Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection

Camilla Casula, Sebastiano Vecellio Salto, Alan Ramponi, Sara Tonelli

{ccasula, svecelliosalto, alramponi, satonelli}@fbk.eu

FONDAZIONE BRUNO KESSLER

Motivation

> **Synthetic data** is now widespread across a variety of applications, including very sensitive ones

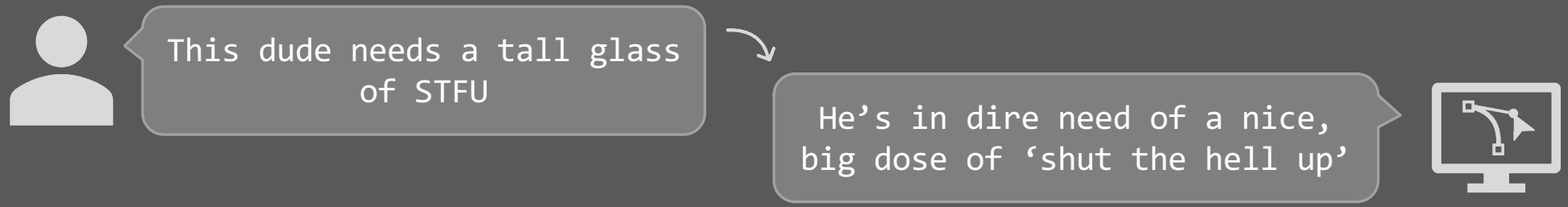
> It can mitigate some issues, e.g. with **privacy**, **identity group representation**, and **negative psychological impact on annotators**

> Previous work reports **mixed results** using synthetic data, especially on **subjective** tasks

> **Goal**: Exploring advantages and risks of synthetic data (for privacy reasons, to improve performance, to reduce human annotation, etc.)

> We work on English (high-resource) → we assume some annotated data is already available

> We *paraphrase* existing data instead of creating data 'from scratch'



Contributions

- An in-depth qualitative analysis of synthetic data
  - **Extrinsic** → we train models on synthetic data produced by 3 LLMs, and compare them with models trained on gold data in and out of distribution
  - **Intrinsic** → we manually annotate 3,500 synthetic examples, evaluating:
    - *Realism*
    - Preservation of *(non) hatefulness*
    - Preservation of *target identity groups*

Creation of Synthetic Data

> We start with the **Measuring Hate Speech** corpus (MHS; Kennedy et al., 2020)

> We create a 'parallel' version of the dataset through paraphrasing

Paraphrase this text: "{text}"

Paraphrased text: "

> We use *Llama-2 Chat 7B*, *Mistral Instruct 7B*, and *Mixtral Instruct 8x7B*

Filtering of Synthetic Data

> We remove:

> Synthetic texts that are too similar to the originals

wow that's great → Wow, that's great!

> Synthetic texts for which a classifier predicts a different *hate* label from that of the original text used for paraphrasing

→ **Classifier filtering**

Extrinsic Evaluation

> *RoBERTa Large* trained on synthetic data *only*

			ID	OOD	
		Test data →	MHS	MDA	HateCheck
		n(train) ↓	M-F_1	M-F_1	M-F_1
Model trained on **gold data**		30k	**.811**	.507	.386
Gen. model	Filter				
Llama-2 Chat 7B	No	28k	.769	.675	.603
	Yes	20k	.805	.539	.346
Mistral 7B Instruct	No	29k	.772	.684	**.665**
	Yes	22k	.808	.526	.371
Mixtral 8x7B Instruct	No	29k	.754	**.687**	**.665**
	Yes	22k	.802	.525	.364

Extrinsic Evaluation

> *RoBERTa Large* trained on synthetic data *only*

			ID	OOD	
		Test data →	MHS	MDA	HateCheck
		n(train) ↓	M-F_1	M-F_1	M-F_1
Model trained on **gold data**		30k	**.811**	.507	.386
Gen. model	Filter				
Llama-2 Chat 7B	No	28k	.769	.675	.603
	Yes	20k	.805	.539	.346
Mistral 7B Instruct	No	29k	.772	.684	**.665**
	Yes	22k	.808	.526	.371
Mixtral 8x7B Instruct	No	29k	.754	**.687**	**.665**
	Yes	22k	.802	.525	.364

Extrinsic Evaluation

> *RoBERTa Large* trained on synthetic data *only*

			ID	OOD	
		Test data →	MHS	MDA	HateCheck
		n(train) ↓	M-F_1	M-F_1	M-F_1
Model trained on **gold data**		30k	**.811**	.507	.386
Gen. model	Filter				
Llama-2 Chat 7B	No	28k	.769	.675	.603
	Yes	20k	.805	.539	.346
Mistral 7B Instruct	No	29k	.772	.684	**.665**
	Yes	22k	.808	.526	.371
Mixtral 8x7B Instruct	No	29k	.754	**.687**	**.665**
	Yes	22k	.802	.525	.364

Intrinsic Analysis: *Realism*

> We provided 2 annotators with 500 texts that are a mix of gold and synthetic, and asked them to tell them apart

> Annotators had an accuracy of 87%, 90%, and 92% in identifying texts generated with Llama-2 Chat, Mistral, and Mixtral respectively

Expert eyes can still spot LLM-written text
even if it is grammatical and plausible

Intrinsic Analysis: *Realism*

> We provided 2 annotators with 500 texts that are a mix of gold and synthetic, and asked them to tell them apart

> Annotators had an accuracy of 87%, 90%, and 92% in identifying texts generated with Llama-2 Chat, Mistral, and Mixtral respectively

Expert eyes can still spot LLM-written text even if it is grammatical and plausible

Kindly halt this conduct characterized by the blending of unconventional gender identities and feminist ideologies

Intrinsic Analysis: *Realism*

> We provided 2 annotators with 500 texts that are a mix of gold and synthetic, and asked them to tell them apart

> Annotators had an accuracy of 87%, 90%, and 92% in identifying texts generated with Llama-2 Chat, Mistral, and Mixtral respectively

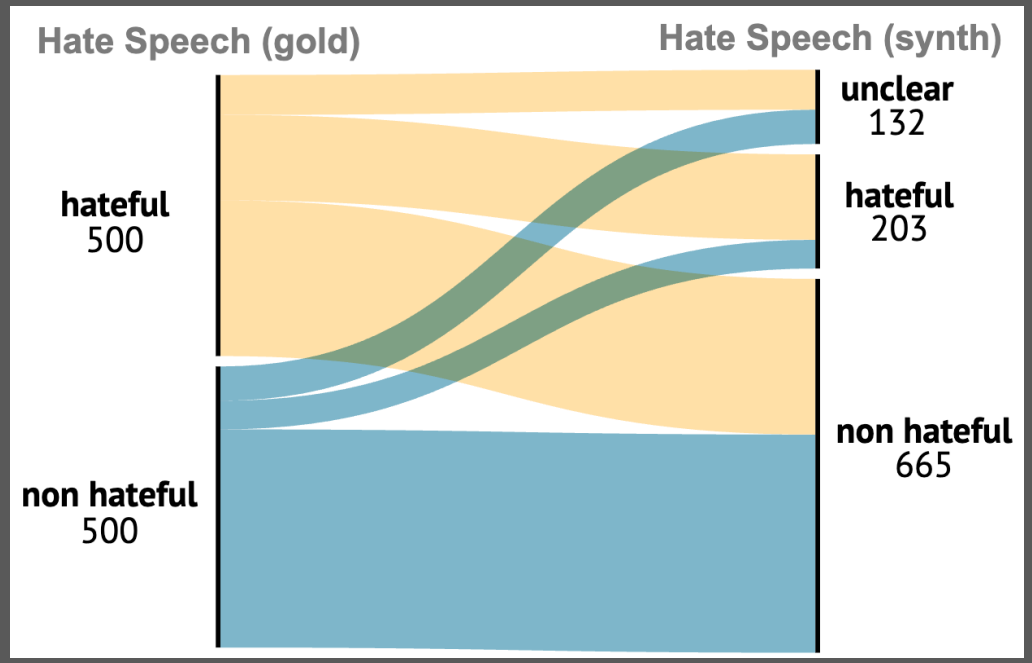
Expert eyes can still spot LLM-written text even if it is grammatical and plausible

Kindly halt this conduct characterized by the blending of unconventional gender identities and feminist ideologies

Paraphrase of:
*please stop this queer feminist bullsh*t*

Intrinsic Analysis: *Label Preservation*

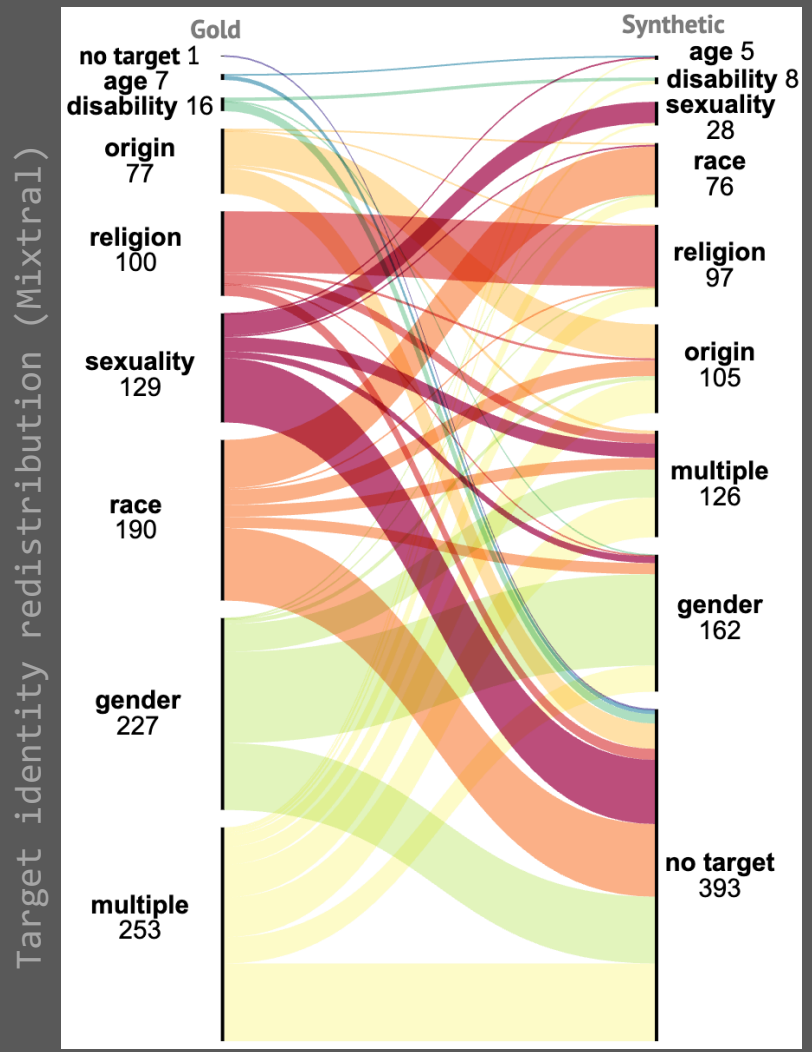
> Annotators labeled 3,000 synthetic texts (1,000 per model) using the same guidelines as the original dataset we paraphrased



Label redistribution (Mixtral)

Paraphrased synthetic texts won't necessarily maintain the same class distribution as the gold data

Intrinsic Analysis: *Identity Preservation*



Paraphrased synthetic texts won't necessarily share the same representation of identity groups as the gold data

Conclusion

> Based on classifier performance, synthetic data can be almost as good as gold data, even better in out-of-distribution scenarios

> A closer look reveals that the preservation of key features in synthetic data should not always be taken for granted

> Although performance shows synthetic data to be potentially useful, it can hide risks we may often be unaware of

> Thank you!

> Questions?

Bonus: Most informative tokens in gold and synthetic data (calculated with VARIATIONIST!)

↪ check out our poster this afternoon!

Target	Subset	Most informative tokens (hateful class)
Age	Gold	f*ck, *ss, b*tch, f*cking, , sh*t, p*ssy, racist, c*nt, kids
	Synthetic	individuals, individual, woman, children, mother, person, people
Disability	Gold	r*tarded, r*tard, f*cking, f*ck, sh*t, *ss, b*tch, r*tards
	Synthetic	individuals, person, foolish, individual, intellectually
Gender	Gold	b*tch, f*ck, *ss, f*cking, c*nt, b*tches, sh*t, p*ssy, wh*re, sl*t
	Synthetic	woman, women, person, individuals, individual, promiscuous
Origin	Gold	f*ck, f*cking, country, sh*t, people, america, *ss, white, b*tch
	Synthetic	individuals, country, people, america, person, individual, return
Race	Gold	n*gga, n*ggas, f*ck, *ss, f*cking, white, sh*t, b*tch, n*gger, 😂
	Synthetic	individuals, people, person, white, individual, racist, black
Religion	Gold	f*ck, jews, f*cking, sh*t, people, jew, muslim, muslims, white
	Synthetic	individuals, people, jewish, individual, jews, muslim, muslims
Sexuality	Gold	f*ggot, f*ck, f*cking, *ss, f*g, sh*t, f*ggots, gay, b*tch, d*ck
	Synthetic	homosexual, person, individuals, gay, individual, term, behavior