Beyond Majority Voting: Agreement-Based Clustering to Model Annotator Perspectives in Subjective NLP Tasks

Tadesse Destaw Belay[1], Ibrahim Said Ahmad[2], Idris Abdulmumin[3], Abinew Ali Ayele[4], Alexander Gelbukh[1], Eusebio Ricárdez-Vázquez[1], Olga Kolesnikova[1], Shamsuddeen Hassan Muhammad[2,5], Seid Muhie Yimam[6]

[1]Instituto Politécnico Nacional, [2]University of Wisconsin–Stevens Point,
[3]University of Pretoria, [4]Bahir Dar University,
[5]Imperial College London, [6]University of Hamburg

tadesseit@gmail.com

Abstract

Disagreement in annotation is a common phenomenon in the development of NLP datasets and serves as a valuable source of insight. While majority voting remains the dominant strategy for aggregating labels, recent work has explored modeling individual annotators to preserve their perspectives. However, modeling each annotator is resource-intensive and remains underexplored across various NLP tasks. We propose an agreement-based clustering technique to model the disagreement between the annotators. We conduct comprehensive experiments in 40 datasets in 18 typologically diverse languages, covering three subjective NLP tasks: sentiment analysis, emotion classification, and hate speech detection. We evaluate four aggregation approaches: majority vote, ensemble, multi-label, and multitask. The results demonstrate that agreement-based clustering can leverage the full spectrum of annotator perspectives and significantly enhance classification performance in subjective NLP tasks compared to majority voting and individual annotator modeling. Regarding the aggregation approach, the multi-label and multitask approaches are better for modeling clustered annotators than an ensemble and model majority vote. The dataset is publicly available in GitHub: https://github.com/Tadesse-Destaw/Beyond-Majority-Voting.

1 Introduction

In supervised machine learning, using multiple annotators to label datasets is a common strategy to improve the quality of training and evaluation data for downstream natural language processing (NLP) tasks (Cabitza et al., 2023). However, annotators' disagreements frequently arise during the annotation process. These disagreements stem not only from random errors but also from systematic differences in task interpretation and understanding. Specifically, sociodemographic factors, such as age, gender, race, educational status, political stance, and living experience, can significantly influence how individuals interpret subjective annotation tasks (Luo et al., 2020; Odbal et al., 2022; Beck et al., 2024). Consequently, annotators often approach the same text from different perspectives, leading to divergent judgments shaped by their personal and cultural connotations (Wan et al., 2023; Lee et al., 2024).

Current approaches for handling annotator disagreement fall into four categories: (1) Aggregating annotations using majority vote, an extra expert as a judge, or Bayesian methods to get a single ground truth label (Paun et al., 2018); (2) Treating each annotator's label as potentially valid while filtering uncertain items due to disagreement (Wang and Plank, 2023); (3) Training models directly on the raw annotations without label aggregation (Daval-Frerot and Weis, 2020); and (4) Leveraging both hard and soft labels from annotations during model training (Rodríguez-Barroso et al., 2024).

Majority voting is the most common practice to decide the final annotation label (Uma et al., 2022; Xu et al., 2024). However, this approach removes genuine disagreement between annotators, thereby marginalizing minority viewpoints that could offer valuable insights (Leonardelli et al., 2023; Rizzi et al., 2025). Although adequate for objective tasks like part-of-speech tagging (where even a single annotator label may suffice), majority voting poses limitations in subjective annotation tasks, where establishing a single ground truth is inherently challenging (Khurana et al., 2024).

In recent years, the practice of only considering the majority vote has been criticized (Sandri et al., 2023). Research has started to advocate for bet-

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1348–1364, 2026. https://doi.org/10.1162/TACL.a.731
Action Editor: Joel Tetreault. Submission batch: 6/2025; Revision batch: 12/2025; Published 7/2026.

ter ways to deal with disagreements between annotators and to preserve all interpretations of annotators rather than eliminate the minority point of view (Uma et al., 2022; Davani et al., 2022; Rodríguez-Barroso et al., 2024). Considering each annotator's perspective is an emerging research direction in computational subjective tasks (Yin et al., 2023), for instance, modeling annotation disagreement, as seen in shared tasks and workshops focusing on disagreement in subjective tasks (Leonardelli et al., 2023; Roth and Schlechtweg, 2025).

A common approach to capturing annotation disagreement while preserving individual perspectives is to model each annotator separately (Davani et al., 2022; Rodríguez-Barroso et al., 2024). However, training a model for each annotator is computationally resource-intensive, especially when the number of annotators is large and the annotation is skewed among annotators. In this work, we propose an agreement-based clustering technique that groups annotators by their agreement patterns while maintaining individual perspectives. Our contributions are:

- **An agreement-based clustering framework** that automatically groups annotators by their agreement patterns, preserving diverse perspectives while significantly reducing computational overhead compared to per-annotator modeling;
- **Large-scale empirical validation** across 40 multilingual datasets covering three challenging subjective tasks (sentiment analysis, emotion classification, and hate speech detection), showing consistent improvements over individual annotator modeling approaches;
- **A comprehensive comparison** of aggregation strategies (ensemble, multi-label, and multitask), establishing baselines against majority voting.

2 Literature Review

In this section, we review the literature relevant to our work, focusing on subjective tasks, the sources of disagreement, and approaches to modeling annotators disagreement in subjective NLP tasks.

2.1 Subjective NLP Tasks

Subjective NLP tasks incorporate diverse answers during annotation due to the differences between the sociodemographic backgrounds of annotators (Röttger et al., 2022). Finding a single true label in such tasks using majority vote can lead to biased results (Uma et al., 2022), and it is crucial to incorporate each perspective of the annotator. Examples of subjective tasks include sentiment analysis (Muhammad et al., 2022, 2023; Parmar and Modh, 2026; Afriyie and Weyori, 2026), hate speech (Kapil and Ekbal, 2024; Beck et al., 2024; Muhammad et al., 2025a), abusive speech (G et al., 2025), humor and sarcasm identification (Simpson et al., 2019), toxicity (van Aken et al., 2018), ironic content (Frenda et al., 2023), good or bad (Martínez Cámara et al., 2018) and emotion classification (Muhammad et al., 2023, 2025b). Obtaining high-quality and reliable annotations in such subjective tasks is challenging. It is common to see low agreement among annotators in such tasks, such as GoEmotions emotion data (Demszky et al., 2020) has 27% Cohen's kappa (considered low), and HateXplain (Mathew et al., 2022) hate speech data has 46% (considered moderate agreement) (Landis and Koch, 1977).

2.2 Sources of Disagreement

Disagreement in the annotations of NLP datasets refers to the absence of a single ground truth label, often arising from genuine differences in annotator interpretation (Röttger et al., 2022; Braun, 2024). Traditionally viewed as noise, such disagreement is now increasingly recognized as a valuable source of information (Fell et al., 2021). While all forms of disagreement contribute to label uncertainty, their underlying reasons may differ. Broadly, three perspectives explain the origins of annotation disagreement: (1) issues related to annotation design, such as unclear guidelines, poorly defined label spaces, lack of contextual information, or low-quality annotator performance (Denton et al., 2021; Parmar et al., 2023); (2) inherent ambiguity in the text itself (Uma et al., 2022); and (3) variation in annotators' perspectives which is influenced by sociodemographic factors (Jiang and de Marneffe, 2022; Niu et al., 2024).

2.3 Modeling Annotation Disagreement

Leveraging annotation disagreement during model training has been demonstrated to serve as a valuable learning signal. Hayat et al. (2022) implemented multitask-based modeling of each annotator separately, which yields better perfor-

mance than the traditional majority-vote approach. Rodríguez-Barroso et al. (2024) proposed federated learning that builds a global model from the client annotator embedding models. Mokhberian et al. (2024) proposed an annotator-aware representation for texts (AART) that combines text with annotator embeddings. Several works (e.g., Davani et al. (2022); Rodríguez-Barroso et al. (2024); Xu et al. (2024)) have proposed ensemble-based approaches that model each annotator individually.

However, training a separate model for each annotator can be computationally expensive and often impractical in real-world settings. For instance, the GoEmotions (Demszky et al., 2020) emotion dataset annotated by over 82 annotators, a sentiment analysis dataset by Diaz et al. (2018) involved more than 1400 annotators, and the hate speech dataset by Cjadams et al. (2019) was constructed with contributions from over 8000 annotators. Moreover, the number of examples annotated per individual varies considerably. In the GoEmotions dataset, for instance, annotator contributions range from as few as 3 to as many as 4800 instances.

Previous studies have examined annotation disagreement at various levels. Venanzi et al. (2014) proposed a probabilistic Bayesian model that jointly learns latent community profiles of crowd workers and estimates both worker reliability and true labels by clustering workers into communities. Similarly, Lakkaraju et al. (2015) proposed a hierarchical Bayesian framework to evaluate the quality of individual annotators and to identify the true labels of items. Their work also diagnoses the types of errors that annotators tend to make and the common characteristics of items on which such errors occur, including clustering annotators who share similar attributes. Weerasooriya et al. (2022) introduced a graphical model that enhances the quality of annotator labels by using clustering to strengthen the signal of noisy data. Finally, Gordon et al. (2022) addressed annotators disagreement through jury learning, which involves selecting a juror from each representative group. This method models each annotator in a dataset and requires sociodemographic metadata such as self-identified gender, race, education, political affiliation, age, parental status, and religiosity to produce a joint jury prediction for classifying unseen examples.

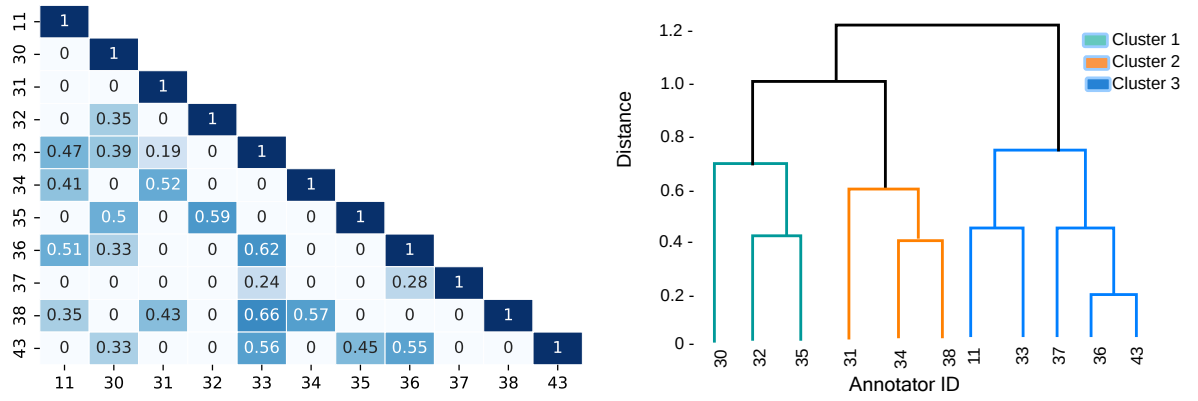
However, these studies primarily focus on capturing the quality and error patterns of individual evaluators during the annotation process (Weerasooriya et al., 2022; Gordon et al., 2022), after annotation when detail annotator attributes are available (Lakkaraju et al., 2015), or by requiring an additional expert (gold) judgments from highly trained editorial annotators to serve as ground truth for evaluating regular annotators (Venanzi et al., 2014). In contrast, our work focuses on the post-annotation stage and aims to model the annotators' perspectives (opinions) to estimate the true labels more accurately before influenced by majority voting. Our approach is designed for already annotated datasets where only the anonymous annotator IDs are available.

Most similar to our work (Fell et al. (2021); Kairam and Heer (2016); Lo and Basile (2023)) proposed clustering or grouping the annotations based on the agreement of annotators. However, these clustering approaches relied on the annotators' metadata, such as culture, demographics, and ethnic backgrounds. This clustering approach works well when the annotators' sociodemographic metadata is available.

Based on the limitations of the approaches implemented in the previous works (Daval-Frerot and Weis, 2020; Fell et al., 2021; Davani et al., 2022; Rodríguez-Barroso et al., 2024), we propose clustering annotators based on their annotation agreements to overcome over-modeling of annotators; more details are in Section §3. In addition, the previous exploration of subjective tasks solely relies on the English language; we covered various low-resource languages, annotation taxonomies, data sources, and subjective tasks. Finally, we evaluate modeling individual annotators versus agreement-based clustered annotators using ensemble, multi-label, and multitask aggregation approaches, with the majority vote as a baseline.

3 Agreement-based Annotator Clustering

A common approach for modeling disagreement is to model each annotator separately (Davani et al., 2022) or a sample of annotators (Yin et al., 2023). While training a model for each annotator has the advantage of preserving individual perspectives, it has two drawbacks: (1) when many annotators participate, annotation distributions become



(a) Annotators (their IDs) and its agreement

(b) Clustering the 11 annotators into $C = 3$

Figure 1: Pairwise Agreement between annotators. This agreement score is used to group annotators and determine which annotator belongs to which cluster. The example is from the Nigerian Pidgin (*pcm*) language sentiment dataset — eleven annotators participated, with each instance annotated by a minimum of 3, and the annotators are clustered as cluster $C1 = 30, 32, 35$, $C2 = 31, 34, 38$, and $C3 = 11, 33, 36, 37, 43$.

highly skewed; and (2) training one model per annotator is computationally intensive and expensive. Instead of modeling each annotator separately, clustering annotators based on their agreements could address this issue.

The clustering approach is motivated by the premise that if two annotators independently assign the same label, such as *hate* in a hate speech annotation task, they demonstrate agreement on that specific instance. This agreement suggests a shared perspective between the annotators, justifying their inclusion within the same cluster. We proposed an agreement-based clustering of annotators into a predefined number of clusters to preserve the perspectives of annotators while clustering. The approach is explained in the following conditions:

Condition 1: For a dataset, if the number of annotation per instance (n) and the total number of participating annotators (N) is equal, we treat each annotator as a distinct cluster. For example, if three annotators annotate each instance in a dataset and there are three participating annotators, then the data naturally form three clusters, with each annotator forming a cluster. In this case, the agreement-based clustering approach is not applied, and the number of clusters C equals the number of annotators N.

Condition 2: For a dataset annotated by N annotators with variable annotation coverage:

1. **Label matrix construction** Let $\mathcal{X} = \{x_1, \ldots, x_M\}$ be the set of data instances. Suppose there are N annotators, $\mathcal{A} = \{a_1, \ldots, a_N\}$. Annotator a_i labels some instances x_j as $l_{i,j}$. We construct an annotation matrix $L \in \mathbb{R}^{N \times M}$, where entry $L_{i,j}$ is the label from a_i for x_j. If a_i did not label x_j, the entry is left empty.

2. **Annotator similarity matrix** For each pair of annotators (a_i, a_k):

   - Identify the set of co-annotated instances: $S_{i,k} = \{x_j \mid l_{i,j} \text{ and } l_{k,j} \text{ exist}\}$
   - Compute an agreement score (Cohen's kappa for multiclass and Jaccard similarity for multi-label comparing sets of labels): $\text{sim}(i,k) = c(\{l_{i,j} \mid x_j \in S_{i,k}\}, \{l_{k,j} \mid x_j \in S_{i,k}\})$ where $c(\cdot,\cdot)$ is the agreement function.

   All $\text{sim}(i,k)$ values form the similarity matrix $A \in \mathbb{R}^{N \times N}$. We convert to a distance matrix:$D_{i,k} = 1 - \text{sim}(i,k)$

3. **Annotator clustering** Apply a clustering algorithm (e.g., k-means clustering) using D as the distance matrix. The number of clusters C is set as: $C = \min_j |\{i : l_{i,j} \text{ exists}\}|$, ensuring each instance is represented by at least one cluster.

   The result is a set of clusters $\mathcal{C}_1, \ldots, \mathcal{C}_C$, with $\mathcal{C}_c \subseteq \mathcal{A}$.

4. **Cluster-level label aggregation** For each instance x_j and cluster $\mathcal{C}_c$:

- Collect all labels for x_j from annotators in $\mathcal{C}_c$.
- If a majority label is available between, assign it as the cluster label for x_j.
- **If there is a tie or multi-label:**
  - If the task is multi-label, such as the multi-label emotion dataset, assign all available labels.
  - Otherwise, find a third or fifth annotator (adjudicator) for the specific tied instance as a tie-breaker and take the majority vote.

So for each x_j, we have C cluster labels - a label(s) after aggregations of annotators.

Number of Clusters We consider conditions to decide the number of clusters. 1) The minimum number of annotators can be found in any annotation condition, including instances that have more annotators. For example, if the minimum number of annotators per instance in a dataset is 3, the cluster can be 3; this covers cases with more annotators, and the number of clusters can be determined based on the availability of computational resources and data factors. 2) At the end, the results of the disagreement modeling approach is evaluated against the majority-voted gold labels. To decide the instance x_j's label(s) within the cluster, the number of clusters is preferably kept odd (e.g., 3, 5, ...) because an odd number of clusters eliminates the possibility of ties, as detailed in Algorithm 1.

Algorithm 1 Annotators Clustering

Require: Label matrix $L \in \mathbb{R}^{N \times M}$
Require: Agreement function $c(\cdot,\cdot) or j(\cdot,\cdot)$
Ensure: Annotator clusters based on agreement
Compute similarity matrix A from L using c
Compute distance matrix $D = 1 - A$
Number of clusters $C = \min_j |\{i : l_{i,j} \text{ exists}\}|$
Cluster annotators into $\mathcal{C}_1, \ldots, \mathcal{C}_C$ using D
for each instance x_j **do**
  for each cluster $\mathcal{C}_c$ **do**
    Aggregate labels for x_j from $\mathcal{C}_c$ members
    Assign cluster label for x_j
  end for
end for
return Annotator clusters

A high label agreement $c(i, j)$ indicates that annotators i, j tend to give similar labels on the items (texts). The lower distance indicates higher agreement based on their distances D to one another; annotators and their labels are clustered into a predefined number of clusters using $k - means$ clustering. After clustering, we evaluate three commonly used **aggregation approaches** in addition to the majority vote baseline: ensemble, multi-label, and multitask learning. These methods operate on non-aggregated annotator labels, preserving the individual perspectives rather than aggregating them into a single label, as is done in the majority vote. Figure 2 illustrates the schematic differences between these approaches.

3.1 Majority Vote

The majority vote is the most commonly used baseline, which aggregates annotations without considering annotator-specific information. It involves training a single-task classifier to predict the aggregated label for each instance, where the gold label is determined by majority voting. We use the majority vote as a baseline for comparison with the proposed approaches.

3.2 Ensemble Approach

This approach trains a separate model for each annotator to predict their individual labels. During inference, the predictions of these individual models are aggregated via majority vote to produce the final label (Akhtar et al., 2020). In our work, we extend this idea to both annotator-level and cluster-based ensembles. We train a model for individual annotators as well as clustered subsets, and evaluate their outputs independently.

3.3 Multi-label Approach

This approach treats each annotation as an individual label and formulates the problem as a multi-label classification task (Davani et al., 2022). The model incorporates a fully connected layer to project the input representation into a vector of annotator-specific label predictions, followed by a sigmoid activation to obtain independent probability scores for each annotator.

3.4 Multitask Approach

The multitask approach models each annotator or cluster as a separate classification task. All tasks share a common encoder to generate sentence representations, while each task has its own task-specific fully connected layer followed by a softmax activation. In contrast to the multi-label ap-

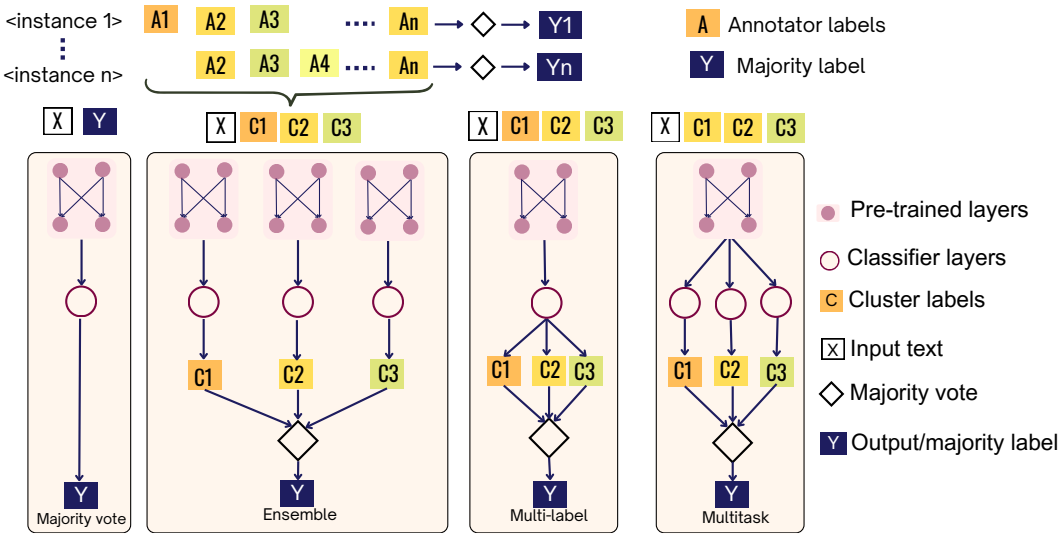


Figure 2: Overview of multi-annotator modeling architectures. Our new contribution is in the clustering of annotators before modeling. In the figure, N annotators, $A1, A2, A3, ...An$, are clustered into C clusters based on annotator label agreement. For visualization, we limit the number of clusters to three.

proach, multitask learning explicitly trains a separate output layer for each task (Liu et al., 2019).

4 Data

This section provides an overview of the datasets used in this study. Although many publicly available datasets exist, annotator-level datasets remain scarce, particularly for languages other than English. For this study, we selected recently available datasets for three highly subjective NLP tasks: hate speech detection, sentiment analysis, and emotion classification. These datasets include only anonymous annotator identifiers (ID numbers), which enable clustering based on annotation behavior. The summary of the dataset is presented in Table 1.

4.1 AfriSenti - Sentiment Analysis

AfriSenti (Muhammad et al., 2023) is a sentiment analysis dataset targeted at 14 African languages. However, annotator-level data is available for 12 of these languages. The data was sourced from X (formerly Twitter), and the classes are either *positive*, *negative*, or *neutral*.

4.2 AfriEmo - Emotion Classification

The SemEval-2025 Task 11 dataset (Muhammad et al., 2025b) is a multilingual emotion dataset covering 32 languages from diverse domains. For this work, we focus on the annotator-level data available for 14 African languages. We refer to this subset as **AfriEmo**. It was annotated based on Ekman's six basic emotions (Ekman et al., 1999) (anger, disgust, fear, joy, sadness, and surprise) in a binary option with a yes or no and intensity scores. We use only the binary emotion annotation.

4.3 AfriHate - Hate Speech Detection

AfriHate (Muhammad et al., 2025a) is a hate speech dataset covering 15 African languages; 14 languages captured annotation-level data and corresponding annotator anonymous IDs. Each instance has been annotated by 3 to 4 annotators, and the classes are either *abuse*, *hate*, or *neutral*. The final gold label was determined by majority voting, i.e., two out of three labels or three out of four labels. In cases of a tie among four annotators (i.e., two annotators selecting one label and two selecting another), as occurs in the `amh`, `orm`, `som`, `tir`, and `yor` datasets, the 'hate' label was prioritized over other classes.

4.4 GoEmotions and GabHate Dataset

We evaluate two widely used English datasets: GoEmotions (Demszky et al., 2020) and the Gab Hate Corpus (GabHate) (Kennedy et al., 2022). GoEmotions was annotated for 27 emotion classes in a multi-label approach. However, our experiment used the six basic emotion classes, following the previous work (Davani et al., 2022). GabHate is annotated for whether the text contains

Task	Language (ISO code)	Train	Dev	Test	Total	#Anno /inst.	#Total Anno.	Labels distribution in %	%full agr.	C. Kappa
Sentiment analysis	Amharic (**amh**)	1,300	200	500	2,000	3	3	Pos (22), Neg (67), Neu (11)	63.55	53.65
	Moroccan Arabic (**ary**)	2,971	458	1,142	4,571	3	9	Pos (30), Neg (34), Neu (36)	52.48	52.55
	Hausa (**hau**)	15,360	2,364	5,907	23,631	3	3	Pos (30), Neg (29), Neu (42)	61.46	61.07
	Igbo (**ibo**)	19,215	2,957	7,391	29,563	3	3	Pos (25), Neg (22), Neu (53)	62.86	59.66
	Kinyarwanda (**kin**)	2,615	403	1,006	4,024	3	3	Pos (26), Neg (33), Neu (40)	38.10	37.96
	Oromo (**orm**)	1,749	270	672	2,691	3	3	Pos (24), Neg (35), Neu (41)	39.58	35.11
	Nigerian Pidgin (**pcm**)	11,207	1,725	4,310	17,242	3	11	Pos (34), Neg (55), Neu (12)	50.78	46.36
	Portuguese-MZ (**ptMZ**)	5,667	872	2,180	8,719	3	6	Pos (16), Neg (18), Neu (66)	45.83	33.82
	Tigrinya (**tir**)	1,560	240	600	2,400	3	3	Pos (29), Neg (49), Neu (21)	65.83	61.81
	Xitsonga (**tso**)	662	111	331	1,104	3	3	Pos (47), Neg (35), Neu (18)	43.12	43.19
	Twi (**twi**)	3,043	469	1,171	4,683	3	3	Pos (47), Neg (38), Neu (15)	49.54	47.16
	Yorùbá (**yor**)	17,807	2,740	6,849	27,396	3	4	Pos (34), Neg (17), Neu (49)	63.00	60.19
Emotion classification	Amharic (**amh**)	3,844	593	1,478	5,915	5	5	26, 28, 2, 12, 16, 4, 12	11.11	39.52
	English (**eng**)	43,410	5,426	5,427	54,263	3.58	82	3, 0.5, 0.5, 8, 1.5, 2.5, 6	31.98	29.40
	Moroccan Arabic (**ary**)	1,746	270	671	2,687	3	3	21, 4, 7, 17, 14, 14, 25	26.87	41.63
	Hausa (**hau**)	2,327	359	895	3,581	5.03	7	16, 12, 12, 12, 25, 13, 9	23.21	54.13
	Igbo (**ibo**)	3,121	481	1,201	4,803	3	3	19, 18, 7, 16, 17, 3, 21	32.50	46.66
	Kinyarwanda (**kin**)	2,656	410	1,023	4,089	3	3	17, 5, 5, 16, 24, 4, 28	50.55	63.27
	Oromo (**orm**)	3,499	539	1,346	5,384	3	6	17, 16, 3, 31, 8, 4, 22	43.59	53.58
	Nigerian Pidgin (**pcm**)	4,041	623	1,554	6,218	3	3	6, 35, 8, 11, 18, 18, 3	13.48	43.03
	Portuguese-MZ (**ptMZ**)	1,243	208	622	2,073	4	4	7, 5, 9, 16, 22, 0, 41	30.05	20.82
	Somali (**som**)	3,229	498	1,242	4,969	3	6	10, 12, 8, 17, 12, 5, 36	30.83	42.53
	Swahili (**swa**)	3,583	552	1,379	5,514	3	3	9, 7, 3, 13, 10, 16, 43	14.51	19.97
	Tigrinya (**tir**)	3,583	554	1,380	5,523	3	6	13, 31, 3, 10, 14, 9, 19	24.21	41.50
	isiXhosa (**xho**)	1,365	228	682	2,275	3	3	4, 0.5, 2, 28, 38, 15, 13	39.74	41.15
	Yorùbá (**yor**)	3,242	500	1,247	4,989	3	3	6, 3, 3, 9, 27, 8, 44	33.09	33.79
	isiZulu (**zul**)	1,899	293	730	2,922	3	3	8, 3, 3, 5, 19, 8, 54	46.17	40.05
Hate speech classification	Amharic (**amh**)	3,132	482	1,205	4,819	3	11	Hate (45), Abuse (28), Neu (27)	53.46	59.67
	English (**eng**)	22,124	–	5,531	27,655	3.13	18	Hate (2,599), non-hate (24,977)	75.38	28.00
	Moroccan Arabic (**ary**)	2,704	416	1,040	4,160	3	3	Hate (15), Abuse (54), Neu (31)	68.68	65.39
	Hausa (**hau**)	5,311	818	2,043	8,172	3	3	Hate (1), Abuse (31), Neu (68)	76.10	64.90
	Igbo (**ibo**)	2,872	442	1,105	4,419	3	6	Hate (8), Abuse (68), Neu (24)	80.86	73.46
	Kinyarwanda (**kin**)	3,068	473	1,180	4,721	3	3	Hate (27), Abuse (24), Neu (49)	80.00	78.97
	Oromo (**orm**)	3,201	493	1,232	4,926	4	9	Hate (46), Abuse (13), Neu (41)	52.70	55.88
	Somali (**som**)	2,822	435	1,085	4,342	4	7	Hate (13), Abuse (26), Neu (59)	35.51	36.08
	Tigrinya (**tir**)	3,026	466	1,164	4,656	4	8	Hate (60), Abuse (22), Neu (18)	45.41	51.14
	Twi (**twi**)	2,534	390	975	3,899	3	3	Hate (11), Abuse (84), Neu (5)	75.46	48.40
	isiXhosa (**xho**)	2,347	362	903	3,612	3	3	Hate (8), Abuse (42), Neu (50)	61.21	54.76
	Yorùbá (**yor**)	2,715	418	1,045	4,178	4	4	Hate (5), Abuse (50), Neu (45)	76.42	70.79
	isiZulu (**zul**)	2,729	420	1,050	4,199	3	3	Hate (5), Abuse (43), Neu (52)	80.71	76.50

Table 1: Summary of datasets. From left to right: 1) **Task** name, 2) **Language** under each task; 3) the number of instances in **Train**, **Test**, **Dev**, and **Total**; 4) Numeber of Annotators per instance, 5) Total number of Annotators involved in the annotation, 5) **Label distribution in** % in the dataset: Sentiment (Pos, Neg, and Neu); Emotion - (anger, disgust, fear, joy, sadness, surprise, and no emotion); and Hate speech (abuse, hate or Neutral), respectively; 6) Percentage (%) of full agreement - all annotators agreed; and 7) overall **Cohen's Kappa** agreement.

hate speech or not.

5 Experimental Setup

This section presents the language models we fine-tuned and the evaluation setups we followed.

5.1 Multilingual Language Models

For the English language experiments, we use two models. In the hate speech task (GabHate), we use XLM-T (Barbieri et al., 2022), an XLM-R model trained on X (Twitter) data. For the GoEmotions data, we evaluate *XLM-Roberta-base* (Conneau et al., 2020). Based on benchmark results reported in the African languages dataset papers (Muhammad et al., 2023, 2025a,b), the best performance is achieved using AfroXLMR, an African-centric language model (Alabi et al., 2022). AfroXLMR is an encoder-only model that extends XLM-R (Conneau et al., 2020) by adding additional pre-training on 76 African languages. Our evaluation for African languages is based on AfroXLMR, as it covers all the target languages mentioned in this work. This specific version of the model can be found in Hugging Face[1].

Training Details The model checkpoint is accessed from the Hugging Face with the transform-

[1] https://huggingface.co/Davlan/afro-xlmr-large-76L

Lang.	Baseline	Individual annotator			Clustered annotators		
	Majority	Ensemble	Multi-label	Multitask	Ensemble	Multi-label	Multitask
Sentiment analysis							
ary	$50.3_{\pm 3.0}$	$42.2_{\pm 1.7}$	$29.5_{\pm 3.9}$	$33.0_{\pm 2.7}$	$\mathbf{53.3}_{\pm 2.4}$	$44.5_{\pm 1.1}$	$52.2_{\pm 3.3}$
pcm	$23.5_{\pm 1.5}$	$24.0_{\pm 0.7}$	$50.0_{\pm 0.9}$	$25.1_{\pm 3.2}$	$51.4_{\pm 1.7}$	$\mathbf{71.8}_{\pm 3.5}$	$62.5_{\pm 2.7}$
ptMZ	$26.5_{\pm 2.1}$	$58.9_{\pm 0.6}$	$26.5_{\pm 3.3}$	$32.9_{\pm 3.3}$	$56.3_{\pm 3.9}$	$\mathbf{66.3}_{\pm 0.9}$	$62.3_{\pm 0.8}$
yor	$51.9_{\pm 1.1}$	$68.2_{\pm 0.2}$	$70.9_{\pm 2.2}$	$70.5_{\pm 2.3}$	$65.7_{\pm .5}$	$\mathbf{72.6}_{\pm 0.5}$	$71.9_{\pm 1.6}$
Emotion analysis							
eng	$46.0_{\pm 2.0}$	$46.9_{\pm 0.0}$	$55.5_{\pm 3.9}$	$45.2_{\pm 2.7}$	$\mathbf{60.4}_{\pm 3.0}$	$50.9_{\pm 2.1}$	$50.1_{\pm 3.7}$
hau	$69.0_{\pm 1.9}$	$66.3_{\pm 0.6}$	$67.0_{\pm 1.7}$	$65.0_{\pm 0.9}$	$66.0_{\pm 2.1}$	$68.0_{\pm 0.7}$	$\mathbf{71.6}_{\pm 1.1}$
orm	$41.7_{\pm 3.2}$	$37.0_{\pm 1.0}$	$61.7_{\pm 1.9}$	$57.77_{\pm 0.6}$	$46.8_{\pm 3.3}$	$\mathbf{61.1}_{\pm 0.8}$	$53.9_{\pm 2.1}$
som	$47.7_{\pm 3.4}$	$24.1_{\pm 4.1}$	$49.6_{\pm 0.3}$	$47.2_{\pm 1.5}$	$43.0_{\pm 1.8}$	$45.0_{\pm 0.9}$	$\mathbf{49.8}_{\pm 1.8}$
tir	$48.0_{\pm 2.6}$	$36.1_{\pm 3.1}$	$51.9_{\pm 2.3}$	$52.4_{\pm 1.9}$	$56.2_{\pm 2.2}$	$\mathbf{53.7}_{\pm 2.7}$	$54.4_{\pm 1.8}$
Hate speech analysis							
amh	$71.6_{\pm 0.8}$	$38.6_{\pm 2.6}$	$71.3_{\pm 3.6}$	$71.5_{\pm 2.2}$	$\mathbf{73.0}_{\pm 3.5}$	$71.4_{\pm 2.9}$	$72.5_{\pm 3.9}$
eng	$47.4_{\pm 4.2}$	$68.5_{\pm 2.0}$	$68.9_{\pm 1.9}$	$71.9_{\pm 2.7}$	$69.5_{\pm 1.5}$	$68.6_{\pm 1.9}$	$\mathbf{73.6}_{\pm 2.9}$
ibo	$83.7_{\pm 4.0}$	$69.8_{\pm 4.2}$	$27.0_{\pm 3.9}$	$\mathbf{88.2}_{\pm 2.3}$	$80.4_{\pm 3.3}$	$87.6_{\pm 1.7}$	$84.0_{\pm 0.5}$
orm	$64.6_{\pm 0.9}$	$62.8_{\pm 1.1}$	$65.2_{\pm 1.3}$	$61.9_{\pm 2.2}$	$70.3_{\pm 0.5}$	$\mathbf{73.6}_{\pm 0.3}$	$71.9_{\pm 0.8}$
som	$50.4_{\pm 1.1}$	$62.7_{\pm 1.3}$	$58.7_{\pm 0.7}$	$61.9_{\pm 0.6}$	$64.9_{\pm 1.8}$	$\mathbf{66.4}_{\pm 1.5}$	$59.4_{\pm 0.5}$
tir	$67.8_{\pm 1.5}$	$71.4_{\pm 1.7}$	$70.4_{\pm 1.5}$	$73.3_{\pm 0.3}$	$64.0_{\pm 0.1}$	$\mathbf{73.9}_{\pm 0.8}$	$71.1_{\pm 1.1}$

Table 2: Performance comparison between **baseline**(majority vote), modeling **individual annotators**, and **clustered annotators** across aggregation methods and datasets. The results are from the number of clusters 3 for all languages, except *hau* = 5. The **boldfaced** results are the best across aggregation approaches for individual annotator modeling and after clustering.

ers library. For training, we fine-tune AfroXLMR for sequence classification using a batch size of 16, a maximum sequence length of 128, over 3 epochs, with a learning rate of 2e-5 consistently across all experiments for reproducibility. Nevertheless, our proposed approaches are not limited to a single model.

5.2 Evaluation Setup

We evaluate model performance using standard text classification metrics: accuracy and macro-F1, we report macro-F1 because of highly label distribution imbalance across the datasets as shown in Table 1. Each experiment is repeated five times, and the reported results are the average across these runs. To evaluate each aggregation approach, the final prediction is obtained by applying a majority vote over the outputs of the cluster models.

6 Experiment Results

In this section, we present results for individual annotator modeling and the proposed clustering approach using the evaluation dataset presented in Table 1.

6.1 Individual Vs. Clustered Annotators

The results in this section compare the languages where the clustering approach was applied. The ensemble approach using individual annotators serves as the baseline for evaluating our proposed clustering method. We report results by comparing models trained on individual annotators with those trained on agreement-based clustered annotators. Table 2 presents the results from the *individual annotator* and the *clustered annotators*.

Modeling Individual Annotator: While the baseline model is trained and tested on the majority vote, individual annotator modeling requires training and evaluating a separate model for each annotator using their respective labels. For example, in the sentiment analysis dataset, the *pcm* language includes 11 annotators, and we train one model per annotator. The final prediction is obtained by ensembling the individual predictions through majority voting.

Modeling Clusters of Annotators: Following the same approach as individual annotator modeling, we train a model for each cluster and aggregate the predictions using majority voting for final evaluation. For the English GabHate dataset, we used individual modeling results from previous work by Rodríguez-Barroso et al. (2024) and replicated the same experiment setup for clustered approaches. For the English GoEmotions dataset, we used the ensemble of individual annotator results implemented in Davani et al. (2022), and we train models using a cluster-based approach.

Lang.	# Anno.	Majo.	Indi.	Clust.
Sentiment analysis				
ary	9	50.3	42.2	**53.3**
pcm	11	23.5	24.0	**51.4**
ptMZ	6	26.5	**58.9**	56.3
yor	4	51.9	**68.2**	65.7
Emotion analysis				
eng	82	46.0	46.9	**60.4**
hau	7	69.0	66.3	**66.0**
orm	6	41.7	37.0	**46.8**
som	6	**47.7**	24.1	43.0
tir	6	48.0	36.1	**56.2**
Hate speech analysis				
amh	11	71.6	36.6	**73.0**
eng	18	47.4	68.5	**69.5**
ibo	6	**83.7**	69.8	80.4
orm	9	64.6	62.8	**70.3**
som	7	50.4	62.7	**64.9**
tir	8	67.8	**71.4**	64.0

Table 3: Zooming macro F1 score results from Table 2 for majority vote (**Majo.**), ensemble of individual annotator (**Indi.**), and ensemble of clustered annotators (**Clust.**).

Table 3 (summarized zoom-in results of Table 2) shows the results from the *Majority vote*, *Individual Annotator* and the *clustered Annotators*.

The results demonstrate that the proposed clustering approach outperforms the majority vote and the individual-annotator modeling across the datasets. Among the aggregation approaches, multi-label and multitask aggregation achieve better performance. One contributing factor to the performance gap is the uneven distribution of annotations among annotators, which results in insufficient representation during individual model training. For instance, in the *pcm* sentiment analysis dataset, 11 annotators participated, but one annotator contributed 7,982 annotations (46.3%), while another provided only 676 (3.9%). This imbalance in annotation distribution leads to a skewed learning signal, limiting the model's ability to effectively generalize each annotator's perspective. As each base model in the individual ensemble is trained solely on the subset of data labeled by its corresponding annotator, annotators with fewer annotations yield weaker models, reducing the overall performance of the ensemble. In contrast, the agreement-based clustering approach mitigates this issue by forming clusters with a larger number of instances while preserving the diversity of annotator perspectives. For the same *pcm* languages, the resulting clusters are *cluster_1*: 9,736, *cluster_2*: 1,266, and *cluster_3*: 8,208, from a total of 11K training instances.

6.2 Best Cluster Aggregation Approaches

Sentiment Analysis Results Table 4(a) presents results for the Sentiment analysis dataset. For the ensemble, multi-label, and multitask models, we evaluate how well the majority vote of predicted labels from individual cluster models aligns with the pre-modeling majority vote. As a result, multi-label and multitask aggregation approaches outperform the ensemble and majority-voted labels. This suggests aggregating annotations before modeling discards valuable annotator-specific perspectives and may introduce label noise. The predicted results are more stable and accurate when clusters are modeled independently and their internal consistency is leveraged. Agreement-based clustering within a multi-label aggregation further enhances performance, achieving state-of-the-art results compared to the majority-voting benchmarks.

Performance correlates with class distribution at the sentiment-class level: more-represented classes yield higher scores. For example, in the *amh* dataset, 67% of the instances are labeled as *Negative*, achieving a macro-F1 score of 84.5%. In contrast, the *Neutral* class, which constitutes only 11% of the data, obtains a lower score of 37.7%.

Emotion Analysis Results Table 4 (b) presents results for the Emotion analysis dataset. The multitask aggregation approach performs better across most languages. This multitask out-performance is consistent with the observation that many an-

a) **Sentiment analysis** results

Aggreg.	amh	ary	hau	ibo	kin	orm	pcm	ptMZ	tir	tso	twi	yor
Maj. vote	53.6	50.3	78.1	72.2	19.2	**36.6**	23.5	26.5	65.6	31.3	33.1	21.9
Ensemble	50.3	**53.3**	77.5	23.2	67.5	31.6	51.4	56.3	61.0	25.4	**43.2**	65.7
Multi-label	**72.9**	44.5	77.1	37.1	64.7	25.3	**71.8**	**66.3**	**68.4**	30.2	30.6	**72.6**
Multitask	26.7	52.2	**78.6**	**73.0**	**69.9**	32.9	62.5	62.3	34.9	**43.7**	40.9	71.9

b) **Emotion analysis** results

Aggreg.	amh	eng	ary	hau	ibo	kin	orm	pcm	ptMZ	som	tir	xho	yor	zul
Maj. vote	**67.7**	46.0	43.1	69.0	42.1	53.9	41.7	**48.4**	58.1	47.7	48.0	28.3	20.4	16.4
Ensemble	46.2	**60.4**	44.2	66.0	45.9	50.3	46.8	43.0	09.1	43.0	56.2	28.1	26.0	17.5
Multi-label	61.6	50.9	32.1	68.0	41.8	53.7	**61.1**	48.1	21.7	45.3	**53.7**	30.5	**33.0**	18.6
Multitask	64.2	50.1	**49.1**	**71.6**	**51.6**	**54.3**	53.9	47.6	**59.0**	**49.8**	54.4	**31.8**	30.7	**20.7**

c) **Hate speech analysis** results

Aggreg.	amh	eng	ary	hau	ibo	kin	orm	som	tir	twi	xho	yor	zul
Maj. vote	71.7	47.4	69.4	48.2	83.7	75.6	64.6	50.4	67.8	60.4	74.1	61.7	21.0
Ensemble	**73.0**	69.5	72.5	35.9	80.4	77.8	70.3	66.9	73.7	57.4	78.6	58.5	82.9
Multi-label	71.4	68.6	37.5	**83.2**	87.6	74.6	**73.6**	**66.4**	**73.9**	**76.7**	76.6	**83.3**	83.3
Multitask	72.5	**73.6**	**75.8**	77.3	**84.0**	**78.9**	71.9	59.4	71.1	60.2	**83.1**	63.1	**89.7**

Table 4: **Sentiment**, **Emoiton**, and **Hate Speech** analysis macro F1 score results. Column names are language codes. We highlight the best aggregation approach results in bold. **Aggreg**.(aggregations) are the cluster aggregation approaches at the end.

notators contribute only a few annotations. For instance, in the *hau* dataset, each instance is annotated by 5 to 7 annotators, but less than 3% of the data includes annotations from the 6th and 7th annotators, indicating sparsity.

At the emotion-class level, performance varies depending on label distribution. In the *swa* dataset, the emotion classes *disgust* and *fear* account for only 7% and 3% of the data, respectively, making them challenging to learn. Similarly, the *xho* dataset performs poorly on *anger*, *disgust*, and *fear*, which comprise just 4%, 0.5%, and 2% of the dataset, respectively. Overall, the emotion classes *fear* and *surprise* consistently yield the lowest performance, while *joy* and *sadness* achieve comparatively higher performance.

Hate Speech Analysis Results Table 4(c) presents results for the hate speech dataset. Consistent with prior findings in sentiment and emotion analysis, the multitask aggregation approach outperforms other methods across most languages. In contrast, the majority vote baseline consistently underperforms. Class-level performance is closely tied to label distribution: classes with more annotated examples yield better results. For instance, in the *yor* dataset, the model struggles to predict the hate class, which comprises only 5% of the data. This trend highlights the challenge of learning from imbalanced class distributions.

Finally, languages with larger training datasets generally achieve better performance. For example, *hau*, which has approximately twice the data size of other languages, achieves superior results, underscoring the importance of balanced label distributions and sufficient data for effective classification.

Summarized Results Table 5 shows the summarized results of the tasks and aggregation approaches at the language level presented in Table 4. Based on the summary, the multitask aggregation approach outperforms other aggregation approaches (18/40 language datasets), with multi-label being the second best (14/40 language datasets). In addition to achieving better results with multitask and multi-label approaches, they are also efficient in terms of computational resources, such as memory and training time. This is because they train a cluster of C models rather than an individual model per annotator.

7 Discussion and Analysis

Individual vs. Agreement-based Clustering: Our agreement-based clustering approach is more robust than individual annotator modeling and ma-

Tasks	Models	Best for	Avg.
Sentiment analysis	Maj. vote	1/12 lang.	42.7
	Ensemble	2/12 lang.	50.5
	Multi-label	5/12 lang.	55.1
	Multitask	4/12 lang.	54.1
Emotion analysis	Maj. vote	2/14 lang.	45.1
	Ensemble	1/14 lang.	41.6
	Multi-label	3/14 lang.	45.7
	Multitask	8/14 lang.	**49.2**
HateSpeech analysis	Maj. vote	0/13 lang.	61.2
	Ensemble	1/13 lang.	69.0
	Multi-label	6/13 lang.	73.6
	Multitask	6/13 lang.	**73.9**

Table 5: Summarized results across modeling disagreement approaches and datasets. The **Best for** column indicates the number of languages (out of the total evaluated) for which a specific model type achieved the best performance. **Avg.** is average macro F1 score across total number languages.

jority vote baselines. By clustering annotators based on their annotation behavior, we capture individual perspectives that improve performance, provide stable training, and reduce computational resources. This advantage arises because some annotators label very few instances, while others label nearly all, making individual annotator-level modeling inefficient and inconsistent.

As expected, individual annotator modeling is the most resource-intensive approach, requiring training N separate models (where N is the number of annotators), each with varying training sizes. In contrast, the clustering approach requires training fewer models (e.g., 3 or 5 in our experiments), resulting in faster and more efficient training. Notably, the multi-label and multitask model training and aggregation approaches offer a favorable trade-off, delivering strong performance without significantly increasing training time.

Performance in Terms of Agreement: We investigate how inter-annotator agreement impacts the performance of different aggregation approaches. Specifically, we examine whether there is a correlation between model performance and overall annotator agreement. Our findings indicate a general trend: higher inter-annotator agreement corresponds to better model performance. For instance, languages such as *hau*, *tir*, and *yor*, which exhibit relatively high Cohen's Kappa scores of 61.5%, 61.8%, and 63.0%, respectively, achieve superior results in the sentiment analysis task. This suggests a direct relationship between annotation consistency and model effectiveness. However, we note that these comparisons are not direct due to variation in data size across languages, which may also influence performance outcomes.

Number of Annotators vs. Agreement: We examine whether the number of annotators involved in the annotation process is related to overall pairwise agreement, as measured by Cohen's Kappa. We observe a trend in which an increase in the number of annotators tends to correspond to a decrease in agreement scores. For instance, in the sentiment analysis dataset, *ptMZ*, *ary*, and *pcm* involved 6, 9, and 11 annotators, respectively, but achieved relatively lower Cohen's Kappa scores of 33.8%, 52.6%, and 46.4%. Similarly, in the hate speech dataset, languages such as *som* (7 annotators), *tir* (8), *orm* (9), and *amh* (11) recorded lower agreement scores of 36.1%, 51.1%, 55.9%, and 59.7%, respectively, lower than other languages with fewer annotators. In terms of performance, datasets with four annotators and a majority vote threshold requiring at least two votes tend to yield better results. This is evident in the AfriHate dataset, where languages such as *amh*, *tir*, *orm*, and *som* exhibit relatively stronger model performance under this setting.

Performance Across Tasks: As presented in Table 5, a comparison across the three tasks reveals that the emotion classification task consistently performs worse than sentiment analysis and hate speech detection tasks. Sentiment task records an average macro F1 score of 55.1% across 12 languages, whereas Emotion and hate speech analysis tasks achieve 49.2% and 73.9%, respectively. This performance gap can be primarily attributed to the inherent complexity of the multi-label emotion classification task.

Computational Resources Individual annotator modeling is computationally very demanding, particularly in terms of memory requirements and training time, since a separate model must be trained for each annotator. For example, the model we used for African languages requires 2.24 GB of memory. Training a separate model for each of the 11 annotators in the `pcm` languages would therefore require 35 minutes (depending on the number of instances per annotator) on a single GPU and approximately 2.24GB×11 of memory size. The

same scaling applies to crowd-sourced datasets such as the English GoEmotions dataset, which involves 82 annotators. To improve efficiency and performance, our proposed clustering of annotators based on their agreement is effective.

8 Conclusion

In this work, we presented an agreement-based annotator clustering approach for modeling annotation disagreement in subjective NLP tasks, providing a more effective alternative to aggregating annotations through majority voting. The multi-task and multi-label aggregation approaches outperform both the majority-vote and individual-annotator modeling approaches. In addition to achieving better performance, these approaches are also more computationally efficient, as they require training only a cluster of C models rather than one model per annotator. The work addressed key limitations highlighted in previous works, which involved training separate models for each annotator. In contrast, our approach efficiently captured diverse annotator perspectives while reducing training overhead and deferring label aggregation until the final prediction stage. This makes it especially useful for real-world applications such as social media monitoring, opinion mining, and others, where understanding subjectivity and disagreement is crucial. Finally, our results across diverse datasets demonstrate that leveraging the full spectrum of annotator input, rather than collapsing annotations into a single majority-voted label, significantly enhances classification performance in subjective NLP tasks while preserving individual annotator perspectives.

Limitations

Our work is not without limitations.

Model Varieties: Due to high computational resources across languages, tasks, and the number of annotators, our experimental setup is limited to a single model evaluation with the least hyperparameter settings. Each downstream NLP task and aggregation approach might need different hyperparameter settings. In the future, this work can be extended to evaluate multiple variants of pre-trained language models and both open-source and proprietary LLMs.

Number of Clusters: Our agreement-based clustering approach groups annotators into C clusters, where the number of clusters C could be determined based on the availability of computational resources and data factors. Our clustering principle works for any number of clusters and can be extended to any custom number of clusters based on the computational resource availability. This is because, as the number of clusters increases, computational resources are also increased. However, we recommend using an odd number of clusters and applying the majority vote rule to the final label for evaluation after modeling each cluster separately. The clustering principle works for any number of clusters and can be extended to a more custom number of clusters based on the training resource availability. For example, if the dataset is annotated by a fixed number of annotators, equally, a different number of clusters may work better. However, in our work, we only used clustering into 3 and 5 clusters.

In a dataset annotated by many annotators, where each instance (text) is annotated equally and requires a custom number of clusters, this method is not applied because the number of annotations for a text equals the number of clusters. Exploring Other clustering approaches, such as Hierarchical and Spectral clustering of annotators, clustering algorithms like Krippendorff's alpha, and clustering methods for annotators that do not have common annotated instances, is an area of open research that will also be an open area for further investigation.

Aggregation Methods: This work, as well as previously conducted works that are mentioned in the related works section, uses a majority vote to decide the final predicted label after modeling the perspectives of annotators. In individual modeling, we train a model for each annotator separately and evaluate using test sets. It does not matter how many annotators annotate each instance, as each annotator uses their own annotation to train an individual model. During clustering, no further individual annotation is provided; annotators are grouped into a cluster, and only cluster-level annotation is provided. We can assume one cluster as an individual annotator. However, other aggregation techniques should be explored, such as using annotator soft labels as gold labels for the final evaluation.

Acknowledgments

We are very grateful to the editors and anonymous reviewers for their constructive comments. The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award. We thank the authors of the dataset used in this paper, who made these subjective NLP task datasets, including individual annotator levels, available. We appreciate and recommend releasing such individual annotator levels for the subjectivity study.

References

Yaw Afriyie and Benjamin A Weyori. 2026. Sentiment analysis based on deep learning approaches for text classification. *Journal of Electrical Systems and Information Technology*, 13(1):14.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 151–154.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Daniel Braun. 2024. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. *Artificial intelligence and law*, 32(3):839–862.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. Kaggle.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Guillaume Daval-Frerot and Yannick Weis. 2020. WMD at SemEval-2020 tasks 7 and 11: Assessing humor and propaganda using unsupervised data augmentation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1865–1874, Barcelona (online). International Committee for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–14. Association for Computing Machinery.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Michael Fell, Sohail Akhtar, and Valerio Basile. 2021. Mining annotator perspectives from hate speech corpora. In *NL4AI@ AI* IA*.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Ratnavel Rajalakshmi. 2025. Overview of the shared task on multimodal hate speech detection in Dravidian languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 114–122, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22(14).

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '16, page 1637–1648, New York, NY, USA. Association for Computing Machinery.

Prashant Kapil and Asif Ekbal. 2024. A Survey on Combating Hate Speech through Detection and Prevention in English. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 485–501, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. Crowdcalibrator: Can annotator disagreement inform calibration in subjective tasks?

Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. 2015. A bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*, pages 181–189. Society for Industrial and Applied Mathematics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Lei Li, Zuying Huang, and Yinan Liu. 2019. Multi-lingual Wikipedia summarization and title generation on low resource corpus. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pages 17–25, Varna, Bulgaria. INCOMA Ltd.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *NLPerspectives@ ECAI*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Eugenio Martínez Cámara, Yudivián Almeida-Cruz, Manuel Carlos Díaz Galiano, Suilan Estévez-Velarde, Miguel Ángel García Cumbreras, Manuel García Vega, Yoan Gutiérrez, Arturo Montejo Ráez, Andrés Montoyo, Rafael Munoz, et al. 2018. Overview of tass 2018: Opinions, health and emotions.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, Samuel Rutunda, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Seid Muhie Yimam, and Nedjma Ousidhoum. 2025a. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2024. Rethinking emotion annotations in the era of large language models.

Odbal, Guanhong Zhang, and Sophia Ananiadou. 2022. Examining and mitigating gender bias in text emotion detection task. *Neurocomputing*, 493:422–434.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Unnati Parmar and Jatin Modh. 2026. Advancements in Sentiment Analysis: A Comprehensive Survey of Techniques, Models, and Real-World Applications. In *Information Systems for Intelligent Systems*, pages 301–308, Cham. Springer Nature Switzerland.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. 2025. Is a bunch of words enough to detect disagreement in hateful content? In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 1–11, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Michael Roth and Dominik Schlechtweg, editors. 2025. *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*. International Committee on Computational Linguistics, Abu Dhabi, UAE.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 155–164, New York, NY, USA. Association for Computing Machinery.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: quantifying annotation disagreement using demographic information. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, and Christopher Homan. 2022. Improving label quality by jointly modeling items and annotators. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 95–99, Marseille, France. European Language Resources Association.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 902–913.