SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

MONITORING CYBERBULLYING THROUGH MESSAGE CLASSIFICATION AND SOCIAL NETWORK ANALYSIS

Elena Cabrio and Serena Villata

UNIVERSITÉ CÔTE D'AZUR

cnrs

Inria

17 novembre 2020

SOPHIA MASTER CLASSES
by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Elena Cabrio
Assistant Professor, HDR
Université Cote d'Azur

Serena Villata
Tenured researcher, HDR
CNRS

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Outline of the Master Class

- **Introduction to hate speech detection**
  - Available datasets
  - State-of-the-art methods
  - Results
  - Open challenges
- **The CREEP Project for cyberbullying detection**
  - Methodology and results
  - Multilinguality and images
  - Demo

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Outline of the Master Class

- **Introduction to hate speech detection**
  - Available datasets
  - State-of-the-art methods
  - Results
  - Open challenges
- **The CREEP Project for cyberbullying detection**
  - Methodology and results
  - Multilinguality and images
  - Demo

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

Disclaimer

17 NOVEMBRE 2020

This presentation contains examples of language that may be offensive to some of you. Of course, they do not reflect the views of the presenter

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Why study hate speech automatic detection?

European Union Commission directives.

- The European Union Commission has been conducting different initiatives for decreasing hate speech.
- Several **programs** are being founded toward the fighting of hate speech (e.g., No Hate Speech Movement by the Council of Europe).
- Another strategy is through **legislation**. The European Union Commission pressured Facebook, YouTube, Twitter, and Microsoft to sign an **EU hate speech code**. This includes the requirement to review the majority of valid notifications for removal of illegal hate speech in less than 24h. Also, European regulators accused Twitter of not being good enough at removing hate speech from its platform.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hate speech detection

- Hate crimes are unfortunately **nothing new** in society.
- **Social media** and other means of online communication have begun playing **an important role in hate crimes**.
  - Example: suspects in several recent hate-related terror attacks had an extensive social media history of hate-related posts, suggesting that social media contributes to their radicalization.
  - Example: social media may play an even more direct role like video footage from the suspect of the 2019 terror attack in Christchurch, New Zealand, was broadcast live on Facebook.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hate speech detection

- Vast online communication forums, including social media, enable users to express themselves **freely**, at times, **anonymously**.
- **Crossroad between the freedom of expression right that should be cherished, and the abuse of this liberty by spreading hate towards another group.**
  - Example: The American Bar Association asserts that in the United States, hate speech is legal and protected by the First Amendment, although not if it directly calls for violence [https://abalegalfactcheck.com/articles/hate-speech.html]: *The U.S Supreme Court has made it clear that governments may not restrict speech expressing ideas that offend.*

SOPHIA MASTER CLASSES
by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hate speech detection

Many online forums such as Facebook, YouTube, and Twitter consider hate speech harmful, and have policies to remove hate speech content.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Facebook

17 NOVEMBRE 2020

12. Content that incites hate

Hate speech is not allowed on Facebook as it creates an environment of intimidation and exclusion and, in some cases, can promote actual violence.

We define hate speech as a direct attack on people on the basis of legally protected aspects, such as race, ethnicity, nationality of origin, religion, sexual orientation, caste, sex, gender or gender identity and disability or serious illness. We offer age-based protection against attacks if this is associated with another protected feature. In addition, we offer certain protections for immigrant status. **We define the attack as violent or dehumanizing speech, harmful stereotypes, declarations of inferiority or incitement to exclusion or segregation.** Attacks are divided into three severities, described below.

Sometimes we share other people's hateful content with the aim of raising awareness or informing other people. In some cases, words or terms that might otherwise violate our standards are used in a self-referential manner or to reinforce a case. Sometimes people express contempt in the context of breaking up a relationship. Other times, they use gender-specific language to check memberships for a health-related or positive-tone support group, such as a women's breastfeeding group. **In all of these cases, we allow the content, but we expect people to clearly indicate their intention, helping us better understand why they shared it. When the intention is not clear,** **we can remove the content.**

Additionally, we believe that using your identity prompts people to be more accountable when sharing these types of comments.

YouTube

17 NOVEMBRE 2020

Hate speech is not allowed on YouTube. 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
- Veteran Status

Here are examples of hate speech not allowed on YouTube.

- *“I’m glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above].”*
- *“[Person with attributes noted above] are dogs” or “[person with attributes noted above] are like animals.”*

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

YouTube

17 NOVEMBRE 2020

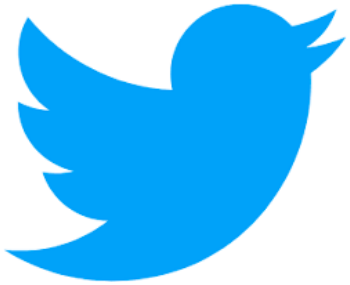
What happens if content violates this policy

If your content violates this policy, we'll remove the content and send you an email to let you know. If this is your first time violating our Community Guidelines, you'll get a warning with no penalty to your channel. If it's not, we'll issue a strike against your channel. **If you get 3 strikes, your channel will be terminated.** You can learn more about our strikes system here.
We may also terminate your channel or account for repeated violations of the Community Guidelines or Terms of Service, as well as due to a single case of severe abuse, or when the channel is dedicated to a policy violation. You can learn more about channel or account terminations here. If we think your content comes close to hate speech, we may limit YouTube features available for that content.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Twitter

17 NOVEMBRE 2020



Hateful conduct policy

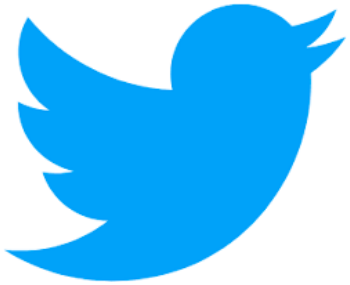
Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Twitter



When this applies

- We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior, whether within Tweets or Direct Messages.
- Violent threats
- Wishing, hoping or calling for serious harm on a person or group of people
- References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims
- Inciting fear about a protected category
- Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone
- Hateful imagery

Source	Hate speech is to incite violence or hate	Hate speech is to attack or diminish	Hate speech has specific targets	Humour has a specific status
EU Code of conduct	Yes	No	Yes	No
ILGA	Yes	No	Yes	No
Scientific paper	No	Yes	Yes	No
Facebook	No	Yes	Yes	Yes
YouTube	Yes	No	Yes	No
Twitter	Yes	Yes	Yes	No

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Why hate speech detection?

- Due to the societal concern and how widespread hate speech is becoming on the Internet, there is strong motivation to study **automatic detection of hate speech**.
- By automating its detection, the spread of hateful content can be reduced.
- **Why human verification is not enough?**
  - It cannot scale with respect to the size of social networks.
- **Human verification is required as a final validation to avoid violating the right of freedom of expression.**

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hate speech: challenges

- Detecting hate speech is a challenging task:
  - there are **disagreements in how hate speech should be defined**. This means that some content can be considered hate speech to some and not to others, based on their respective definitions.
  - competing definitions provide **challenges for evaluation of hate speech detection systems**;
  - existing datasets differ in their definition of hate speech, leading to datastets that are **not only from different sources**, but also capture **different information**.

- Some recent approaches found promising results for detecting hate speech in textual content.
- The proposed solutions employ machine learning techniques to classify text as hate speech.
- One limitation of these approaches is that **the decisions they make can be opaque and difficult for humans to interpret why the decision was made.** This is a practical concern because systems that automatically censor a person's speech likely need a manual appeal process.
- Some of the existing approaches use **external sources**, such as a hate speech lexicon, in their systems. This can be effective, but it requires **maintaining these sources and keeping them up to date** which is a problem in itself.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Take home messages

- Automatic hate speech detection is **technically difficult**;
- Some approaches achieve **reasonable performance**;
- **Specific challenges** remain among all solutions;
- Without societal **context**, systems cannot generalise sufficiently.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Defining hate speech

- **The definition of hate speech is neither universally accepted nor are individual facets of the definition fully agreed upon.**
- [Ross et al., 2016] believe that a clear definition of hate speech can help the study of detecting hate speech by making annotating hate speech an easier task, and thus, making the annotations more reliable.
- **However, the line between hate speech and appropriate free expression is blurry, making some wary to give hate speech a precise definition.**
- The American Bar Association does not give an official definition, but instead asserts that speech that contributes to a criminal act can be punished as part of a hate crime.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Definitions of hate speech

17 NOVEMBRE 2020

- **Encyclopedia of the American Constitution**: “Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity.”
- **Facebook**: “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.”
- **Twitter**: “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.”
- **Davidson et al.**: “Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.”
- **de Gilbert et al.**: “Hate speech is a deliberate attack directed towards a specific group of people motivated by aspects of the group’s identity.”
- **Fortuna et al.:** “Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.”

SOPHIA MASTER CLASSES
by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Facebook examples for its workers

Message	Evaluation
Don't trust boys!	Violating—delete
Refugees should face the figuring squad!	Violating—delete
Fucking Muslims!	Violating—delete
Fucking migrants!	Non-violating—Ignore
Migrants are filthy cockroaches that will infect our country	Violating—delete
I'm such a faggot people call me diva!	Non-violating—Ignore
The French are alcoholics	Violating—delete
All English people are dirty!	Violating—delete
Don't try to explain—Irish Catholics are just idiots	Violating—delete
Migrants are scum!	Violating—delete
People should stop to use the word nigger.	Non-violating—Ignore
I hate migrants!	Non-violating—Ignore
Don't trust boys who say they love you!	Non-violating—Ignore
Tall girls are just freaks!	Non-violating—Ignore
American shitheads!	Violating—delete
Migrants are so filthy!	Non-violating—Ignore
Refugees! More like rape-fugees!	Violating—delete
Asylum seekers out!	Violating—delete
Group for blacks only!	Non-violating—Ignore

- In some of the definitions above, **a necessary condition is that it is directed to a group**.
- This differs from the Encyclopedia of the American Constitution definition, where an attack on an individual can be considered hate speech.
- A common theme among the definitions is that **the attack is based on some aspect of the group or peoples identity**.
- While in de Gilbert's definition the identity itself is left vague, some of the other definitions provide specific identity characteristics, e.g., in Davidson et al. and Facebook definitions.
- Fortuna et al.'s definition specifically calls out variations in language style and subtleties. This can be challenging, and **goes beyond what conventional text-based classification approaches** are able to capture.

- Fortuna et al.’s definition is based on an analysis of the following characteristics from other definitions:
  - **Hate speech is to incite violence or hate**
  - **Hate speech is to attack or diminish**
  - **Hate speech has specific targets**
  - **Whether humor can be considered hate speech**

- A particular problem not covered by many definitions relate to **factual statements**.
  - For example, "*Jews are swine*" is clearly hate speech by most definitions (it is a statement of inferiority), but "*Many Jews are lawyers*" is not.
  - In the latter case, to determine whether each statement is hate speech, we would need to check whether the statement is factual or not using **external sources of knowledge**.
  - This type of hate speech is difficult because it relates to **real-world fact verification**.
  - To evaluate validity, we would initially need to define **precise word interpretations**, namely, is "many" an absolute number or by relative percentage of the population, further complicating the verification.

- Another issue that arises in the definition of hate speech is the potential **praising of a group that is hateful**.
  - For example, praising the KKK is hate speech, however praising another group can clearly be non-hate speech.
  - It is important to know what groups are hate groups and what exactly is being praised about the group. For example, *the Nazis were very efficient in terms of their "Final Solution"*.
  - Praise processing alone is, at times, difficult.

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Datasets

- Collecting and annotating data for the training of automatic classifiers to detect hate speech is challenging.
- Specifically, **identifying and agreeing whether specific text is hate speech** is difficult, as there is no universal definition of hate speech.
- [Ross et al., 2016] studied the reliability of hate speech annotations and suggest that annotators are unreliable. Agreement between annotators, measured using Krippendorff's α, was very low (up to 0.29).

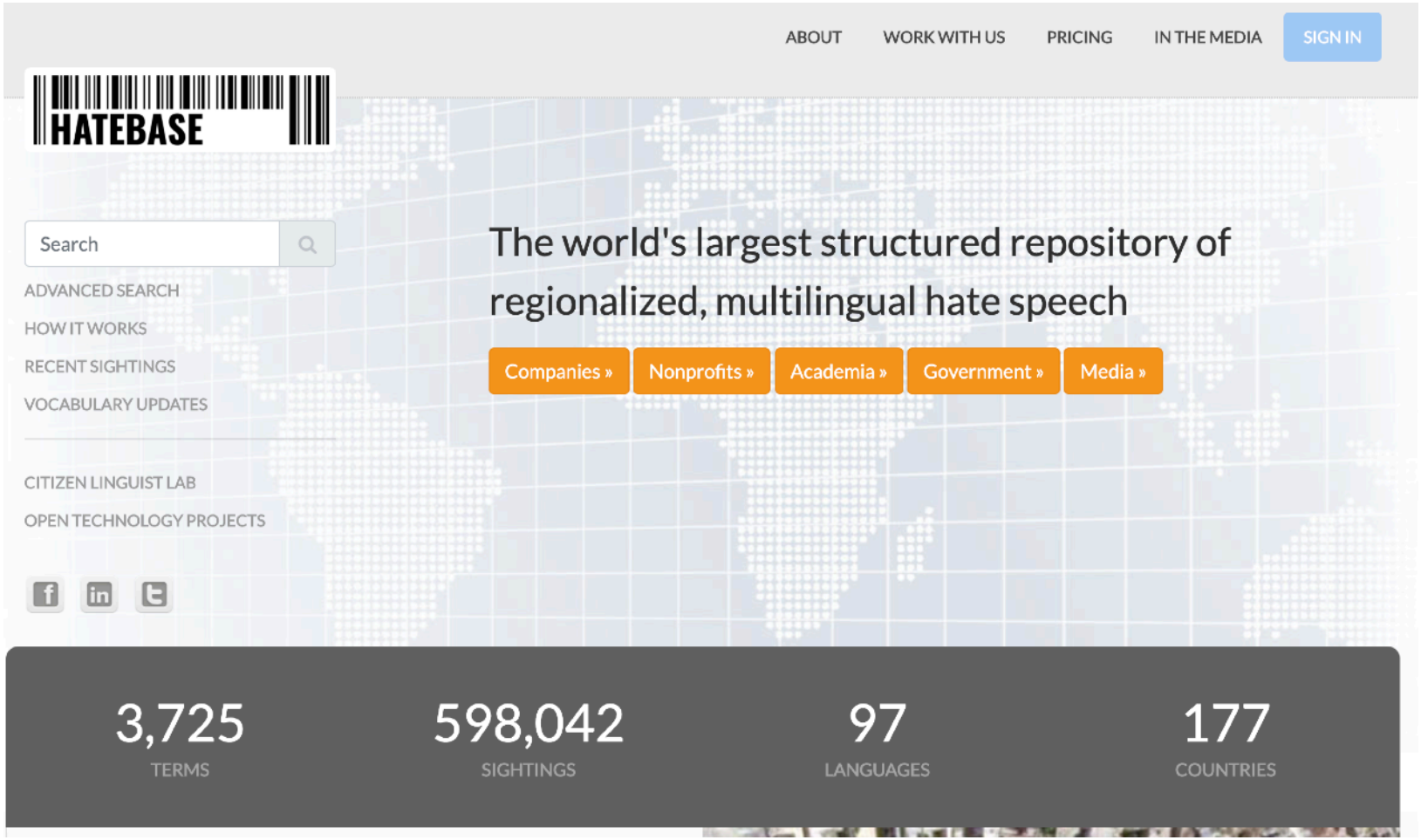
- **Social media platforms are a hotbed for hate speech, yet many have very strict data usage and distribution policies.**
- This results in a relatively **small number of datasets available to the public to study**, with **most coming from Twitter** (which has a more lenient data usage policy).
- While the Twitter resources are valuable, their **general applicability is limited** due to the unique genre of Twitter posts; the character limitation results in terse, short-form text.
- Posts from other platforms are typically longer and can be part of a larger discussion on a specific topic, e.g., Facebook, Instagram, TikTok. This provides additional context that can affect the meaning of the text (including images and videos).
- Another challenge is that **there are not many publicly-available, curated datasets that identify hateful, aggressive, and insulting text**.

SOPHIA MASTER CLASSES
by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hatebase

https://hatebase.org/



SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

HateBase Twitter

- Twitter dataset of 24,802 tweets provided by [Davidson et al., 2017].
- **Procedure** for creating the dataset:
  - They took a hate speech lexicon from Hatebase and searched for tweets containing these terms, resulting in a set of tweets from about 33,000 users.
  - They took a timeline from all these users resulting in a set of roughly 85 million Tweets.
  - From the set of about 85 million tweets, they took a random sample, of 25k tweets, that contained terms from the lexicon.
  - Via **crowdsourcing**, they annotated each tweet as **hate speech, offensive (but not hate speech), or neither hate speech nor offensive**.
  - If the agreement between annotators was too low, the tweet was excluded from the set.
  - A commonly-used subset of this dataset is also available, containing 14,510 tweets.

HatebaseTwitter

	Hate	Offensive	Neither
Unigram	racist	c*nts	mock
	queer	n*ccas	oreo
	spic	n*ggah	colored
	f*g	c*nt	brownies
	f*gs	sh*t	birds
	white	h*e	trash
	n*ggers	h*es	bird
	f*ggots	p*ssy	yellow
	n*gger	b*tches	charlie
	f*ggot	b*tch	yankees
2-gram	ugly d*ke	b*tch i'm	derek jeter
	black people	ass b*tch	rt yankees
	you're f*cking	p*ssy http	like trash
	biggest f*ggot	yo b*tch	charlie sheen
	stupid n*gger	b*tch ass	planet apes
	n*gger music	bad b*tches	trash talk
	f*ggot ass	h*es http	charlie crist
	you're f*ggot	h*e http	charlie brown
	f*cking f*ggot	bad b*tch	early bird
	white trash	like b*tch	flappy bird
3-gram	vanessa f*ckin f*ggot	b*tch ass n*gga	yellow http t
	vinniepolitan kill coons	don't love h*es	brownies http t
	f*ggot rant night	f*ck right p*ssy	new york yankees
	amarierubio ch*nk eyed	like http t	early bird catches
	n*ggah know wassup	b*tches http t	bird catches worm
	runuldorants f*ggot sack	p*ssy http t	look like trash
	n*ggas retarded lmfaoo	h*es http t	bird http t
	happy birthday f*ggot	h*e http t	early bird gets
	creepy ass cracker	h*es ain't loyal	bird gets worm
	south white trash	b*tch http t	trash http t

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Waseem's Datasets https://github.com/zeerakw/hatespeech

- **Waseem and Hovy** provide a dataset from Twitter, consisting of 16,914 tweets labeled as **racist, sexist, or neither**.
  - They first created a corpus of about 136,000 tweets that contain slurs and terms related to **religious, sexual, gender, and ethnic minorities**.
  - From this corpus, the authors themselves annotated 16,914 tweets and had a gender studies major review the annotations.
- Waseem creates another dataset by sampling a new set of tweets from the 136,000 tweet corpus.
  - Waseem recruited **feminists and anti-racism activists along with crowdsourcing** for the annotation of the tweets.
  - The labels therein are **racist, sexist, neither or both**.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Hateful symbols or hateful people? [Waseem and Hovy, 2016]

	All	Racism	Sexism	Neither
Men	50.08%	33.33%	50.24%	50.92%
Women	02.26%	0.00 %	02.28%	01.74%
Unidentified	47.64%	66.66%	47.47%	47.32%

Table 1: Distribution of genders in hate-speech documents.

Sexism	Distribution	Racism	Distribution
not	1.83%	islam	1.44%
sexist	1.68%	muslims	1.01%
#mkr	1.57%	muslim	0.65%
women	0.83%	not	0.53%
kat	0.57%	mohammed	0.52%
girls	0.48%	religion	0.40%
like	0.42%	isis	0.38%
call	0.36%	jews	0.37%
#notsexist	0.36%	prophet	0.36%
female	0.34%	#islam	0.35%

Table 2: Distribution of ten most frequently occurring terms

SOPHIA MASTER CLASSES by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Stormfront Dataset

- de Gilbert et al. provide a dataset from posts **from a white supremacist forum, Stormfront**.
  - They annotate the posts at sentence level resulting in 10,568 sentences labeled with **Hate, NoHate, Relation, or Skip**.
  - Hate and NoHate labels indicate presence or lack thereof, respectively, of hate speech in each sentence.
  - The label Relation indicates that **the sentence is hate speech when it is combined with the sentences around it**.
  - The label **Skip** is for sentences that are non-English or not containing information related to hate or non-hate speech.
  - They also capture the **amount of context** (i.e., previous sentences) that an annotator used to classify the text.

SOPHIA MASTER CLASSES
by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Stormfront

	Hate	Not Hate
Unigram	jews black n*gro scum race white ape africa asian place	youtube thank welcome pm check idea sf link happy join
2-gram	non white race mixing crippin n*gga black people like blacks white woman race traitors white countries blacks asians	welcome sf watch tv sounds like thank posting good luck years ago like minded home school yankee jim
3-gram	homosexuals stay closet way advantage state liberals care diversity reality money charade pakis forcing culture think exempt rules i'm winnipeg cesspool wonder races achieve maximum resistance zog white genocide project	like minded people love big dog treading ice jimmy like comment link readily happy welcome sir thomas lawrence 30 years ago think write book hope talk later camellia idea great

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

TRAC Dataset

https://sites.google.com/view/trac1/shared-task

- The 2018 Workshop on Trolling, Aggression, and Cyberbullying (TRAC) hosted a **shared task focused on detecting aggressive text in both English and Hindi**.
- Aggressive text is often a component of hate speech.
- The dataset from this task is available to the public and contains **15,869 Facebook comments** labeled as **overtly aggressive, covertly aggressive, or non-aggressive**.
- There is also a small **Twitter dataset**, consisting **of 1,253 tweets**, which has the same labels.

SOPHIA MASTER CLASSES
by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

TRAC(Facebook)

	NAG	CAG	OAG
Unigram	invest	black	islam
	good	bike	worst
	buy	bjp	fool
	hi	burnol	terrorist
	cnbc	mother	hell
	tata	cash	idiots
	nifty	old	bloody
	proud	modiji	stupid
	market	men	idiot
	anuj	reservation	shame
2-gram	hi sonia	poor people	dont forget
	time buy	sonu right	shame u
	news jansatta	ha ha	cheap publicity
	royal enfield	common people	u people
	ratan tata	old man	like u
	short term	common man	gone mad
	mukesh ambani	indian express	wrong decision
	hi anuj	political parties	indian express
	bank nifty	modi ji	agent bjp
	long term	black money	sonu nigam
3-gram	reason hcltech fall	akhlaq killer draped	seriously need education
	advice bank nifty	killer draped tricolor	u seriously need
	bank nifty npa	black money holders	hate sonu nigam
	hi anuj sonia	indian express mind	old man useless
	soul rest peace	banned banned religions	shame indian express
	real surgical strike	good sonu nigam	powerful man world
	anuj sonia view	common people suffering	man world dont
	good time buy	surgical strike lol	9th powerful man
	long term view	owaisi mamta begum	world dont forget
	cnbc tv 18	trained kejriwal d	vote bank politics

SOPHIA MASTER CLASSES
by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

HatEval@SemEval 2019 — Task 5

https://competitions.codalab.org/competitions/19935

- This dataset is for competition on **multilingual detection of hate targeting to women and immigrants in tweets**.
  - It consists of several sets of labels.
  - The first indicates **whether the tweet expresses hate towards women or immigrants**, the second, **whether the tweet is aggressive**, and the third, **whether the tweet is directed at an individual or an entire group**.
  - Note that targeting an individual is not necessarily considered hate speech by all definitions.

SOPHIA MASTER CLASSES
by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

HatEval

	Hate	Not Hate
Unigram	h*e	immigrant
	nodaca	men
	wh*re	ram
	maga	k*nt
	buildthewall	son
	b*tches	calling
	illegal	h*
	womensuck	stand
	buildthatwall	rohingya
	b*tch	thank
2-gram	b*tch f*ck	immigrant children
	b*tch h*e	immigrant families
	ass b*tch	anti immigrant
	nodaca noamnesty	men
	women stupid	migrants https
	trump maga	men women
	stupid b*tch	rohingya refugees
	illegal aliens	men like
	illegal immigrants	men men
	illegal alien	immigrant parents
3-gram	senkamalaharris hysterical woman	migrants https t
	build wall buildthatwall	refugees https t
	need wall buildthatwall	immigration https t
	speech time https	says https t
	free speech time	life https t
	buildthewall lockthemup enddaca	children https t
	trump maga rednationrising	woman accused nelly
	realdonaldtrump buildthewall lockthemup	accused nelly rape
	b*tch https t	today https t
	potus realdonaldtrump buildthewall	unitednations https t

SOPHIA MASTER CLASSES
by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Kaggle Dataset

https://www.kaggle.com/c/detecting-insults-in-social-commentary

- Kaggle.com hosted a shared task on **detecting insulting comments**.
  - The dataset consists of 8,832 social media comments labeled as insulting or not insulting.
  - While not necessarily hate speech, **insulting text may indicate hate speech**.

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

GermanTwitter Dataset

- Ross et al. created a Twitter dataset **in German for the European refugee crisis**.
- It consists of **541 tweets in German**, labeled as **expressing hate or not**.

SOPHIA MASTER CLASSES
by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Dataset	Labels and percents in dataset	Origin Source	Language
HatebaseTwitter [9]	Hate 5% Offensive 76% Neither 17%	Twitter	English
WaseemA [17]	Racism 12% Sexism 20% Neither 68%	Twitter	English
WaseemB [18]	Racism1 1% Sexism 13% Neither 84% Both 1%	Twitter	English
Stormfront [14]	Hate 11% Not Hate 86% Relation 2% Skip 1%	Online Forum	English
TRAC (Facebook) [19]	Non-aggressive 69% Overtly agg. 16% Covertly agg. 16%	Facebook	English & Hindi
TRAC (Twitter) [19]	Non-aggressive 38% Overtly agg. 29% Covertly agg. 33%	Twitter	English & Hindi
HatEval [20]	Hate 43% / Not Hate 57% Agg. / Not agg. roup / Individual	Twitter	English & Spanish
Kaggle [21]	Insulting 26% Not Insulting 74%	Twitter	English
GermanTwitter (Expert 1 annotation) [11]	Hate 23% Not Hate 77%	Twitter	German

https://doi.org/10.1371/journal.pone.0221152.t001

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Issues about the datasets

- These datasets **vary considerably in their size, scope, characteristics of the data annotated, and characteristics of hate speech considered**.
- The most common source of text is Twitter, which consists of short-form online posts.
- Corpora constructed from social media and websites other than Twitter are rare, making analysis of hate speech difficult to cover the entire landscape.
- There is also **the issue of imbalance in the number of hate and not hate texts within datasets**.
- On a platform such as Twitter, hate speech occurs at a very low rate compared to non-hate speech.
- Although datasets reflect this imbalance to an extent, they do not map the actual percentage due to training needs, e.g., in the Waseem and Hovy dataset, 20% of the tweets were labelled sexist, 11.7% racist, and 68.3% neither. There is still an imbalance in the number of sexist, racist, or neither tweets, but it may not be as imbalanced as expected on Twitter.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Automatic approaches to hate speech detection

- Most social media platforms have established user **rules that prohibit hate speech**
- Enforcing these rules requires copious manual labor to review every report.
- Some platforms, such as Facebook and Twitter, recently increased the number of **content moderators**.
- **Automatic tools and approaches could accelerate the reviewing process and allocate the human resource to the posts that require close human examination.**

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Keyword Based Approaches

- By using an **ontology** or **dictionary**, text that contain potentially hateful keywords are identified.
- For instance, Hatebase maintains a database of derogatory terms for many groups across 95 languages.
- Such well-maintained resources are valuable, as terminology changes over time.
- However, **simply using a hateful slur is not enough to detect hate speech.**

SOPHIA MASTER CLASSES by Telecom Valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Keyword Based Approaches

- Keyword-based approaches are fast and straightforward to understand.
- However, they have **severe limitations**:
  - Detecting only racial slurs would result in a **highly precise system but with low recall** (where precision is the percentage of relevant from the set detected and recall is the percent of relevant from within the global population).
  - It would not identify hateful content that does not use these terms.
  - Including terms that could but **are not always hateful** (e.g., “trash”, “swine”, etc.) would create too **many false alarms**, increasing recall at the expense of precision.
- Keyword-based approaches cannot identify hate speech that does not have any hateful keywords (e.g., figurative or nuanced language).
  - Slang such as “build that wall” literally means constructing a physical barrier (wall). However, with the political context, some interpret this is a condemnation of some immigrates in the United States.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Source Metadata

- Additional information from social media can help further understand the characteristics of the posts and potentially lead to a better identification approach, i.e., **demographics of the posting user**, **location**, **timestamp**, or even **social engagement on the platform**.
- However, **this information is not often readily available to external researchers as publishing data with sensitive user information raises privacy issues**.
- **Risk of bias**: a system trained on these data might naturally bias towards flagging content by certain users or groups as hate speech based on incidental dataset characteristics.
- **Using user information potentially raises some ethical issues:** bias against certain users and frequently flag their posts as hateful even if some of them are not.
- Relying too much on demographic information could miss posts from users who do not typically post hateful content.
- Flagging posts as hate based on user statistics could create a chilling effect on the platform and eventually **limit freedom of speech**.

SOPHIA MASTER CLASSES
by Telecom Valley | Animateur Azuréen Numérique
17 NOVEMBRE 2020

Machine Learning Approaches to Hate Speech Detection

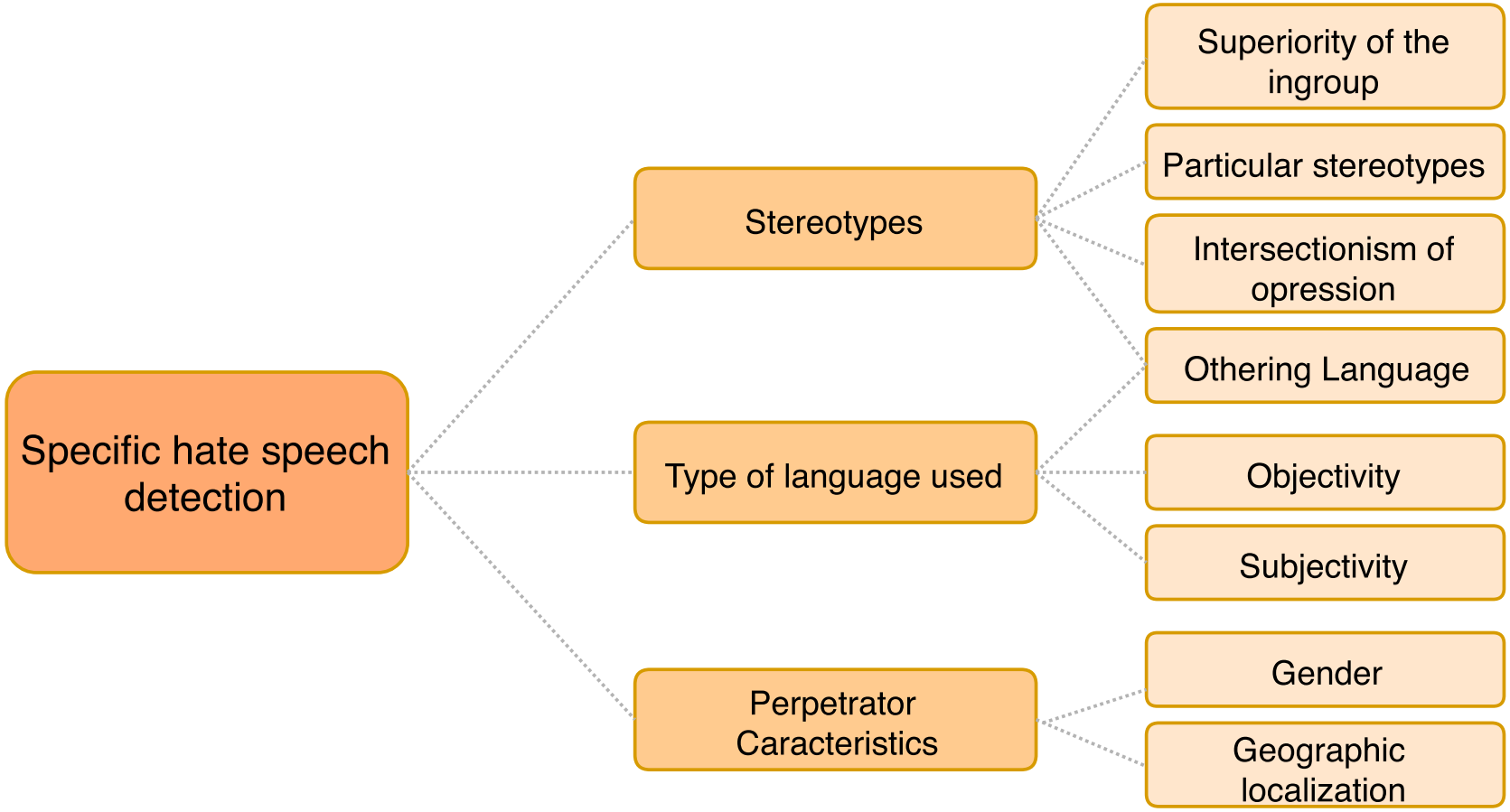
Content preprocessing and feature selection.

- To classify user-generated content, text features indicating hate must be extracted.
- Standard features are **individual words or phrases** (**n-grams**, i.e., sequence of n consecutive words).
- To improve the matching of features, words can be stemmed to obtain only the root removing morphological differences.
- **Bag-of-words: a post is represented as a set of words or n-grams without any ordering.** Various ways to assign weights to the terms that may be more important, such as TF-IDF.
- **Distributional features, word embeddings**, i.e., assigning a vector to a word, such as word2vec, are common when applying deep learning methods in Natural Language Processing.
- **Sentiment and emotions**.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Features for hate speech



SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Some Hate Speech Detection Approaches and Baselines

Naïve Bayes, Support Vector Machine and Logistic Regression.

- These models are commonly used in text categorization.
- Naïve Bayes (NB) models label probabilities directly with the assumption that the features do not interact with one another.
- Support Vector Machines (SVMs) and Logistic Regression are **linear classifiers** that predict classes based on a combination of scores for each feature.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Some Hate Speech Detection Approaches and Baselines

[Davidson et al.,2017] proposed a state-of-the-art feature-based classification model that incorporates **distributional TF-IDF features**, **part-of-speech tags**, and other linguistic features using **SVMs**.

The incorporation of these linguistic features helps identify hate speech by distinguishing between different usages of the terms.

Still suffers from some subtleties, such as when typically offensive terms are used in a positive sense: queer in "*He's a damn good actor. As a gay man, it's awesome to see an openly queer actor given the lead role for a major film*.", from Hatebase Twitter dataset

Some Hate Speech Detection Approaches and Baselines

Neural Ensemble

[Zimmerman et al., 2018] propose an ensemble approach, which combines the decisions of ten **convolutional neural networks (CNN)** with different weight initializations, with convolutions of length 3 pooled over the entire document length.

The results of each model are combined by averaging the scores.

SOPHIA MASTER CLASSES
by Telecom valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

Some Hate Speech Detection Approaches and Baselines

- C-GRU, a Convolution-GRU Based Deep Neural Network proposed by [Zhang et al., 2018] combines Convolutional Neural Networks (CNN) and Gated Recurrent Networks (GRU) to **detect hate speech on Twitter**.
- In the HatebaseTwitter dataset, they treat both Hate and Offensive as Hate resulting in a binary classification task instead of its original multi-class task.

SOPHIA MASTER CLASSES by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Some Hate Speech Detection Approaches and Baselines

- **FastText** is an efficient classification model proposed by researchers in Facebook. The model produces **embeddings of character n-grams** and provides predictions of the example based on the embeddings.
- **BERT** is a recent **transformer-based pre-trained contextualized embedding model** extendable to a classification model with an additional output layer. State-of-the-art performance in text classification, question answering, and language inference without substantial task-specific modifications. Further BERT-based language models: ROBERTA, ALBERT, DistilBERT.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Results

Stormfront

Method	Accuracy All	Accuracy Hate	Accuracy Not Hate	Macro F_1
Naïve Bayes	0.6423	0.6164	0.6828	0.6378
SVM	0.7469	0.7438	0.7500	0.7469
Logistic Regression	0.7218	0.7155	0.7280	0.7217
FastText	0.6506	0.6406	0.6622	0.6502
Davidson et al. [9]	0.7364	0.7344	0.7384	0.7364
Neural Ensemble [10]	0.8033	0.7736	0.8404	0.8027
mSVM (ours)	0.8033	0.8251	0.7843	0.8031
BERT [26]	0.8201	0.8255	0.8148	0.8201
C-GRU [33]	0.6297	0.5969	0.6962	0.6188

HatEval

Method	Accuracy All	Accuracy Hate	Accuracy Not Hate	Macro F_1
Naïve Bayes	0.6800	0.6253	0.7208	0.6730
SVM	0.7190	0.6594	0.7694	0.7152
Logistic Regression	0.7340	0.6801	0.7776	0.7297
FastText	0.7380	0.6846	0.7812	0.7338
Davidson et al. [9]	0.7390	0.6869	0.7806	0.7346
Neural Ensemble [10]	0.7470	0.6867	0.7996	0.7441
mSVM (ours)	0.7590	0.7143	0.7933	0.7543
BERT [26]	0.7480	0.6866	0.8023	0.7452
C-GRU [33]	0.6670	0.6399	0.6802	0.6471

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019)
Hate speech detection: Challenges and solutions. PLoS ONE 14(8)

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Results

17 NOVEMBRE 2020

TRAC(Facebook)

Method	Accuracy				Macro
	All	NAG[1]	CAG[2]	OAG[3]	F_1
Naïve Bayes	0.4758	0.8306	0.2111	0.2640	0.4080
SVM	0.5714	0.8326	0.2444	0.4762	0.5050
Logistic Regression	0.5556	0.8450	0.2421	0.4438	0.5001
FastText	0.5626	0.8326	0.2246	0.4800	0.4879
Davidson et al. [9]	0.5604	0.8428	0.2259	0.4235	0.4875
Neural Ensemble [10]	0.5358	0.8575	0.2306	0.4647	0.4945
mSVM (ours)	0.6121	0.8479	0.2589	0.5202	0.5368
BERT [26]	0.5809	0.8538	0.2516	0.4881	0.5234
C-GRU [33]	0.4769	0.7436	0.1818	0.2156	0.3696

[1] not aggressive [2] covertly aggressive [3] overtly aggressive

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019)
Hate speech detection: Challenges and solutions. PLoS ONE 14(8)

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Results

17 NOVEMBRE 2020

HatebaseTwitter

Method	Accuracy All	Hate[1]	Off.[2]	N[3]	Macro F_1
Naïve Bayes	0.8297	0.2571	0.8418	0.7940	0.5138
SVM	0.9092	0.5429	0.9326	0.8282	0.6788
Logistic Regression	0.9149	0.5000	0.9409	0.8356	0.6914
FastText	0.9068	0.5100	0.9346	0.8220	0.6930
Davidson et al. [9]	0.9007	0.6098	0.9270	0.8033	0.6877
Neural Ensemble [10]	0.9213	0.5179	0.9453	0.8628	0.7218
mSVM (ours)	0.9108	0.4961	0.9585	0.8251	0.7704
BERT [26]	0.9209	0.4857	0.9499	0.8917	0.7609
C-GRU [33]	0.8588	0.5556	0.9065	0.6550	0.5651

[1] hate speech [2] offensive language [3] neither

TRAC(Facebook)

Method	Accuracy All	NAG[1]	CAG[2]	OAG[3]	Macro F_1
Naïve Bayes	0.4758	0.8306	0.2111	0.2640	0.4080
SVM	0.5714	0.8326	0.2444	0.4762	0.5050
Logistic Regression	0.5556	0.8450	0.2421	0.4438	0.5001
FastText	0.5626	0.8326	0.2246	0.4800	0.4879
Davidson et al. [9]	0.5604	0.8428	0.2259	0.4235	0.4875
Neural Ensemble [10]	0.5358	0.8575	0.2306	0.4647	0.4945
mSVM (ours)	0.6121	0.8479	0.2589	0.5202	0.5368
BERT [26]	0.5809	0.8538	0.2516	0.4881	0.5234
C-GRU [33]	0.4769	0.7436	0.1818	0.2156	0.3696

[1] not aggressive [2] covertly aggressive [3] overtly aggressive

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: Challenges and solutions. PLoS ONE 14(8)

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

Beyond hate speech

17 NOVEMBRE 2020

- **Hate speech**: targets individual or groups on the basis of their characteristics; demonstrates a clear intention to incite harm, or to promote hatred; may or may not use offensive or profane words.

  Assimilate? No they all need to go back to their own countries. #BanMuslims Sorry if someone disagrees too bad.

- **Abusive language**: bears the purpose of insulting individuals or groups, and can include hate speech, derogatory and offensive language.

  All you perverts (other than me) who posted today, needs to leave the O Board.

  I spend my money how i want bitch its my business

- **Bullying**: has the purpose to harass, threaten or intimidate typically individuals rather than groups.

  Our class prom night just got ruined because u showed up. Who invited u anyway?

SOPHIA MASTER CLASSES
by Telecom valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Beyond hate speech

Concept	Definition of the concept	Distinction from hate speech
Hate	Expression of hostility without any stated explanation for it [68].	Hate speech is hate focused on stereotypes, and not so general.
Cyberbullying	Aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend him or herself [10].	Hate speech is more general and not necessarily focused on a specific person.
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment [69].	Hate speech is a form of discrimination, through verbal means.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [35]	Hate speech can occur in any context, whereas flaming is aimed toward a participant in the specific context of a discussion.
Abusive language	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity [58].	Hate speech is a type of abusive language.
Profanity	Offensive or obscene word or phrase [23].	Hate speech can use profanity, but not necessarily.
Toxic language or comment	Toxic comments are rude, disrespectful or unreasonable messages that are likely to make a person to leave a discussion [43].	Not all toxic comments contain hate speech. Also some hate speech can make people discuss more.
Extremism	Ideology associated with extremists or hate groups, promoting violence, often aiming to segment populations and reclaiming status, where outgroups are presented both as perpetrators or inferior populations. [55].	Extremist discourses use frequently hate speech. However, these discourses focus other topics as well [55], such as new members recruitment, government and social media demonization of the in-group and persuasion [62].
Radicalization	Online radicalization is similar to the extremism concept and has been studied on multiple topics and domains, such as terrorism, anti-black communities, or nationalism [2].	Radical discourses, like extremism, can use hate speech. However in radical discourses topics like war, religion and negative emotions [2] are common while hate speech can be more subtle and grounded in stereotypes.

SOPHIA MASTER CLASSES by Telecom Valley Animateur Azuréen Numérique

17 NOVEMBRE 2020

Beyond hate speech

- **Knowledge-Based features** such as messages mapped to stereotypical concepts in a knowledge base
- **Multimodal information** such as **image captions**, **pixel features** and **videos** are used in cyberbullying detection.
- **Author profiling for abusive language detection** [Mishra et al., 2018]: incorporate community-based profiling features of Twitter users, outperform state-of-the-art results.
- **Counter narratives to fight hate speech** online [Guerini et al., 2020]

References

SOPHIA MASTER CLASSES by Telecom Valley | Animateur Azuréen Numérique

17 NOVEMBRE 2020

- Fortuna, P. and S. Nunes. “A Survey on Automatic Detection of Hate Speech in Text.” ACM Computing Surveys (CSUR) 51 (2018): 1 - 30.
- MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: Challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152
- Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: The 3rd Workshop on Natural Language Processing for Computer-Mediated Communication @ Conference on Natural Language Processing; 2016.
- Wermiel SJ. The Ongoing Challenge to Define Free Speech. Human Rights Magazine. 2018;43(4):1–4.
- Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM 2017.
- Waseem Z, Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: SRW@HLT-NAACL; 2016.
- Waseem Z. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Proceedings of the first workshop on NLP and computational social science; 2016. p. 138–142.
- de Gibert O, Perez N, Garc’ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive Language Online @ EMNLP; 2018.
- Kumar R, Ojha AK, Malmasi S, Zampieri M. Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). ACL; 2018. p. 1–11.
- Zhang Z, Robinson D, Tepper J. Detecting hate speech on twitter using a convolution-gru based deep neural network. In: European Semantic Web Conference. Springer; 2018. p. 745–760.
- Zimmerman S, Kruschwitz U, Fox C. Improving Hate Speech Detection with Deep Learning Ensembles. In: LREC; 2018.