Cross-Dataset Generalization for COVID-Related Hate Speech Detection

Student: Xuanyu Su

Supervisor: Isar Nejadgholi

Text Analytics team

Table of Content

Introduction

- Motivation
- Challenges
- Research problems

Datasets

- Data source
- Ratio of positive class
- Labels

Results

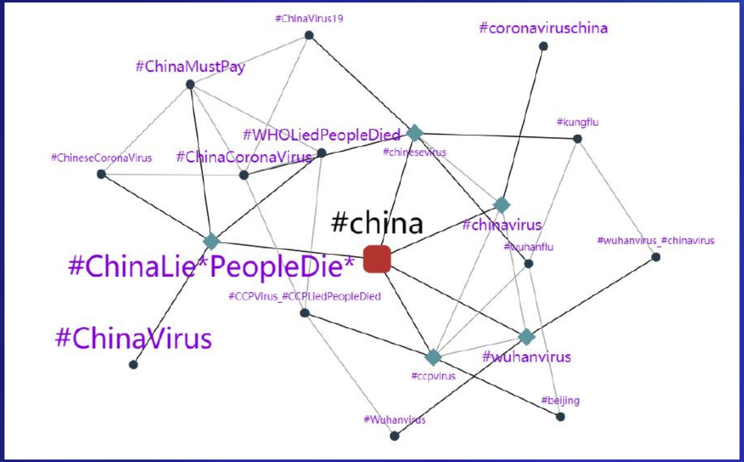
- Introduction of classifiers
- Results of different models

Conclusion

- Future direction
- Conclusions

Motivation

- **Online hate:** Hateful language occurs on the Internet and leads to severe harm for the targeted group and the society in large. NLP is often used to detect this type of language.
- **Online hate linked with Covid-19**: Detecting hateful comments with more specific themes; Anti-Asian racism during COVID-19.



Neighboring keywords for the target word ‘China’ on Twitter. (Rabiul, et al.,2020).

Research Question

- Data annotation is laborious, costly and time consuming.
- **Cross-Dataset Generalization:**
  - Can we use the datasets annotated for other types of abusive behaviour to detect COVID-related hate speech?
- **Train Datasets:**
  - 3 Previously annotated datasets for other types of abusive behaviour
  - 1 dataset annotated for COVID-related hate speech
- **Test Set :Covid Racism dataset**
  - too small for splitting into train, valid and test sets

Challenges

Datasets are created with different task formulations and definitions.

Examples of previously used definitions:

- ➢ **Attack:** Violent, harmful, or destructive comment against other people.
- ➢ **Toxic**: Makes people leave the platform.
- ➢ **Hate Speech**: Hateful words based on race, ethnicity, nationality, religion and etc..
- ➢ **East Asian Prejudice**: Prejudice against East Asians.

Train Datasets

Attack Comment (binary)

- Source: English Wikipedia
- Ratio of positive class: 0.10
- 15362(pos): 159686(neg)
- Labels: Normal / Personal attack

Toxicity (binary)

- Source: English Wikipedia (larger than AC)
- Ratio of positive class: 0.15
- 232055(pos): 1366234(neg)
- Labels: Normal /Toxic

GAB

- Source: from social website: gab.ai
- Ratio of positive class: 0.13
- 11249(pos): 75280(neg)
- Labels: Normal / Hate /(13 subcategories)

East Asian (covid-related)

- Source: collected from Twitter
- Ratio of positive class: 0.27
- 5331(pos): 14669(neg)
- Labels: Normal / hostile/critical/ counterhate/discussion

Test Dataset

Covid Racism

- Source: Twitter
- Ratio of positive class: 0.29
- 678(pos): 1641(neg)
- Labels: Hate/ Counter-hate/ Neutral/ others

Binarizing the labels

Database	Positive	Negative
Toxicity	Toxic	Normal
Attack Comment	Personal attack	Normal
Gab	Vulgarity or Offensive language Hate based on race or ethnicity Hate based on nationality/regionalism	Others
Esat Asian	Entity_directed_hostility Entity_directed_critism	None_of_above, Counter_speech, Discussion_of_eastasian_prejudice
Covid Racism (test dataset)	Hate	Counter-hate Neutral Others

Classifiers

BERT + Linear

Pretraining Data:

- ❖ Wikipedia
- ❖ Books Corpus

RoBERTa + Linear

Pretraining Data:

- ❖ Wikipedia
- ❖ Books Corpus
- ❖ News
- ❖ Other texts from Web

Classifiers

BERT + Linear

Hyperparameters:

- Epochs: 5
- Learning Rate: le-5
- Weight-decay: 0.01
- Batch size: 16
- Optimizer: AdamW
- Criterion: CrossEntropy()

Evaluations:

- ROC-AUC score
- F1 scores
- Macro Avg

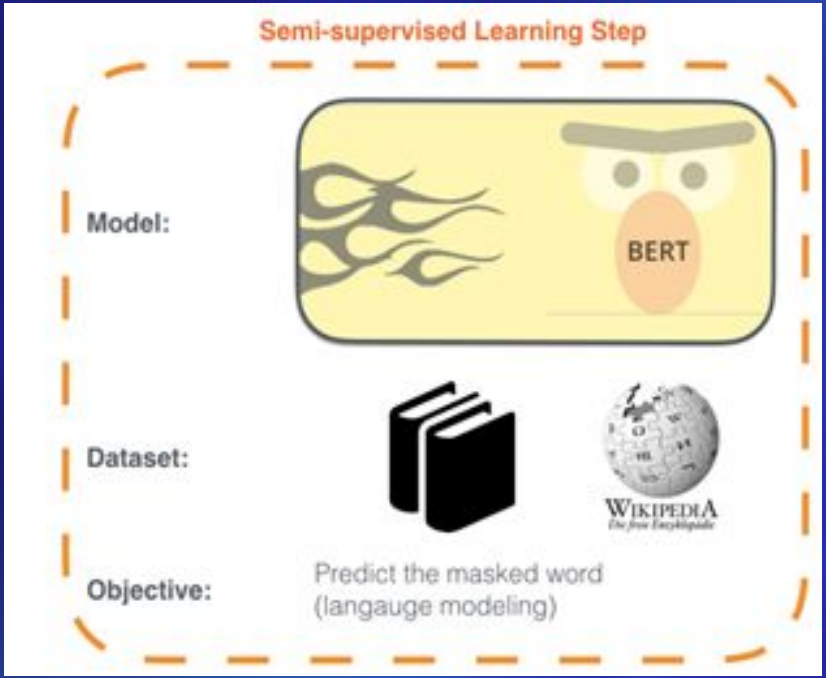


Figure 2: The pretrained datasets and objective of BERT model

Classifiers

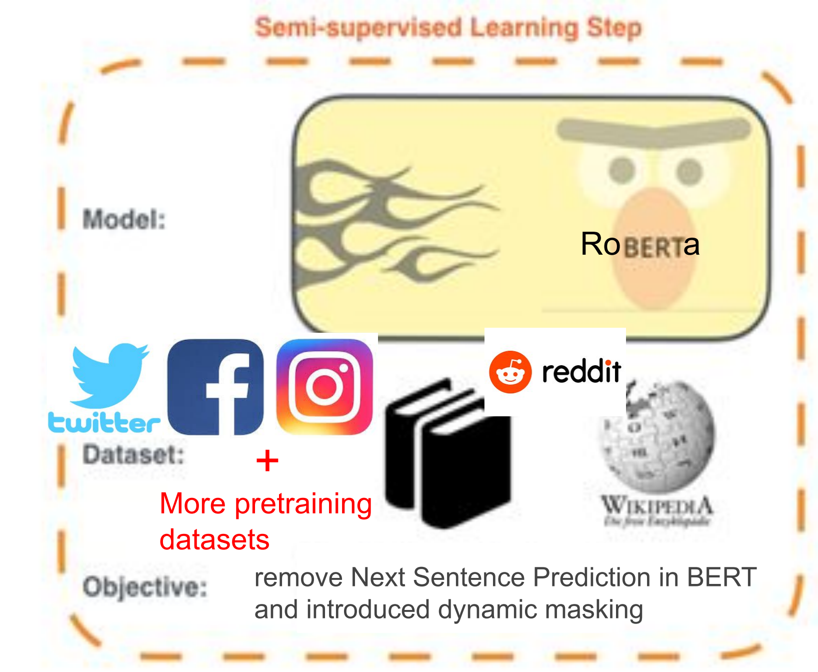


Figure 3: The pretrained datasets and objective of RoBERTa model

RoBERTa + Linear

Hyperparameters:

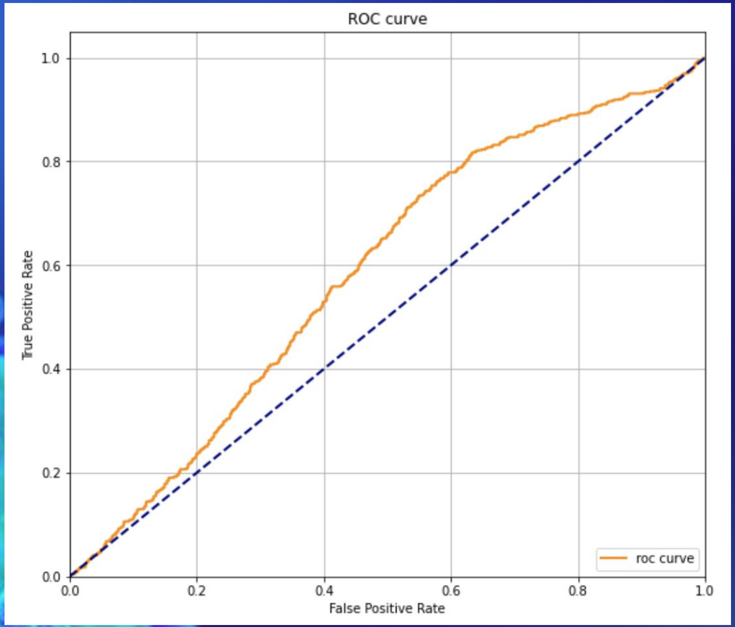
- Epochs: 3
- Learning Rate: le-5
- Weight-decay: 0.01
- Batch size: 16
- Optimizer: AdamW
- Criterion: CrossEntropy()

Evaluations:

- ROC-AUC score
- F1 scores
- Macro Avg

Example

E.g.: RoBERTa model trained with EA data, tested on Covid Racism data:



Classification Report:

	precision	recall	f1-score	support
1	0.6142	0.7419	0.6720	678
0	0.8833	0.8074	0.8437	1641
accuracy			0.7883	2319
macro avg	0.7487	0.7747	0.7578	2319
weighted avg	0.8046	0.7883	0.7935	2319

The classification report of model

The ROC curve of model

Results

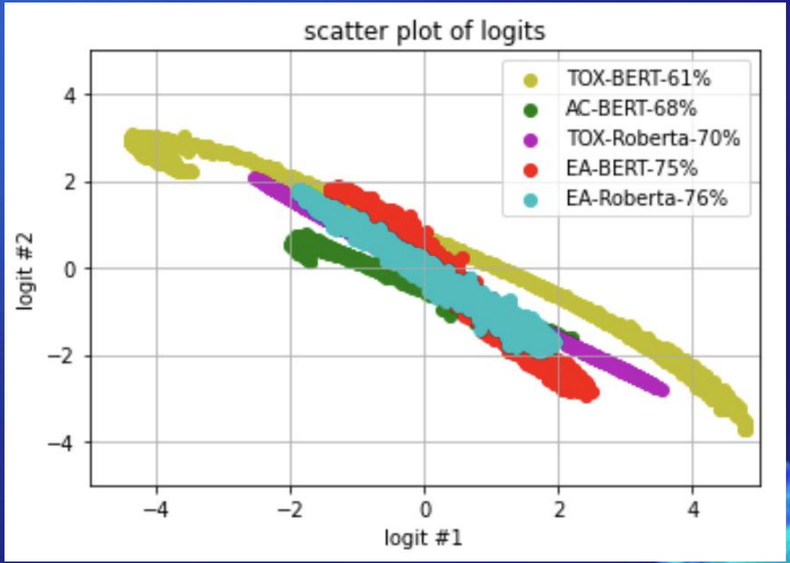
Accuracy per class and macro-averaged F1-score on the covid-racism dataset for different models and training datasets

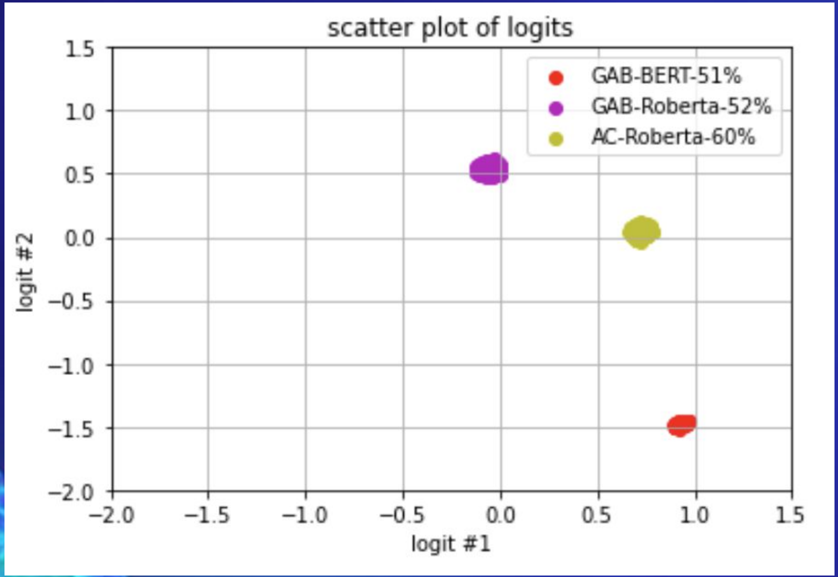
Training dataset / classifier	Attack Comment	Gab	EA	Toxicity
BERT+Linear	0.6460(pos) 0.7228(neg) 0.6844(MA)	0.2439(pos) 0.7876(neg) 0.5157(MA)	**0.6683(pos)** **0.8242(neg)** **0.7462(MA)**	0.5240(pos) 0.7060(neg) 0.6150(MA)
RoBERTa+Linear	0.5278(pos) 0.6804(neg) 0.6041(MA)	0.2676(pos) 0.8023(neg) 0.5294(MA)	**0.6720(pos)** **0.8437(neg)** **0.7578(MA)**	0.6173(pos) 0.7974(neg) 0.7060(MA)

Future Direction

We visualized the distribution of classifier output for each dataset.

Is the distribution of classifier output an indicator of generalizability?





Conclusion

- The EA data set generalized best to the Covid Racism data (both detect attack languages towards Asians during COVID-19).
- For the Toxicity, Gab, EA data sets, the RoBERTa model performs better than Bert.
- Overall, the Roberta+Linear model trained with EA dataset is the best performing one.
- More adequate preprocessing of data (selection and classification of hashtags) could help the model achieve better results in the prediction phase.

Knowledge learned during this internship

- ❖ **Academic:**
  - ➢ The structure and use of Bert and RoBERTa model
  - ➢ Steps and operations of data preprocessing
  - ➢ Practical methods to prevent overfitting: weight-decay and dropout
  - ➢ How to evaluate the imbalanced class, etc..
- ❖ **Work and Daily:**
  - ➢ Professional vocabularies in text writing and oral expression
  - ➢ Regular work summary could help more efficient learning and task completion

Special thanks to my supervisor:
Isar Nejadgholi
and other colleagues for their
help during this internship!

Thank you!

Xuanyu Su
Master student from University of Ottawa
xsu072@uottawa.ca

Questions?