Angry or Sad ? Emotion Annotation for Extremist Content Characterization

V. Dragos, Delphine Battistelli, A Etienne, Y. Constable

Angry or Sad ? Emotion Annotation for Extremist Content Characterization

V. Dragos, D. Battistelli, A. Etienne, Y. Constable
ONERA - The French Aerospace Lab, France
University of Paris Nanterre, France
valentina.dragos@onera.fr, delphine.battistelli@parisnanterre.fr,
aline.etienne@parisnanterre.fr, yolene.constable@onera.fr

Abstract

This paper examines the role of emotion annotations to characterize extremist content released on social platforms. The analysis of extremist content is important to identify user emotions towards some extremist ideas and to highlight the root cause of where emotions and extremist attitudes merge together. To address these issues our methodology combines knowledge from sociological and linguistic annotations to explore French extremist content collected online. For emotion linguistic analysis, the solution presented in this paper relies on a complex linguistic annotation scheme. The scheme was used to annotate extremist text corpora in French. Data sets were collected online by following semi-automatic procedures for content selection and validation. The paper describes the integrated annotation scheme, the annotation protocol that was set-up for French corpora annotation and the results, e.g. agreement measures and remarks on annotation disagreements. The aim of this work is twofold: first, to provide a characterization of extremist contents; second, to validate the annotation scheme and to test its capacity to capture and describe various aspects of emotions.

Keywords: Emotion annotation, social data analysis, extremist content

1. Introduction

Social platforms play an increasingly important role in the propagation of extremist ideas. As the growth of extremism online continues, the ability to detect this harmful content is paramount to restrain the spread of messages. Emotions associated to this type of content reflect the affiliations and aptitudes of users towards entities, events but also ideas and give clues about their activities online.

A major barrier to the development of accurate models for extremist content detection is the limited number or relevant corpora and the absence of a reliable protocol for data annotation. More specifically, the need for corpora annotated by combining both sociological knowledge as to why the content is extremist has also emerged.

The paper focuses on extremism analysis in French textual contents collected on social platforms and adopts an emotion annotation scheme to investigate emotions expressed in extremist and non-extremist bodies of texts. In this work, we show that extremist contents capture valuable emotional linguistic signals revealing, for instance, when extremist attitudes are associated with fear or sadness.

Building a labeled corpus for extremism detection is not trivial, since there is no consensus among researchers in the field of sociology on the definition of extremist content or the identification of its main characteristics (Alava et al., 2020). In addition, detecting emotion in text is a difficult task even for humans (Öhman, 2020). In order to cope with those difficulties, we developed an approach which consists of two main phases: (1) the development of a semi-automatic procedure guided by knowledge from sociology was implemented to create two distinct data sets containing extremist and non-extremist contents, respectively; and (2) the use of an annotation scheme designed to manually annotate emotions, where annotators identify linguistic markers that convey an emotion among a pre-defined set of emotions.

Based on this combination of sociological knowledge and emotion analysis, we are able to build a new data set, which provides interpretable associations of emotions that highlight differences between extremist and non-extremist contents.

Overall, this paper makes three main contributions. First, we integrate both sociological knowledge and emotion annotations and show that emotions annotation correlate well with human judgments of extremist or non-extremist contents. Second, we provide fine-grained annotated data sets in french, which can be widely used for training data for machine learning approaches and can be considered as gold standards in text classification tasks. Finally, we carry out a manual validation of emotion annotation and a cross-analysis of annotation categories.

The remainder of the paper is organized as follows. Next section discusses several related approaches and section 3. gives a brief overview of the corpora used for this work and explains the procedures adopted to build it. Section 4. elaborates on the annotation scheme used for emotion analysis of texts and provides examples of annotations. Section 5. discusses annotation validation and remarks on agreement measures. Section 6. concludes and presents directions for future work.

2. Related work

The analysis of affective states, including sentiments, emotions or opinions, received attention from several research communities, including natural language processing, sociolinguistics and machine learning, to the extent that building labeled data sets for training raises questions related to the nature of emotions and the representation scheme suitable to describe them. The section discusses several data sets labeled with emotion types and information related to their extremist nature, respectively.

2.1. Emotion labeled data sets

A large number of research efforts developed data sets annotated according to different emotion schema. Those re-

sources can be classified into two main categories: data sets created to linguistically describe emotions (and sometimes also their causes) and data sets built to investigate how emotions correlate with other factors.

The first category includes EmoBank, a data set released by Buechel and Hahn (Buechel and Hahn, 2017). EmoBank consists of 10k sentences in English, manually annotated according to the valence-arousal-dominance model (Mauss and Robinson, 2009). This model describes emotions according to there dimensions: valence or polarity, arousal, a concept capturing the degree of calmness or excitement, and dominance, which is to say the perceived degree of control over a situation. EmoBank builds on multiple genres and domains and the annotation highlights both the emotion expressed by the writer, and the emotion perceived by the readers.

EmoInt is another labeled data set built by Mohammad and Bravo-Marquez (Mohammad and Bravo-Marquez, 2017) in order to associate paragraphs with various intensities of emotions. The collection focuses on social media and gathers 7,097 tweets altogether in English. The tweets were annotated via crowdsourcing with various intensities of four emotions: *anger, joy, sadness,* and *fear*, although most tweets are annotated with one emotion.

The Emotion-Stimulus dataset was published by Ghazi and colleagues (O'Reilly et al., 2016) in order to predict the causes of emotions in the text. The resource consists of 1,549 sentences in English annotated with emotions and 820 enriched sentences annotated with emotions and their causes. The set of labels includes the list of basic emotions (*anger, contempt, disgust, enjoyment, fear, sadness, surprise*) to which *shame* is added.

The second category includes research efforts intended to investigate the relations between emotions and other factors, such as context, news, events or actions. Hence, the Stance Sentiment Emotion Corpus, is a data set in English released by Schuff and colleagues (Schuff et al., 2017). The collection consists of 4,868 tweets annotated thank to a hybrid approach mixing linguistic expert annotations and crowd sourcing. Each tweet was labeled with multiple emotion labels following the Plutchik's wheel of emotions (Donaldson, 2017). The resource has several annotation layers, allowing the analysis of relationships among emotion types.

Following a similar research line, Grounded-Emotions is an emotion tagged data set in English published by Liu and Mihalcea (Liu et al., 2017) in order to correlate emotions with other factors including weather, news and social aspects. The collection was built on social media and consists of 2,557 single labeled instances published by 1,369 unique users. The set of labels includes only two emotions: *happy* and *sad*. The resource was used in experiments showing the role played by contextual factors in predicting emotions.

Taking a step forward, GoodNewsEveryone is a corpus built by Bostan and colleagues (Oberländer et al., 2020) in order to tackle emotions from a structured learning perspective. The collection consists of 5000 English news headlines, annotated via crowdsourcing by associating emotion types, semantic roles capturing emotion causes, experiencers and targets, as well as the reader's perception. By adding those annotation layers, different types of associations can be further inferred, such as correlations of emotions and their causes but also differences between the emotions as related by the author and perceived by the reader. Other resources develop even more complex structures to describe emotions in text, and for example AffectVec (Raji and De Melo, 2020) matches English words to numerical vectors, in which a given dimension quantifies the degree of association of that specific word with a specific emotion.

2.2. Data sets labeled for extremism detection

As shown in the previous section, annotation of emotions in texts has extensive literature, but there are not numerous studies considering emotions in contents dealing with extremist ideologies. More specifically, building labeled data sets for extremism detection received little attention.

De Gibert and colleagues developed a labeled data set for White Supremacy detection (de Gibert et al., 2018). The resource is in English and consists of data collected from the StormFront Website, a main portal of white supremacists. Although data was collected from an extremist source, the annotation focused mainly on hate detection, and several annotation rules were defined. For example, one rule labeled a text as *Hate* if an attack against a specific group of individuals was mentioned; another rule annotates as *Hate* any content explicitly supporting extremist ideas. The overall collection was annotated with a set of three labels: *Hate, NoHate,* and *Skip* (for undetermined contents).

Another extremist data set in English was built by scraping extremist content from Twitter feeds by using known white supremacist hashtags like *white-privilege*, and *its-ok-to-be-white* (Alatawi et al., 2021). Three different annotators annotated the collection with a set of four labels: *ExplicitWhiteSupremacist, ImplicitWhiteSupremacist, OtherHateSpeech*, and *Neutral*. Other contributions considered the binary *(Extremist/NonExtremist*) annotation of online profile, as it is the case in this study (Hartung et al., 2017) carried out by Hartnung and colleagues to detect right-wing extremism in German Twitter profiles.

Regarding the analysis of French contents, a first data set for sexism detection in French is presented in (Chiril et al., 2020). The collection gathers about 115,000 tweets among which 12,274 are manually annotated with two main categories *Sexist* and *NonSexist*. The annotation procedure also provides a deeper characterization of sexist contents, by making the distinction between tweets that are directly addressed to a target (a woman or women in general) and tweets that report or denounce sexism experienced by women. A review of methods and annotated corpora for online extremist detection is presented in (Gaikwad et al., 2021).

With this paper, we provide a data set in French annotated with both linguistic and sociological expert knowledge. Data was collected from Twitter and the resource provides information on right-wing extremist categories and on emotion types. Thus it allows us to explore the relationships between those two distinct annotation layers.

3. Data collection and characterization

Collecting data sets for the specific purposes of usage-driven approaches relies on one major hypothesis, namely that the content of corpus is representative of the phenomenon under scrutiny. The representativeness of corpus raises methodological and practical questions, and for this work, concepts and dynamics from sociology guided the construction of data sets. Extremist content can be gleaned after explicitly defining the notion of *extremist* or by selecting several sources generally accepted as being extremist.

Defining extremism in general as a process or concept, suffers from a lack of consensus within the sociology field itself (Thorburn et al., 2018). If we consider the specific case of right-wing extremism in Europe, the features of this ideology change considerably from one country or region to another. Several authors pointed in to the fluid nature of extremism and its multifaceted nature including a verity of discourses from revisionists, racists, skinheads and extremist hooligans, nationalists or identity-based groups, paramilitary, xenophobic or anti-migrant groups (Araque and Iglesias, 2021). In order to avoid the difficulty of defining the concept of *extremism* we restricted the application area of this study to the analysis of extremism in French blogosphere and we also focused on information sources. The data collection step consists at the first run on selecting several sources considered as extremist and collecting streams of data released by those sources. A source can be a user, a hashtag, a keyword or a combination of those three elements.

Data was collected from December 2019 to December 2020 by an interdisciplinary team of researchers in sociology, linguistics and computer science. The sets of data have been scraped from various online platforms using the MediaCentric tool [1]. The corpus was built iteratively in an effort to select only texts which were clearly connected with a certain ideology. The data set includes data from Twitter, forums, and online sources.

Streams collected from different sources were merged together. Merging those contents, although provided by distinct sources, allows us to build a homogeneous corpus according to the principle of *homophily* (Oussalah et al., 2018) stating that users within a certain group setting have a tendency to develop a similar use of language when developing the ideas of the group. In spite of using keywords, the subject matter and the homogeneity of the corpora were difficult to establish after the collection step, because online users user can write about any subject.

After collecting rows of online date, several validation and characterization steps were performed on the entire collection to ensure that data sets are relevant and to detect finer categories of extremist contents. Those steps are described in fig. 1.

First, the initial collection gleaned on social platforms was manually explored by two experts in sociology (one is a senior researcher in education sciences and the other one is a post doc in sociology with a background in social communication) in order to discard non-relevant contents. The initial corpus and was further roughly divided into two sub-

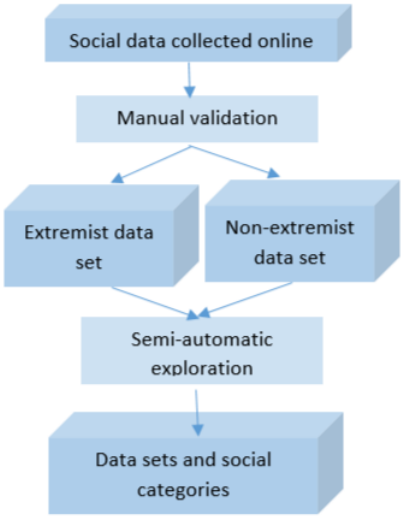

Figure 1: Data sets validation and exploration

sets, having respectively extremist and non-extremist contents as shown in fig. 2.

Extremist	Non extremist
Textual unit 1 : La France doit rester la France, notre patrie sacrée. (France must remain France, our sacred homeland.) **Textual unit 2 :** Etes-vous prêts à prendre le contrôle du pays ? (Are you ready to take control over the country?) **Textual unit 3 :** Parce que le temps est *à* l'urgence nous avons fait des choix... (Because the time is up, we have made choices ...) **Textual unit 4 :** Nul ne peut ignorer les dangers qui menacent l'Europe. (No one can ignore the dangers threatening Europe) **Textual unit 5 :** Nous nous inscrivons dans un combat radical et sans concession. (We are part of a radical and uncompromising fight.) **Textual unit 6 :** Ils militent pour un journalisme factuel et traquent les informations erronées. They campaign for fact-based journalism and hunt down misinformation. **Textual unit 7 :** Nous devons nous prendre en charge par des actes de résistance. We shall take care of ourselves through acts of resistance	**Textual unit 1 :** Ils veulent que vous restiez pauvres. (They want you to stay poor.) **Textual unit 2 :** Les femmes ne sont pas des victimes. (Women are not victims.) **Textual unit 3 :** Si notre pays est aujourd'hui malade du mondialisme, le nationalisme sera son remède. (If our country is sick with globalism today, nationalism will be its cure) **Textual unit 4 :** La nature a été violemment arraisonné par la technique et le développement économique (Nature has been violently boarded by technology and economic development.) **Textual unit 5 :** Le monde moderne fabrique un modèle dominant d'antihéros. (The modern world creates a dominant model of antiheroes.) **Textual unit 6 :** Défendre sa liberté d'expression ne revient aucunement à défendre ses actes. Defending your freedom of expression is not the same as defending your actions. **Textual unit 7 :** Les écrans, désormais omniprésents, sont des outils de destruction massive. (Screens, now ubiquitous, are tools of mass destruction.)

Figure 2: Extremist and non-extremist textual units

Then, extremist and non-extremist data sets were further semi-automatically explored in the light of sociological knowledge in order to derive specific extremist categories from empirical data.

[1] https://www.chapsvision.fr/data

Using statistical tools for textual analysis (IRAMUTEQ [2]) and tagging collected from 56 information sources we identified categories of texts based on internal similarity (lexical similarity estimated by the tool) and external distances (semantic distances estimated by the tool).
The identification of social categories was carried out by the same team of two researchers in sociology who have a sufficient understanding of online platforms. They were provided with a set of symbols, hashtags and notions associated to extremist ideologies. This input was used as a baseline to decide of which content is considered extremist and which is not. Researchers carefully studied both the information comprised in texts and about the texts (namely the source or keywords used to collect the tweet) to infer the main category associated to the texts. They also identified textual units carrying relevant extremist or non-extremist contents. The length of textual units ranges from one to eight sentences (clearly, one sentence only is often not enough to identify extremist data).
Keywords and topics that seemed to characterize the content were highlighted and discussed. When textual units corresponded to several categories, each category would be indicated accordingly. When this was not the case, new categories would be proposed inductively. By following this approach, four right-wing extremist categories were generated, see tab. 1.

Category	Frequent key-words
Fundamentalists	Faith, Identify, Culture, Origin
Defenders	Victim, Danger, Threat, Manipulation
Nostalgics	Value, Nation, New order
Fighters	Fight, Conflict, Planning, Action

Table 1: Categories of right-wing extremism

At the est of this process, each textual unit was labeled either as extremist (1), non-extremist (0) or unknown (x) by each person involved. For the final collection, only the extremist and non-extremist textual units were selected.
The total number of textual units in the final data set is 1728, out of which 1129 were labeled extremist and the remaining 599 were labeled non extremist. For each textual unit we stored the text together with the right-wing extremist categories associated to it and the binary extremist/ non extremist distinction.

4. Emotion annotation

The annotation scheme used for this work was developed independently of the task of extremism detection (Étienne and Battistelli, 2021). This paragraph introduces the elements of the scheme and the annotation procedure.

4.1. Description of annotation scheme

The annotation scheme uses *SitEmo* units to describe *emotional situations*, which is to say emotions expressed by linguistic markers and thus spatially and temporally anchored. *SitEmo* units are characterized by using 4 elements shown in tab. 2:

[2]http://www.iramuteq.org/

Expression Mode	Labelled, Displayed, Suggested, Behavioral
Type of emotion	Basic, Complex
Category of emotions	Anger, Joy, Fear, Pride. etc.
Emotion trigger	Core of emotional expression

Table 2: Elements of *SitEmo* units

Emotion Mode indicates how the emotion is linguistically expressed in the textual segment. The scheme considers that emotions are expressed in four main ways: through an emotional label, displayed by the linguistic characteristics of an utterance, suggested by a situation or inferred from behavior, thus values of *Mode* fit into four sub-types[3]: *Labelled, Displayed, Suggested* and *Behavioral*.
Labelled emotions are explicitly mentioned through an emotional label, such as *happy, anger, etc.*. This expression mode corresponds to the use of emotional lexicon's words (Creissen and Blanc, 2017; Micheli, 2014).
Displayed emotions are shown directly through the characteristics of statements, which occurs when the enunciator experiences an emotion at the time of the utterance. The statement then includes linguistic markers, which show that the speaker felt an emotion and on which the reader relies to infer the emotional state of the enunciator. According to (Micheli, 2014), *Displayed* emotions can be shown by syntactic structures, lexical marks (interjections, judgments etc.) or typographical signs (exclamation points, etc).
Suggested emotions are inferred by the reader from the description of socio-cultural conventions associated to feelings. The emotion is therefore neither directly presented, nor translated by a behavior, nor visible in the structure of the statement. The reader will analyze the overall picture, and this analysis serves as a support to build the emotion: the emotion felt by the character (or Narrator/Writer) is inferred by the reader based on the situation described. (Creissen and Blanc, 2017; Micheli, 2014)
Behavioral emotions are expressed by using the description of behaviors or physical manifestations of the character who feels the emotion. Affective states are often associated to a variety of behavioral and emotional reactions, but sometimes only behaviors are reported, and the nuances of emotions will then be inferred from their descriptions (Creissen and Blanc, 2017).
The *Type* of emotion (basic or complex) and the emotional category (fear, joy, etc.) also characterize *SitEmo* units. More specifically, the annotation scheme introduces 11 categories of emotions. The list of emotion categories comprises the 6 basic emotions introduced by (Ekman, 1992) and 4 complex emotions taken from (Blanc and Quenette, 2017) and (Davidson, 2006). We added a fifth complex category, *Admiration*, to better balance basic and complex emotions. Values of *Type* and *Category* elements are corre-

[3]The scheme was developped in French. We translate the sub-types as follows: *Désigné* is *Labelled*, *Montré* is *Displayed*, *Suggéré* is *Suggested* and *Comportemental* is *Behavioral*.

lated and each of the 11 emotional categories is previously associated to the *basic* or *complex* types, as shown in table 3.

Basic emotions	Complex emotions
Anger	Admiration
Disgust	Guilt
Joy, Fear	Embarrassment
Surprise	Pride
Sadness	Jealousy

Table 3: Associations of emotional types and categories

Those categories are defined in order to regroup several more specific emotions. For example, the category *Anger* can be used to annotate textual segments expressing not only *Anger* but also *annoyance, fury, indignation* and *disapproval*.
The set of emotional categories implemented in the scheme is not exhaustive. In order to keep the annotations in line with the content under analysis, the unit *Other* can be used to annotate segments expressing an emotional category not foreseen by the annotation scheme.
The last element of *SitEmo* units is the *Trigger*, understood as the core of the annotated emotional expression. The *trigger* of emotions is the shortest segment (term or group of terms) which, within the annotated segment, constitutes the most salient emotional element, the one that focuses the emotional meaning of the textual segment and thus motivates the annotation. In some practical situations, several triggers can be identified for the same *SitEmo* unit.

4.2. Annotation procedure

The annotation scheme was designed to be implemented within the Glozz annotation platform (Widlöcher and Mathet, 2012). Glozz was selected because it can be customized by plugging specific annotation schemes and allows linguistic annotations at different levels (words, grammatical categories, sentences, groups of sentences) . The annotation starts with distinguishing a seed from a whole sentence in terms of emotion, see fig. 3.

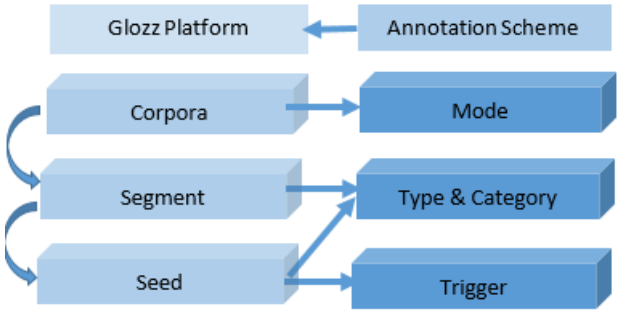

Figure 3: Annotation process

The basic annotation segment can be a word or a row of several words in the sentence, including the seed, see fig. 4. Every single seed expresses a single emotion and triggers the creation of a SitEmo unit, although a sentence can bear

Parce que le *temps est à* ***l'urgence*** nous avons fait des choix…
(Because the time is up, we have made choices ...)
SitEmo <Fear> {Mode: Suggested Type: Basic, Category: Fear, Trigger : urgence}
Les nations s'engouffrent toutes dans ce *grand tourbillon* ***destructeur***.
Nations are all rushing into this great vortex of destruction.
SitEmo <Sadness> {Mode: Suggested Type: Basic, Category: Sadness, Trigger: destructeur}

Figure 4: Examples of ***seeds*** and *annotation units*

multiple seeds and several SitEmo can be associated to the same sentence.
Emotions are described from the perspective of the writer, and the annotators need to be able to identify such emotions within the content and to pinpoint precise textual segments carrying emotions as basic SimEmo units.
In many cases, emotions can be clearly specified, see fig. 5 for an explicit example of *Pride* annotation, but in some cases emotions are hidden and their identification is not that simple.

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}

Figure 5: Labelled annotation of type *Pride*

Fig. 6 shows an example of implicit annotation of type *Sadness*. The type is not directly indicated in the sentence but rather inferred from the seed *catastrophique (catastrophic)*.

Notre pays est dans une *situation* ***catastrophique.***
(Our country faces a catastrophic situation)
SitEmo <Sadness> {Mode: Suggested Type: Basic, Category: Sadness, Trigger: catastrophique}

Figure 6: Suggested annotation of Type *Sadness*

Emotion clues are also detected by observing the behavior described in texts, either in an euphoric manner, i.e. combat radical (radical fight) or in an angry manner (i.e. ils militent (they campaign), nous réfutons (we reject)), see fig. 7.
The annotation procedure identifies distinct emotions within the sentence and explicitly annotates their modes, types and categories. Used jointly, the annotation procedure and scheme allows to built complex descriptions of emotions, such as:

- Mode: Suggested
- Type: Basic

Nous nous inscrivons dans un ***combat radical*** *et sans concession.*

(We are part of a radical and uncompromising fight.)

SitEmo <Anger> {Mode: Behavioral Type: Basic, Category: Anger, Trigger: combat radical}

Ils ***militent*** pour un journalisme factuel et traquent les informations erronées.

They campaign for fact-based journalism and hunt down misinformation.

SitEmo <Anger> {Mode: Behavioral Type: Basic, Category: Anger, Trigger: militent}

Toujours en première ligne, nous ***refusons*** la fatalité.

(Always on the front line, we reject fate.)

SitEmo <Anger> {Mode: Behavioral Type: Basic, Category: Anger, Trigger: refusons}

Figure 7: Examples of behavioural annotations

- Category: Fear

The annotation procedure creates a fine-grained enriched corpus.

5. Validation and first results

5.1. Validation of annotations

The task of annotating the entire corpus in terms of emotions was given to three annotators: two of them authored the paper (A1 and A2) and the other is a master student in computational linguistics (A3). The annotation started with a set of 267 annotations created by A2.

Two validation procedures were carried out, focusing on two different aspects. The first question was whether annotators agreed on values assigned to categories of emotions. This procedure is intended to validate the annotation scheme and the number and/or categories of emotions, as sometimes disagreement means adding emotions that are not included in the annotation scheme. For the first procedure, the initial annotation set was reviewed by annotator A1 in order to check the degree of agreement regarding the category of emotions. This validation identified 37 disagreements. When a1 and a2 are the sets of anchors annotated by A1 and A2, respectively, the recall of A2 with respect to A1 is given by :

$$Recall(A2||A1) = \frac{|a1 \cap a2|}{|a2|} \quad (1)$$

For emotion category, the value of Recall is 0.86 and fig. 8 illustrates cases that annotators disagreed on.

As shown in the examples in fig. 8, the disagreement is related to the limitations of the annotation scheme, and adding additional categories or clearly indicating secondary emotions covered by categories in the scheme can improve the agreement.

The second procedure investigated whether annotators would recognize the same linguistic units as seeds, and therefore triggers of SitEmo units. For this second validation, we started again with the set of annotations created

Les fondements de notre civilisation sont ***attaqués de tous parts***.

The pillars of our civilization are under attack from all sides.

SitEmo <Sadness> {Mode: Behavioral Type: Basic, Category: Sadness, Trigger: ***attaqués de tous parts***}

Suggested category: Threat

Nul ne peut ignorer ***les dangers qui menacent*** l'Europe.

No one can ignore the dangers threatening Europe.

SitEmo <Fear> {Mode: Behavioral Type: Basic, Category: Fear, Trigger ***les dangers qui menacent***}

Suggested category: Distress

Figure 8: Examples of disagreements on Category

by A2, but in this case the corpora was reviewed by annotator A3. During this revision, some anchors will be deleted, some anchors will be validated and kept as such while other anchors will be created if annotator A3 detects new seeds within the corpora. The initial anchor set $a2$ has 267 annotations, the final set $a3$ is composed of 563 annotations and the number of matching $a2 \cap a3$ is 198.

For the second validation procedure, we calculate F-measure as the mean of $Recall(A3||A2)$ and $Recall(A2||A3)$ to estimate the agreement of annotators about linguistic emotion units and triggers.

Recall(A3$\|\|A2$)	Recall(A2$\|\|A3$)	F-Measure
0.35	0.74	0.54

Table 4: Agreement on emotion seeds

As shown in tab. 4, there is an important asymmetry of recalls $Recall(A3||A2)$ and $Recall(A2||A3)$, because annotator A3 created a much larger set of anchors. The value of the F-measure is low, and this value can not be improved by adjusting the annotation scheme, since it reflects what annotators consider as subjective expressions in the corpora.

Nous devons nous prendre en charge par des ***actes de résistance***.

We shall take care of ourselves through acts of resistance.

SitEmo <Anger> {Mode: Behavioral Type: Basic, Category: Anger, Trigger: ***actes de résistance***}

Suggested trigger: prendre en charge

Le chaos découle des sociétés multiculturelles, et donc multi conflictuelles.

Chaos stems from multicultural, and therefore multi-conflictual, societies.

SitEmo <Fear> {Mode: Behavioral Type: Basic, Category: Fear, Trigger: ***Le chaos découle***}

Suggested trigger: multi-conflictuelles

Figure 9: Examples of disagreements on seeds (Trigger)

Fig. 9 illustrates disagreements on seeds, caused by subjec-

tive interpretations of the text by annotators.

5.2. Cross-analysis and remarks

The cross analysis of results was carried out by highlighting the distribution of SitEmo units in extremist and non-extremist data, respectively. As shown in fig. 10 , extremist data contains mainly emotions of *Anger* (45%). The second predominant emotion is *Fear* (19%) then, to a lesser extent, we detected emotions of *Sadness* (13%).

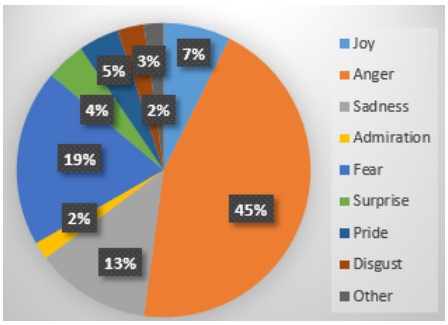

Figure 10: Categories of emotions in extremist data

The most representative emotion for non-extremist data is *Sadness*, see fig. 11 followed by *Anger* and *Fear*.

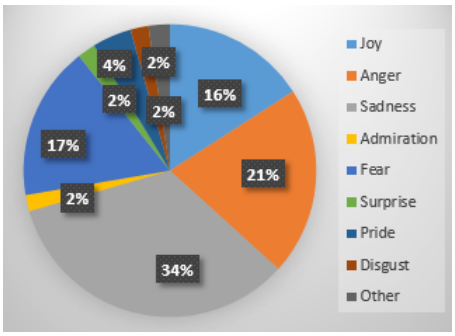

Figure 11: Categories of emotions in non-extremist data

Regarding the distribution of basic vs. complex emotions, the main complex emotion detected in non-extremist data is *Pride*. However, the percentage of *Pride* in extremist data (5%) is slightly higher than in non-extremist data (4%). Moreover, *Joy* is approximately twice as represented in non-extremist data (16% vs. 7% that in extremist data) and *Fear* is less dominant (17 % for non-extremist data vs. 19% in extremist data).

6. Conclusion and perspectives

This paper explores some of the issues that are faced when exploiting emotions correlated to extremist contents released on social platforms. The paper presents a French annotated corpus in which extremist categories were originally constructed by exploring extremist contents and emotion labels were provided by a linguistic annotation scheme defined independently of this work. The manual labelling of data sets is a labor-intensive task and semi-automatic procedures are needed (Canales et al., 2019). For this work, the annotation task might be made easier by first considering clusters of data sets according to sociology-driven perspectives and then labelling all textual units within a given cluster based on the manual review of a limited set of examples within the cluster. Ontologies of appraisal categories (Dragos et al., 2018) or hate (Battistelli et al., 2020) can also be used to guide the annotation.

Future work will address training machine learning algorithms for automatic classification of extremist contents and analyze the relevance of different features for extremist content detection. We will also extend the data sets by including new contents gleaned with additional keywords. The annotation of this new content could give a good overview of the robustness of the emotion annotation scheme and can help us verify if the annotation procedure is easily adaptable to other contents.

7. Acknowledgements

This document has been produced in the context of the FLYER project [4] funded by ASTRID, a research program supported by ANR (The French National Research Agency) and DGA (The French Defence Procurement Agency). The authors are grateful to Aurore Lessieux for her contribution to the annotation task.

8. Data availability statement

Data sets presented in this article are not available because of the nature of the material collected and analyzed.

9. Bibliographical References

Alatawi, H. S., Alhothali, A. M., and Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

Alava, S., Chaouni, N., and Charles, Y. (2020). How to characterise the discourse of the far-right in digital media? interdisciplinary approach to preventing terrorism. *Procedia Computer Science*, 176:2515–2525.

Araque, O. and Iglesias, C. A. (2021). An ensemble method for radicalization and hate speech detection online empowered by sentic computing. *Cognitive Computation*, pages 1–14.

Battistelli, D., Bruneau, C., and Dragos, V. (2020). Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.

Blanc, N. and Quenette, G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans: quelle méthodologie pour quels résultats? *Enfance*, (4):503–511.

Buechel, S. and Hahn, U. (2017). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Canales, L., Daelemans, W., Boldrini, E., and Martínez-Barco, P. (2019). Emolabel: Semi-automatic methodology for emotion annotation of social media text. *IEEE Transactions on Affective Computing*.

[4] https://anr.fr/Projet-ANR-19-ASTR-0012

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in french tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403.

Creissen, S. and Blanc, N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10ans ? apports d'une étude multimédia. *Psychologie Française*, 62(3):263–277. Cognition et multimédia : les atouts du numérique en situation d'apprentissage.

Davidson, D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30(3):232–242.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Donaldson, M. (2017). Plutchik's wheel of emotions—2017. update.

Dragos, V., Battistelli, D., and Kelodjoue, E. (2018). Beyond sentiments and opinions: exploring social media with appraisal categories. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1851–1858. IEEE.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Étienne, A. and Battistelli, D. (2021). Annotation manuelle des émotions dans des textes écrits avec la plateforme Glozz. Research report, MoDyCo ; Université Paris Nanterre, June.

Gaikwad, M., Ahirrao, S., Phansalkar, S., and Kotecha, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404.

Hartung, M., Klinger, R., Schmidtke, F., and Vogel, L. (2017). Identifying right-wing extremism in german twitter profiles: A classification approach. In *International conference on applications of natural language to information systems*, pages 320–325. Springer.

Liu, V., Banea, C., and Mihalcea, R. (2017). Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483. IEEE.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

Micheli, R. (2014). *Les émotions dans les discours. Modèle d'analyse, perspectives empiriques*. Champs linguistiques. De Boeck Supérieur, Louvain-la-Neuve.

Mohammad, S. M. and Bravo-Marquez, F. (2017). Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Oberländer, L. A. M., Kim, E., and Klinger, R. (2020). Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Öhman, E. (2020). Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.

Oussalah, M., Faroughian, F., and Kostakos, P. (2018). On detecting online radicalization using natural language processing. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 21–27. Springer.

O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., Bölte, S., Baron-Cohen, S., and Lundqvist, D. (2016). The eu-emotion stimulus set: a validation study. *Behavior research methods*, 48(2):567–576.

Raji, S. and De Melo, G. (2020). What sparks joy: The affectvec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Thorburn, J., Torregrosa, J., and Panizo, Á. (2018). Measuring extremism: Validating an alt-right twitter accounts dataset. In *International conference on intelligent data engineering and automated learning*, pages 9–14. Springer.

Widlöcher, A. and Mathet, Y. (2012). The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 171–180.

Alatawi, H. S., Alhothali, A. M., and Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

Alava, S., Chaouni, N., and Charles, Y. (2020). How to characterise the discourse of the far-right in digital media? interdisciplinary approach to preventing terrorism. *Procedia Computer Science*, 176:2515–2525.

Araque, O. and Iglesias, C. A. (2021). An ensemble method for radicalization and hate speech detection online empowered by sentic computing. *Cognitive Computation*, pages 1–14.

Battistelli, D., Bruneau, C., and Dragos, V. (2020). Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.

Blanc, N. and Quenette, G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans: quelle méthodologie pour quels résultats? *Enfance*, (4):503–511.

Buechel, S. and Hahn, U. (2017). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Canales, L., Daelemans, W., Boldrini, E., and Martínez-Barco, P. (2019). Emolabel: Semi-automatic methodology for emotion annotation of social media text. *IEEE Transactions on Affective Computing*.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in french tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403.

Creissen, S. and Blanc, N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10ans ? apports d'une étude multimédia. *Psychologie Française*, 62(3):263–277. Cognition et multimédia : les atouts du numérique en situation d'apprentissage.

Davidson, D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30(3):232–242.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Donaldson, M. (2017). Plutchik's wheel of emotions—2017. update.

Dragos, V., Battistelli, D., and Kelodjoue, E. (2018). Beyond sentiments and opinions: exploring social media with appraisal categories. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1851–1858. IEEE.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Étienne, A. and Battistelli, D. (2021). Annotation manuelle des émotions dans des textes écrits avec la plateforme Glozz. Research report, MoDyCo ; Université Paris Nanterre, June.

Gaikwad, M., Ahirrao, S., Phansalkar, S., and Kotecha, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404.

Hartung, M., Klinger, R., Schmidtke, F., and Vogel, L. (2017). Identifying right-wing extremism in german twitter profiles: A classification approach. In *International conference on applications of natural language to information systems*, pages 320–325. Springer.

Liu, V., Banea, C., and Mihalcea, R. (2017). Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483. IEEE.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

Micheli, R. (2014). *Les émotions dans les discours. Modèle d'analyse, perspectives empiriques*. Champs linguistiques. De Boeck Supérieur, Louvain-la-Neuve.

Mohammad, S. M. and Bravo-Marquez, F. (2017). Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Oberländer, L. A. M., Kim, E., and Klinger, R. (2020). Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Öhman, E. (2020). Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.

Oussalah, M., Faroughian, F., and Kostakos, P. (2018). On detecting online radicalization using natural language processing. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 21–27. Springer.

O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., Bölte, S., Baron-Cohen, S., and Lundqvist, D. (2016). The eu-emotion stimulus set: a validation study. *Behavior research methods*, 48(2):567–576.

Raji, S. and De Melo, G. (2020). What sparks joy: The affectvec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Thorburn, J., Torregrosa, J., and Panizo, Á. (2018). Measuring extremism: Validating an alt-right twitter accounts dataset. In *International conference on intelligent data engineering and automated learning*, pages 9–14. Springer.

Widlöcher, A. and Mathet, Y. (2012). The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 171–180.