ToKen : Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection

Badr AlKhamissi, Srini Iyer, Faisal Ladhak, Ves Stoyanov ,Zornitsa Kozareva, Xian Li ,Pascale Fung Lambert Mathias, Asli Celikyilmaz, Mona Diab

Meta AI

Year of publication: May 2022

Number of citations:0

Overview

1. Introduction
2. Research Problem
3. Datasets
4. Task Decomposition & Knowledge Infusion
5. Results
6. Conclusion

Introduction

What is hate speech?

We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. [1]

- Facebook –

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. [2]

- Twitter -

"We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran Status"[3]

- YouTube -

[1] "Community standards." [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

[2] "twitters policy on hate help." [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] "Hate speech policy - ful conduct | twitter youtube help." [Online]. Available: https://support.google. com/youtube/answer/2801939?hl$=$en

Introduction cont.

- Growth of social-media usage, raises a platform to a new kind of social dilemma namely cyberbullying
- HS detection is still a challenge for the research community and policy makers as humans find loopholes to trick those algorithms [5]
- Challenge of detecting hate speech within online user communication due to its vast scope and the complexity
- Secure the freedom of speech [7]

[5] A. Oboler, “Solving antisemitic hate speech in social media through a global approach to local action,” in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2

[6] “Twitter hate accounts targeting meghan and harry, duke and duchess of sussex,” Oct 2021. [Online]. Available: https://www.msn.com/en-us/news/politics/ organized-campaign-targeted-harry-and-meghan-on-twitter-report/ar-AAQ180P

[7] T. M. Massaro, “Equality and freedom of expression: The hate speech dilemma,” Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

Introduction - Research Objective

- Benchmark algorithm for detecting Hate Speech using the semantic oppositeness metric.

1. Comprehensive comparative analysis on different datasets on English HS using existing models generated for languages other than English.
2. Comprehensive comparative analysis on different datasets on English HS using different word embedding techniques.
3. Comprehensive comparative analysis for the same model on different HS datasets available.
4. Determining the gap between normal domain embedding an HS domain embedding.
5. Comprehensive comparative analysis on datasets for oppositeness measure.

Research Problem

- Hate speech detection is complex
- Difficulty in collecting a large-scale hate speech annotated dataset.

Contributions:

- Frame this problem as a **few-shot learning task** and show significant gains with decomposing the task into its "constituent" parts.
- Improve performance by infusing knowledge from reasoning datasets. Eg: ATOMIC$_{20}$

Datasets

- SBIC - Social Bias Inference Corpus

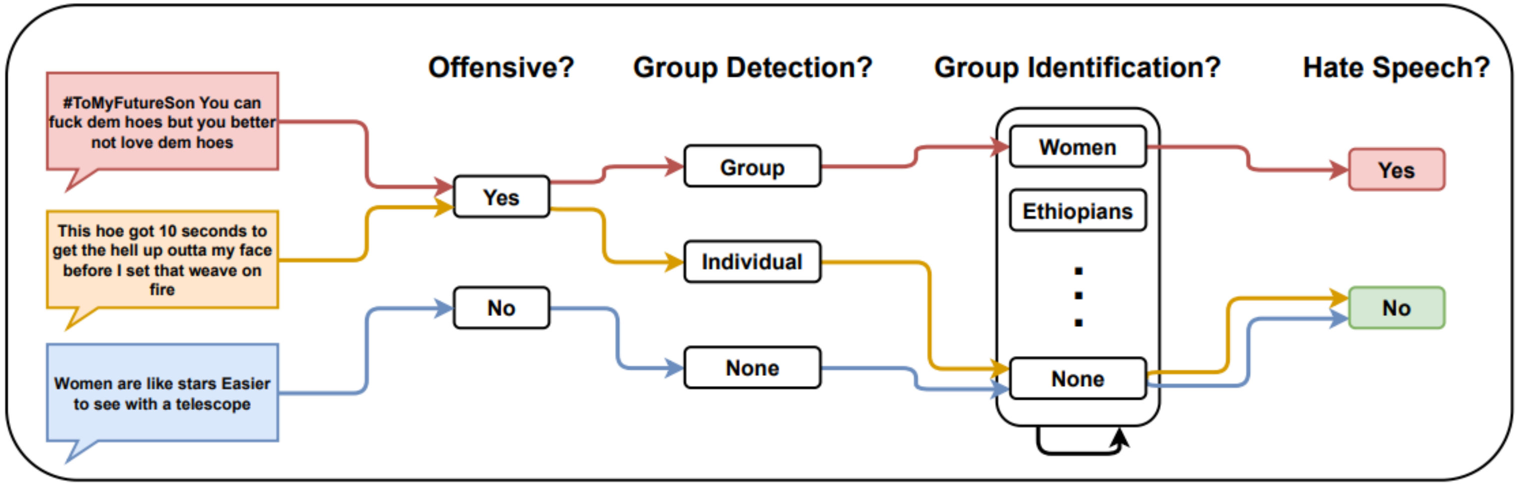
Reddit, Twitter

compiled by researchers at Stanford, the University of Washington, and the Allen Institute for Artificial Intelligence

annotations for the offensiveness, targeted group

Stereotype is being implied by the post

Does not have explicit labels for Hate/ Non-hate

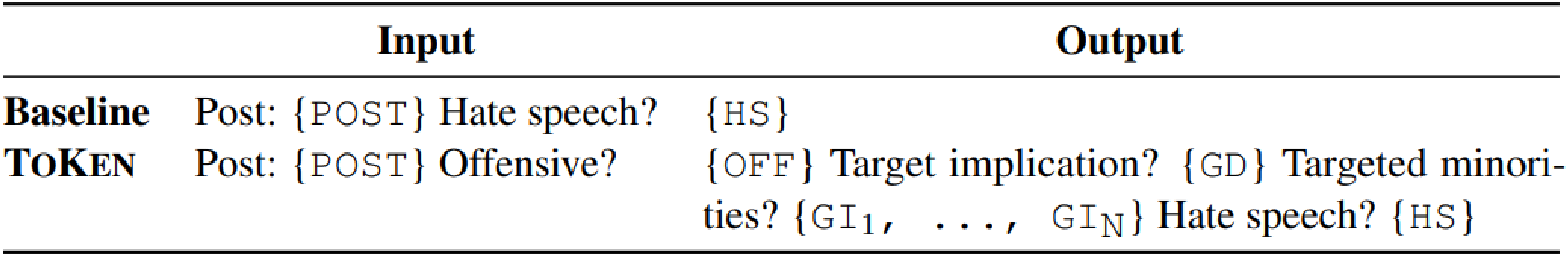

Offensive?
Group Detection?
Group Identification?
Hate Speech?
#ToMyFutureSon You can fuck dem hoes but you better not love dem hoes
This hoe got 10 seconds to get the hell up outta my face before I set that weave on fire
Women are like stars Easier to see with a telescope
Yes
No
Group
Individual
None
Women
Ethiopians
None
Yes
No

Datasets

- ATOMIC20 - a commonsense knowledge graph containing 1.33M inferential knowledge tuples in textual format
- StereoSet
  - developed to measure stereotype bias in language modeling
  - contains 17k sentences that measure biases across four different domains: gender, profession, race and religion
  - Used a subset of StereoSet

  - HateXplain - includes posts from Twitter and Gab along with the HS labels (i.e. hate, offensive or normal)
  - HS18 - posts on Stormfront
  - Ethos - YouTube and Reddit data with binary labels for HS

Task Decomposition



	Input	Output
Baseline	Post: {POST} Hate speech?	{HS}
TOKEN	Post: {POST} Offensive?	{OFF} Target implication? {GD} Targeted minorities? {GI_1, ..., GI_N} Hate speech? {HS}

- Linearization scheme for the Baseline and the TOKEN models.
- Given the post, the Baseline predicts whether it is HS or not; whereas the task-decomposed model does the prediction for Offensiveness, Group Detection and Group Identification before predicting the HS label. HS and OFF are binary labels (i.e. either Yes or No).
- GD can be one of { Group , Individual , None }. Finally, GII is a group identity (e.g. Women).

Knowledge Infusion

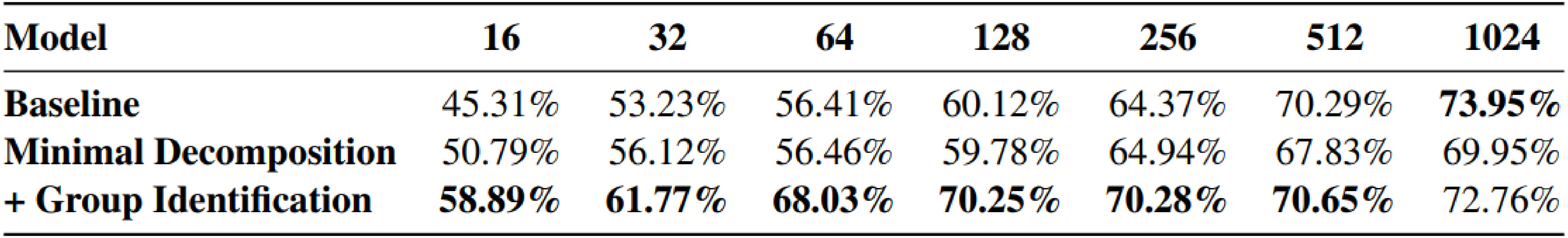
COMET

- Fine tune BART on the ATOMIC20 20 dataset, where each tuple is converted into a natural language statement using human readable templates.
- Resulting COMET model achieves similar performance.

StereoSet

- Finetune both BART and our trained COMET on stereotypical sentences from the StereoSet dataset.

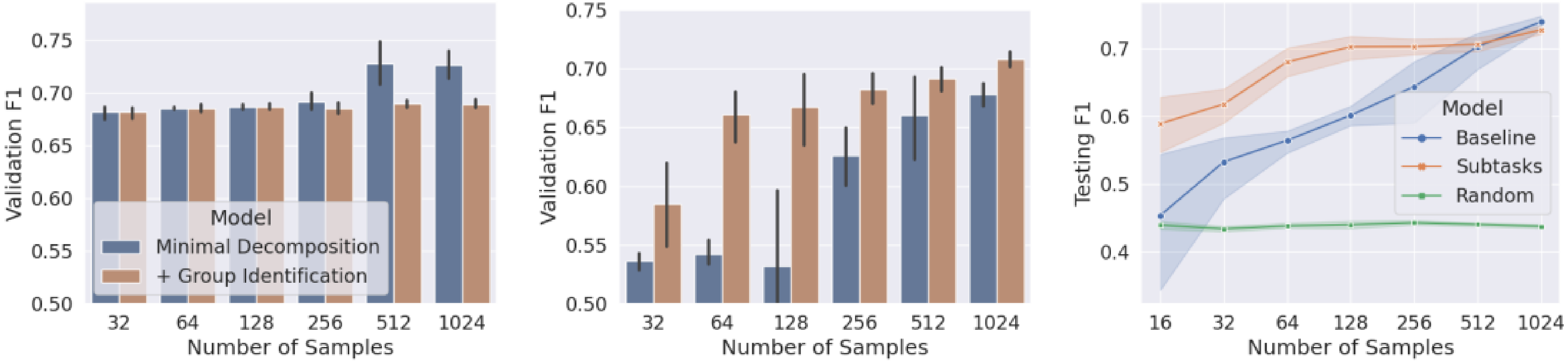
Results - Task Decomposed Model



Model	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
Baseline	45.31%	53.23%	56.41%	60.12%	64.37%	70.29%	**73.95%**
Minimal Decomposition	50.79%	56.12%	56.46%	59.78%	64.94%	67.83%	69.95%
+ Group Identification	**58.89%**	**61.77%**	**68.03%**	**70.25%**	**70.28%**	**70.65%**	72.76%

- Results of the Task Decomposed Model on the Hate Speech detection task.
- Baseline predicts only whether the input post is HS or not.
- Minimal Decomposition additionally predicts whether the post is offensive or not and the group detection.
- + Group Identification additionally predicts the minority groups the post is targeting if any.

Results - Model performance



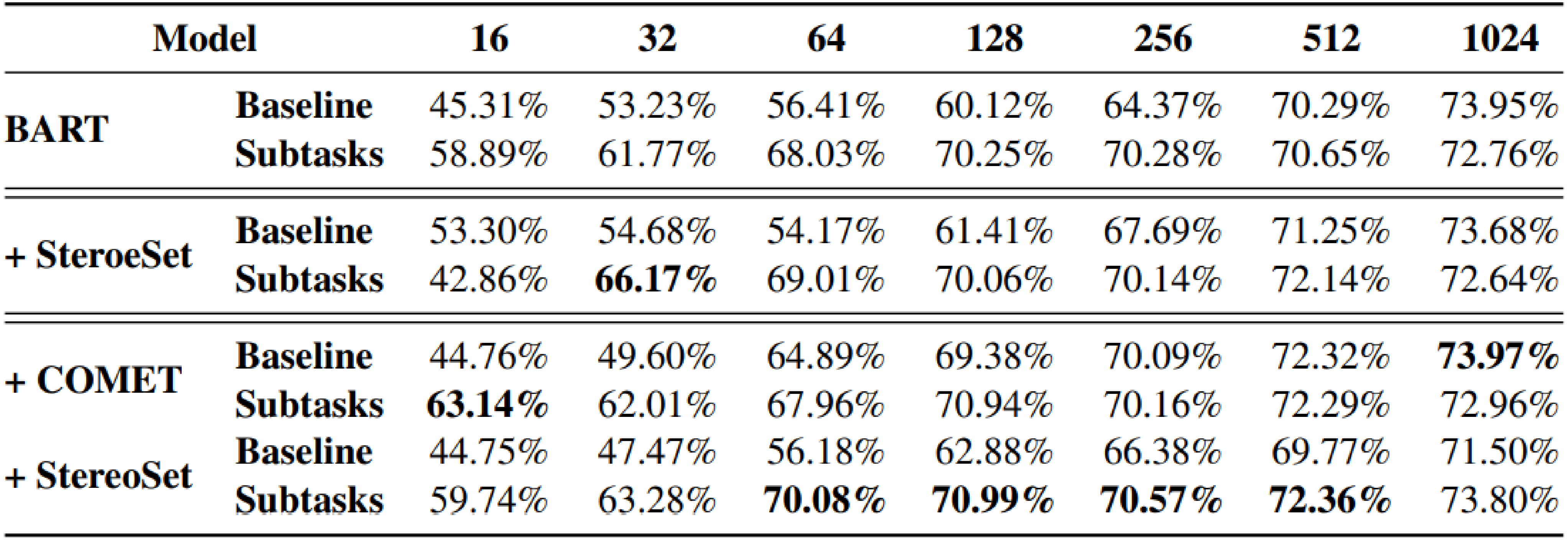
(a) Offensiveness Performance

(b) Group Detection Performance

(c) Hate Speech Performance

(a) The validation F1-score of the Offensiveness subtask for the minimal decomposition and task decomposed models.
(b) The validation F1-score for the Group Detection subtask. It can be seen that adding the Group Identification subtask improves the performance dramatically.
(c) (c) Testing Performance of Baseline vs the Task Decomposed model. Random performance plotted for reference.

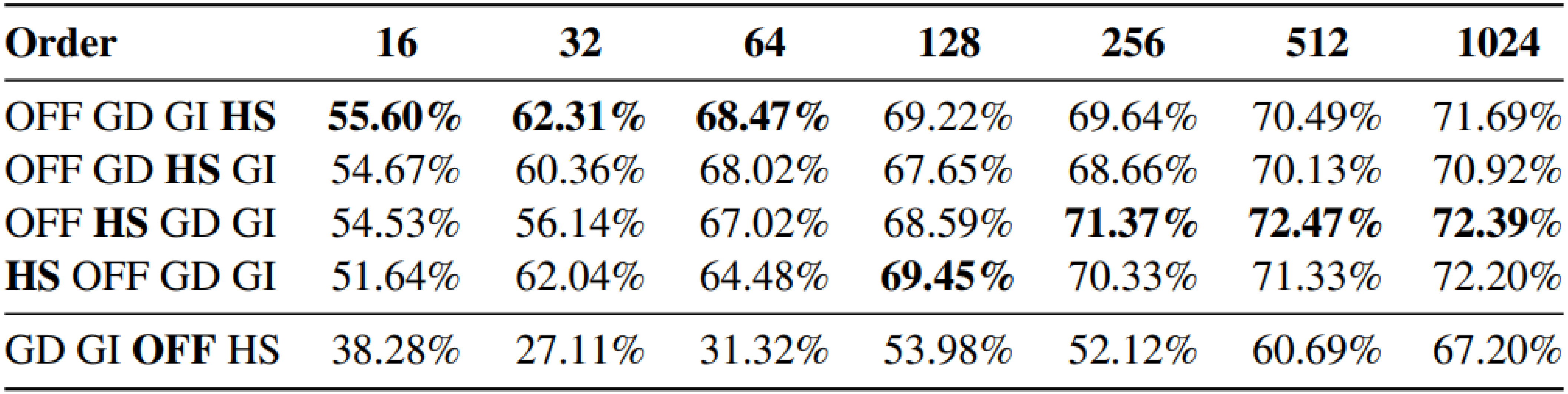
Results – Knowledge Infusion



Model		16	32	64	128	256	512	1024
BART	Baseline	45.31%	53.23%	56.41%	60.12%	64.37%	70.29%	73.95%
	Subtasks	58.89%	61.77%	68.03%	70.25%	70.28%	70.65%	72.76%
+ SteroeSet	Baseline	53.30%	54.68%	54.17%	61.41%	67.69%	71.25%	73.68%
	Subtasks	42.86%	**66.17%**	69.01%	70.06%	70.14%	72.14%	72.64%
+ COMET	Baseline	44.76%	49.60%	64.89%	69.38%	70.09%	72.32%	**73.97%**
	Subtasks	**63.14%**	62.01%	67.96%	70.94%	70.16%	72.29%	72.96%
+ StereoSet	Baseline	44.75%	47.47%	56.18%	62.88%	66.38%	69.77%	71.50%
	Subtasks	59.74%	63.28%	**70.08%**	**70.99%**	**70.57%**	**72.36%**	73.80%

- Binary F1-score on the SBIC testing set for the HS detection task using models with different degrees of knowledge infusion.
- In each row, compared the corresponding baseline and subtasks models.
- Results in bold show the best overall model in each few-shot setting.

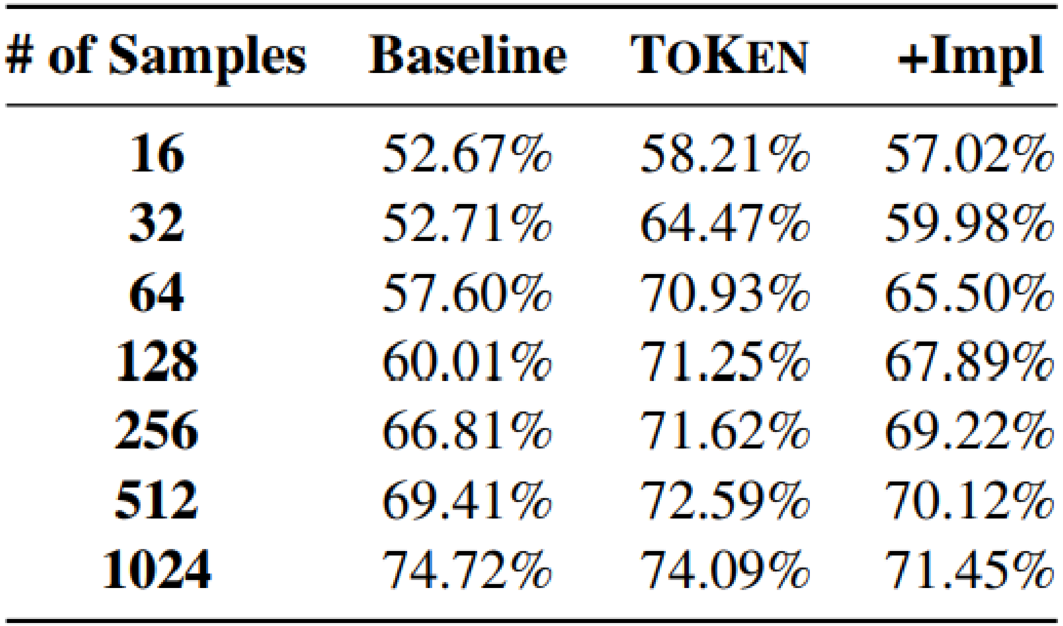
Results – Validation performance



Order	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
OFF GD GI **HS**	**55.60%**	**62.31%**	**68.47%**	69.22%	69.64%	70.49%	71.69%
OFF GD **HS** GI	54.67%	60.36%	68.02%	67.65%	68.66%	70.13%	70.92%
OFF **HS** GD GI	54.53%	56.14%	67.02%	68.59%	**71.37%**	**72.47%**	**72.39%**
HS OFF GD GI	51.64%	62.04%	64.48%	**69.45%**	70.33%	71.33%	72.20%
GD GI **OFF** HS	38.28%	27.11%	31.32%	53.98%	52.12%	60.69%	67.20%

- The validation performance of the best model on the HS detection task as a function of the position of the HS label in the sequence of subtasks across different number of training samples.

Results - Implications Results



# of Samples	Baseline	TOKEN	+Impl
16	52.67%	58.21%	57.02%
32	52.71%	64.47%	59.98%
64	57.60%	70.93%	65.50%
128	60.01%	71.25%	67.89%
256	66.81%	71.62%	69.22%
512	69.41%	72.59%	70.12%
1024	74.72%	74.09%	71.45%

- The HS detection performance of the Baseline in comparison with the Subtasks models before and after adding the implication to the subtasks across 5 runs for a given training set size.

Conclusion

TOKEN models generalize better to three out-of-distribution datasets in the few-shot setting.

Task decomposition not only improves the performance, but also allows for fine-grained inspection of the model’s behavior.

TOKEN models generalize better to three out-of-distribution datasets in the few-shot setting and significantly more robust to training setups.

Reference

[1] “Community standards.” [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

[2] “twitters policy on hate help.” [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] “Hate speech policy - ful conduct | twitter youtube help.” [Online]. Available: https://support.google.com/youtube/answer/2801939?hl$=$en

[4] AlKhamissi, B., Ladhak, F., Iyer, S., Stoyanov, V., Kozareva, Z., Li, X., Fung, P., Mathias, L., Celikyilmaz, A. and Diab, M., 2022. ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection. arXiv preprint arXiv:2205.12495.

[5] A. Oboler, “Solving antisemitic hate speech in social media through a global approach to local action,” in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2

[4] “Twitter hate accounts targeting meghan and harry, duke and duchess of sussex,” Oct 2021. [Online]. Available: https://www.msn.com/en-us/news/politics/ organized-campaign-targeted-harry-and-meghan-on-twitter-report/ar-AAQ180P

[7] T. M. Massaro, “Equality and freedom of expression: The hate speech dilemma,” Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

Thank you!

Any Questions?

You can find me at: dinuja.21@cse.mrt.ac.lk