Hate and Abuse Detection

An Overview

Agostina Calabrese

calabrese.a@di.uniroma1.it
SapienzaNLP
Sapienza University of Rome

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

How do we define "hate speech" and "abuse"?

• bn:00629957n • NOME • Concetto • Categorie: Censorship, Ethically disputed political practices, Freedom of speech, Hate crime...

EN hate speech

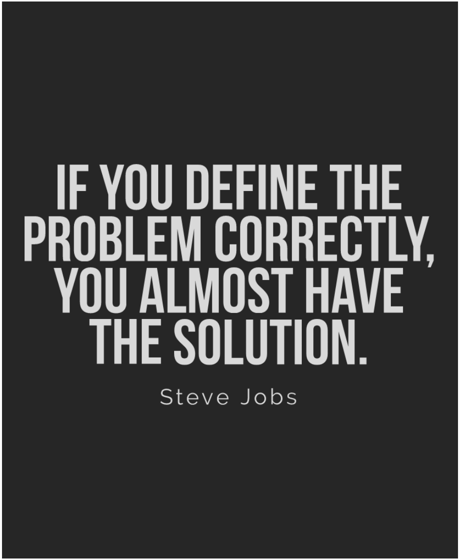
Hate speech is speech which attacks a person or group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. *Wikipedia* *Più definizioni*

• bn:00000521n • NOME • Concetto • Categorie: Abuse, Bullying, Psychological abuse

EN abuse · insult · revilement · contumely · vilification

A rude expression intended to offend or hurt *WordNet* *Più definizioni*

SAPIENZA NLP



Difficulties in annotating abuse:

- Lack of standard definitions
- Differences in annotators' cultural background
- Ambiguity in the annotation guidelines

SAPIENZA NLP

Issues in data annotation

Figure: "Il bambino con lo zerbino" by Federico Clapis

SAPIENZA NLP

Issues in data annotation

Figure: "Il bambino con lo zerbino" by Federico Clapis

- Is this artwork racist?

Figure: "Il bambino con lo zerbino" by Federico Clapis

- Is this artwork racist?
- **Only the target gets to decide**

Issues in data annotation (2)

Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

SAPIENZA NLP

Issues in data annotation (2)

Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

- We need:
  - Context

SAPIENZA NLP

Issues in data annotation (2)

Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

- We need:
  - Context
  - Information about author and target

Issues in data annotation (2)

Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

- We need:
  - Context
  - Information about author and target
- Good luck with ethical issues..

Table of Contents

SAPIENZA NLP

But we must have something! ..right?

- Datasets characterised by:
  - **Source:** Twitter, Facebook, Reddit, Wikipedia
  - **Composition:** e.g., racism and sexism, or personal attack and racism, or hate speech and profanity
  - **Language:** English, followed by German, Hindi and Dutch.

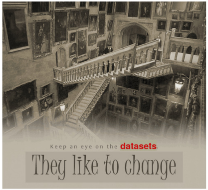
name	publication	source	microposts	%abusive
Kaggle†	(Wulczyn et al., 2017)	Wikipedia	312,737	9.6
Founta	(Founta et al., 2018)	Twitter	59,357	14.1
Razavi	(Razavi et al., 2010)	diverse	1,525	31.9
Warner	(Warner and Hirschberg, 2012)	diverse	3,438	14.3
Waseem	(Waseem and Hovy, 2016)	Twitter	16,165	35.3
Kumar	(Kumar et al., 2018)	Facebook	15,000	58.1

Figure: Table from [Wiegand et al., 2019]

SAPIENZA NLP

But we must have something! ..right?

- Datasets characterised by:
  - **Source:** Twitter, Facebook, Reddit, Wikipedia
  - **Composition:** e.g., racism and sexism, or personal attack and racism, or hate speech and profanity
  - **Language:** English, followed by German, Hindi and Dutch.



name	publication	source	microposts	%abusive
Kaggle†	(Wulczyn et al., 2017)	Wikipedia	312,737	9.6
Founta	(Founta et al., 2018)	Twitter	59,357	14.1
Razavi	(Razavi et al., 2010)	diverse	1,525	31.9
Warner	(Warner and Hirschberg, 2012)	diverse	3,438	14.3
Waseem	(Waseem and Hovy, 2016)	Twitter	16,165	35.3
Kumar	(Kumar et al., 2018)	Facebook	15,000	58.1

Figure: Table from [Wiegand et al., 2019]

SAPIENZA NLP

The Problem of Biased Datasets [Wiegand et al., 2019]

- There are issues in data collection as well!
  - Can't rely on random sampling: only 0.1% to 3% of posts are abusive

name	publication	source	microposts	%abusive	sampling	%explicit*
Kaggle†	(Wulczyn et al., 2017)	Wikipedia	312,737	9.6	boosted random sampling	76.9
Founta	(Founta et al., 2018)	Twitter	59,357	14.1	boosted random sampling	75.9
Razavi	(Razavi et al., 2010)	diverse	1,525	31.9	boosted random sampling	64.7
Warner	(Warner and Hirschberg, 2012)	diverse	3,438	14.3	biased sampling	51.3
Waseem	(Waseem and Hovy, 2016)	Twitter	16,165	35.3	biased sampling	44.4
Kumar	(Kumar et al., 2018)	Facebook	15,000	58.1	biased sampling	32.7

Figure: Table from [Wiegand et al., 2019]

SAPIENZA NLP

The Problem of Biased Datasets [Wiegand et al., 2019]

- There are issues in data collection as well!
  - Can't rely on random sampling: only 0.1% to 3% of posts are abusive
- What about **focused sampling**?
  - Boosted random sampling: random sampling + heuristics (fails at capturing implicit abuse)
  - Biased sampling: manual selection of query words and topics

name	publication	source	microposts	%abusive	sampling	%explicit*
Kaggle†	(Wulczyn et al., 2017)	Wikipedia	312,737	9.6	boosted random sampling	76.9
Founta	(Founta et al., 2018)	Twitter	59,357	14.1	boosted random sampling	75.9
Razavi	(Razavi et al., 2010)	diverse	1,525	31.9	boosted random sampling	64.7
Warner	(Warner and Hirschberg, 2012)	diverse	3,438	14.3	biased sampling	51.3
Waseem	(Waseem and Hovy, 2016)	Twitter	16,165	35.3	biased sampling	44.4
Kumar	(Kumar et al., 2018)	Facebook	15,000	58.1	biased sampling	32.7

Figure: Table from [Wiegand et al., 2019]

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words
  - ..but query words correlate (PMI) with the classes of the dataset!

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words
  - ..but query words correlate (PMI) with the classes of the dataset!
  - Distribution of abusive tweets is highly skewed towards 3 authors

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words
  - ..but query words correlate (PMI) with the classes of the dataset!
  - Distribution of abusive tweets is highly skewed towards 3 authors

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

→ Focused sampling introduces **data bias**!

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words
  - ..but query words correlate (PMI) with the classes of the dataset!
  - Distribution of abusive tweets is highly skewed towards 3 authors

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

→ Focused sampling introduces **data bias**!

→ High classification scores are likely to be the result of modeling the bias in the datasets

SAPIENZA NLP

The Problem of Biased Datasets (2) [Wiegand et al., 2019]

- Waseem and Hovy [2016]:
  - Most used dataset
  - Extract tweets matching query words
  - ..but query words correlate (PMI) with the classes of the dataset!
  - Distribution of abusive tweets is highly skewed towards 3 authors

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

→ Focused sampling introduces **data bias**!

→ High classification scores are likely to be the result of modeling the bias in the datasets

→ We won't talk about SOTA

Table of Contents

SAPIENZA NLP

Hate and Abuse Detection Systems

- Systems are meant to alert and support human moderators:
  - Precision: FPs turn into infringement of free speech
  - Recall: FNs correspond to victims which the system fails to protect

Hate and Abuse Detection Systems

- Systems are meant to alert and support human moderators:
  - Precision: FPs turn into infringement of free speech
  - Recall: FNs correspond to victims which the system fails to protect

- **Rule-based** Approach [Spertus, 1997]
  - Reading is an experience: *awk* and *sed* scripts for hate detection

SAPIENZA NLP

Hate and Abuse Detection Systems

- Systems are meant to alert and support human moderators:
  - Precision: FPs turn into infringement of free speech
  - Recall: FNs correspond to victims which the system fails to protect

- **Rule-based** Approach [Spertus, 1997]
  - Reading is an experience: *awk* and *sed* scripts for hate detection
  - Rules: noun phrases used as appositions ("you flamers"), imperative statements, bad-words, condescending statements ("isn't it?"), etc.

SAPIENZA NLP

Hate and Abuse Detection Systems

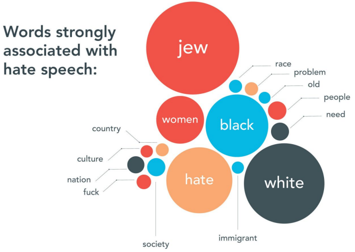
- Systems are meant to alert and support human moderators:
  - Precision: FPs turn into infringement of free speech
  - Recall: FNs correspond to victims which the system fails to protect

- **Rule-based** Approach [Spertus, 1997]
  - Reading is an experience: *awk* and *sed* scripts for hate detection
  - Rules: noun phrases used as appositions ("you flamers"), imperative statements, bad-words, condescending statements ("isn't it?"), etc.
  - Limitations: sarcasm, complex sentences and errors in spelling, punctuation and grammar

SAPIENZA NLP

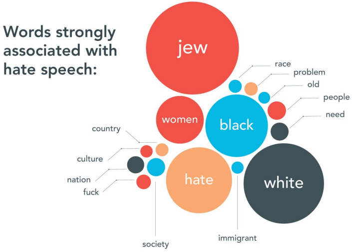
Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact



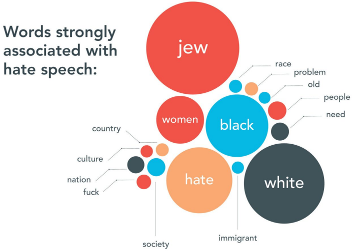
Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries



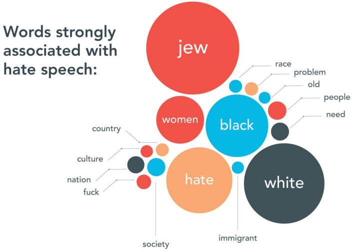
Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier



Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier

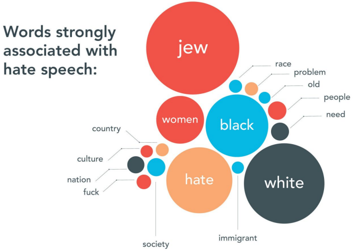


- Gitari et al. [2015]:
  - Isolate subjective sentences through rule-based approach

SAPIENZA NLP

Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier

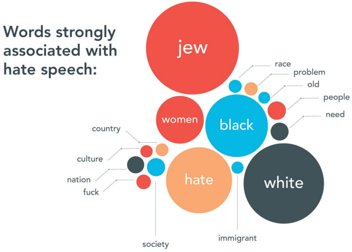


- Gitari et al. [2015]:
  - Isolate subjective sentences through rule-based approach
  - Build lexicon using negative polarity, hate verbs and theme-based grammatical patterns (i.e., religion, race, nationality)

SAPIENZA NLP

Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier

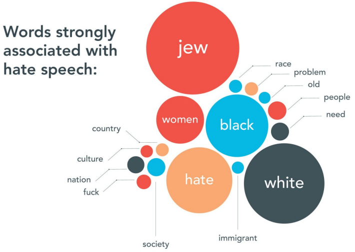


- Gitari et al. [2015]:
  - Isolate subjective sentences through rule-based approach
  - Build lexicon using negative polarity, hate verbs and theme-based grammatical patterns (i.e., religion, race, nationality)
  - Rule-based classifier: number of negative opinionated words

SAPIENZA NLP

Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier

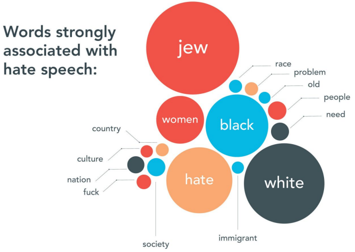


- Gitari et al. [2015]:
  - Isolate subjective sentences through rule-based approach
  - Build lexicon using negative polarity, hate verbs and theme-based grammatical patterns (i.e., religion, race, nationality)
  - Rule-based classifier: number of negative opinionated words
- Wiegand et al. [2018]:
  - Automated framework for generating hate speech lexicons

SAPIENZA NLP

Lexicon-Based Approaches

- Razavi et al. [2010]:
  - IALD: Collect 2700 words and phrases and associate them with weights indicating their abusive impact
  - Features: frequency, max weight and normalised avg weight of IALD entries
  - Naïve Bayes classifier



- Gitari et al. [2015]:
  - Isolate subjective sentences through rule-based approach
  - Build lexicon using negative polarity, hate verbs and theme-based grammatical patterns (i.e., religion, race, nationality)
  - Rule-based classifier: number of negative opinionated words
- Wiegand et al. [2018]:
  - Automated framework for generating hate speech lexicons
- Work well on explicit posts, **limitations on implicit abuse**

SAPIENZA NLP

Exploiting User Profiling [Mishra et al., 2019a]

- Employ NN to extract features for users instead of manually leveraging gender, location, etc.

Exploiting User Profiling [Mishra et al., 2019a]

- Employ NN to extract features for users instead of manually leveraging gender, location, etc.
- Capture structure of online communities and linguistic behaviour of users
  - *Homophily*: people tend to cluster with those who appear similar to themselves

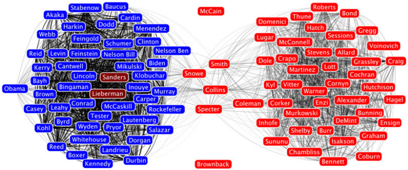


Figure: Graph by Human-computer Interaction Lab: nodes are senators (red for Republicans, blue for Democrats) and edges indicate the similarity of voting records

Exploiting User Profiling (2) [Mishra et al., 2019a]

- Data:
  - 16,202 tweets out of the 16,907 from Waseem and Hovy [2016]
  - 1,875 unique users
  - Classes: 12% racist, 19.4% sexist, 68.6% clean

Exploiting User Profiling (2) [Mishra et al., 2019a]

- Data:
  - 16,202 tweets out of the 16,907 from Waseem and Hovy [2016]
  - 1,875 unique users
  - Classes: 12% racist, 19.4% sexist, 68.6% clean
- Graphs:
  - Homogeneous **community graph**: nodes are the authors, undirected edges connect two authors if either one follows the other on Twitter
  - Heterogeneous **extended graph**: additionally contains nodes representing tweets, each tweet is connected to his author

SAPIENZA NLP

Exploiting User Profiling (2) [Mishra et al., 2019a]

- Data:
  - 16,202 tweets out of the 16,907 from Waseem and Hovy [2016]
  - 1,875 unique users
  - Classes: 12% racist, 19.4% sexist, 68.6% clean
- Graphs:
  - Homogeneous **community graph**: nodes are the authors, undirected edges connect two authors if either one follows the other on Twitter
  - Heterogeneous **extended graph**: additionally contains nodes representing tweets, each tweet is connected to his author
- Graph-based **semi-supervised** problem:
  - Only tweet nodes have labels!
  - The model should distribute gradient information from the supervised loss on the labeled nodes
  - How? **Graph Convolutional Network**

SAPIENZA NLP

Exploiting User Profiling (3) [Mishra et al., 2019a]

- Graph Convolutional Networks (informal):
  - Classic NNs would process nodes independently: $O = \sigma(F\,W)$
  - Can't use Convolutional layers: nodes have different number of neighbours
  - GCNs take into account the adjacency matrix A: $O = \sigma(A\,F\,W)$

SAPIENZA NLP

Exploiting User Profiling (3) [Mishra et al., 2019a]

- Graph Convolutional Networks (informal):
  - Classic NNs would process nodes independently: $O = \sigma(F\,W)$
  - Can't use Convolutional layers: nodes have different number of neighbours
  - GCNs take into account the adjacency matrix A: $O = \sigma(A\,F\,W)$
- Authors apply a 2-layer GCN to the extended graph:
  - $O = softmax(A\ ReLU(A\,F\,W^1)\,W^2)$ ..but how do we represent the input (F)?

SAPIENZA NLP

Exploiting User Profiling (3) [Mishra et al., 2019a]

- Graph Convolutional Networks (informal):
  - Classic NNs would process nodes independently: $O = \sigma(F\,W)$
  - Can't use Convolutional layers: nodes have different number of neighbours
  - GCNs take into account the adjacency matrix A: $O = \sigma(A\,F\,W)$
- Authors apply a 2-layer GCN to the extended graph:
  - $O = softmax(A\ ReLU(A\,F\,W^1)\,W^2)$ ..but how do we represent the input (F)?
  - Tweet as binary BOW, author as sum over all composed tweets

SAPIENZA NLP

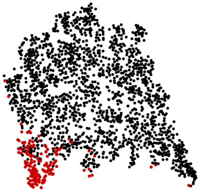
Exploiting User Profiling (3) [Mishra et al., 2019a]

- Graph Convolutional Networks (informal):
  - Classic NNs would process nodes independently: $O = \sigma(F\,W)$
  - Can't use Convolutional layers: nodes have different number of neighbours
  - GCNs take into account the adjacency matrix A: $O = \sigma(A\,F\,W)$
- Authors apply a 2-layer GCN to the extended graph:
  - $O = softmax(A\ ReLU(A\,F\,W^1)\,W^2)$ ..but how do we represent the input (F)?
  - Tweet as binary BOW, author as sum over all composed tweets
  - Extract **node embeddings**: $E = \sigma(A\,F\,W^1)$

SAPIENZA NLP

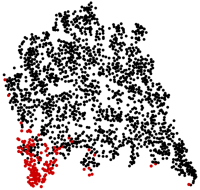
Exploiting User Profiling (4) [Mishra et al., 2019a]

- How do they classify tweets?
  - **GCN**: assign the label provided by the GCN



Exploiting User Profiling (4) [Mishra et al., 2019a]

- How do they classify tweets?
  - **GCN**: assign the label provided by the GCN
  - **LR + GCN**: author embedding is appended onto the tweet's character n-gram representation for training a Linear Regression classifier



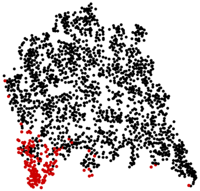
SAPIENZA NLP

Exploiting User Profiling (4) [Mishra et al., 2019a]

- How do they classify tweets?
  - **GCN**: assign the label provided by the GCN
  - **LR + GCN**: author embedding is appended onto the tweet's character n-gram representation for training a Linear Regression classifier

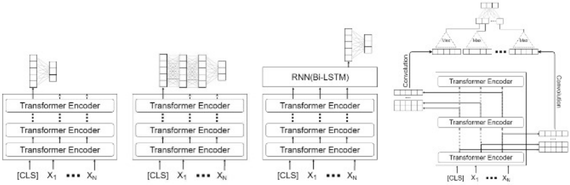
- LR + GCN performs better (or better models the bias in the dataset..)

Method	Racism			Sexism			Overall		
	P	R	F_1	P	R	F_1	P	R	F_1
GCN†	74.12	64.95	69.23	82.48	**82.22**	82.35	81.90	79.42	80.56
LR + GCN†	79.08	**79.90**	**79.49**	**88.24**	80.95	**84.44**	**86.23**	**84.73**	**85.42**



SAPIENZA NLP

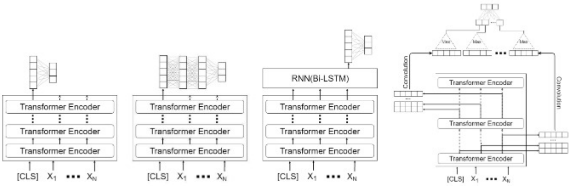
Where is BERT? [Mozafari et al., 2019]



(a) $\text{BERT}_{\text{base}}$ fine-tuning (b) Insert nonlinear layers (c) Insert Bi-LSTM layer (d) Insert CNN layer

- Pre-trained $\text{BERT}_{\text{base}}$ model
- Fine-tuning datasets: Waseem and Hovy [2016] and Davidson et al. [2017]

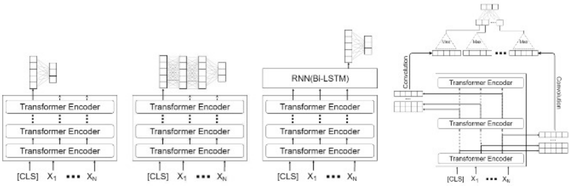
Where is BERT? [Mozafari et al., 2019]



(a) $\text{BERT}_{\text{base}}$ fine-tuning (b) Insert nonlinear layers (c) Insert Bi-LSTM layer (d) Insert CNN layer

- Pre-trained $\text{BERT}_{\text{base}}$ model
- Fine-tuning datasets: Waseem and Hovy [2016] and Davidson et al. [2017]
- Best performing model: BERT + CNN

Where is BERT? [Mozafari et al., 2019]

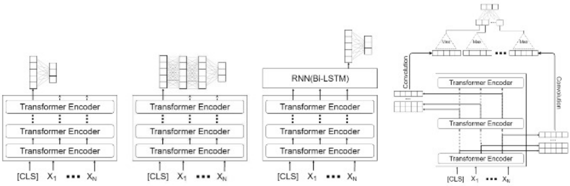


(a) $\text{BERT}_{\text{base}}$ fine-tuning (b) Insert nonlinear layers (c) Insert Bi-LSTM layer (d) Insert CNN layer

- Pre-trained $\text{BERT}_{\text{base}}$ model
- Fine-tuning datasets: Waseem and Hovy [2016] and Davidson et al. [2017]
- Best performing model: BERT + CNN
- Authors show model's ability to detect dataset's bias

SAPIENZA NLP

Where is BERT? [Mozafari et al., 2019]



(a) $\text{BERT}_{\text{base}}$ fine-tuning (b) Insert nonlinear layers (c) Insert Bi-LSTM layer (d) Insert CNN layer

- Pre-trained $\text{BERT}_{\text{base}}$ model
- Fine-tuning datasets: Waseem and Hovy [2016] and Davidson et al. [2017]
- Best performing model: BERT + CNN
- Authors show model's ability to detect dataset's bias
- "It can be a valuable clue in using pre-trained BERT model for debiasing hate speech datasets in future studies"

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Open Problems [Mishra et al., 2019b]

- **Ethical challenges** in user profiling:
  - Is profiling based on identity traits of users or on their online behaviour?
  - Is the training procedure likely to induce bias against users with certain traits?
  - Can users observe how they or other have been profiled?

Open Problems [Mishra et al., 2019b]

- **Ethical challenges** in user profiling:
  - Is profiling based on identity traits of users or on their online behaviour?
  - Is the training procedure likely to induce bias against users with certain traits?
  - Can users observe how they or other have been profiled?
- Systems embody the morals of their creators and annotators
  - Critical issue: automated systems can invalidate abusive experiences

Open Problems [Mishra et al., 2019b]

- **Ethical challenges** in user profiling:
  - Is profiling based on identity traits of users or on their online behaviour?
  - Is the training procedure likely to induce bias against users with certain traits?
  - Can users observe how they or other have been profiled?
- Systems embody the morals of their creators and annotators
  - Critical issue: automated systems can invalidate abusive experiences
- Continuous **evolution of the Internet jargon**:
  - Contextual features may become irrelevant over time

SAPIENZA NLP

Open Problems [Mishra et al., 2019b]

- **Ethical challenges** in user profiling:
  - Is profiling based on identity traits of users or on their online behaviour?
  - Is the training procedure likely to induce bias against users with certain traits?
  - Can users observe how they or other have been profiled?
- Systems embody the morals of their creators and annotators
  - Critical issue: automated systems can invalidate abusive experiences
- Continuous **evolution of the Internet jargon**:
  - Contextual features may become irrelevant over time
- Abuse is inherently **contextual**:
  - Sophisticated techniques are needed to capture the history of the conversation and the behavior of the users as it develops over time

Open Problems (2) [Mishra et al., 2019b]

- **Need for domain-specific learning**:
  - Performance of state of the art classifiers decreases substantially when tested on data drawn from domains different to those in the training set

Open Problems (2) [Mishra et al., 2019b]

- **Need for domain-specific learning**:
  - Performance of state of the art classifiers decreases substantially when tested on data drawn from domains different to those in the training set
- **Multimodality**:
  - Posts on social media often include data of multiple modalities

Open Problems (2) [Mishra et al., 2019b]

- **Need for domain-specific learning**:
  - Performance of state of the art classifiers decreases substantially when tested on data drawn from domains different to those in the training set
- **Multimodality**:
  - Posts on social media often include data of multiple modalities
- **Implicit abuse**:
  - Requires improvements in modeling of figurative language and sarcasm detection

Open Problems (2) [Mishra et al., 2019b]

- **Need for domain-specific learning**:
  - Performance of state of the art classifiers decreases substantially when tested on data drawn from domains different to those in the training set
- **Multimodality**:
  - Posts on social media often include data of multiple modalities
- **Implicit abuse**:
  - Requires improvements in modeling of figurative language and sarcasm detection
- **Explainability**:
  - Establish intent of abuse or lack of it
  - Highlight instances of abuse if present, be they explicit or implicit
  - Identify the target(s) of abuse

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Our Project

- Goal:
  - Develop the first model for **cross-domain explainable** abuse detection..

Our Project

- Goal:
  - Develop the first model for **cross-domain explainable** abuse detection..
  - ..(hopefully) without modeling the dataset's bias..

SAPIENZA NLP

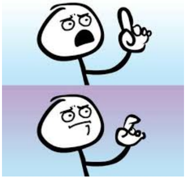
Our Project

- Goal:
  - Develop the first model for **cross-domain explainable** abuse detection..
  - ..(hopefully) without modeling the dataset's bias..
  - ..and with no annotated explanations available

SAPIENZA NLP

Our Project

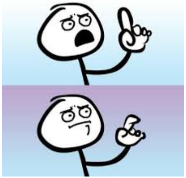
- Goal:
  - Develop the first model for **cross-domain explainable** abuse detection..
  - ..(hopefully) without modeling the dataset's bias..
  - ..and with no annotated explanations available



SAPIENZA NLP

Our Project

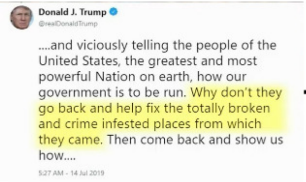
- Goal:
  - Develop the first model for **cross-domain explainable** abuse detection..
  - ..(hopefully) without modeling the dataset's bias..
  - ..and with no annotated explanations available



- Proposed solution:
  - Introduce **evidence-based** abuse detection
  - Evidences: collection of documents (from, e.g., social media guidelines, blog posts about social issues, etc.)

SAPIENZA NLP

Our Project (2)



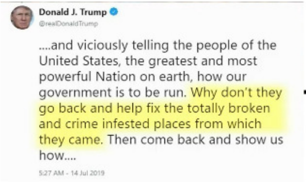
Donald J. Trump
@realDonaldTrump

....and viciously telling the people of the United States, the greatest and most powerful Nation on earth, how our government is to be run. Why don't they go back and help fix the totally broken and crime infested places from which they came. Then come back and show us how....

5:27 AM - 14 Jul 2019

SAPIENZA NLP

Our Project (2)



Donald J. Trump
@realDonaldTrump

....and viciously telling the people of the United States, the greatest and most powerful Nation on earth, how our government is to be run. Why don't they go back and help fix the totally broken and crime infested places from which they came. Then come back and show us how....

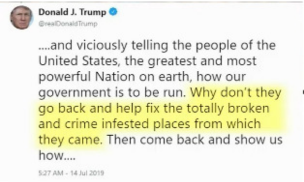
5:27 AM - 14 Jul 2019

Inciting fear about a protected category

We prohibit targeting individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., "all [religious group] are terrorists".

SAPIENZA NLP

Our Project (2)

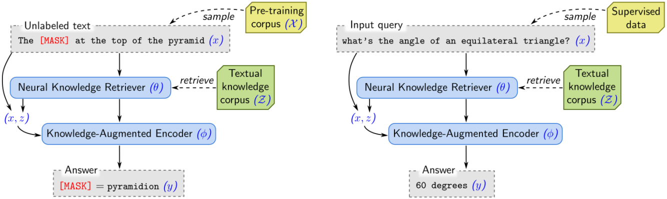

Donald J. Trump
@realDonaldTrump

....and viciously telling the people of the United States, the greatest and most powerful Nation on earth, how our government is to be run. Why don't they go back and help fix the totally broken and crime infested places from which they came. Then come back and show us how....

5:27 AM - 14 Jul 2019

Inciting fear about a protected category
We prohibit targeting individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., "all [religious group] are terrorists".

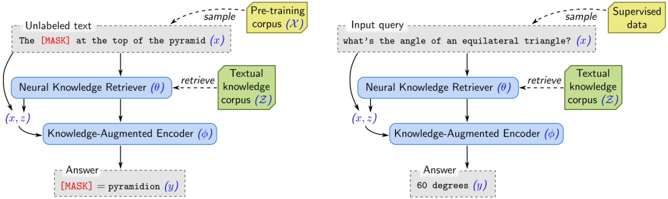
Our Project (3)



- How? Adopt and adapt REALM [Guu et al., 2020]:

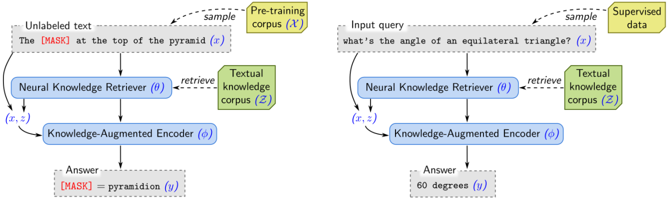
SAPIENZA NLP

Our Project (3)



- How? Adopt and adapt REALM [Guu et al., 2020]:
  - Need to define suitable pre-training tasks

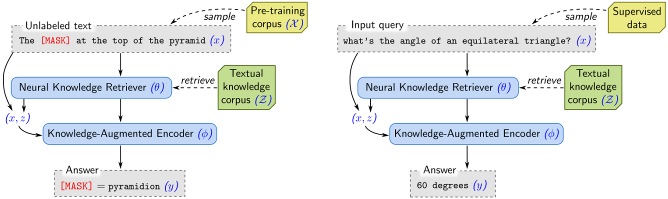
Our Project (3)



- How? Adopt and adapt REALM [Guu et al., 2020]:
  - Need to define suitable pre-training tasks
  - Ideas:
    - Given a tweet, find tweet with same hashtag
    - Given a tweet, retrieve the linked document
    - Evidence-based fact checking
    - Detect *possibly sensitive* tweets
    - Stance detection on social issues

SAPIENZA NLP

Our Project (3)



- How? Adopt and adapt REALM [Guu et al., 2020]:
  - Need to define suitable pre-training tasks
  - Ideas:
    - Given a tweet, find tweet with same hashtag
    - Given a tweet, retrieve the linked document
    - Evidence-based fact checking
    - Detect *possibly sensitive* tweets
    - Stance detection on social issues
- Evaluation through adversarial attacks, as in [Schiller et al., 2020]

SAPIENZA NLP

Thank you! Questions?

References I

Jill Burstein, Christy Doran, and Thamar Solorio, editors. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019. Association for Computational Linguistics. ISBN 978-1-950737-13-0. URL `https://www.aclweb.org/anthology/volumes/N19-1/`.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017. ISBN 978-1-57735-788-9. URL `https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665`.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909, 2020. URL `https://arxiv.org/abs/2002.08909`.

References II

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. Abusive language detection with graph convolutional networks. In Burstein et al. [2019], pages 2145–2150. ISBN 978-1-950737-13-0. doi: 10.18653/v1/n19-1221. URL https://doi.org/10.18653/v1/n19-1221.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024, 2019b. URL http://arxiv.org/abs/1908.06024.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer, 2019. ISBN 978-3-030-36686-5. doi: 10.1007/978-3-030-36687-2_77. URL https://doi.org/10.1007/978-3-030-36687-2_77.

SAPIENZA NLP

References III

Amir Hossein Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive language detection using multi-level classification. In Atefeh Farzindar and Vlado Keselj, editors, *Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence, Canadian, AI 2010, Ottawa, Canada, May 31 - June 2, 2010. Proceedings*, volume 6085 of *Lecture Notes in Computer Science*, pages 16–27. Springer, 2010. ISBN 978-3-642-13058-8. doi: 10.1007/978-3-642-13059-5_5. URL https://doi.org/10.1007/978-3-642-13059-5_5.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118, 2017. URL http://arxiv.org/abs/1701.08118.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Stance detection benchmark: How robust is your stance detection? *CoRR*, abs/2001.01565, 2020. URL http://arxiv.org/abs/2001.01565.

SAPIENZA NLP

References IV

Ellen Spertus. Smokey: Automatic recognition of hostile messages. In Benjamin Kuipers and Bonnie L. Webber, editors, *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97, IAAI 97, July 27-31, 1997, Providence, Rhode Island, USA*, pages 1058–1065. AAAI Press / The MIT Press, 1997. ISBN 0-262-51095-2. URL http://www.aaai.org/Library/IAAI/1997/iaai97-209.php.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics, 2016. ISBN 978-1-941643-81-5. doi: 10.18653/v1/n16-2013. URL https://doi.org/10.18653/v1/n16-2013.

SAPIENZA NLP

References V

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words - a feature-based approach. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1046–1056. Association for Computational Linguistics, 2018. ISBN 978-1-948087-27-8. doi: 10.18653/v1/n18-1095. URL `https://doi.org/10.18653/v1/n18-1095`.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In Burstein et al. [2019], pages 602–608. ISBN 978-1-950737-13-0. doi: 10.18653/v1/n19-1060. URL `https://doi.org/10.18653/v1/n19-1060`.

SAPIENZA NLP