Federated Learning for Exploiting Annotators' Disagreements in Natural Language Processing

Nuria Rodríguez-Barroso,[1] **Eugenio Martínez Cámara,**[3] **Jose Camacho Collados,**[4] **M. Victoria Luzón**[2] and **Francisco Herrera**[1]

[1]Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain
rbnuria@ugr.es, fherrera@decsai.ugr.es

[2]Department of Software Engineering, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain
luzon@ugr.es

[3]Department of Computer Science, University of Jaén, Spain
emcamara@ujaen.es

[4]Cardiff University, Cardiff, United Kingdom
CamachoColladosJ@cardiff.ac.uk

Abstract

The annotation of ambiguous or subjective NLP tasks is usually addressed by various annotators. In most datasets, these annotations are aggregated into a single ground truth. However, this omits divergent opinions of annotators, hence missing individual perspectives. We propose FLEAD (Federated Learning for Exploiting Annotators' Disagreements), a methodology built upon federated learning to independently learn from the opinions of all the annotators, thereby leveraging all their underlying information without relying on a single ground truth. We conduct an extensive experimental study and analysis in diverse text classification tasks to show the contribution of our approach with respect to mainstream approaches based on majority voting and other recent methodologies that also learn from annotator disagreements.

1 Introduction

Artificial intelligence (AI) and in particular natural language processing (NLP) are dominated by data-driven approaches that often require datasets with human judgments (Uma et al., 2022). The difficulty of annotating data is magnified in NLP due to the inherent ambiguity of text (Basile et al., 2021) and the subjectivity concerned in the evaluation of its meaning, which often depends on the interpretation of individual annotators (Sandri et al., 2023b). In other words, the annotation may be a reflection on some private state caused by emotions, sentiments, hate, or opinions of the author (Wiebe, 1990). This subjectivity is present in many NLP tasks such as sentiment analysis (Pang et al., 2008; Kenyon-Dean et al., 2018), offensive language detection (Basile, 2021) and hate speech analysis (Kocoń et al., 2021), to name a few. The participation of more than one human annotator per data instance is a common strategy to mitigate the ambiguity and subjectivity of language (Sandri et al., 2023b). Then, each item is adjudicated a gold label. This implies that a ground truth exists, which does not usually fit the real practice of text annotation in which disagreements are frequent among annotators (Plank et al., 2014b; Uma et al., 2022; Leonardelli et al., 2023).

There are several approaches to adjudicate a gold label overcoming the disagreement among annotators (Uma et al., 2022): (1) approaches which simply aggregate crowd annotations into (typically, one) gold label for each instance; (2) approaches which assume a gold label for each item but consider disagreement to filter or weigh items when the true label is uncertain; (3) approaches for directly learning a classifier from crowd annotations; and (4) approaches that train a classifier by combining both hard labels and soft labels obtained from crowd annotations. In this paper, we argue that integrating disagreement into the learning process brings about clear benefits, as the data perspectivism paradigm advocates (Basile

Transactions of the Association for Computational Linguistics, vol. 12, pp. 630–648, 2024. https://doi.org/10.1162/tacl_a_00664
Action Editor: Micha Elsner. Submission batch: 8/2023; Revision batch: 12/2023; Published 5/2024.

et al., 2023). Instead of relying on a single aggregated label, we can exploit all annotator' opinions by leveraging disagreement among annotators. To this end, we propose FLEAD (Federated Learning (FL) for Exploiting Annotators' Disagreements), a methodology that separately model each annotator and consolidates all these annotator-specific models into a global model that integrates all the annotators perspectives. Our solution relies on federated learning to model each annotator's behavior and to summarize them into a global model. Hence, our methodology does not rely on a single ground truth, but instead exploits the information provided by each annotator.

Our FL-based methodology requires datasets with the labels of different annotators, which is not a common practice in NLP, as the Perspectivist Data Manifesto[1] highlights. The emergence of this paradigm has led to the construction of datasets where individual annotator information is provided. Nonetheless, the availability of resources is mainly skewed to the English language. Thus, as an additional contribution of this paper we have created and annotated the multilingual sentiment analysis dataset SentiMP for English, Spanish, and Greek.

Finally, we evaluate our methodology on this and other datasets from subjective NLP tasks, and we compare it with other approaches taking into account all the annotators' information (Davani et al., 2022). The results highlight the benefits of our approach with respect to dominant paradigms and previous work on modelling disagreement. To better understand the behavior of the FLEAD methodology, we perform an extensive analysis, including targeted ablations on the components of our proposed methodology.

2 Related Work

The quality of supervised learning models in NLP depends on the quality of the annotation of the datasets. This paradigm requires a human interpretation of annotation guidelines and text content that may cause disagreements among the annotators (Basile et al., 2021; Parmar et al., 2022; Jiang and Marneffe, 2022; Pavlick and Kwiatkowski, 2019). In many cases, the disagreements are considered to be noise, and they tend to be filtered out by adjudicating a single gold label to each item. However, the use of disagreement as learning signal has been proved useful in NLP and other areas of artificial intelligence (Uma et al., 2022). Sandri et al. (2023a) went a step further defining and providing a taxonomy of different types of disagreements among annotators, including an offensive language classification case study on the MD-Agreement dataset (Leonardelli et al., 2021).

There are various strategies to learn from crowd annotations. The prevalent method in the literature consists of the aggregation of the annotators judgments into a single label (Paun et al., 2018), for example via a majority vote. Depending on the type of the annotation, other heuristics to reduce the noise of the annotations are the following: weighting labels according to probability distributions (Jamison and Gurevych, 2015; Peterson et al., 2019); re-annotation (Sheng et al., 2008); filtering hard items (Reidsma and op den Akker, 2008); adapting the original labels to probabilistic ones based on the labels of all annotators (Sakaguchi and Van Durme, 2018; Chen et al., 2020; Plank et al., 2014a); adding a specific layer in an end-to-end model to learn the individual behavior of the annotators (Rodrigues and Pereira, 2018; Sullivan et al., 2023; Shahriar and Solorio, 2023); or adding information about the annotators and labels into the model (Yin et al., 2023). In our case, and building on the success of language models, we propose a single model that learns from all the non-aggregated annotations, without altering the language model or the individual labels. However, our proposed FLEAD methodology can be applied to any context with multiple annotations, regardless of the task and the underlying learning model.

Most similar to our methodology are the approaches presented by Davani et al. (2022), who built their proposal upon different aggregation strategies: (1) *ensemble*, where a model is learned for each annotator and models predictions are aggregated at the end; (2) *multi-label*, in which all possible labels are processed by a single model, effectively converting the problem into multi-label classification; and (3) *multi-task*, where the labels of each annotator are considered as an independent classification task. In all cases, the last step is based on a majority vote, which resembles the traditional practice of adjudicating a gold label. We show an overview of the baselines and our proposed methodology, which we will explain in more detail in the following section, in Figure 1.

[1] https://pdai.info/.

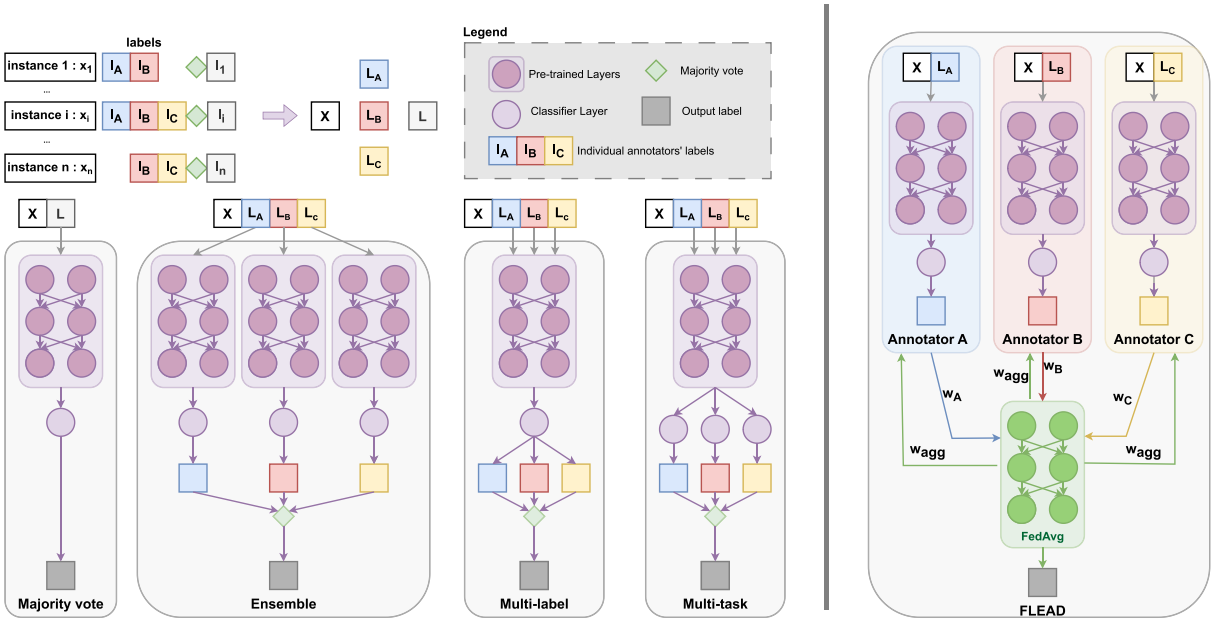


Figure 1: Overview of the text classification baselines on the left. Figure inspired by Davani et al. (2022). On the right, illustration of the FLEAD methodology with three clients and a central server. For simplicity, in the figure we assume three annotators: A, B, and C. Each annotator is shown in a different color and each data instance x_i refers to a classification training example and its corresponding annotator labels l_i. Finally, each client is depicted by a language model.

3 Methodology

The FLEAD methodology is built upon FL with the objective of learning from the disagreement among annotators, building a global model that integrates the perspectives of each of them. While the FL-based methodology is flexible to be applied to other tasks, in this paper our goal is to build a text classification model. In the following we formally define FL (Section 3.1) and present the details of the FLEAD methodology (Section 3.2).

3.1 Federated Learning

Federated learning is a distributed learning paradigm that preserves data privacy by orchestrating the independent training of learning models in data silos and the iterative aggregation of those local models in a global model (Kairouz et al., 2021). FL also stands out from better fit and handle heterogeneous or non-iid data distributions (McMahan et al., 2017), as long as the data distribution is not highly skewed (Zhao et al., 2022). The annotation of items by several annotators resembles the setting of FL, where each annotator matches with a federated client and the disagreement in the annotation matches a soft non-iid distribution focused on the label distribution. Likewise, FL builds a global model from the local models, which implies a real and independent integration of the evaluations of each annotator that allows to exploit the disagreement information.

More formally, given a set of clients or data owners $\{C_1, \ldots, C_n\}$ with their respective local training data $\{D_1, \ldots, D_n\}$. Each of these clients C_i, which implies the learning of n local learning models $\{L_1, \ldots, L_n\}$. FL aims at learning a global learning model G, using the scattered data across clients through an iterative learning process known as a *round of learning*. For that purpose, in each learning round t, each client trains its local model over their local training data D_i^t, resulting in the update of the local parameters L_i^t to $\hat{L}_i^t$. Thereafter, the global parameters G^t are computed by aggregating the trained local parameters $\{\hat{L}_1^t, \ldots, \hat{L}_n^t\}$ using a fixed federated aggregation operator Δ, and the local learning models are updated with the aggregated parameters:

$$\begin{aligned} G^t &= \Delta(\hat{L}_1^t, \hat{L}_2^t, \ldots, \hat{L}_n^t) \\ L_i^{t+1} &\leftarrow G^t, \quad \forall i \in \{1, \ldots, n\}. \end{aligned} \tag{1}$$

Updates among the clients and the server are repeated for the learning process until a given stop criteria is met. Thus, the final value of G will sum up the knowledge modeled in the clients.

FL for Text Classification The annotation of a text classification dataset by several annotators can be formally defined as: Given D as the entire dataset to annotate, $A_i \in \{A_1, A_2, \ldots, A_n\}$ as the annotator i out of n annotators and (D_i, L_i) as the set of k_i instances labeled by the annotator A_i, where $D_i = \{d_1, d_2, \ldots, d_{k_i}\}$ represent the instances of the dataset labeled by annotator A_i and $L_i = \{l_1, l_2, \ldots, l_{k_i}\}$ their corresponding labels. Finally, we denote $l^j = \{l^j_{i_1}, l^j_{i_2}, \ldots l^j_{i_m}\}$ as the m labels assigned to the instance j of D where each l^j_i represent the label assigned by the annotator i to this instance j.

3.2 FLEAD Methodology

We propose the Federated Learning for Exploiting Annotators' Disagreement (FLEAD) methodology, which is grounded in FL for learning from disagreement among annotators. It is based on the use of all the evaluations of the annotators by the training of a global model upon the aggregation of as many learning models as annotators. Broadly speaking, it consists of matching each annotator A_i with a federated client and independently training their data D_i in a local learning model LM_i. Formally, we define the t-th round of learning as

$$\begin{aligned} &LM_i^t \leftarrow \textit{train}(LM_i^{t-1}; (D_i, L_i)), \qquad \forall i \in \{1, 2, \ldots, n\} \\ &LM^t \leftarrow \textit{aggregation}(LM_i^t)_{i \in \{1,2,\ldots,n\}} \\ &LM_i^t \leftarrow LM^t, \qquad \forall i \in \{1, 2, \ldots, n\} \end{aligned} \tag{2}$$

where the training is conducted on the clients and the aggregation on the server. As aggregation we rely on FedAvg (McMahan et al., 2017), since its prominence in the literature (Zhao et al., 2022). FedAvg consists of the average of all parameters as follows:

$$\text{FedAvg}(LM_i^t)_{i=\{1,2,\ldots n\}} = \frac{\sum_{i=1}^{n} LM_i^t}{n}. \tag{3}$$

The result of the aggregation is a global learning model (G) that summarizes the partial information from the local learning models (LM). In this case, the global model integrates the perspectives of all annotators, instead of relying on other aggregation approaches like majority voting as other works do (Davani et al., 2022).

In Figure 1 we depict the FLEAD methodology. The figure shows an example in the case of three annotators per dataset instance. In blue, red, and yellow we represent the local training of each of the models of each client, which matches with each annotator. After the local training, the clients share their model weights w_i with the server, who aggregates the weights resulting in w_{agg}, which is shared with the clients to be the start point in the next round of learning.

Beyond the distributed aggregation provided by FL, it also lends: (1) *data-privacy*: since the data never leaves the local devices; (2) *robustness*: since it converges to a common solution based on the clients' partial solutions; and (3) *leverage all information*: since all annotator evaluations are used to train the local models. Divergent opinions thus come into play in training, not being disregarded during the adjudication of a gold label. Moreover, this matching of each annotator with a federated client is more effective than other assignment strategies due to it takes full advantage of all available information (see Section 6.2).

4 Experimental Framework

In this section we present the text classification experimental setup to test our FLEAD methodology.

4.1 Data

While there exist many publicly available text classification datasets, the availability of NLP datasets with the annotations of all the annotators is scarce, especially in languages other than English. In Section 4.1.1, we present a compilation of existing datasets which include individual annotator's information that we use for evaluation.

In addition to these existing datasets, we create our own dataset (*SentiMP* henceforth, Section 4.1.2) with the aim of including a diverse set of languages and controlling all the stages of the annotation process under consistent conditions, which can provide insights into the strengths and limitations of the FLEAD methodology. In particular, the SentiMP dataset enables us to better understand the relation between annotator agreement and performance (Section 6.2), and to perform a targeted error analysis (Section 6.4).

4.1.1 Datasets from the Literature

Since the main purpose of the paper is dealing with disagreement, we focus on subjective tasks as they represent a major challenge to annotators resulting in higher disagreement. In particular, we use the following datasets:

EmoEvent (Plaza del Arco et al., 2020) It is a multilingual (English and Spanish) collection of tweets about different events. The tweets are

annotated according to the six Ekman's basic emotions plus the ''neutral or other emotions'' label (EmoEvent-emotion) and as offensive or not offensive (EmotionEvent-offensive). We use the following splits:

1. *EmoEvent multiple*: multi-class classification task of the six Ekman's basic emotions.
2. *EmoEvent binary*: binary classification task of deciding whether a tweet is neutral or shows other emotions.
3. *EmoEvent offensive*: binary classification of tweets as offensive or not offensive.

TASS18 - GoodOrBad (Martínez-Cámara et al., 2018) The dataset is aimed at modeling the task of automatically selecting an adequate news for posting ads in online newspapers. Hence, it annotates the positive (SAFE) or negative (UNSAFE) emotion that a news article raises in a reader, which can be viewed as a type of stance classification. The dataset is composed of several online newspaper articles written in different Spanish language varieties used in diverse countries (Spain, Cuba, U.S.A., among others).

GabHate (Kennedy et al., 2022) It consists of English posts from gam.com designed for identifying ''hate-based rhetoric''. All the items are annotated by at least three annotators from a team of 18 annotators who participated in the annotation.

ConvAbuse (Cercas Curry et al., 2021) It is the first corpus on abusive language towards three conversational AI systems. It is annotated by multiple annotators but each of them only labels a sample of the dataset. The main challenge of this dataset is its marked unbalancedness across labels.

4.1.2 SentiMP Dataset

Given the lack of sentiment analysis datasets with individual annotations, we decided to construct SentiMP. The SentiMP dataset is a multilingual sentiment analysis dataset on the politics domain. Due to its controversial nature, this domain leads to divergent interpretations among the annotators. In the context of politics, social media, and, in particular Twitter, is the main place where politicians, and specifically members of parliament (MPs), communicate with their voters and general citizens. Hence, Twitter can act as a thermometer of the politicians' sentiment with respect to a specific period or topic. Indeed, these tweets are generally covered by mass media and arguably represent the main means of communication nowadays both from the government and opposition parties.

Data Collection The SentiMP dataset contains tweets written by members of parliament in Greece, Spain, and United Kingdom in 2021. We collected 500 tweets per language using the tweet collection provided by Antypas et al. (2022). For each country, tweets from the data collected were randomly sampled and annotated based on their sentiment. All tweets were anonymized by removing non-verified user information and removing URLs.[2]

Annotation We follow a three-level opinion meaning annotation schema, adding an *indeterminate* label for those tweets whose sentiment meaning is not evident, ambivalent or lacking context. Specifically, the annotation labels are:

- *Positive (1)*: Tweets which express happiness, praise a person, group, country or a product, or applaud something.
- *Negative (−1)*: Tweets which attack a person, group, product or country, express disgust, criticism or unhappiness towards something.
- *Neutral (0)*: Tweets which state facts, give news or are advertisements. In general those which do not fall into the above 2 categories.

Each set of tweets was annotated by a group of native speakers, namely, five annotators for the Spanish subset, and three for the English and Greek sets. The annotators were a mix of university students, faculty, and professionals with gender parity and enough knowledge on politics to conduct the annotation. The annotators were advised to consider only information available in the text, e.g., to not follow links, and in cases where a tweet includes only news titles or similar, to assess the sentiment of the item being shared. Tweets annotated as *indeterminate (X)* by one annotator were discarded in the experimental evaluation.

[2]We release the three language sets of the SentiMP dataset. The English set is available at https://huggingface.co/datasets/rbnuria/SentiMP-En, the Spanish one at https://huggingface.co/datasets/rbnuria/SentiMP-Sp and the Greek set at https://huggingface.co/datasets/rbnuria/SentiMP-Gr.

	Ann. 1	Ann. 2	Ann. 3	Gold
''In 1984 only 12% of engineering students were women. Nearly forty years later, the dial has barely moved to 14%.''	NEG	NEU	POS	**NEG**
''The news channels available to MPs and their staff include propaganda like RT_com. But as of yet, we can't get the new entrant in proper broadcasting GBNEWS. I am sure this will be put right soon''	POS	NEG	NEU	**NEG**
''Lots of constituents signed the important petitions being debated in Parliament on #Israel #Palestine, but in an oversubscribed debate, frustratingly I've not been selected. Rest assured, I'll continue to speak out whenever possible on the human rights abuses and violence there''.	NEU	POS	NEG	**NEG**

Table 1: Example of SentiMP-En instances with ties between annotators. The table shows the labels given by each annotator (Ann. 1, Ann. 2, and Ann. 3), and the final label decided in a joint discussion (Gold).

	# Tweets					avg. length				MTLD			
	NEG	NEU	POS	DIS	ALL	NEG	NEU	POS	ALL	NEG	NEU	POS	ALL
Spanish	206	75	137	82	500	39.83	31.04	37.15	37.37	99.6	87.7	101.62	98.44
English	129	98	244	29	500	43.95	34.61	37.14	38.51	153.77	139.92	163.22	155.18
Greek	213	129	149	9	500	37.43	29.71	39.48	36.02	185.58	141.05	162.01	176.19

Table 2: Statistics of the SentiMP Dataset: Number of tweets, average length, and linguistic diversity (in terms of MTLD) of negative (NEG), neutral (NEU), positive (POS), and discarded (DIS) tweets.

To break the ties, all the annotators met and discussed their positions, arriving at a common decision. The number of such tie-break cases was relatively small, with 8 cases for the Greek subset, 14 for the Spanish, and 9 for UK. Note that the discussion is only performed for ties in order to decide a single gold standard label. We show some examples of ties and the final gold label in Table 1.

SentiMP Statistics We show the corpus statistics in terms of number and lengths of tweets, and linguistic diversity by means of the measure of textual lexical diversity (McCarthy, 2005, MTLD), of each of the classes among the different datasets in Table 2. As can be observed, the distribution of tweets differ across languages, with the English subset being the most unbalanced in terms of polarity (244 positive and 129 negative tweets). We also computed the percentage of tweets in which there is at least one annotator who labels it as positive and another as negative (i.e., those with opposite annotations), and the percentage in the datasets of the three languages is among 5 and 10 percent (9.6 for Spanish, 6.8 for English, and 6.2 for Greek). In contrast to the datasets described in the previous section, we do not pre-define a train/test split for SentiMP. Instead, we run our experiments based on five-fold cross-validation, which is a more statistical robust evaluation method.[3]

4.1.3 Data Statistics

We present a summary of the data statistics of all datasets in Table 3. According to the strength of Cohen's Kappa agreement, most datasets present a moderated or fair agreement (Landis and Koch, 1977). In the case of the Greek and English subsets of the SentiMP dataset, this agreement is substantial. This difference in terms of annotator agreement represents the diversity of the experimental setup, which makes the evaluation more complete in terms of conclusions drawn with respect to varying levels of agreement.

4.2 Baselines

We compare the FLEAD methodology with several baselines. First, we include the *majority vote* baseline, where the gold label is adjudicated by a majority vote, and the language model is fine-tuned over those labels. This is the main comparison baseline, which is the most common approach used in the literature. In addition to this aggregated baseline, we also compare with three approaches for learning from disagreement proposed by Davani et al. (2022).

[3]For the sake of reproducibility, we have made available the 5 folds used in our experiments in the SentiMP website.

Dataset	Language	Train	Test	Val	Total	#Labels	#Annotators	Task	SM?	C. Kappa
SentiMP	**Spanish**	418	–	–	418	3	5 (5)	Sentiment analysis	yes	54.48
	English	471	–	–	471	3	3 (3)			64.94
	Greek	491	–	–	491	3	3 (3)			70.03
EmoEvent-emotion	**Spanish**	5723	1656	844	8223	7	3 (3)	Emotion classification	yes	38.36
	English	5112	1447	744	8049	7	3 (3)			27.16
EmoEvent-offensive	**Spanish**	5723	1656	844	8223	2	3 (3)	Offensive language identification	yes	54.67
	English	5112	1447	744	8049	2	3 (3)			25.79
Tass18 - GoodOrBad	**Spanish**	1250	500	250	2000	2	2 (2)	Stance classification	no	59.00
GabHate	**English**	22124	5531	–	27655	2	18 (3.13)	Hate speech detection	yes	28.00
ConvAbuse	**English**	4785	1026	1026	6837	5	8 (3.24)	Nuanced abuse detection	no	46.92

Table 3: Statistics of the datasets used in the evaluation. From left to right we include: (1) the language of the dataset (Language); (2) the amount of instances of train (Train), test (Test), validation (Val), and the total amount (Total); (3) the number of labels (#Labels); (4) the number of annotators (#Annotators) specifying the total amount of annotators which participate in the annotation process and, between parentheses the average of annotations per instance; (5) the task addressed (task); (6) whether the text is from social media or not (SM?); and (7) the inter-annotator agreement according to Cohen's Kappa (C. Kappa).

- *Ensemble*: The partitions are created with the instances labeled by each annotator with their respective labels. Then, we train a language model over each of these subsets. After that, we use the ensemble of these models by performing a majority vote over the classes predicted by the models.

- *Multi-label*: All labels for each instance are considered in a multi-label classification model where each label denotes individual annotators' labels. The model first adds a fully connected layer to get a vector of the dimension of the number of annotators, and then apply a sigmoid function.

- *Multi-task*: It considers the labels of each annotator as independent classification tasks, all sharing encoder layers to generate the same representation of the input sentence, each with its separate fully connected layer and softmax activation. Compared with the multi-label baseline, the multi-task approach includes a fully connected layer explicitly fine-tuned for each annotator.

An overview of these baselines and our proposal can be found in Figure 1. In contrast to our approach, all these baselines require a majority vote layer. We re-implemented the ensemble, multi-label, and multi-task baselines based on the configuration expressed in Davani et al. (2022).[4]

[4]Training details in Section 4.4.

Finally, we also compare with a *majority class* naive baseline, which does not rely on any model training.

4.3 Language Models

While the FLEAD methodology could be applied to any supervised model, in this paper we focus on transformer-based language models given their state-of-the-art performance in NLP tasks (Wolf et al., 2020). For practical reasons and due to computational limitations, we decided to perform our main experiments with base-size models (see Section 6.3 for an analysis using models of different size). We use the following language models:

Multilingual Language Model (XLM) Depending on whether the dataset is based on social media texts or not, we use two different multilingual models: (1) for *Social media XLM*, we use the *cardiffnlp/twitter-xlm-roberta-base* model (Barbieri et al., 2022), a XLM-roberta-base model trained on tweets; (2) for *No social media XLM*, we use the *xlm-roberta-base* model (Conneau et al., 2019).

Monolingual Language Model (MLM) We also carry out experiments using language models trained on the target language. We use different language models depending on whether the dataset is from social media: (1) for *English datasets*, we use the *cardiffnlp/twitter-roberta-base* (Barbieri et al., 2020) model for the social media datasets and *roberta-base* (Liu et al., 2019) for the

	Epochs	$\text{Epochs}_{\text{FLEAD}}$	$\text{Rounds}_{\text{FLEAD}}$	LR	Batch
SentiMP En	250	25	10	$5e^{-5}$	16
SentiMP Sp	250	25	10	$5e^{-5}$	16
SentiMP Gr	250	25	10	$5e^{-5}$	16
EE-Sp off.	300	20	15	$5e^{-5}$	32
EE-Sp bin.	300	20	15	$5e^{-5}$	64
EE-Sp mul.	300	20	15	$5e^{-5}$	64
EE-En off.	300	20	15	$5e^{-5}$	32
EE-En bin.	300	20	15	$5e^{-5}$	64
EE-En mul.	300	20	15	$5e^{-5}$	64
GabHate	200	10	20	$5e^{-4}$	32
ConvAbuse	200	10	20	$5e^{-4}$	64
TASS18	200	20	10	$5e^{-6}$	32

Table 4: Training hyperparameters of the FLEAD methodology and baselines.

others; (2) for *Spanish datasets*, we use the *daveni/twitter-xlm-roberta- emotion-es* model (Vera et al., 2021) for social media datasets and *dccuchile/bert-base-spanish-wwm-cased* (Cañete et al., 2020) for the rest; and (3) for the *Greek dataset*, we utilize the *gealexandri/palobert-base-greekuncased-v1* model (Alexandridis et al., 2021).

4.4 Training Details

Table 4 shows the configuration of each learning model to ease the reproducibility of the experimental setup, and, in particular: epochs in the baselines, epochs and learning rounds (FL Epochs and Rounds, respectively) in the experiments following the FLEAD methodology, and learning rate (LR) and batch size, which are common to all the experiments. The hyperparameters utilized were standard for each task, and the slight variations were decided on a small validation task from each training set for the majority base baseline, and kept for all models. We use as stop criteria a fixed amount of epochs and learning rounds. Notice that, for a fair comparison between FL experiments and the baselines $Epochs = Epochs_{FLEAD} \times Rounds_{FLEAD}$. This way, we make the total number of rounds of learning during which all models are trained the same number of epochs, both in the FLEAD and baselines experiments. Each model is trained five times and the final results are averaged across the five different runs.

5 Experimental Results

In this section we present the results with the aim of evaluating our FLEAD methodology in a multi-annotation context with both the standard single label evaluation protocol and an additional setting in which disagreement between annotators is taken into account in the evaluation. We use standard text classification evaluation metrics (Accuracy and Macro-F1) in Section 5.1, and metrics specifically designed for disagreement between annotators in Section 5.2.

5.1 Majority-based Single Label Evaluation

In order to compare with mainstream approaches that do not model disagreement, we perform an evaluation using standard Accuracy and Macro-F1 metrics. Since both Accuracy and Macro-F1 metrics require a single gold-standard label on which to evaluate the models, we follow the methodology widely used in the literature, which consists of deciding this label by majority vote among the annotators' labels.

Table 5 shows that the FLEAD methodology outperforms all the baselines according to Macro-F1. In contrast, the FLEAD methodology does not return the highest result according to Accuracy (see top part of Table 5) on unbalanced datasets, which are widely known to be skewed toward the majority class. Indeed, the best performing baseline in those cases is the *majority class*. If we compare the monolingual and multilingual language models, performance, we find that the results are very similar, with a slight superiority of the multilingual language models.

The results of the FLEAD methodology on the Spanish and Greek datasets are similar to the ones reached on the English datasets, as Table 6 shows. FLEAD is only slightly outperformed by the multilabel baseline according to Accuracy. Regarding the comparison between monolingual and multilingual language models, the multilingual language models achieve slightly superior results. This difference with the results in Table 5 may be due to the small number of high-quality language models in languages other than English.

5.2 Class Probabilities as Gold Label Evaluation

For the evaluation in the previous section, we used standard evaluation metrics that relied on a single test label. This has the shortcoming of depending strongly on such a gold test label, without taking into consideration the disagreement information among annotators. For example, an instance labeled with $\{1, 1, 0\}$ is assigned with the final label 1, similarly to an instance labeled with

		SentiMP		EmoEvent off.		EmoEvent bin.		EmoEvent mul.		GabHate		ConvAbuse	
		XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM
Accuracy	**Maj. class**	51.8	51.8	**92.9**	**92.9**	54.7	54.7	45.2	45.2	**90.7**	**90.7**	**86.4**	86.4
	Maj. vote	82.0	75.1	92.8	92.8	61.6	62.1	58.4	54.5	90.2	90.2	85.4	83.4
	Ensemble	81.9	76.3	83.0	88.4	63.9	62.9	57.6	57.3	88.2	87.7	**86.4**	82.9
	Multilabel	83.0	72.6	82.2	89.1	65.1	64.2	59.1	58.8	88.7	87.5	83.4	**83.6**
	Multitask	84.1	77.1	82.8	93.2	72.7	72.6	59.3	59.1	89.7	88.8	82.3	83.3
	FLEAD	**87.1**	**84.5**	84.0	89.5	**79.1**	**78.4**	**60.3**	**59.4**	89.9	89.3	81.4	82.1
Macro-F1	**Maj. class**	22.7	22.7	48.1	48.1	35.3	35.3	8.9	8.9	46.2	46.2	46.3	46.3
	Maj. vote	78.2	68.4	48.1	48.1	57.6	58.3	38.9	38.1	47.4	47.7	56.3	48.3
	Ensemble	77.8	70.1	52.9	59.6	59.1	59.2	39.6	38.9	68.5	68.0	66.3	59.3
	Multilabel	79.2	73.1	53.9	62.8	58.9	59.1	40.2	39.8	68.9	67.1	69.0	68.8
	Multitask	83.3	75.6	56.4	65.4	59.3	59.4	40.6	40.3	71.9	69.5	70.9	69.5
	FLEAD	**86.8**	**79.0**	**68.4**	**66.3**	**60.5**	**60.8**	**41.2**	**40.9**	**72.1**	**71.6**	**72.8**	**71.9**

Table 5: Results on the English datasets according to Accuracy and Macro-F1. We use the XLM and MLM language models for each dataset. Best results for each model are highlighted in bold.

		SentiMP Sp		SentiMP Gr		EmoEvent off.		EmoEvent bin.		EmoEvent mul.		TASS18	
		XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM	XLM	MLM
Accuracy	**Maj. class**	49.3	49.2	43.3	43.3	91.6	91.6	52.6	52.6	47.3	47.3	60.8	60.8
	Maj. vote	77.0	71.5	74.7	71.5	91.9	91.2	70.4	82.2	60.5	58.1	79.0	80.1
	Ensemble	76.3	72.2	73.9	72.7	92.5	91.5	68.3	75.6	59.5	57.6	80.3	78.3
	Multilabel	**84.2**	75.1	79.0	74.6	89.0	90.0	66.7	72.3	60.9	59.4	82.4	77.1
	Multitask	83.4	75.9	80.3	79.1	92.6	91.7	67.6	70.5	61.2	60.3	82.4	79.3
	FLEAD	84.1	**82.8**	**88.7**	**79.9**	**94.8**	**92.3**	**78.9**	**80.1**	**63.4**	**62.7**	**94.1**	**90.1**
Macro-F1	**Maj. class**	22.0	22.0	20.1	20.1	47.8	47.8	34.4	34.4	9.1	9.1	37.8	37.8
	Maj. vote	71.6	64.4	75.7	68.5	47.9	47.7	70.3	65.4	44.1	40.5	79.0	78.4
	Ensemble	72.3	67.8	77.1	68.9	69.3	47.7	63.4	60.5	46.2	42.5	80.9	76.5
	Multilabel	78.8	68.3	78.0	72.1	70.1	69.9	62.2	61.9	47.8	44.2	83.2	75.9
	Multitask	78.5	70.4	78.9	76.9	69.2	68.9	64.2	65.3	52.1	50.9	83.4	77.1
	FLEAD	**85.4**	**77.1**	**86.8**	**77.4**	**74.8**	**73.9**	**72.7**	**70.8**	**55.2**	**53.1**	**85.0**	**80.4**

Table 6: Results on the Spanish and Greek datasets according to Accuracy and Macro-F1. We use the XLM and MLM language models for each dataset. Best results for each model are highlighted in bold.

$\{1, 1, 1\}$. In this case, if the classifier model labels both instances with 0, it is a mistake in both cases according to the metrics used. However, the *error* is arguably less pronounced in the first instance than in the second one.

In this section, we replace gold labels by the probability distribution of each label according to the annotation of each item (Baan et al., 2022). For instance, if we consider the labels $\{1, 1, 0\}$, the vector of probabilities over the three possible labels $\{-1, 0, 1\}$ would be $\{0, 0.33, 0.67\}$. We use the following metric *DistCE* proposed in Baan et al. (2022):

$$\text{DistCE}(x) = \text{TVD}(f(x), \pi(x)) \quad (4)$$

where $f(x)$ is the vector of class probabilities, $\pi(x)$ the probabilities predicted by the classifier and $\text{TVD}(y, z) = (||y - z||_1)^{1/2}$. In essence, this metric measures how close the probability distribution returned by the model is to the probability distribution over the labels of all annotators, so the closer to zero the better.

Table 7 shows the results of the comparison between the standard classification models based on a single gold label for training (i.e., the *majority vote* baseline) and the FLEAD methodology in terms of the DistCE metric (see Equation 4). The results highlight that the label probabilities returned by the FLEAD methodology are more similar to the objective annotation distribution

	Maj. vote	FLEAD
SentiMP En	0.824	**0.383**
SentiMP Sp	1.127	**0.466**
SentiMP Gr	1.051	**0.468**
EmoEvent En Off	0.486	**0.285**
EmoEvent En Bin	0.671	**0.325**
EmoEvent En Mul	0.671	**0.367**
EmoEvent Sp Off	0.501	**0.291**
EmoEvent Sp Bin	0.568	**0.317**
EmoEvent Sp Mul	0.849	**0.567**
TASS 18	0.530	**0.450**
ConvAbuse	1.305	**0.761**

Table 7: DistCE results of the baseline Majority Vote and the FLEAD methodology.

than the probabilities returned by the *majority vote* baseline. This implies that the FLEAD methodology fits better to the annotators behavior and, as we will see in Section 6.4, the errors are more easily explainable by the subjectivity of the task.

6 Analysis

In this section, we analyze the source of the FLEAD methodology improvement through ablation studies in Section 6.1, the effect of the disagreement between annotators in the gain of performance of each approach in Section 6.2, and the effect of the learning model used in Section 6.3. We also perform a qualitative study in Section 6.4 with human annotators that supports the operation of the FLEAD methodology.

6.1 Ablation Study

In order to analyze in which parts of the FLEAD methodology the performance improvements lie, we devise simple baselines to compare with.

Random Clients The FLEAD methodology matches each annotator with a client in the federated scheme. We test if this match is really essential, or if distributing the different annotations among different models is enough. For that, we design the baseline *FL-random*, in which we simulate a federated scenario with as many clients as annotators in the dataset. However, we distribute the labels of the annotators randomly among the clients instead of matching them with each client. We find (see the rows *FL-random* in Table 8) that the random distribution of the labels among clients improves the results of the baseline *majority vote* but not the results of the FLEAD methodology, highlighting the value of matching each annotator with a client.

Multitask + Aggregation In general, the combination of different models provides more robust results (Rokach, 2005). We aim to test if the improvements are simply due to this aggregation, and not to the matching of each annotator to a federated client and the FL operation. Thus, we design the baseline *multi-agg*, which combines the best baseline (multitask) with an aggregation every few rounds of learning, similarly to what is done in the FLEAD methodology. It consists of training as many multitask models (as described in Section 4.2) as annotators and aggregating the weights of the models into a single one. This process is repeated the same number of times as in the FLEAD methodology (see Section 3 for more details). The row *Multi-agg* in Table 8 shows better results than the Multitask baseline, which confirms that the aggregation of different models produces better and more robust results in general. However, this is not the only factor leading to the FLEAD performance gains, as FLEAD still reaches a higher performance in all the datasets. Hence, the performance improvements are not only due to the aggregation conducted every few epochs but also to the FL operation.

Best Single Annotator The annotation of subjective tasks may evidence that some annotators are more accurate than others, i.e., their evaluations lead to more accurate learning models. In this analysis we train different models for each annotator and report the best one. We refer to this baseline as *best annotator*.

In Table 8 we show that the results are significantly worse than the baselines in which all the available information is used, thus confirming our claim that leveraging all the labels of the annotators improves results.

6.2 Performance in Terms of Agreement

We explore the effect of the annotators in the gain of performance of each approach with respect to the *majority vote* baseline in terms of the agreement. For that purpose, we formally define the relative performance gain (*rpg*) of each model following the expression:

$$\text{rpg(model)} = \frac{MacroF1(\text{model})}{MacroF1(\text{majority vote})} \quad (5)$$

		S.MP En	S.MP Sp	S.MP Gr	E.En.Off	E.En.bin	E.En.mul	E.Sp.Off	E.Sp.bin	E.Sp.mul	TASS18	G.Hate	C.Abuse
Accuracy	**Maj. vote**	82.0	77.0	74.7	**92.8**	61.6	58.4	91.9	70.4	60.5	79.0	**90.2**	85.4
	Multitask	84.1	83.4	80.3	82.8	72.7	59.3	92.6	67.6	61.2	82.4	89.7	82.3
	FLEAD	**87.1**	**84.1**	**88.7**	84.0	**79.1**	**60.3**	94.8	**78.9**	**63.4**	**94.1**	81.4	81.4
	FL-random	85.3	82.7	85.5	85.5	78.9	60.1	**95.5**	78.8	62.1	93.9	88.2	81.8
	Multi-agg	84.3	83.5	82.2	83.3	74.5	59.2	92.5	73.4	63.3	87.9	89.9	83.3
	Best annotator	85.1	80.2	85.2	90.2	58.4	59.2	93.8	67.2	60.9	91.3	80.1	**85.7**
Macro-F1	**Maj. vote**	78.2	71.6	75.7	48.1	57.6	38.9	47.9	70.3	44.1	79.0	47.4	56.3
	Multitask	83.3	78.5	78.9	56.4	59.3	40.6	69.2	64.2	52.1	83.4	71.9	70.9
	FLEAD	**86.8**	**85.4**	**86.8**	**68.4**	**60.5**	**41.2**	**74.8**	**72.7**	**55.2**	**85.0**	**72.8**	**72.8**
	FL-random	83.5	83.2	84.1	65.2	58.8	40.5	71.3	69.4	55.1	84.4	70.9	71.1
	Multi-agg	83.7	78.8	80.1	62.1	60.2	40.9	72.5	67.8	53.4	84.5	71.9	71.5
	Best annotator	81.2	80.1	80.9	61.3	53.1	38.7	70.5	65.6	54.3	82.9	45.6	55.2

Table 8: Results of the ablation study according to Accuracy and Macro-F1. We show the results of the FLEAD methodology, the majority vote baseline, and all the new baselines proposed in the ablation study. We only use the XLM model for each dataset. We highlight in bold the best results.

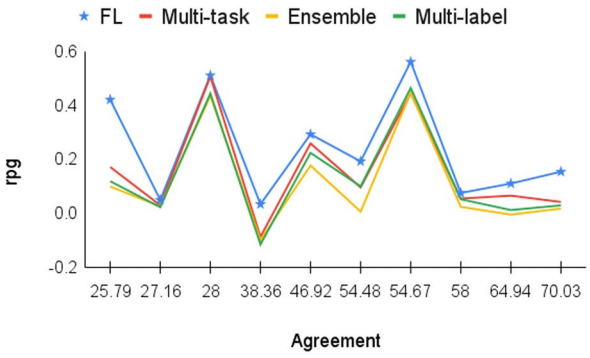


Figure 2: Relative performance gain with respect to agreement (Kappa) in all datasets.

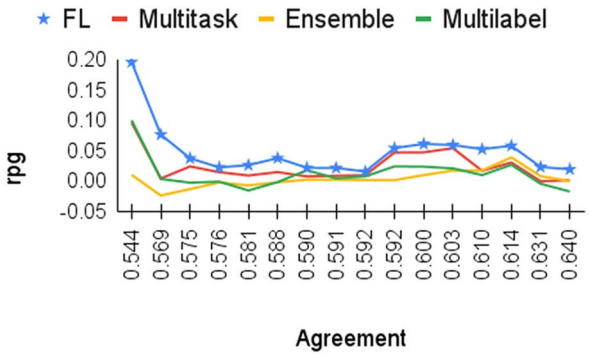


Figure 3: Relative performance gain with respect to agreement (Cohen's Kappa) in SentiMP Sp.

Our goal is to analyze whether there is any correlation between the gain of performance and the decrease of agreement (in terms of Cohen's Kappa). In Figure 2 we sort all the datasets according to the agreement in an increasing order and show the relative performance gain. We find that there is no tendency in the graph, so there appears to be no direct relationship between inter-annotator agreement and approach's performance. However, in this analysis we are comparing different tasks. Accordingly, we chose the SentiMP Spanish dataset, which is the only dataset where all samples are labeled by more than 3 annotators. In order to have different datasets with different agreements between annotators, we create all subsets of the SentiMP Spanish dataset considering three and four annotators. Note that in these newly generated datasets only the gold labels change, not the task. In Figure 3 we show the results of this analysis, sorting the new datasets according to agreement.

Although there is no solid trend, we observe how the relative gain of the FLEAD methodology is higher when agreement is lower (agreement $<$ 0.58). Hence the FLEAD methodology appears to be more useful when there is less agreement among the annotators.

6.3 Language Model Size

In principle, FLEAD can be used with any learning model, and in particular with any learning model size if computational resources are available. Accordingly, we evaluate the FLEAD methodology with language models of different sizes. We evaluate our methodology with Roberta multilingual large (Liu et al., 2019), base (Liu et al., 2019), and distill (Sanh et al., 2019) models in the English datasets.

Table 9 shows the results of all models using Majority vote and FLEAD. FL-Large models reach the highest results in all configurations that

		SentiMP		EE. off.		EE. bin.		EE. mul.		GabHate		Convab.	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Maj. Vote	Large	87.1	86.5	83.0	53.3	76.5	59.4	55.7	39.3	90.2	57.8	86.5	47.5
	Base	82.1	78.3	92.8	48.1	65.3	58.5	55.2	38.2	90.3	47.6	83.4	48.3
	Distill	78.8	74.2	92.8	48.1	64.9	57.8	53.6	36.7	89.4	46.1	86.4	46.3
FLEAD	Large	87.3	87.1	89.5	67.8	78.2	61.1	60.3	42.1	–	–	–	–
	Base	87.2	86.8	89.6	67.3	77.9	60.7	59.5	40.8	89.3	71.5	82.1	71.9
	Distill	87.0	84.1	88.8	65.2	77.2	60.1	58.3	38.7	90.1	68.0	83.2	69.5

Table 9: Comparison among large, base, and distill models using Majority Vote (Maj. Vote) and FLEAD. FL-Large results for GabHate and ConvAbuse are missing due to the large number of annotators (18 and 8, respectively) and high computational requirements.

we could run because of computational resource restrictions (see Section 7). However, the differences between the FL-Large and FL-Base models are quite marginal compared to the difference in computational capacity, with the large model being approximately 3 times the base model, which is multiplied by the number of clients in the case of the federated model. In fact, if we compare the results of the Large model and the FL-Base model, whose computational requirements are similar, we notice that the FLEAD methodology used in FL-Base always achieves considerably better results. This shows that although the use of large models in FL can be a constraint depending on the number of clients, it is not such a strong issue because competitive results can be reached with smaller models. These results imply that the FLEAD methodology provides a more efficient learning process, as small language models can reach similar results to large language models.

6.4 Error Analysis

In this section we perform qualitative analyses on whether the errors made by the FLEAD methodology can be more understandable or explained than those made by the other baselines. To this end, in Section 6.4.1 we analyze how many of the original annotators agree with the output of each model, and in Section 6.4.2 we perform an additional analysis to understand whether new annotators agree with the system outputs.

6.4.1 How Many Annotators Agree?

We propose a simple metric applied to the classification errors to measure how many the annotators agree with the output of the models. We refer to this metric as *any_annotator* and it is defined by equation 6:

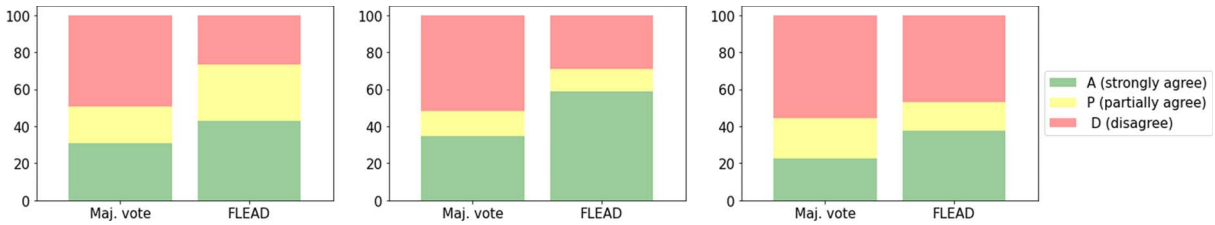
$$any\_annotator = \frac{\#annotator\_agrees}{\#mistakes} \quad (6)$$

where $\#annotator\_agrees$ represents the amount of mistakes in which at least one annotator gave the label predicted by the classifier and $\#mistakes$ is the total amount of mistakes produced by the classifier. What this metric tries to measure is the degree to which any of the annotators agreed with the label provided by the model. In other words, even if the model fails to get the label chosen by a majority vote, it may still choose a reasonable label according to the original annotators.

In Table 10 we compare the *majority vote* baseline and the FLEAD methodology according to this metric in both the own mistakes of each proposal and the overlap of mistakes between

	All mistakes%		Overlap%	
	Maj. vote	FLEAD	Maj. vote	FLEAD
SentiMP En	28.9	**85.3**	34.4	**61.2**
SentiMP Sp	33.9	**89.5**	41.1	**72.3**
SentiMP Gr	29.8	**77.5**	39.2	**62.1**
EmoEvent En Off	49.8	**85.9**	52.4	52.4
EmoEvent En Bin	33.2	**72.1**	57.3	57.3
EmoEvent En Mul	27.1	**67.2**	29.9	**57.7**
EmoEvent Sp Off	52.3	**88.1**	58.2	58.2
EmoEvent Sp Bin	32.7	**66.2**	49.3	49.3
EmoEvent Sp Mul	24.1	**67.9**	35.4	**66.2**
TASS18	30.1	**65.8**	45.1	45.1
ConvAbuse	41.7	**88.1**	51.6	**79.3**

Table 10: Results in terms of *any_annotator* metric in the mistakes of each approach (columns 2 & 3) and in the overlap of mistakes (columns 4 & 5) in all datasets except for GabHate, whose annotator labels in the test partition are not available.



(a) SentiMP Spanish. (b) SentiMP English. (c) SentiMP Greek.

Figure 4: Comparison in terms of percentage of agreement of the two annotators in the re-annotation qualitative analysis among the Majority Vote baseline and the FLEAD methodology in the overlap of mistakes. We respectively represent in green, yellow, and red the three levels of agreement defined.

	Gold Label	FLEAD	Maj. Vote
''I thought you better than this, Adam. John is a kind, honest and gracious man who is respected across the political spectrum. Who were you hoping to engage with a tweet such as this? Really disappointing.''	NEG	NEU	POS
''Fire and rehire is immoral - the fightback starts now. I'll be showing my solidarity and support for the strikes and actions starting tomorrow!''	NEU	POS	NEG
''Boris Johnson's diabolical record in 2020 - please watch, share, organise.''	NEU	NEG	POS

Table 11: Examples of SentiMP-En test instances misclassified by both FLEAD and the Maj. Vote baseline. In the first one the external annotators disagree with the predicted labels by both models, and in the second and third examples both external annotators strongly agree with FLEAD.

both approaches. We find that the *any_annotator* accuracy percentage achieved by FLEAD is considerably higher than the percentage reached by the *majority vote* baseline. This indicates that a large portion of the instances mislabeled by our federated approach are reasonable mistakes given that a human agrees with that predicted label, even if it does not match the gold standard.

6.4.2 How Much Does a New Annotator Agree?

In this section we analyze how much a new human annotator agrees with the output of the different methodologies (majority vote and FLEAD). To this end, we designed an analysis that consists of providing new annotators with the misclassified instances with their predicted label. Then, each annotator is asked to consider whether this new level is correct or not. We perform this analysis in the SentiMP datasets, since we designed and know the annotation guidelines.

We define three levels of agreement: (1) **A (strongly agree)**: if the label set by the learning model is the one the re-annotator would assign, (2) **P (partially agree):** if this is not the label re-annotator would assign, but she partially agrees with it, and (3) **D (strongly disagree)**: the re-annotator disagrees with the label set by the model.

In Figure 4 we show the mean results of the two external annotators in all SentiMP datasets in the overlap of mistakes between the FLEAD methodology and the *majority vote* baseline, respectively. We see that in the three datasets the external annotators agree that the labels generated by the FLEAD methodology make more sense than the labels generated by the *majority vote* baseline, which may be mostly due to the subjectivity of the task and not due to models' error.

Table 11 shows some examples of misclassified SentiMP tweets in which both external annotators concur with respect to the predicted labels. In general, the labels predicted by FLEAD appear to be more reasonable by the annotators than those predicted by the *majority vote* baseline.

7 Limitations

We have identified five main limitations of our proposal and evaluation:

1. **Computational Resources.** Using our methodology for a large set of annotators is computationally very demanding, especially when it comes to memory requirements of

language models that increase linearly with respect to the number of annotators. For our FLEAD methodology, a separate model is required for each annotator. As we analyzed in Section 6.3, there may be a trade-off between model size and our methodology when the number of annotators is large (e.g., in crowdsourcing annotation schemes). It is likely that as the complexity of the task increases, larger models may be necessary to improve performance. Therefore, new efficient techniques to address this issue would need to be explored, especially when the number of annotators is large. For instance, there may be techniques to perform the federated aggregation individually for each model, alleviating memory issues.

2. **Aggregation Methods.** Most of our experiments are based on the majority vote to decide the gold label. However, as we argued throughout the paper, other aggregation techniques such as the one proposed by Baan et al. (2022) and analyzed in Section 5.2 should be more thoroughly analyzed.

3. **Federated Aggregation.** We use FedAvg for all the experiments because of its prominence in the literature and competitive performance (Zhao et al., 2022). However, it would be interesting for future work to analyze the influence of the aggregator in the methodology.

4. **Annotator Diversity and Type of Disagreement.** Each dataset has a different degree of annotator diversity. In the case of our newly constructed dataset, SentiMP, both original (see Section 4.1.2) and external annotators (see Section 6.4) of SentiMP have some similar demographic characteristics, which may affect some of the conclusions drawn from this dataset and the qualitative analysis. The agreement between annotators of SentiMP is among moderated and substantial (see Table 3), which is larger than other datasets. To mitigate this potential limitation, we have performed an additional analysis with respect to the impact of disagreement in Section 6.2. Moreover, in this paper we do not focus on the type of disagreement and this is modelled jointly by the federated model. It is possible that different types of disagreement affect the model differently, but this is not explored or explicitly analysed in this work.

5. **Task Variety.** We have only focused on text classification and do not explore other NLP tasks, partially due to the lack of datasets with individual annotations. We believe that our methodology is not specific to text classification and can be applied to other NLP tasks, or even other machine learning related applications where we can find similar disagreement issues (Albarqouni et al., 2016; Beyer et al., 2020; Cabitza et al., 2019, 2020), but we leave this extended analysis for future work.

8 Conclusions

In this paper, we proposed a text classification method to leverage the information from all annotators separately. To this end we put forward FLEAD, a methodology based on FL and considering each annotator that participate in the annotation of a dataset as a federated client. Thus, the labels of each annotator are independently learned and aggregated in a global or final model. In general, our methodology shows promising results and prove that FL can be used beyond protecting data privacy, in this case to learn from the disagreements among annotators in subjective tasks.

Finally, we performed an in-depth evaluation and analysis to understand the different component of our methodology, with the following conclusions: (1) The results on several multilingual datasets of subjective text classification tasks show that leveraging information from all the annotators is indeed beneficial and enhances the classification performance; (2) our ablation analysis highlights that the improvements are largely due to the FL operation, in addition to other side benefits that our methodology offers; (3) the qualitative analysis shows that the external annotators generally agree more with the errors made by our FLEAD-based model, in comparison to the models trained on a single ground truth.

Acknowledgments

This work was partly supported by the grants PID2020-119478GB-I00, PID2020-116118GA-I00, and TED2021-130145B-I00 funded by MCIN/AEI/10.13039/501100011033. Jose Camacho-Collados is supported by a UKRI Future Leaders Fellowship.

We acknowledge the support of Dimosthenis Antypas for obtaining the SentiMP data, and the additional help of Dimitra Mavridou, Mairi Antypas, David Owen, Matthew Redman, Gabriel Wong, Miguel López Campos, Daniel Jiménez López, and Alberto Argente Garrido in the SentiMP annotation.

References

Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. 2016. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321. https://doi.org/10.1109/TMI.2016.2528120, PubMed: 26891484

Georgios Alexandridis, Iraklis Varlamis, Konstantinos Korovesis, George Caridakis, and Panagiotis Tsantilas. 2021. A survey on sentiment analysis and opinion mining in Greek social media. *Information*, 12(8):331. https://doi.org/10.3390/info12080331

Dimosthenis Antypas, Alun Preece, and Jose Camacho Collados. 2022. Politics and virality in the time of twitter: A large-scale cross-party sentiment analysis in Greece, Spain and United Kingdom. *CoRR*, abs/2202.00396. https://doi.org/10.2139/ssrn.4166108

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.124

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of Findings of EMNLP*, pages 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

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.

Valerio Basile. 2021. It's the end of the gold standard as we know it. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. https://doi.org/10.1007/978-3-030-77091-4_26

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. https://doi.org/10.18653/v1/2021.bppf-1.3

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with imagenet? *CoRR*, abs/2006.07159.

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Medical Informatics and Decision Making*, 20(1):1–21. https://doi.org/10.1186/s12911-020-01224-9, PubMed: 32917183

Federico Cabitza, Angela Locoro, Camilla Alderighi, Raffaele Rasoini, Domenico Compagnone, and Pedro Berjano. 2019. The elephant in the record: On the multiplicity of data recording work. *Health Informatics Journal*, 25(3):475–490. https://doi.org/10.1177/1460458218824705, PubMed: 30666882

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Practical Machine Learning for Developing Countries at ICLR 2020*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7388–7403. https://doi.org/10.18653/v1/2021.emnlp-main.587

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.774

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116. https://doi.org/10.18653/v1/2020.acl-main.747

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. https://doi.org/10.1162/tacl_a_00449

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297. https://doi.org/10.18653/v1/D15-1035

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. https://doi.org/10.1162/tacl_a_00523

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecná, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and open problems in federated learning. *Foundations and Trends®in Machine Learning*, 14(1–2):1–210. https://doi.org/10.1561/9781680837896

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108. https://doi.org/10.1007/s10579-021-09569-x

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, pages 1886–1895. https://doi.org/10.18653/v1/N18-1171

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643. https://doi.org/10.1016/j.ipm.2021.102643

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174. https://doi.org/10.2307/2529310

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 The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.314

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.822

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

Eugenio Martínez-Cámara, Yudivián Almeida-Cruz, Manuel C. Díaz-Galiano, Suilan Estévez-Velarde, Miguel Á. García-Cumbreras, Manuel García-Vega, Yoan Gutiérrez, Arturo Montejo Ráez, Andrés Montoyo, Rafael Muñoz, Alejandro Piad-Morffis, and Villena-Román Julio. 2018. Overview of tass 2018: Opinions, health and emotions. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172, pages 13–27.

Philip M. McCarthy. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135. https://doi.org/10.1561/9781601981516

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions. *CoRR*, abs/2205.00415. https://doi.org/10.18653/v1/2023.eacl-main.130

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. https://doi.org/10.1162/tacl_a_00040

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625. https://doi.org/10.1109/ICCV.2019.00971

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751. https://doi.org/10.3115/v1/E14-1078

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511. https://doi.org/10.3115/v1/P14-2083

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin.

2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16. https://doi.org/10.3115/1611628.1611631

Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1161–1168. https://doi.org/10.1609/aaai.v32i1.11506

Lior Rokach. 2005. Ensemble methods for classifiers. In *Data Mining and Knowledge Discovery Handbook*, pages 957–980. https://doi.org/10.1007/0-387-25465-X_45

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1020

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023a. 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. https://doi.org/10.18653/v1/2023.eacl-main.178

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023b. 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 2420–2433. https://doi.org/10.18653/v1/2023.eacl-main.178

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sadat Shahriar and Thamar Solorio. 2023. SafeWebUH at SemEval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 94–100, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.12

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622. https://doi.org/10.1145/1401890.1401965

Michael Sullivan, Mohammed Yasin, and Cassandra L. Jacobs. 2023. University at Buffalo at SemEval-2023 task 11: MASDA–modelling annotator sensibilities through DisAggregation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.135

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *Journal Artificial Intelligence Research*, 72:1385–1470. https://doi.org/10.1613/jair.1.12752

D. Vera, O. Araque, and C. A. Iglesias. 2021. GSI-UPM at IberLEF2021: Emotion analysis of Spanish tweets by fine-tuning the XLM-RoBERTa language model. In *Proceedings of the Iberian Languages Evaluation Forum*.

Janyce Wiebe. 1990. Identifying subjective characters in narrative. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*. https://doi.org/10.3115/997939.998008

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf,

Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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. https://doi.org/10.1609/icwsm.v17i1.22198

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2022. Federated learning with non-iid data. *CoRR*, abs/1806.00582.