# Modelling Political Aggression on Social Media Platforms

**Akash Rawat[1], Nazia Nafis[1], Dnyaneshwar Bhadane[1],**
**Diptesh Kanojia[1,2], Rudra Murthy[1,3]**
[1]Indian Institute of Information Technology Lucknow, India.
[2]Surrey Institute for People-Centred AI, University of Surrey, United Kingdom.
[3]IBM Research, Bangalore, India.
[1]{mcs21014,mcs21004,mcs21010}@iiitl.ac.in
[2]d.kanojia@surrey.ac.uk, [3]rmurthyv@in.ibm.com

## Abstract

Recent years have seen a proliferation of aggressive social media posts, often wreaking even real-world consequences for victims. Aggressive behaviour on social media is especially evident during important sociopolitical events such as elections, communal incidents, and public protests. In this paper, we introduce a dataset in English to model political aggression[1]. The dataset comprises public tweets collated across the time-frames of two of the most recent Indian general elections. We manually annotate this data for the task of aggression detection and analyze this data for aggressive behaviour. To benchmark the efficacy of our dataset, we perform experiments by fine-tuning pre-trained language models and comparing the results with models trained on an existing but general domain dataset. Our models consistently outperform the models trained on existing data. Our best model achieves a macro F1-score of 66.66 on our dataset. We also train models on a combined version of both datasets, achieving the best macro F1-score of 92.77, on our dataset. Additionally, we create subsets of code-mixed and non-code-mixed data from the combined dataset to observe variations in results due to the Hindi-English code-mixing phenomenon. We publicly release the anonymized data, code, and models for further research.

## 1 Introduction

In recent years, social media has risen as one of the most popular ways in which people share opinions with each other (Pelicon et al., 2019). On such platforms, anonymity is a major factor that impacts user behavior (Bernstein et al., 2011; Postmes et al., 1998), and the possibility of posting anonymously on platforms such as Twitter and Reddit has changed the way people communicate (Décieux et al., 2019). This has given rise to a significant amount of aggressive behavior- including but not limited to the use of snide remarks, abusive words, and personal attacks, going as far as rape threats (Hardaker and McGlashan, 2016). Modern definitions of human aggression establish it as *any behavior enacted with the intention of harming another person who is motivated to avoid that harm* (Anderson et al., 2002; Bushman and Huesmann, 2014). Aggression is now defined as social behavior patterns, and several studies have noted the proliferation of abusive language and an increase in aggressive content on social media (Mantilla, 2013; Suzor et al., 2019). Such behavior begets the automated analysis of social media content for aggressive behavior, lying at the intersection of Natural Language Processing (NLP) and Computational Social Sciences (CSS).

NLP research community has proposed various tasks to analyze aggressive behavior, some of which are well-known, *viz.,* offensive language identification (Zampieri et al., 2020), hate speech detection (Warner and Hirschberg, 2012; MacAvaney et al., 2019; Paz et al., 2020), aggression detection (Kumar et al., 2018b,a), cyber bullying (Dadvar et al., 2013; Kumar et al., 2020), and so on. Various shared tasks have been organized for these NLP sub-areas, motivating us to investigate the phenomenon of aggression on social media. Aggression is displayed not only by unnamed and anonymous troll accounts but, on occasion, by known personalities who can influence thousands of followers (O'Toole et al., 2014). However, we investigate this problem in the context of political trolling and aggression displayed on social media close to the government election.

In this paper, we investigate the task of aggression detection on a social media platform, *i.e.,* Twitter, in the context of Indian elections. We curate a set of political-themed tweets from the user handles of known personalities ($\sim$ 110 in number) and perform manual annotation to create a dataset for the task. Our annotation schema aligns with

[1] https://doi.org/10.5281/zenodo.7540489

*Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 497–510
July 14, 2023 ©2023 Association for Computational Linguistics

|  |  |
|---|---|
| **OAG** | He will kill 22000, will abolish NREGA, nullify food security. NO control on the present. [MASK] baba derives satisfaction in being astro-baba. |
| **CAG** | Also at 9pm: Did you know our Parliament has a record number of MPs facing criminal cases? What does that tell you about our democracy? |
| **NAG** | We wont be detrimental to the development. We are partners in development & Progress. |

Table 1: Examples of Overtly, Covertly, and Non-aggressive tweets from our dataset. [MASK] token is to avoid naming an individual in this example.

the existing aggression detection datasets where text sequences are labelled as overtly-aggressive, covertly-aggressive and non-aggressive as shown in Table 1 with an example for each class. We also collected the datasets released at TRAC-2018 (Kumar et al., 2018a) and TRAC-2020 (Kumar et al., 2020) shared tasks to benchmark task performance. With the help of pre-trained language models, we perform a topic analysis along with various experiments to perform the task of aggression detection and discuss the obtained results in terms of precision, recall, and macro F1 scores. We also perform transfer learning-based experiments to observe the cross-dataset performance.

While these datasets are mostly in the Latin script, many words belong to one of the Indian languages, such as Hindi but are transliterated into the Latin script. This led us to label our data instances as code-mixed *vs.* non-code-mixed using a known heuristics-based approach, and we performed additional experiments on these data sub-sets. Our **key contributions** are:

- We release an English tweet dataset to model political aggression along with our code and models[2].
- Experimental analysis of aggressive behavior with multiple subsets of our dataset.
- Evaluation of task performance using language models, including observations over the presence of Hindi-English code-mixing.

## 2 Related Work

The earliest approaches to the task of classifying derogatory messages used decision trees (Spertus, 1997). Manual rules with syntactic and semantic text features were the basis of these models. Since then, much of the focus has been on feature engineering the text which includes features like Bag-of-Words (BOW) (Kwok and Wang, 2013; Liu et al., 2019a), N-grams in the word level (Pérez and Luque, 2019; Liu and Forss, 2014; Watanabe et al., 2018), N-grams in character level (Gambäck and Sikdar, 2017; Pérez and Luque, 2019), typed dependencies (Burnap and Williams, 2016), part-of-speech tags (Davidson et al., 2017), dictionary-based approaches (Tulkens et al., 2016) and other lexicons (Burnap and Williams, 2016; Alorainy et al., 2019).

Later, word-embedding-based approaches for automatic extraction of semantic features reigned as state-of-the-art approaches (Nobata et al., 2016; Zhang et al., 2018; Badjatiya et al., 2017; Kshirsagar et al., 2018; Orăsan, 2018; Pratiwi et al., 2019; Galery et al., 2018). Approaches using Deep Neural Networks have also been explored in the literature (Nina-Alcocer, 2019; Ribeiro and Silva, 2019). Use of Convolutional Neural Networks (Gambäck and Sikdar, 2017; Roy et al., 2018; Huang et al., 2018), Long-Short Term Memory (LSTM) (Badjatiya et al., 2017; Pitsilis et al., 2018; Nikhil et al., 2018) and Gated Recurrent Unit (GRU) (Zhang et al., 2018; Galery et al., 2018) or a combination of different Deep Neural Network architectures in an ensemble setting (Madisetty and Sankar Desarkar, 2018) have been explored for obtaining better feature representation and thereby improving the aggression detection performance.

Recently, the use of contextual embedding-based approaches like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have become state-of-the-art approaches in terms of performance on aggressive language identification tasks (Bojkovský and Pikuliak, 2019; Ramiandrisoa and Mothe, 2020; Mozafari et al., 2019). Particularly, the use of BERT-based approaches is gaining traction within shared tasks on abusive language detection for performance improvement. This can be observed in SemEval-2019 Task 6 (Zampieri et al., 2019) for English tweets, and TRAC (Kumar et al., 2018a) for Hindi and English tweets and Facebook comments. This motivates us to explore the use of several pre-trained language models for this work.

[2] https://github.com/surrey-nlp/political-aggression-detection

## 3 Datasets

For our experiments, we collected two datasets as described below.

### 3.1 D1 (TRAC Dataset)

The TRAC dataset used for our experiments is a collated and pruned version of two shared task datasets released with Trolling, Aggression, and Cyber-bullying (TRAC) 2018 (Kumar et al., 2018a) and 2020 (Kumar et al., 2020) workshops. This data has been crawled from Facebook, Twitter, and YouTube comments and was mainly collected from pages containing issues concerning the Indian population. As described in their papers, this dataset contains English and Hindi-English code-mixed data. However, upon manual observation, we noticed some Hindi instances too and pruned them using steps mentioned in section Code-mixed Data 3.3.1. The remaining instances only contain English and Hindi-English code-mixed data, and the number of instances for each class is shown in Table 2.

### 3.2 D2 (Our Dataset)

The objective of creating this dataset was to analyze aggression on social media, specifically in the context of Indian general elections. We could scrape approximately 10,000 tweets made through the public Twitter handles of 110 most influential Indian personalities. These tweets were made when general elections were held in India in 2014 and 2019. The ratio of tweets collected in the pre and post-election time frames was about 3 to 2. These personalities **belong to the following domains**:

- Political figures and official handles of political parties like Indian National Congress (INC), Bharatiya Janta Party (BJP), Aam Aadmi Party (AAP), and so on.
- Journalists, independent and affiliated with mainstream news organizations, followed by many people.
- Other prominent personalities who hold a massive follower count and make political tweets, such as actors, sports persons, *etc.* which can be considered influential.

We filtered these tweets based on the context, as we aim to model for political aggression, using some of the popular election-related keywords (such as “EVMs”, “rallies”, “election results”, *etc.*). This keyword-based manual pruning reduced the number of data instances to 2000. Data were sampled during collection based on language, including only English and some Hindi-English code-mixed data. We labeled it manually with the help of two annotators. Both our annotators are graduate students who are native speakers of Hindi, with proficiency in English and an understanding of the political context in which the tweets were made. We also assess the inter-annotator agreement using **Cohen’s Kappa** score and discuss it in a subsection below. The statistics for the datasets we use for experimentation are shown in Table 2.

|  | OAG | CAG | NAG | Total |
|---|---|---|---|---|
| **D1 (TRAC)** | 2715 | 4093 | 5436 | 12244 |
| **D2 (Our)** | 489 | 519 | 992 | 2000 |
| **D3 (Combined)** | 3204 | 4612 | 6428 | 14244 |
| **Code-mixed** | 943 | 1364 | 2670 | 4977 |
| **Non-Code-mixed** | 2261 | 3248 | 3758 | 9267 |

Table 2: Statistics for the different datasets used in our experiments.

To perform experiments over a combined dataset, *we also concatenate both D1 and D2 to create D3 - a combined dataset* (Table 2). Our experiments include applying our aggression detection approaches to this dataset as well.

### 3.3 Code-mixing: A Challenge

Code-mixing is the intermixing of units like words or phrases from one language (embedded language) within a second or primary language (matrix language) (Sitaram et al., 2019). Some of the most prevalent instances of such types of sentences can be observed in Hinglish (Hindi-English) (Srivastava and Singh, 2021) and Spanglish (Spanish-English) (Bullock et al., 2019) datasets. Although such text can be considered informal, with the increasing number of multilingual speakers, its usage has become quite the norm today. Thus, it has become essential to study the opinions and mindsets of people using code-mixing to express their views, especially when investigating data from social media platforms The most popular platform for observing code-mixing nowadays is social media. With people expressing their innate views, understanding and analyzing such data has garnered interest from different research communities.

Since the data is not exclusive to a single language, there are challenges associated with handling it. Each language has its own set of rules. Standardizing text that deviates from a canonical form happens at the token level or even at the semantic level (Çetinoğlu et al., 2016; Parikh and Solorio, 2021). Parsing poses yet another problem due to the syntactic rules that apply to one language but not to the other and the fact that errors may propagate from the previous layer (Çetinoğlu et al., 2016). Further, language identification poses challenges when languages are closely related and have common false friends (semantically different words sharing the same ancestor language).

Such challenges make code-mixing data harder to work with, as compared to working with monolingual data or even datasets containing well-separated monolingual instances from multiple languages.

### 3.3.1 Code-mixed Data

Tweets posted in the Indian political context are known to contain code-mixed data, *i.e.,* the presence of transliterated Hindi words (written in the Latin script). Such data presents challenges even for pre-trained multilingual language models, as they do not encounter code-mixed data during pre-training.

We obtained two separate sub-parts from D3 (Combined) to address the challenges presented by code-mixed data - *code-mixed* and *non-code-mixed* (Table 2). To obtain this separation, we perform some initial pre-processing, use a heuristics-based approach and utilize a Language Identification (LID) Model (Nayak and Joshi, 2022) as follows:

1. All the punctuation marks, special symbols, and their respective words, for entities like @mentions, and #hashtags were removed from the sentences.

2. If the sequence length obtained after punctuation or special mention removal became null, those sentences were omitted (classified as non-code-mixed).

3. We provide each data instance after following the above steps as input to the LID model to obtain token-level labels.

4. For classifying the sentences into Hindi, English, and Hindi-English code-mixed categories, a range of thresholds from 2%-20% were applied. Finally, after observing the number of sentences that fell into each language across these thresholds, 12% was chosen as a filter for categorizing language for each sentence. For example, in a sentence containing 36 words, if 5 or more words (equivalent to greater than or equal to 12%) were identified as Hindi, it would be labeled as code-mixed; else, it would be counted as English.

5. There were instances where even the complete sentences were in Hindi. Such sentences were also removed as we were dealing primarily with English data, with instances of Hindi code-mixing.

6. We also mask all usernames using [MASK] in the tweets to avoid biasing our models.

Table 7 in the appendix section reports the language-wise statistics obtained after these steps. It is to be noted here that the above steps were performed only for the separation of code-mixed data *from* non-code-mixed data. No pre-processing was performed for the aggression detection task.

## 3.4 Dataset Validation and Analysis

D2 (Our Dataset) was curated from over $10,000$ tweets, and as discussed, keyword-based manual pruning with the help of annotators reduced the final data instances to $2,000$. This data was collected from tweets posted four months before the Indian elections and two months after the declaration of election results. Out of these $2,000$ tweets, $1,200$ were collected in the 'pre-election' period, and 800 were obtained in the 'post-results' period.

Two annotators labelled this data manually, and we obtained an inter-annotator agreement score of 0.76 (Cohens' Kappa), which indicates "substantial agreement" ($p < 0.05$). Our annotators belonged to different political ideologies, and substantial agreement was obtained on the aggression label. Given a disagreement on any instance, we obtained a class label on such instances with the help of a third annotator.

**Dataset Analysis:** We also perform topic modeling-based analysis on D2 (our dataset) using BERTopic (Grootendorst, 2022). In Figures 1, 2, and 3, we show the most frequently occurring topic-wise token distribution for top-k words (k = 5). We make the following observations:

1. **Overtly Aggressive**: Among the topics clustered by BERTopic, as seen in Figure 1, the

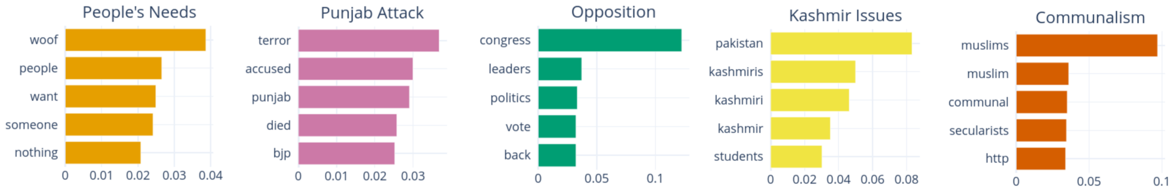

Figure 1: **Topics in the OAG category.** We note that people on Twitter are the most overtly aggressive about the political opposition, Kashmir and other internal security issues, and communal topics, whereas people's actual needs, and a terror attack that took place in the state of Punjab, take a backseat.

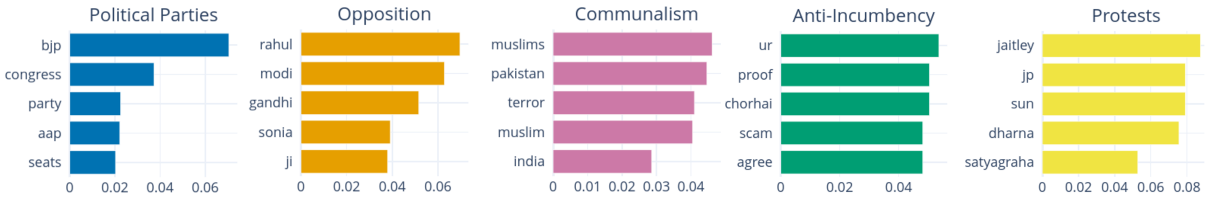

Figure 2: **Topics in the CAG category.** We note that most covert aggression is reserved for the protests going on around in the country for various issues, and an overall anti-incumbency sentiment against the ruling central government. Aggression is also prevalent against the political opposition as well as other political parties in general.

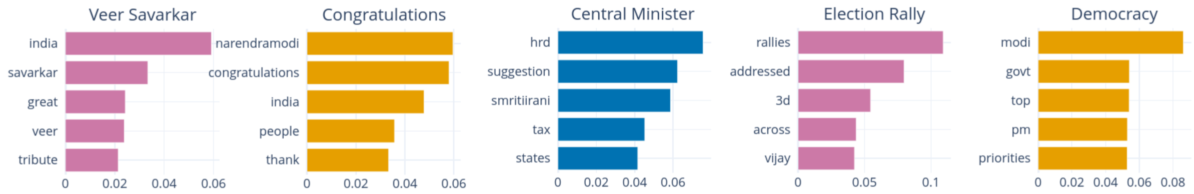

Figure 3: **Topics in the NAG category.** Most non-aggressive tweets correspond to the congratulatory messages sent to members of the party that emerged victorious in the elections. Discussions also take place around key cabinet roles and their possible contenders, election rallies, and the overall state of democracy in general.

data suggests that the most discussed topics where people were overtly aggressive in their tweets are *people's needs*, *Punjab attack*, *opposition*, *Kashmir issues*, and *communalism*. We note that political, religious, and national security issues are more aggressively discussed and debated online than people's needs for food and housing.

2. **Covertly Aggressive**: Similarly, the data (Figure 2) also suggests that there was minor aggressive behavior displayed when topics like *political parties*, *opposition*, *anti-incumbency*, *protests*, and *communalism* are concerned. Aggressiveness against the ruling party, the opposition, and all political parties, in general, is observed. This subsection also consists of tweets made with regard to various public protests that were being carried out in the run-up to the elections.

3. **Not Aggressive**: However, the data in Figure 3 shows us that social media discussions were non-aggressive when topics like *Veer Savarkar*, *congratulations*, *central minister*, *election rally*, and *democracy* are concerned. It includes congratulatory messages extended to the winning party members. Similarly, contenders for cabinet ministry posts are speculated, and the overall state of democracy is pondered upon.

We also perform additional topic modeling-based analysis for these tweets by segregating them into the 'pre-election' and 'post-result' periods (please see Appendix A.1). We choose four months before the election since this time is sensitive, and *exit polls* in the mainstream media start creating the election buzz. However, post-results, as observed from the data, the political scenario becomes rather concentrated on congratulating the winning party, diminishing data on development-related issues.

| | TRAC (D1) | Ours (D2) | Combined (D3) | Code-Mixed | Non Code-Mixed |
|---|---|---|---|---|---|
| BERT$_{base}$ | 67.17±0.53 | 58.89±2.42 | 65.44±0.70 | 66.40±1.13 | 63.84±1.42 |
| RoBERTa$_{base}$ | 69.05±0.57 | **66.66±3.82** | 66.85±1.23 | 65.16±2.06 | 65.11±1.08 |
| ALBERT$_{base-v2}$ | 66.03±0.89 | 54.61±4.28 | 64.71±0.76 | 62.15±3.89 | 59.97±2.60 |
| XLM-RoBERTa$_{base}$ | 67.73±2.02 | 61.08±2.21 | 62.88±3.08 | 64.52±2.56 | 60.97±2.73 |
| MURIL$_{base}$ | 66.64±1.08 | 60.62±2.00 | 65.47±0.83 | 66.71±1.37 | 62.33±0.88 |
| XLM-RoBERTa$_{large}$ | 68.00±1.29 | 66.38±1.84 | **67.95±1.37** | 67.83±2.52 | 64.92±0.97 |
| Hing-BERT | **69.37±0.96** | 62.41±3.02 | 67.48±1.91 | **68.50±1.35** | 65.13±1.62 |
| Hing-mBERT | 67.41±1.06 | 57.65±2.36 | 65.70±0.66 | 65.84±1.71 | **65.84±1.40** |
| HingRoBERTa | 68.85±1.28 | 64.81±2.79 | 66.95±1.43 | 68.36±1.71 | 63.11±1.85 |

Table 3: Mean macro F1-Score (F) from various pre-trained language models on TRAC (D1), Our Dataset (D2, Combined (D3), code-mixed and non-code-mixed subsets of D3; reported in percentage points. The values in **bold** highlight the best-performing language model on each dataset.

## 4 Approach

Recently, sequence classification via fine-tuning of pre-trained language models has become a standard approach for performing various NLP tasks. We take a similar approach and fine-tune various pre-trained language models for the task of aggression detection to report the results below. We select some monolingual, some multilingual, and some pre-trained language models specific to Hindi-English code-mixing.

Every sentence/tweet containing a sequence of words is tokenized into a sequence of sub-words using the model-specific tokenizer. The input to the model is a sequence of sub-word tokens that pass through the Transformer encoder layers. The output from the transformer is an encoder representation for each token in the sequence. We take the encoder representation of the [CLS] token in the case of BERT or the last encoder hidden states for other models. The output layer is a linear layer followed by *softmax* function, which takes in the above representation. The model is trained by optimizing for the cross-entropy loss value.

## 5 Experimental Setup

We fine-tune various pre-trained languages (both monolingual and multilingual) for the task of aggression detection and use the following pre-trained language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), ALBERT (Lan et al., 2020) which are pre-trained over English data. We also include XLM-RoBERTa (Conneau et al., 2020), both base and large variants, which are trained over multilingual data (containing both English and Hindi) and MURIL (Khanuja et al., 2021), a multilingual language model specifically built for Indian language. Finally, HingBERT, HingMBERT, and, HingRoBERTa (Nayak and Joshi, 2022) are also included as they are pre-trained over code-mixed Hindi-English data.

**Data Split and Evaluation** We report macro F1 Score on TRAC Dataset (D1), Our dataset data set (D2), combined dataset (D3) along with code-mixed and non-code-mixed subsets of data as discussed in Section 3. For the train/validation/test split sizes, we choose uniform 80% / 10% / 10% from each dataset to perform our experiments. We additionally report results on the subset of data containing code-mixed instances extracted from the combined dataset. To demonstrate the efficacy of our dataset, we also perform zero-shot domain transfer experiments. We evaluate the model trained on the TRAC dataset and tested it on our dataset and vice-versa to report zero-shot domain transfer results in Table 5.

**Experiment Settings** We perform experiments using the Huggingface Transformers library (Wolf et al., 2020). We monitor the *validation* set Macro-F1 to find the best hyperparameter values. We use the following range of values for selecting the best hyperparameter:

- **Batch Size:** 8, 16, 32
- **Learning Rate:** 1e-5, 1e-6, 3e-5, 3e-6, 5e-5, 5e-6

We repeat each training five times with different random seeds and report the mean macro F1-score along with its standard deviation. Our experiments were performed using 2 x Nvidia RTX A5000 and a single training run usually takes approximately 1 hour, on the combined dataset. For other datasets, however, the runtime is approximately 30 minutes. We generate various models during our experiments where the number of trainable parameters varies from 100M to 200M depending on the language model used.

**Custom Weighted Loss** As the dataset exhibits class imbalance, we use weighted cross-entropy loss in all our experiments. We assign a weight to the loss of every instance depending on the class label. We find the percentage of examples by class belonging to each class from the train split. We take the inverse of the probability values as the weight for the particular class. In this way, we provide more importance to the instances belonging to the minority class.

## 6 Results and Discussion

We report the results obtained via fine-tuning pre-trained language models in this section. Table 3 reports the Test set F1-Score from various pre-trained language models on the TRAC dataset, our dataset, and combined results. In addition to this, we also present the scores on code-mixed and non-code-mixed subsets of the entire data. We observe that Hing-BERT model outperforms other pre-trained language models on the TRAC dataset, achieving the highest macro F1-score of 69.37 across all combinations. On our dataset, however, we observe that RoBERTa$_{base}$ outperforms other pre-trained language models. For clarity in resultant observations, we provide a separation between monolingual, multilingual, and language models pre-trained on the code-mixed data.

### Multi-Dataset Fine-Tuning

From Table 4 we can observe that training on the combined dataset (D3) results in significant performance improvements on both our dataset and TRAC dataset. On the TRAC dataset we observe an increase in the best F-Score from 69.37 to 93.51. Similarly, we observe an increase in best F-Score from 66.66 to 92.77 on our dataset across models.

### Code-mixed

As both the TRAC dataset and our dataset contain code-mixed instances, we fine-tune and report F-Score on these subsets of instances (Table 3). As expected, we get the best F1 score on code-mixed instances with *Hing** models. This may be attributed to the fact that *Hing-** models have been *pre-trained on millions of code-mixed Hindi-English sentences*. However, to our surprise, the Hing-mBERT model outperforms other monolingual and multilingual models on non-code-mixed data as well. This result may be attributed to the fact that a significant amount of code-mixed data used in the pre-training of the *Hing** models comes from the social-media domain.

| **Models** | **D3 –>D1** | **D3 –>D2** |
|---|---|---|
| BERT$_{base}$ | 86.12±7.95 | 81.49±11.73 |
| RoBERTa$_{base}$ | 90.56±2.71 | 90.01±3.51 |
| ALBERT$_{base-v2}$ | 75.40±6.62 | 75.59±6.68 |
| XLM-RoBERTa$_{base}$ | 78.14±11.05 | 73.09±15.13 |
| MURIL$_{base}$ | 84.11±3.62 | 81.58±4.62 |
| XLM-RoBERTa$_{large}$ | 87.87±2.74 | 88.02±6.18 |
| HingBERT | 89.57±6.56 | 87.86±10.19 |
| Hing-mBERT | 88.71±6.34 | 86.28±8.51 |
| Hing-RoBERTa | **93.51±1.14** | **92.77±1.17** |

Table 4: Zero-Shot Test Set F1-Score from various language models trained on D3; D1 represents the TRAC dataset, D2 is our manually curated dataset, and D3 is the combined dataset.

| **Models** | **D1 –>D2** | **D2 –>D1** |
|---|---|---|
| BERT$_{base}$ | 48.82±2.55 | 50.55±1.33 |
| RoBERTa$_{base}$ | 46.29±3.60 | 55.33±1.53 |
| ALBERT$_{base-v2}$ | 46.32±2.58 | 47.14±1.23 |
| XLM-RoBERTa$_{base}$ | 47.32±2.28 | 52.53±1.19 |
| MURIL$_{base}$ | 48.77±3.42 | 52.49±0.68 |
| XLM-RoBERTa$_{large}$ | 47.67±2.84 | 55.77±0.98 |
| HingBERT | 47.08±2.38 | 54.34±1.12 |
| Hing-mBERT | 43.06±3.38 | 52.09±1.87 |
| Hing-RoBERTa | 49.30±3.43 | 52.12±0.71 |

Table 5: Zero-Shot Test Set F1-Score from language models trained on D1 and D2 respectively. D1 represents the TRAC dataset, D2 is our manually curated dataset, and D3 is the combined dataset.

### Zero-Shot Transfer Learning

Table 5 presents the results from our transfer learning setup. Columns D1 –>D2 and D2 –>D1 present a zero-shot setup from which we observe the performance of models fine-tuned on the D1 (TRAC) dataset and tested on D2 (our data) and vice-versa, respectively. From here, we observe that models trained on our dataset consistently obtain better F1-

| **Tweet** | **GT** | **M1** | **M2** | **M3** | **Error Type** |
|---|---|---|---|---|---|
| As per Zee News 405 for seats for BJP in UP. Total constituency is 403. Two seats given by Zee News on free of cost. | **CAG** | NAG | NAG | NAG | **Sarcasm** |
| Finally paused the video . It's so nice now lol | **CAG** | NAG | NAG | NAG | **Sarcasm** |
| Do you know Malda. ?? | **CAG** | NAG | CAG | NAG | **Short sequence** |
| Oh really | **CAG** | NAG | NAG | NAG | **Short sequence** |
| ek problem hai Main parents ke saath nahi dekh payunga. | **NAG** | CAG | NAG | NAG | **Code-Mixing** |
| Jay hind Pakistan me jabrdast Hamla Kare Hmari Sena jbab dena jaruri h | **CAG** | NAG | CAG | CAG | **Code-Mixing** |

Table 6: Prediction on test set examples from some of the fine-tuned models. **GT**: Ground Truth label, **M1**: RoBERTa$_{base}$, **M2**: XLM-RoBERTa$_{large}$, **M3**: Hing-BERT.

Score compared to models trained on the TRAC dataset. This performance benchmark is surprising, given the dataset size of the TRAC data is larger compared to our data; and given approximately similar underlying class balance ratio for both datasets.

### Discussion: Error Analysis

For error analysis, we pick the best-performing models on a combined dataset from the monolingual, multilingual, and code-mixed categories which were RoBERTa$_{base}$, XLM-RoBERTa$_{large}$ and Hing-BERT respectively.

Upon going through examples, we encountered various examples which were part of the TRAC dataset, where we found a disagreement with the annotated labels. For instance, the following sentences are labeled 'Not Aggressive', even though they have some amount of aggression:

- "Oh yeah cave civilisation can claim that.. After all u r their illegal creation"
- "He is modi dog. Godi media not usefull in India."

Leaving such disputed annotations aside, we report some of the most common error patterns in Table 6. Instances carrying sarcasm were quite often not recognized correctly by the three models, since it is not an easy task to recognize the latent intent or in this case, the aggression in such a sense. Another common error we noticed included very short sequences. Such types of sentences are quite common on social media, where these often carry some hidden context or a backstory. But the models find it difficult to predict the exact category for such examples. Finally, since the data contains some amount of code-mixing, we see a monolingual model, RoBERTa performing relatively worse than multilingual and code-mixed models like XLM-R and Hing-BERT which have seen more such kind of data while pre-training.

## 7 Conclusion and Future Work

In this paper, we curate a novel dataset to model political aggression. We analyze this dataset using various approaches like topic modeling, aggression detection, and report results. To benchmark our performance, we also perform the aggression detection task with the help of an existing dataset. Our results and analyses also take into account the code-mixing phenomenon observed on social media platforms. The zero-shot cross-dataset experiments show the efficacy of our dataset, which consistently outperforms the approaches used with existing data. While political aggression is subtle occasionally, we observe that some data instances show overtly aggressive behavior. It is important to note the limitations of such a study and we discuss them in the next section. We release any data, code, and models produced during this study (including any raw data, but keeping user handles anonymous) publicly for further research by the community. We license this release under CC-BY-NC-SA 4.0.

In the future, we aim to collect more data from multiple social media platforms and release it to model aggressive behavior. We plan to perform similar experiments on a large dataset while benchmarking and comparing our current models' performance. We also plan to investigate online or active learning for the same. Finally, we also aim to expand on the theoretical underpinnings of sublime aggression and offense by attempting to identify these within other more tangential domains, *viz.*, comedy.

## Limitations

Our work can be considered to have the following possible limitations:

1. The dataset we introduce and use to perform analysis contains 2000 tweets sampled from a specific time frame over a single social media platform. However, we aim to extend this work by collecting more political data across various social media platforms and using it to model aggressive behavior. Please do note that these tweets have been manually filtered from a larger set of 10,000 tweets while manually labelling them and ensuring that they are relevant to the political domain.

2. The number of user handles that we scrape tweets from for this study is around 110. This number might not be reflective of a large political space considering the plethora of politically active personalities in India. However, it is noteworthy that each of these 110 user handles has a minimum of $100,000$ followers on Twitter, on the basis of which we consider them to be influential on a social media platform.

## Ethics Statement

Our dataset of tweets was obtained by scraping Twitter. We also obtain a subset of data from existing aggressive detection datasets cited in this paper, complying with the terms of use of each of these datasets. All datasets were anonymized, no tweet-ids or Twitter usernames or any of their demographics are included in the data used to train our models. We plan to release only the tweet ids as part of our dataset, along with the labels, in the final version.

## References

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L. Williams. 2019. "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3).

Craig A Anderson, Brad J Bushman, et al. 2002. Human aggression. *Annual review of psychology*, 53(1):27–51.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press.

Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 50–57.

Michal Bojkovský and Matúš Pikuliak. 2019. STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 464–468, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Barbara Bullock, Wally Guzmán, and Almeida Jacqueline Toribio. 2019. The limits of Spanglish? In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 115–121, Minneapolis, USA. Association for Computational Linguistics.

Peter Burnap and Matthew Leighton Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *Epj Data Science*, 5.

Brad J Bushman and L Rowell Huesmann. 2014. Twenty-five years of research on violence in digital games and aggression revisited: A reply to elson and ferguson (2013).

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. *arXiv preprint arXiv:1610.02213*.

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

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515.

Jean Philippe Décieux, Andreas Heinen, and Helmut Willems. 2019. Social media and its role in friendship-driven interactions among young people: A mixed methods study. *Young*, 27(1):18–31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thiago Galery, Efstathios Charitos, and Ye Tian. 2018. Aggression identification and multi lingual word embeddings. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 74–79, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Claire Hardaker and Mark McGlashan. 2016. "real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80–93.

Qianjia Huang, Diana Inkpen, Jianhong Zhang, and David Van Bruwaene. 2018. Cyberbullying intervention based on convolutional neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 42–51, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors. 2020. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018b. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 1621–1622. AAAI Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L. Williams. 2019a. A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, 6(2):227–240.

Shuhua Liu and Thomas Forss. 2014. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - Volume 1: SSTM, (IC3K 2014)*, pages 530–537. INSTICC, SciTePress.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karla Mantilla. 2013. Gendertrolling: Misogyny adapts to new media. *Feminist studies*, 39(2):563–570.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media.

Ravindra Nayak and Raviraj Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. In *Proceedings of The WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. 2018. LSTMs with attention for aggression detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 52–57, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Victor Nina-Alcocer. 2019. HATERecognizer at SemEval-2019 task 5: Using features and neural networks to face hate recognition. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 409–415, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Constantin Orăsan. 2018. Aggressive language identification using word embeddings and sentiment features. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 113–119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Moderator: Mary Ellen O'Toole, Participants: Eugene V Beresin, Alan Berman, Steven M Gorelick, Jacqueline B Helfgott, and Chuck Tobin. 2014. Celebrities through violence: The copycat effect and the influence of violence in social media on mass killers. *Violence and gender*, 1(3):107–116.

Dwija Parikh and Thamar Solorio. 2021. Normalization and back-transliteration for code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124, Online. Association for Computational Linguistics.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Andraž Pelicon, Matej Martinc, and Petra Kralj Novak. 2019. Embeddia at semeval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 604–610.

Juan Manuel Pérez and Franco M. Luque. 2019. Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48:4730–4742.

Tom Postmes, Russell Spears, and Martin Lea. 1998. Breaching or building social boundaries? side-effects of computer-mediated communication. *Communication research*, 25(6):689–715.

Nur Indah Pratiwi, Indra Budi, and Ika Alfina. 2019. Hate speech detection on indonesian instagram comments using fasttext approach. In *2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, 2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018, pages 447–450, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2018 IEEE.; 10th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018 ; Conference date: 27-10-2018 Through 28-10-2018.

Faneva Ramiandrisoa and Josiane Mothe. 2020. Aggression identification in social media: a transfer learning based approach. In *TRAC*.

Alison Ribeiro and Nádia Silva. 2019. INF-HatEval at SemEval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Arjun Roy, Prashant Kapil, Kingshuk Basak, and Asif Ekbal. 2018. An ensemble approach for aggression identification in English and Hindi text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 66–73, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, page 1058–1065. AAAI Press.

Vivek Srivastava and Mayank Singh. 2021. HinGE: A dataset for generation and evaluation of code-mixed Hinglish text. In *Proceedings of the 2nd Workshop on*

*Evaluation and Comparison of NLP Systems*, pages 200–208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicolas Suzor, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2019. Human rights by design: The responsibilities of social media platforms to address gender-based violence online. *Policy & Internet*, 11(1):84–103.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835. Publisher Copyright: © 2018 IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Z Zhang, D Robinson, and J Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In A Gangemi, R Navigli, M-E Vidal, P Hitzler, R Troncy, L Hollink, A Tordai, and M Alam, editors, *The Semantic Web: Proceedings of the 15th European Semantic Web Conference (ESWC 2018), Heraklion, Crete, Greece, 3-7 June 2018*, volume 10843 of *Lecture notes in computer science*, pages 745–760. Springer, Cham, Switzerland.

# A Appendix

In this appendix, we provide additional details which could not be included in the paper. We start by showing the language-wise class distribution in our dataset, in Table 7.

| | OAG | CAG | NAG | Total |
|---|---|---|---|---|
| **English** | 2263 | 3254 | 3756 | 9273 |
| **Hindi(excluded from D3)** | 8 | 21 | 96 | 125 |
| **Code-Mixed** | 941 | 1358 | 2672 | 4971 |
| **Total** | 3212 | 4633 | 6524 | 14369 |

Table 7: Language-wise class distribution

## A.1 Additional Dataset Analysis

We create two subsections of D2 (our dataset) by categorizing tweets that were made *before* the conduct of the elections (both 2014 and 2019) and *after* the declaration of results (both 2014 and 2019). We individually perform topic modeling on these subsections using BERTopic to get an insight into what issues were prominent before and after the elections.

### Pre-Elections

1. **Overtly Aggressive**: Among the topics that were discussed online before elections, the most overtly aggressive debates happened on the tussle between *journalists* and the *ruling party*. *Casteism*, *communalism*, and national security including the *Kashmir issue* also find their way amongst the overtly aggressive tweets.

2. **Covertly Aggressive**: Our data analysis results also suggest that the top topics in the covertly aggressive category were *political parties*, *terrorism*, *development* and *unemployment* related issues.

3. **Not Aggressive**: The social media discussions were non-aggressive when topics like *martyrs*, and people's *rights* and the overall situation of *democracy* were being discussed. The presence of *religion* and *communalism* in this section also suggests that the peacemakers are equally active on this social media platform, as are the notorious aggressive tweeters.

### Post-Results

We generally observe a stark decline in the number of aggressive tweets (OAG+CAG) when the post-

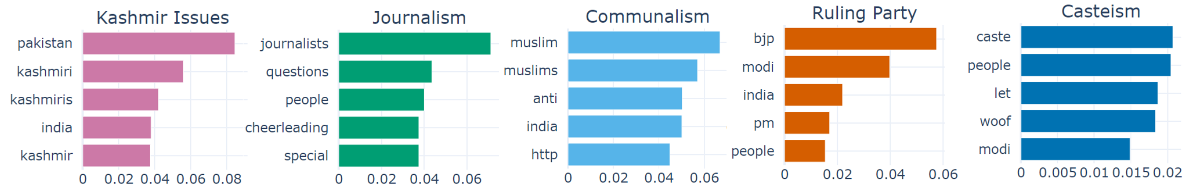

Figure 4: **Topics in the OAG category** in the 'Pre-Elections' sub-section of dataset D2. We note that there is a tussle between journalists in popular media and the ruling party. Aggressive tweets are also shared on topics of casteism, communalism, and issues related to Kashmir and the internal security of India.

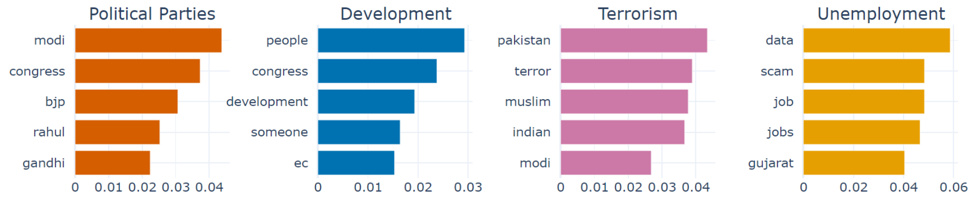

Figure 5: **Topics in the CAG category** in the 'Pre-Elections' sub-section of dataset D2. The tweets in this class see covertly aggressive debates on political parties, development, and social and security challenges such as unemployment and terrorism.

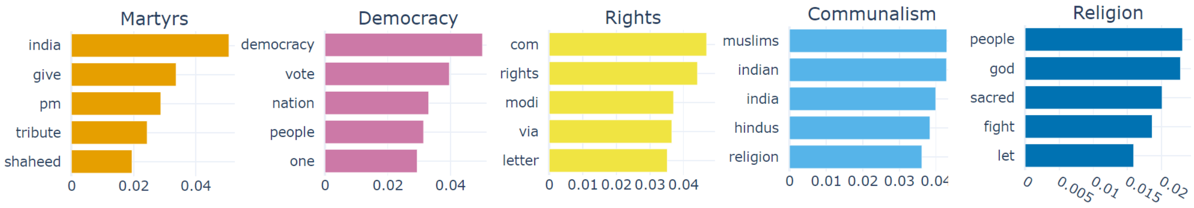

Figure 6: **Topics in the NAG category** in the 'Pre-Elections' sub-section of dataset D2. This includes tweets paying respect to the martyrs, and discussions on socio-political rights and democracy in general. It also sees non-aggressive discussions on religion and the problem of communalism.

elections data is taken into consideration. Compared to 709 out 1200 (59.33%) for "pre-elections" the ratio of aggressive tweets "post-results" comes down to 298 out of 800 (37.50%).

1. **Overtly Aggressive**: The overtly aggressive class in the after-elections category saw discussions on *political parties* in general and the *ruling party* in particular. It also saw heated debates on the issue of *religion*.

2. **Covertly Aggressive**: Tweets belonging in the covertly aggressive class include *activism*, *post-election address* made by victors, and calls to *democracy*.

3. **Not Aggressive**: Non-aggressive tweets saw *tributes* offered to veteran political leaders, and speeches made by the *ruling party* which contained mentions of *development*, *law enforcement*, and a *New India*.

Additionally, we have a full hyperparameter table which we are omitting due to space constraints; to be added to the camera-ready on acceptance. If accepted, we will try to add it to the camera-ready copy of our paper; along with the visualizations from the 'Pre-elections' and 'Post-elections' topic modeling discussed here.

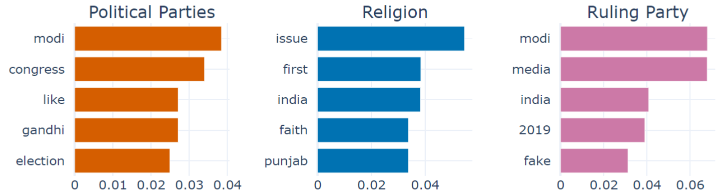

Figure 7: **Topics in the OAG category** in the ‘Post-Results’ sub-section of dataset D2. BERTopic gives clusters for tweets on the ruling party and other political parties, apart from an omniscient presence on the topic of religion.

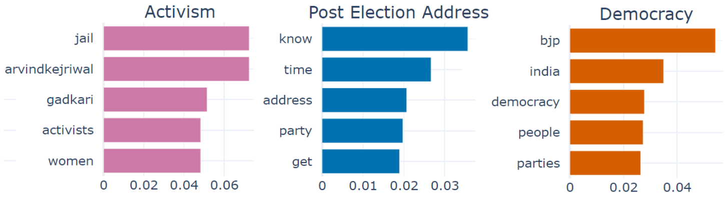

Figure 8: **Topics in the CAG category** in the ‘Post-Results’ sub-section of dataset D2. It includes tweets on activism and post-election addresses including references to democracy.

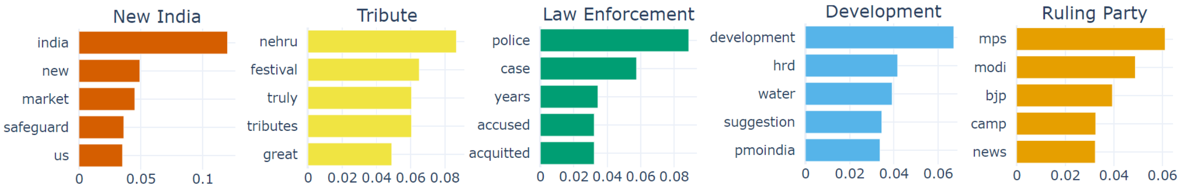

Figure 9: **Topics in the NAG category** in the ‘Post-Results’ sub-section of dataset D2. We note that the topics here are related to the ruling party that emerged victorious once again, tokens of tribute to veteran political leaders, and mentions of development, law enforcement, and a New India.