Generating Counter Narratives against Online Hate Speech: Data and Strategies

Marco Guerini
Fondazione Bruno Kessler

Disclaimer

This presentation contains examples of offensive language; they do not represent the views of the authors. Islamophobia is the topic of the EU project **Hatemeter** under which this study has been conducted.

The Context

Hate Speech on Social Media

Hate Speech (HS) refers to “expressions that attack or diminish, that incite violence or hate against groups, based on specific characteristics such as religion, ethnicity, sexual orientation, gender or other”.

With the rapid growth of social media platforms, abusive and offensive language can spread quickly and is difficult to monitor.

Hate Speech on Social Media

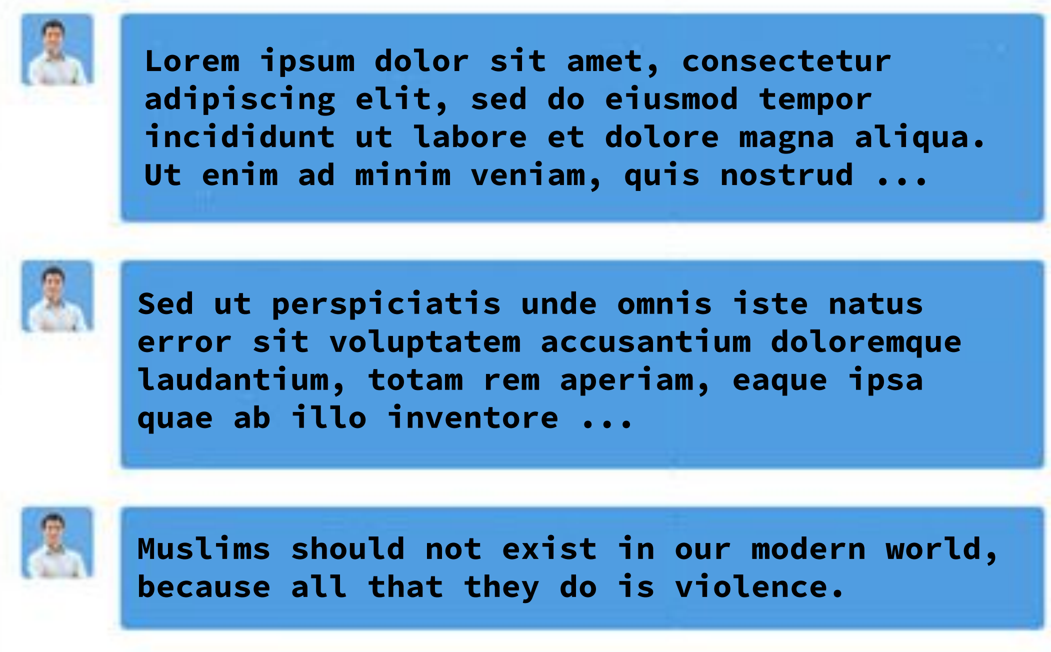
HS has real-life consequences: it can lead to depression or suicide, promote the use of violence, encourage discrimination, and increase societal divisions.

An unprecedented effort to provide adequate responses in terms of laws and policies to hate content on social media platforms.

HS Countering - Standard Approaches

Standard policies are based on **identify-and-sanction** strategies.

- Content deletion
- User suspension
- Shadowbanning

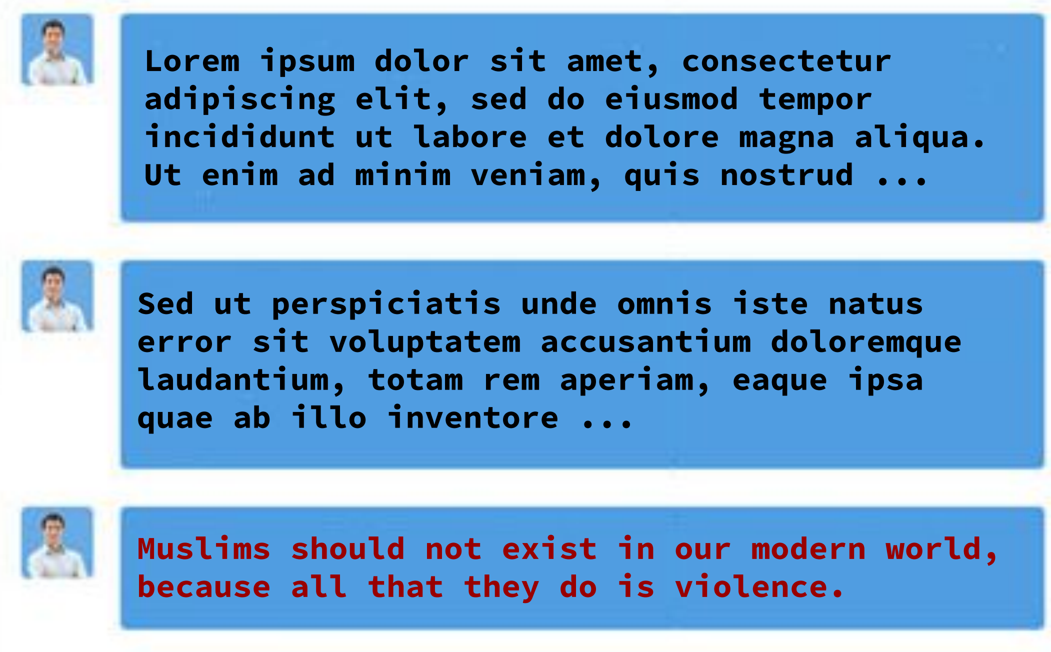


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud ...

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ...

Muslims should not exist in our modern world, because all that they do is violence.

First step: hate identification

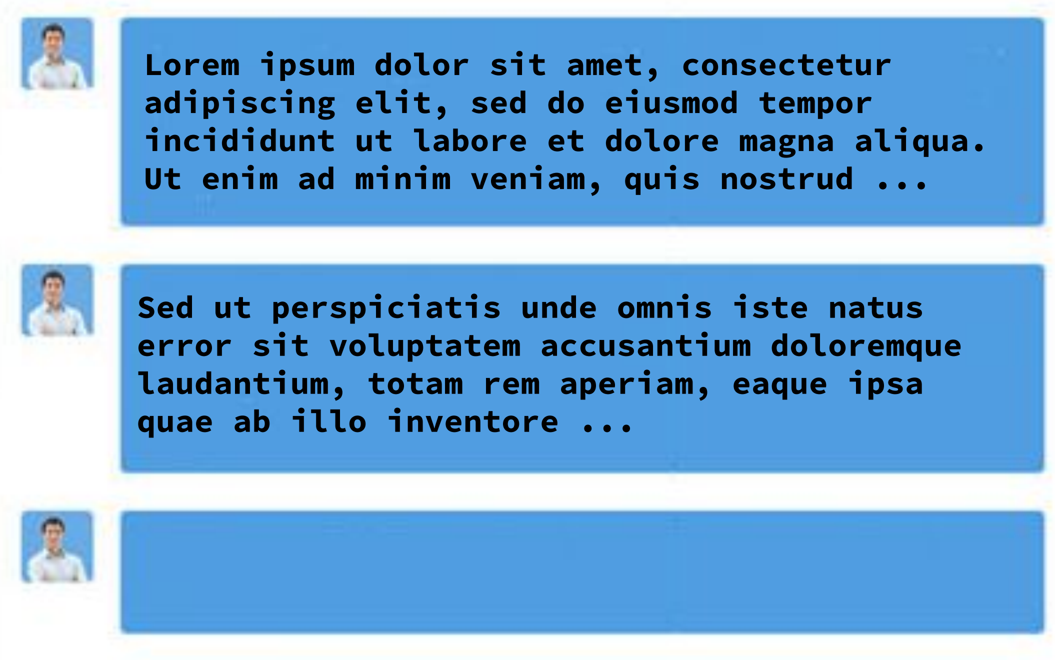


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud ...

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ...

Muslims should not exist in our modern world, because all that they do is violence.

Content Removal

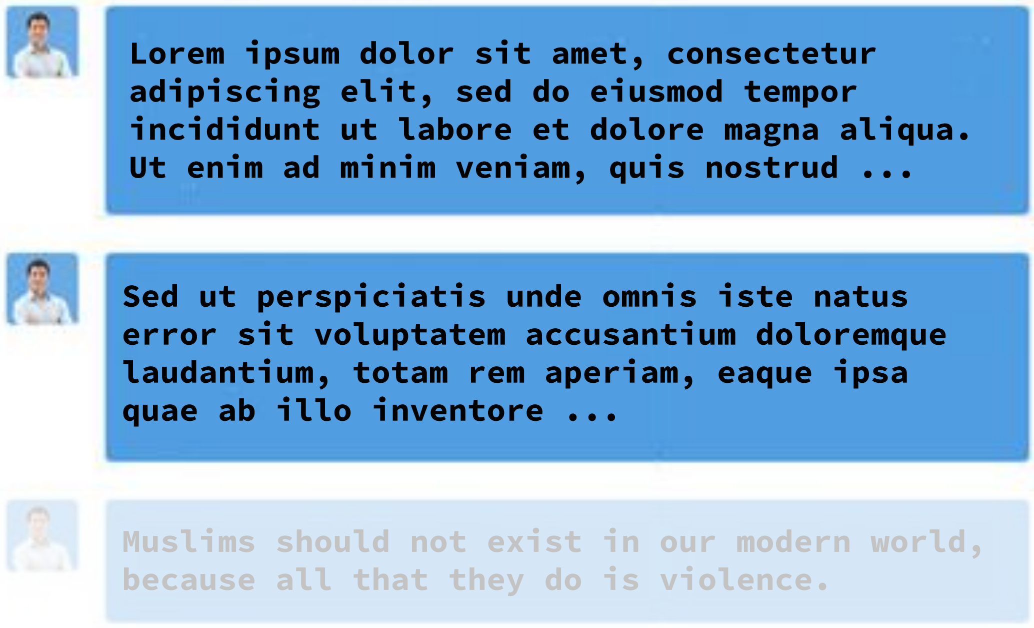


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud ...

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ...

User Suspension

Shadowbanning



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud ...

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ...

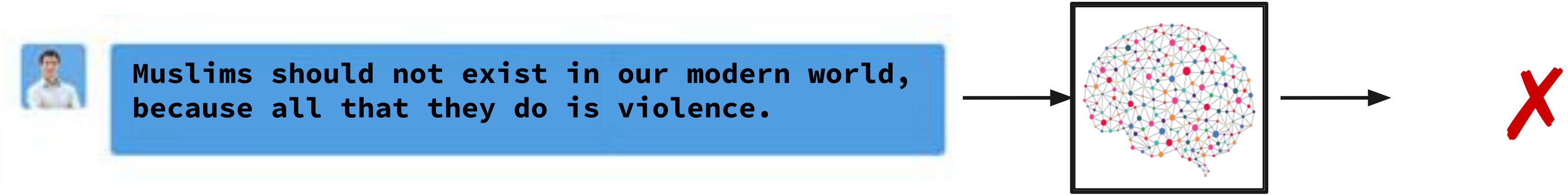
Muslims should not exist in our modern world, because all that they do is violence.

HS Countering - Contribution of AI

UP TO NOW: systems that are able to classify content (i.e. to tell if it is hateful or not)

HS Countering - Contribution of AI

UP TO NOW: systems that are able to classify content (i.e. to tell if it is hateful or not)

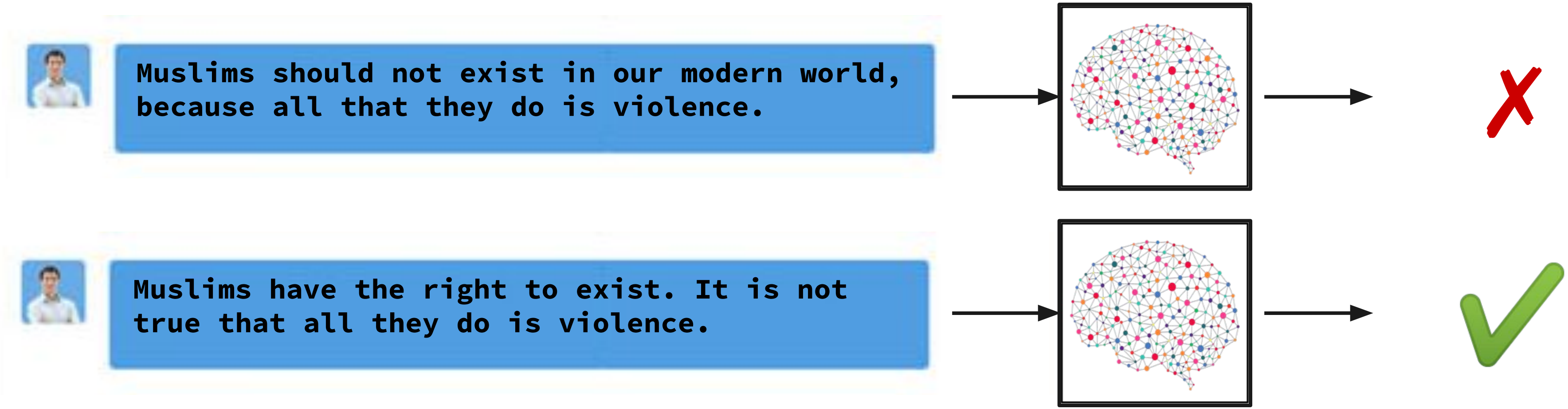


Muslims should not exist in our modern world, because all that they do is violence.

✗

HS Countering - Contribution of AI

UP TO NOW: systems that are able to classify content (i.e. to tell if it is hateful or not)



Muslims should not exist in our modern world, because all that they do is violence. ✗

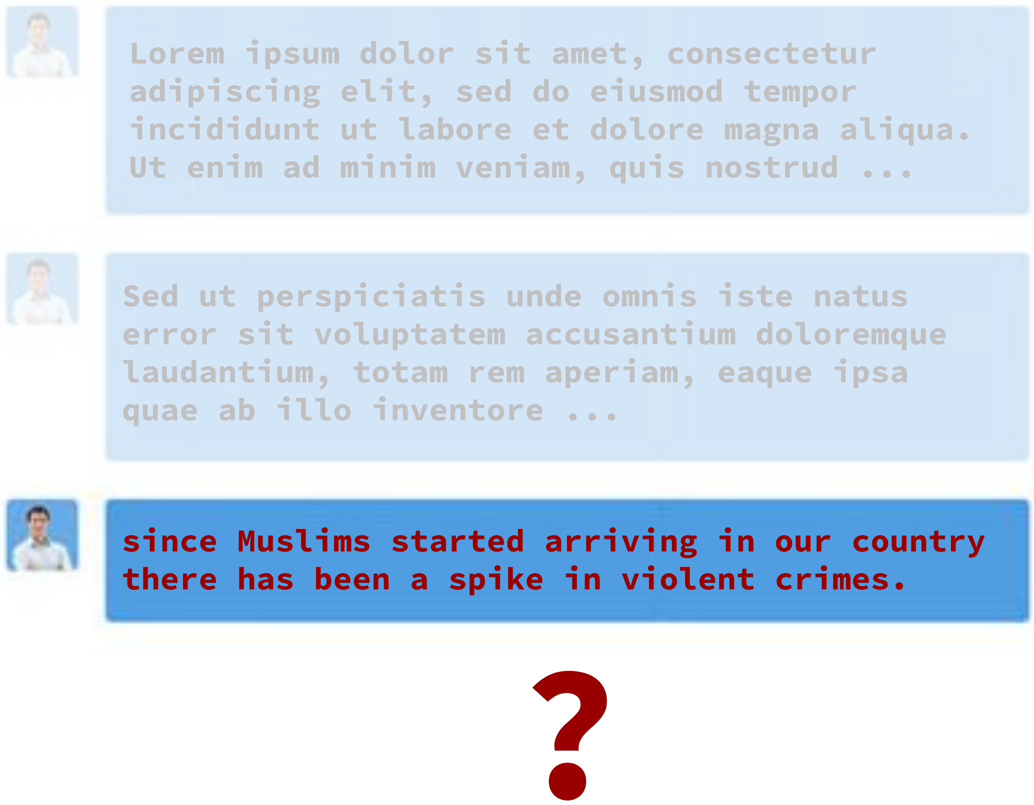
Muslims have the right to exist. It is not true that all they do is violence. ✓

Standard Approaches - Limitations

These approaches can be charged with censorship and overblocking. They can hinder Freedom of Speech.

They cannot be applied to **dangerous speech** - i.e. content that is stirring up hatred and divisiveness but does not fall into a formal definition of HS.

True - but spurious correlation



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud ...

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ...

since Muslims started arriving in our country there has been a spike in violent crimes.

?

The Idea

HS Countering - Counter Narratives

One alternative strategy - little attention so far by the AI community - is to actually oppose hate content with counter-narratives.

Counter Narrative: **direct intervention in the discussion** to withstand hate messages using a non-aggressive textual response that offers feedback through fact-bound arguments.

Counter Narratives Advantages

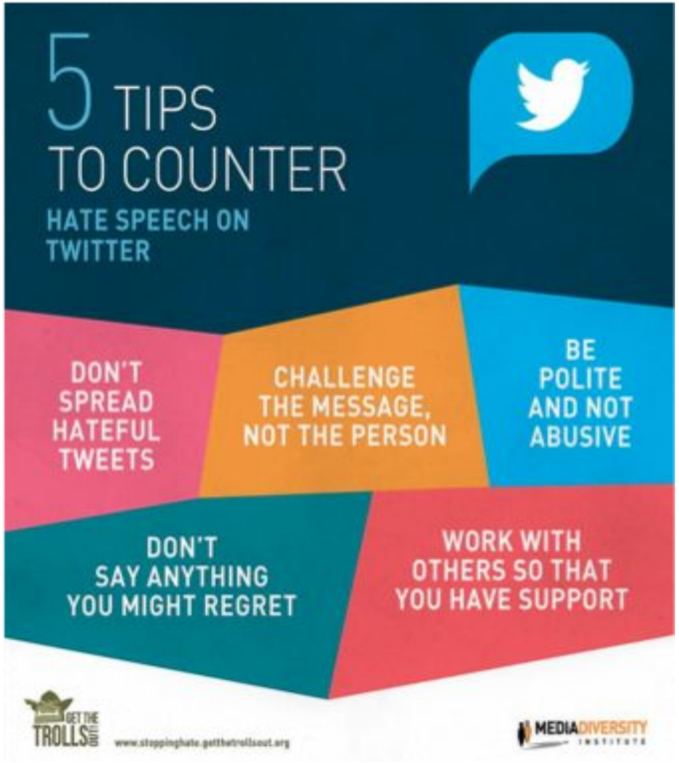
Preserve the right to freedom of speech

Counter stereotypes with credible evidence

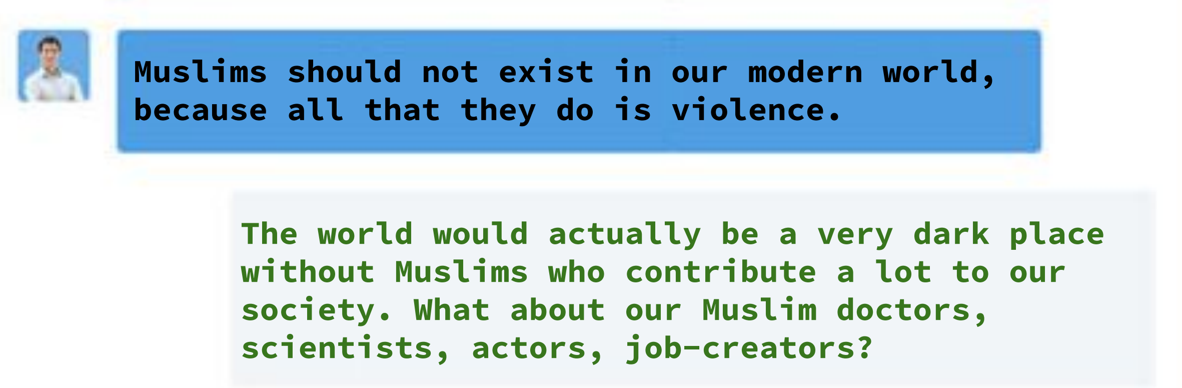
Alter the viewpoints of haters and bystanders

Encourage mutual understanding

Help de-escalating the conversation


5 TIPS TO COUNTER
HATE SPEECH ON TWITTER
DON'T SPREAD HATEFUL TWEETS
CHALLENGE THE MESSAGE, NOT THE PERSON
BE POLITE AND NOT ABUSIVE
DON'T SAY ANYTHING YOU MIGHT REGRET
WORK WITH OTHERS SO THAT YOU HAVE SUPPORT
GET THE TROLLS OUT
www.stoppinghate.getthetrollsout.org
MEDIA DIVERSITY INSTITUTE

Counter Narrative



Muslims should not exist in our modern world, because all that they do is violence.

The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

HS Countering - Actors

There are NGOs training volunteers/operators to intervene with CNs.

ITALIA
AMNESTY
INTERNATIONAL

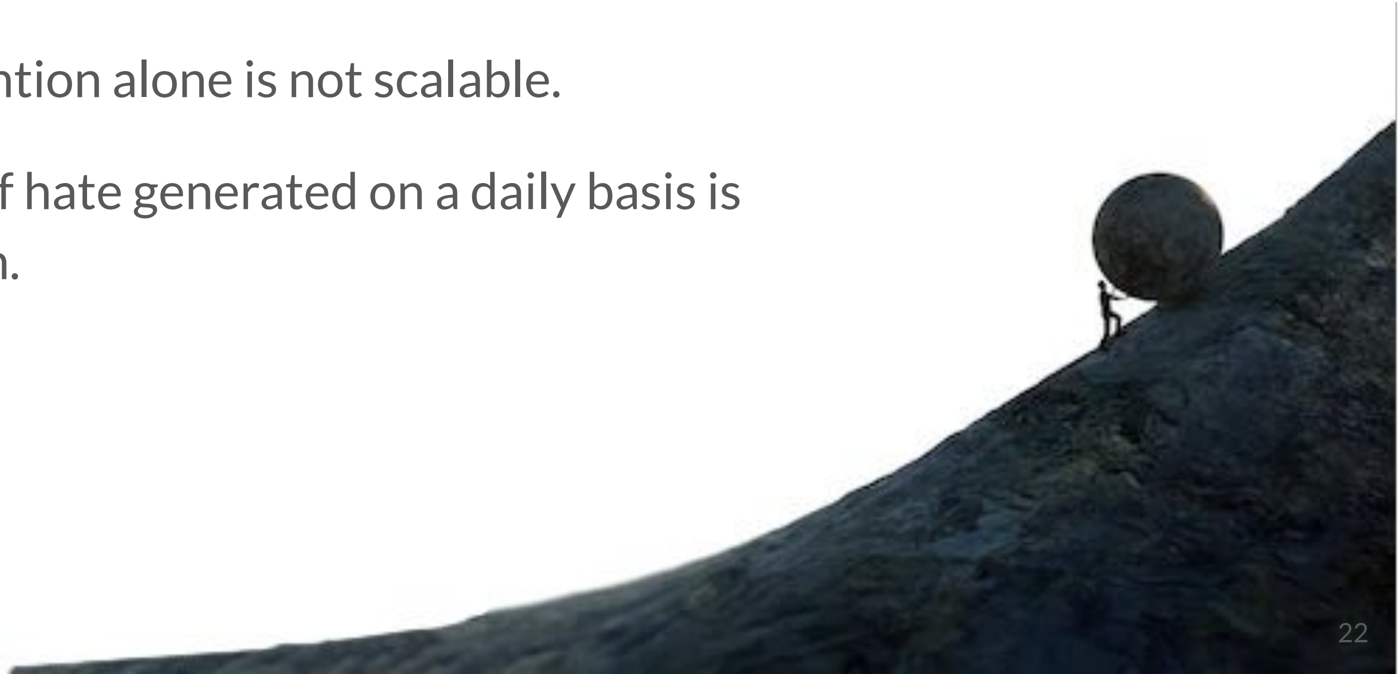
STOP HATE UK

CCIF
COLLECTIF CONTRE
L'ISLAMOPHOBIE
EN FRANCE

Limitations of Counter Narratives

Manual intervention alone is not scalable.

Sheer amount of hate generated on a daily basis is simply too much.



CN suggestion Tools

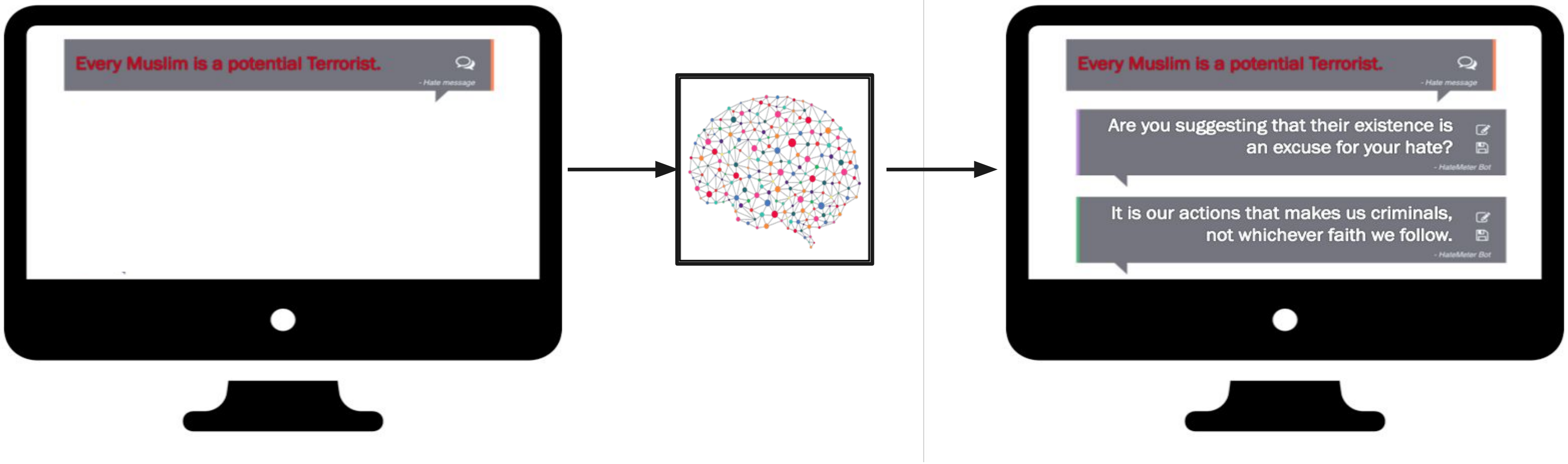
Computer Supported - Human activity

Tools that assist NGO operators in fighting HS.

Partial Automation of CN writing through NLG suggestions.

Drastic reduction in time needed to compose a CN.

Suggestion Tool



Every Muslim is a potential Terrorist.

- Hate message

Every Muslim is a potential Terrorist.

- Hate message

Are you suggesting that their existence is an excuse for your hate?

- HateMeter Bot

It is our actions that makes us criminals, not whichever faith we follow.

- HateMeter Bot

Building the Tools

- ALGORITHMS
- DATA

1. Algorithms

Algorithms for CN generation

1. Information Retrieval
2. Generation using Neural Networks
3. Generation using pre-trained Generative Language Models
4. Generation using External Knowledge

Information Retrieval

YL Chung, SS Tekiroglu, S Tonelli, M Guerini **Empowering NGOs in Countering Online Hate Messages: An Evaluation Study.** IEEE Transactions on Affective Computing (submitted)

Information Retrieval

Build a term document matrix starting from your database

Every Muslim is a potential terrorist.	Every Muslim is also a potential peacemaker, doctor, philanthropist... What’s your point?
The veil is contrary to secularism.	On the contrary, secularism allows every citizen to freely profess his faith.
…	…

Information Retrieval

Build a term document matrix starting from your database

DOC	Muslim	Terrorist	Secularism	Potential	Contrary
DC_1	2	1	-	1	-
DC_2	-	-	2	-	1
…	…	…	…	…	…

How it works

Muslim	Terrorist	Contrary	Secularism
2	**1**	3	-
-	-	-	1
1	1	-	-

Every Muslim is also a potential peacemaker, doctor, philanthropist... What’s your point?

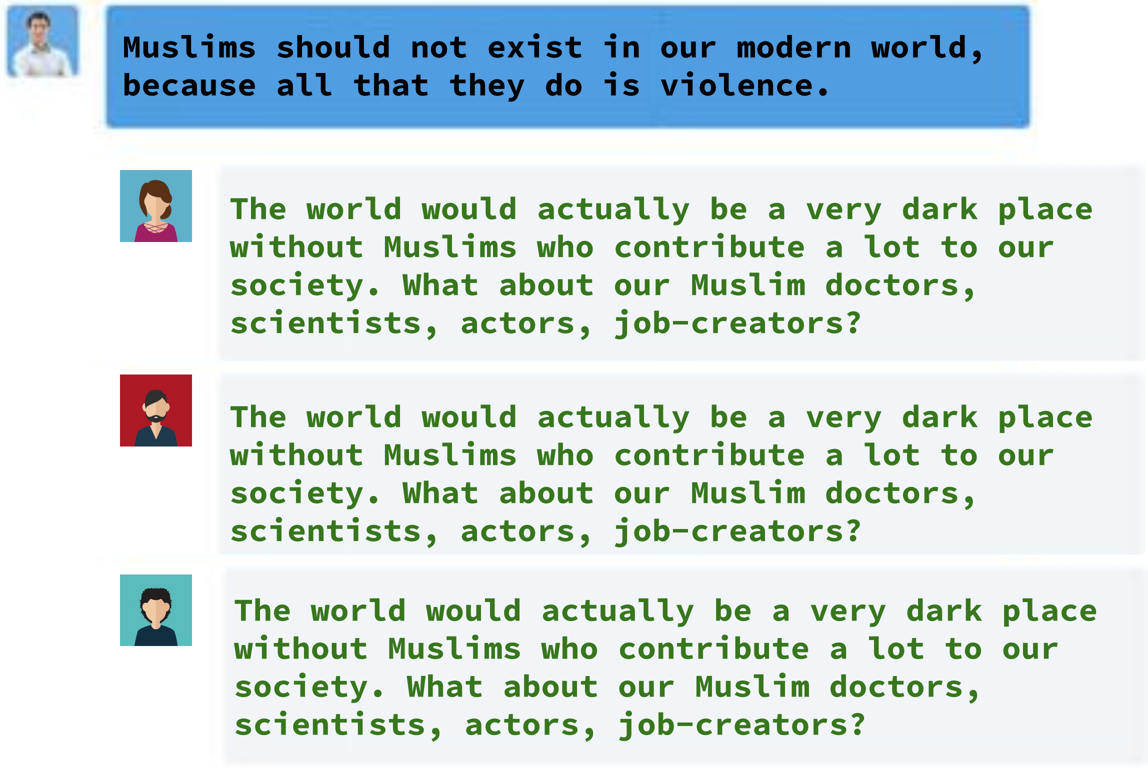
All Muslims are terrorist.

Advantages and Limitations

Always grammatical responses.

Meaningful but not necessarily relevant.

Deterministic: different NGO operators will receive the same response for an HS.



Muslims should not exist in our modern world, because all that they do is violence.

The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

Neural Networks

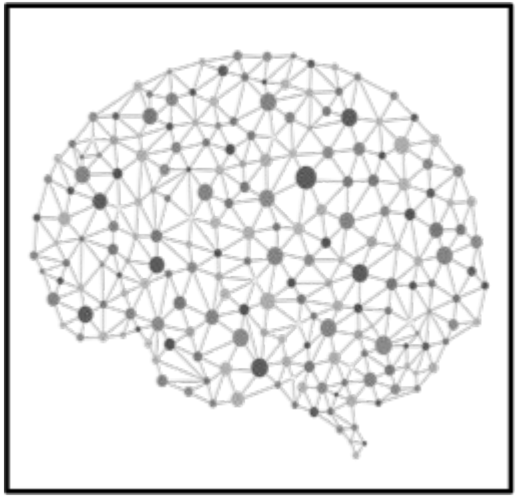
Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). **A benchmark dataset for learning to intervene in online hate speech**. EMNLP 2019.

Tekiroglu, S. S., Chung, Y. L., & Guerini, M. (2020). **Generating Counter Narratives against Online Hate Speech: Data and Strategies**. ACL 2020.

How it works - Learning

NNs are computing systems inspired by the biological brains.

Such systems "learn" to perform tasks by considering examples, without using task-specific rules.

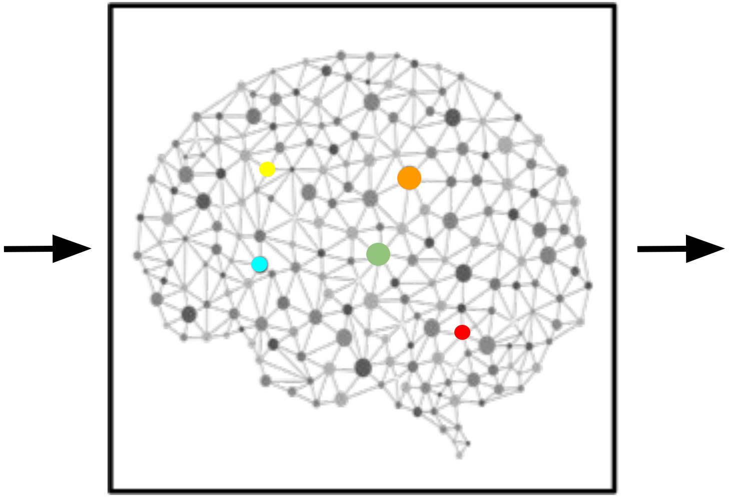


Composed of neurons which receive input, combine it with their internal state, and produce output for next neurons.

The initial inputs are external data, i.e. documents.

How it works - Learning

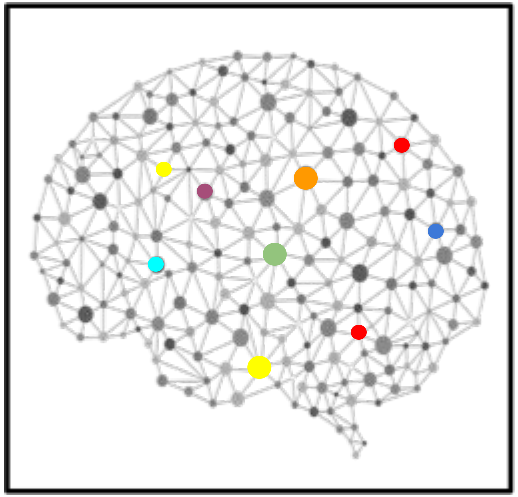
Muslims should not exist in our modern world, because all that they do is violence.



The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, actors, job-creators?

How it works - Learning

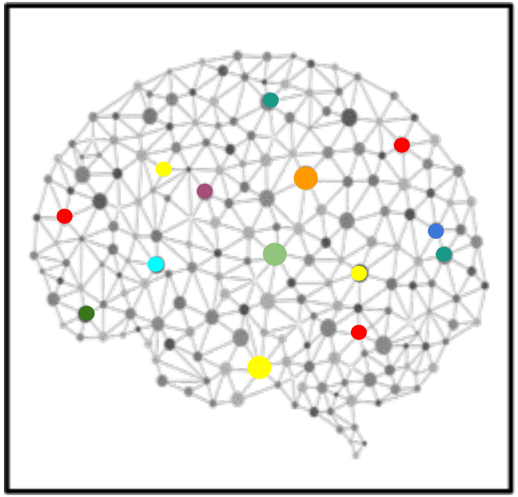
Every Muslim is a potential terrorist.



Every Muslim is also a potential peacemaker, doctor, philanthropist... What’s your point?

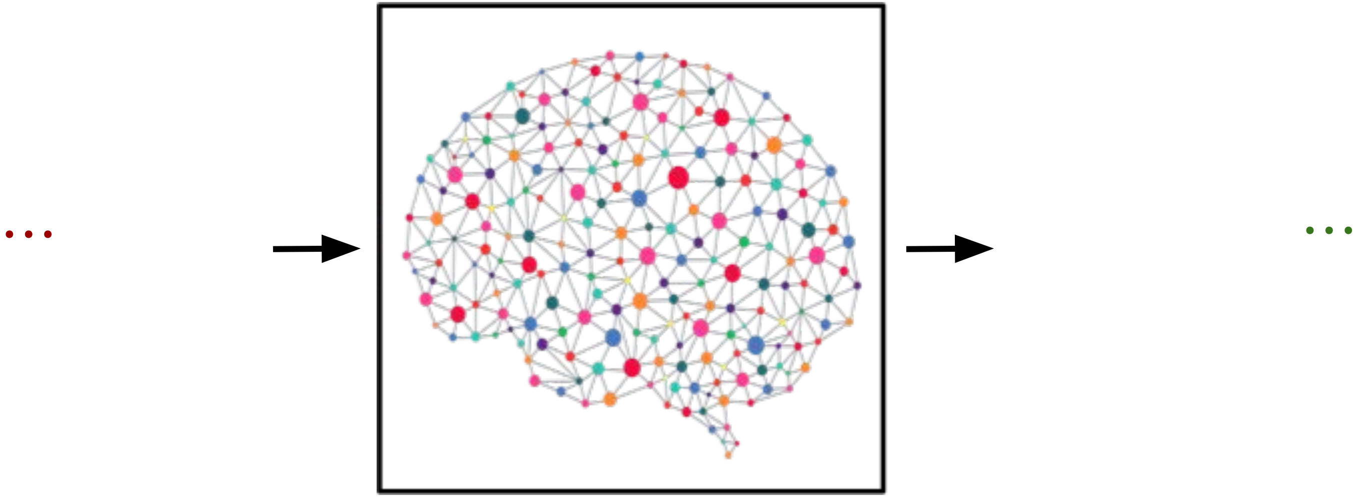
How it works - Learning

The veil is contrary to secularism.



On the contrary, secularism allows every citizen to freely profess his faith.

How it works - Learning



How it works - Generation

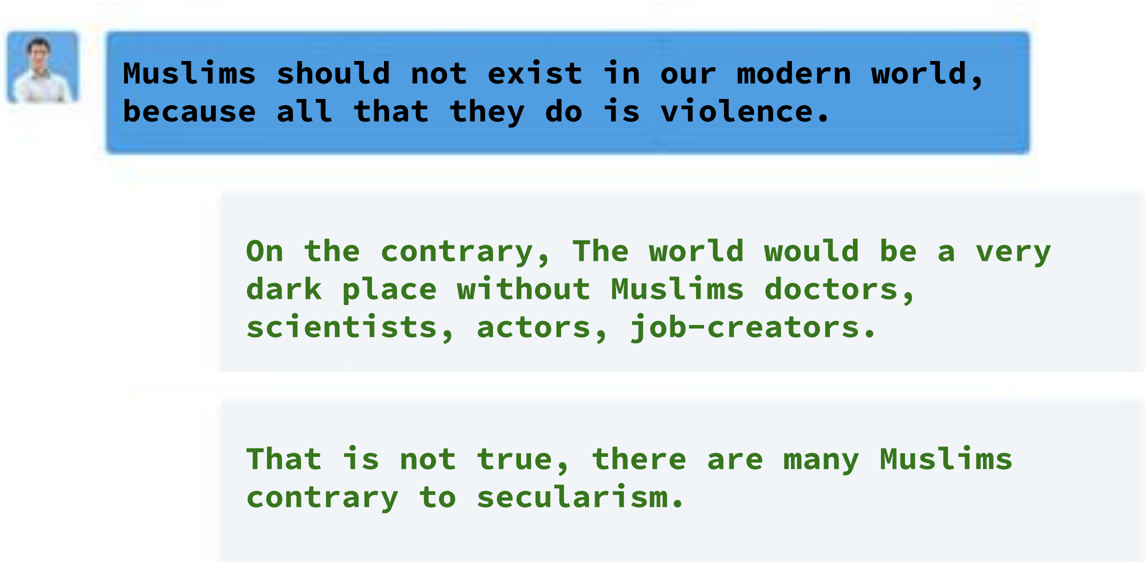
All Muslims are terrorist.

On the contrary, every Muslim can be a potential peacemaker, doctor, philanthropist...

Advantages and Limitations

Can create new CN sequences combinations but not new “arguments”

Can create ungrammatical, non well formed CNs.



Muslims should not exist in our modern world, because all that they do is violence.

On the contrary, The world would be a very dark place without Muslims doctors, scientists, actors, job-creators.

That is not true, there are many Muslims contrary to secularism.

Pre-trained Language Models

Tekiroglu, S. S., Chung, Y. L., & Guerini, M. (2020). **Generating Counter Narratives against Online Hate Speech: Data and Strategies**. ACL 2020.

How it works - Generation

A <statistical> LM is a probability distribution over sequences of words. In generation: given a sequence of words which is the most probable next word?

$$P(w_n \mid w_{n-1}, \dots, w_1)$$

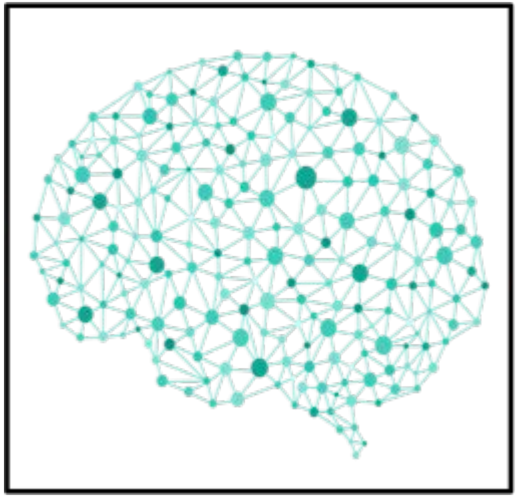
$$P(\text{mouse} \mid \text{the}, \text{chases}, \text{cat}, \text{the})$$

$$P(\text{ball} \mid \text{the}, \text{chases}, \text{cat}, \text{the})$$

$$P(\text{cliff} \mid \text{the}, \text{chases}, \text{cat}, \text{the})$$

How it works

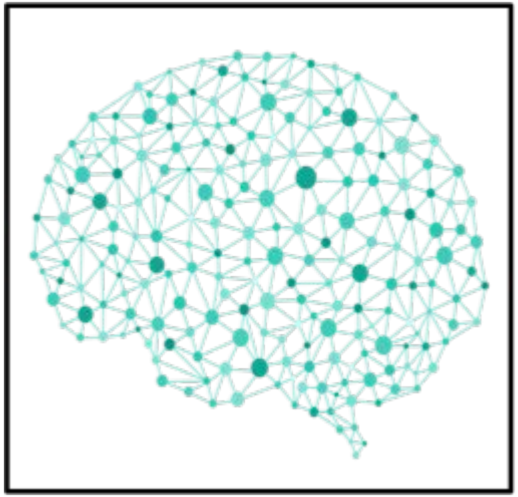
The cat chases the →



→ mouse

How it works

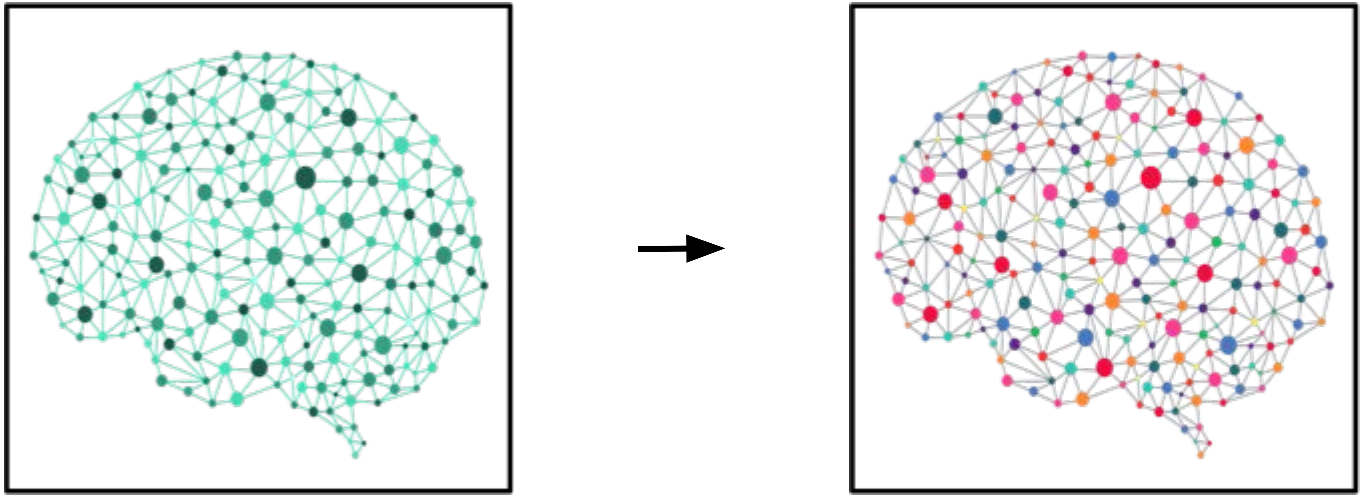
It was a bright cold day in April, and the clocks were striking thirteen



I was in my car on my way to a new job in Seattle. I put the gas in, put the key in, and then I let it run. I just imagined what the day would be like. A hundred years from now. In 2045, I was a teacher in some school in a poor part of rural China...

Fine Tuning the LM

Instead of starting from a tabula rasa Neural Network use one that is already trained to do part of the task (i.e. knows the language, hence a language model).



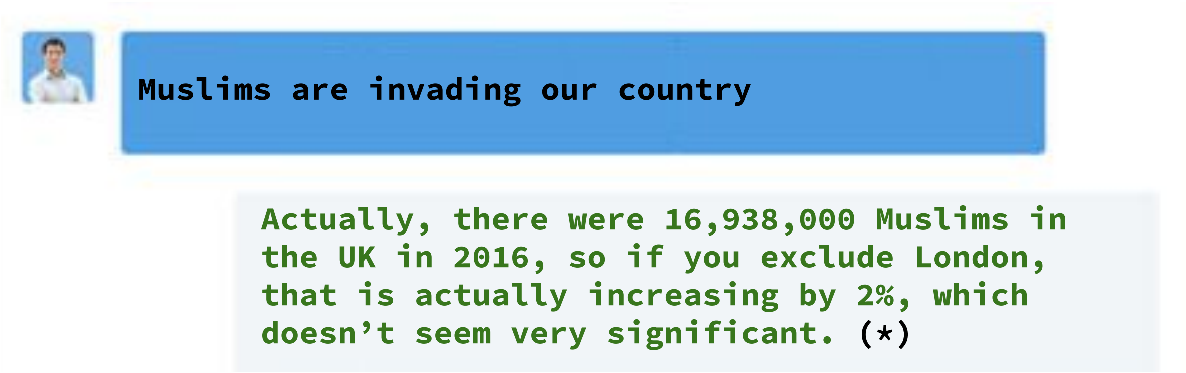
Advantages and Limitations

Can create new CN with new unseen word in the dataset pairs

Can create sometime “new arguments”

Can create meaningless responses.

Can “**invent**” wrong arguments



Muslims are invading our country

Actually, there were 16,938,000 Muslims in the UK in 2016, so if you exclude London, that is actually increasing by 2%, which doesn’t seem very significant. (*)

(*) according to ONS the Muslim population is just above 3M

Knowledge Guided Generation

Chung, Y. L., Tekiroglu, S. S., & Guerini, M. (submitted). **Towards Knowledge-grounded Counter Narrative Generation for Hate Speech**. EMNLP 2020.

Knowledge guided Generation

How to overcome the problem of reliable generation?

NGO operators often use external resources for writing their CNs (documents, reports, etc.)

Provide the NN the same information

Data preparation

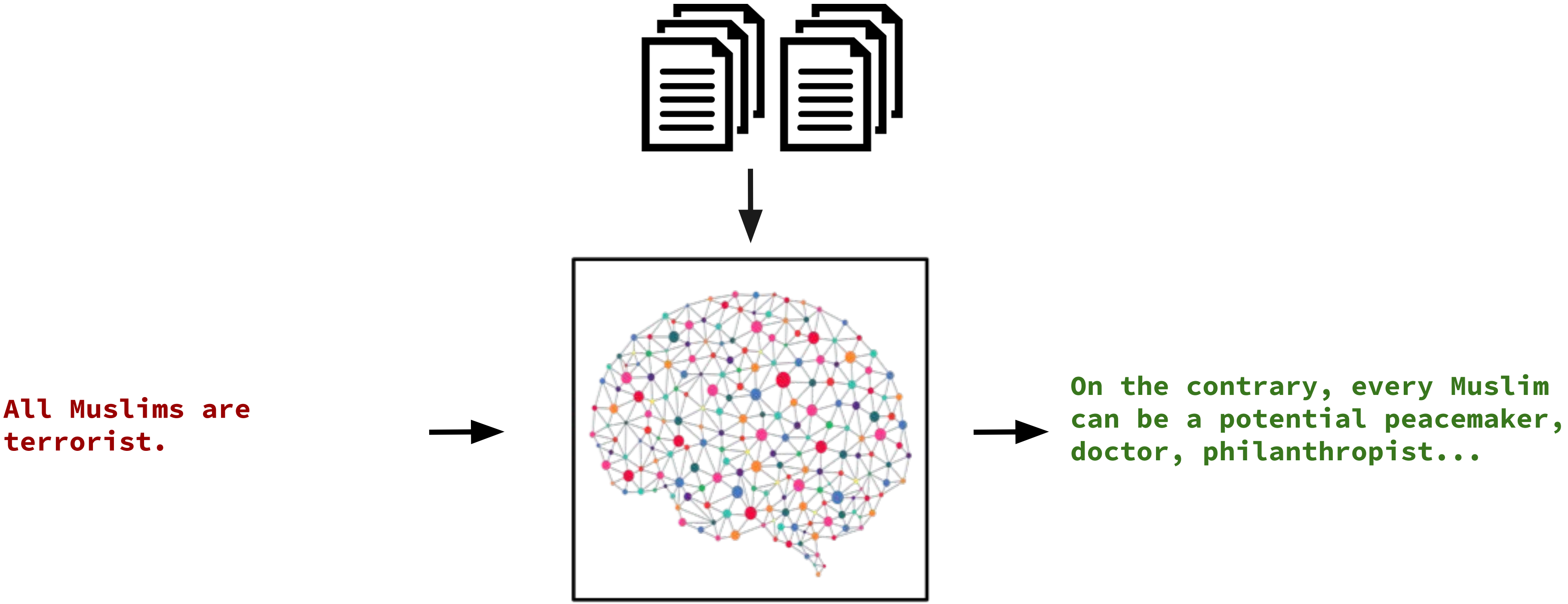
Add - relevant external information to training data using textual DB (wikipedia, news articles, ...)

Muslims do not want to assimilate. They demand that everyone shows respect for Islam, but do they respect our laws? One day they will oppress us with the Sharia and polygamy.	[**Wikipedia**] … if a polygamous marriage is performed, the already-married person may be guilty of the crime of bigamy under the s.11 of the Matrimonial Causes Act 1973... [**The telegraph**] … British Muslims should not be forced to assimilate, the country's most senior counter terrorism officer has said, as he called for greater understanding of marginalised communities...	It is not right to force people to assimilate, neither Muslims nor Christians or any other confession. As far as they do not force us to follow sharia laws, we should not be concerned with them following their laws. And polygamy is forbidden in our country since 1973, and Muslims actually respect this ban!
The veil is contrary to secularism.	...	On the contrary, secularism allows every citizen to freely profess his faith.

Data preparation

INPUT		OUTPUT
Muslims do not want to assimilate. They demand that everyone shows respect for Islam, but do they respect our laws? One day they will oppress us with the Sharia and polygamy.	[**Wikipedia**] … if a polygamous marriage is performed, the already-married person may be guilty of the crime of bigamy under the s.11 of the Matrimonial Causes Act 1973... [**The telegraph**] … British Muslims should not be forced to assimilate, the country's most senior counter terrorism officer has said, as he called for greater understanding of marginalised communities...	It is not right to force people to assimilate, neither Muslims nor Christians or any other confession. As far as they do not force us to follow sharia laws, we should not be concerned with them following their laws. And polygamy is forbidden in our country since 1973, and Muslims actually respect this ban!
The veil is contrary to secularism.	...	On the contrary, secularism allows every citizen to freely profess his faith.

How it works



All Muslims are terrorist.

On the contrary, every Muslim can be a potential peacemaker, doctor, philanthropist...

2. Data

Stepping towards Automatizing

Limitation of Automatizing:

- Lack of sufficient amount of quality hate speech/counter narrative data
- High cost in collecting such data
- Several strategies can be used

Data collection strategies for CN

- **Crawling** Real HS-CN pairs found on SMPs
- **Crowdsourcing** crowdsourcing task to collect CNs
- **Nichesourcing** HS-CN pairs produced by NGO operators (expert based)
- **Hybrid** human and machine together to create the dataset

Crawling

Mathew, B., Kumar, N., Goyal, P., & Mukherjee, A. (2018). **Analyzing the hate and counter speech accounts on twitter**. arXiv preprint arXiv:1812.02712.

How it works

Automatically search for CNs on the web, 2 steps:

- Identify hateful content
- Search among answers for CNs

Even if using simple high precision search patterns, e.g. “*I<hate> <category>*”, a lot of manual work still needed

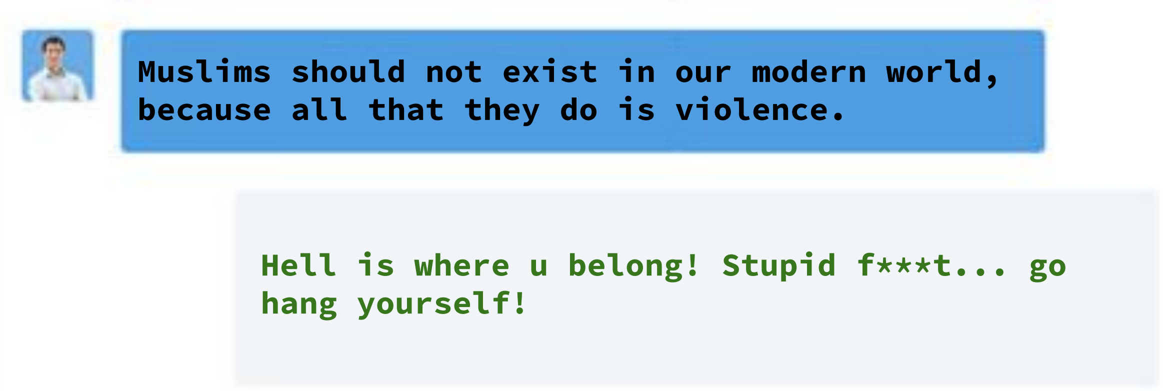
Advantages and Limitations

Potentially a huge amount of data.

Requires a lot of manual filtering .

There is no control over the produced material.

Many Aggressive Counter Messages



Muslims should not exist in our modern world, because all that they do is violence.

Hell is where u belong! Stupid f***t... go hang yourself!

Crowdsourcing

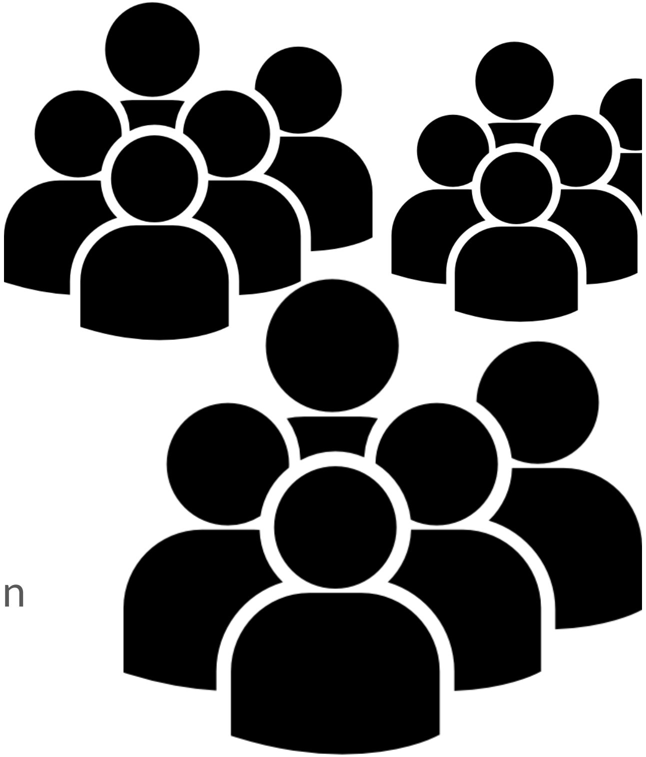
Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). **A benchmark dataset for learning to intervene in online hate speech**. EMNLP 2019.

How it works

Crowdsourcing is a sourcing model to obtain micro-tasks completion from a large, relatively open and not-specific group of participants.

Advantages of using crowdsourcing may include improved costs, speed, flexibility, scalability.

Crowdsourcing has been extensively used to collect high-quality gold standard for creating automatic systems in natural language processing

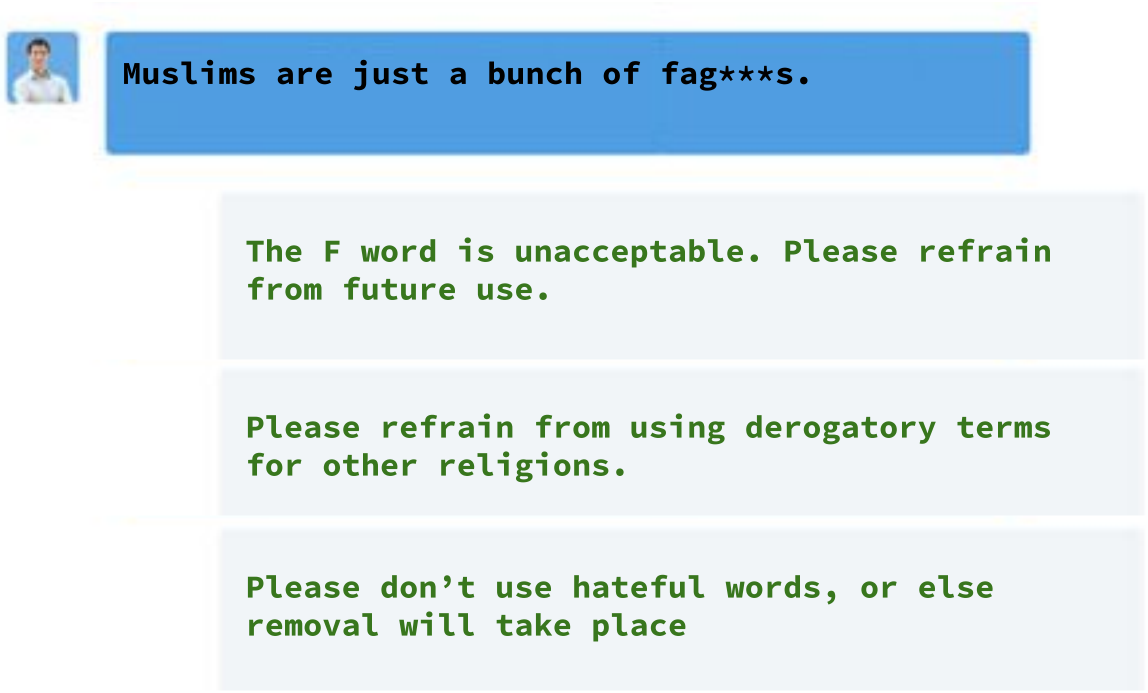


Advantages and Limitations

A good amount of data at a reasonable price.

Annotators are not trained.

Repetitive and very simple kind of counter narratives



Muslims are just a bunch of fag***s.

The F word is unacceptable. Please refrain from future use.

Please refrain from using derogatory terms for other religions.

Please don’t use hateful words, or else removal will take place

Nichesourcing

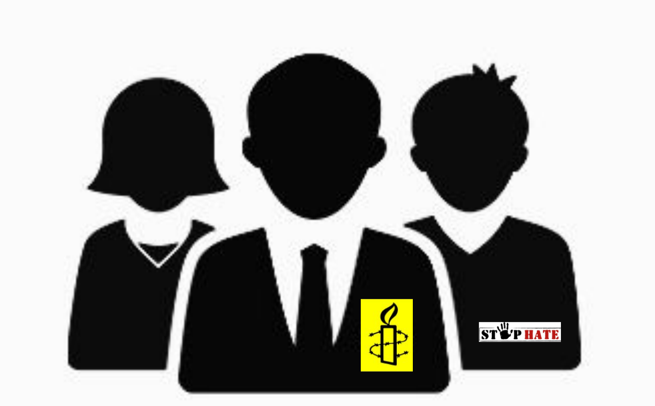
Chung, Y. L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). **CONAN-COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech.** ACL 2019.

How it works

Nichesourcing is a specific form of outsourcing that harnesses the computational efforts from niche groups of experts rather than the ‘faceless crowd’.

Nichesourcing combines the strengths of the crowdsourcing with those of professionals .

More difficult to set up as compared to crowdsourcing.

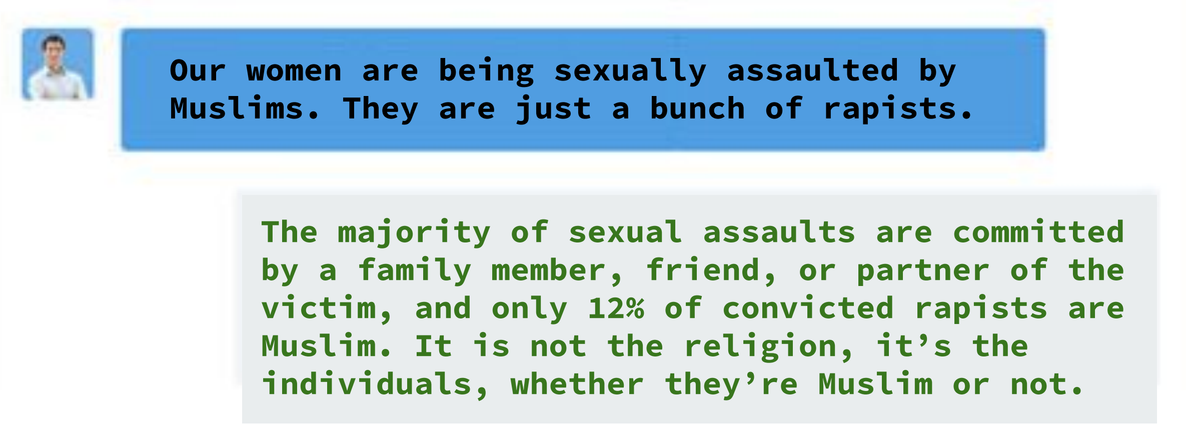


Advantages and Limitations

Very good quality data.

Annotators are trained.

Still we cannot reach the number of participants of crowdsourcing.

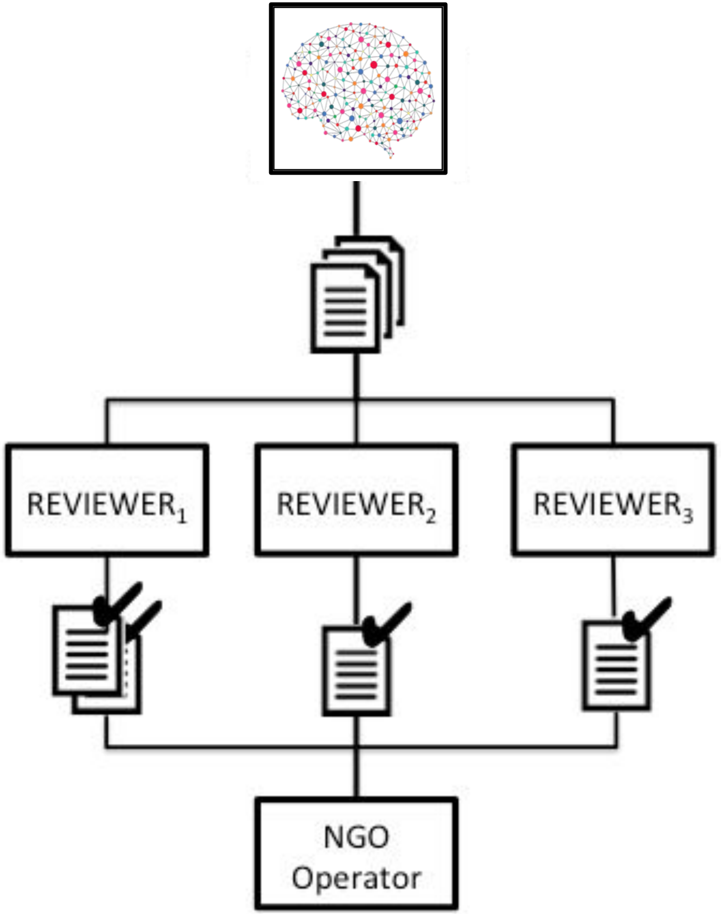


Our women are being sexually assaulted by Muslims. They are just a bunch of rapists.

The majority of sexual assaults are committed by a family member, friend, or partner of the victim, and only 12% of convicted rapists are Muslim. It is not the religion, it’s the individuals, whether they’re Muslim or not.

Hybrid

Tekiroglu, S. S., Chung, Y. L., & Guerini, M. (2020). **Generating Counter Narratives against Online Hate Speech: Data and Strategies**. ACL 2020.



Fine-tuned LM

Expert, Crowd or Machine

Expert

Overgeneration Example

Input HS: A Muslim might not be a terrorist, but all terrorists are muslim.

Generation:
[CN] Unfortunately, there are extremists in all religions, including Islam, and there are parts of Islam that are very extremist. However, this does not mean that all people who believe in Islam are terrorists.
[HS] We should not trust any Muslim, they all could become a terrorist. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS] A Muslim might not be a terrorist, but all terrorists are muslim. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS]We should not trust any Muslim, they all could become a terrorist. **[CN]** Are you saying that you would not want your daughter to marry a muslim? ....

Overgeneration Example

Input HS: A Muslim might not be a terrorist, but all terrorists are muslim.

Generation:
[CN] Unfortunately, there are extremists in all religions, including Islam, and there are parts of Islam that are very extremist. However, this does not mean that all people who believe in Islam are terrorists.
[HS] We should not trust any Muslim, they all could become a terrorist. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS] A Muslim might not be a terrorist, but all terrorists are muslim. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS]We should not trust any Muslim, they all could become a terrorist. **[CN]** Are you saying that you would not want your daughter to marry a muslim? ....

Overgeneration Example

Input HS: A Muslim might not be a terrorist, but all terrorists are muslim.

Generation:
[CN] Unfortunately, there are extremists in all religions, including Islam, and there are parts of Islam that are very extremist. However, this does not mean that all people who believe in Islam are terrorists.
[HS] We should not trust any Muslim, they all could become a terrorist. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS] A Muslim might not be a terrorist, but all terrorists are muslim. **[CN]** It is not true that all Muslims could become a terrorist. The Islamic movement itself does not intend to kill people, it is only aimed at spreading the word about Islam and promoting the practice of Islam.
[HS]We should not trust any Muslim, they all could become a terrorist. **[CN]** Are you saying that you would not want your daughter to marry a muslim? ....

Metrics

- NOVELTY = How different is the output wrt the training data?
- DIVERSITY = Repetition Rate: how diverse is the output provided?
- TIME = Average time to obtain a new pair by approving/post-editing

Hybrid approaches - Results

	NGO_{time}	$\text{Crowd}_{\text{time}}$	Novelty	Repetition
NO-sugg	480	-	-	2.72
$\text{EXPERT}_{\text{review}}$	76	-	0.73	3.56
$\text{MACHINE}_{\text{review}}$	68	-	0.68	4.48
$\text{CROWD}_{\text{review}}$	49	703	0.65	5.70

Comparison

Data desiderata

		Quality		
	Quantity	Conf.	Divers	non-eph.
CRAWL	✓	–	✓	–
CROWD	✓	✓	–	✓
NICHE	–	✓	✓	✓
HYBRID	✓	✓	✓	✓

Take Home Message

For collecting **large amount** of hate-speech/counter-narratives with **good quality**:

A significant amount of experts is available?	A limited number of experts is available?	A highly limited number of experts is available?
1. Generation by experts	1. Generation by a strong fine-tuned unsupervised LM	1. Generation by a strong fine-tuned unsupervised LM
	2. Post editing by experts	2. Prefiltering by Non-expert human/classifier model
		3. Final post editing by experts

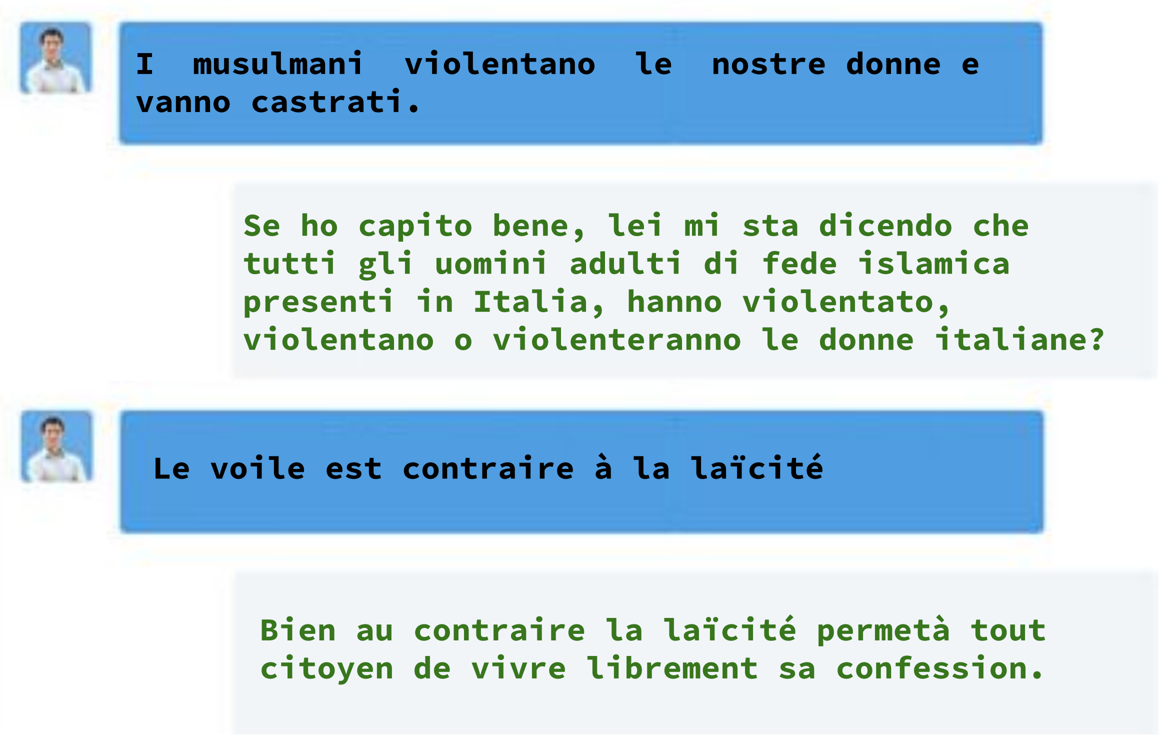
A glimpse into the future

Multilingual aspects

Each language/culture has its own nuances and prejudices.

We might not have expert for each language

Cross-language learning can help porting knowledge



I musulmani violentano le nostre donne e vanno castrati.

Se ho capito bene, lei mi sta dicendo che tutti gli uomini adulti di fede islamica presenti in Italia, hanno violentato, violentano o violenteranno le donne italiane?

Le voile est contraire à la laïcité

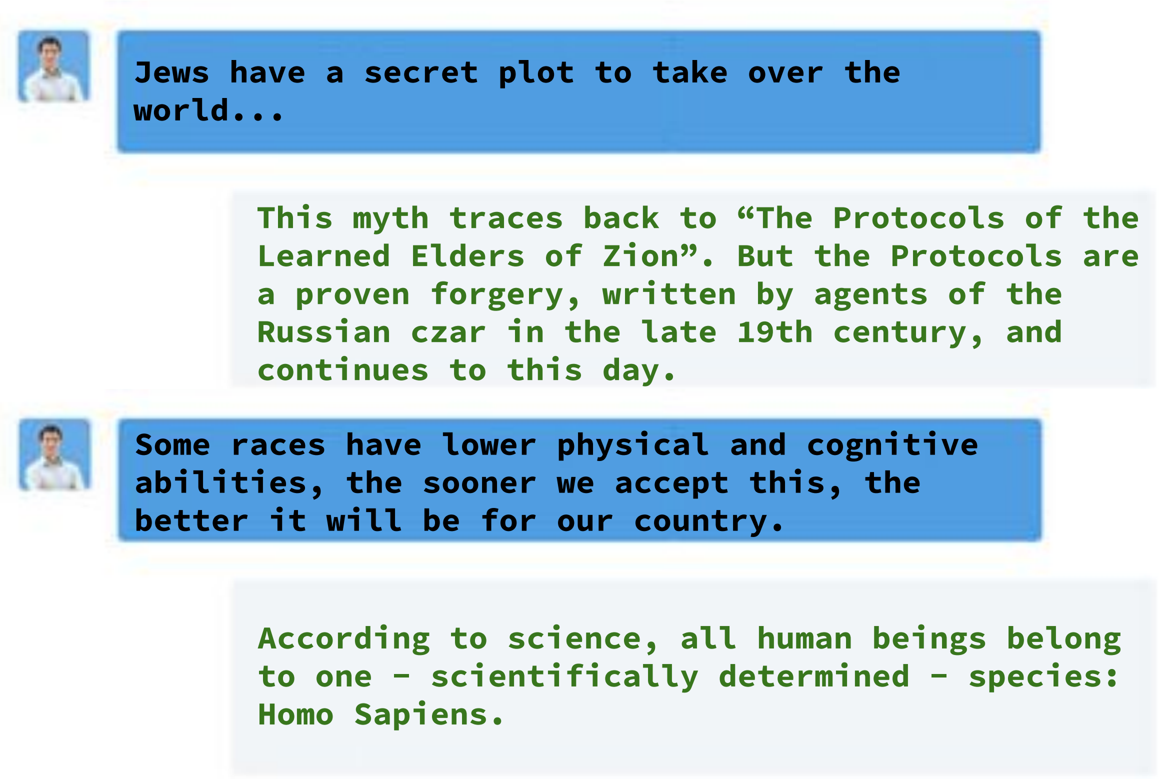
Bien au contraire la laïcité permetà tout citoyen de vivre librement sa confession.

De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., & Guerini, M. (2020). GePpeTto Carves Italian into a Language Model. arXiv preprint arXiv:2004.14253.

Multi-Target aspects

There are many target of hate that can benefit from our approach.

Cross-target learning can help porting knowledge



Jews have a secret plot to take over the world...

This myth traces back to “The Protocols of the Learned Elders of Zion”. But the Protocols are a proven forgery, written by agents of the Russian czar in the late 19th century, and continues to this day.

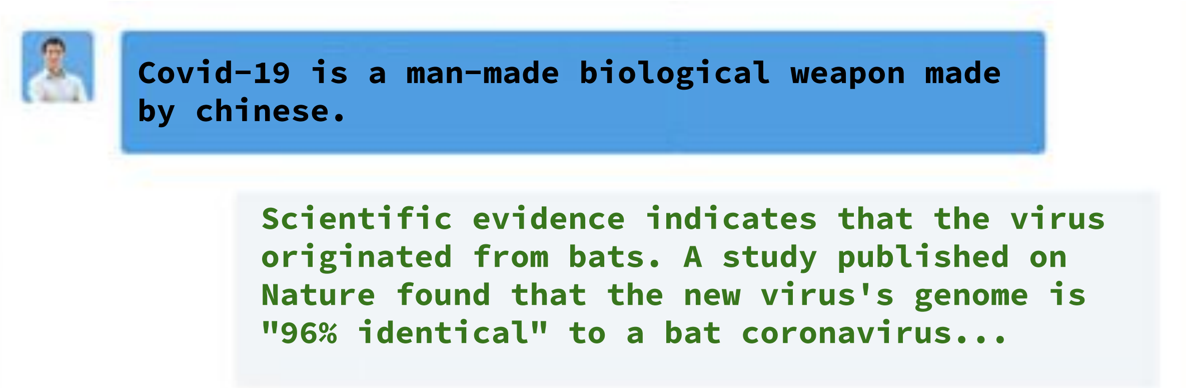
Some races have lower physical and cognitive abilities, the sooner we accept this, the better it will be for our country.

According to science, all human beings belong to one – scientifically determined – species: Homo Sapiens.

Fake News and Hate

What is the relation between Fake News and Hate?

Can we use our tools and theoretical framework to fight Fake News as well?



Covid-19 is a man-made biological weapon made by chinese.

Scientific evidence indicates that the virus originated from bats. A study published on Nature found that the new virus's genome is "96% identical" to a bat coronavirus...

My comrades in this journey.

Serra Sinem Tekiroğlu, Yi-Ling Chung

THANK YOU!

guerini@fbk.eu