Vio-Lens: A Novel Dataset of Annotated Social Network Posts Leading to Different Forms of Communal Violence and its Evaluation

Sourav Saha † [1], **Jahedul Alam Junaed** † [1], **Maryam Saleki** [2],
Arnab Sen Sharma [3], **Mohammad Rashidujjaman Rifat** [4], **Mohamed Rahouti** [2]
Syed Ishtiaque Ahmed[4], **Nabeel Mohammad** [5], **Ruhul Amin** [2]
[1] Shahjalal University of Science and Technology, Bangladesh,[2]Fordham University, USA,
[3] Northeastern University, USA, [4] University of Toronto, Canada,
[5] North South University, Bangladesh,
{sourav95,jahedul25}@student.sust.edu ,*mamin17@fordham.edu,

Abstract

This paper presents a computational approach for creating a dataset on communal violence in the context of Bangladesh and West Bengal of India and benchmark evaluation. In recent years, social media has been used as a weapon by factions of different religions and backgrounds to incite hatred, resulting in physical communal violence and causing death and destruction. To prevent such abusive use of online platforms, we propose a framework for classifying online posts using an adaptive question-based approach. We collected more than 168,000 YouTube comments from a set of manually selected videos known for inciting violence in Bangladesh and West Bengal. Using both unsupervised and later semi-supervised topic modeling methods on those unstructured data, we discovered the major word clusters to interpret the related topics of peace and violence. Topic words were later used to select 20,142 posts related to peace and violence of which we annotated a total of 6,046 posts. Finally, we applied different modeling techniques based on linguistic features, and sentence transformers to benchmark the labeled dataset with the best-performing model reaching ∼71% macro F1 score.

Category	Sub-Category	Example
Direct Violence	Kill/Attack	তার গর্দান কেটে ফেলা হোক (Let his neck be cut)
	Re/Desocialization/Oppression	হিন্দুদের ভারতে পাঠিয়ে দাও (Send the Hindus to India)
Passive Violence	Passive/Justification	সরকারের দোষ, সরকারের দালালি বন্ধ কর (Blame the government, stop the government brokering)
Non-Violence	Social-Rights	যে হামলা হয়েছে তার তীব্র প্রতিবাদ জানাচ্ছি এবং এই ঘটনার সঠিক বিচার চাই (Strongly protesting the attack and I want a fair trial of this incident)
	Non-Violence	ধর্ম এসব শিক্ষা আমাদের দেয় না বরং আমাদের উচিত মিলেমিশে থাকা (Religion doesn't teach us these things but we should live together with harmony)

Table 1: The Table depicts examples of different categories: Direct Violence, Passive Violence, and Non-Violence. We also show the English translation using Google Translator service.

1 Introduction

With the rise of social media users, different kinds of toxic behavior have been climbing sharply, including hate speech (Silva et al., 2016; Romim et al., 2022), online abuse (Nobata et al., 2016; Huang et al., 2014), and even for terrorist purposes. Previous analyses focusing on the in-group and out-group community relationships and conflicts in Southeast Asia highlighted the role of perceived relative deprivation, economic inequalities, and competitions as the precursor for such communal violence (Tausch et al., 2009) which is now taking place on social media in a larger scale. In recent years, social media has become a vehicle for inciting violence against minority and underrepresented communities, especially, based on ethnicity, religion, and even nationality around the world, not to mention increasingly in Southeast Asia. Even though there exist different approaches to detect whether an online post has negative sentiment (Islam et al., 2021), or expresses hatred (Romim et al., 2022), and in some cases, the veracity of content (Hossain et al., 2020), there is a lack of computational approach to identify violence inciting posts for instigating in-group factions to perform harmful activities on out-group communities by targeting them on social media. Most importantly, there is a scarcity of a well-annotated dataset representing different degrees of online violence.

Violence is rather a much-studied topic in social sciences, especially in Peace Studies[1] (Galtung, 1969). The term *violence* can be characterized by a broad spectrum - from a minimalist approach of an intentional act of excessive or detrimental force to an infringement of rights (Bufacchi, 2005; Mider, 2013). Preeminent author Galtung in his seminal work argued that violence inhibits

* Authors have equal contributions

[1] https://en.wikipedia.org/wiki/Peace_and_conflict_studies

individuals from realizing their full physical and mental potential, resulting in a gap between what could have been achieved and what actually transpires (Galtung, 1969). Recent studies show that *indirect* or *structural violence*, e.g. racism, sexism, heterosexism, xenophobia, and even elitism, can be observed more frequently on social media (Djuric et al., 2015). This kind of violence includes the use of political or economic power to commit violent acts or constrain/restrict an individual or a specific group of people. Even though those non-physical acts on social media seem unharmful, these activities related to structural violence more often than not translated to physical conflict in Southeast Asian societies (Mirchandani, 2018). Therefore, we focus on preparing a dataset on *violence incitement* by collecting online posts that perpetrated real-life violence across ethnic or communal space, including its detection in Bangla.

To the best of our knowledge, no existing research has developed a dataset for detecting Bangla text that incites violence, based on events leading to significant fatalities and extensive property damage. This paper contributes the following:

- A novel framework for annotating online communal violence-inciting comments in Bangla.
- A novel dataset of 6,046 annotated social media posts for detecting different forms of communal violence taking place online. We present one example for each class label in Table 1.
- Benchmark evaluation of the dataset using linguistic features, and pre-trained sentence transformer models.

2 Background and Motivation

Drawing from Galtung's foundation research work on peace and violence (Galtung, 1969), violence can be understood as any barrier that hinders individuals from reaching their maximum personal and cognitive development, creating a gap between their possible potential and their lived experiences. Numerous instances from our everyday lives can help to elucidate this concept. One particularly poignant incident from Bangladesh is the 2021 Cumilla Durga Puja violence that started with a Facebook post (Rahman, 2022). With a staggering 38,005 instances recorded, this event exemplifies how external forces, especially those fueled by socio-political conflicts and religious tensions, can inhibit the growth and well-being of numerous individuals. Such events not only disrupt the immediate safety and security of the people involved but also alter the course of their lives, casting a long shadow on their future prospects. In the subsequent sections, we will explore the ideas of both *peace and violence*, understanding their definitions, manifestations, and significance in our broader comprehension of societal life.

Bangla Comment	Peaceful Posts
এই রকম পোসিটিভ আলোচনা সত্যিই প্রয়োজন। সবাই একে অপরের সাথে সহযোগিতা এবং সহযোগিতা দিয়ে এগিয়ে যেতে পারে। (Positive discussions like this are truly needed. Everyone can move forward with cooperation and collaboration.)	Express support for peaceful discussion.
যতোটুকু আমরা সহিষ্ণুতা প্রদর্শন করি, ততটুকু আমরা সবাইকে একত্রে আনতে পারি। সত্যিই অসাধারণ প্রেরণা! (The more tolerance we show, the more we can bring everyone together. Truly inspiring!)	Express solidarity for empathy.
যখন আমরা কথা বলি, সেখানে আমরা সবচেয়ে বড় প্রভাব তৈরি করি। ধর্ষণ ছাড়াই সহমতি অনুসন্ধান। (When we communicate, we make the biggest impact. Seeking consensus without aggression.)	Supporting the need of dialogues.

Table 2: Bangla comments from YouTube videos expressing support for peaceful resolution in different scenarios.

2.1 Peace/Non-violence

In many discussions, the term *peace* is frequently invoked to lend support to various ideas, even when these ideas may not inherently contribute to harmony. Using the term *peace* in a broad and generalized manner to imply unity can sometimes obscure the underlying issues of conflict and suffering. As elaborated by Galtung in his seminal work on the subject (Galtung, 1969), a deeper and more nuanced understanding of *peace* is needed, one that transcends the simplistic notion of the absence of violence.

Peace encapsulates a condition of equilibrium and well-being in which individuals, communities, and nations coexist peacefully, fostering an environment of serenity, cooperation, and mutual respect. This deeper understanding of *peace* empowers individuals to engage in constructive dialogue, empathize with others, and seek non-violent resolutions to conflicts. Because of its dynamic nature, *peace* involves the pursuit of justice, equality, and social harmony, as well as the promotion of human rights and the rule of law (see Table 2). In such a context, *peace* becomes a catalyst for progress, development, and the betterment of humanity.

To truly harness the power of *peace* in discus-

sions and policy-making, it is crucial to understand that achieving *peace* is a fundamental human aspiration. It requires continuous efforts to address the root causes of conflicts, whether they be economic disparities, cultural misunderstandings, or political disputes. Thus *peace*, in particular, is not actually a passive state, but rather refers to an active endeavor that includes dialogue and negotiation to resolve conflicts through peaceful means.

2.2 Violence

Violence is not limited to physical harm or injury; a narrow interpretation of violence would inaccurately deem many harmful social constructs as benign. *Violence* manifests in various forms, each with distinct impacts on individuals and society (Roy et al., 2023). These forms can range from overt acts of aggression to more subtle forms of oppression, such as discrimination or systemic inequality (Galtung, 1990). We discuss two major categories of *violence* below:

2.2.1 Direct Violence

Historically, *direct violence* was primarily conceptualized as physical confrontations. However, with the digital revolution and the subsequent rise of social media, the definition of *direct violence* has broadened to include more covert and insidious forms of harm (Kaufhold and Reuter, 2019). *Direct violence* in the context of social media refers to any form of aggressive or harmful behavior that is explicitly targeted at an individual or group through online platforms. This type of *violence* is characterized by its overt and deliberate nature, as it involves direct actions or expressions aimed at causing harm, distress, or fear. Facebook, for example, played a crucial role in facilitating communication among political protesters during the Arab Spring (Kaufhold and Reuter, 2019). Both Facebook and Twitter (currently, **X**) are still being used by terrorists to spread extremist ideologies. While social bots are being used to skew social and political narratives by the nationalists and industrialists in their favor (Lazer et al., 2018).

Understanding how *direct violence* takes place in social media encompasses delving into both the means of harm and the depth of participant engagement. Table 3 presents some examples of direct somatic violence, categorizing its various forms according to their effects on human anatomy. From this table, we identify the *crushing* form of *violence* which involves the application of significant force on the body leading to injuries through pressure or impact, *piercing* form of *violence* refers to actions that penetrate skin and tissue leaving wounds often caused by tools like knives or bullets, and the *denial of movement* which encompasses both the physical restriction using barriers or devices like chains including the more intangible methods affecting the mind, such as brainwashing techniques to adopt radical beliefs by force.

Bangla Comment	**Somatic Direct Violence**
পুলিশ যে মানুষগুলোকে গুলি করে মারল এর বিচার করতে হবে। (The police who shoot and kill people must be held accountable.)	Piercing - by the means of shooting.
ছাত্র নামের এসব সন্ত্রাসীকে জেলে এনে রিমান্ডে ডিম থেরাপি দেওয়া হোক এবং নাহিদকে যারা পিটিয়ে মেরেছে তাদেরকে ক্রসফায়ারে হত্যা করা হোক। (Bring all these terrorists with student names to jail, give them egg therapy in remand, and let those who have beaten and killed Nahid be killed by crucifixion.)	Piercing, tearing, and crushing - by force and execution.
ইসলামে হিজাব বাধ্যতামুলক। হিজাব, নিকাব পড়তেই হবে। সেজন্য ইসলামি দেশগুলোতে হিজাব না পড়লে মেয়েদের কঠোর শাস্তি দেওয়া হয়। তো হিজাব স্বাধীনতা হয় কিভাবে? ইসলাম না জেনে কেবল কিছু মহাউস্মাদ মাথাপাগলরা একে স্বাধীনতা বলে। (In Islam, wearing a hijab is mandatory. Hijab and niqab must be worn. That's why in Islamic countries if women don't wear hijab, they are subjected to severe punishments. So, how is hijab freedom? People who know nothing about Islam just call it freedom.)	Denial of the movement of women in the name of Islam - by brainwashing techniques, i.e., forcing to adopt radical beliefs.

Table 3: YouTube video comments in Bangla offering a lens into public comments, reflecting the real-world implications of *direct somatic violence.*

2.2.2 Passive Violence

The increasing number of social media users has seen a corresponding uptick in various toxic behaviors. Hate speech, as highlighted by Silva et al. (2016) and Romim et al. (2022), has become a pervasive issue on these platforms. Online abuse, documented by Nobata et al. (2016) and Huang (2014), further showcases the extent of the problem. Beyond these individual-centered issues, there's also the concerning trend of social media platforms being exploited for extremist propaganda and terrorism.

Based on Galtung's research (Galtung, 1969), *passive violence* can be correlated to a concealed threat in our digital age. While we might not always witness overt acts of aggression, the rise in toxic behaviors in social media is a testament to this concept. The surge in hate speech and online abuse is an indicator of the underlying *passive violence*. Even if these toxic behaviors aren't always aggressive actions, they represent an unstable environment where harmful acts can quickly escalate.

One of the key features of *passive violence* is

Bangla Comment	Passive Violence
আরব দেশগুলোকে বলব ভারতের সাথে সব ব্যবসা বাণিজ্য বন্ধ করে দেন যারা হিন্দু ব্যবসায়ী আছে তাদের সাথে সব বন্ধ করে দেয়া উচিত। (I would tell the Arab countries to stop all trade with India, especially with the Hindu businessmen; it would be appropriate to sever ties with them.)	Express religious hate towards a nation.
ছাত্ররা বিভিন্ন অপমানের মুখোমুখি হয়; এটি গ্রহণ করা যাবে না। সমস্ত ছাত্র একত্রিত হওয়া এবং এই অত্যাচারের বিরুদ্ধে দাঁড়ানো উচিত। (The students face various insults; this cannot be accepted. All students should unite and stand against this atrocities.)	Instigating student protest leading to violent outcomes.
মালালা মুসলিমদের জন্য ভালো চায় না, আমি তাকে ঘৃণা করি। (Malala does not wish well for the Muslims, I despise her)	Expressing hate to Nobel Laureate Malala for her liberal activities.

Table 4: Bangla comments from YouTube videos related to various violent incidents that showcase passive violence.

its role in normalizing negative online behavior. When individuals passively accept or engage with harmful content or behaviors without objection, it sends a message that such behavior is acceptable, thereby perpetuating a cycle of toxicity. *Passive violence* often thrives in environments where individuals are not held accountable for their actions or silence. Inaction, indifference, or apathy can contribute to the persistence of online conflicts and harassment. Over time, *passive violence* can erode the overall culture of respect, empathy, and constructive dialogue on social media platforms. It can lead to polarization, division, and the silencing of marginalized voices. Table 4 presents some examples of passive violence in the context of Bangladesh.

3 Dataset Creation

3.1 Data Collection

We used YouTube platform to collect user posts, those expressing different forms of violence and also those urging for peaceful resolution, since it made the data easily accessible via the publically available YouTube API. [2] To prepare the dataset, we first cataloged the 9 violent communal incidents that originated from social media posts causing loss of lives and properties from 2012 to 2022 (Table 5). For all incidents, a set of 184 YouTube videos were selected manually based on the date of the video posts, their content in support of the violence, and the count of views. Then we used YouTube API to collect 168,232 comments from those video posts.

[2] https://developers.google.com/youtube/v3/docs/commentThreads/list

Event	Instances	Year
Ramu Incident	149	2012
Blogger Avijit Murder	8,624	2015
Nasirnagar Violence	1,052	2016
Election	14,181	2018, 2021, 2022
Political Clashes	12,491	2018, 2020, 2022
Hartal	5,288	2018, 2020, 2022
Cumilla Durga Puja Incident	38,005	2021
India Hijab Incident	57,437	2022
Dhaka College Vs New Market	31,005	2022
Total Instances	**168,232**	

Table 5: The table shows the number of comments collected from the YouTube videos related to various violent incidents that took place in Bangladesh in the last decade. For more details see Appendix A.1 and A.2.

3.2 Data Processing Pipeline

We detail the data processing pipeline using the methods of traditional topic modeling (Hong and Davison, 2010) for data pre-processing, content understanding, and related content filtering in three steps as discussed below. As social media comments for a video include discussions on many tangential issues, these steps deemed necessary to confirm that we will be able to select posts related to peace and violence in the context of Bangladesh. This pipeline is also depicted in the Figure 1.

- **Data Pre-processing and Deanonymization:** We removed all comments that included any code-mixed data, URLs, spam, or non-Bangla texts and removed comments that solely consisted of emojis without any accompanying text. Then we removed personally identifying information e.g., names, phone numbers, user mentions and addresses from the comments. This process left a total of 80,185 comments.

- **Unsupervised Topic Modeling for Content Understanding:** To understand main themes that are prevalent within this large collection of posts, we performed unsupervised topic modeling. We observed five major clusters of words based on the optimal coherence score. Following the work of Galtung (1990), we could map four of the clusters to Kill/Attack, Resocialization/Desocialization/Oppression, Passive Violence/Justification, and Peace/Non-violence. The fifth cluster of words contained terms like “demand”, “rights”, “protest”, “freedom”, etc., and thus we considered it to be the fifth topic for “Social Rights.”

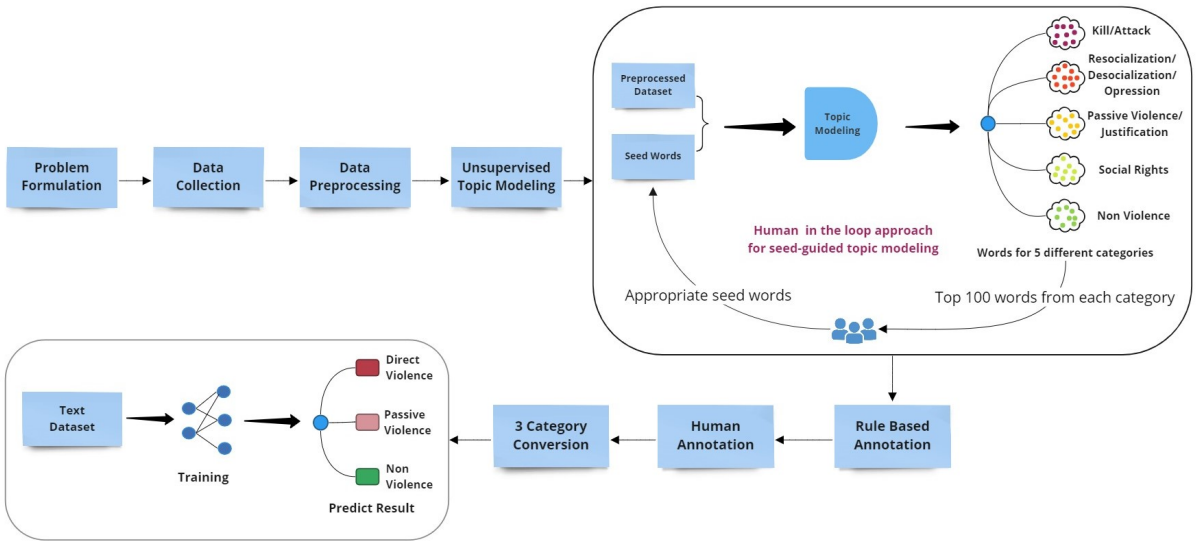

Figure 1: In this figure, we depict the workflow involving data pre-processing, content understanding by using unsupervised topic modeling, followed by the process of Guided LDA with human in the loop for content filtering, then annotation by human annotators, and finally dataset benchmarking.

- **Guided-LDA with Human in the Loop for Content Filtering:** We selected most relevant words for each of the five topics, and using those five sets of seed-words, we performed seed-guided semi-supervised topic modeling (Guided-LDA) on 80,185 data with two human experts in the loop to discover only the relevant terms for each topic (Tasnim et al., 2021). At the end of each iteration, we selected 100 top frequent terms from each cluster and then Both of our experts discussed and agreed to each term before its inclusion to extend the respective see word lists. Both the seed and final word lists for each categories are presented in the Table 11 of Appendix.

Finally, we used these extended seed word lists to filter out the posts that contained those specific seeds to select posts for each category with a higher chance (shown as rule-based annotation in the Figure 1). This process left us with 20,142 posts out of total 80,185 posts. We then randomly selected 6,046 comments for human annotation.

3.3 Data Labeling Framework

To create the framework for data labeling, we followed the research work of Anastasopoulos and Williams (2019) on violent protest and made all necessary changes related to our dataset. To keep the focus on creating a dataset for communal violence only, we selected a random sample of ∼100 posts for each of the nine events mentioned in Table 5 from the filtered 20,142 posts. In the next step, we manually checked and categorized each comment into five categories, four of which are as suggested by Galtung (1990), i.e. Kill/Attack, Resocialization/Desocialization/Deportation, Passive Violence/Justification, and Peace/Non-Violence, and the newly discovered fifth category for "Social Rights." Finally, we assessed each categorized post manually in a group of 3 persons to create an adaptive question-based framework to categorize any social media posts in the 5 categories defined earlier. We list the questions below:

- **Question 1:** Does the post call for or justify any form of violence against a person or community? Question 1 decides if the post represents any violence or not. For a positive response, we consult Question 2; otherwise, we consult Question 4.
- **Question 2:** Does the comment call for direct violence (Kill/Attack, Resocialization/Desocialization/Deportation) or rather indirect violence (Passive Violence/Justification)? For a positive response we consult Question 3; otherwise, the post is categorized as "Passive Violence/Justification" which is later used as a label.
- **Question 3:** Does the post reflect a call for Kill/Attack against a person or community? For a positive response, it's the "Direct Phys-

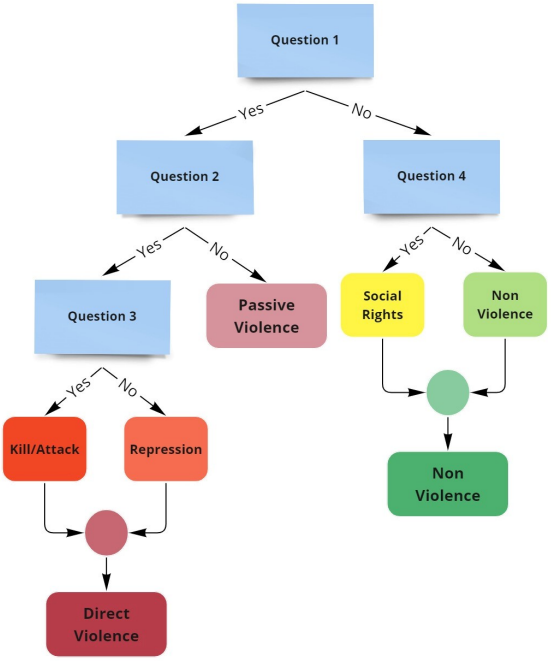

Figure 2: This figure illustrates a decision tree representing the adaptive framework employed for categorizing an online post. The decision process starts with one question at each level to help ramifications into a sub-tree based on types of violence. In this process an annotator has to answer at most 4 questions.

ical Violence;" Otherwise, it is "Repression." Later, both of these categories were merged to present "Direct Violence" label.

- **Question 4:** Does the post reflect the urge for any kind of social rights? For a positive response, we categorize the post in "Social Rights;" otherwise, the post is related to "Peace/Non-violence." These two categories were mereged into a single label "Non-violence" for labeling peaceful posts.

We present the adaptive question-based post-categorization framework as a decision tree in Figure 2. Through the application of adaptive questioning and the accompanying decision tree, our annotators could systematically categorize each comment, which we later aggregated into three classes: "Direct Violence" (by merging posts representing kill/attack and repression), "Passive Violence," and "Non-violence" (by merging posts referring to social rights and non-violence). We visualized the word clouds for each of the five categories of posts in the Appendix A.5 and also provided a few examples of using the proposed framework for categorizing/labeling an online post in Table 12 of the Appendix.

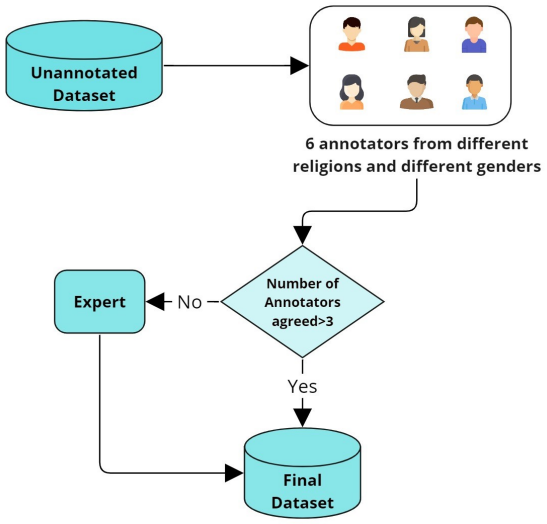

Figure 3: Dataset labeling process by six annotators.

3.4 Data Annotation

As the posts in our dataset are very sensitive for different genders, races, and ethnic communities, we had to employ a diverse set of data annotators to avoid any in-group biases during the annotation. We trained 6 annotators from different gender (2 females, 4 males), religious (3 Muslims and 3 Hindus), and political backgrounds (2 liberals, 2 conservatives, and 2 centrists) on the proposed framework to categorize any social media post into one of 5 categories and then subsequently into 3 labels as discussed in the previous section. After the annotation, one expert validated the annotated data with major disagreements (i.e. $agreements \leq 3$). Our six annotators labeled 6,046 samples independently using the proposed framework to categorize and label the data. The inter-annotator agreement (Fleiss-Kappa) is 0.7040, indicating a substantial agreement between them. We found that more than 3 annotators disagreed on 365 data, which is $\tilde{6}\%$ of our total samples. To resolve this disagreement, an expert was employed to arbitrate the final decision. We discuss each of the data labels below:

- **Direct Violence:** Direct violence is the combination of the Kill/Attack and Resocialization/Desocialization/Deportation category. This category encompasses explicit threats directed towards individuals or communities, including actions such as killing, rape, vandalism, deportation, desocialization (threats urging individuals or communities to abandon their religion, culture, or traditions), and re-socialization (threats of forceful conversion).

Model Name	Direct			Passive			Non-Violence			Macro		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Random Baseline	0.3302	0.3352	0.3132	0.3302	0.3352	0.3132	0.5183	0.3369	0.4084	0.3302	0.3352	0.3132
Majority Voting	0.1812	0.3333	0.2348	0.1812	0.3333	0.2348	0.1812	0.3333	0.2348	0.1812	0.3333	0.2335
Unigram (U)	0.6571	0.4577	0.5396	0.7422	0.4645	0.5714	0.6942	0.9033	0.7851	0.6979	0.6085	0.6320
Bigram (B)	0.7778	0.1045	0.1842	0.6310	0.1474	0.2390	0.5744	0.9544	0.7172	0.6610	0.4021	0.3801
Trigram (T)	0.0000	0.0000	0.0000	0.4138	0.0334	0.0618	0.5475	**0.9781**	0.7020	0.3204	0.3372	0.2546
U+B	0.6555	0.3881	0.4875	0.7487	0.4061	0.5266	0.6656	0.9151	0.7706	0.6899	0.5698	0.5949
B+T	0.6593	0.9215	**0.7686**	0.7533	0.3950	0.5182	0.6262	0.3333	0.4351	0.6796	0.5500	0.5740
U+B+T	0.5682	**0.9653**	0.7153	0.6493	0.1210	0.2040	0.7500	0.0746	0.1357	0.6558	0.3870	0.3517
Char-1-gram (C1)	0.4595	0.1692	0.2473	0.6152	0.3380	0.4363	0.6257	0.8832	0.7325	0.5668	0.4634	0.4720
Char-2-gram (C2)	0.6241	0.4378	0.5146	0.7133	0.4534	0.5544	0.6883	0.8905	0.7765	0.6753	0.5939	0.6152
Char-3-gram (C3)	0.6923	0.4478	0.5438	0.7473	0.4729	0.5792	0.7016	0.9161	0.7946	0.7137	0.6122	0.6392
Char-4-gram (C4)	0.7615	0.4129	0.5355	0.7724	0.4437	0.5636	0.6841	0.9325	0.7892	0.7393	0.5964	0.6294
Char-5-gram (C5)	0.8171	0.3333	0.4735	0.7892	0.4061	0.5363	0.6650	0.9489	0.7820	**0.7571**	0.5628	0.5972
Char-6-gram (C6)	**0.8226**	0.2537	0.3878	0.7877	0.3561	0.4904	0.6458	0.9599	0.7721	0.7520	0.5232	0.5501
C2+C3	0.6884	0.4726	0.5605	0.7473	0.4854	0.5885	0.7073	0.9106	0.7962	0.7143	0.6229	0.6484
C2+C3+C4	0.7143	0.4478	0.5505	0.7646	0.4743	0.5854	0.7015	0.9243	0.7976	0.7268	0.6154	0.6445
C2+C3+C4+C5	0.7391	0.4229	0.5380	0.7664	0.4701	0.5828	0.6966	0.9279	0.7958	0.7340	0.6070	0.6388
C2+C3+C4+C5+C6	0.7706	0.4179	0.5419	0.7778	0.4673	0.5838	0.6929	0.9325	0.7950	0.7471	0.6059	0.6403
Multilingual Bert (MBERT)	0.4835	0.6567	0.5570	**0.8091**	0.4186	0.5518	0.7104	0.8887	0.7896	0.6677	0.6547	0.6328
Xlm-RoBERTa (Base)	0.4568	0.7363	0.5638	0.7899	0.5021	0.6139	0.7587	0.8549	0.8039	0.6685	0.6978	0.6606
DistilBERT	0.3735	0.6169	0.4653	0.6813	0.4965	0.5744	0.7353	0.7783	0.7562	0.5967	0.6306	0.5986
BanglaBERT	0.4669	0.8408	0.6004	0.7968	**0.6273**	**0.7019**	**0.8327**	0.8266	**0.8297**	0.6988	**0.7649**	**0.7107**

Table 6: The table shows the outcomes classification using baselines, linguistic features, and pre-trained language models for the test set. All the experiments used the same dataset and parameters for a fair evaluation. We observe that BanglaBERT achieved the best F1-score for most of the individual classes and overall dataset.

- **Passive Violence:** In this category, instances of violence are represented by the use of derogatory language, abusive remarks, or slang targeting individuals or communities. Additionally, any form of justification for violence is also classified under this category.

- **Non-Violence:** The contents falling under this category pertain to non-violent subjects, such as discussions about social rights or general conversational topics in support of lawful activities that do not involve any form of violence.

This led to the creation of our final annotated "Vio-Lens" dataset.

3.5 Data Statistics

In our dataset, about 7.78% of posts are related to the Kill/Attack category, while 5.19% are related to Resocialization/Desocialization/ Deportation/Repression/Oppression category. Both of these categories together constitute "Direct Violence" class, accounting for approximately 13% of the dataset. About 34.04% of posts are related to "Passive Violence" class. From the rest of the data, 12.84% represents Social Rights, and 40.12% belongs to Peace/Non-violence. When these two categories are combined, 52.96% of the datataset falls into "Non-violence" class. The details statistics about Direct, Passive, and Non-violence are provided in table 7.

	Direct Violence	Passive Violence	Non-Violence	Total
Train	389	922	1389	**2700**
Dev	196	417	717	**1330**
Test	201	719	1096	**2016**
Total	**786**	**2058**	**3202**	**6046**

Table 7: Statistics of the online posts in the Train, Dev, and Test dataset.

4 Baseline Creation

To establish a violence detection benchmark we explore three different types of modeling techniques in comparison to the baseline method. We discuss the evaluation methods below:

- **Baselines:** We defined two baselines for our work: 1) random baseline and 2) majority baseline.
- **Linguistic Features:** For each post, we extracted word n-grams (n=1, 2, 3), and character n-grams (n=2, 3, 4, 5, 6). We then trained SVMs for classification tasks.
- **Pre-trained Language Models:** We employed three different sentence transformer models, such as Multilingual BERT

(MBERT) [3] (Devlin et al., 2019), DistillBert(Sanh et al., 2019)[4] and XLM-RoBERTa (Liu et al., 2019)[5], and monolingual BanglaBERT (Bhattacharjee et al., 2022)[6] . We used Hugging Face transformers (Wolf et al., 2019) to finetune the models on our dataset.

5 Experiments and Results

We split our dataset into the train set (2700 samples or $\tilde{4}5\%$), the dev set (1330 samples or $\tilde{2}2$ %), and the test set (2016 samples or $\tilde{3}3\%$) so that nearly 2/3rd of the data is provided for both train set and dev set and the rest 1/3rd of the data is provided for the test set to ensure a good number of data is available for test set prediction. We applied Hugging Face Transformers (Wolf et al., 2019), Skilit-Learn Tool (Pedregosa et al., 2011), and the PyTorch Framework (Paszke et al., 2019) to carry out our studies. The configurations for the models are discussed in the Suppl. Table 10 and dev set results can be found in the Suppl. Table 13.

We present the test set results of our experiments in Table 6, highlighting the best performance both for individual classes and whole classes. Most of the models perform significantly worse in predicting two types of violence: direct and passive violence while overperforming in the Non-violence category. Among all the experiments, BanglaBERT (Bhattacharjee et al., 2022) showed the best performance with macro F1 scores of 0.71 for the test set.

Error Analysis: For the *Direct Violence* category, out of 201 test instances, 84.08% was predicted correctly, while 4.48% misidentified as *Passive Violence*, and 11.44% were misclassified as *Non-Violence*. The *Passive Violence* test set comprises 719 samples. Of those, 62.73% were correctly classified, while 15.16% were befuddled with *Direct Violence*, with the rest erroneously categorizing it under *Non-Violence*. For the *Non-Violence* category, which had 1,096 samples in the test set, an impressive 82.66% were correctly categorized by all the teams. A minor 7.66% samples were incorrectly identified as *Direct Violence*, with the remaining misclassified as *Passive Violence*. More details can be found in Figure 4. Thus, it can

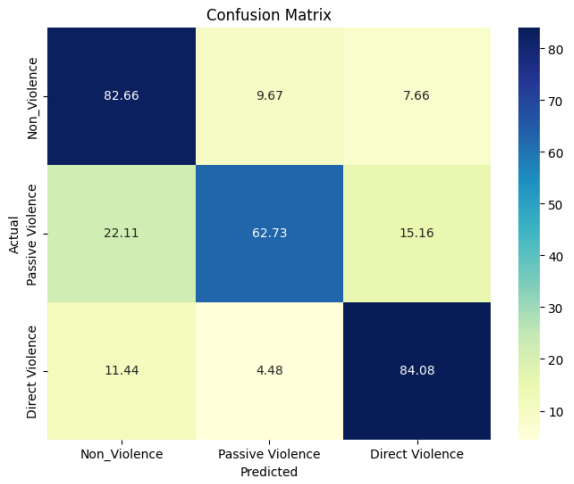

Figure 4: Confusion matrix illustrating category distribution predicted by best performing BanglaBERT model. In this representation, the columns depict the predicted label percentages for each classification type (rows)

be inferred from the confusion matrix that the best performing BanglaBERT although correctly classified *Direct Violence* and *Non-Violence* most of the time, has trouble predicting Passive Violence with a significant number of samples overlapped with both Direct-Violence and Non-Violence.

6 Conclusion

In this paper, we propose Vio-Lens, the first-ever dataset and adaptive categorization procedure of communal violence. Through our investigation, we find that BanglaBERT (Bhattacharjee et al., 2021) performs better for our case. We find that BanglaBERT performs the best with an F1 score of 71.07. The dataset and annotation is only applied to the Bangla language and incidents and source are limited to the region Bangladesh and West Bengal of India. Therefore, a good direction for our future work will be to gather violence-related data from different regions and different languages and create a baseline from that multilingual dataset. We would also like to expand towards a real-time violence detection model.

Limitations

The study has some potential limitations. One of the potential limitations is that our dataset is comprised of informal data from social media which is usually very noisy and contains misspellings, and slang words creating challenges to the machine learning model. Moreover, our dataset consists of

[3] huggingface.co/bert-base-multilingual-cased
[4] huggingface.co/distilbert-base-multilingual-cased
[5] huggingface.co/xlm-roberta-base
[6] huggingface.co/csebuetnlp/banglabert

roughly 6K and from specific regions data leaving the scope for extension of the dataset in the future across multiple languages and regions.

Ethical Considerations

Dataset Release The Copy Right Act. 200015 of The People's Republic of Bangladesh allows copyright material reproduction and public release for non-commercial research proposals. We will release our Vio-Lens dataset under a non-commercial license. Publicizing other supplementary materials like codes won't cause any copyright infringements.

Violent Content: The dataset contains different kinds of threats, attacks, and vulgar and derogatory comments against persons, communities, religions, and nations.

Annonators Compensation All the annotators' and experts were paid for their service according to the standard laws of the local market.

Quality Assurance of the Dataset All the annotations were done by native Bangla speakers. The Fleiss Kappa score of our dataset showed very substantial agreement, ensuring the quality of our dataset. To further ensure the quality the annotators were taken from diverse races and gender and an expert resolved the disagreements.

References

Lefteris Jason Anastasopoulos and Jake Ryland Williams. 2019. A scalable machine learning approach for measuring violent and peaceful forms of political protest participation with social media data. *Plos one*, 14(3):e0212834.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad Uddin, Kazi Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL*.

Vittorio Bufacchi. 2005. Two concepts of violence. *Political studies review*, 3(2):193–204.

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

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 2930, New York, NY, USA. Association for Computing Machinery.

Johan Galtung. 1969. Violence, peace, and peace research. *Journal of peace research*, 6(3):167–191.

Johan Galtung. 1990. Cultural violence. *Journal of peace research*, 27(3):291–305.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789*.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, SAM '14, page 36, New York, NY, USA. Association for Computing Machinery.

Ting-Hao Kenneth Huang. 2014. Social metaphor detection via topical analysis. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Marc-André Kaufhold and Christian Reuter. 2019. Cultural violence and peace in social media. *Information Technology for Peace and Security: IT Applications and Infrastructures in Conflicts, Crises, War, and Peace*, pages 361–381.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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 preprint arXiv:1907.11692*.

Daniel Mider. 2013. The anatomy of violence: A study of the literature. *Aggression and Violent Behavior*, 18(6):702–708.

Maya Mirchandani. 2018. Digital hatred, real violence: Majoritarian radicalisation and social media in india. *ORF Occasional Paper*, 167:1–30.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Mohammad Javed Kaisar Ibne Rahman. 2022. Religious nationalism in digitalscape: An analysis of the post-shahbag movement in bangladesh. *Open Journal of Social Sciences*, 10(5):201–218.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. *arXiv preprint arXiv:2206.00372*.

Sajal Roy, Ashish Kumar Singh, et al. 2023. Sociological perspectives of social media, rumors, and attacks on minorities: Evidence from bangladesh. *Frontiers in Sociology*, 8:1067726.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media*.

Nazia Tasnim, Md Istiak Hossain Shihab, Moqsadur Rahman, Sheikh Rabiul Islam, and Mohammad Ruhul Amin. 2021. Exploring the scope and potential of local newspaper-based dengue surveillance in bangladesh. *arXiv preprint arXiv:2107.14095*.

Nicole Tausch, Miles Hewstone, and Ravneeta Roy. 2009. The relationships between contact, status and prejudice: An integrated threat theory analysis of hindu–muslim relations in india. *Journal of Community & Applied Social Psychology*, 19(2):83–94.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendices

A.1 Events

The specific sources containing the description of violent incidents are detailed in the Table 8.

Event	Sources
Cumilla Durga Puja Incident	https://w.wiki/4Eti
India Hijab Incident	https://w.wiki/6FAb
Ramu Incident	https://w.wiki/6FAd
Blogger Avijit Murder	https://w.wiki/6FAf
Nasirnagar Violence	https://w.wiki/6FAh
Dhaka College Vs New Market	https://www.thedailystar.net/

Table 8: This table provides different sources containing the description of violent incidents based on which the proposed dataset was created.

A.2 Sources

We have analyzed data pertaining to online comments on popular YouTube news channels from Bangladesh and India. The specific number of comments collected from each channel is presented in Table 9.

Source	Number of Instances
Somoy Tv	28,241
Ekattor Tv	10,114
Independent Television	18,333
BBC News Bangla	14,339
ATN News	1,759
RTV News	1,717
Jamuna TV	64,853
India Today	3,314
Hindustan Times	16,922
Republic World	8,266
Zee 24 Ghanta	374
Total Instances	**168,232**

Table 9: The table presents the number of comments collected from various YouTube news channels that broadcasted videos on the violent incidents cited in this paper.

A.3 Model Hyperparameter

We have fine-tuned the pre-trained language model using a set of hyperparameter values. These values are presented in Table 10

Hyperparameter	Value
learning rate	1e-5
train batch size	8
evaluation batch size	8
epochs	50
evaluation steps	250
early stopping patience	5

Table 10: The table depicts the hyperparameter of the fine-tuned pre-trained language model

Figure 5: Social Rights

A.4 LDA Seeds

This section contains some final seeds used in the LDA which are provided in Table 11.

A.5 Word Cloud

In order to gain insight and potentially discover useful information, a word cloud analysis was conducted on each incident which are provided in Figure 5 to 9

Figure 6: Non-violence

Classification	Words
Kill/Attack	**Seed Word List:** হামলা, ভাঙচুর, হত্যা, মারা, আঘাত, ভাংচুর, ধ্বংস, দাংগা, মারার, যুদ্ধ, ভেঙে **English Translation:** Assault, vandalize, kill, kill, hurt, vandalize, destroy, riot, kill, fight, break
	Extended Word List:হত্যা, ধ্বংস , মারা, মেরে, খুন, যুদ্ধ, ধর্ষণ, রক্ত, জিহাদ, হামলা, সংঘাত, কেটে, বোমা, ধর্ষন, , যুদ্ধের, আক্রমণের, জিহাদেরধর্ষণের , জ্বালিয়ে, গণধর্ষণ, সংঘর্ষের , পুড়িয়ে, ভেংগে , ভাঙ , ধোলাই, গর্দান, গজব **English Translation:** Kill, Destroy, Kill, Kill, Kill, Fight, Rape, Blood, Jihad, Attack, Conflict, Cut, Bomb, Rape, Battle, Attack, Jihad, Rape, Burn, Gang Rape, Clash, Burn, Break, Break, Hit, Neck Chopping, Curse
Re/Desocialization/ Repression/ Oppression/ Deportation	**Seed Word List:**অত্যাচার, নির্যাতন, অন্যায়, জোর, জুলুম,,নির্যাতন,গ্রেফতার **English Translation:** Torture, torture, injustice, force, oppression, torture, arrest
	Extended Word List:বন্ধ ,ভয়, নির্যাতন , বয়কট , তান্ডব , অন্যায়ের , চাপিয়ে , শোষণ , চাপানোর , ক্রিতদাস , কৃতদাস , আটক , বাইন্ধা , বেঁধে, বর্বরোচিত, নির্যাতনের, হুমকি **English Translation:** Stop, Fear, Torture, Boycott, Tyranny, Injustice, Impose, Exploit, Oppress, Enslaved
Passive Violence/ Justification	**Seed Word List:**গুজব, নোংরামি, উচিত, জঙ্গি, নাস্তিক, উগ্রবাদী, জায়েজ, দালাল, দালালি, অবমাননা **English Translation:** Rumour, Filth, Should, Militant, Atheist, Extremist, Legitimate, Broker, Broker, Contempt
	Extended Word List: নোংরামি , অবমাননা , দালাল , পাগল , গুজব , মিথ্যা , বাজে , চোর, নোংরা , কাফের , সন্ত্রাসীদের , দায় , বাটপার , সাম্প্রদায়িকতা , উস্কানি , ব্যভিচারের ,জঙ্গিদের , জালিম , রাজাকার , ধামাচাপা , চামচা , কটাক্ষ , জালেম , কাফির , দালালরা , কুলাঙ্গারদের , উগ্রবাদীদের , বেহায়া , কুলাঙ্গাররাই **English Translation:** Scumbag, Scumbag, Scumbag, Lunatic, Rumour, Liar, Rascal, Thief, Filth, Infidel, Terrorists, Liability, Fraud, Communalism, Incitement, Adultery, Militants, Oppressors, Extortionists, Cover-up, Scoundrels, Slanderers, Tyrants, Infidels, Brokers, Curmudgeons, Extremists, Perverts, Curmudgeons
Social Rights	**Seed Word List:**প্রতিবাদ, অধিকার, স্বাধীনতা, দাবি, বিচার, আন্দোলন, স্বাধীন, মিছিল **English Translation:** Protest, Rights, Freedom, Demand, Trial, Movement, Independent, March
	Extended Word List:বিচার, স্বাধীনতা, অধিকার, আন্দোলন , স্বাধীন, তদন্ত, সমর্থন , বিচারের, নিরাপত্তা , গ্রেফতার , অভিযোগ , মিছিল , প্রতিরোধ , প্রতিবাদী ,আন্দোলনের , আন্দোলনে, গ্রেপ্তার, মর্যাদা, মানবাধিকার, জাগ্রত, গনতন্ত্র, হরতালে, বিক্ষোভ, চেতনা, আইনি, জবাবদিহি **English Translation:** Trial, Freedom, Rights, Movement, Independent, Investigation, Support, Trial, Security, Arrest, Complaint, March, Resistance, Protest, Movement, Agitation, Arrest, Dignity, Human Rights, Vigil, Democracy, Strike, Protest, Consciousness, Legal, Accountability
Non-Violence	**Seed Word List:**ধন্যবাদ, সম্মান, শান্তি, সৃষ্টি, সুন্দর, জন্ম, **English Translation:** Thanks, Honor, Peace, Creation, Beautiful, Birth
	Extended Word List:শিক্ষা, ধন্যবাদ, পবিত্র , সৃষ্টি, রক্ষা , সুন্দর, ভাল, জন্ম, আশা, চিন্তা, খুশি , একমত , প্রিয়, নিরপেক্ষ , পছন্দ, দুঃখজনক , শান্তিতে , মানবতা , সুযোগটাও , নিরপেক্ষতা, ভাই, সুস্থ , কল্যাণ , সত্যটা , আশ্রয় , রক্ষার , ভদ্র , গর্ব , সৌন্দর্য **English Translation:** Teaching, Thanking, Holy, Creating, Protecting, Beautiful, Good, Born, Hoping, Thinking, Happy, Agree, Dear, Neutral, Like, Sad, At Peace, Humanity, Opportunity, Impartiality, Brother, Health, Welfare, Truth, Shelter, Protection, Polite, Pride, Beauty

Table 11: The table presents each seed word list followed by respective final word list extended by Guided LDA with human in the loop for five different categories: Kill/Attack, Resocialization/Desocialization/Oppression/Deportation, Passive Violence, Social Justice and Peace/Non-Violence.

Bangla Comment	Question 1	Question 2	Question 3	Question 4	Label
ছাত্রদের আন্দোলন সঠিক,ব্যবসায়ীদেরকে কঠিন শাস্তি দেওয়া হোক,মারো আরো জোরে মারো ব্যবসায়ী মাগির পোলারা ডাকাত,মার আরো জোরে মার (The students' movement is correct; the businessmen should be severely punished. Beat them harder; beat those corrupted businessmen)	yes	yes	yes	-	Direct Violence
পূজা মণ্ডপে হামলা করার উদ্দেশ্যে বা পূজা উৎসবকে বানচাল করার উদ্দেশ্যে পরিকল্পিতভাবে এই কাজটি করা হয়েছে।বিরোধী দলগুলো কোন ইসু খুঁজে পাচ্ছে না সরকারকে ঘায়েল করার জন্য তাই জনগণকে ধর্মীয় সুরসুরি দিয়ে ঘোলা পানিতে মাছ শিকার করা যায় কিনা (This act was done deliberately to attack the puja pandal or to disrupt the puja festival. Opposition parties can't find any issues to blame the government)	yes	no	-	-	Passive Violence
নিউমার্কেটে দোকানের করমচারীরা মেয়েদের ইভটিজিং প্রতিদিনের ঘটনা এর আগেও ছাত্র/ছাত্রী দের সাথে এমন হয়েছে শক্ত স্টেপ না নিলে এসব দোকানের করমচারীদের সন্ত্রাসী মূলক কারজকলাপ বন্ধ হবে (Every day in New Market, the shop employees are eve-teasing the girls. This has happened with students before. If strict measures are not taken, these shop employees will continue their terrorist activities)	-	-	-	yes	Non-Violence
একদম মেরে ছাত্রদের হাড্ডি গুড়া করে দে।এরা ছাত্র না এরা আগামী দিনের সন্ত্রাস (Completely break the students' bones. They are not students; they are terrorists of the future)	yes	yes	yes	-	Direct Violence
এটা কোনো কথা হোলো সবাই দেখেছে কি হোয়েছে আর তোমরা বলছো গুজব আমার মনে হয় কোরআনের সব চেয়ে বড় শত্রু তোমরা (Is this a joke? Many people have seen what happened, and you are saying it's a rumor. I believe the biggest enemies of the Quran are people like you)	yes	no	-	-	Passive Violence
পুলির= চোর। এরা ব্যবসায়িদের পক্ষকেই বেছে নিবে। কারন ব্যবসায়িরা তো টাকা দিবে। ছাত্ররাতো আর টাকা দিতে পারবেনা। (Police = Thieves. They will always favor businessmen because businessmen will give money. Students, on the other hand, won't be able to)	-	-	-	no	Non-Violence

Table 12: The table displays Bangla comments from YouTube videos pertaining to various incidents, along with their labels determined by answers to four specific questions as presented in the data annotation framework. The decision process starts with one question at each level, leading to ramifications into a sub-tree based on types of violence.

Figure 7: Kill/Attack

Figure 8: Resocialization, Deportation or Opression

Figure 9: Passive Violence

Model Name	Direct			Passive			Non-Violence			Macro		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Random Baseline	0.1395	0.3162	0.1935	0.3435	0.3416	0.3426	0.4983	0.3233	0.3921	0.3271	0.3270	0.3094
Majority Voting	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5391	1.0000	0.7005	0.1797	0.3333	0.2335
Unigram (U)	0.7159	0.3214	0.4437	0.7087	0.5659	0.6293	0.6942	0.8801	0.7761	0.7063	0.5891	0.6164
Bigram (B)	0.7692	0.1020	0.1802	0.6232	0.2062	0.3099	0.5823	0.9470	0.7212	0.6583	0.4184	0.4038
Trigram (T)	0.6667	0.0102	0.0201	0.4364	0.0576	0.1017	0.5472	**0.9707**	0.6998	0.5501	0.3462	0.2739
U+B	0.7162	0.2704	0.3926	0.6896	0.5540	0.6144	0.6851	0.8801	0.7705	0.6970	0.5681	0.5925
B+T	0.8667	0.0663	0.1232	0.6404	0.1751	0.2750	0.5754	0.9637	0.7205	0.6941	0.4017	0.3729
U+B+T	0.7042	0.2551	0.3745	0.7006	0.5276	0.6019	0.6772	0.8926	0.7702	0.6940	0.5584	0.5822
Char-1-gram (C1)	0.5667	0.1735	0.2656	0.5917	0.5108	0.5483	0.6604	0.8382	0.7388	0.6063	0.5075	0.5176
Char-2-gram (C2)	0.7778	0.3571	0.4895	0.6554	0.6067	0.6301	0.7248	0.8633	0.7880	0.7193	0.6091	0.6359
Char-3-gram (C3)	0.8090	0.3673	0.5053	0.7046	0.6235	0.6616	0.7317	0.8898	0.8030	0.7484	0.6269	0.6566
Char-4-gram (C4)	0.8182	0.3214	0.4615	0.7478	0.6187	0.6772	0.7159	0.9066	0.8000	0.7606	0.6156	0.6462
Char-5-gram (C5)	**0.8519**	0.2347	0.3680	0.7219	0.5540	0.6269	0.6851	0.9135	0.7830	0.7530	0.5674	0.5926
Char-6-gram (C6)	0.8478	0.1990	0.3223	0.7355	0.4868	0.5859	0.6637	0.9331	0.7757	0.7490	0.5396	0.5613
C2+C3	0.7789	0.3776	0.5086	0.7008	0.6235	0.6599	0.7326	0.8828	0.8008	0.7375	0.6280	0.6564
C2+C3+C4	0.8353	0.3622	0.5053	0.7216	0.6403	0.6785	0.7349	0.8968	0.8078	0.7639	0.6331	0.6639
C2+C3+C4+C5	0.8182	0.3214	0.4615	0.7228	0.6379	0.6777	0.7299	0.9010	0.8065	0.7570	0.6201	0.6486
C2+C3+C4+C5+C6	0.8219	0.3061	0.4461	0.7210	0.6259	0.6701	0.7263	0.9066	0.8065	0.7564	0.6129	0.6409
Multilingual Bert (MBERT)	0.6752	0.5408	0.6006	0.7331	0.5731	0.6433	0.7400	0.8745	0.8018	0.7162	0.6628	0.6819
Xlm-RoBERTa (Base)	0.6882	0.6531	0.6702	0.7241	0.6859	0.7044	0.7957	0.8312	0.8131	0.7360	0.7234	0.7292
DistilBERT	0.5455	0.5510	0.5482	0.6300	0.6451	0.6374	0.7773	0.7643	0.7707	0.6509	0.6535	0.6521
BanglaBERT	0.7577	**0.7500**	**0.7538**	**0.7449**	**0.7842**	**0.7640**	**0.8580**	0.8340	**0.8458**	**0.7869**	**0.7894**	**0.7879**

Table 13: The table shows the outcomes classification using baselines, linguistic features, and pre-trained language models for the development set. All the experiments used the same dataset and parameters for a fair evaluation. We observe that BanglaBERT achieved the best F1-score for most of the individual classes and overall dataset.