RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

modyco
Modèles, Dynamiques, Corpus
UMR 7114
Sciences du langage.

Exploring the Emotional Dimension of French Online Toxic Content

LREC COLLING 2024
May, 2024

Valentina Dragos
Delphine Battistelli
Fatou Sow
Aline Etienne

Summary

- Goals and motivation
  - Understanding the nature of French toxic content
- Description of corpora
  - Right wing extremism, Sexism; Online Hate
- Data processing and emotion annotation
  - Overview of the general architecture
  - Annotation schema and process for emotions
- Experiments
  - Comparative analysis of emotion distribution
- Conclusion and future work

Toxic content on French social media

- Social platforms and new forms of communication and interaction
  - Numerous and various users
  - Variety of topics
  - Content release and spread
- Toxic content
  - Extremism, Online Hate, Sexism, etc.
  - Danger to individuals and society as a whole
- What is the nature of toxic content?
  - Focus on emotional dimension
  - Previous studies showing that emotions trigger information spread and propagation

Corpora of toxic content in French (1/2)

- Corpus of right wing extremisms
  - Source: created by a previous national project investigating automatic methods to detect right wing extremism on social platforms
  - Content: tweets and messages collected on discussion forums
  - Size: 1576 textual paragraphs (several sentences)
  - Classes: extremist vs. non extremist, slightly unbalanced

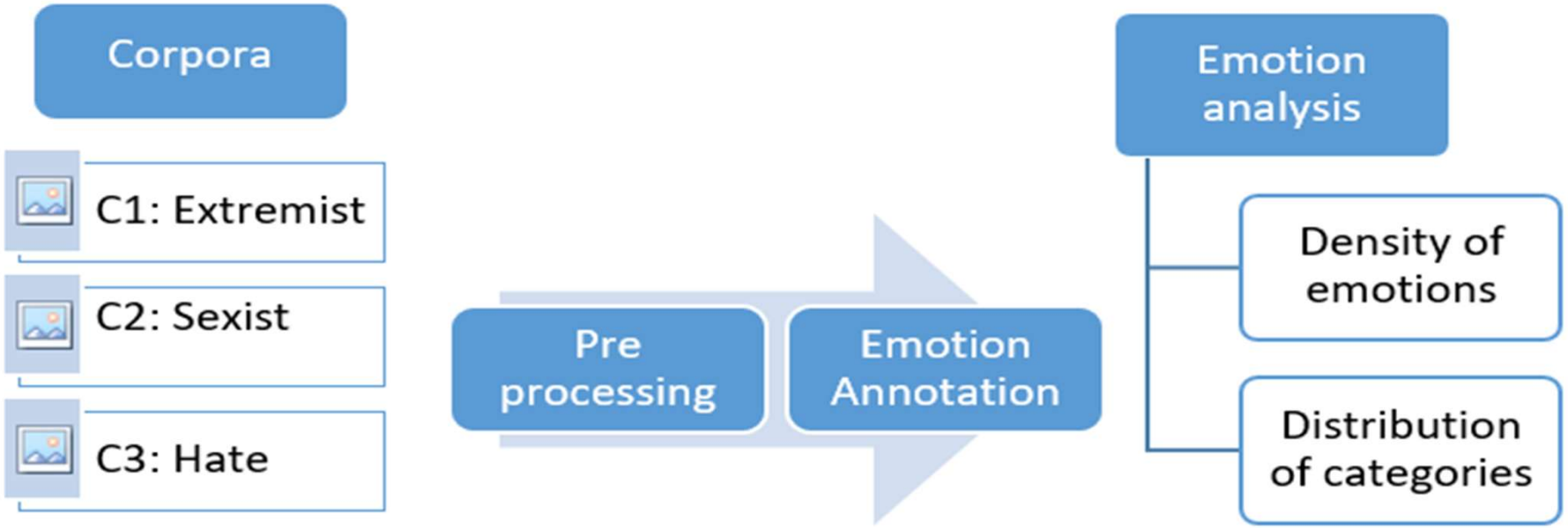
- Corpus of sexism
  - Source: created by a previous national project investigating automatic methods to detect sexism on social platforms
  - Content: tweets
  - Size: 12 000 tweets
  - Classes: sexist vs. non sexist, slightly unbalanced

Corpora of toxic content in French (2/2)

- Corpus of online hate
  - Source: created by the previous national project investigating automatic methods to detect extremism on social platforms
  - Content: tweets
  - Size: 600 tweets
  - Classes: hateful vs. non hateful, balanced

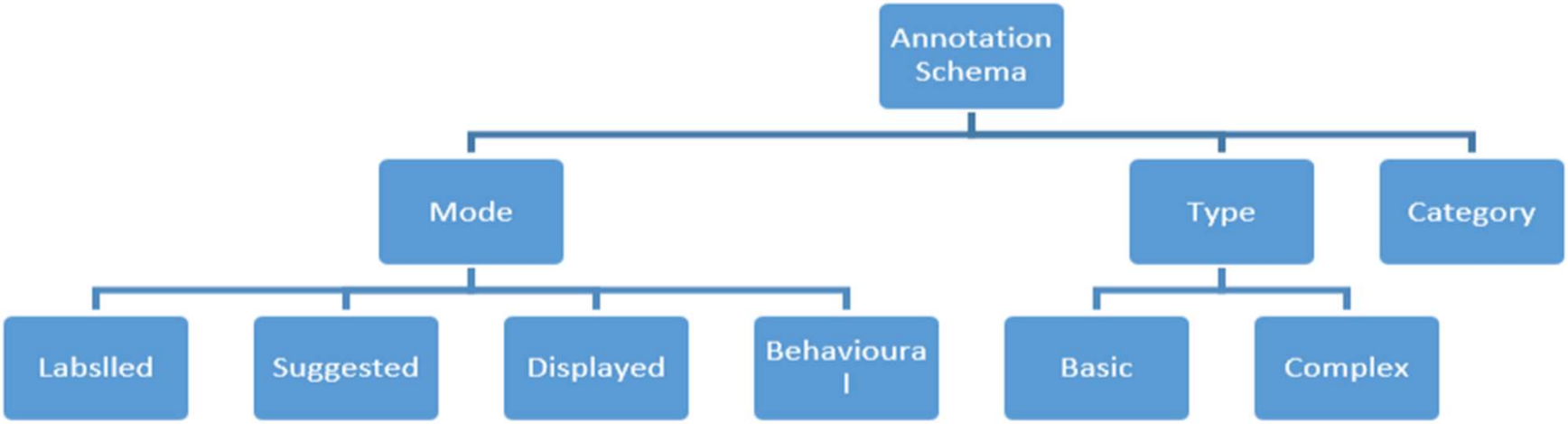
- Remarks
  - Corpora of various types of toxicity
  - Consist of user-generated content collected on social platforms
  - Manual annotation of classes
  - Comparison possible at corpora level (at class level) but also inter-corpora

Data processing and emotion annotation (1/4)



- Steps
  - Pre processing : corpora cleaning, lemmatization, tokenization
  - Emotion annotation
  - Analysis of emotions (density of emotions and distributions of emotional categories)

Data processing and emotion annotation (2/4)



- Emotion annotation
  - Annotation schema for situated emotions
  - Detection of emotional units (SitEmo)
- Annotation values:
  - EmotionMode (Labelled, Displayed, Suggested, Behavioral)
  - EmotionType (Basic, Complex), Category (Joy, Anger, Fear, etc.)
  - EmotionTrigger (Seed): link to textual data

Data processing and emotion annotation (3/4)

Fiers *de notre héritage* et confiants dans notre destin.

(Proud of our heritage and confident in our destiny.)

SitEmo <Pride> {Mode: Labelled Type: Complex,

Category: Pride, Trigger: fiers}

Labelled annotation of type *Pride* *Seed: Fiers (Proud)*

Notre pays est dans une *situation* ***catastrophique.***

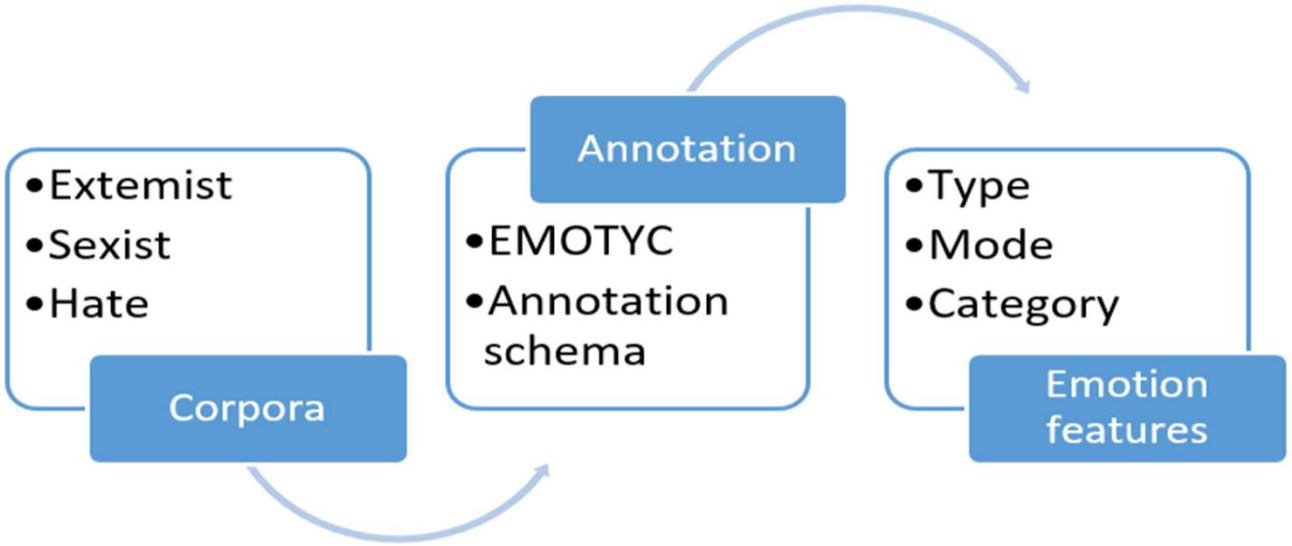
(Our country faces a catastrophic situation)

SitEmo <Sadness> {Mode: Suggested Type: Basic,

Category: Sadness, Trigger: catastrophique}

Suggested annotation of Type *Sadness* *Seed: catastrophique (catastrophic)*

Data processing and emotion annotation (4/4)



- Annotation procedure
  - Emotyc : an automatic parser of emotions in text
  - Tasks of Emotyc:
    - Detection of emotions
    - Detection of type, mode and category
    - No indication of the seed

Experiments and case studies (1/2)

- Question1 : do we have different emotion density in corpora ?
  Question 2: do we have different emotion distribution within various corpora and their classes?
- Manual analysis of SitEmo distribution after Emotyc annotation
- Case study 1: density of emotions

Corpus	Annotations	Density of emotions
C1	972	56.25
C2	6816	56.80
C3	534	89.00

- Extremist and sexist corpora have similar density of emotions
- Hate corpus has a higher density of emotions

Experiments and case studies (2/2)

- Case study 2: distribution of emotions

Anger	Admiration	Sadness	Surprise	Fear
213	45	44	14	13
61.5	13	12.71	4.04	3.75

Distribution of emotion Categories for Hateful class

Admiration	Surprise	Anger	Joy	Sadness
73	36	32	21	16
35.78	17.64	15.68	10.29	7.84

Distribution of emotion Categories for Non Hateful class

- Remarks:
  - Hateful and non hateful classes are the most contrasted in terms of emotion categories
  - Extremis and non extremist classes have the most similar distributions of emotions
  - Anger is the dominant emotion in all corpora, with just one exception; the non hateful class

Conclusion and future work

- Exploration of toxic content in French online data
  - Three types of toxic content: extremist, sexist and hateful
  - Investigation of emotional dimension
  - Vector for content propagation
- Emotion detection and annotation
  - Annotation schema for situated emotions : Type, Mode and Category
  - Emotyc: an automatic parser for emotion detection
- Experiments
  - Classes of the hate corpus are contrasted
  - Anger is dominant emotion sf toxic content
- Perspectives:
  - Development of learnings models integrating the emotions
  - Fine tuning of Emotyc to improve the accuracy of automatic annotation

Q&A

Thank you for your attention!

valentina.dragos@onera.fr