Hybrid Intelligence

#15: Generalizability of NLP Experiments

Urja Khurana[1], Antske Fokkens[1] ,Eric Nalisnick[2], Ivar Vermeulen[1]

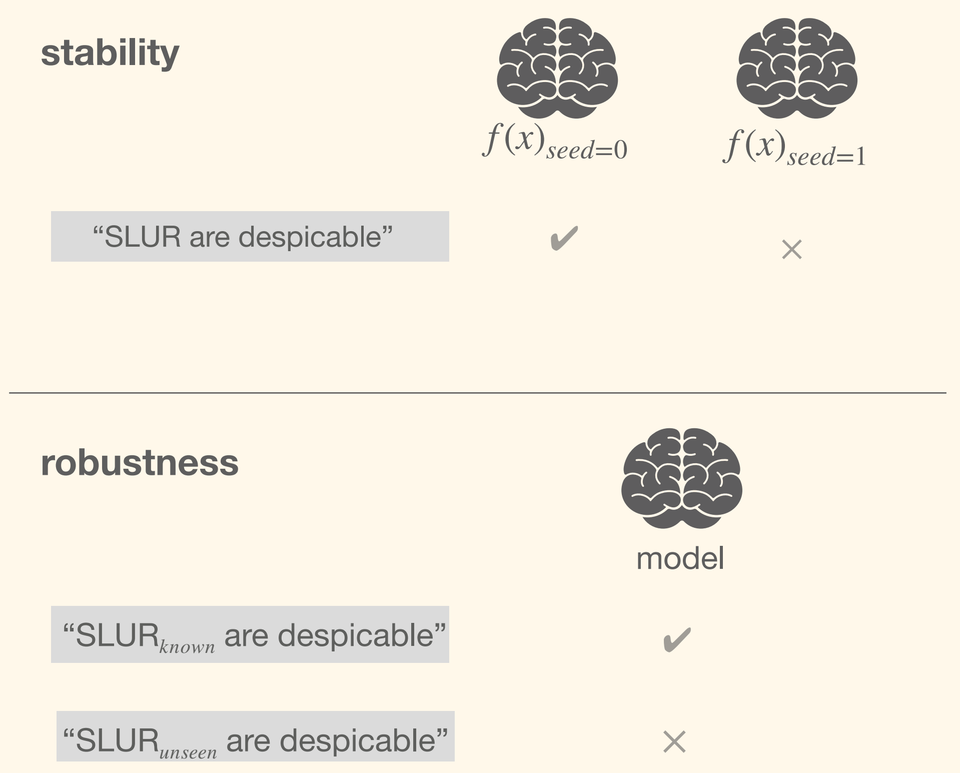
Disclaimer: Due to the nature of the subject, this poster might contain strong language. This does not reflect the opinion of the authors.

1. Stability and Robustness

Current Large Language Models (LLMs) in a fine-tuning setup achieve high scores on data that resembles the training set.

Prone to lack of **stability** and **robustness** when trained on noisy data!

Problematic for safety-critical applications e.g. hate speech



2. Definitional Building Block: Hate Speech Criteria (HSC)

1. **Target:** Person or group from specific

■Gender ■Disability ■Race
■Nationality ■Sexual Orientation ■Religion
■Color □Language □Class
■Ethnicity

2. **Target:**
Are dominant groups also considered alongside non-dominant groups:

○No ○Yes ○Yes, but depends on severity*

*Elaborate................

3. **Perpetrator:**
Are perpetrator characteristics taken into account?
○Yes ○No

If **YES**, which aspects:
□ The dominance of the group
□ Societal role
□ Member of target group itself

4. **Presence of explicit reference to group through:**
▶Stereotype
▶Group Characteristic
▶Slur
related to above specified target(s)

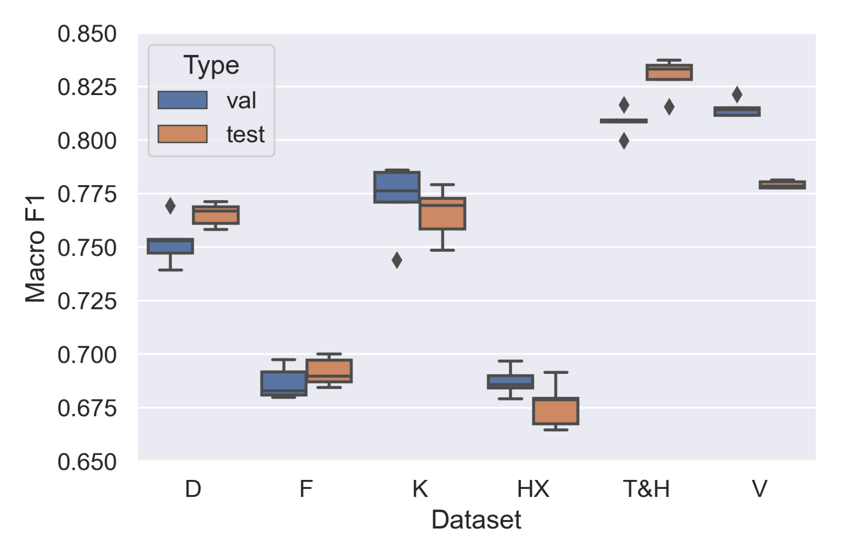
5. □ **Insults group**
□ **Incites:**
■Violence
■Hate
□Discrimination

3. Explicit vs. Implicit Definitions

Can we use a dataset's **definition** as a proxy for what can be expected from **model behavior** and the **generalization capabilities** of a model?

Imprecise definitions can be a spoilsport!

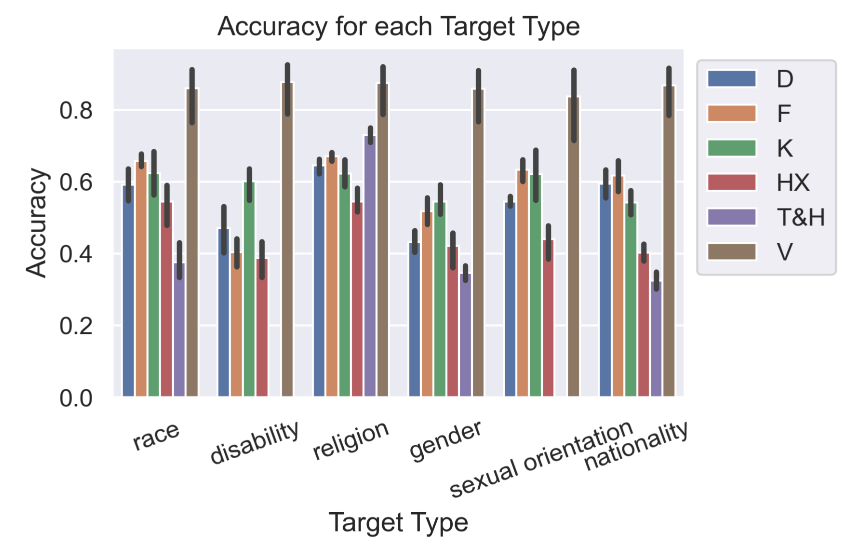
1. **Identify** for 6 different Hate Speech datasets which **aspects** of HSC are **present in definition**
2. **Link** challenge set instances to **aspects from HSC** and **investigate overlap** between expectations and actual performance
3. **Analyze** if similar definitions perform well in **cross-dataset** evaluation

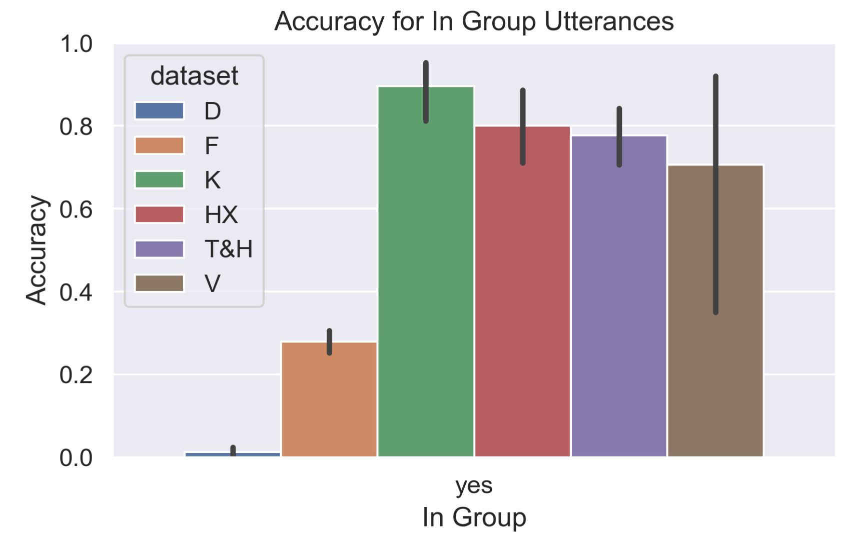


4. Definitions Reflected in Model Behaviour?

No consistency between expected aspects from definition and actual model behavior.

More **precise definitions** tend to perform **better on the challenge set** though!





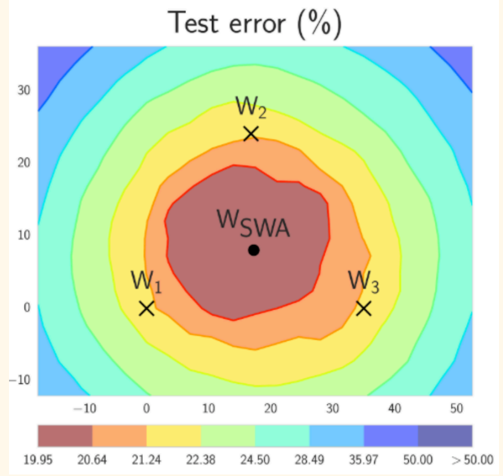
5. Weight Ensembling and Training Data Properties

Underspecification: trajectory of solution in **loss landscape influenced by inherent randomness** of training highly parameterized models with SGD

Ensembling may cause randomness to **cancel out**

Can Stochastic Weight Averaging (**SWA**) improve stability and robustness in case of noisy data?

How do different ensemble members contribute to the final model?



Source: Izmailov et al. (2018)

6. Data Manipulations and Evaluation

Noise in data can stem from different sources

Apply different **data manipulations** to two tasks and see for **which** ones **SWA** shows **improvement**:

1. Class (Im)balance
2. Dataset Size
3. Spurious Correlations
4. Shortcuts

Evaluate agreement (**stability**) with Fleiss' Kappa between models

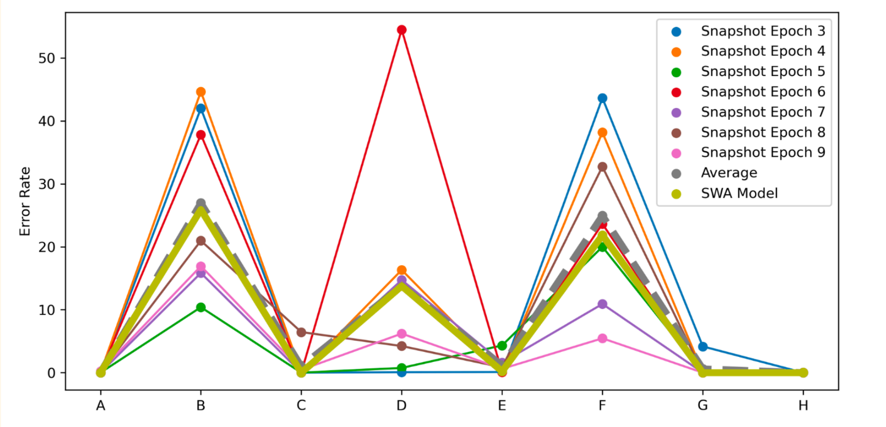
- Annotator : Models
- Annotations : Predictions

Use challenge set to evaluate **robustness**

7. Influence of SWA

No consistent **improvement** in performance nor influenced by data manipulations

Compositionality: SWA largely follows **majority vote**



(b) CoLA.

8. Future Directions

More analysis on loss landscape and how this impacts the influence of SWA

What do spurious correlations and shortcuts mean for robust hate speech detection?

Establish a link between **what** is in the data, **what** is captured by the model, and **what** is similar between the training data and unseen data

Where does the mismatch between definition and model behavior stem from? E.g. annotation differences, lack of coverage in data?

How can this mismatch be corrected?