orthography and code-mixing [8], [6]. In the case of native-script Bangali, Unicode normalization, morphology-aware sub wording, and sensitive treatment of diacritics and compound forms make sparsity less and semantics more robust [10]. In both environments, multi-label heads using sigmoid activation, class weighting and focal loss have become standard options to deal with skew, which is complemented by calibrated decision rules and per-label threshold tuning to trade-off precision and recall [2], [4], [7], [14], [11].

The practices of evaluation have taken new forms to deal with imbalance head-on. Big English benchmarks (OLID/OffensEval; HatEval) both spurred method development and indicated multi-granular labeling (offensive vs. target vs. type), as well as showing cross-dataset brittle behavior [19], [1]. It is also shown in dataset studies that the policies of label, sample, and instruction can change the balance of classes and model behaviour and encourage more extensive documentation and post-hoc auditing [18], [6], [17], [2]. Behavioural/functional tests (HateCheck) also stress-test robustness of templated constructions and identity terms [16]; CheckList-style testing generalizes this to perturbation and capability matrices in NLP more generally [15]. Identity-sensitive metrics (e.g. subgroup AUCs and unintended-bias scores) and explicit calibration tests have now become standard suggestions to ensure equitable deployment, as aggregate micro-F1 or accuracy may obscure differences in error rates across the protected groups [2], [4], [7], [11]. Training and validation on datasets that reflect real-world skew, such as BOISHOMMO, can expose brittleness that is obscured by artificially balanced splits, and so it is desirable to promote approaches that make explicit long-tail attention [10], [6], [2].

The literature is permeated with ethical and curatorial directions, such as platform policies and adherence, sensitive content, and care- are domains where BAAD and similar datasets describe procedures and caveats that can be transferred into text datasets [9], [6], [3], [1]. There is high consensus on explicit abuse, as well as under coverage of euphemisms, sarcasm and contextual targeting, which drives repetitive collection protocols and active learning to elicit borderline and implicit cases [18], [6], [16], [15], [4]. Reproducibility and principled reuse are supported by documentation of collection windows, platform sources, label definitions, and sampling frames, which are now standard in major releases, and make it possible to perform second analyses of both bias and coverage gaps [6], [19], [1], [13], [2]. Such openness is especially pertinent in low-resource contexts where smaller data sets may accidentally encode regional or dialectal biases that influence downstream moderation [10], [8], [13], [17], [2].

Cross-resource synthesis of Bangali proposes a strategic research agenda: use the native-script imbalance profile of the BOISHOMMO system together with the transliteration realism of BanTH to investigate the impact of scripts and adaptation mechanisms; use audio of text and BAAD to investigate multimodal fusion as a method of higher recall in noisy conditions; and investigate domain-shift robustness across

platforms (Facebook vs. YouTube) and time-periods to learn evolving abusive vocabularies [10], [6], [8], [13]. Experimentally, data-level benchmarking [e.g., class-balanced sampling, mixup/EDA variants], cost-sensitive algorithm-level benchmarking [e.g., focal loss, reweighting], and post-hoc calibration and thresholding should be benchmarked to provide fair per-label performance under deployment constraints [2], [4], [7], [14], [11]. Confusion analyses between neighbouring labels (e.g., Behaviour vs. Gendered insults), sensitivity to tokenization of both native script and transliteration, as well as ablations on label-correlation modules should also be reported [10], [8], [16], [1], [14].

Enduring issues are to capture subtle, context-dependent harms, to mitigate over-reliance on explicit slurs, and to deal with code-switching at intra-word and intra-sentence levels typical of South Asian social media [18], [6], [13], [17]. The addition of context-thread history, temporal metadata, linked media could enhance performance at the cost of increased privacy and access, which must be carefully governed in line with platform and ethical rules [18], [6], [19], [1], [2]. Temporal drift requires semi-automatic updating of models based on weak supervision with human-in-the-loop evaluation to maintain model utility, in particular, at politically sensitive times when discourse evolves quickly [18], [6], [15], [2]. In the case of low-resource environments, cross-lingual transfer and translation-prompting provide practical solutions, although great caution is needed to ensure that performance improvements do not come at the expense of dialectal coverage [6], [13], [17], [2].

Altogether, the literature is coalescing around multi-label, imbalance-sensitive, script-sensitive, and ethically edited models as the requirements of effective hate speech identification in low-resource languages like Bengali. BAAD generalizes to audio offensiveness as part of dialectal variability [6]; BanTH generalizes to explicit imbalance and strong agreement with explicit abuse as part of explicit annotation using native-script multi-labeling [8]. Altogether, these datasets and tools make it possible to support context-sensitive moderation in low-resource, multi-label NLP: surveys that map the problem space [5], [18], [12]; Bangla-specific data sets [9], [8], [10], [6], [3]; modeling techniques that focus on fairness and robustness [16], [15], [3], [1], [13]; and more recent evaluation systems [17], [2], [4], [7], [14], [11].

3 Methodology

Fig. 1. shows the five main steps that went into making the BOISHOMMO dataset. These will entail data capture and cleaning, annotation, agreement, and assemblage. All the steps guarantee the dataset's quality, reliability, and usefulness in identifying hate speech in Bangla.

Fig. 1. Dataset Creation Methodology

3.1 Data Collection

A total of three popular public Facebook news sources were sampled to gather Bangali comments in 2020-2021: Prothom Alo, Jugantor and Kaler Kantho. These were chosen based on their socio-political topicality, since discussions within these spaces occur online often where one may expect hot-tempered debates and hate speech.

The dataset consists of 2499 comments in pure Bangla writing. Every entry contains the original Bangla text, the English translation of the text, and binary annotations of the ten hate categories is given on **Fig. 2**.

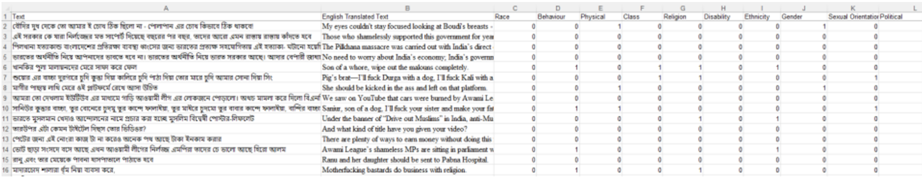

Fig. 2. Annotated Comments

Table 1 Gives a general overview of the distribution of the labels among the different hate categories in the dataset. Each category has an associated count and percentage, showing the distribution of hate content types across time. The most significant part of entries is categorized as "Behaviour" entries (19.3%), even more than "Religion" (11.6%) and "Ethnicity" (14.0%). This distribution informs the frequency and the nature of hate speech types recorded in the dataset and is critical for informing the underlying dataset characteristics and subsequent analysis. So, the dataset is heavily imbalanced, with 'Behaviour' (19.3%) dominating and 'Disability' (5.3%) being the minority class.

Table 1. Label Distribution Summary

Hate Category	Count	Percentage
Behaviour	482	19.3%
Ethnicity	350	14.0%
Religion	290	11.6%
Political	230	9.2%
Race	227	9.1%
Sexual Orientation	217	8.7%

Gender	205	8.2%
Class	192	7.7%
Physical	170	6.8%
Disability	136	5.3%
Total	**2499**	**100%**

Fig. 3 shows the distribution of the categories of hate speech throughout the 2,499 comments of the BOISHOMMO dataset. The statistics show considerable inequality of classes, as the most prevalent one is Behaviour (482 instances). On the other hand, there is no overrepresentation of categories like Disability and Physical with only 136 and 143 instances, respectively. This distribution is very relevant to the true situation in the field studied with regard to the prevalence of various forms of hate speech.

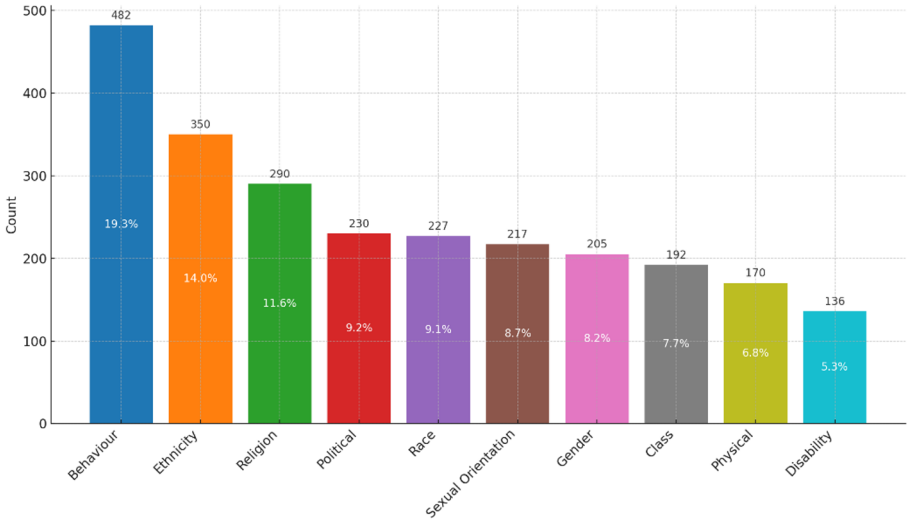

Fig. 3. Distribution of Hate Speech Categories in the BOISHOMMO Dataset.

3.2 Data Preprocessing

The preprocessing of raw social media text was according to a carefully specified pipeline, hence ensuring reproducibility. The entire procedure was done in Python with the help of pandas, regex, emoji, and Bangali-normalizer libraries. The main activities are outlined as follows. The preprocessing pipeline had four different stages.

Eradication of Non-Linguistic Artifacts.

In this study we completely removed every non-linguistic element of the corpus and used a set of pattern-based filters which are used to leave only semantically rich linguistic material.

- URLs: detected and removed using Python re (regex pattern: r'http\S+|www\.\S+').
- User mentions (@ username): removed using r'@\w+'
- Hashtags: removed the # symbol; retained the tag text only if it contained Bangla characters.
- Emojis and emoticons: removed with the Python emoji library, which is supplemented with a set of emoji-regex, and thus cleaning up the data of affective symbols.
- Character repetition and punctuation: standardized through rules of regular expressions that minimized redundancy thus limiting noise and improving the quality of the text.
- Unnecessary whitespace: trimmed using .strip() and re.sub(r'\s+', ' ', text).

This step ensured that only linguistically relevant text remained.

Handling Code-Switching and Non-Bangla Tokens

The social media information often captures English-based lexical objects enmeshed with Bangla, thus requiring a stringent method in order to maintain the linguistic purity in subsequent analyses:

- In an attempt to ensure reproducibility, we outlined Bangla text using its Unicode range, namely, U + 0980-U + 09FF.
- As such, any token representation of characters outside this given Unicode gamut was cut out of the dataset.
- The English letters, Romanized forms of the Bangali language (e.g. “ami valo”), numeracy, and alphanumeric slang were also cleansed instead of being transliterated.
- This will make the resulting corpus made up of Bangla lexical items only.

Text Normalization (Bangla Unicode Consistency)

Bangla text frequently contains visually similar but technically different Unicode forms. To address this, we applied:

- Unicode normalization (NFC)
- Normalization of variant characters using standard Bangla NLP rules: Normalizing “ড়/ঢ়” variants, fixing broken diacritics, Standardizing vowel signs (e.g., split "ো" sequences), Removing zero-width joiners
- Normalization library: bnunicodenormalizer

Bangla Stop-Word Removal

We utilized a filtered list of the Bangali stop words. The rules were:

- Some words were functional (e.g., এবং, কিন্তু, যে, এই), removed afterwards so that information cannot be lost in the comments.
- Comments that were very short (the ones less than 5 words) were filtered minimally to maintain meaning.

Fig. 4 illustrates the key stages involved in the preprocessing of the dataset.

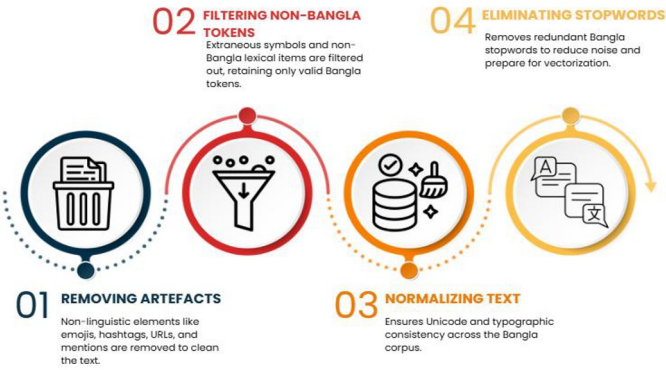

Fig. 4. Data Processing Steps

3.3 Data Annotation

Each comment was translated into English and annotated in ten categories of hate-speech, including: Race, Behaviour, Physical, Class, Religion, Disability, Ethnicity, Gender, Sexual Orientation and Political. A multi-label scheme was used, in which every category was coded (1 presence, 0 absence) and hence one comment was allowed to be related to several categories. All of the 2,499 comments were annotated by three native speakers of the Bangali language separately in Google Sheets, following the elaborate instructions that included a definition and an example of context of each category. Ambiguous cases were solved with the help of ordinary conversations and final labels were created with the help of the majority voting, which required at least two annotators. This approach provided annotations that were consistent and reliable besides covering the overlapping aspects of hate speech.

3.4 Statistical Analysis of Annotation

In this research, the procedure of determining the best statistical models is used to determine the validity of manual annotations in the BOISHOMMO corpus. The data is specifically meant to classify hate speech (that has multiple labels), and it is annotated by three native speakers of Bengali (who are independent); thereby ensuring that inter-rater consistency was of primary concern, before making the corpus available to downstream machine-learning uses. In that regard, two well-established metrics of agreement, such as the κ of Cohen and the κ of Fleiss, were selected due to their suitability

to categorical data and their wide usage in the natural language processing evaluation guidelines.

Cohen's kappa (κ_K). used to measure pairwise agreement between annotators of each of the ten categories of hate-speech, thus measuring the degree to which the observed concordance exceeded that which would be likely given randomness. The κ coefficient goes as:

$$k_k = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the observed agreement between two annotators and p_e is the hypothetical probability of chance agreement. This model was chosen because it directly measures reliability for binary labels and is interpretable using standard thresholds (e.g., slight, moderate, substantial, or almost perfect agreement).

Fleiss' Kappa (κ_F) was used to estimate overall agreement across all three annotators simultaneously, which is more appropriate for fixed sets of multiple raters. It is calculated as:

$$k_F = \frac{\bar{p} - c}{1 - \bar{p}_e}$$

where $\bar{p}$ is the mean proportion of agreement across all items and $\bar{p}_e$ is the expected agreement by chance. This allows assessing consistency at the corpus level rather than only pairwise.

In a bid to critically determine the reliability of annotation, we determined Cohen Kappa (κ) on a pairwise agreement and Fleiss Kappa on multi-rater consistency with Python 3.9 (scikit-learn and statsmodels). The annotated corpus was created in binary vectors as a structure of 2,499 comments. The computation of pairwise κ was done on all combinations of annotators with ten categories to determine category specific ambiguity. On the corpus-wide agreement, Fleiss κ was used on the entire dataset (24,990 total ratings) to be able to identify agreement. These statistical scores affirm the stability of the annotations as a gold standard in downstream predictive tasks and all the scripts are available in a way that can ensure reproducibility.

Scores followed standard thresholds:

- 0.00–0.20: Slight agreement
- 0.21–0.40: Fair agreement
- 0.41–0.60: Moderate agreement
- 0.61–0.80: Substantial agreement
- 0.81–1.00: Almost perfect agreement

Table 2 Represents Cohen's Kappa (κ) scores of agreements between three annotators of the ten hate speech categories in the BOISHOMMO. The κ values range from 0.89 to 0.99, indicating high to almost perfect agreement, which confirms strong consistency

among annotators in identifying hate categories. These high agreement scores reflect the binary nature of the labels and the clear annotation guidelines provided to annotators. Although the κ values may appear unusually strong, they are supported by the presence of explicit and unambiguous instances of hate speech that are frequently found in social media discourse.

Categories such as Religion (κ = 0.9914), Gender (κ = 0.9881), and Political (κ = 0.9874) show the highest consistency across annotators, suggesting more precise boundaries and stronger annotator consensus. In contrast, Disability (κ = 0.8999) and Race (κ = 0.9271) yield comparatively lower κ values, indicating more subjective interpretation or ambiguity.

To ensure label reliability, a majority voting strategy was employed to finalize category labels rather than relying on individual annotators. This approach strengthens the annotated dataset's overall credibility and objectivity.

However, a key limitation must be acknowledged: due to data filtering and the focus on clear-cut examples of hate, more subtle or implicit cases of hate speech may be underrepresented. This may affect the generalizability of models trained on the dataset when applied to nuanced or borderline content.

Table 2. Pairwise Inter-Annotator Agreement (Cohen's Kappa)

Hate Category	k(Annotator 1 VS 2)	k(Annotator 1 VS 3)	k(Annotator 2 VS 3)	Average k
Race	0.9142	0.9487	0.9185	0.9271
Behaviour	0.9933	0.9825	0.9825	0.9861
Physical	0.9624	0.9771	0.9697	0.9697
Class	0.9857	0.9886	0.9857	0.9867
Religion	0.9900	0.9928	0.9914	0.9914
Disability	0.8947	0.9504	0.8545	0.8999
Ethnicity	0.9896	0.9794	0.9794	0.9828
Gender	0.9822	0.9923	0.9897	0.9881
Sexual Orientation	0.9945	0.9755	0.9755	0.9818
Political	0.9862	0.9849	0.9912	0.9874

An inter-annotator agreement analysis was conducted of each hate speech type to measure the consistency and reliability of the annotations throughout the dataset. The proportion of agreement between the three annotators is quite extreme in all ten classes. Over 97 percent of the samples in each category received complete agreement (3/3 annotators), and the Class and Gender categories recorded the highest correspondence (both at 99.7 percent). In contrast, Race was the lowest, reaching 97.8%. The portion of majority agreement (2/3 annotators) was moderate, ranging between 0.3 and 2.2 percent.

In contrast, there was no disagreement (1/3 annotators) at all, testifying to the strength and the definiteness of the annotation guidelines that were provided. The level of partial disagreement was somewhat elevated at 2.2% and 1.0%, respectively, in the Race and Behaviour categories; however, the values remained within an acceptable range of annotation differences. These scores provide rather convincing evidence that the annotation process has high inter-rater reliability, indicating both the quality of the

labelling schema and the success of the guidelines provided to the annotators. The categorical names were assigned as the final rule in a majority-voting scheme, which further proves the quality and reliability of the data.

Fig. 5 illustrates the high consistency of the annotation process, where 'Full Agreement' (3/3 annotators) exceeds 97% across all categories. This dominance of unanimous agreement validates the clarity of the annotation guidelines and the reliability of the final labels.

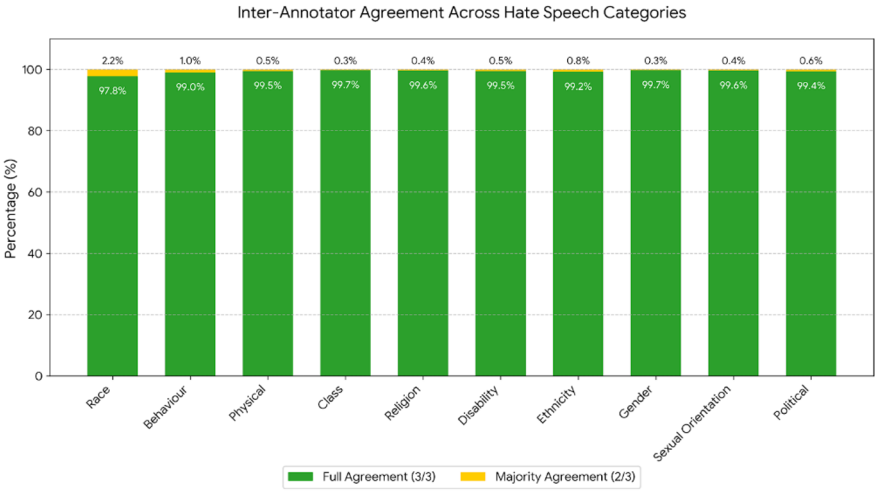

Fig. 5. Inter-Annotator Agreement Across Hate Speech Categories.

Overall Multi-Rater Agreement with Fleiss' Kappa.

In determining the general consistency of the three annotators, we computed Fleiss Kappa in all the ten categories. The calculated score was 0.9807 (based on 24989 annotation points) and this is an indication that it is very close to perfect agreement. This finding demonstrates that the annotation guidelines were always used with a very low level of ambiguity, which supports BOISHOMMO as a credible ground-truth source of low-resource hate speech studies.

4 Benchmark Experiments

4.1 Baseline Model Training

In an attempt to determine the validity and performance of the BOISHOMMO corpus, we engaged in a methodological evaluation using two well-known paradigms of machine-learning that are commonly used to detect text classification.

Experimental Setup. The raw textual data were systematic cleansed, i.e., non-Bengali orthographic characters, Arabic digits, and words that were marked by a special Bengali stop-word compendium were removed. After that, a representation in the form of a Term Frequency -Inverse Document Frequency (TF-IDF) was created, limited by a maximum of 5,000 unique features, which converted the corpus into a numerical vector space. The obtained dataset was split into an 80 % training component and 20 % testing component; the partition was stratified in such a way that the original distribution of the class labels across partitions was maintained.

Algorithms. Due to the multi-label structure of the corpus, where each comment can be a member of a number of annotation categories, we have chosen a One-vs-Rest (OvR) scheme. Under this framework two baseline classifiers were instantiated, namely: Logistic Regression, which was selected because of its probabilistic interpretability and computational efficiency; and the Linear Support Vector Classifier (Linear -SVC), which was selected because of its demonstrated effectiveness on high-dimensional and sparsely populated feature spaces characteristic of textual data. Cases of label coding which were non binary (not equal to 0 or 1) were carried to 0 therefore making the learning process even in terms of binary premise.

4.2 Benchmark Evaluation

We were assessing the models with reference to Exact Match Accuracy, Hamming Loss and Macro F1-Score. Since the ratio between the classes in the dataset is severe, the Macro F1-Score will be of the most use since it does not differentiate the classes based on their frequency.

Results Analysis. Table 3 shows the performance of the two models in comparison to each other. The Linear SVC had a Macro F1 -score of 0.2101, which was significantly higher than the Logistic Regression model, which scored 0.1203.

Table 3. Baseline Classification Results

Metric	Logistic Regression	Linear SVC
Exact Match Accuracy	22.20%	20.40%
Hamming Loss	0.1346	0.1418
Macro F1-Score	0.1203	0.2101

Class-wise Performance. The in-depth classification reports demonstrate the effect of the imbalance of the dataset towards the long-tail.

Majority Classes. Those that received more support like Behaviour and Political fared fairly well. Linear SVC had an F1-score of 0.49 for Behaviour and 0.43 for Political.

Minority Classes. Both models had a lot of difficulties with underrepresented categories. As an example, Physical (30 samples), Disability (11 samples), and Race (64 samples) have F1 -scores of 0.00 or close to zero under Logistic Regression.

Model Comparison. While Logistic Regression achieved a slightly higher exact match accuracy (22.2%), it did so by biasing towards the majority class (predicting 0 often). Linear SVC was more sensitive to minority classes with non-zero values in F1 scores of Race (0.08) and Sexual Orientation (0.18), but the Logistic Regression did not detect these groups at all (0.00).

These baseline scores allow confirming that the dataset of BOISHOMMO is not an easy one. Even simple classical models cannot reveal the specifics of hate speech in low-resource, imbalanced environments, which is why it is necessary to conduct further research on the more advanced deep learning systems and data augmentation approaches.

To have a more granular diagnosis of the behavior of the classifiers, a heatmap of the classification reports is shown in **Fig. 6**. This figure analysis shows how harsh the long-tail distribution of the dataset is to the operation of models.

The obvious distinction between the dark stippled blocks (denoting the majority classes such as Behaviour and Political), and the non-stippled blocks (denoting the minority classes such as Disability and Physical) shows the bias of the models towards often used labels. Despite the fact that both of the classifiers are adversely affected by data sparsity, which has been identified by the heatmap, Linear SVC (Right) is more sensitive than Logistic Regression (Left). Specifically, Linear SVC reaches non-zero F1-scores in difficult categories, like Sexual Orientation and Race, but Logistic Regression loses all evidence of such information, returning a null-value. This visualization goes directly to demonstrate that aggregate measures hide critical malfunctions in the detection of intersectional, low resource hate speech.

Fig. 6. Comparative heatmaps of class-wise performance metrics (Precision, Recall, F1-Score) for Logistic Regression and Linear SVC.

5 Conclusion

We presented BOISHOMMO, a carefully curated multi-labeled corpus of Bengali hate-speech on social media (2,499 comments on ten overlapping categories). The reliability and consistency of the annotation protocol are ensured by high scores on Cohen and Fleiss Kappa. Moreover, trials on baseline experiments with Linear SVC set a low Macro F1-score (0.2101), which quantitatively reflects how difficult it was to deal with the harsh impact of the imbalance in the problem of classes. The much larger gap in performance of majority (e.g., Behaviour) and minority classes (e.g., Disability) confirms the incapability of the standard classifiers to cope with the realistic long-tail distribution. BOISHOMMO can, therefore, be regarded as a critical testbed that the future research regarding imbalance-sensitive deep learning models and fairness-aware content moderation in low-resource languages can be based on.

References

1. Basile, V. et al.: SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter.
2. Borkan, D. et al.: Nuanced metrics for measuring unintended bias with real data for text classification. In: The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019. pp. 491–500 Association for Computing Machinery, Inc (2019). https://doi.org/10.1145/3308560.3317593.

3. Davidson, T. et al.: Automated Hate Speech Detection and the Problem of Offensive Language. (2017).
4. Dixon, L. et al.: Measuring and Mitigating Unintended Bias in Text Classification. In: AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 67–73 Association for Computing Machinery, Inc (2018). https://doi.org/10.1145/3278721.3278729.
5. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text, (2019). https://doi.org/10.1145/3232676.
6. Founta, A.-M. et al.: Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. (2018). https://doi.org/10.1609/icwsm.v12i1.14991.
7. Guo, C. et al.: On Calibration of Modern Neural Networks. (2017).
8. Haider, F. et al.: BanTH: A Multi-label Hate Speech Detection Dataset for Transliterated Bangla. (2025).
9. Hossain, M.F. et al.: BAAD: A multipurpose dataset for automatic Bangla offensive speech recognition. Data Brief. 48, (2023). https://doi.org/10.1016/j.dib.2023.109067.
10. Al Kafi, A. et al.: BOISHOMMO: Holistic Approach for Bangla Hate Speech. (2025).
11. Lin, T.Y. et al.: Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell. 42, 2, 318–327 (2020). https://doi.org/10.1109/TPAMI.2018.2858826.
12. Malmasi, S., Zampieri, M.: Detecting hate speech in social media. In: International Conference Recent Advances in Natural Language Processing, RANLP. pp. 467–472 Incoma Ltd (2017). https://doi.org/10.26615/978-954-452-049-6_062.
13. Ousidhoum, N. et al.: Multilingual and Multi-Aspect Hate Speech Analysis. (2019).
14. Read, J. et al.: Classifier chains for multi-label classification. Mach Learn. 85, 3, 333–359 (2011). https://doi.org/10.1007/s10994-011-5256-5.
15. Ribeiro, M.T. et al.: Beyond Accuracy: Behavioral Testing of NLP Models with CheckList.
16. Röttger, P. et al.: HATECHECK: Functional Tests for Hate Speech Detection Models.
17. Sap, M. et al.: The Risk of Racial Bias in Hate Speech Detection. Association for Computational Linguistics.
18. Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: Garbage in, garbage out, (2021). https://doi.org/10.1371/journal.pone.0243300.
19. Zampieri, M. et al.: SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval).