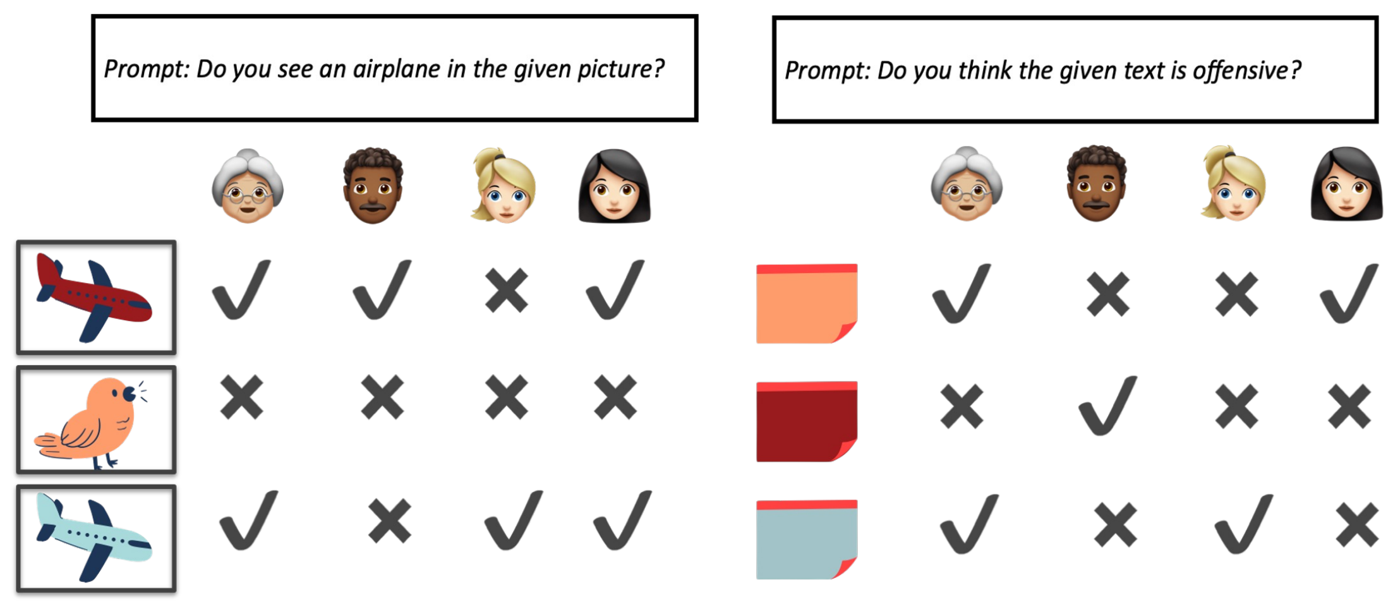
Understanding The Relation Between Noise And Bias In Annotated Datasets

Abhishek Anand, Anweasha Saha, Prathyusha Naresh Kumar,
Ashwin Rao, Zihao He, Negar Mokhberian
University of Southern California

MOTIVATION

- **Bias in Annotation:** Annotator differences in **subjective tasks** introduce **bias in their annotations**, especially in sensitive domains like hate speech recognition, stemming from diverse backgrounds and perspectives.
- **Misinterpretation of Bias as Noise:** Minority votes are often considered **outliers** by models. This causes models to perceive them as **noise**, leading to biased predictions favoring **majority vote.**
- In this project, we'll explore if **perspectivist classification models** effectively utilize valuable insights from instances labeled as noisy by noise-detection techniques.


Prompt: Do you see an airplane in the given picture?
Prompt: Do you think the given text is offensive?
✓ ✓ ✗ ✓
✗ ✗ ✗ ✗
✓ ✗ ✓ ✓
✓ ✗ ✗ ✓
✗ ✓ ✗ ✗
✓ ✗ ✓ ✗

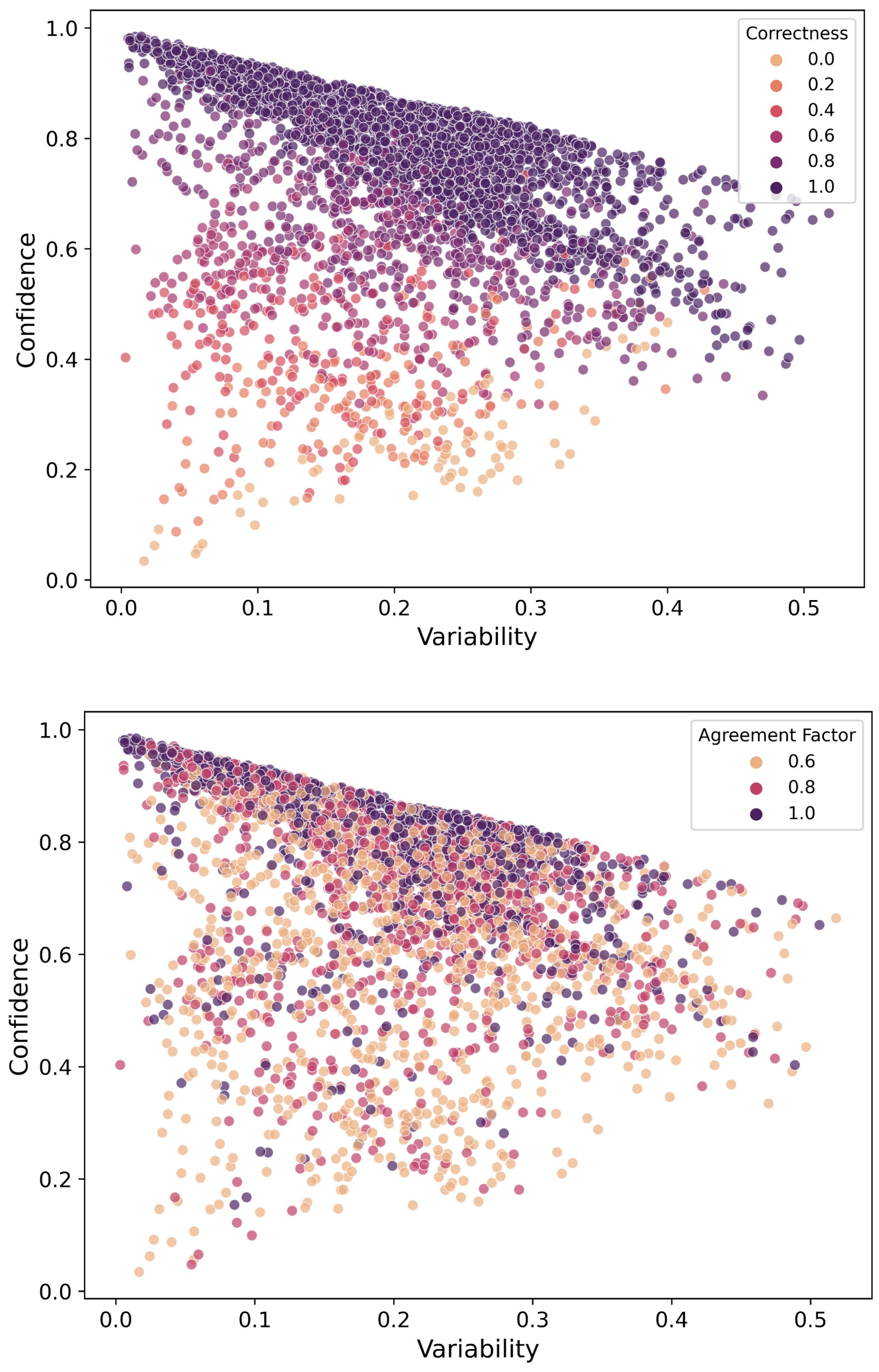
METHODS

- **Data Cartography** summarizes training dynamics for all samples as
  - ○ **Confidence:** Mean of probabilities for gold label across epochs.
  - ○ **Variability:** Standard Deviation of probabilities for gold label across epochs.
- **Multi-annotator models** leverages the diverse viewpoints brought by different annotators. They learn to **predict the labels each annotator would provide** for each instance in the dataset. These models get an instance id and an annotator id as input.

DATASETS

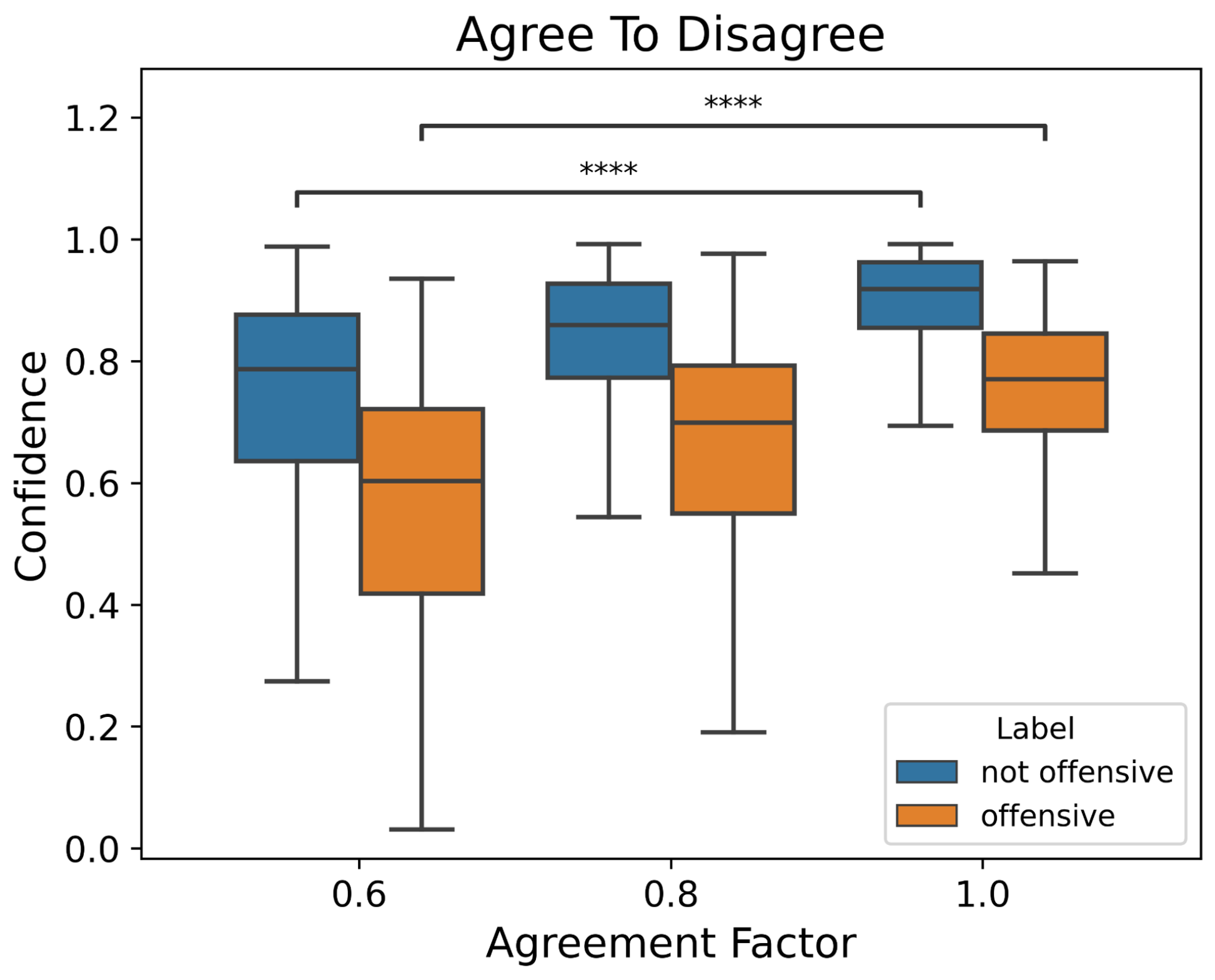
	Toxicity or Hate speech			Emotion
	Attitudes [1]	Kennedy [4]	Agree To Disagree [3]	SemEval [2]
# Annotators	184	7,912	819	183
# Annotations per annotator	18.8±25.6	17.1±3.8	63.7±139	476.4±1079
# Unique texts	627	39,565	10,440	11,090
#Annotations per text	5.5±0.8	2.3±1.0	5	7.8±3.0

RESULTS - Data Cartogram

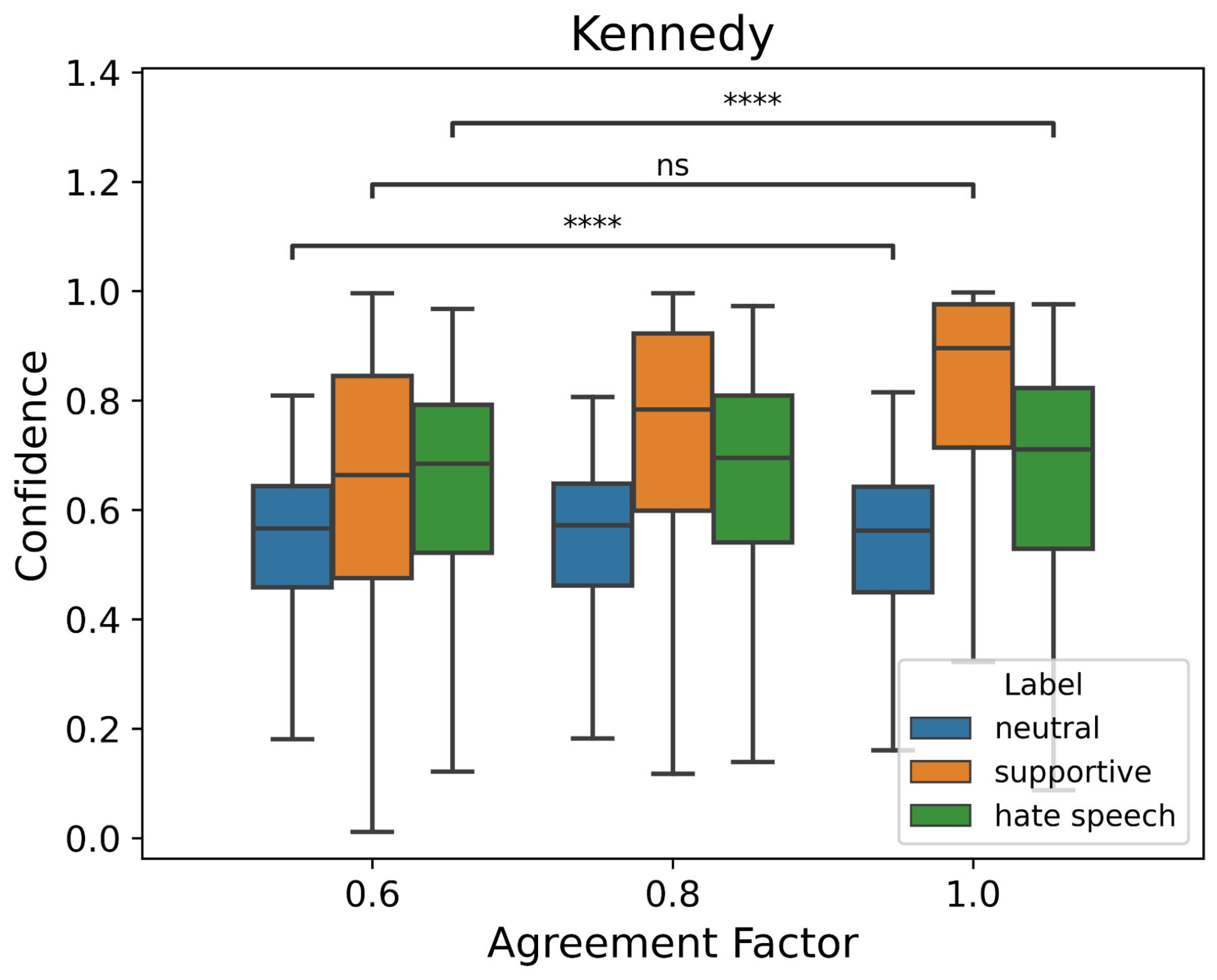


Dataset Cartography for instances of Agree To Disagree dataset with hue depicts **(a)** the correctness of the prediction from the trained model , and **(b)** the human annotators agreement on gold label (the majority vote is considered as final gold label for instances)

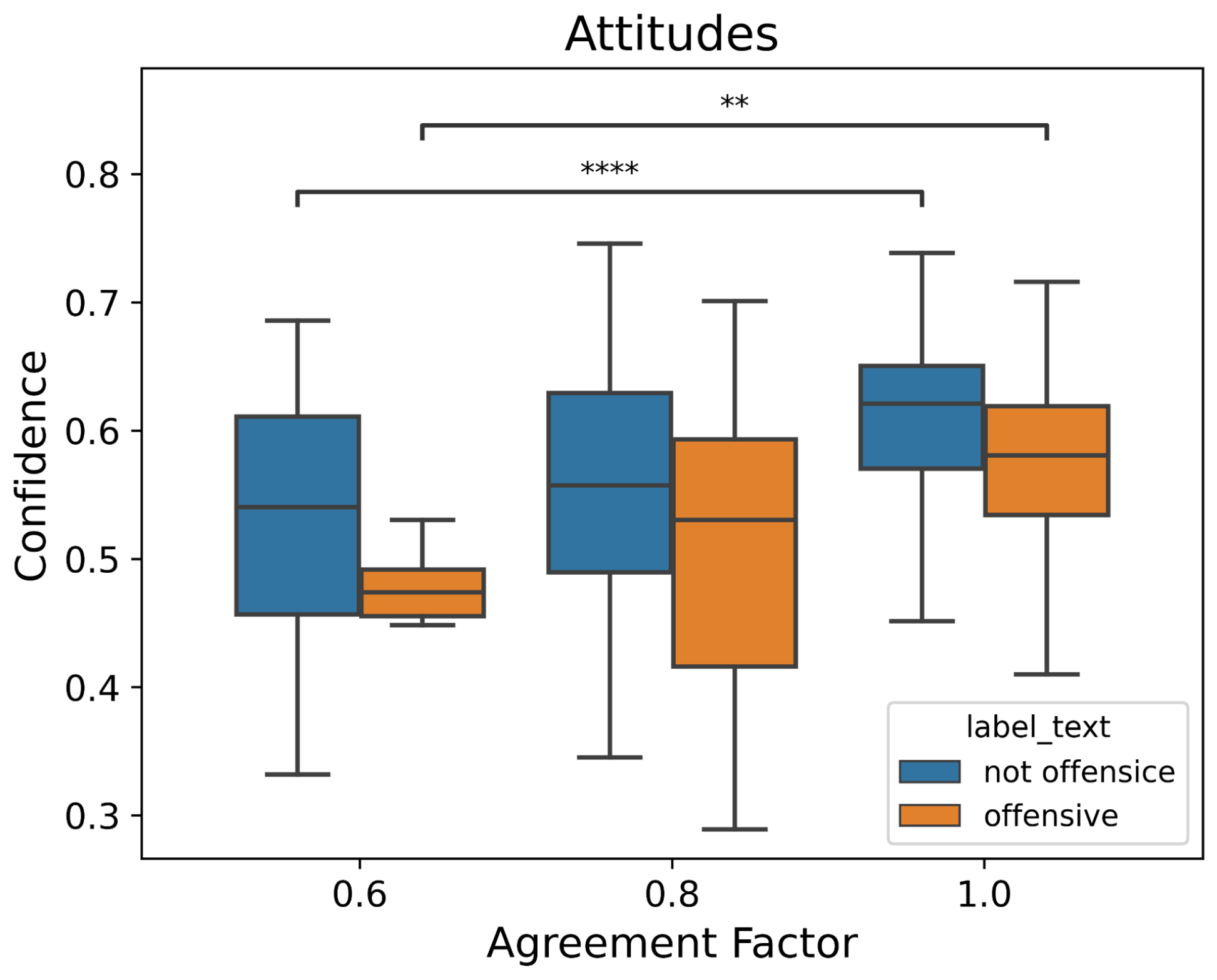
RESULTS - Correlation Between Annotator Agreements and Model Confidence



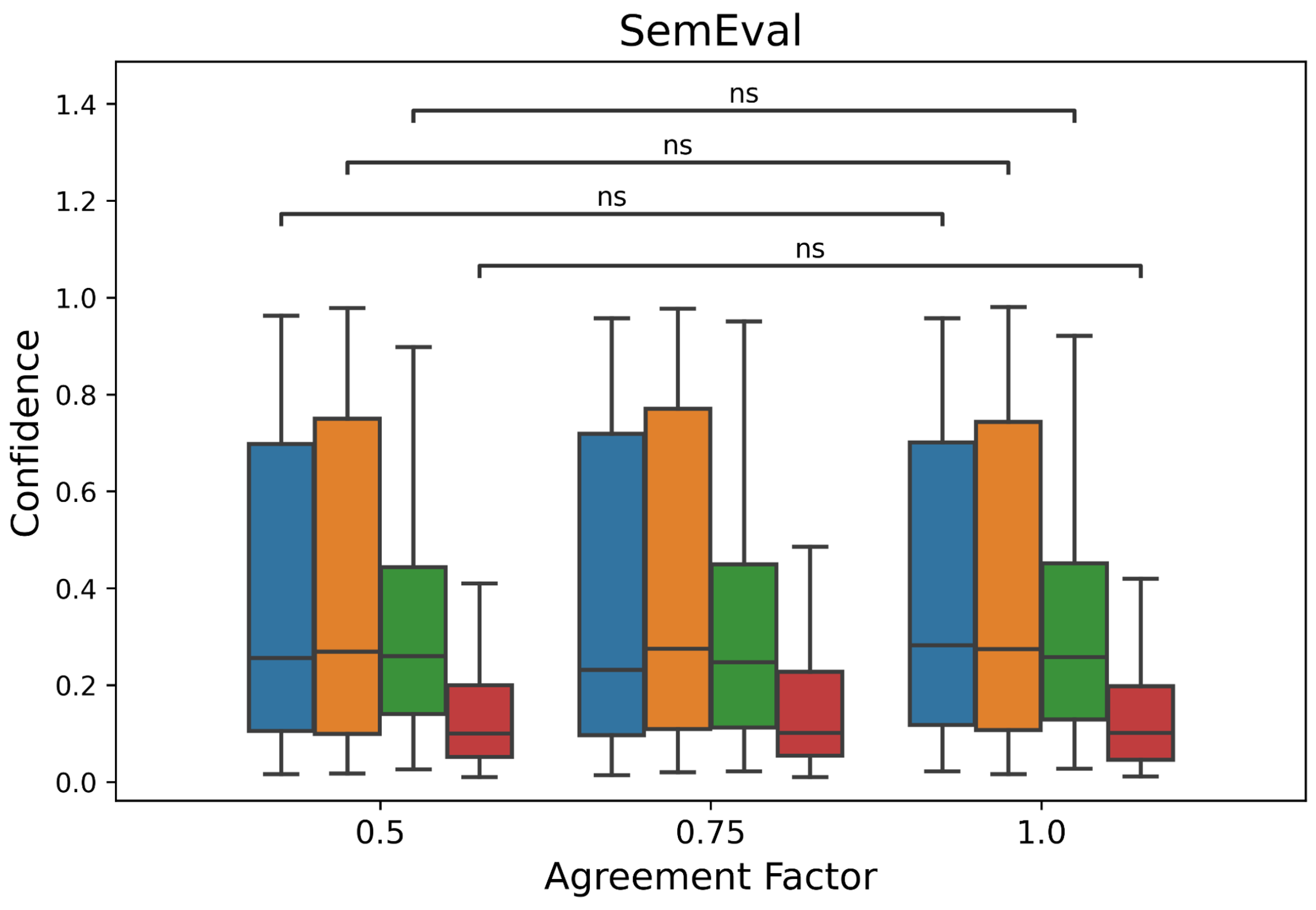
Pearson's R	P-value
0.38	3.07e-136



Pearson's R	P-value
0.20	1.35e-318



Pearson's R	P-value
0.25	8.52e-10

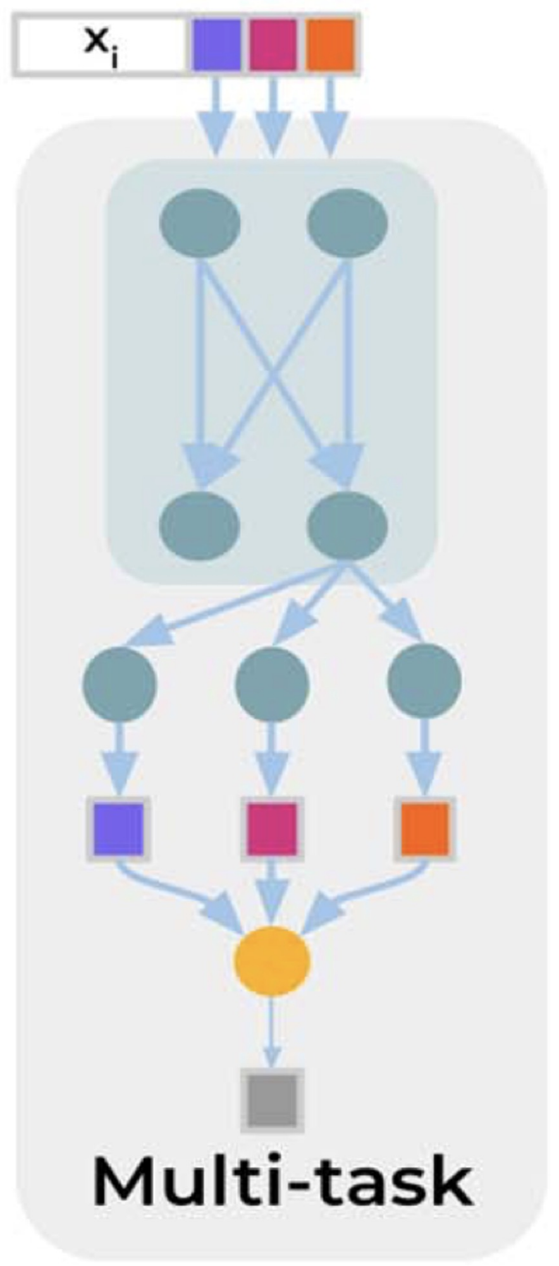


Emotion	Pearson's R	P-value
Anger	0.01	0.61
Joy	0.00	0.84
Sadness	-0.01	0.76
Fear	0.01	0.63

Focusing on the correlation between human agreement and the model's confidence over the instances we observe:

- In two of the datasets (Agree To Disagree and Attitudes) there is a significant correlation.
- In Kennedy dataset for "neutral" and "hate speech" the correlation is significant. However, for the instances that their majority vote is "supportive" there is no significant correlation.
- In SemEval which is a multi-label dataset on detecting emotions, we couldn't find any significant correlations.

NEXT STEPS



Multi-annotator models are supposed to overcome the limitations caused by simple label aggregation. These models try to learn different perspectives among the annotators based on their annotation behaviors We will train multi-annotator classifiers on our datasets.

RQ1. Do multi-annotator models give better confidence for instances that have been recognized as noisy by the baseline?

RQ2. Does the confidence remain high for non-noisy instances?

Figure shows a multi-task architecture proposed by Davani et al., 2022 [6] for a given instance learns to predict the labels each annotator would assign to it.

REFERENCES

1. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection (Sap et al., NAACL 2022)
2. SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., SemEval 2018)
3. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement (Leonardelli et al., EMNLP 2021)
4. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application (Kennedy et al., 2020)
5. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics (Swayamdipta et al., EMNLP 2020)
6. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations (Mostafazadeh Davani et al., TACL 2022)