DEMOGRAPHIC-SPECIFIC SOFT LABELS FOR TOXIC SPEECH DETECTION

By Shane Kaszefski-Yaschuk

Introduction

- Label variation among annotators is commonly seen in many NLP tasks
- This variation is even more pronounced for subjective tasks such as hate speech detection and irony detection
- In the past, any disagreement among annotators was thought of as undesirable ‘noise’
  - Researchers would instead take the majority vote among all annotators for a given text

Introduction

- Recently, the practice of aggregating labels into a singular majority vote label has come into question for several reasons:
  - Different sociodemographic groups sometimes annotate differently
  - Any minority voices which do not agree with the majority are silenced
  - Many tasks lack a single objective ‘truth’, so removing disagreement removes potentially useful information
- Recent research has instead been shifting towards the creation of more ‘perspectivist’ models (i.e., models which leverage any inherent differences in annotator perspectives)

Research Questions

- This project seeks to answer two main questions:
  - How well do toxic speech detection models trained on only one demographic attribute (e.g., only annotations from people aged 35-44, etc.) perform? Is there any noticeable difference in performance?
  - Can soft labels be leveraged to model disagreement among annotators of the same demographic group meaningfully? Which metrics work the best for this?

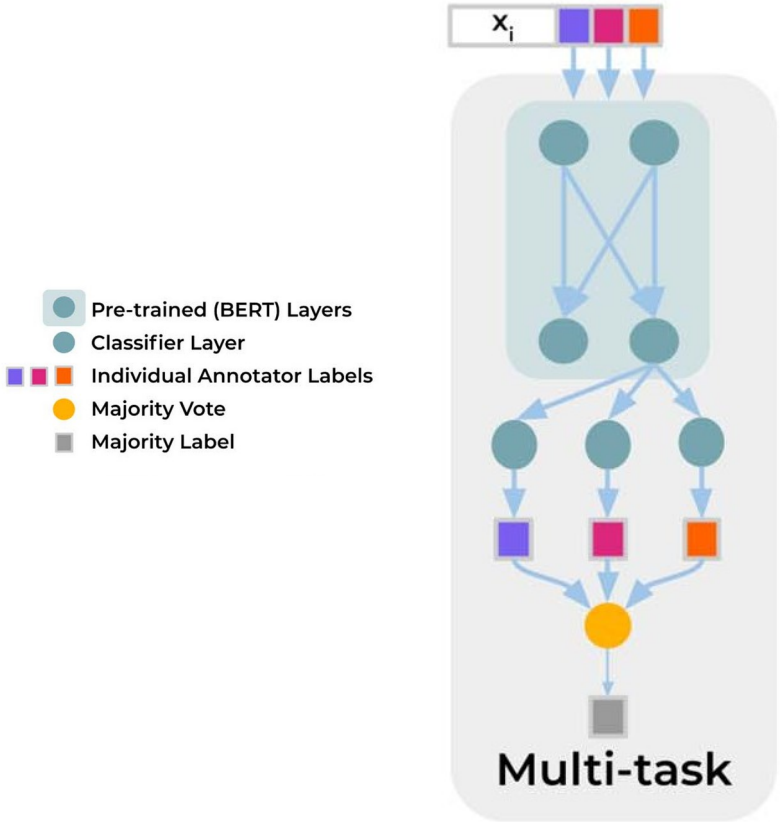
LEVERAGING DISAGREEMENTS AND DEMOGRAPHICS

Modelling Inter-Group Disagreement

- Akhtar et al. (2020) trained a hate speech detection model on two separate groups of annotators
- These groups were selected in a way which maximizes intra-group agreement and minimizes inter-group agreement
- Both group-specific models outperformed the simpler majority-vote baseline model
- An ensemble of both models further improved classification performance

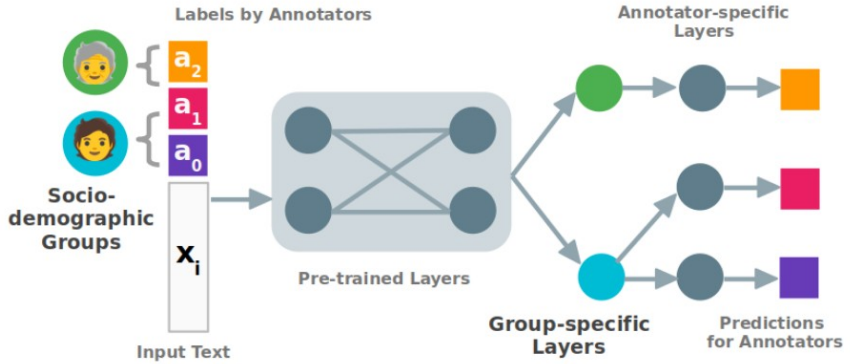
Modelling Intra-Group Disagreement

- Davani et al. (2022) introduced a multi-task model with annotator-specific classification heads
  - For a given comment, each annotator's label is learned by the model
  - These labels are then aggregated into a singular label during the final prediction stage



Leveraging Annotator Demographics

- Orlikowski et al. (2023) sought to determine the impact of explicitly accounting for the sociodemographic attributes of annotators when training a toxic speech detection model
- Building off the multi-task architecture proposed by Davani et al. (2022), this work included demographic group-specific classification heads before the annotator-specific ones
  - Contrary to previous research, they found that explicitly accounting for sociodemographics in this way did NOT improve performance



SOFT LABELLING & EVALUATION

Soft Labels

- Galstyan and Cohen (2007) define *hard* and *soft* labels as such:
  - Hard labels are *binary*: either the data point is a member of a class or it isn't
  - Soft labels are *probabilistic*: given a data point and a class, a soft label gives the probability of said data point belonging to that class
- For this project, they can also be defined as such:
  - Hard labels are the majority vote among all annotators; some text either belongs to class 0 or class 1 in the binary case (but *never* both)
  - Soft labels are defined by each annotator; if three annotators give the label '0' to some text, and two give the label '1', then the soft label would take the form **[0.6, 0.4]**

Why Soft Labels?

- The most common evaluation metrics for classification require any reference labels to binarized
- Unfortunately, converting labels to a binary format also removes any potentially useful information regarding inter-annotator disagreement
- Soft labels, on the other hand, keep any intra-group disagreements intact
- However, Rizzi et al. (2024) note that common evaluation metrics (e.g., macro F1, accuracy, etc.) are not sufficient when examining the performance of soft labels

Evaluation Metrics

- Rizzi et al. (2024) tested the performance of several “soft evaluation” metrics for both binary and multi-class classification:
  - Cross-Entropy
  - Manhattan Distance
  - Euclidean Distance
  - Jensen-Shannon Divergence
- They identified 6 different desirable properties for these metrics to have
- They found that in the binary classification case, using Manhattan distance and Euclidean distance as evaluation metrics satisfies all 6 of these properties, while Cross-Entropy and Jensen-Shannon Divergence did not

“Fuzzy” Metrics

- Rather than adopting new evaluation metrics, Harju and Mesaros (2023) instead sought out ways to calculate precision, recall, and F1-scores without the need to binarize labels beforehand
- They proposed ‘fuzzy’ equivalents of these three metrics:

$$\begin{aligned}
\text{Precision} &= \frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i \hat{y}_i}, \\
\text{Recall} &= \frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i y_i}, \\
\text{F}_1 - \text{score} &= 2\frac{\sum_i \min(\hat{y}_i, y_i)}{\sum_i (\hat{y}_i + y_i)}.
\end{aligned}$$

EXPERIMENTAL SETUP

Project

- This project leverages both soft labels to capture intra-group disagreement as well as demographic-specific groupings to (potentially) capture inter-group disagreement
- With soft labels, disagreement among annotators can be leveraged as a source of information
  - Even without explicitly modelling each annotator individually
  - Grouping together annotators based on their sociodemographic attributes increases the representation of each individual voice

Dataset

- This project uses the Kumar et al. (2021) dataset for toxic speech detection
- Contains 107,620 comments annotated by 17,280 annotators for a total of 549,058 annotations
- Each comment was rated from 0 – 4 based on how toxic they are
  - 0 being ‘not at all toxic’ and 4 being ‘extremely toxic’
- Extensive information about each annotator was also collected:
  - Demographic groups (age, education level, gender, etc.)
  - Behaviours (how often they use social media, which platforms they prefer, etc.)
  - Attitudes (e.g., whether they think toxic speech online is an issue or not, political leanings, etc.)

Dataset Preprocessing

- The dataset was preprocessed before training
- The toxicity scores were first binarized:
  - Scores from 0-1 were turned into a 0 (‘non-toxic’)
  - Scores from 2-4 were turned into a 1 (‘toxic’)
  - Final class ratio: ~70% non-toxic, ~30% toxic
- Four sociodemographic attributes were selected:
  - Age
  - Gender
  - Education Level
  - Political Leaning

Train and Test Splits

- With the pre-processed dataset, specific train and test splits were generated
- 90% of the data was used for the train (and validation) split
- The remaining 10% was used for the test split
  - Any comments found in the train split were removed from the test split to ensure only unseen data is used at test time

Experimental Setup – Training Configurations

- Four different configurations were used when fine-tuning BERT:
  - Demographic-specific models with soft labels
  - Demographic-specific models with hard labels (i.e., the majority vote for that group)
  - Baseline model (no demographic splits) with soft labels
  - Baseline model (no demographic splits) with hard labels (i.e., the majority vote among *all* annotators for each comment)
- Initially, the labels were calculated from *all* annotators (i.e., no demographic splits)
- However, the test labels always match the label configuration under examination

Experimental Setup – Metrics and Loss

- Three evaluation metrics were selected:
  - Regular Macro-F1 (hereafter referred to as the ‘hard’ F1 score)
  - ‘Fuzzy’ Macro-F1
  - Manhattan Distance
- L1 Loss (i.e., the Manhattan distance) was selected as the loss function rather than Cross-Entropy for two main reasons:
  - Initial trial runs showed better performance for all splits
  - The Manhattan distance is also used as an evaluation metric, so the L1 loss directly optimizes for this metric

Preliminary Results v1.0

- Initial results (and the last group meeting) revealed a few issues

Baseline	Hard Macro F1	Fuzzy Macro F1	Manhattan Dist	# of Train Instances
Baseline (Soft)	**0.67**	**0.82**	22.80	24,000
Baseline (Hard)	0.43	0.76	**16.71**	24,000

Gender	Hard Macro F1	Fuzzy Macro F1	Manhattan	# of Train Instances
Female (Soft)	0.63	**0.75**	90.78	24,000
Female (Hard)	0.42	0.71	57.09	24,000
Male (Soft)	**0.65**	**0.75**	44.47	24,000
Male (Hard)	0.42	0.72	123.1	24,000
Non-Binary (Soft)	0.53	0.73	**35.01**	2,065
Non-Binary (Hard)	0.42	0.72	160.43	2,065
Prefer not to say (Soft)	0.54	0.73	49.44	2,244
Prefer not to say (Hard)	0.43	0.72	109.93	2,244

Current Directions

- Pairwise comparisons between different demographic groups (train split for one demographic, test split for another)
  - Baseline: entire dataset w/o splits for both train and test (soft labels only)
  - 2nd Baseline: train the entire dataset w/o splits and test on demographic-specific test sets
- Implemented the ‘normalized’ Manhattan distance metric
  - 1 – (Sum of all Manhattan distances divided by the number of labels)

Preliminary Results v2.0

Baseline	Hard F1	Fuzzy F1	Normalized Manhattan
Baseline	**0.67**	**0.82**	**0.64**
Bachelor's	0.62	0.72	0.45
High School	0.62	0.70	0.39
Master's	0.59	0.65	0.30
Some College	0.65	0.72	0.44
Associate's	0.64	0.71	0.41
Less than HS	0.59	0.68	0.36
Doctorate	0.64	0.71	0.41
Professional (JD/MD)	0.58	0.67	0.33
Other	0.58	0.65	0.31
Prefer Not To Say	0.61	0.66	0.31

Future Directions

- Further examination of evaluation metrics and loss functions for soft labels
- Ensemble models
- Intersectionality