An Approach to Detect Cyberbullying on Social Media

Fatemeh SAJADI ANSARI, Mahmoud BARHAMGI,
Aymen KHELIFI, and Djamal BENSLIMANE

CyberBullying definition

- *“Willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices. “*
- Cyberbullying can be characterized as
  - A deliberate act, carried out by the perpetrator in a repeated fashion through the use of digital means with the objective to inflict harm to the victim.

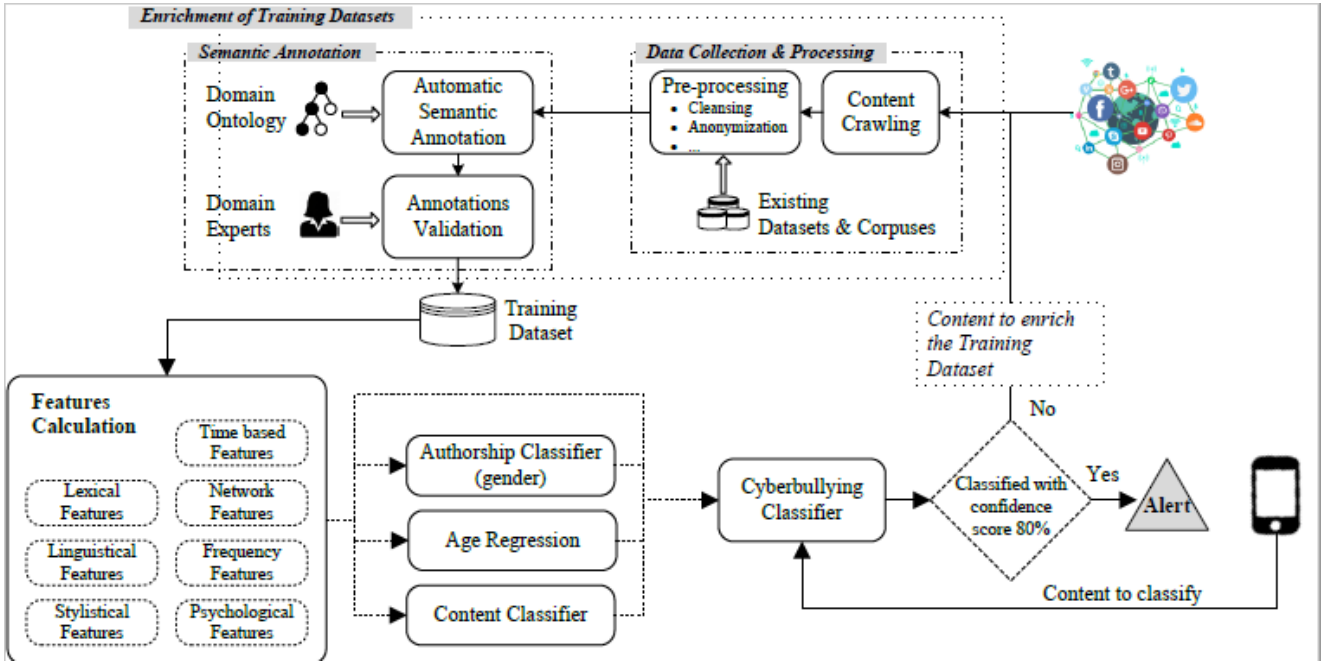
Existing approaches

- Machine Learning Based Solutions
- Lexicon-Based Solutions
- Rules-Based Solutions
- Hybrid Solutions

Approach overview

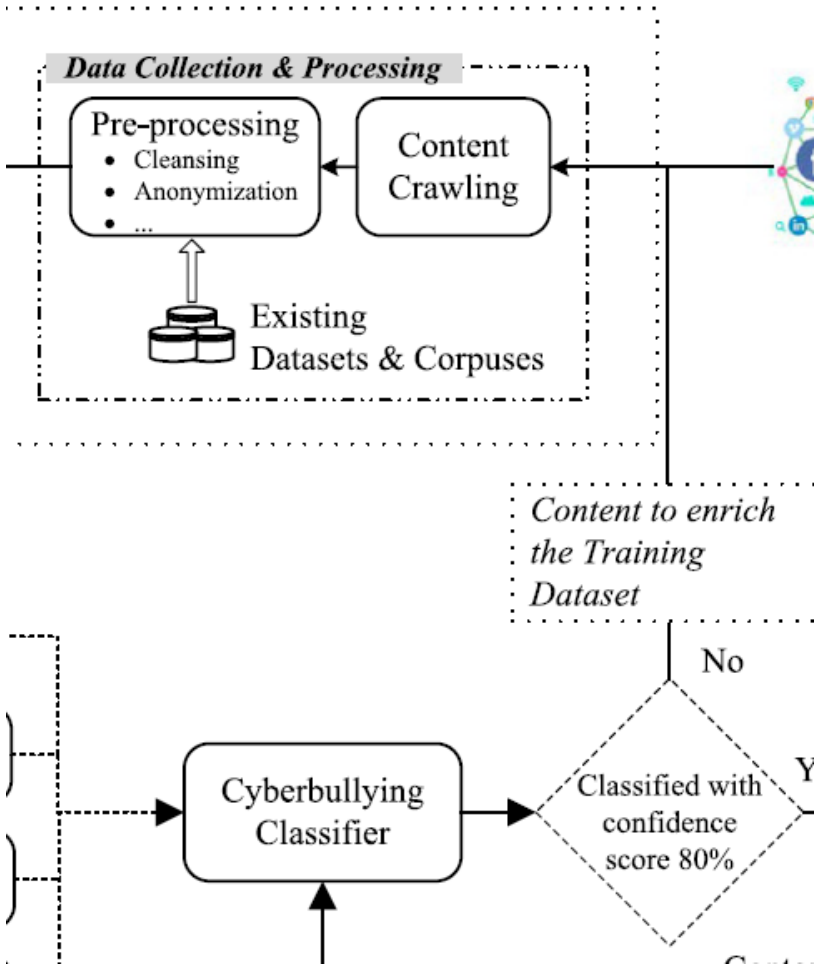
The approach in based on three inter-related groups od processes

- Construction and enrichment of training datasets
- Feature calculation and content classification
- Cyberbullying detection

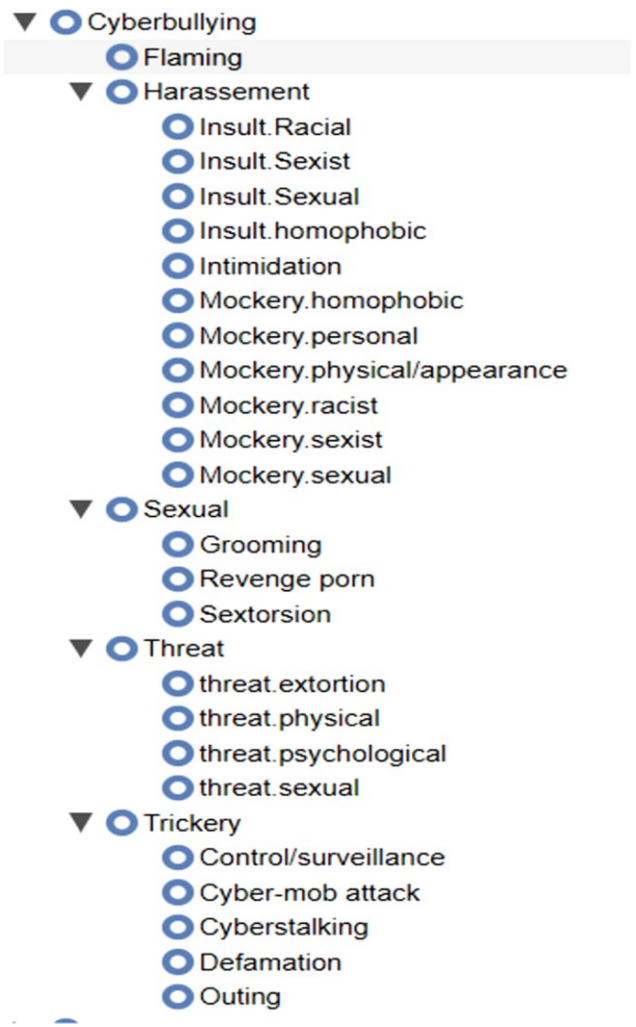


Training datasets construction & Enrichment

- Existing datasets and corpuses that are used by the scientific community
- Real data collected from social medias including Twitter and Facebook
  - Cleaned & anonymized
  - Annotated by semi automatically manner

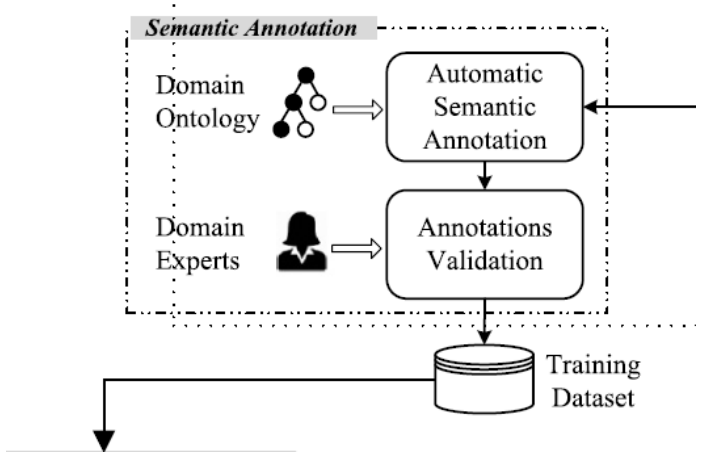


Cyberbullying Ontology



Semantic Annotation

1. Apply a set of syntactic and linguistic rules to detect messages with toxic content.
2. These messages are marked with the corresponding category.
3. The semantic annotations are validated manually by domain experts
4. New terms that could appear in a toxic message are automatically added to the representative terms of appropriate category.



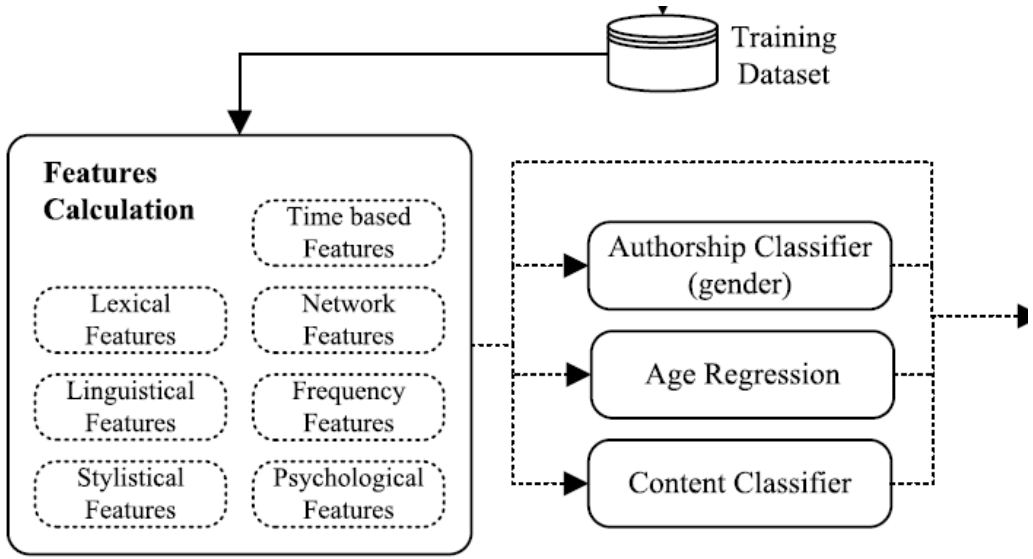
Rule-1	
Preconditions:	Occurrences of: [*imperative/indicative verb with negative meaning*], [second person], [racial offense][proper noun] {0,1}
Annotations:	Insult. Racial
Examples:	*"Niggers and their liberal friends steal everything not tied down, just like the presidency here with acorn with its liberal defenders, FUCK YOU NIGGER OBOAMA!"*

Rule-2	
Preconditions:	Occurrences of:[second person/third person pronoun] [*state verb*], [body organ]{0,1}, [derogatory content]
Annotations :	mockery/appearance
Examples:	"Wikipedia is not the proper place for you to abuse your powers just because you're unsatisfied in life. It ain't my fault you're ugly, sista"

Feature calculation

The approach involves various categories of indicators :

- Content related indicators
  - linguistical indicators
  - lexical indicators
  - stylistically indicators
- Time-based indicators
- Network related indicators
- Psychological indicators



Prototype, Experiments and Results

Datasets Descriptions

Corpus / Labels	T_{t1}	T_{t2}	F_{t3}	*All*
insult	2252	6041	84	8377
mockery	429	1380	28	1837
threat	215	552	8	775
sexual content	68	50	5	123
defamation	113	306	7	426
advertising	479	54	19	552
No Bullying	1521	1010	464	2995
nan (other languages)	81	40	266	387

- Total, 15472 messages were extracted
  - 5158 extracted messages from Twitter at time *t*1
  - Then 9433 messages extracted from Twitter at time *t*2
  - Finally, 881 messages from Facebook at time *t*3

Textual and extra-textual indicators

- Textual and extra-textual indicators operated to classify an exchange Ex(i) between two or more individuals
  - Textual Indicators :
    - Textual surface indicators
    - Conversational indicators
    - Lexical indicators
    - LIWC indicators
  - extra-textual indicators
    - Frequency indicators
    - Temporary meta data
    - Profile Indicators

Classifiers and Conducted Experiments

- Six classifiers were implemented
  - Multiclasses :
    - Toxic comments classifier based on ontology classes (CamemBERT)
    - Age detection
  - Binary :
    - Toxic, non-toxic CamemBERT tweets classification
    - Gender classification
- Personality analysis which is a binary classification of five personality traits

Natural language	Model objective	Class labels	ML models
English	Toxic comments detection	Toxic, severe toxic, insult, obscene, identity hate	Bert
English	Gender prediction from text	Male, female	SVM, Bert
English	Gender prediction from name	Male, female	LSTM, CNN
English	Age prediction	Adolescent, young adult, adult	SVM, Bert
English	Personality analysis	Big 5 labels	SVM, Bert
French	Toxic tweets classification	Toxic, non toxic	CamemBERT
French	Toxic tweets classification	Cyberbullying ontology categories	CamemBERT

Overview of classifiers performance

Big 5 / *Value*	Op	Co	Ag	Ex	Ne
boolean value	true	true	true	false	false
probability	0.65	0.56	0.66	0.56	0.29

(a) Personality analysis from text

	precision	recall	F score
Female	0.89	89	0.89
Male	0.81	0.8	0.81

(d) Gender classifier

precision	recall	F score
88.1%	82.4%	85.1%

(b) Toxicity detection

Conclusion

- Our approach combines several data mining methods
- It models cyber harassment on the time axis under its different dimensions such as lexical, linguistical, and psychological
- It relies on a detailed analysis of the different categories of cyber harassment in order to attribute an appropriate level of severity to each detected risk situation

Thank You for your attention