Learning from the Worst: Dynamically Genearated Datasets to improve Online Hate Detection

park chaewon

Abstract

- present a **human-and-model-in-the-loop** process for **dynamically generating datasets**
- provide **a new dataset** of ~40, 000 entries, generated and labelled by trained annotators **over four rounds** of dynamic data creation. includes ~15, 000 **challenging perturbations**.
- Hateful entries make up 54% of the dataset, which is substantially higher than comparable datasets.
- Models trained on **later rounds** of data collection **perform better** on test sets

1. Introduction

1. Introduction

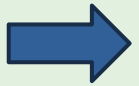
- **importance of hate speech detection** : minimizing the risk that harm will be inflicted on victims and making online spaces more accessible and safe.
- **difficult of detecting online hate** : performance, robustness, generalisability and fairness of even stateof-the-art models
- **To address these challenges** : a human-and-model-in-the-loop, 4 rounds of data generation and model training

1. Introduction - 4 rounds

1. We first trained a classification model using previously released hate speech datasets.
2. We then tasked annotators with presenting content that would trick the model and yield misclassifications.
3. At the end of the round we trained a new model using the newly presented data.
4. In the next round the process was repeated with the new model in the loop for the annotators to trick.

1. Introduction - 4 rounds

새로운
분류 모델
생성



이전 라운드
데이터의 조합으로
분류모델 학습

속일 수 있을 만한
content enter
& 실시간 피드백을 받아
데이터 생성

ex) R2에서 새롭게 분류 모델 생성(M2) -> R0, R1의 데이터로 M2 학습
-> M2를 이용해 어려운 content 생성(R2의 데이터) -> R3에서 새롭게 분류 모델 생성(M3)
-> R0, R1, R2의 데이터로 M3 학습 …

1. Introduction

- **contrast sets** : manipulate the original text just enough to flip the label
- **inspiration from real-world** : to make content as adversarial, realistic, and varied as possible.
- **lower accuaracy from later rounds** : content becomes more adversarial.
- **learn from the worst** : become increasingly accurate in detecting hate & annotators have to provide more challenging content

1. Introduction - contribution

make 3 contribution to online hate classification research

1. present a human-and-model-in-the-loop process
2. present a dataset of 40, 000 entries fine-grained
3. present high quality and robust hate detection models

2. Background

2. Background - Benchmark dataset

Several benchmark datasets have been put forward for online hate classification

Numerous problems with hate speech training dataset

- lacking linguistic variety
- being inexpertly annotated
- degrading over time

In addition, many datasets are formed with **bootstrapped sampling**, such as keyword searches

2. Background - Model limitation

Systems trained on existing datasets have been shown to lack accuracy / robustness / generalisability / **creating a range of false positives and false negatives**

- **False positives (non-hateful -> hateful misclassified)**

overfit on the use of slurs and pejorative terms, treating them as hateful irrespective of how they are used (ex/ counter speech)

- **False negatives (hateful -> non-hateful misclassified)**

Subtle and implicit forms of hate speech can also create false negatives (ex/ sarcasm, irony, rhetorical questions)

2. Background - Dynamic benchmarking and contrast sets

Addressing the flaws of models is a difficult task.
problem may partly lie in :

1. the use of static benchmark datasets
2. fixed model evaluations

solution :

1. dynamic dataset creation setup
2. creating ‘contrast sets’ of perturbations

these approaches remain under-explored for hate speech detection
no prior work has combined the two approaches

3. Dataset labels

3. Dataset labels

Previous research shows the limitations of using only a binary labelling schema.

there are few established taxonomies and standards in online hate research.

The **hierarchical taxonomy** we present aims for a balance between granularity versus conceptual distinctiveness and annotation simplicity

Hate / Non Hate

‘Hate’ secondary labels -> type, target

3. Dataset labels - Type of hate

- Derogation
  : derogatory, attacks, insult
- Animosity
  : implicit, subtle
- Threatening language
  : inflicting harm
- Support for hateful entities
  : supports hateful action
- Dehumanization
  : describing groups as insect

3. Dataset labels - Targets of hate

Target	Number of entries
Black people	2,278
Women	2,192
Jewish people	1,293
Muslims	1,144
Trans people	972
Gay people	875
Immigrants	823
Disabled people	575
Refugees	533
Arabs	410

target - vulnerable, marginalized
or discriminated-against group

provided annotators with a non-exhaustive list of 29 identities
(e.g., women, black people, Muslims, Jewish people and gay)

* identities considered out-of-scope for hate : men, white people, and heterosexuals.

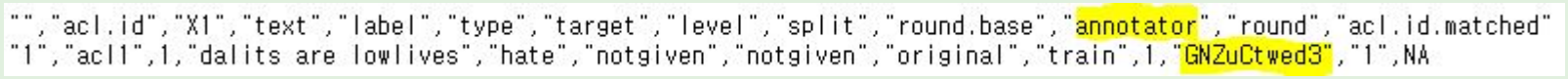
4. Annotation

4. Annotation

- Annotation guidelines - created at the start of the project and then updated after each round
- annotators (20) - received extensive training and feedback during the project.
- small pool of annotators - driven by the logistical constraints of hiring and training them to the required standard and protecting their welfare given the sensitivity and complexity of the topic

  -> raises the potential for bias.

  -> provide an annotator ID to enable further research into this issue.



5. Dataset formation

5. Dataset formation

dataset was generated over four rounds.

each round involved ~10,000 entries.

final dataset comprise 41,255 entries

after each round, the data was split into training(80%),

dev(10%), test(10%).

Approximately half of the entries in the test sets are produced by annotators who do not appear in the training and dev sets.

: This makes the test sets more **challenging** and minimizes the risk of annotator bias given our relatively small pool of annotator

5. Dataset formation - Target model implementation

target model - Every round has a model in the loop
model architectual - RoBERTa / Transformers library
train data - a combination of data collected in the previous round(s). ex) M2 <- R0, R1 data

identify the best sampling ratio of previous rounds' data using the dev sets.
M1 <- R0 data
M2 <- R0, upsampled R1 data
M3 <- R0, upsampled R1, upsampled R2 data
M4 <- R0, upsampled R1, upsampled R2, one lot of the R3 data

5. Dataset formation - Round 1

train data - 11 English language training datasets for hate and toxicity taken from hatespeechdata.com

annotators were instructed to enter synthetic content into
the model that would trick M1

All entries were validated by one other annotator
marked as incorrect -> sent for review by expert annotators
decided the final label / made minor adjustments to the text.

final dataset comprises 11,157 entries ‘Hate’ (65%), ‘Not Hate’ (35%)

5. Dataset formation - Round 2

pivot - guide annotator work

10 hateful, 12 not hateful

train data - half comprises originally entered content

other half comprises perturbed contrast sets.

guidance - realistic perturbed/ meet the criteria ...

original entries which fooled model <- 3~4 annotator validate

every perturbation <- 1 annotator validate

inccorect/flagged original&pertubations -> reviewed by expert.

Krippendorf’s alpha = 0.815.

5. Dataset formation - Lessons from R2

1. several ‘template’ statements were entered by annotators.
2. in attempting to meet the ‘pivots’ they were assigned, some annotators created unrealistic entries
3. the pool of 10 trained annotators is large for a project annotating online hate but annotator biases were still produced.

5. Dataset formation - R3 & R4

In R3, annotators were tasked with finding realworld hateful online content to inspire their entries.

In R4, each annotator was given a target identity to focus on,

	R3	R4
entered entries	9950	10152
Krippendorf's alpha	0.55	0.52
reviewed entries	981	967

6. Model performance

6. Model performance - model error rate

Round	Total	Not	Hate	Animosity	Dehuman -ization	Derogation	Support	Threatening
R1	54.7%	64.6%	49.2%	-	-	-	-	-
R2	34.3%	38.9%	29.7%	40.1%	25.5%	28.7%	53.8%	18.4%
R3	27.8%	20.5%	35.1%	53.8%	27.9%	29.2%	59.6%	17.7%
R4	27.7%	23.7%	31.7%	44.5%	21.1%	26.9%	49.2%	18.3%
All	36.6%	35.4%	37.7%	46.4%	24.8%	28.3%	55.4%	18.2%

Table 3. Error rate for target model

6. Model performance - Test set performance

target model (upsampled)

training only one round train set

not upsampled

Model	R1	R2	R3	R4
M1 (R1 Target)	44.84±1.1	54.42±0.45	66.07±1.03	60.91±0.4
M2 (R2 Target)	90.17±1.42	66.05±0.67	62.89±1.26	60.87±1.62
M3 (R3 Target)	91.37±1.26	77.14±1.26	**76.97±0.49**	74.83±0.92
M4 (R4 Target)	**92.01±0.6**	**78.02±0.91**	75.89±0.62	**75.97±0.96**
M(R1 only)	**92.20±0.55**	62.87±0.63	47.67±1.04	52.37±1.27
M(R2 only)	80.73±0.4	76.52±0.7	**77.43±0.51**	**74.88±0.85**
M(R3 only)	72.71±1.05	**78.55±0.71**	74.14±1.5	73.16±0.58
M(R4 only)	72.26±1.3	76.78±1.65	77.21±0.43	69.6±0.6
M(R0+R1)	88.78±0.89	66.15±0.77	67.15±1.11	63.44±0.26
M(R0+R1+R2)	91.09±0.37	74.73±0.95	74.73±0.46	71.59±0.59
M(R0+R1+R2+R3)	**91.17±0.99**	77.03±0.72	74.6±0.48	**73.94±0.94**
M(R0+R1+R2+R3+R4)	90.3±0.96	**77.93±0.84**	**76.79±0.24**	72.93±0.56

Table 4. macro F1 of models

6. Model performance - HateCheck

- **HATECHECK -** suite of functional tests for hate speech detection models
- To better understand the weaknesses of the target models from each round, apply them to **HATECHECK**

- Performance is better than all four models evaluated by Rottger et al.(2020)
- performance of M4 is consistent across both ‘Hate’ and ‘Not Hate’, achieving 95% and 93% respectively.

7. Discussion

7. Discussion

- Dynamic data creation systems offer several **advantages** for training better performing model.
- However, this approach also presents some **challenges**.

8. Conclusion

8. Conclusion

- presented a human-and-model-in-the-loop process for training an online hate detection system.
- They have fine-grained annotations for the type and target of hate, and include perturbations to increase the dataset difficulty.
- demonstrated that the models trained on these dynamically generated datasets are much better at the task of hate speech detection

Thank
you