Crowdsourcing the curation of the training set for harmful content classifiers used in social media

A pilot study on political hate speech in India

November 2023

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on 'Social Media Governance'. The report reflects the personal opinions of the GPAI Experts and External Experts involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the 'Social Media Governance' project, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Responsible AI Expert Working Group. The GPAI Responsible AI Expert Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Alistair Knott[*], School of Engineering and Computer Science, Victoria University of Wellington
Dino Pedreschi[*], Department of Computer Science, University of Pisa

The report was written by: **Tapabrata Chakraborti**[‡], Alan Turing Institute, University College London, University of Oxford; **Avigyan Bhattacharya**[‡], Department of Computer Science, Jadavpur University, India; **Subhadip Basu**[‡], Department of Computer Science, Jadavpur University, India; **Alistair Knott**[*], School of Engineering and Computer Science, Victoria University of Wellington; **Dino Pedreschi**[*], Department of Computer Science, University of Pisa; **Raja Chatila**[*], Sorbonne University; **Susan Leavy**[*], School of Information and Communication Studies, University College Dublin; **David Eyers**[†], School of Computing, University of Otago; **Paul D. Teal**[†], School of Engineering and Computer Science, Victoria University of Wellington; and **Przemyslaw Biecek**[*], Warsaw University of Technology.

We would like to thank the other members of GPAI's Social Media Governance project, as well as Matt Farrington, for helpful comments on a draft of this manuscript. (Remaining errors are of course our own.) We are grateful to CMATER and the Department of Computer Science and Engineering, Jadavpur University India, for providing necessary infrastructural support during the project conceptualisation stage. Thanks are also due to team Infomaticae, for coordinating the smooth execution of the project. We particularly thank our annotators for their work.

GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Responsible AI Working Group. We are grateful, in particular, for the support of **Stephanie King** and **Lama Saouma** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Catherine Régis**[*] and **Raja Chatila**[*].

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2023c. Crowdsourcing the curation of the training set for harmful content classifiers used in social media: A pilot study on political hate speech in India, Report, November 2023, Global Partnership on AI.

Content Note

This report, given its subject of hate speech, contains quotes from content that many readers will find offensive. Of course, none of this quoted content in any way reflects the opinion of the report authors. Readers who may be disturbed by such content should avoid, in particular, Section 6.

Table of Contents

List of Figures

List of Tables

1. Introduction: Harmful Content Classifiers, and two concerns about their training

The world's dominant social media platforms all run active **content moderation** programmes. They take seriously their responsibility to moderate the content of their users' posts, to keep their community of users safe. Moderation involves checking for 'harmful content' of various kinds, and taking various actions when it is found. Some content is removed if it is found to be illegal or if it violates the company's published standards. Other content is left in place, but is moderated in less draconian ways, perhaps by flagging it with messages advising caution, or by downranking it in the platform's recommender algorithm. This latter type of content is often termed **borderline content**, and has been the subject of much discussion.

The scale of social media platforms means that content moderation processes must use *automated tools*, as well as human effort. AI tools are central in moderation processes, so content moderation is an important topic for our group at GPAI, which focusses on Social Media Governance.

Typically, for a given category of harmful content, an AI **classifier** is trained to recognise items of this category. Different classifiers are needed for different content modalities–images, texts, videos, and so on. Different categories of harmful content often require different classifiers too. Each individual classifier must be trained on a large **training set** of annotated examples. One job for the company's human content moderators is to construct these training sets, by labelling (or **annotating**) selected content items with categories relevant to moderation. After training, a classifier can contribute to the company's moderation processes. Typically, a trained classifier has a way of reporting the **confidence** it has in its verdict for a given item. If confidence is high, the appropriate moderation action is sometimes performed without any human intervention. Otherwise, the item can be passed to a human moderator for further consideration.

The project we describe in this report addresses two problems with existing methods for building harmful content classifiers in social media contexts. The first problem relates to *transparency*. We (the public) don't know much about how companies create and evaluate their classifiers. We don't know what methods are used to create the training sets they are trained on, or how the annotation process works. We don't know much about how content items are selected for training sets, and how annotators are selected to label these items. We don't know how training sets are kept up to date. We don't know what role user reporting of harmful content plays in training set curation. (When a user reports an item as harmful, they don't learn if the item was added to a training set, for instance.) Most importantly, we don't know how well the trained classifiers *perform*. It is standard practice for classifiers to be evaluated on examples held out from training: a classifier's performance on unseen examples can be reported as a percentage accuracy. (For instance, one classifier might be found to correctly classify 90% of unseen examples–another might be found to correctly classify 75% of unseen examples. The errors may also be usefully separated into false-negatives and false positives.) Companies typically report content moderation performance using another score, the 'proactive detection rate', which is the percentage of violating content found by its classifiers *before it is reported by users*. This measure of performance is useful–but a simple percentage accuracy score would also be useful, and is currently conspicuous by its absence in companies' transparency reporting. It is vital for us to know more about automatic content moderation, because each moderation action takes a position on a critical moral trade-off: between the need to avoid the *proliferation of harm* (if there are many false negatives), on the one hand, and the need to respect individuals' *freedom of speech* (and hence limit false positives) on the other. A trade-off must of course be made between these two vital principles: that is the essence of content moderation. But we would like to know more about how companies' AI systems are making this trade-off.

The second problem we address in our project concerns *consistency across platforms*. Different platforms have different definitions of harmful content, and have different policies on what to do with such content. At one extreme, some platforms (such as Parler and Gab) have no content moderation at all. But we believe that some moderation is essential to create workable information ecosystems–and this belief is shared by all the mainstream social media companies. In fact, the mainstream platforms adopt somewhat similar definitions of harmful content categories, and are likely to implement somewhat similar content moderation policies on these categories. (Of course we don't know the details, because classifiers are built behind closed doors, as already noted.)

The proposal we explore in this project is that companies' content moderation systems should implement *the same* practical definitions of harmful content, in some given locality or jurisdiction. We are thinking first and foremost of content which is *illegal* (in a given jurisdiction). There is no reason why companies should not train the classifiers for a given type of illegal content in exactly the same way, as they must all implement the same definition of illegal content. But at present we don't know if they do: they each build their own in-house training sets, and we have little information about how this is done. But even for 'borderline content' that should be moderated in ways other than removal, there is often no good reason why definitions of harmfulness should vary between companies. For instance, whether item *A* is more harmful than item *B*, and should be downranked more, is a decision to be made by assessing *the content items themselves*: we don't expect the decision to vary from one platform to another.

In summary, we see two problems with companies' current practices in developing harmful content classifiers. The first is that these practices are not *transparent*: they happen behind the closed doors of companies. The second is that they are not *consistent*: there is nothing to prevent different companies implementing different definitions in the training sets they create, both for categories of illegal content (where there is arguably a legal requirement for consistency) and for categories of borderline content.

In this project, we explore an alternative model, which addresses both problems. We will introduce this model in the next section. A summary of the work undertaken is shown in Figure 1.1.

2. A way of addressing the concerns: Curation of a Training Set in a semi-public domain

Our proposal is that in a given locality or jurisdiction, and for a given category of harmful content, social media companies operating in that jurisdiction should all use *the same training set* to train their classifiers for this content category. In our proposed system, this training set should be developed externally to companies, in a semi-public domain. We use the term 'semi-public' since the public should be informed about the processes by which the training set is created, and updated: specifically, they should know how content items are chosen for inclusion in the training set, and how annotators are selected to perform the labelling. However, there are limits to what the public should know. They should not know the identity of the annotators—because this would make annotators vulnerable to coercion of various kinds. They should not have access to the annotated dataset—because this would allow 'adversarial' methods to be deployed, by purveyors of harmful content who wish to avoid detection. (Such actors could train their own classifiers, and use these to finesse content items that escape detection.)

Our proposal has a number of advantages, which we will now enumerate. For the sake of concreteness, we will consider one particular category of harmful content in what follows: **hate speech**.

- Firstly, our proposal aligns with a fundamental principle of *public consultation*: definitions of harmful content in a given place should be determined by 'the people' in that place. Consider hate speech, for instance. The experts in deciding what counts as hate speech in a given place are the people in that place. (Those who are the *targets* of hate should arguably have a particular say in defining what counts as hate: we will explore that idea in Section 4.) Often, local expertise involves linguistic expertise too: each language needs its own classifiers, with their own dedicated training sets. Hate speech often also references particular groups in a given locality: many of the relevant groups are local to a given country or region. Our proposal implements a principle of *local governance* of social media platforms, at least as regards harmful content moderation. This is a very different model from those companies currently use, but we think it has much to recommend it. Our proposal also connects strongly with existing work on harmful content annotation and classification in the public domain—in particular, with meetings organized around public datasets and shared tasks, such as the OFFENSEVAL tracks of the SEMEVAL meeting, and the HASOC tracks of the FIRE meeting (see Jahan and Oussalah, 2023 for a recent review).

- Secondly, our proposal provides a far more subtle way of articulating definitions of harmful content than the definitions provided in black-letter law, or company policy. These latter definitions are paragraphs of text that express certain high-level generalisations. Such definitions leave many open questions about specific content items, and much room for interpretation. An annotated dataset, on the other hand, provides a detailed assessment of many specific items: it *includes* the interpretation step. (Note that there is no requirement that annotators agree on their interpretations and assessments: there is ample room for disagreement, as we discuss in Section 3.1. In fact, we argue that disagreement can helpfully *inform* content moderation practices, and provide a quantitative basis for some of these.) Note that legal systems often use juries in defamation trials; this process explicitly assumes that a group of citizens is well placed to collectively determine the degree of harm caused to a given individual by a given assertion (see e.g. Buck, 2022; Bennett, 2023 for relevant discussion). There is even an interesting argument to be made that a law on harmful content can *reside* in a training set of examples, annotated by a large group of citizens. We will outline this argument in Section 4.

- Thirdly–and moving to more pragmatic considerations–creating a single training set for use by all companies is an *efficient use of resources*. The more items a training set contains, the better the trained classifier will be. If companies can pool resources in the creation of a training set, the result will be a better classifier. The resourcing argument is of particular relevance to the (many) jurisdictions where companies are not able to devote large resources to content moderation, for instance for regions of the developing world, or for low-resource languages.

- Fourthly—and centrally—our proposal makes the process of training content classifiers more *transparent* than it currently is. There are limits to this transparency, as already noted, but the method whereby training sets are created would be a matter of public record. The method could even be specified by law, in a given jurisdiction.

- Finally, our proposal makes it possible for companies' harmful content classifiers to be more directly *evaluated* than is currently the case. We envisage that companies will continue to build their own content classifiers: these would remain behind closed doors, and the IP of the companies. But these classifiers would now all be evaluated *on the same dataset*. This paradigm for building and evaluating classifiers has in fact been central to all AI machine learning research for the last 30 years. In a given domain, a 'shared task' is defined, for which a training set is made available. A *competition* is then organised around this shared task: research teams compete to build the best system for that task, that learns most effectively from the shared training set. In competition, trained systems are assessed on unseen data held out from the training set. If the training set for harmful content is curated in a semi-public domain, and used by all companies, this will allow companies to compete against one another *on the quality of their harmful content classifiers*—that is to say, on a metric that goes directly to the public good.

In this report, we present a project that pilots the process of creating a training set for a social media hate speech classifier in a (semi-)public domain. Our pilot project runs in India, and focusses on political hate speech. In Section 3, we introduce the annotation scheme we have developed for hate speech, which is specially configured for use in a social media platform, and describe how we envisage classifiers being trained on the annotated datasets we gather. In Section 4, we outline the general methodology we propose for selecting annotators to perform the annotation exercise. In Section 5 we make some suggestions about how the methods we are trialling could be evaluated if they were put into use. In the remainder of the report, we describe the small pilot study we have conducted so far, and present some initial results.

A graphical summary of the new scheme we are piloting is shown in Figure 2.1. This figure is intended as a preview: details will be presented in the relevant sections. In relation to the **dataset of content items** selected to train the content moderation system:

- Companies currently assemble their own datasets, through methods the public knows little about.
- Our proposal is to assemble *a single dataset of items*, stored by an independent regulatory body, through participation by companies and the public. The public will know about the methods used to assemble this dataset, but the dataset itself will remain private, to prevent adversarial exploitation.

In relation to the **annotation protocol**:

- Companies currently administer their own in-house annotation processes, often adopting subtly different annotation schemes, again with little public transparency.
- Our proposal is to provide *a single annotation scheme* for use by all companies, and then extensively sample members of the public to annotate the single dataset of content items. This single annotated dataset is then to be used by all companies, to train their in-house content classifiers.

In relation to **content moderation**:

- The public currently knows little about how companies use AI systems in content moderation processes, and little about how well they perform in their allocated roles.
- In our proposal, the way moderation systems are trained and the way they inform the moderation process are a matter of public record; measures of their performance are also made available to the public.

In relation to **evaluation** of the resulting content moderation scheme:

- The public currently has little information about this process. (Often we don't even know if there is an evaluation process.)
- In our proposal, evaluation happens by A/B testing alternative moderation schemes, and asking users to report on how well the platform optimizes the crucial trade-off between preserving freedom of speech, and keeping the platform free from harmful content.

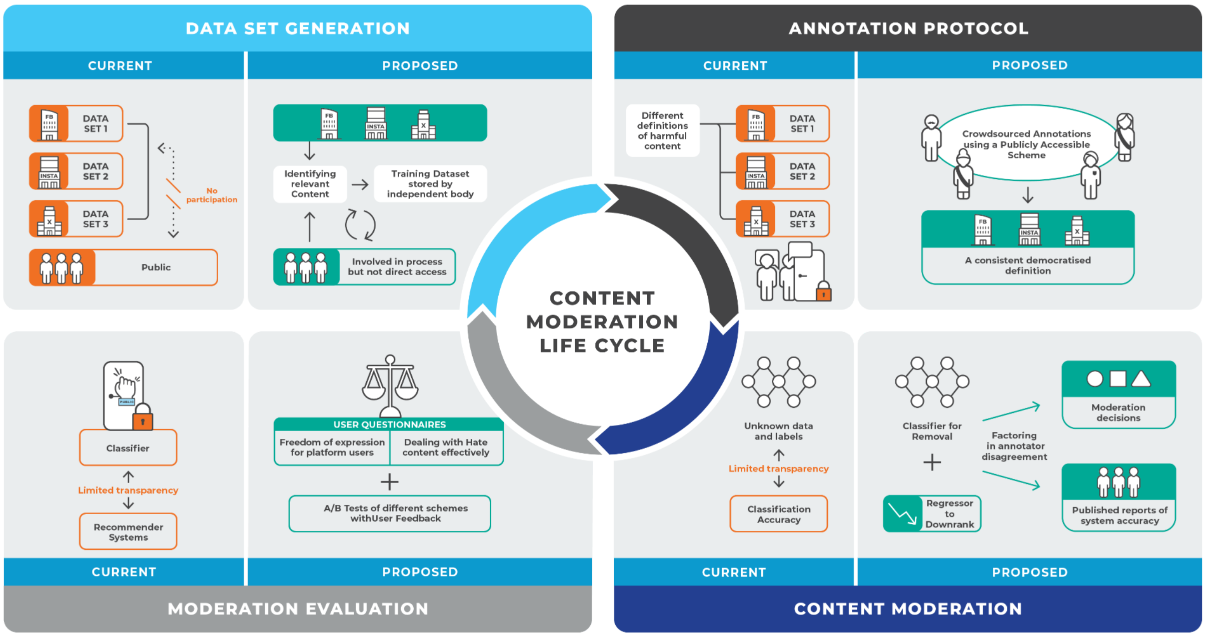

DATA SET GENERATION
CURRENT
PROPOSED
FB
DATA SET 1
INSTA
DATA SET 2
X
DATA SET 3
No participation
Public
Identifying relevant Content
Training Dataset stored by independent body
Involved in process but not direct access
ANNOTATION PROTOCOL
CURRENT
PROPOSED
Different definitions of harmful content
DATA SET 1
DATA SET 2
DATA SET 3
Crowdsourced Annotations using a Publicly Accessible Scheme
A consistent democratised definition
CONTENT MODERATION LIFE CYCLE
PUBLIC
Classifier
Limited transparency
Recommender Systems
USER QUESTIONNAIRES
Freedom of expression for platform users
Dealing with Hate content effectively
A/B Tests of different schemes withUser Feedback
CURRENT
PROPOSED
MODERATION EVALUATION
Unknown data and labels
Limited transparency
Classification Accuracy
Classifier for Removal
Regressor to Downrank
Factoring in annotator disagreement
Moderation decisions
Published reports of system accuracy
CURRENT
PROPOSED
CONTENT MODERATION

Figure 2.1: Graphical summary of the content moderation scheme we are piloting, and contrasts with the status quo

3. A proposed Annotation Scheme for Hate Speech, and a proposed training method for Hate Speech Classifiers

In this section, we introduce and motivate the annotation scheme for hateful speech that we use in our pilot study.

There are two key design principles for our scheme. The first principle is that the scheme should be directly informed by the repertoire of *content moderation actions* that social media companies have at their disposal. Harmful content can be straightforwardly *removed* from a platform, but it can also be *downranked* in the recommender algorithm, so that it is disseminated less than it would otherwise have been. It can also be left on the platform but *flagged* in various ways. Companies have a sophisticated range of moderation actions at their disposal; we suggest the annotation scheme should make direct reference to these actions, so annotations of a given content item directly convey intuitions about how platforms should deal with it. We call this an **operational** annotation scheme. There are two reasons we propose an operational scheme. One is that annotators' judgments can be directly reflected in moderation actions. Another is that each annotation requires the annotator to express a concrete attitude towards *free speech*, as well as a value judgment about content. As noted in Section 1, the central dilemma for content moderation is the trade-off between the need to limit harmful content and the need to respect freedom of expression. In an operational annotation scheme, each individual annotator must take a position on this dilemma for each of the problematic items they annotate: this means the annotated dataset as a whole expresses the detailed views of the selected annotators on this crucial issue. We will present and further motivate our annotation scheme in Section 3.1.

Our second principle is that *disagreement* between annotators provides valuable information about moderation actions, which should be explicitly incorporated into the process of training classifiers. When annotators are making value judgements, we *expect disagreement* between them: there is no a priori expectation of 'inter-annotator agreement' as there is in more factual annotation domains. In fact, we make a virtue of disagreement in our content proposed moderation scheme: we argue it carries valuable information about the appropriate moderation action. This has consequences for the scheme we use to train our classifiers. We want classifiers to learn *how much disagreement* there is amongst annotators about the content items in our training set, as well as first-order judgements about harmfulness. We will present our approach to training classifiers, and to annotator disagreement, in Sections 3.2 and 3.3.[1]

3.1. A proposed 'operational' annotation scheme for hate speech

As just noted, companies can take two different types of action on harmful content. Firstly, if it is harmful enough, or crosses some legal threshold for harmfulness, it is (or should be) straightforwardly removed from the platform. If not, it can remain. This is a *discrete* criterion: each item must either be removed or left in place. Second, for those items that remain on the platform, there can be items that have some sub-threshold degree of harmfulness—so-called 'borderline' harmful content. Borderline content is defined on a *continuous* scale: there are degrees of harmfulness. For this content, there are moderation actions that are defined on a similarly

[1] We should note that the classifiers we envisage can be used either to assist human moderators, or to take moderation actions automatically: we don't currently take a view on which of these roles they play.

continuous scale. In particular, the action of *downranking* an item in the platform's recommender algorithm is inherently continuous: an item can be downranked to any continuous degree.[2]

Items that don't need *any* downranking in fact constitute another discrete content category: an item can be 'removed', or 'downranked', or *'untouched'* [by moderation actions]. The boundary between the discrete categories of 'downranked' and 'untouched' content corresponds to the zero point on the continuous scale of downranking.[3] In fact, we can define one further discrete category: content which should be *upranked* by the recommender algorithm, because it has positively good value to an ongoing conversation. Upranking is a very different kind of moderation action. Companies certainly make use of upranking—for instance, in paying particular attention to health content from authoritative sources. We will include this category in our discrete scheme, even though our focus is on moderation of harmful content.

We would like our annotation scheme to directly inform both discrete moderation decisions, and, for borderline content, continuous moderation decisions (specifically about downranking). We therefore propose a *two-pass* annotation process.

In the **first pass**, annotators are asked to make a *discrete* judgment about content items. Should a given item be 'removed', or 'downranked', 'untouched', or 'upranked'? The annotated dataset from this pass can be used to train a **content classifier** that outputs decisions about discrete classes of content items.

In the **second pass**, annotators are asked to make a *continuous* judgment about content items: how harmful is a given item? The annotated dataset from this pass can be used to train a **content scorer** that gives a continuous 'harmfulness score' for each content item. (In machine learning terms, this scorer is a **regression model**, rather than a classifier.) In our current method, we present *all* content items to annotators in the second pass, including those that are agreed to need no action, or are candidates for upranking.

In the second pass, rather than asking annotators to directly score items, we present annotators with pairs of items, and ask them to *rank* these pairs. We use this method because there is some evidence in psychological research that ranking produces a more valid rating system than absolute judgements (see e.g., Goffin and Olson, 2011). Ranking has also been specifically advocated in schemes for annotating harmful content datasets for social media platforms—see e.g. Kiritchenko and Nejadgholi (2020). The intuition behind a ranking scheme can be conveyed by the educational task of marking students' essays. Markers often perform a ranking exercise as a first step in marking, to provide a partial order over the items to be marked, before quantitative scores are given. Given ranking judgements from many annotators on many pairs, there are good methods for generating absolute scores for each individual item, as we will discuss in Section 3.3.

[2] Flagging borderline content can also be done in different ways, depending on its severity. But we will focus for now on the operation of downranking, which is prototypically continuous.

[3] Defining the boundary between the discrete categories of 'downrank' and 'remove' is more open to interpretation. It could be seen as the point on the continuous downranking scale where an item is downranked so heavily it is not disseminated at all. Or it could be defined as the point which guarantees the item is at the end of users' feeds. Or it could be defined simply as a system-specific constant, whose effects vary over contexts.

3.2. Training a discrete classifier: a ‘soft labels’ model for encoding disagreement

The best-performing text classifiers at present are neural network models, that use pretrained **large language models** (**LLMs**) to derive rich and informative representations of input texts (see e.g., Chang et al., 2023). The best-performing image classifiers are also neural network models, which make similar use of large pretrained models (see e.g., ImageNet, 2023). We assume the dominant content classifiers used by platforms will be neural networks for the foreseeable future. A neural network classifier is trained by a loss (or error) term computed for each training item by comparing the network’s actual output for that item with the output it *should* have produced (see e.g., Sathyanarayana, 2014). The overall loss over the whole training set can then by minimised, using standard neural network learning techniques.

The loss term used to train a classifier is typically defined in the terminology of probability distributions. The output layer of a classifier consists of *N* units, one for each class. If the activity over these units for a given input item is scaled to sum to 1, it can be interpreted as a *discrete probability distribution* for the class of this given item. The output the network *should* have produced is likewise given as a probability distribution. The loss term for a given training item is then given as a comparison between these two probability distributions, so the network is encouraged to make its output distributions as close as possible to the ‘target’ output distributions. The standard comparison measure used is **cross-entropy**, which is a measure of the similarity of two probability distributions (see e.g., Gordon-Rodriguez et al., 2020 for an introduction).

In most classification paradigms, there is good agreement amongst annotators about what the network ‘should’ have produced for each training input. Accordingly, the ‘target’ distribution is typically presented as a one-hot distribution, with all the probability mass allocated to a single output category. But it is also possible for target distributions to extend to several categories, with probability mass allocated to several different categories, or potentially all of them. A cross-entropy loss term is just as meaningful in this case. But now, the network is being trained to learn full distributions for each input item. This training paradigm is often referred to as ‘soft labels’: see e.g., Uma et al. (2021) for a recent survey.

A ‘soft labels’ training model allows a classifier to *learn about disagreement* amongst annotators. This is particularly helpful in training a classifier to identify when it can be *confident* in its output for a given item, and when it should not be confident. The confidence of a classifier is measured in many ways, but a particularly good way is to measure the **entropy** of the classifier’s output probability distribution (see e.g., Tornetta, 2021). Entropy is a measure of how much certainty (or knowledge, or information) about the output class is contained in the output distribution. If probability mass is concentrated in one category, entropy is low, and confidence is high; if it is spread over the distribution, entropy is high, and confidence is low. The entropy over the distribution of annotators’ judgements for a given item directly measures the amount of disagreement between annotators for that item: high entropy is high disagreement, low entropy is high agreement.

Even if a classifier is trained on one-hot target distributions, after training, it will still generate a distribution over categories. The entropy of this distribution can still be computed, and this entropy term can still function as a measure of the classifier’s confidence in its output for a given input. But it is likely that the entropy term will be a better measure of annotator disagreement if the network is trained on ‘soft labels’—and that is our intention in the current project. (In due course we will also test the hypothesis that training with soft labels provides a better measure of annotator disagreement than training with one-hot measures of the ‘majority’ annotator decisions.)

Our trained classifier can be used to make discrete moderation decisions. There are various possible options here. One option is simply to take the most active category as the moderation decision. But it would also be possible to use the classifier’s entropy to *adjust* decisions. In particular, if the strongest output category is ‘remove’, but *entropy is high*—suggesting annotators disagree about this item—the decision can be adjusted to ‘downrank’, to acknowledge the plurality of opinions about this item, and keep the item available for discussion.

3.3. Training a regression model to inform downranking: a role for disagreement metrics

Our regression model for downranking requires each content item to be associated with a continuous ‘harmfulness score’. As noted above, the meaningful extrema for this score can be identified by computing the average harmfulness score for items on the border between ‘remove’ and ‘downrank’ (for the highest value of downranking), and the border between ‘downrank’ and ‘untouched’ (for the lowest value of downranking). The regression model itself will again be a neural network, whose first layers are a pretrained LLM; a regression head can be added at the end of these layers.

As noted above, we ask our annotators to rank pairs of items in the training set, rather than directly provide absolute scores. We then need to convert our set of ranked pairs into a score for each item. We will use the established method for doing this: the Bradley-Terry model (Bradley-Terry, 2023).

After we have trained our regression model, it will output continuous scores, which can be used to inform downranking. But again, we suggest it is useful to *modulate* these scores for a given item using information about disagreement. We can again use the entropy of the discrete classifier for the item in question to estimate likely disagreement between annotators. We can then modify the downranking action, so that items likely to engender high disagreement are downranked a little less.

3.4. Relationship to existing harmful content annotation schemes

Many annotation schemes have been used for harmful content—see again Jahan and Ousallah (2023) for a review. Most schemes in common use require annotators to place items into discrete categories. Sometimes categories provide a coarse-grained analysis of the continuous ‘badness’ of content—for instance, a scheme might include categories such as ‘strong hate’ and ‘weak hate’. Our combined discrete and continuous annotation scheme diverges somewhat from existing schemes; whether or not it is practical remains to be seen.

For hate speech annotation, a common strategy is to use a hierarchical annotation scheme, allowing different types of offense to be identified, through a structured sequence of decisions. A commonly used scheme is that of Zampieri et al. (2019). In this scheme, annotators must first determine whether an item is ‘offensive’ or ‘not offensive’. For offensive items, a second distinction must be made, between ‘targeted’ and ‘untargeted’ offensiveness. (‘Untargeted’ offensiveness is just rudeness.) For targeted offensiveness, a further distinction must be made, as to whether the target is an individual or a group. If the target is an individual, the item can be identified as ‘bullying’. If the target is a group, the item can be identified as ‘hate speech’. This serial decision process is useful in guiding annotators towards different categories of offensiveness. Our operational annotation scheme focusses on annotations that directly inform moderation actions, rather than on distinguishing types of harmfulness, according to a prespecified taxonomy. Again, whether it's practical to leave annotators to make their own decisions about types of harm remains to be seen.

4. A Methodology for choosing Annotators

A very big question for our project is how the annotators who create the training sets for content classifiers and regression systems are *selected*. Grounding our approach in principles of deliberative democracy, we propose that the training sets used in a given location should be built locally, by people in that location and that consultation within that location should be wide, and should feature all groups within the community. Deliberative methods include citizen's assemblies such as that used in Ireland where a representative sample of citizens were selected to form a 'mini-public'. This group met to deliberate and form recommendations to the government on contentious societal issues including abortion, electoral reform and climate change. We propose building on such models in developing an approach to annotation in this project.

For hate speech, we additionally suggest that sampling should be skewed towards those who are the *targets* of hate. Thus the annotators for systems moderating hate speech against women should preferentially sample women; the annotators for systems moderating hate speech against the LGBTQIA+ community should preferentially sample people from the LGBTQIA+ community, and so on. How strong these preferential skews should be is of course a big question; we will briefly discuss how decisions can be made on such matters in Section 5.

A final point is worth mentioning here, in connection with citizens' juries. A dataset of content items annotated with assessments about harmfulness (of one kind or another) can be thought of as showing examples of how a law about harmful content can be interpreted. But it can also be thought of, more interestingly (and radically), as an *expression* of a law about harmful content. To articulate this idea, it's useful to refer to a fundamental distinction in law, between 'statutory' law and 'case law'. A statutory law is a textual document that provides a general definition of some prescribed or proscribed practice. A judge considering a given case must then decide how this generalisation applies in the case at hand. Case law works differently: here, the law resides in a collection of judgements made in the past by many judges, in some given domain. To decide how to rule in a new case, a judge has to abstract over all of these specific judgements, and work out how they generalise to the new case at hand.

We note that a collection of annotated content items bears an interesting resemblance to a body of case law. It is a set of judgements, made by many 'judges', about many particular items. As such, it carries vastly more information about the relevant categories of harmful content than is conveyed by a general textual definition. The individual decisions about individual cases collectively define vastly more subtle and nuanced definitions of these categories. If an annotated dataset of this kind functions as a (rich) body of case law, the role of the 'judge' who must abstract over these individual decisions is played by an AI content classifier (or regression system), rather than a human judge. In some ways, at least, the AI system has superhuman abilities to learn a system of generalisations over these training examples, that can be used to rule on new content items. If the AI system can do this accurately—which is an empirical question—then maybe laws about harmful content can *reside* in annotated datasets, rather than in concise generic statements.

5. Evaluating a Content Moderation Pipeline

A big question that arises for our proposed method concerns how a content moderation process should be *evaluated*. There are two subquestions here. A ‘narrow’ evaluation question concerns how the performance of a classifier and regression model can be evaluated against the datasets they were trained on. This question is easy to answer: we can simply hold out portions of these datasets from training, and then evaluate the trained systems on these held-out portions. This is precisely the kind of evaluation that is missing from the transparency reports currently provided by companies, as discussed in Section 2. As noted there, a great benefit of bringing the annotation process into the public domain is to have all companies compete on the ‘shared task’ of training systems that reproduce annotators’ collective decisions. Naturally, training data (and evaluation data) should be refreshed periodically, so evaluation would happen regularly, to keep up with changes in the form of harmful content over time.

There is also a broader evaluation question, which is much more challenging: how good is a given evaluation pipeline *overall*? The fundamental measure of goodness here relates to a whole social media platform, and its effects on users. For content moderation, this measure must centrally involve a tradeoff, between retaining freedom of expression for platform users, and dealing effectively with harmful content. Every step in the moderation pipeline can impact both measures in the tradeoff. They are affected by how a content annotation scheme is selected, how it is ‘operationalised’ in moderation decisions, how annotators are sampled from the public, how training items are gathered, and how classifier and regression models are trained. The ‘operationalising’ step involves quantitative decisions about how entropy values for items will influence boundary decisions (for classifiers) and downranking decisions (for regression models). What are the right decisions?

Note that companies must already have some way of evaluating their moderation pipelines. The evaluation measures they report publicly mainly focus on the ‘narrow’ question of classifier performance. The broader question is not confronted, as far as we know. We believe that bringing content moderation into a semi-public domain provides an opportunity to tackle the broader evaluation question for the first time.

There do seem to be suitable ways for tackling the broader evaluation question, using methods available within companies. In particular, there could be **A/B tests** of different versions of a moderation pipeline. In an A/B test, groups of users are randomly selected, and presented with different versions of some aspect of platform functionality.[4] We suggest the ultimate evaluation for a given pipeline should come from the user community, in just the same way that definitions of harmful content are sourced collectively from the public in the annotation method we propose here. For evaluation purposes, users participating in the A/B test can be asked for their opinions on harmful content proliferation on the platform, and on tolerance for free speech. Different versions of the moderation pipeline would likely lead to different responses—and these would be helpful in informing a decision about the most suitable pipeline.

Of course the pilot we present in our current report is a very small step in the direction of building (and evaluating) a complete moderation pipeline. Building a complete pipeline would require collaboration between companies and the kind of publicly-facing initiatives we are piloting in this study. We present our initial pilot study in a spirit of active exploration of this very interesting and important area.

[4] (Our project has worked a lot on A/B tests for studying the effects of recommender systems on platform users; see e.g. GPAI 2021; 2022.)

6. The Indian pilot study: a preliminary report

We are running a study in India, to trial and explore the proposal and methods outlined so far. Our decision to conduct our initial trial in India aims to highlight that hate speech is a global phenomenon, which needs attention everywhere, including in all GPAI countries, North and South. There is also evidence that hate speech is a particularly serious problem in India; for instance, India ranked in the top five countries for the number of data removal requests issued to Twitter in 2021, on several measures (Statista, 2021). The study is being run in collaboration with Jadavpur University (JU), a leading research institute based in Kolkata. The study is being led by Prof Subhadip Basu based in the Dept of Computer Science, JU.

There is already an active programme of research on the annotation and classification of offensive content in India. A forum for much of this research is the FIRE (Forum for Information Retrieval Evaluation) conference, which has a track called HASOC, dedicated to 'Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages', that has run annually since 2019. Hate speech datasets have been gathered for many relevant languages and contexts in these meetings, including datasets for Sinhala, Gujarati, Bengali, Assamese, Bodo, Hindi, Marathi, and code-mixed variants of these (including variants with English); for details, see e.g. Dowlagar et al, 2021; Mandl et al., 2019; 2020; Kumar et al., 2018; Bohra et al., 2018; Chakravarthi et al., 2020; Saroj and Pal, 2020. We hope our work will add to the work conducted in this forum.

Most datasets in HASOC are annotated using discrete schemes: either a simple 'hate'/'not-hate' scheme, or a variant on the hierarchical discrete scheme of Zampieri et al. (2019) discussed in Section 3.4. As noted in Section 3.4, the main distinctive feature of our annotation scheme is that it is set up to directly inform moderation actions available on a social media platform, rather than to position content items within a predefined taxonomy of types of harmfulness. In particular, our scheme includes a discrete category of 'borderline content' to be downranked, and a method for measuring the degree of harm of content items on a continuous scale, to inform downranking. So far, we have only completed a small pilot of the first pass of the annotation process, involving discrete content categories. In this section, we will describe the work done in this pilot to date.

The content for our initial pilot study is tweets taken from Indian political discussions, during the lead-up to two recent political elections. We focussed on these political discussions, because the elections were the occasion for considerable political violence, including fatalities [citation]. Our study recruited 10 annotators from a broad range of backgrounds. We selected two groups of tweets for annotation: 1000 themed tweets taken from the 2019 Indian general election, and 400 additional tweets taken from the 2022 state elections. Each of the tweets were annotated by all 10 annotators, on a web-based annotation platform that was purpose-built for the current project.

We begin in Section 6.1 with definitions of the discrete content categories our annotators used. We describe the annotation platform in Section 6.2. We describe the selection of annotators in Section 6.3, and give some summary information. In Section 6.4 we describe how we selected the tweets to be annotated. Statistics that explore the annotations recorded are presented and discussed in Section 6.5.

6.1. Definitions of discrete content categories for the Indian context

As discussed in Section 3.1, our discrete annotation scheme involves four content categories: ‘remove’, or ‘downrank’, ‘leave untouched’, or ‘uprank’. When we introduce these categories to annotators, we need to provide definitions that make sense to them, while accurately conveying the moderation actions (or lack thereof) associated with each category. Our annotators are drawn from a range of backgrounds, so definitions cannot be complex. Here are the definitions we provided for the key categories that relate to hate speech moderation proper.

- **Remove**: These are tweets which are derogatory, discriminatory, abusive or offensive in nature and need to be removed so that it doesn't influence other users in an extremely negative sense.
- **Downrank**: These are tweets that convey a defaming or upsetting message to the reader and provoke a sense of division at some level should not be suggested but rather be down-ranked by a recommender system.
- **Neutral**: These are the tweets which pass a generic message to the reader. It is mostly informative or declarative where the tweets mainly represent events or a statement or feeling of the user without targeting or discriminating against anyone.

As also noted in Section 3.1, we included a fourth category of content, which is not related to hateful content, but to its opposite: namely content that encourages mature debate, and reconciliation between conflicting groups. (Chakravarthi et al.’s 2021 concept of ‘hope speech’ identifies a similar category in the Indian YouTube space.) Content of this kind can be given special attention by social media platforms, in a positive attempt to bring opposing groups together. Again the recommender system is a key instrument: content of this kind can be ‘upranked’, so it is disseminated more widely than it would otherwise have been. Some platforms experiment with this kind of social engineering: Polis is a well-known example (see e.g., Small et al., 2023 for a recent discussion). Our project is focussed on responding to harmful content, but we include a category related to upranking, as an experimental part of our initial pilot. Here is the definition we gave our annotators for this category.

- **Uprank**: [definition]These tweets should have an uplifting/motivational message. Such messages can be generic and likely to invoke a positive emotion from the reader.

Recall from Section 3.1 that because our content categories are defined ‘operationally’, with explicit reference to moderation actions, they elicit annotators’ intuitions on the topic of free speech as well as their value judgments about content. The key labels for our content categories are moderation actions, to keep their operational nature in the forefront of annotators’ minds.

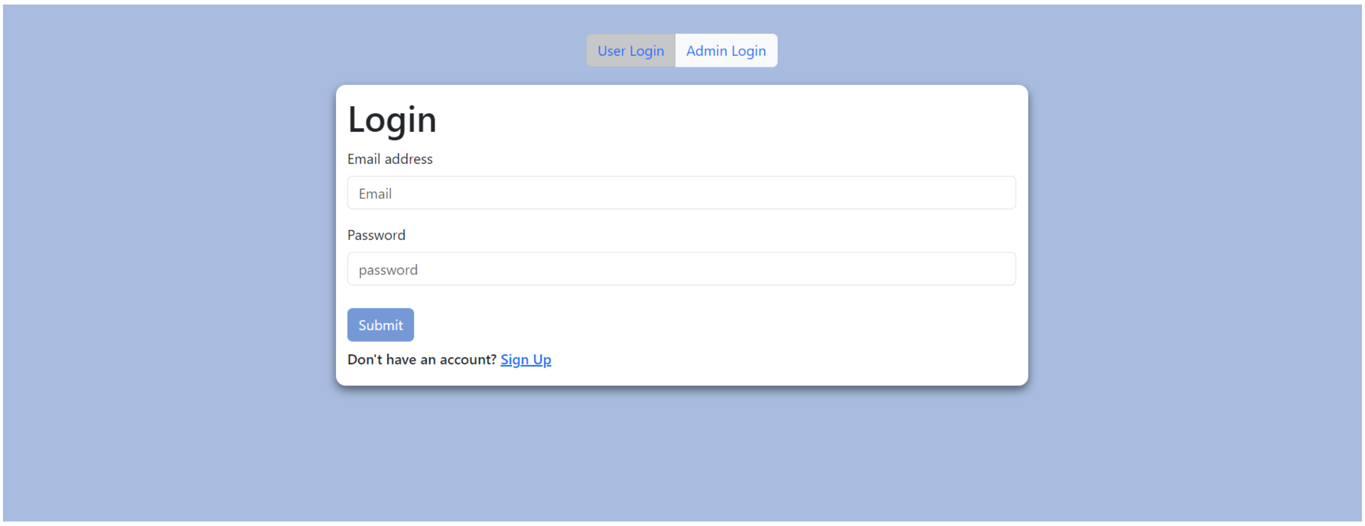
The annotators were provided with a 2-page ‘guide’ on the landing page of the annotation platform, that contains the definitions of the actions, plus few illustrative examples of scenarios for each action. (We include this guide as an appendix.) The guide is also quickly accessible from each page of the annotation form through a quick click of a button. Additionally, hovering over the action options for each tweet on the annotation form pops up a short definition for the action for quick reference.

6.2. An Annotation platform for the project

To give us maximum flexibility in delivering an annotation system, we developed a web based interface that was used by all the annotators. In this section we show the functionality of this interface, first from the perspective of users (Section 6.2.1) and then from the perspective of administrators/analysts (Section 6.2.2).

6.2.1. The User Interface

The user *login screen* is shown in Figure 6.1.



Login

Email address

Email

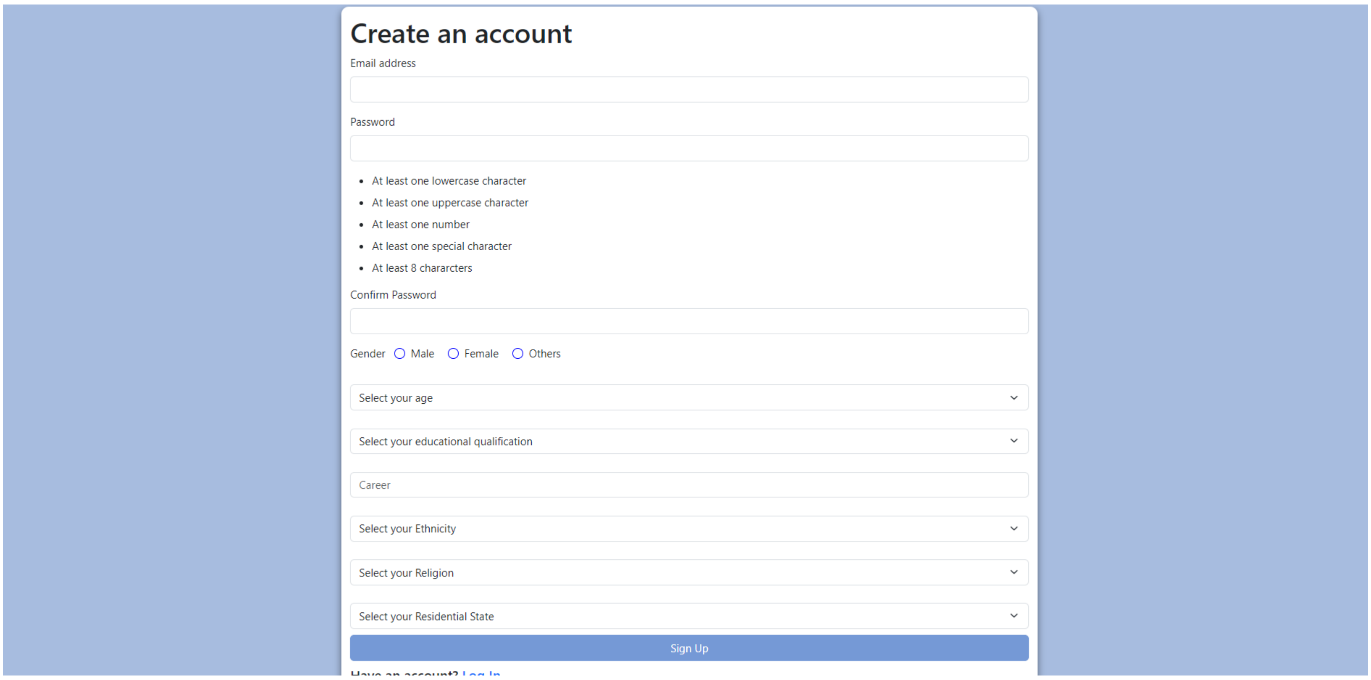
Password

password

Don't have an account? Sign Up

Figure 6.1: Login Page

The user *sign-up page* is shown in Figure 6.2. Required fields are those mentioned in the annotator table in Section 6.3.



Create an account

Email address

Password

- At least one lowercase character
- At least one uppercase character
- At least one number
- At least one special character
- At least 8chararcters

Confirm Password

Gender ○ Male ○ Female ○ Others

Select your age

Select your educational qualification

Career

Select your Ethnicity

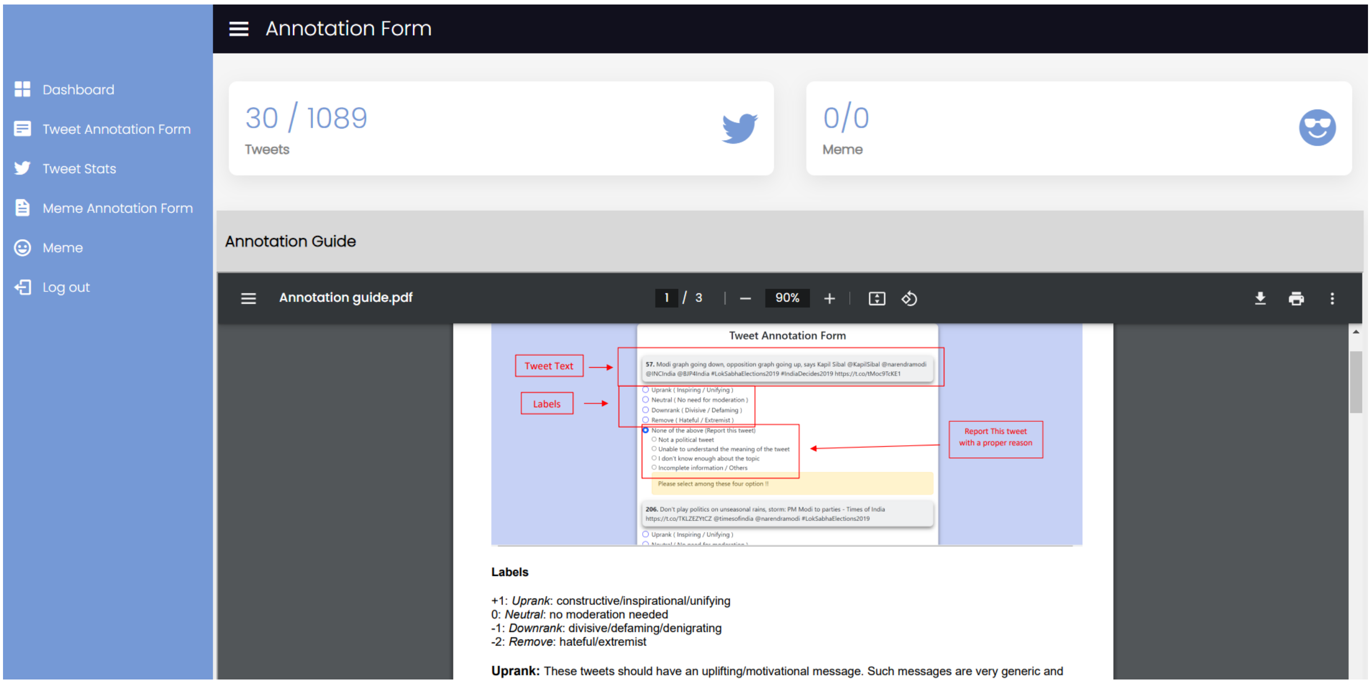
Select your Religion

Select your Residential State

Have an account? Log In

Figure 6.2: Sign-up Page

The user *dashboard/homepage* is shown in Figure 6.3. This is the landing page of the annotation interface. It contains an annotation guide, that provides the definitions of each of the discrete labels, plus several illustrative examples, to help out a first-time user. (The annotation guide is included in Appendix 1.) The dashboard/homepage also shows the total number of tweets in the database, with a running count of how many the user has annotated till now.



57. Modi graph going down, opposition graph going up, says Kapil Sibal @KapilSibal @narendramodi @INCIndia @BJP4India #LokSabhaElections2019 #IndiaDecides2019 https://t.co/tMoc9TcKE1

○ Uprank (Inspiring / Unifying)
○ Neutral (No need for moderation)
○ Downrank (Divisive / Defaming)
○ Remove (Hateful / Extremist)
● None of the above (Report this tweet)
○ Not a political tweet
○ Unable to understand the meaning of the tweet
○ I don't know enough about the topic
○ Incomplete information / Others

Please select among these four option !!

206. Don't play politics on unseasonal rains, storm: PM Modi to parties - Times of India https://t.co/TKLZEZYtCZ @timesofindia @narendramodi #LokSabhaElections2019

○ Uprank (Inspiring / Unifying)

Labels

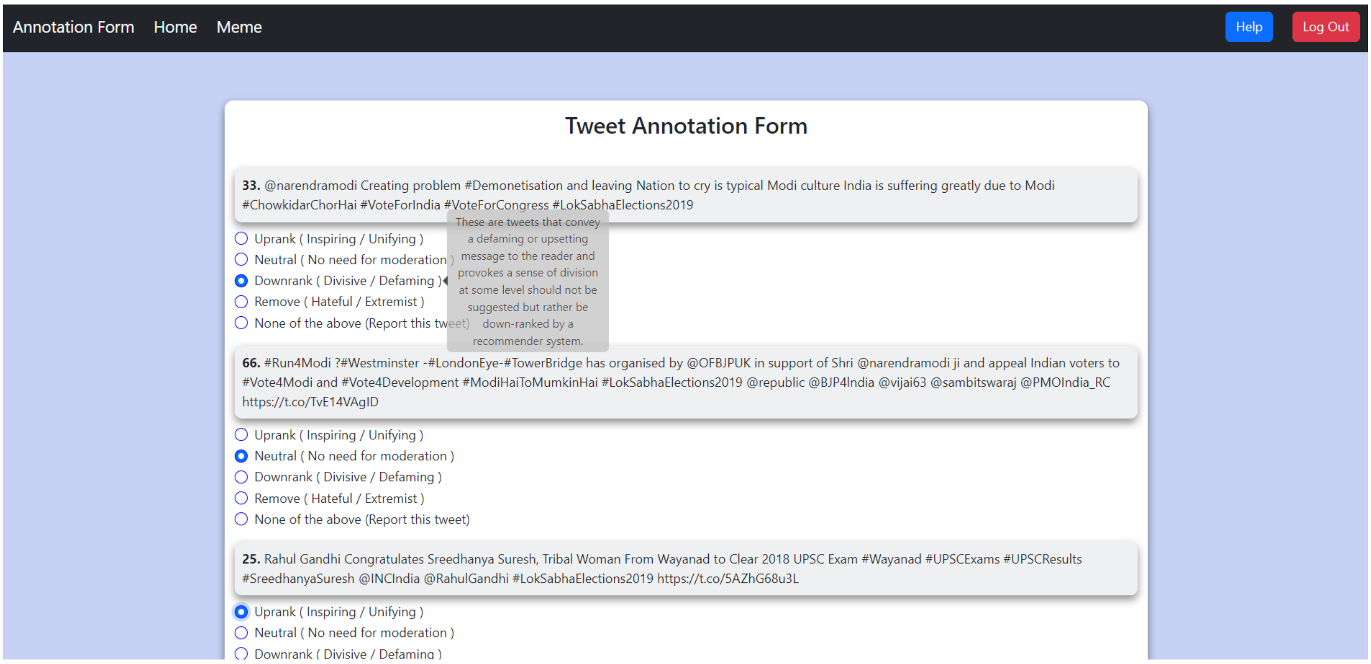
+1: *Uprank*: constructive/inspirational/unifying
0: *Neutral*: no moderation needed
-1: *Downrank*: divisive/defaming/denigrating
-2: *Remove*: hateful/extremist

Uprank: These tweets should have an uplifting/motivational message. Such messages are very generic and

Figure 6.3: User Dashboard

The user's *annotation form page* is shown in Figure 6.4. A total of 10 tweets appear on a single page. Our intention here is not to overcrowd the page and overwhelm the annotator.

Note that the discrete annotation options in our tool include an option of '***None of the above'.*** This is available to the user for cases in which he/she is not sure of the label for various reasons. In this case, they can 'report' the item (in the terminology we use). If a user 'reports' an item they are required to choose from a number of suboptions which are detailed in the annotation guide: for instance, that the tweet is 'not political', or that the user does not understand its meaning. A full list of options is given in our appendix. The user guide is also available to the user from this page, if they click on the *Help* button on the top right.



Tweet Annotation Form

33. @narendramodi Creating problem #Demonetisation and leaving Nation to cry is typical Modi culture India is suffering greatly due to Modi #ChowkidarChorHai #VoteForIndia #VoteForCongress #LokSabhaElections2019

- ○ Uprank (Inspiring / Unifying)
- ○ Neutral (No need for moderation)
- ● Downrank (Divisive / Defaming)
- ○ Remove (Hateful / Extremist)
- ○ None of the above (Report this tweet)

These are tweets that convey a defaming or upsetting message to the reader and provokes a sense of division at some level should not be suggested but rather be down-ranked by a recommender system.

66. #Run4Modi ?#Westminster -#LondonEye-#TowerBridge has organised by @OFBJPUK in support of Shri @narendramodi ji and appeal Indian voters to #Vote4Modi and #Vote4Development #ModiHaiToMumkinHai #LokSabhaElections2019 @republic @BJP4India @vijai63 @sambitswaraj @PMOIndia_RC https://t.co/TvE14VAgID

- ○ Uprank (Inspiring / Unifying)
- ● Neutral (No need for moderation)
- ○ Downrank (Divisive / Defaming)
- ○ Remove (Hateful / Extremist)
- ○ None of the above (Report this tweet)

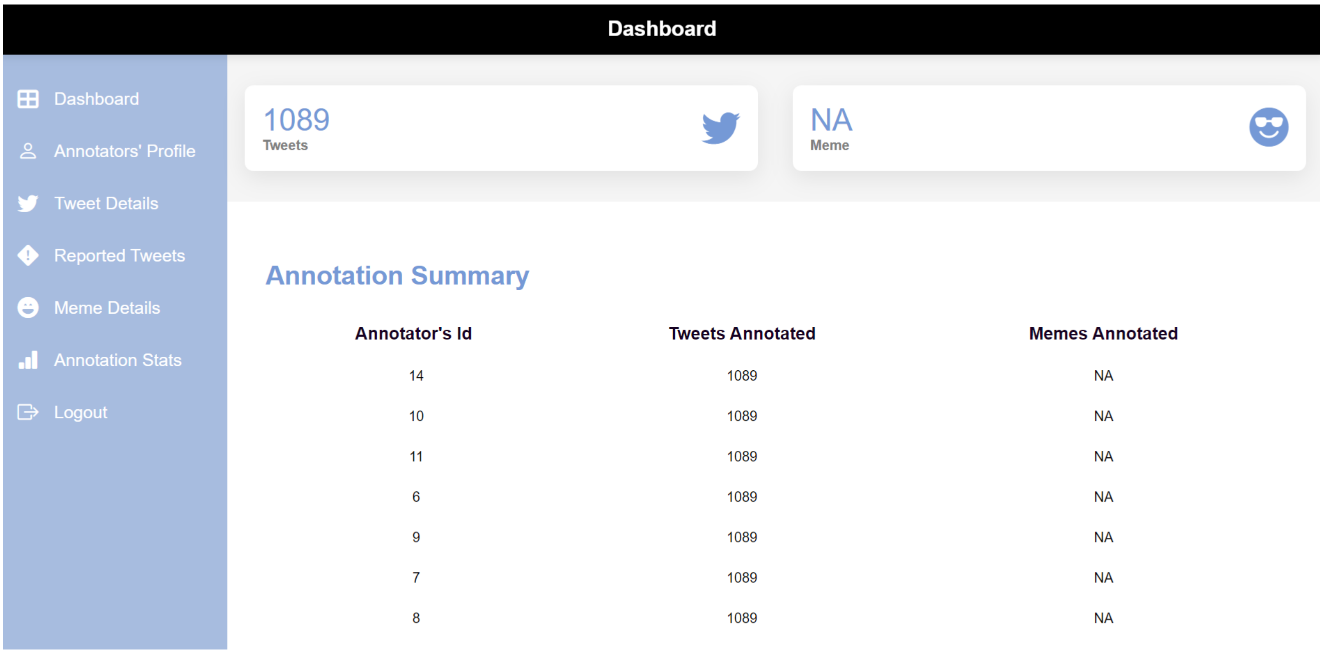
25. Rahul Gandhi Congratulates Sreedhanya Suresh, Tribal Woman From Wayanad to Clear 2018 UPSC Exam #Wayanad #UPSCExams #UPSCResults #SreedhanyaSuresh @INCIndia @RahulGandhi #LokSabhaElections2019 https://t.co/5AZhG68u3L

- ● Uprank (Inspiring / Unifying)
- ○ Neutral (No need for moderation)
- ○ Downrank (Divisive / Defaming)

Figure 6.4: Annotation Form Page

6.2.2. The Administrator Interface

The administrator *dashboard/homepage* is shown in Figure 6.5. This page shows the total number of tweets and memes that are present in the database and the number of total tweets annotated by each annotator. Annotators are referred to by an anonymous (but unique) ID, so their identity is not disclosed to the administrator.



1089 Tweets

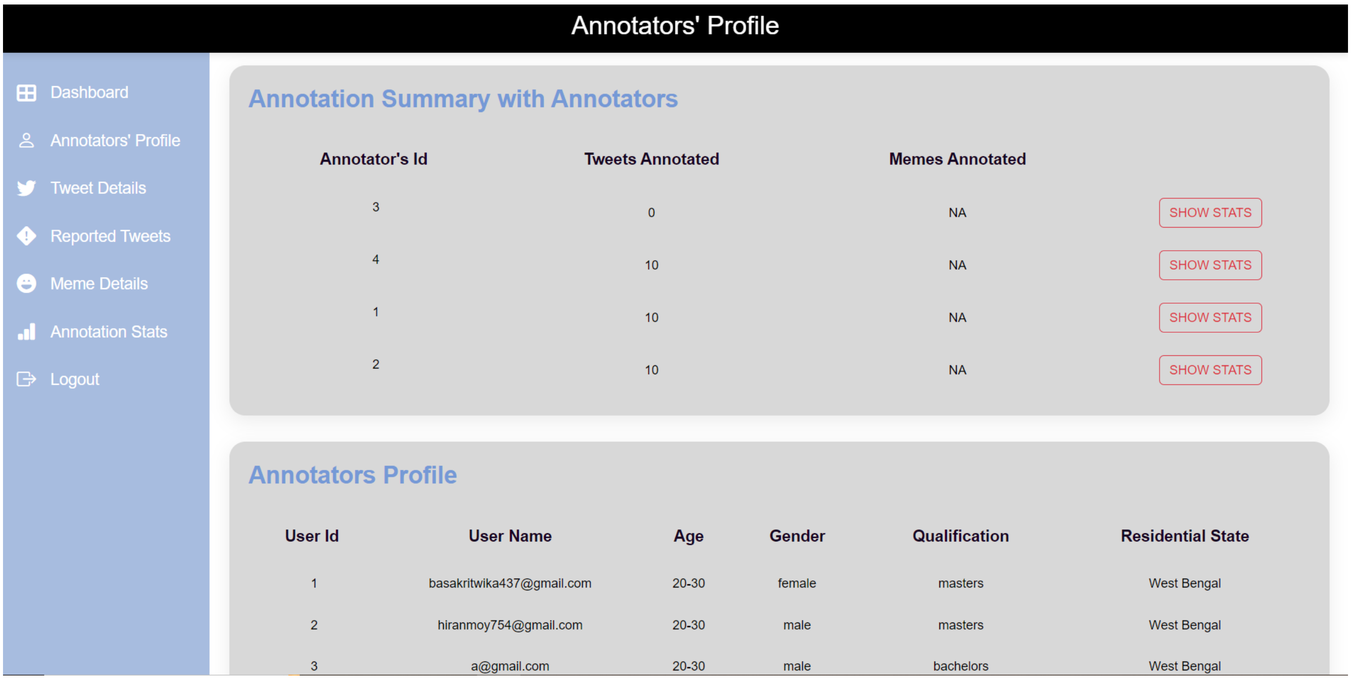
NA Meme

Annotation Summary

Annotator's Id	Tweets Annotated	Memes Annotated
14	1089	NA
10	1089	NA
11	1089	NA
6	1089	NA
9	1089	NA
7	1089	NA
8	1089	NA

Figure 6.5: Admin Dashboard

The *annotators' profile page* is shown in Figure 6.6. This page shows the details of all the annotators that have registered so far (on the User Signup page shown above). The page also allows the administrator to check the individual annotation statistics of each annotator by clicking on the associated *Show Stats* link.


Annotators' Profile

Annotation Summary with Annotators

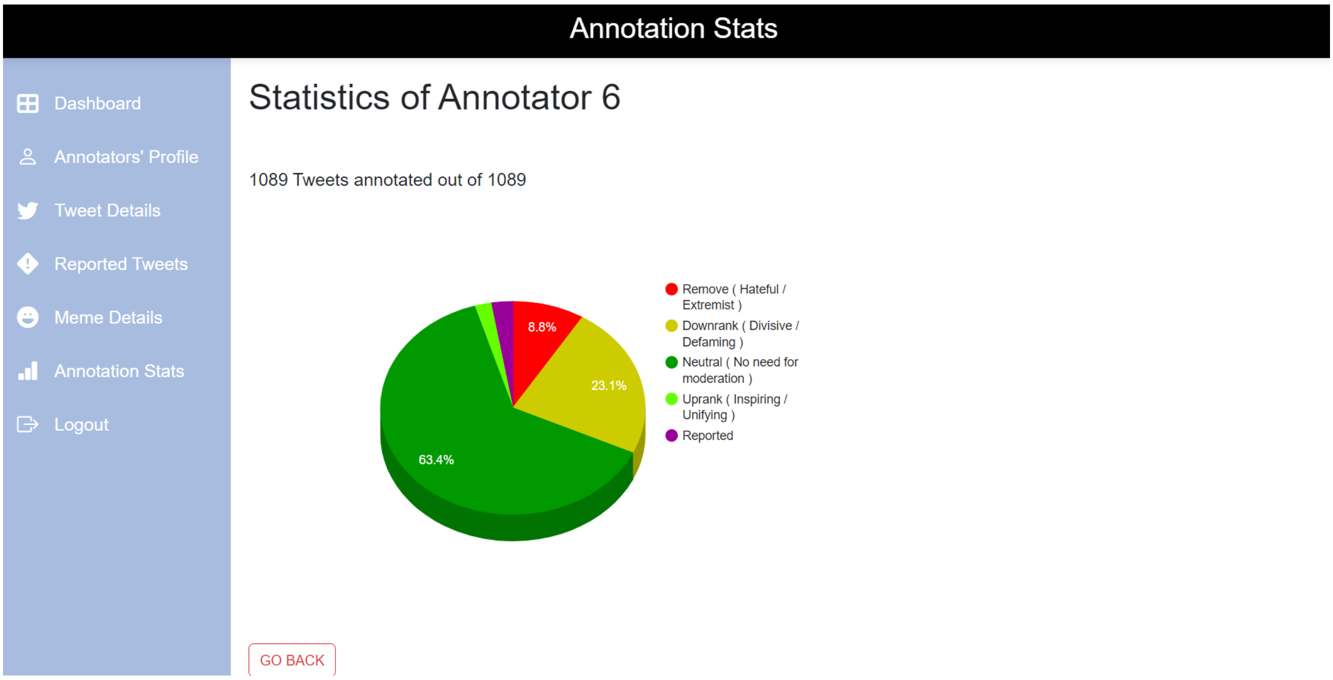
Annotator's Id	Tweets Annotated	Memes Annotated	
3	0	NA	SHOW STATS
4	10	NA	SHOW STATS
1	10	NA	SHOW STATS
2	10	NA	SHOW STATS

Annotators Profile

User Id	User Name	Age	Gender	Qualification	Residential State
1	basakritwika437@gmail.com	20-30	female	masters	West Bengal
2	hiranmoy754@gmail.com	20-30	male	masters	West Bengal
3	a@gmail.com	20-30	male	bachelors	West Bengal

Figure 6.6: Annotators' Profile Page

The *individual stats page* for a selected annotator is shown in Figure 6.7. This page represents the annotations given by a particular annotator in the form of a pie chart.


Annotation Stats

Dashboard
Annotators' Profile
Tweet Details
Reported Tweets
Meme Details
Annotation Stats
Logout

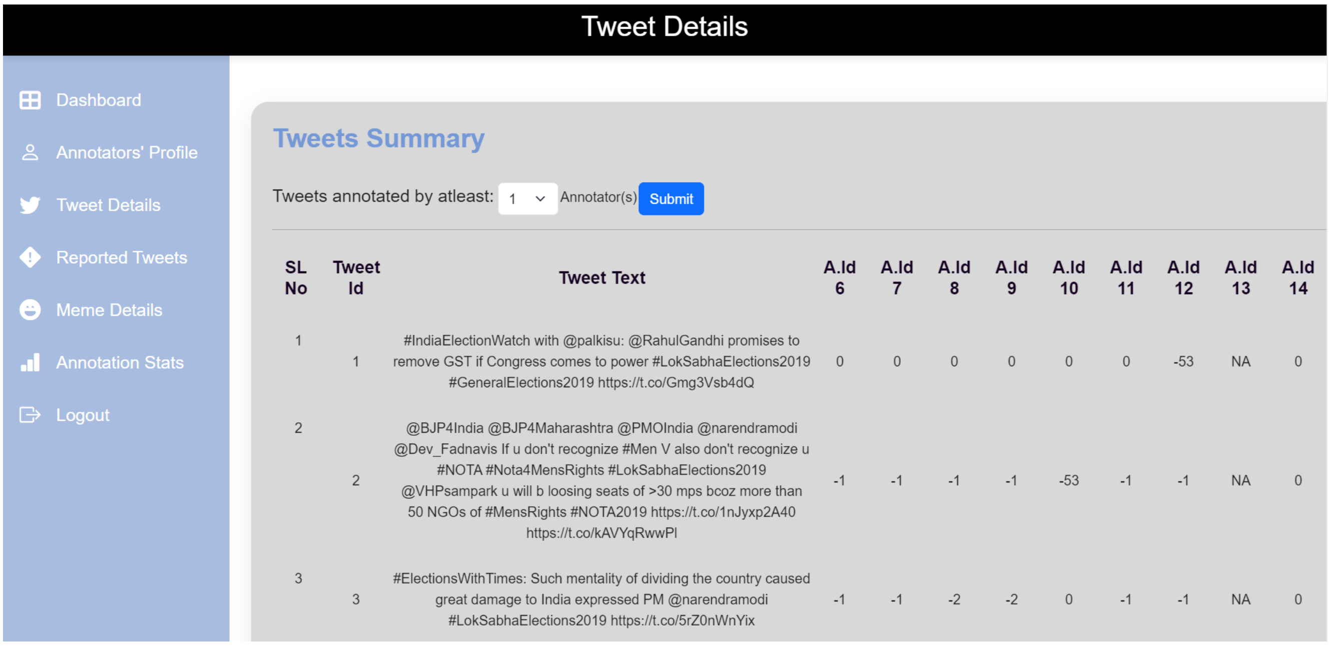
Statistics of Annotator 6

1089 Tweets annotated out of 1089

8.8%
23.1%
63.4%
● Remove (Hateful / Extremist)
● Downrank (Divisive / Defaming)
● Neutral (No need for moderation)
● Uprank (Inspiring / Unifying)
● Reported
GO BACK

Figure 6.7 Individual Annotator Stats

The *tweet details* page is shown in Figure 6.8. This page shows the administrator the labels provided for each tweet by each annotator. Annotators are denoted by ID, as before. Tweets are denoted by ID, and by content.



Tweet Details

Tweets Summary

Tweets annotated by atleast: 1 Annotator(s)

SL No	Tweet Id	Tweet Text	A.Id 6	A.Id 7	A.Id 8	A.Id 9	A.Id 10	A.Id 11	A.Id 12	A.Id 13	A.Id 14
1	1	#IndiaElectionWatch with @palkisu: @RahulGandhi promises to remove GST if Congress comes to power #LokSabhaElections2019 #GeneralElections2019 https://t.co/Gmg3Vsb4dQ	0	0	0	0	0	0	-53	NA	0
2	2	@BJP4India @BJP4Maharashtra @PMOIndia @narendramodi @Dev_Fadnavis If u don't recognize #Men V also don't recognize u #NOTA #Nota4MensRights #LokSabhaElections2019 @VHPsampark u will b loosing seats of >30 mps bcoz more than 50 NGOs of #MensRights #NOTA2019 https://t.co/1nJyxp2A40 https://t.co/kAVYqRwwPl	-1	-1	-1	-1	-53	-1	-1	NA	0
3	3	#ElectionsWithTimes: Such mentality of dividing the country caused great damage to India expressed PM @narendramodi #LokSabhaElections2019 https://t.co/5rZ0nWnYix	-1	-1	-2	-2	0	-1	-1	NA	0

Figure 6.8: Tweet Details Page

6.3. Choice of Annotators

In the current pilot our group of annotators is extremely small, so we're not in a position to apply a fully-specified sampling methodology, of the kind outlined in Section 4. Our current concern was simply to pick a set of annotators from different ages, genders, educational levels, career positions, ethnicities, religions and geographic regions.

Demographics of the ten annotators who participated in the pilot study are shown in Table 6.1. Again, annotator names are redacted for anonymity.

Table 6.1: Summary demographics for the ten annotators in the pilot study

Annotator. ID	Age	Sex	Education	Career Status	Ethnicity	Religion	State
A1	22	M	BE	Student	General	Hinduism	West Bengal
A2	24	F	BSc	Student	General	Hinduism	West Bengal
A3	24	M	BSc	Student	SC	Hinduism	West Bengal
A4	29	F	MTech	Mid Career IT Professional	General	Hinduism	West Bengal
A5	45	M	BSc	Mid Career Content/ Digital Marketeer	General	Hinduism	Punjab
A6	46	M	BE	Business	General	Hinduism	Karnataka/UttarPradesh
A7	20-30	M	Mtech	Mid career IT professional	General	Hinduism	India/USA
A8	20-30	M	Bachelors	Software Engineer	General	Islam	Karnataka
A9	20-30	M	Bachelors	Tech support	General	Christian	Karnataka
A10	30-40	M	Masters	MBA	General	Hinduism	Haryana

6.4. Choice of Datasets to annotate

India being the largest democracy in the world, the elections in this country serve as pivotal moments that reflect the collective voice of more than a billion citizens. The electoral system comprises two major types of elections: the general elections, also known as Lok Sabha elections, and the state elections, referred to as Vidhan Sabha elections. The Lok Sabha elections determine the composition of the lower house of the Indian Parliament, where Members of Parliament (MPs) are elected to represent the diverse constituencies across the country. These elections are conducted every five years, allowing citizens to participate in shaping the central government. On the other hand, Vidhan Sabha elections are held at the state level to elect members to the legislative assemblies. Each state in India has its own Vidhan Sabha, and these elections occur periodically, determining the governance of individual states. Both types of elections play a crucial role in shaping the political landscape of India, providing citizens with the opportunity to voice their preferences and contribute to the democratic governance of the country.

In this initial pilot study, we have focussed our attention on content written in the lead-up to both types of election, to assess these distinct yet interconnected electoral arenas. Our focus spans the 2019 Lok Sabha elections (the most recent general elections with the next one coming up in 2024), a crucial moment shaping the central government, and the 2022 Vidhan Sabha elections, a granular examination spanning five diverse states. Our current pilot focusses on content taken from Twitter. Our basic process is to analyse the hashtags which are most likely to yield hateful or controversial content, and select tweets from these hashtags.

In this preliminary pilot study, we consider individual tweets in isolation, without considering their context. There are good arguments that context is important, for annotation and classification; see for instance Nagar et al. (2022). Our future work is likely to incorporate some analysis of context.

6.4.1. The 2019 Lok Sabha Elections

Our first dataset[5] from which we have curated 607 out of the 1007 tweets is centered around the 2019 Lok Sabha Elections covering all of the 29 states and 2 out of the 7 union territories. (The remaining 5 union territories had negligible data, less than 15 tweets.) Apart from the main field of tweet text we also have fields recording the number of likes and number of retweets for each tweet (but note that this information was not presented to the annotators). These additional fields can help us interpret and analyze the extent to which each tweet was influential, and the degree by which each tweet was in circulation giving us valuable insights into the mindsets and ideologies of citizens at the digital forefront of Indian political discussions.

In order to curate the tweets for our study we extracted and studied the distribution of the hashtags (‘#’) as well as the tags (‘@’) present in all the tweet texts as these demonstrate the context in which a user had written the tweet.

[5] Sourced from a Kaggle dataset titled Indian Political Tweets 2019 (Feb to May) which gives us the tweets tweeted everyday at a specific time of the day in the mentioned time period just before the declaration of the election results.

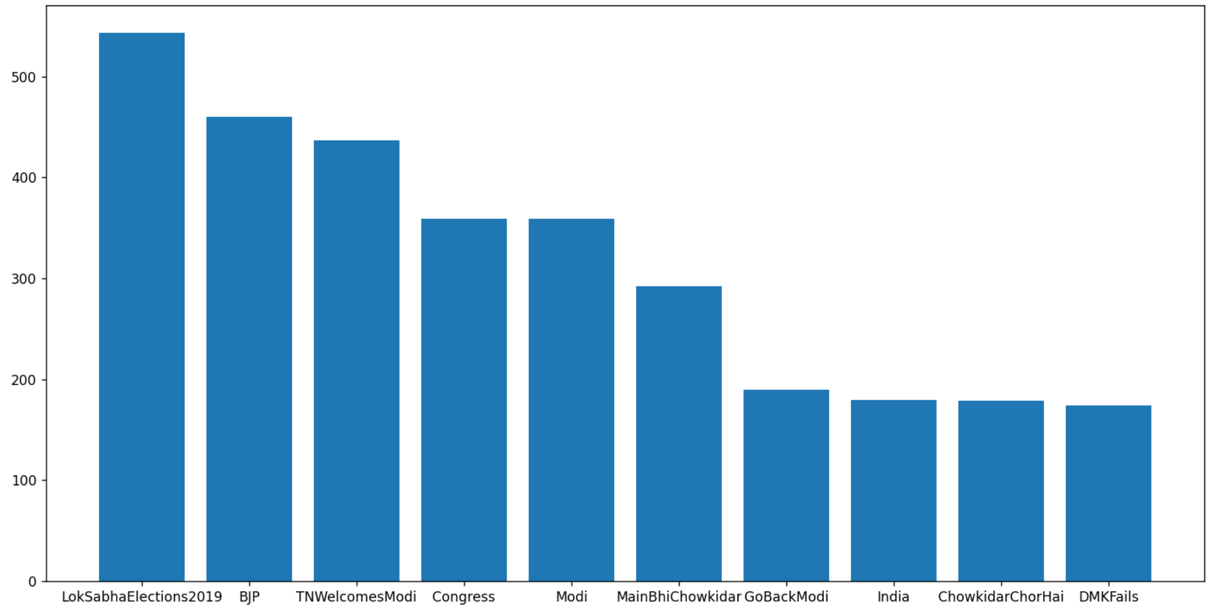


Figure 6.9: Hashtag frequency distribution in the first set

As shown in the histogram in Fig 6.9, we study a total of 10 hashtags ('#') which occur with the highest frequencies in the tweet texts. We consider this distribution and carefully select appropriate hashtags to deliver the most relevant content: this is our first stage of filtering.

As a starting point, we consider the leading hashtag in terms of frequency in the dataset, ***"LokSabhaElections2019"***. The total number of tweets which have this particular hashtag is around 550. For the second stage of filtering, we extract and analyze the distribution of the tags('@') in this particular set of about 550 tweets. The distribution is shown in Figure 6.10.

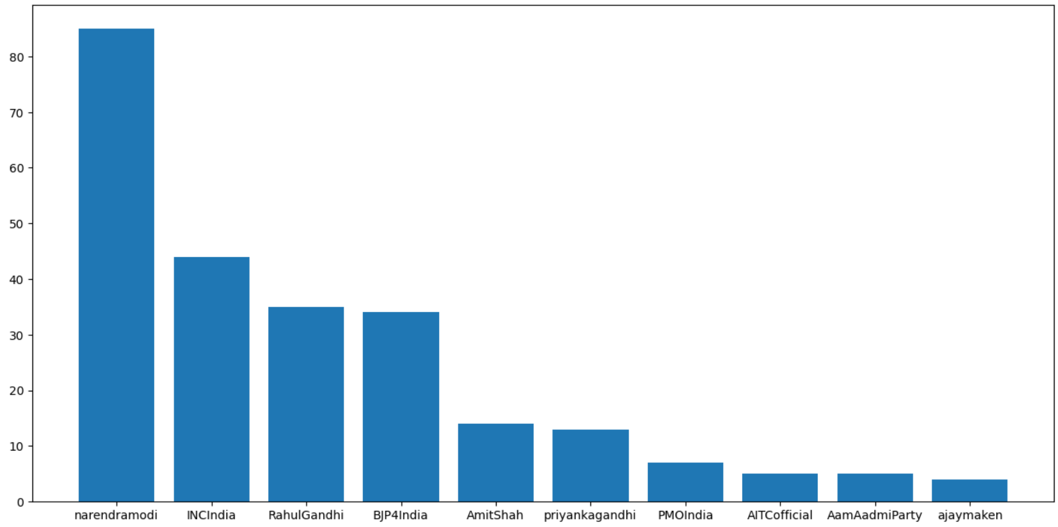


Figure 6.10: Tag frequency distribution in the first filtered subset

An interesting although predictable trend is visible in Fig 6.10: the tags (official Twitter handles) with highest frequencies are those of the two most prominent political parties (BJP and Congress), along with the three main political leaders of each: in India at the central level who have historically been in power in the Parliament and have shaped India's political narrative. Therefore we consider the leading 6+1=7 tags (since the 7th most used tag in the data is technically equivalent to the 1st) to extract the first subset which comes out to be 150 tweets in count.

For the rest of the tweets in the 2019 Lok Sabha Elections set, we turn our attention to three very relevant hashtags which mainly concern the respected prime minister of India - ***"MainBhiChowkidar"*** , ***"GoBackModi"*** and ***"ChowkidarChorHai"*** with these three hashtags coming in at 6th, 7th and 9th among the top 10 leading hashtags. Combining these and removing any tweet already included in the first subset as detailed above, we are left with around 600 tweets. Next, we curate two different subsets from these 600 tweets using two different filtering procedures.

As mentioned above for the first of the two (and second overall) subsets, on similar lines to the previous subset we extract and study the distribution of the tags ('@') which is as illustrated below -

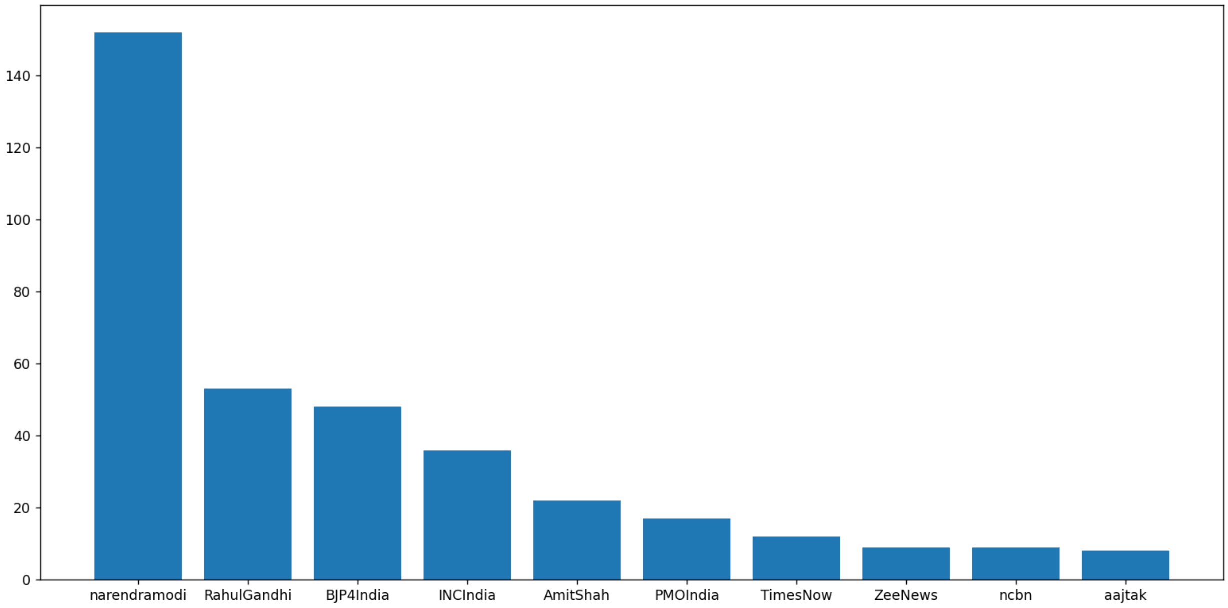


Figure 6.11: Tag frequency distribution in the second filtered subset

Analogous to Fig 6.10, the histogram of tags ('@'), Fig 6.11 also demonstrates that among the 10 leading tags sorted by frequency the top 6 are those of the aforementioned political parties and their prominent leaders. Thus, similar to our previous "tag filtering strategy", we consider these 6 tags present in the tweets and after this filtering the count of tweets comes out to be 207 which becomes our second subset.

Finally for the last subset pertaining to the 2019 Lok Sabha Elections, we look at the same three hashtags as the last subset considering their relevance to our objective. As the second stage of filtering instead of looking at the tags, we turn our attention to a measure of influence of those tweets, the number of retweets that a tweet receives. This data is plotted in Figure 6.12, in the form of a histogram where the horizontal axis is the number of retweets while the vertical axis is the count of the tweets with a particular number of retweets. As illustrated in this figure, the majority of tweets have a retweet count of under 4000, but a few exceed that figure. Our selected set of tweets thus includes a number of quite influential tweets, as measured by retweet count.

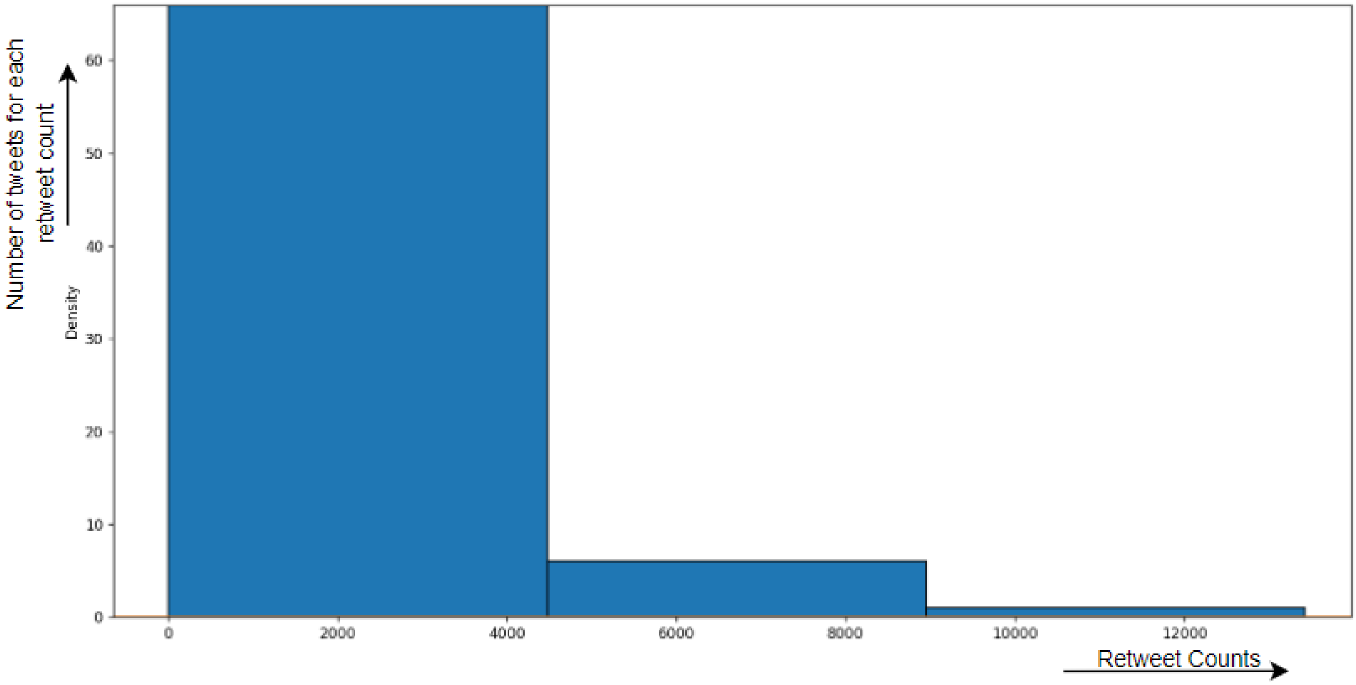


Figure 6.12: Plot for the number of tweets as a function of retweet count

As a final filter, we sort all the tweets (excluding those already included in the previous subsets) according to the retweet count in descending order, and choose the top 250 ones for our annotation set.

In summary: we curated a total of 607 tweets in the context of the 2019 Lok Sabha elections. These were divided into three subsets of 150, 207 and 250 with each subset created using distinct filtering strategies.

6.4.2. Indian State Elections 2022

Our second dataset,[6] from which we have drawn 400 tweets from our total set of 1007, is centered around various state elections which took place in 2022. Tweets pertaining to the state elections of a total of 5 states have been included. We chose 5 states to try to ensure some regional diversity: the selected states cover a wide geographic area spanning from *Punjab, Uttarakhand* and *Uttar Pradesh* in the north to *Goa* in the south-west and *Manipur* in the north-east. All the tweets for each statewere extracted from Twitter using the hashtag *"#StateElections2022"* where State is one of the 5 states, e.g. *"#GoaElections2022"* or *"#PunjabElections2022"*. To curate the 400 tweets we used a similar filtering strategy as for the 2019 Lok Sabha Elections data. For each state we used a 2-stage filtering using tags ('@') first and then hashtags ("#") and thereafter selected 80 tweets from each state's set appropriately. Since in this section we have diverted our attention from the central level to the state level we have included tweets in languages other than English, with a substantial number of tweets in Hindi and a handful in Bengali as well.

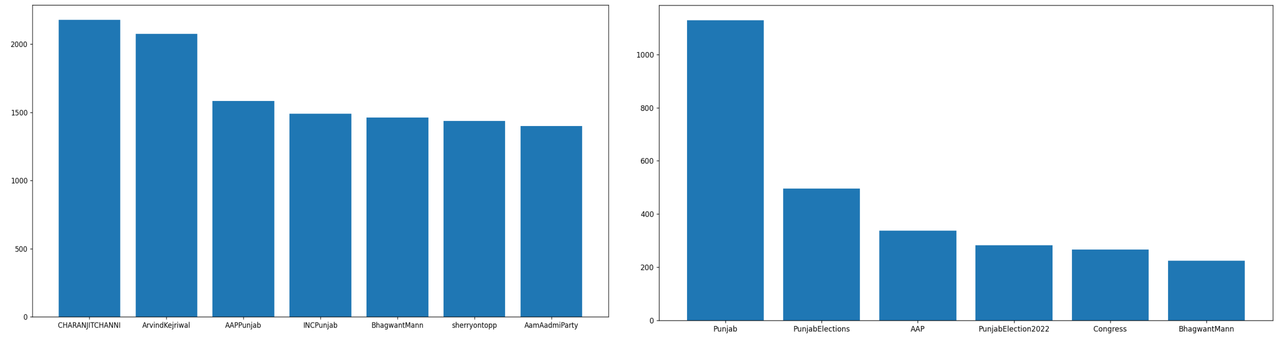
For every state, we researched the respective elections and focussed on the leading 2–3 political parties that secured the highest number of votes. As the first filtering stage for curation, for each state we considered the tags ('@') of the official Twitter handles of the political parties in leading contention for the seat of the Chief Minister of that state, along with their main representatives.

[6] Sourced from a Kaggle dataset titled Indian State Elections 2022 Twitter Dataset which gives us the tweets tweeted in the context of the state elections of 5 different states from 1/11/2021 to 9/3/2022, i.e., up to 1 day before the actual results were finalised.

These tags are also among the leading tags in terms of their frequency in the tweet texts. As the second stage of filtering we chose relevant hashtags, which range from political parties to social media accounts and news channels which are quite influential in the respective states.

In the remainder of this section we present histograms of tag distributions for each of the five states we studied. In each histogram, the horizontal axis denotes the tags/hashtags represented by the texts and the vertical axis shows the count of tweets with the respective tag/hashtag.

a) **Punjab**. As shown in Fig 6.13, first we analyze the tag distribution and select 4 of the leading tags, *'CHARANJITCHANNI', 'AAPPunjab', 'INCPunjab', 'BhagwantMann'*. Next, we choose two relevant hashtags, *“AAP”* and *“Congress"* after which we are left with around 500 tweets out of which we select 80 appropriate ones manually.

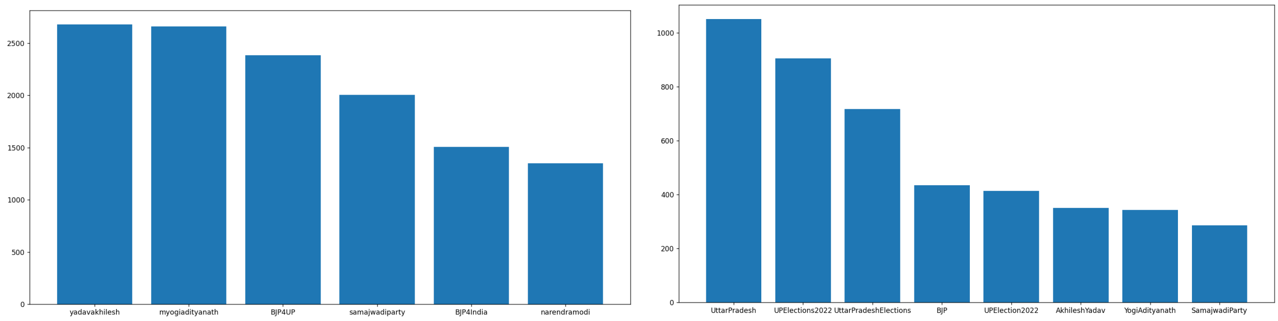


Tag Distribution

Hashtag Distribution after Tag Filtering

Figure 6.13: Two stage filtering for the tweets on state of Punjab

b) **UttarPradesh**. As shown in Fig 6.14, first we analyze the tag distribution and select 4 of the leading tags, *'yadavakhilesh', 'myogiadityanath', 'BJP4UP', 'samajwadiparty'*. Next, we choose two relevant hashtags, *“BJP”* and *“SamajwadiParty”* after which we are left with around 650 tweets out of which we select 80 appropriate ones manually.

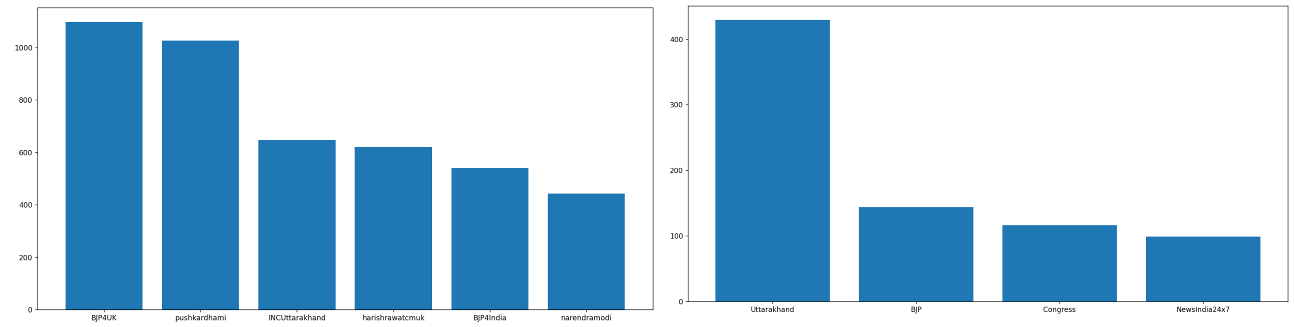


Tag Distribution

Hashtag Distribution after Tag Filtering

Figure 6.14: Two stage filtering for the tweets on state of UttarPradesh

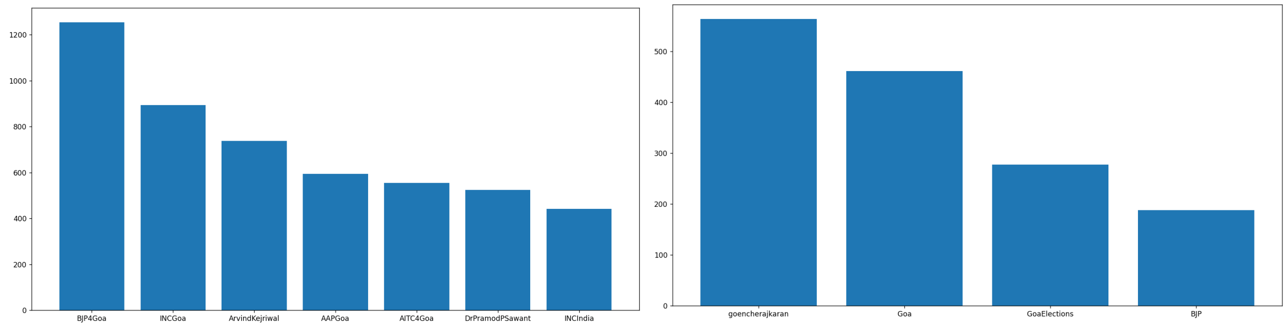
c) **Uttarakhand**. As shown in Fig 6.15, first we analyze the tag distribution and select 4 of the leading tags, *'BJP4UK', 'pushkardhami', 'INCUttarakhand', 'harishrawatcmuk'*. Next, we choose two relevant hashtags, *'BJP', 'Congress'* after which we are left with around 220 tweets out of which we select 80 appropriate ones manually.



Tag Distribution Hashtag Distribution after Tag Filtering

Figure 6.15: Two stage filtering for the tweets on state of Uttarakhand

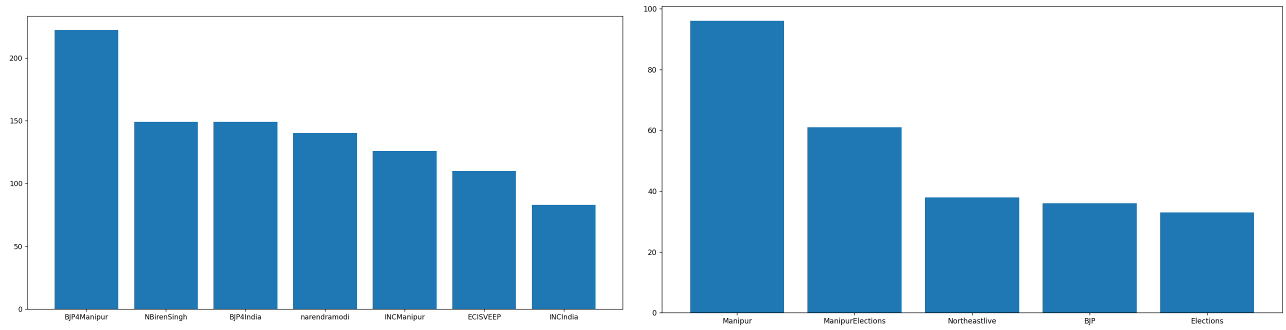
d) **Goa**. As shown in Fig 6.16, first we analyze the tag distribution and select 5 of the leading tags, *'BJP4Goa', 'INCGoa', 'AAPGoa', 'AITC4Goa', 'DrPramodPSawant'*. Next, we choose the leading single hashtag, *“goencherajkaran”* which is a social media account for Goan politics, after which we are left with around 500 tweets out of which we select 80 appropriate ones manually.



Tag Distribution Hashtag Distribution after Tag Filtering

Figure 6.16: Two stage filtering for the tweets on state of Goa

e) **Manipur**. As shown in Fig 6.17, first we analyze the tag distribution and select 3 of the leading tags, *'BJP4Manipur', 'INCManipur', 'NBirenSingh'*. Next, we choose three relevant and leading hashtags, *'Manipur', 'ManipurElections', 'Northeastlive'* the last one being a news channel dedicated for the north east region, after which we are left with around 140 tweets out of which we select 80 appropriate ones manually.



Tag Distribution Hashtag Distribution after Tag Filtering

Figure 6.17: Two stage filtering for the tweets on state of Manipur

In summary: we have curated a dataset comprising a total of 1007 tweets, covering both the central as well as state elections.

6.5. Results of the annotation study

Our initial pilot study is obviously too small to provide meaningful evidence about attitudes of the Indian public towards the tweets in our dataset. In our analysis we have two more modest goals. Firstly, we would like to gauge how much *disagreement* we can expect if we sample a larger group of annotators. We would like to know two things:is there much *variance* over tweets in the amount of disagreement between annotators? And what are the upper and lower bounds on disagreement? We are interested in early information on these points, because the moderation scheme we proposed in Section 3 treats content items differently depending on the amount of disagreement the classifier predicts they would incur amongst annotators. Secondly, we would like a reality check about the collective ability of a group of annotators to provide appropriate moderation decisions. Does the group collectively make sensible decisions? Of course, there is no objective gold standard here—but individual readers and assessors can nonetheless take a view on this.

6.5.1. Brief annotation results on the tweets of 2019 Lok Sabha Elections

The first part of our training dataset contains **607 tweets** (see Section 6.4.1). It was annotated by a total of 10 annotators and the findings are analysed in the current section. Our analysis focuses on the disagreement that was found between annotators. We will analyse this disagreement in a number of ways, finishing with the entropy measure discussed in Section 3.2.

For analysis purposes, we first separate out the tweets that were ‘reported’ (i.e., flagged as unclassifiable) by a majority of annotators. There were 9 of these. The remaining tweets were further analysed into three discrete classes, based on the amount of agreement between the annotators:

- Unanimous Agreement (109 tweets)
- Disagreement by a degree of 1 (326 tweets)
- Disagreement by a degree of 2 (145 tweets)
- Disagreement by a degree of 3 (18 tweets).

These results are depicted in Figure 6.18.

As obvious from the above numbers, and from the figure, there are varying *degrees* of disagreement across this dataset—including some tweets for which there is significant disagreement. This result is not unexpected across a team of annotators, but it is useful in providing preliminary evidence that the amount of disagreement will vary significantly over content items.

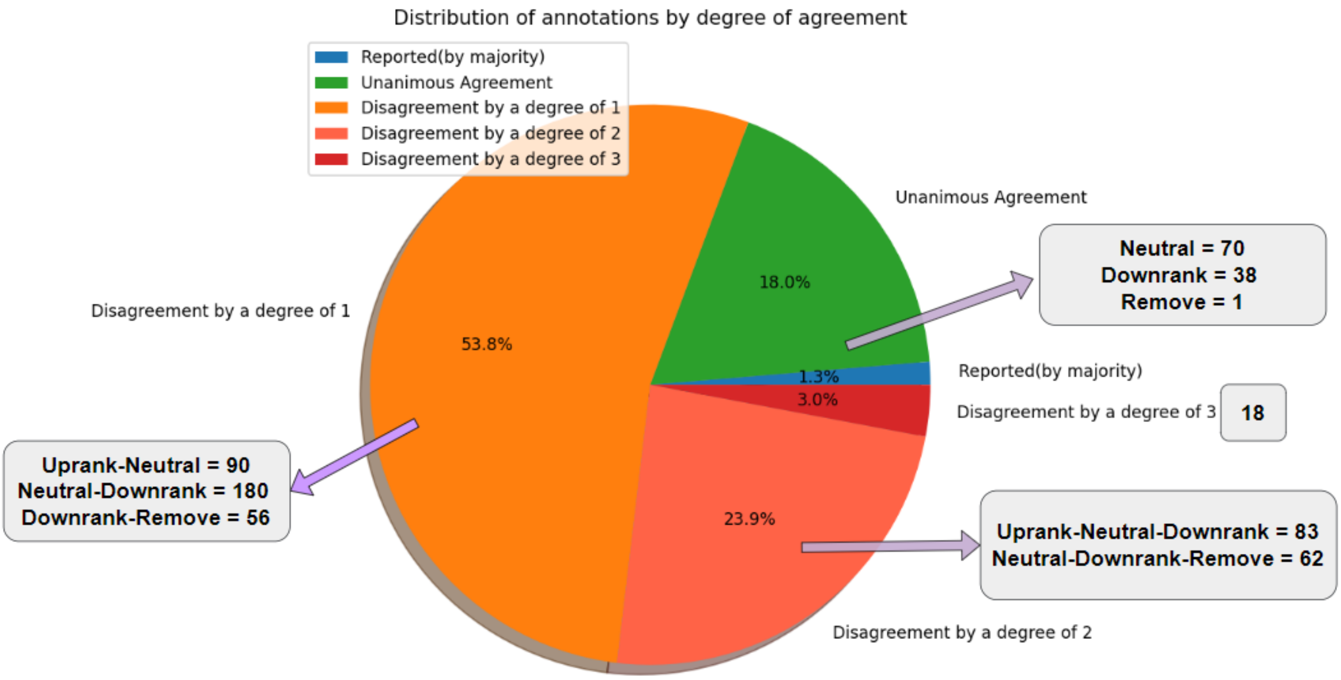


Figure 6.18: Distribution of annotations of 2019 Lok Sabha Election tweets by degree of agreement

Figure 6.19 shows the distribution over annotations for which annotators are in full agreement. These annotations make up 18% of the total dataset.

- Neutral = 70
- Downrank = 38
- Remove = 1

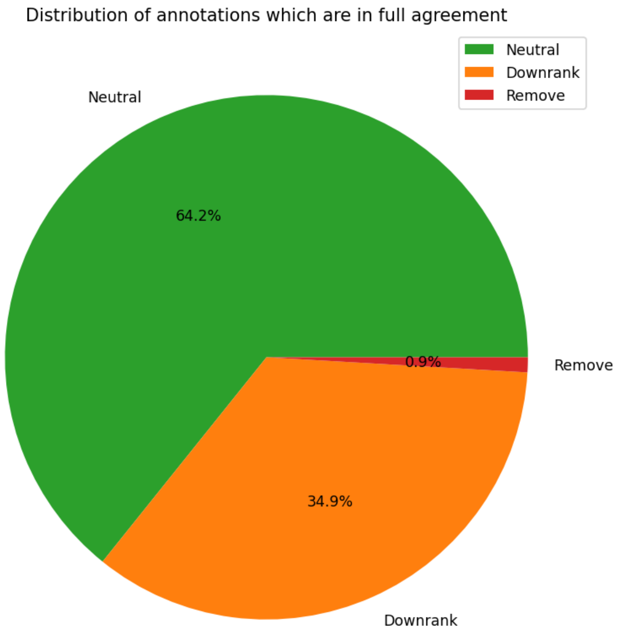


Figure 6.19: Distribution over annotations for which there was full agreement

To take a reality check on the collective decisions of our annotators, it is useful to consider some examples of tweets where there was unanimity amongst our (small) group of annotators. Here is the tweet which was unanimously annotated as *Remove*:

> " BJP Kerala State President says “Muslims can be identified by removing their clothes” #LokSabhaElections2019 @CMOKerala @INCKerala @ShashiTharoor @INCIndia @RahulGandhi @INCMinority @vijayanpinarayi https://t.co/CRIf9gPsh2 "

Here are some selected tweets that were unanimously annotated as *Downrank:*

> " @BJP4India @narendramodi You have created???? Did BJP disbursed those loans did BJP given those jobs??? Did BJP will recover those loans when borrower will not repay his loan?? दलाली छोङ के कुछ काम भी कर लो कब तक दूसरो की मेहनत का खाते रहोगे दलालो?? #corruption #ChowkidarChorHai "
>
> “ Crony Sevak's #SuitBootLootKiSarkar waived 3 50 000 Crores for just 15 rich cronies in last 5 Yrs. But when it comes to farmers Modi Sarkar says it's not their policy to waive farm loans: @RahulGandhi #NYAYForIndia #BJP_भगाओ_देश_बचाओ #ChowkidarChorHai https://t.co/CNSbpRM9vv “
>
> “#RafaleScam files stolen. Some are in a bathroom in Goa. This government is a joke to save #ChowkidarChorHai . #WhoAteTheRafalePie Mr #Modi? Who? AA? Who else? “

We now consider the distribution over annotations for which annotators were in disagreement by a degree of 1. These are shown in Figure 6.20. These annotations make up 53.8% of the total dataset.

- Uprank-Neutral = 90
- Neutral-Downrank = 180
- Downrank-Remove = 56

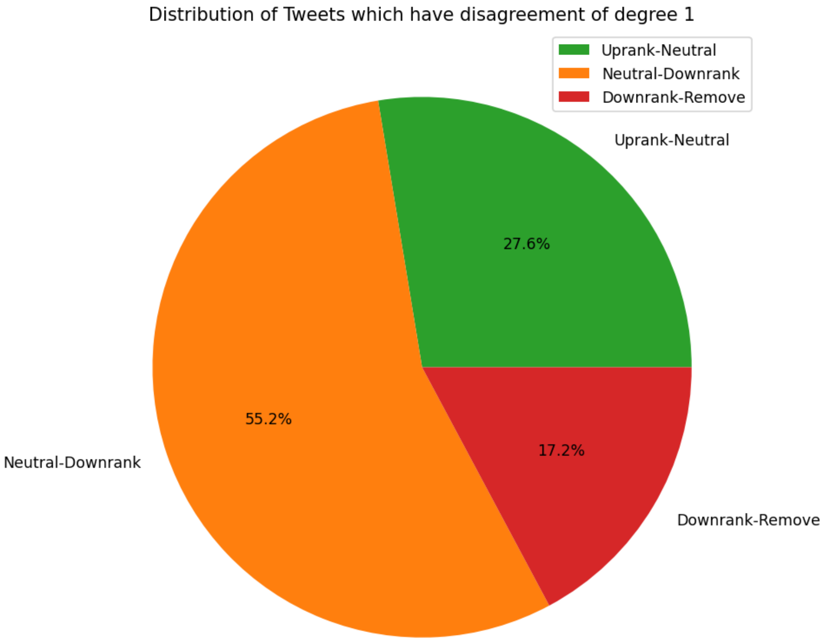


Figure 6.20: Distribution over annotations for which annotators were in disagreement by a degree of 1

As a further reality check, here are some examples of *Neutral-Downrank* tweets:

> " @MinhazMerchant @MahimaShastri @RahulGandhi @narendramodi Nothing comes free. Fact-checkers work only for clients who pay them. But what about \Independent\" media and public intellectuals of high integrity? That's precisely what @narendramodi asked @rahulkanwal on #ModiOnAajTak #LokSabhaElections2019 #NarendraModi #Modi"
>
> “Why every malfunctioning EVM is favouring @BJP4India not any other party? #Elections2019 #LokSabhaElections2019 #ioceu @INCTharoorian @INCIndia https://t.co/fTTgNEcwvA”

Here are some examples of *Downrank-Remove* tweets:

> " #ChowkidarNarendraModi lets #vijaymallya #NiravModi #choksi ESCAPE & lets #JaisheMohammed terrorists ENTER to carry out #PulwamaAttack #uriattack - IS AN INSULT to millions of poor hardworking #Chowkidar of #india . #ChowkidarChorHai #ChowkidarHiChorHai #Chowkidaar @RahulGandhi "
>
> “ #DalalModi #ChowkidarChorHai 135 Crore Indians trust on @narendramodi was sold for 30 000 Cr & Kickbacks. #Namo is a fraud who leaked #NationalSecurity related info 2 #AnilAmbani.Office of @PMOIndia's credibility is undoubtedly compromised. @RahulGandhi https://t.co/brHP1min4L “

We next consider the distribution over annotations for which annotators were in disagreement by a degree of 2. These annotations make up 24% of the total dataset. These are shown in Figure 6.21.-

- Uprank-(Neutral)-Downrank = 83 (13.6%)
- Neutral-(Downrank)-Remove = 62 (10.2%)

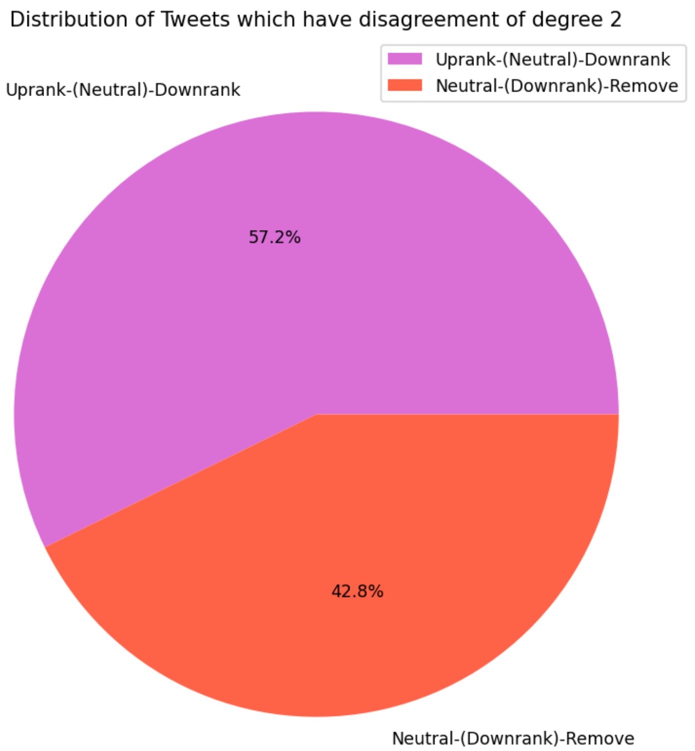


Figure 6.21: Distribution over annotations for which annotators were in disagreement by a degree of 2

Here are some examples of *Uprank-Neutral-Downrank* tweets:

> " #MainBhiChowkidar respected p.m. Our organization has been chowkidar since 2001 but @TelanganaDGP gives us a tag #ChowkidarChorHai #justice4ebiz We millions of people tweeted you many times but everytime we saw no response chowkidar chaukanna hain ki nhi @narendramodi "
>
> “ If you want to know what is the level of patience @ncbn is having you should first know how much time he waited for @narendramodi to full fill his promises. No need to work in alliance with people who makes false promises. #GoBackModi #APDharmaporatam #ModiBetrayedAP “

Here are some examples of *Neutral-Downrank-Remove* tweets:

> " Whatelse can it be other than \demonitisation money\"mystery box of @BJP4India & Mr @narendramodi oh sorry d so called poor #chowkidar #chaiwala with no money ??but such blackboxes of ?how pathetic s India became? #ChowkidarChorHai #modihataodeshbachao #LokSabhaElections2019 https://t.co/UNLZxNT8Wy" "
>
> “ #GoBackModi #csatvictims #ewsjumla #compensatoryAttempts Modi CHeated Hindi STUDENTS Cheated RURAL YOUTH CHEATED #CSATvictims Cheated General category poor. No age relaxation in #ews reservation like OBC reservation @PMOIndia @IYC @RahulGandhi @AmitShah @INCIndia “

We next show the distribution of annotations which have the highest degree of disagreement, that is, 3. (In these cases, while one or more annotator has felt it should be upranked, some other annotator(s) have felt that it is abusive or offensive and needs to be removed completely.) We present our analysis here in the form of a stacked bar plot (Figure 6.22), to give a more detailed understanding of division in labeling for each of the 18 tweets (3% of all) that fall in this category.

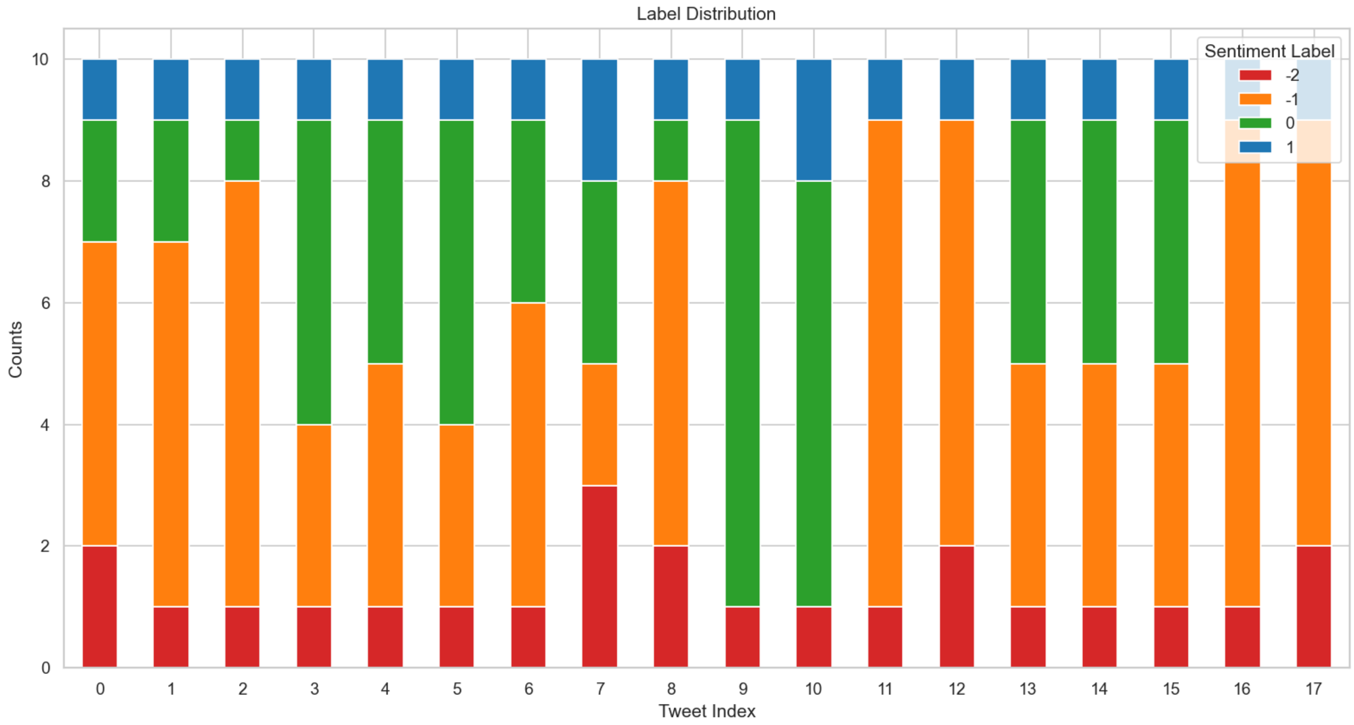


Figure 6.22: Label counts for each annotation having disagreement of degree 3

Again we give some examples from this category of maximally controversial tweets. These tweets have indeces 2, 6 and 7 respectively in Figure 6.22.

> " @narendramodi has lost his mental stability after seeing reports of @BJP4India going down the drain in the ongoing #LokSabhaElections2019 . He has forgotten all the sacrifices made by great heroes like Mahatma Gandhi Nehru Rajiv Gandhi etc and blabbering ill abt them. "
>
> “ @RahulGandhi French Ex-President Hollande?repeated his statement that Chowkidar INSISTED on deal being gifted to AnilBhai's brand new company in exchange for the Rafale Contract #ChowkidarChorHai so he never contradicted or denied what Hollande said #RafaleScam https://t.co/sPH7tPYBtm “
>
> “ @SheilaDikshit @bansalsurinder1 @ArvindKejriwal Ma'am I intensely hate @INCIndia ! Im staunch #MainBhiChowkidar ModiBhakt! Yet I like u for ur grace and intelligence! If @AamAadmiParty lose security deposit in all 7seats ur contribution towards it can't be denied. “

Finally, we report an entropy analysis of the annotations for this dataset. Figure 6.23 shows a histogram of entropy ranges for the dataset. The tweets with most disagreement shown in Figure 6.22 have entropies from the highest entropy ranges in Figure 6.23.

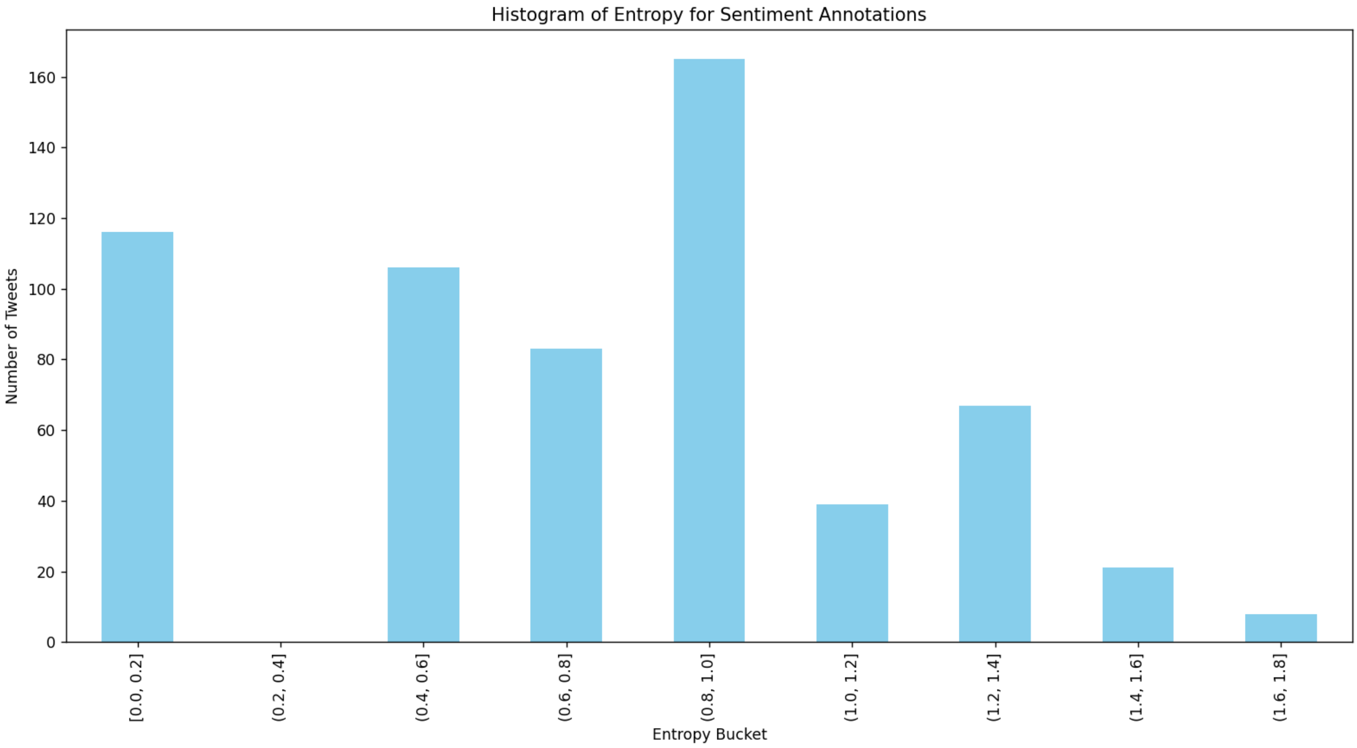


Figure 6.23: Histogram of the bucketed entropies of the annotations of 2019 Lok Sabha elections dataset

6.5.2. Brief annotation results on the tweets of 2022 State Elections

The second part of our training dataset contains **400 tweets**, details of which are given above. Similar to the first part, this was also annotated by a total of 10 annotators and the findings are visualized below along with the same statistics as the first part to help us study the disparity in annotations.

Again we begin by separating out tweets that were 'reported' by a majority of annotators. There are 8 of these. The remaining tweets are then classified into three classes, based on the degree of agreement between the annotators. The results are shown in Figure 6.24.

- Unanimous Agreement = 149
- Disagreement by a degree of 1 = 191
- Disagreement by a degree of 2 = 50
- Disagreement by a degree of 3 = 2

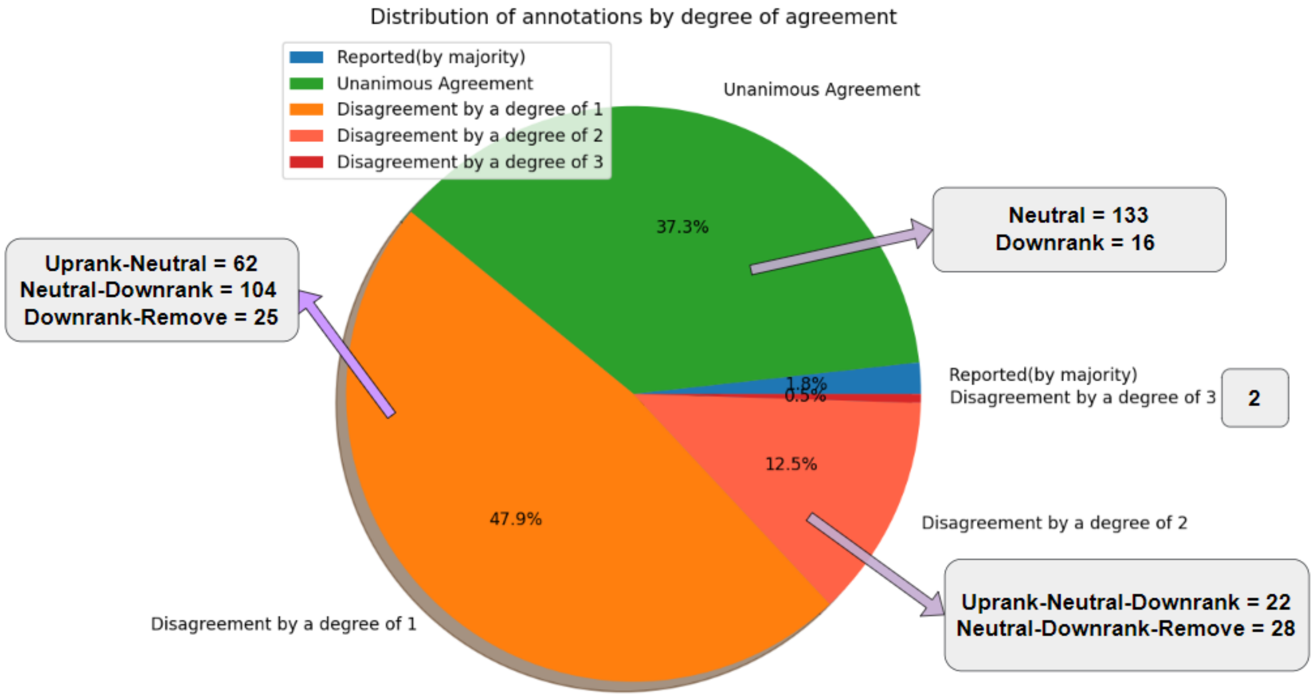


Figure 6.24: Distribution of annotations of 2019 Lok Sabha Election tweets by degree of agreement

We now consider the distribution of annotations for which annotators are in full agreement. These annotations make up 37.3% of the total dataset. These are shown in Figure 6.25.

- Neutral = 133
- Downrank = 16

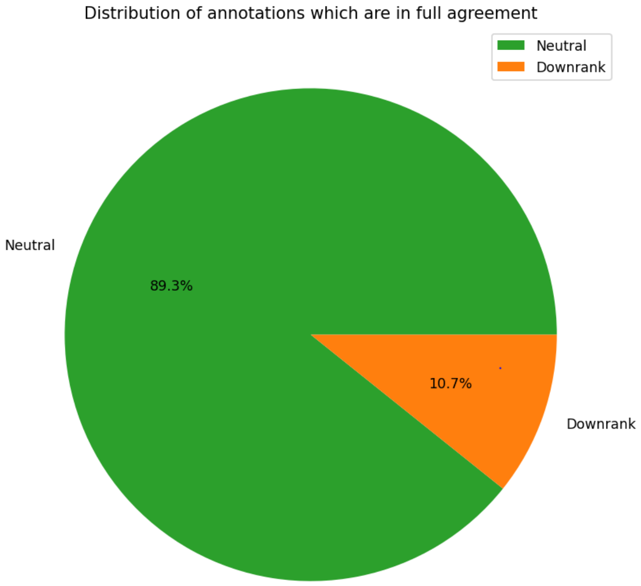


Figure 6.25: The distribution of annotations which are in full agreement

For a reality check, here are some examples of Unanimous *Downrank* tweets:

> " @ANI #SamajwadiParty @yadavakhilesh ji - Please leave the EVM free for one more day. Otherwise whom will you blame on result day? Yes exit polls are creating a

> perception of BJP is winning. May be actual results will covert this perception to a reality. #UttarPradeshElections2022 "
>
> “ Arvind Kejriwal has no work nowadays in Delhi. He should pay attention to people of Delhi. They are not going to get anything by making allegations & counter-allegations in #GoaElections2022. He is just replacing Congress here: CM @DrPramodPSawant #goencherajkaran #GoaElections https://t.co/MimEjAntpZ “
>
> “ #Congress leader #RahulGandhi alleged that the #BJP and #RSS have undermined” #Manipur's traditions and culture. #ManipurElections2022 #ManipurElections #Elections2022@RahulGandhi@RGWayanadOffice@INCManipur https://t.co/qe3sIVcf1n“

We next consider the distribution of annotations for which annotators were in disagreement by a degree of 1. These annotations make up 48% of the total dataset. These are shown in Figure 6.26.

- Uprank-Neutral = 62
- Neutral-Downrank = 104
- Downrank-Remove = 25

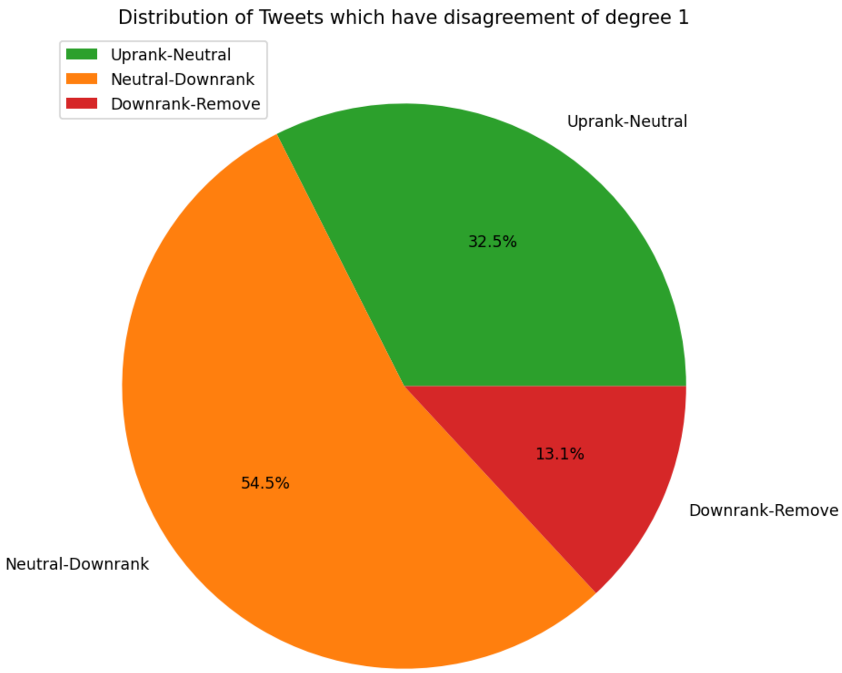


Figure 6.26: The distribution of annotations which are in disagreement by a degree of 1

Here are some examples of *Neutral-Downrank* tweets:

> " Ahead of the Uttarakhand assembly polls state forest minister Harak Singh Rawat on Sunday was dismissed from the state cabinet and expelled from the primary membership of the BJP for six years. @BJP4UK @drhsrawatuk #BJP #UttarakhandElections2022 https://t.co/CkUon16dnw "

> “#ManipurElections2022 | Congress alleges BJP of pre-poll violence seeks fair & peaceful elections @INCIndia @INCManipur @RahulGandhi @priyankagandhi @SoniaGandhi_FC @BJP4India @BJP4Manipur @Jairam_Ramesh #PreePollViolence #Elections #Northeastlive https://t.co/bexBhIqYAS ”

Here are some examples of *Downrank-Remove* tweets:

> " Mamata has made Bengal as mini Pakistan. There are no schemes in Bengal but @AITC4Goa promise them in Goa. Goa can be developed only by @BJP4Goa says Union minister @shripadynaik #goencherajkaran #GoaElections2022 https://t.co/pOba3aEslI "
>
> “ उत्तराखंड के पूर्व CM ने @harishrawatcmuk को बताया ब्राह्मण विरोधी कहा- लालकुआँ में होगी राजनीतिक मौत #UttarakhandElections2022 #BJP #congress #uttarakhand https://t.co/OxyataFqzO… “
>
> “ Sorry @sherryontopp ji Aapko Darshani Ghoda bana rkha h Aur Reta chorr @CHARANJITCHANNI gamdhe ko aapke upr rkha @INCIndia ne Very bad ???? #AAP #PunjabElections2022 #LataDidi https://t.co/7maQ08ODMg”

We next consider the distribution of annotations for which annotators are in disagreement by a degree of 2. These annotations make up 12.5% of the total dataset. These are shown in Figure 6.27.

- ○ Uprank-(Neutral)-Downrank = 22
- ○ Neutral-(Downrank)-Remove = 28

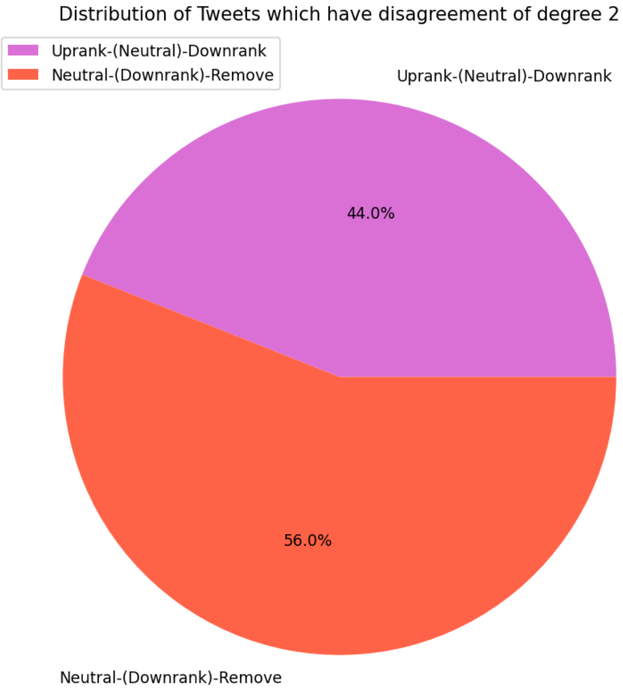


Figure 6.27: The distribution of annotations which are in disagreement by a degree of 2

From this set, here are examples of *Uprank-Neutral-Downrank* tweets:

> " Congress candidates are packed together in a resort at Bambolim ahead of the counting. The party has abandoned the plan to shift the candidates outside the state. @INCGoa is doing so to avoid defection : Sources #goencherajkaran #GoaElections2022 #AssemblyElections2022"
>
> “ Shiv Sena maintained friendly relations with BJP & was not as active in Goa to avoid loss to @BJP4Goa . But from now on @Shivsena4Goans will be fully active in Goa & Shivsainik will fight bravely in every election says MH minister @AUThackeray. #goencherajkaran #GoaElections2022 https://t.co/RqEUSAYXnj “

Here are some examples of *Neutral-Downrank-Remove* tweets:

> " Digambar Kamat govt was the most corrupt Govt He has no right to call our Govt corrupt ask him about 'Acche Din' Tell him to see all days before 2012 than you'll come to know what are 'Acche Din' : @DrPramodPSawant @digambarkamat #goencherajkaran #GoaElections2022 https://t.co/RBpIPIMz89 "
>
> “I am leaving delhi soon because of him and he is calling people from Canada ??????. #PunjabElections2022 @ArvindKejriwal @AtishiAAP #BJP #Congress @AamAadmiParty @AAPDelhi @AAPPunjab https://t.co/M2UfurDBDR”

Finally, we show the distribution of annotations for which annotators have the highest degree of disagreement of 3. Again we use a stacked bar plot here: see Figure 6.28. There are only 2 tweets (0.5%) that fall in this category: we reproduce the text of both tweets below.

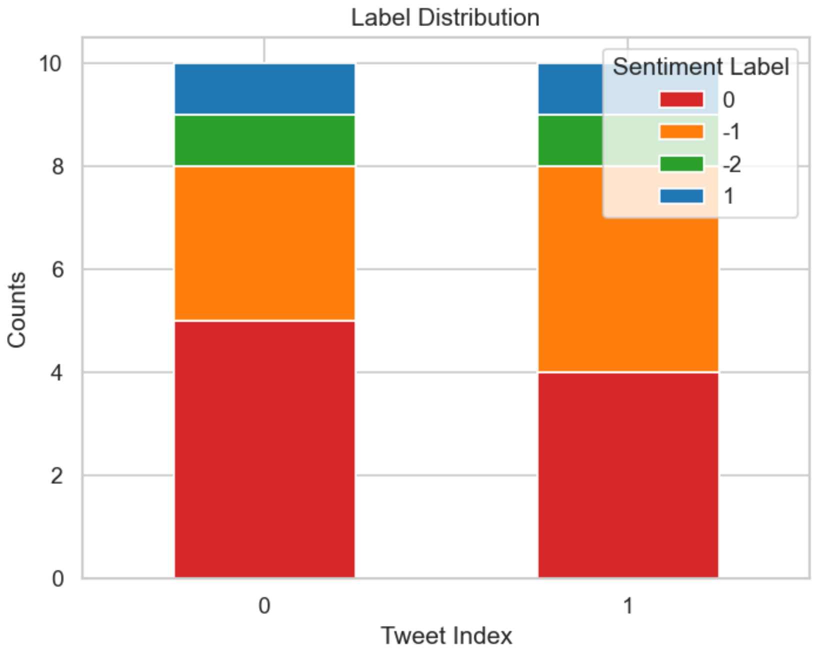


Figure 6.28: Label counts for each annotation having disagreement of degree 3

The two *Uprank-Neutral-Downrank-Remove* tweets (indices 0 and 1 respectively in the above plot) are as follows:

> " Several incidents of pre-poll violence have already rocked #Manipur in the past couple of weeks. #ManipurElections2022 @NBirenSingh @manipur_police @ECISVEEP @CeoManipur https://t.co/lc0u0fIgfp "
>
> “ BJP must analyse ticket of existing Chandausi MLA - sheer waste and good for nothing ! Astonishingly also holds ministerial portfolio.What has been done for the city residents? #UttarPradeshElections2022 #BJP #YogiAdityanath @myogiadityanath @swatantrabjp @BJP Uttar Pradesh“

Again we also computed the entropy over annotator judgements for these tweets. Figure 6.29 shows a histogram of entropies, again grouped into buckets of constant size. Again, the tweets with highest disagreement are also those with highest entropy and tweets having least disagreement are those with the least entropy.

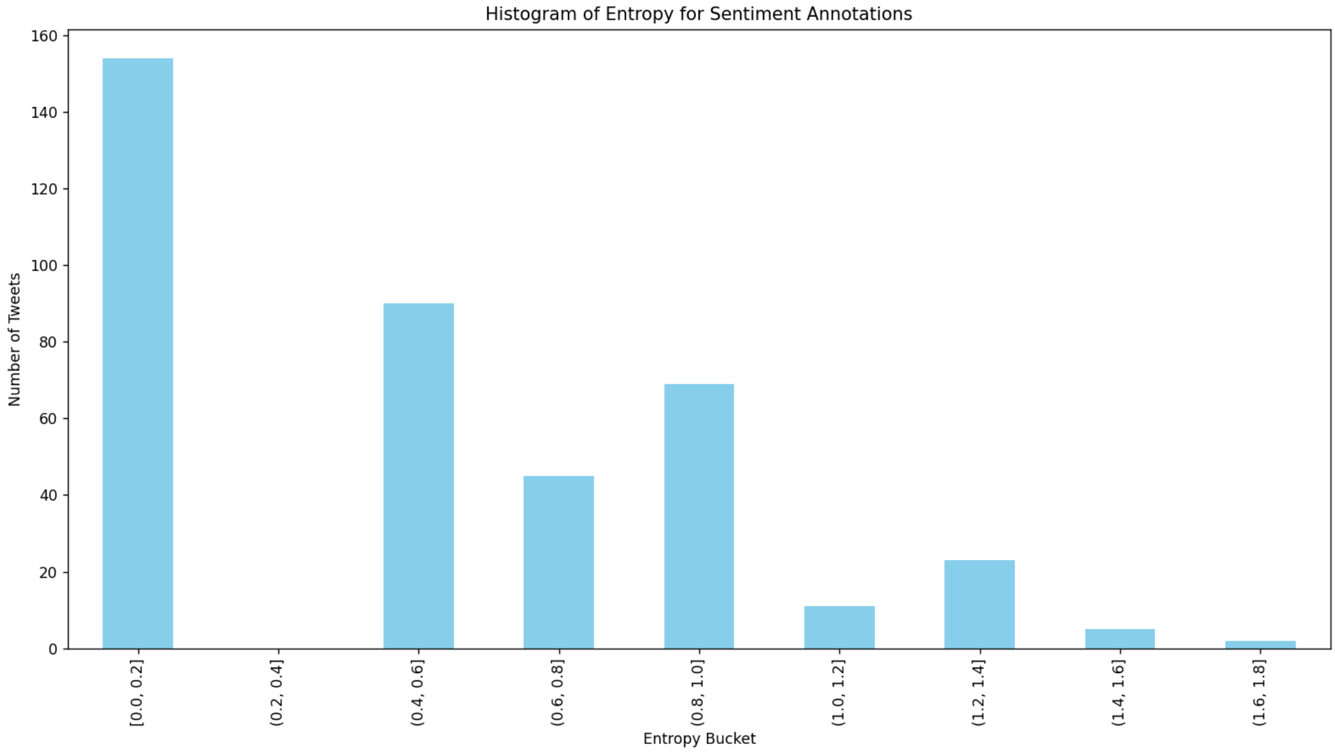


Figure 6.29: Histogram of the bucketed entropies of the annotations of 2022 State elections dataset

6.6 Discussion

The main purpose of this preliminary pilot is to get a first indication of whether judgements by ‘the crowd’ of annotators are usable as the basis for moderation decisions. Readers with familiarity with the Indian political context from which our annotations are taken can take their own view on this, based on the example tweets we have presented. Within our group of authors, those of us who have local familiarity feel that the annotators are providing useful information, both in the majority decisions they take, and in their disagreements. For instance, the tweets that are unanimously classed as ‘remove’ feel to us worse than those that are unanimously classed as ‘downrank’. And if we consider tweets which have the same majority label, but different amounts of disagreement, there is some case to be made that the ones with more disagreement should be treated more leniently in moderation.

Obviously these are very preliminary results, but we feel they are interesting enough that it is worth extending to a larger annotation study, with more annotators and more content items. In a larger study, many other interesting analysis methods become possible. In particular, clustering analyses will identify *groups* of annotators of several different kinds. Recall that our sampling methodology calls for certain groups to be given a particular voice. Identifying these groups in advance may not be easy—but clustering may provide valuable information about their identity.

7. Summary and future work

This report has two parts. In the first part (Sections 1–5), we outline a general proposal for a new way of training the harmful content classifiers that operate in social media platforms, using training sets constructed in a semi-public domain, rather than behind the closed doors of companies. We present arguments for this general approach in Sections 1 and 2. In Section 3, we define a two-pass annotation scheme for harmful content that is operationalised to support content moderation actions (Section 3.1), and processes for training classifiers and regression models using this scheme (Sections 3.2 and 3.3). In Section 4 we suggest some very broad principles for how to select a group of annotators. Taken together, these proposals define a *methodological pipeline* for a content moderation process operating in a semi-public domain. In Section 5, we discuss how this pipeline could be evaluated, if it were implemented in a social media platform.

Our proposals are very preliminary: we are at an early stage in exploring these ideas. One important way of assessing them is to conduct *pilot projects* that try out the methods being proposed. The second part of our report describes an initial pilot project, in which a small group of annotators annotate a small set of content items, using the discrete scheme we propose as the first pass of content annotation. The pilot is far too small to draw solid conclusions about actual content categories—but as a very preliminary reality check, annotators' collective judgments about tweets seem to be a useful resource. It is particularly interesting to see that disagreement between annotators varies quite considerably over items. Our proposed method for content moderation envisages a specific role for annotator disagreement; the pilot study suggests that there will be sufficient variability in disagreement over items for our proposed method to provide value.

We have two objectives for followup work. Most immediately, we will pilot the second pass of our annotation scheme, using the annotators and content items described in the current report. The second pass asks annotators to rank pairs of items: the analysis of these judgments will allow items to be placed on a continuous scale of 'hatefulness'. With the results of this second pass, we will be able to make informative connections between the discrete and continuous annotation schemes. In particular, we will be able to identify scores at the boundary between 'neutral' and 'downrank' items, and the boundary between 'downrank' and 'remove' items, which respectively indicate the scores for minimum and maximum downranking operations. We will also be able to explore the use of entropy in adjusting moderation actions, both in decisions to remove, and in downranking operations.

In the longer term, our intention is to *scale up* the annotation exercise, drawing on lessons learned in the piloting process. Scaling up will provide us with meaningful and valuable information about perceptions of harmful content in our chosen domain of Indian political discussions. This information will be of great interest in its own right, as a reflection on the nature of the discussions taking place, and of the public observing and contributing to these discussions. But it will also contribute to our proposed methodology, in allowing us to train effective content classifiers and regression models on the annotated datasets we obtain. (Our scaling up exercise will be targeted at this practical objective.) When we have trained classifiers and regression models, we can pilot the remaining parts of our proposed pipeline, exploring possible ways of adjusting moderation actions using the entropy of classifier outputs, and proposing a specific content moderation process that could be deployed by social media companies in this chosen domain.

References

Bennett, T. D. (2023). Interpretation is opinion: realigning the fact/opinion distinction in English defamation law. Journal of Media Law, 1-28.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media (pp. 36-41).

Bradley-Terry (2023). Bradley-Terry model. Encyclopedia of Mathematics. http://encyclopediaofmath.org/index.php?title=Bradley-Terry_model&oldid=22181

Buck, N. (2022). The use of juries in defamation proceedings in America and Australia. Kennedys Law. https://kennedyslaw.com/en/thought-leadership/article/the-use-of-juries-in-defamation-proceedings-in-america-and-australia/

Chakravarthi, B. R., Anand Kumar M, McCrae, J. P., Premjith, B., Soman, K. P., & Mandl, T. (2020, December). Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In FIRE (Working notes) (pp. 112-120).

Chakravarthi, B. R., & Muralidaran, V. (2021, April). Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In Proceedings of the first workshop on language technology for equality, diversity and inclusion (pp. 61-72).

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., ... & Xie, X. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Dowlagar, S., & Mamidi, R. (2021, December). A survey of recent neural network models on code-mixed Indian hate speech data. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 67-74).

GPAI (2021). Responsible AI for Social Media Governance: A proposed collaborative method for studying the effects of social media recommender systems on users. Report, November 2021, Global Partnership on AI.

GPAI 2022. Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions. Report, November 2022, Global Partnership on AI.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. Perspectives on Psychological Science, 6(1), 48-60.

Gordon-Rodriguez, E., Loaiza-Ganem, G., Pleiss, G., & Cunningham, J. P. (2020). Uses and abuses of the cross-entropy loss: Case studies in modern deep learning.

Imagenet (2023). Image Classification on ImageNet https://paperswithcode.com/sota/image-classification-on-imagenet

Kiritchenko, S., & Nejadgholi, I. (2020). Towards ethics by design in online abusive content detection. arXiv preprint arXiv:2010.14952.

Kumar, R., Reganti, A. N., Bhatia, A., & Maheshwari, T. (2018). Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in

Indo-European languages. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (pp. 14-17).

Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2020, December). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 29-32).

Nagar, S., Gupta, S., Bahushruth, C. S., Barbhuiya, F. A., & Dey, K. (2022). Hate speech detection on social media using graph convolutional networks. In Complex Networks & Their Applications X: Volume 2, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10 (pp. 3-14). Springer International Publishing.

Saroj, A., & Pal, S. (2020, May). An Indian language social media collection for hate and offensive speech. In Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language (pp. 2-8).

Sathyanarayana, S. (2014). A gentle introduction to backpropagation. Numeric Insight, 7, 1-15.

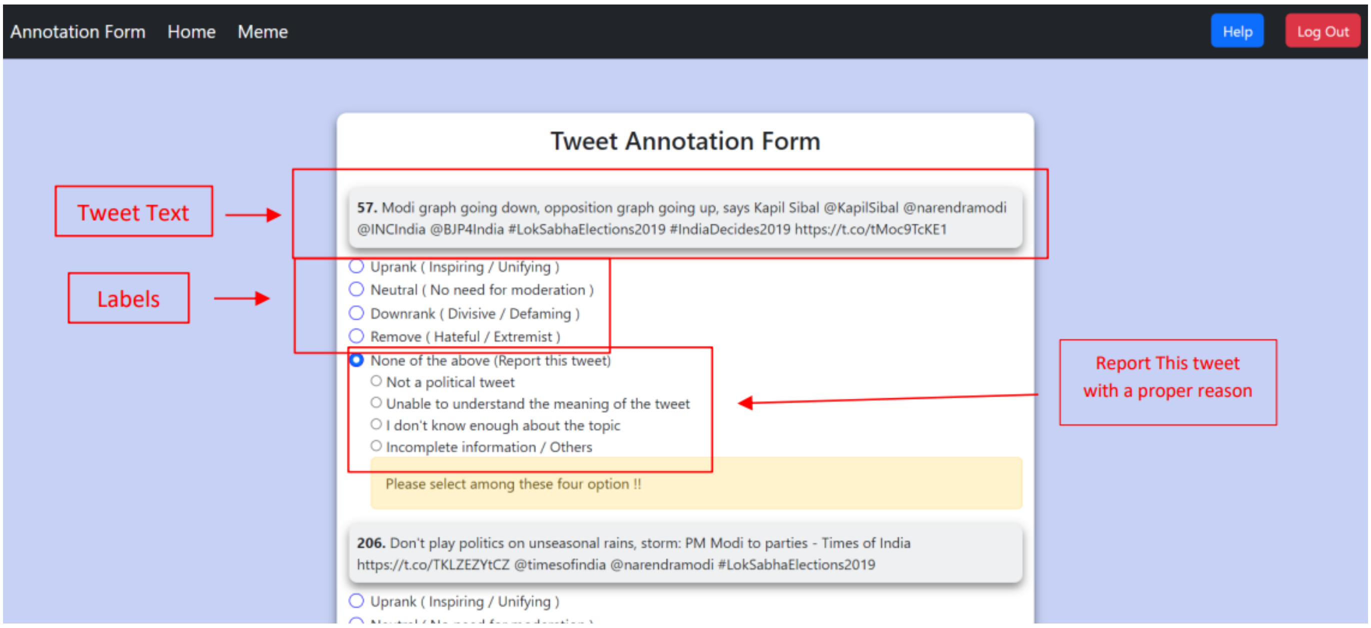
Small, C. T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., ... & Megill, C. (2023). Opportunities and Risks of LLMs for Scalable Deliberation with Polis. arXiv preprint arXiv:2306.11932.

Statista (2023). Number of data removal requests issued to X (formerly Twitter) from June to December 2021, by country and institution. Statista report. Retrieved October 2023 from https://www.statista.com/statistics/234858/number-of-requests-for-data-removal-from-twitter/

Tornetta, G. N. (2021). Entropy methods for the confidence assessment of probabilistic classification models. arXiv preprint arXiv:2103.15157.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.

Appendix: Annotation Guide



Tweet Annotation Form

57. Modi graph going down, opposition graph going up, says Kapil Sibal @KapilSibal @narendramodi @INCIndia @BJP4India #LokSabhaElections2019 #IndiaDecides2019 https://t.co/tMoc9TcKE1

○ Uprank (Inspiring / Unifying)
○ Neutral (No need for moderation)
○ Downrank (Divisive / Defaming)
○ Remove (Hateful / Extremist)
● None of the above (Report this tweet)
  ○ Not a political tweet
  ○ Unable to understand the meaning of the tweet
  ○ I don't know enough about the topic
  ○ Incomplete information / Others

Please select among these four option !!

206. Don't play politics on unseasonal rains, storm: PM Modi to parties - Times of India https://t.co/TKLZEZYtCZ @timesofindia @narendramodi #LokSabhaElections2019

○ Uprank (Inspiring / Unifying)

Labels
+1: *Uprank*: constructive/inspirational/unifying
0: *Neutral*: no moderation needed
-1: *Downrank*: divisive/defaming/denigrating
-2: *Remove*: hateful/extremist

Uprank: These tweets should have an uplifting/motivational message. Such messages are very generic and likely to invoke a positive emotion from the reader.

- "Embracing unity and diversity, we stand stronger together as a nation. Let's celebrate our differences and work towards a brighter future for all! #UnitedWeRise #InclusivityMatters 🤝 #ChangeMakers #UnityInAction"
- "In the face of challenges, we rise. United, we are invincible! 💪 #TogetherWeStand #StrongerTogether"
- "Education is the key to unlocking endless possibilities. Let's invest in our schools and empower our students to reach for the stars. Every child deserves a quality education. 🎓📖 #EducationMatters #FutureLeaders

Remove: These are tweets which are derogatory, discriminatory, abusive or offensive in nature and need to be removed so that it doesn't influence other users in an extremely negative sense.

- "XYZ community is responsible for all the crime in our city! They are a burden on society. #Discrimination #HateSpeech"
- "I can't stand people from ABC country. They're all lazy and unintelligent. #Prejudice #Xenophobia"
- "I hope all members of this political party suffer and fail miserably. They are destroying our country! #PoliticalHate #Hostility"

- "Those who believe in XYZ religion are brainwashed fools. They should be eradicated from society. #ReligiousHate #Intolerance"
- "I wish harm upon people who support the ABC cause. They deserve nothing but misery. #Vitriol #Malice"

Neutral: These are the tweets which pass a generic message to the reader. It is mostly informative or declarative where the tweets mainly represent events or a statement or feeling of the user without targeting or discriminating against anyone.

- "Discussing potential reforms to ensure our electoral system remains fair, transparent, and accessible to all citizens. Let's work together to uphold the integrity of our democratic processes. 🗳️🗂️ #ElectoralReform #Democracy"
- "🌍 Climate change is a global challenge that demands collaborative efforts. Let's explore innovative strategies to reduce our carbon footprint and protect the environment for future generations. 🌎🌿 #ClimateAction #Sustainability"
- "Congratulations to the winning team in the championship! They played exceptionally well. #Sports #Champions"
- "💼 Balancing fiscal responsibility with social programs is a continuous deliberation. Let's find middle ground solutions that address the needs of our citizens while maintaining a stable economy. 💰🤝 #FiscalPolicy #SocialPrograms"

Downrank: These are tweets that convey a defaming or upsetting message to the reader and provokes a sense of division at some level should not be suggested but rather be down-ranked by a recommender system.

Example: "The environment and economy tug us in opposite directions. Balancing these competing interests resembles walking on a tightrope. Can we truly harmonize growth and conservation? 🌱🏗️ #EnvironmentVsEconomy #DelicateBalance"
Example: "There's no hope for the future. Everything is going downhill. #Negativity #Pessimism"
Example: "The politician @PoliticianName is a liar and a cheat. We can't trust anything they say. #CharacterAssassination #Negativity"

None of the Above: If suppose, it is confusing for an annotator to choose among the four given labels, he/she can choose this option and then select one of the suboptions which are self-explanatory-

- Not a political tweet
- Unable to understand the meaning of the tweet
- I don't know enough about the topic
- Incomplete information / Others

Example

1. "Lost in the pages of an enthralling novel, transported to another world where imagination knows no bounds. 📚✨ #BookLover #EscapeThroughReading"
2. "Rhythmic raindrops compose a soothing melody, creating an atmosphere of tranquility that nourishes the soul. 🌧️🎶 #RainyDayBliss #NatureSymphony"

3. "Unlocking new levels, conquering challenges, and embracing the thrill of virtual adventures – a world of gaming wonders awaits! 🎮🕹️ #GamerLife #VirtualExplorer"

Political Tweet (Non-Neutral, Non-Hateful, Non-Uprank/Downrank):

1. "Bold proposals spark fiery debates as visions for the future collide. Amidst the fervor, perspectives emerge that challenge conventions and reshape the discourse. Let's explore uncharted territories of thought. 🌄🗣️ #RedefiningDebate #UnveilingVisions"