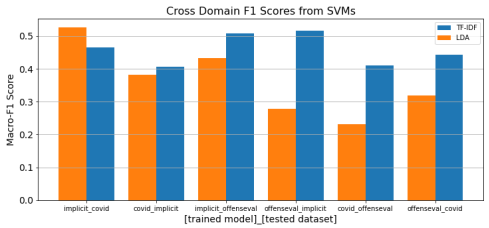
Apples-to-Apples: Comparing the Performance of Hate Speech Detection Models in Context

Seema Baddam, Richard Huang, Kai McKeever (CS 6471: Spring 2022)

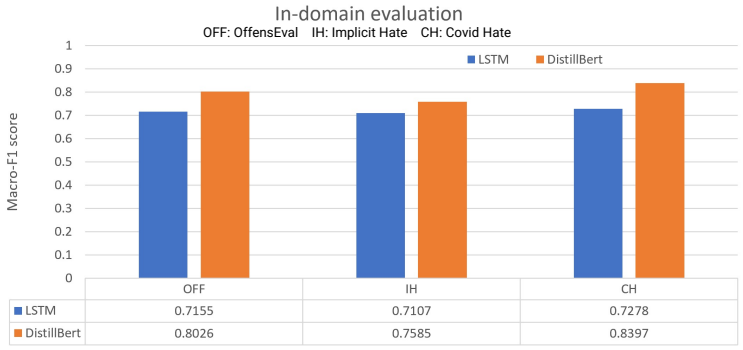
Github Repo

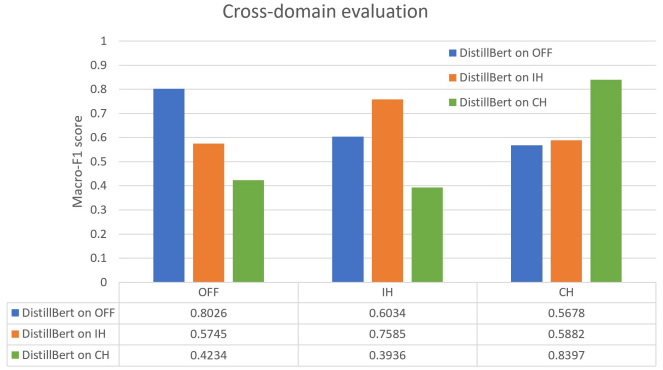
LDA vs TF-IDF as input features for SVM



LDA features underperform in most cross-domain tests and TF-IDF features also have poor performance.

Generalization evaluation on LSTM and BERT





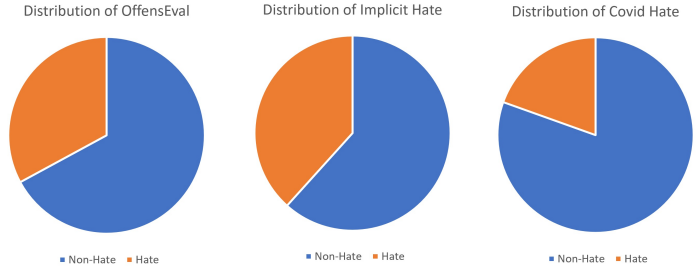
Problem/Research Question

How well can different hate speech detection models perform across varying topics in Twitter-based datasets?

Data analysis

Property	OffensEval	Implicit Hate	Covid Hate
Num. Documents	14,100	21,480	2,290
Avg. Document Size (chars)	127	89	173
Num. Labels (Non-hate/hate/others)	2	3	3

Table 1: Datasets Summary



Methods

- Topic Modeling with LDA and TF-IDF
- Training/Gridsearch pipeline for SVM, LSTM, BERT
- Cross-domain evaluation
  - Between SVM models
  - Between BERT models
- XAI analysis with Integrated Gradients on BERT
  - Qualitative study on the sentence inputs
  - Quantitative study on attribution scores

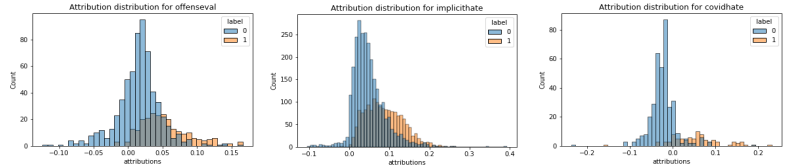
Interpretation of BERT's inferences with XAI

XAI algorithm: Integrated Gradients on DistillBert for OffensEval

Legend: Neutral Hate

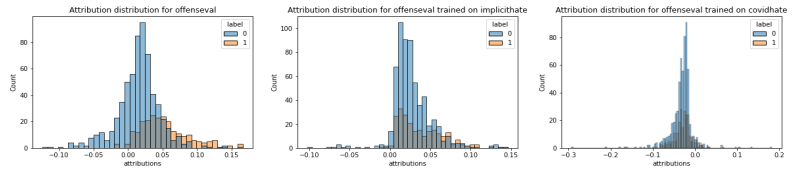
Original Index	True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
433	Hate	Hate (0.95)	Hate	1.95	[CLS] @ user damn i felt this shit . why you so loud lo ### [SEP]
730	Hate	Hate (0.95)	Hate	1.61	[CLS] and she has a pet ? ? ? fucking disgusting ur ### [SEP]
259	Hate	Hate (0.95)	Hate	1.50	[CLS] ! ! ! ! bitch i ' m fucking coming back ur ### [SEP]
406	Hate	Hate (0.95)	Hate	1.68	[CLS] alright let me get right with god bc mother nature is like fuck humans ur ### [SEP]

Distribution of attribution scores on all the datasets



F1: 80.26% F1: 75.85% F1: 83.97%

Distribution of attribution scores on OffensEval



OFF | F1: 80.26% IH | F1: 57.45% CH | F1: 42.34%

Takeaways & Future work

Takeaways:

- Neat and tidy topics hardly exist for hate speech
- LDA features can't substitute for other embeddings
- Larger neural network models perform better
- Data matters for explainability and accuracy

Future work:

- Resampling datasets for data imbalance
- Concatenating features for training