Exploring the Emotional Dimension of French Online Toxic Content

Valentina Dragos, Delphine Battistelli, Fatou Sow, Aline Etienne

Exploring the Emotional Dimension of French Online Toxic Content

Valentina Dragos, Delphine Battistelli, Fatou Sow, Aline Etienne
ONERA - The French Aerospace Lab, Palaiseau, France
University of Paris Nanterre, Nanterre, France
valentina.dragos@onera.fr,
delphine.battistelli@parisnanterre.fr,
fatou.sow@onera.fr,
aline.etienne@parisnanterre.fr

Abstract

One of the biggest hurdles for the effective analysis of data collected on social platforms is the need for deeper insights on the content of this data. Emotion annotation can bring new perspectives on this issue and can enable the identification of content–specific features. This study aims at investigating the ways in which variation in online toxic content can be explored through emotions detection . The paper describes the emotion annotation of three different corpora in French which all belong to toxic content (extremist content, sexist content and hateful content respectively). To this end, first a fine-grained annotation parser of emotions was used to automatically annotate the data sets. Then, several empirical studies were carried out to characterize the content in the light of obtained emotional categories. Results suggest that emotion annotations can provide new insights for online content analysis and stronger empirical background for automatic toxic content detection.

Keywords: Emotion annotation, corpora exploration in French, toxic and harmful content, online data

1. Introduction

Social platforms enable new forms of interactions between users from different countries and cultures and play a central role in shaping the society (Domínguez et al., 2019). More specifically, those networks have a profound impact on propagating toxic or harmful content, such as extremist, hateful or violent messages and ideas (Castaño-Pulgarín et al., 2021). There is a need to understand the nature of online content in order to develop a response that accounts for the threat the propagation of those messages may pose to the society. As empirical evidence shows that emotions, and especially negative ones, trigger content spread (Berger, 2016), the question of emotion density in toxic content arises.

This paper explores the emotional density of three corpora of online data in French having (potentially) harmful content (see section 3 for the description those three corpora). Moreover, the paper also looks into the way these three corpora diverge in terms of categories of emotions conveyed. To investigate to what extent emotions can help the characterization of corpora falling under the umbrella of toxic content (Thomas et al., 2021) is pf particular interest. Furthermore, our special concern is to investigate to what extent categories of emotions (e.g. Joy, Sadness, etc.) are distributed in a comparable way (or not) in different corpora. Intuitively, it can be doubted for example that emotions are distributed in a comparable way in sexist data and in hateful data.

Most approaches for online data analysis focus on building models for automated detection of specific toxic attitudes , such as sexism (Kumar et al., 2021), misogyny (Pamungkas et al., 2020b) or hate (Ketsbaia et al., 2020). For those models, detection is cast as a binary classification task (e.g. hateful vs. non hateful, sexist vs. non sexist). However this binary classification fails to capture the nuances of content and there is a need to develop annotated corpora that can shed a better light on the features that allow making the distinction, such as emotions as investigated in this paper. Currently, corpora annotated with emotions are scarce, and most of the resources have been created for English. This paper addresses those challenges, and has the following main contributions:

- A new task is proposed in the field of online discourse characterization, focusing on the emotional analysis of three types of toxic data in French collected online;
- The construction of three textual corpora having two layers of characterization, indicating both the type of toxic attitude and the categories of conveyed emotions. Those corpora can be used by the research community for further experiments;
- The paper contributes to the development of resources for languages different from English and focuses on building emotionally annotated corpora in French.

LREC-COLING 2024, pages 6945–6954
20-25 May, 2024.

The reminder of the paper is structured as follows : Section 2 discusses related approaches addressing the characterization of toxic content and the emotional annotation of social data. Section 3 offers an overview of corpora, while data processing is presented in section 4, with emphasis on emotional annotation. Case studies and experiments are presented in section 5 and the last section concludes the paper and sketches directions for future work.

2. Related Approaches

The analysis of online data has attracted noticeable attention in the last years in the field of natural language processing (NLP). To ground the study, this section presents related approaches along two dimensions: the first discusses computational methods to detect online hate, sexism, violence and other toxic content in social media posts; the second investigates more specifically approaches taking into account the emotions as they are expressed in languages.

2.1. Detection of toxic discourse on social media

Sexism detection is a topic largely addressed by research efforts tackling online data analysis. An exhaustive approach for sexism detection in Spanish is presented in (Rodríguez-Sánchez et al., 2020). The study aims to understand how sexism is expressed in social networks conversations and first builds a corpus of sexist expressions as identified in tweets. This resource is further used to train and test several traditional and deep learning approaches. By using a variety of features, the authors compare those approaches and shows that BERT (Devlin et al., 2019) provide the best detection results and reaches an accuracy of around 74% for sexism detection. The study also provides an analysis of errors, and shows that the accuracy of detection is affected by both linguistic phenomena (irony, sarcasm) and the limited size of the corpus.

Detection of online hate across several language is discussed in (Corazza et al., 2020). In this paper, a robust neural architecture is developed and several experiment are carried out to evaluate the robustness of this architecture when processing online data in three languages: Italian, English and German. Most specifically, the authors investigate the contribution of several components including text features such as word embeddings, unigrams and bigrams, but also emotions, hashtags and emojis. Conclusions are summarized in the form of findings: for example, using domain-specific embeddings improves the performance of approaches for all languages, although taking into account the set of emojis yields the best accuracy for English, but has no impact on German. The summary of findings can guide the development of novel approaches for online hate detection.

Detection of swearing, a particular type of offensive language is tackled in (Pamungkas et al., 2020a). The main contribution of the paper is the development of the SWAD corpus, composed of tweets that are manually annotated for swearing at word level. This corpus is then used to train learning models to perform automatic detection of swearing, and the authors show that the fine-grained annotation improves the accuracy of detection. In addition, an in-depth analysis of results provides new insights on the most predictive features allowing the detection. Results show that bigrams, emotion and syntactic features all improve the classification performance, while the Twitter features (hashtags) have a detrimental effect decreasing the accuracy of learning models.

Those approaches discuss methods developed to detect different types of toxic content by ignoring linguistic phenomena such as irony and sarcasm. However, several studies show that for instance, sarcasm can also be a source of abusive and harmful content (Frenda et al., 2022).

2.2. Emotion annotation for social media analysis

As shown above, detection of toxic content is an established task for NLP research community, although the recognition of linguistic phenomena like expressions of emotions within this type of content received less attention. Some contributions to emotion investigation for social data analysis independently of the question of toxic content are listed below.

EmoEvent is a corpus manually annotated with seven labels for emotion categorisation: six basic emotions from the Ekman model (Ekman et al., 1999) (*Anger, Disgust, Fear, Joy, Sadness, Surprise*) and an additional *neutral or other* emotion for unspecified cases (Plaza-del Arco et al., 2020). The corpus comprises tweets in English and Spanish and it is based on events, such as festivals, incidents political unrest or global strikes. The corpora was used mainly to observe the differences in how people express their emotions, and the paper also presents preliminary results on using thus corpus for automatic detection of emotions in social posts. The authors also performed preliminary experiments on emotion detection, and results show that classifiers trained on the corpora identify more accurately *Joy* and *Sadness*, while the other emotions (*Fear, Anger, Disgust* and *Surprise*) remain difficult to detect for the classifier.

The integration of several corpora annotated with emotion according to various guidelines and annotation schemata is presented in (Oberländer and

Klinger, 2018). After analysing all annotation formats and comparing annotated corpora, the authors developed a unified format for emotion annotation, taking into account the features and advantages of each annotation schema. Then, all corpora have been mapped to this format in order to build an integrated resource. The aggregated corpora can be used for enhanced experiments, as it offers a variety of corpora augmented with annotations layers. The benchmark allows to various users to select the most suitable corpus for a specific domain and opens up the possibility of enhanced experiments, such as transfer learning and domain adaptation, although there are no further indications about how predictive the emotion annotations can be.

In order to simplify the emotion annotation task, EmolLabel, a semi-automatic annotation methodology is presented in (Canales et al., 2019) . EmoLabel consists of two phases; first, the preannotation used both supervised and unsupervised approaches in order to label sentences by using emotion categories; then, a manual refinement allows validating the annotation labels. Experiments show that the global methodology, combining automatic preannotation and expert validation is more effective and less time consuming. Most specifically, the supervised preannotation provides better results in terms of inter annotator agreement.

In the field of NLP, automatic analysis of emotions in written texts is generally addressed exclusively through the notion of emotional category (e.g. *Joy, Fear*, etc.) - often with a focus on a sole linguistic mean to express emotions, the emotional lexicon. However, as pointed out by linguistics (Micheli, 2014), psycho-linguistic works (Creissen and Blanc, 2017) and NLP very recent works (Troiano et al., 2023), (Cortal et al., 2023), this is not sufficient to explore, and then to identify, emotions in their diversity of modes of expressions in texts. Moreover, from a strictly NLP and/or information extraction point of view, there is a need to consider the huge diversity of emotion expressions (thus, not only the strictly lexical ones) in order to better quantitatively capture the emotions. For example, emotions can be expressed by interjections, as described in (Fraisse and Paroubek, 2015) or emojis (Battistelli et al., 2023). They can also be revealed by appraisals as described in (Klinger, 2023), behaviours or suggestions as presented in (Etienne et al., 2022).
(Etienne, 2023).

The work presented in this paper is closely related to those approaches. The originality stems from the nature of data used, namely toxic content in French and the compared analysis of those corpora with respect to their emotional dimension.

3. Overview of Corpora

The work uses a collection of three distinct corpora for which we analyzed the emotional dimension. Twitter was used as a data source for this research as it allows quick and relatively easy access to people's views and opinions. In addition, views shared on social media often complement data collected through traditional methods (e.g. questionnaires in sociology) and may capture original views that are possibly underrepresented using other methods of data gathering. Corpora were collected on social platforms and are categorized by indicating several types of toxic attitudes: sexist, hateful, extremist, as described hereafter.

Corpus of right-wing extremist attitude (henceforth named C1) This corpus was created by a previous project investigating the nature of extremism online [1] NOTE BAS FLYER. It comprises Tweets and messages collected on discussion forums (Dragos et al., 2022). Data was collected by using a combination of hashtags and keywords that are specific to right-wing extremism. The corpus 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 validate the content. Thanks to this expert validation, the corpus was certified as conveying radical attitude and was divided into two categories, composed of radical extremist and non-radical extremist data, respectively. Although the entire collection is composed of extremist content, those finer categories (radical and non-radical) have a practical meaning, from a sociology-specific perspective and highlight differences of sources, hashtags and keywords. The corpus comprises 1728 textual units (composed of several sentences), out of which 1129 were labeled as radical extremist and the remaining 599 were annotated as non-radical extremist. This corpus is slightly imbalanced.
Sentences (E1) and (E2) show non-radical and radical examples.

E1: *Ils veulent que vous restez pauvres.* (They want you to stay poor.)
E2: *La France doit rester la France, notre patrie sacrée.* (France must remain France, our sacred homeland.

Corpus of sexist attitude (henceforth named C2) This corpus was created by a previous study dedicated to sexism detection in online data (Chiril et al., 2020) in French. The collection comprises around 12 000 tweets collected online and manually annotated with two labels indicating whether a tweet is sexist or non-sexist. The distribution

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

of tweets is not balanced, and the collection comprises 7789 non-sexist tweets and 4046 tweets labelled as sexist.

The following examples illustrate sexist (E3) and non sexist (E4) tweets:

E3: *Vous vous dites femmes vous savez même pas faites des pâtes, bande de connasse* (You say you're women you don't even know how to make pasta, you bitch)

E4: *Les bleues font un très, très, très bon match ! #FRANOR @X* (The blue ones make a very, very, very good game! #FRANOR @X)

Corpus of hateful attitude (henceforth named C3) This corpust comprises around 600 Tweets manually collected to highlight hateful and non hateful attitude (Battistelli et al., 2020). The main goal of this resource was to facilitate the binary classification of hateful/ non hateful Tweets and thus the corpus is annotated with hateful/ non-hateful labels. The annotation was carried out manually by taking into account the main characteristics of online hateful content, as highlighted by definitions largely adopted in the literature (Malecki et al., 2021). This data set is well balanced, and the two classes (hateful and non-hateful) contain almost the same number of tweets.

The following examples illustrate hateful (E6) and non hateful (E5) tweets:

E5: *Quand il s'agit d'entretenir les étrangers parasites on est toujours sur de trouver les gauchistes.* (When it comes to nurture the parasitic foreigners we are always sure to find the leftists.)

E6: *J'espère que tu vas avoir un cancer et mourir!* (I hope you get cancer and die!)

Remarks: all corpora contain user-generated data that was collected online in the frame of different projects and by different research teams. The collections have different size AND nature /// and annotations///, as shown in Table 1 and Table 2.

Corpus	Nature	Balanced
C1	Radical vs. NonRadical	NON
C2	Sexist vs.NonSexist	NON
C3	Hateful vs.NonHateful	YES

Table 1: Characteristics of the three corpora

Corpus	Size	Type
C1	1728 text units	Tweets, forums
C2	12 000 tweets	Tweets
C3	600 Tweets	Tweets

Table 2: Size and type of corpus C1, C2 and C3

Data was collected with specific keywords and hashtags, which is to say related to extremism, sexism and hate, but without using emotion-specific markers. These corpora were chosen for two main reasons: first, all data were collected online, and therefore one can detect emotions as often people share data on social platforms by adding their own thoughts and feelings; toxic content is also released, as online anonymity allows users to express themselves without reservation. second, corpora reveal different attitudes, which makes it possible to test the robustness of the annotation schema in different contexts.

4. Data Processing and Emotion Annotation

4.1. General architecture

The general architecture developed for this work includes three steps: pre-processing of data, emotional annotation, and analysis of emotions distributions, as shown in Fig. 1.

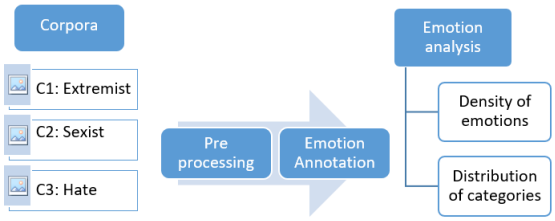

Figure 1: Overview of the general architecture

Pre-processing filters out URLs and performs lemmatization and tokenization.

The next steps of the methodology are presented hereafter: section 4.2 discusses the annotation tool, while the last two steps will be detailed in section 4.2, as they are specific to use cases.

4.2. Annotation Tool for Emotion Detection

The annotation tool adopted for this work was developed independently of the task of online toxic content characterization. This paragraph describes the tool, the underlying annotation schema and the tool evaluation.

Emotion annotation schema This work used Emotyc, the parser of emotions in French developed by (Battistelli et al., 2022). Emotyc was trained on a French genre-diversified corpus of texts (three genres have been considered: fictional, encyclopedic and journalistic) manually annotated with different kinds of emotional labels as they are described in an annotation schema (Etienne, 2023). The annotation schema integrates relevant notions from linguistic and psycho-

linguistic perspectives to characterize emotions and describes emotional textual units by considering three main features, see Fig. 2:

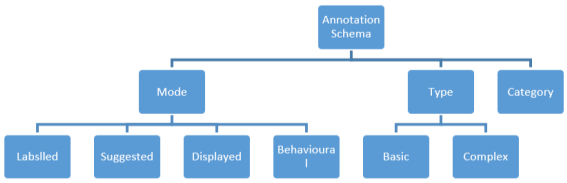

Figure 2: Emotions annotation schema

Emotion mode: this feature indicates how the emotion is expressed. Emotions can be: labeled, displayed, suggested or behavioural.

Labeled emotions are directly designated trough an emotional label, i.e. an emotional lexicon word, such as ”happy”, ”anger”or ”afraid” (Creissen and Blanc, 2017), (Micheli, 2014). Fig. 3 shows an example of *Pride* annotation, where the emotion is clearly specified in the text.

***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 3: Labelled annotation of type *Pride*

Displayed emotions are conveyed by various linguistic characteristics of utterances, that occur when the speaker feels an emotion at the time of utterance. These characteristics show that the speaker was experiencing an emotion. The reader/interlocutor relies on them to infer the emotional state of the speaker. Markers that display an emotion take many forms, for instance words like interjections (e.g. “oh”), syntactic structures like nominal sentences (e.g. “So many presents !”), or typographic marks (e.g. “!”) (Micheli, 2014).

Suggested emotions are expressed through the description of situations associated with emotions by social conventions. Thanks to these conventions, the reader/interlocutor infers the emotion from the depicted situation (Creissen and Blanc, 2017); (Micheli, 2014). For instance, in many western European societies, receiving a present is usually associated with a positive emotion, like *Joy*. Describing this type of situation can thus convey *Joy*.
Fig. 4 shows an example of suggested annotation of type *Sadness*. For this example, the emotion 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 4: Suggested annotation of Type *Sadness*

Behavioral emotions are indicated by descriptions of emotional behaviors, for example “crying” or “smiling” (Creissen and Blanc, 2017). The reader relies on the depicted behavior to infer the emotion felt by the character.

Emotion type and category. Those features deal with the emotional category and the type of emotions, as expressed by the annotated markers. The annotation schema introduces eleven categories, that can be basic or complex, see table 3.

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

Table 3: Types and categories of emotions

The six basic emotions are those introduced in (Ekman et al., 1999): *Anger, Disgust, Joy, Fear, Surprise* and *Sadness.* Four of the complex categories were taken from (Blanc and Quénette, 2017) and (Davidson, 2006) and include *Guilt, Embarrassment, Pride* and *Jealousy. Admiration* was added as a fifth complex category to better balance emotion types in the schema. Each of the eleven categories corresponds to more specific identifications of emotions. For instance, *Anger* regroups anger but also *Annoyance, Rage* and *Fury*. Since the schema’s eleven categories are not sufficient to account for the diversity of emotions, an additional unit called *Other* was defined. It is to be used to annotate markers expressing other emotions not captured by the previous eleven categories such as *Disdain, Love* or*Hate*. When a textual unit conveys several emotions at the same time, the features Category2 and Type2 can be used to tell which second emotion is expressed. When only one emotion is detected, the default value of those features is “None”.

Emotyc, an automatic parser of emotions in texts The manual application of the annotation schema on a corpus of more than 1,500 French texts diversified in genres allowed creating an

emotion-annotated corpus. This corpus was used to develop the Emotyc classifier, a tool for the automatic analysis of emotions in texts. Emotyc was developed using deep learning techniques and more specifically the CamemBERT model (Martin et al., 2019). Emotyc performs automatic annotation, as shown in fig. 5 and carries out emotion analysis at several incremental levels: [task A] predicting whether or not a sentence contains an emotion; [task B)] if so, which type of mode of expression is used; [task C] whether it is a basic or complex emotion; and [task D] which emotional category it falls into.

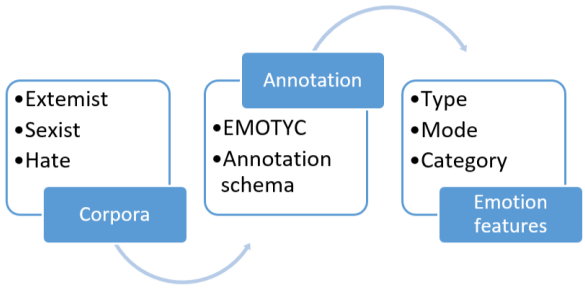

Figure 5: Annotation procedure

[2]

The Emotyc classifier efficiently finds emotional information and achieves good performances for example: whether a sentence conveys emotions (value of F1 = 0.74), identification of emotion mode (value of F1 = 0.80), detection of the basic type of emotions (value of F1 = 0.71). For emotion category identification, the classifier achieves surprise the values of (F1 0.71 for *Surprise* and F1 =0.67 for *Fear* battistellichaine.

Evaluation of Emotyc on social data The quality of Emotyc annotations was evaluated by comparing the set of Emotyc annotations against a set of manual annotations carried out by linguistic experts on the extremist corpus only.

The evaluation considers two Emotyc tasks: [task A] predicting whether or not a sentence contains an emotion and [task D] which emotional category it falls into.

The measures of Precision, Recall and F-Measure were estimated. Tables 5 and 4 show the values of those measures for Tasks A and D.

Those tables show high values of Precision but low values of Recall, for both tasks A and D.

[2]The Emotyc classifier is integrated into the automatic text processing chain of the TextToKids project through Emotyc to extract emotional descriptors. The tool can be tested online at the following address: http://texttokids.ortolang.fr.

Measure	Precision	Recall	F-Measure
Values	0.85	0.34	0.49

Table 4: Detection of emotional sentences (task A)

Measure	Precision	Recall	F-Measure
Values	0.80	0.25	0.38

Table 5: Detection of emotion categories (task D)

Those values indicate that Emotyc provides annotations of good quality, while having a low capacity to detect all the emotions conveyed by the corpora.

Analysis of errors and ambiguous cases In order to understand the low values of Recall, 50 annotated sentences have been randomly selected and analysed by a human annotator and the system. First, the evaluation highlighted an intrinsic drawback of Emotyc, namely a majority of false positives for *Admiration*. For example, examples E7 and E8 illustrate erroneous annotations, as *Admiration* was detected by Emotyc, although there is no such emotion according to human analysis:

E7: Mais attention de ne pas penser que les MGTOW ne veulent absolument pas fonder une famille ou vivre en couple. (But be careful not to think that MGTOWs do not want to have a family or live as a couple.)

E8: Inutile de vouloir redresser les choses à grande échelle. (There is no point trying to scale up.)

Since Emotyc is a CamemBERT-based model, the over detection of *Admiration* can be due to bias in the training of the initial language model. Other errors occur when Emotyc fails to detect emotional sentences, as shown in the following example, annotated with *Anger* by expert annotation but not identified as conveying emotion by Emotyc.

E9: *Elles bafouent publiquement les lois françaises.* (They publicly flout the French laws).

The evaluation also identified several ambiguous cases, when both human annotation and Emotyc detect emotions but they are different, and more specifically there is no clear distinction between the types of emotions. The following example shows an ambiguous case, manually annotated with *Pride* and annotated with *Joy* by Emotyc.

E10: *Fiers de notre héritage et confiant en notre avenir, nous ne reculerons pas!* (Proud of our heritage and confident in our future, we will not step back).

Erroneous and ambiguous cases highlighted by the qualitative analysis have also an impact on the values of Precision and Recall.

Emotyc was used for emotion annotation of the three corpora conveying toxic attitudes and described in section 3. Results are presented hereafter.

5. Experiments and case studies

Experiments were carried out by in the following settings: first, Emotyc is used for automatic annotation of all corpora; then the distribution of emotion annotations is analysed manually.

Analysis of Emotion Distribution This paragraph details the results of annotation and discus some interesting findings related to the results of the annotation task.
As a global indicator, the density of emotion annotations was calculated for all corpora by dividing the number of emotions annotations by the size of the corpus, as shown in table 6.

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

Table 6: Density of emotions for corpora C1, C2 and C3

The values of emotion density show that corpora C1 and C2 have similar percentages of emotions, in spite of their distinct size, nature and collection mode. In addition, the density of emotions in corpus C3 is very high, although the corpus is composed on hateful and non hateful messages. A more detailed analysis of emotion annotation at corpus level is described hereafter.

Emotion annotations in C1 corpus Tables 7 and 8 show the distribution of emotion categories within the two classes of C1 corpus: the radical and non-radical classes, respectively.

Anger	Admiration	Fear	Sadness	Surprise
188	144	99	75	52
30.12	23.07	15.86	12.01	8.33

Table 7: C1: NonRadical: distribution of emotion categories

Anger	Admiration	Fear	Sadness	Surprise
108	91	62	45	35
27.76	23.39	15.93	11.56	7.45

Table 8: C1: Radical: distribution of emotion categories

As illustrated above, the two classes have a very similar distribution of emotions, with *Anger*, *Admiration* and *Fear* being the most prominent emotions detected. Not only the top five emotions are identical, but the values of their proportions are very similar as well.
This result can be explained by the nature of the corpus: although radical and non-radical classes have been identified as such by researchers in sociology, the whole corpus gathers extremist data. From a practical standpoint, the annotation with emotions is not able to highlight the features that are specific to each classes, given the similarity of their content. The fine-grained granularity of classes requires humane expertise in order to make a distinction.

Emotion annotations in C2 corpus Tables 9 and 10 show the distribution of emotions within the two classes (sexist and non-sexist) of the sexist corpus.

Anger	Admiration	Sadness	Surprise	Fear
1131	698	411	260	222
35.85	22.13	13.03	8.24	7.03

Table 9: C2: Sexist: distribution of emotion categories

Anger	Admiration	Joy	Surprise	Sadness
1261	978	501	415	407
30.49	23.65	11.11	10.03	9.84

Table 10: C2: NonSexist: distribution of emotion categories

The tables show a distinct distribution of emotions for sexist and non-sexist content, but not as clear as the distinction between hate and hon date classes. Thus, the first two relevant emotions are identical for both classes: *Anger* and *Admiration*. Moreover, *Anger* is more frequent in the sexist class, having 35.85% of all emotion occurrences, while the percentage for the non sexist class is slightly lower (30.49%). In addition, differences occur only for the third, fourth and sixth emotions detected and annotated.

The distribution of emotions shows that, although emotions are different in the sexist and non-sexist classes, the differences are not very clearly emphasized by emotion annotation.

Emotion annotations in corpus C3 The distribution of emotions within the classes of the online hate dots set is shown in tables 11 and 12.

The distribution of emotions highlights a clear distinction between the hate and non-hate classes.

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

Table 11: C3: Hate: distribution of emotion categories

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

Table 12: C3: NonHate: distribution of emotion categories

While *Anger* is the most relevant emotion for the hate class, followed by *Admiration* and *Sadness*, the non hate class exhibits *Admiration*, *Surprise* and *Anger* as specific emotions. The values are also significantly different: while *Anger* represents 61.5% of the overall emotions identified in the hate class, but only 15.68 of emotions detected in non-hate class. Moreover, *Admiration* covers 35.78 of emotions in non-hate class, but only 13% for hate class.

The comparison of emotions and their values as detected in both classes shows a clear distinction between hate and non-hate contents that can be captured thanks to emotion annotation.

6. Conclusion and Future Work

This paper investigates the emotional dimensions of data collected on French social platforms. The study compared, though the lens of emotions, several corpora having toxic content. Annotation was carried out automatically, by using the Emotyc tool which is based on a rich annotation schema. Results indicate that emotions can highlight differences of and capture variation in online toxic content. More specifically, several categories of emotions including *Anger, Joy, Surprise, Fear* and *Sadness* are identified as the most prevalent within each corpus.

Case studies described in this paper were conducted on corpora having related topics: sexism can be considered as a particular type of online hate, and several connections between extremism and online hate have been highlighted in the literature. One practical question was whether there are similarities of emotion annotations at corpora level, not only between the classes of the same corpus. Hence, a global analysis of emotion distribution shows that *Anger* is the most frequent emotion detected in five cases out of six. The non-hate is the only class not having *Anger* as the prominent emotion. These results are in line with previous studies that demonstrated that *Anger* is a prominent emotion characterising extremist contents (Dragos et al., 2022).

The main direction for future work aims at training accurate models for toxic content detection that are able to take into account emotion labels. Moreover, fine-tuning of the Emotyc model in order to cope with the intrinsic limitations highlighted by this study is also envisioned ,

Data availability statement

Annotated corpora are available for research purposes upon request.

Ethical considerations and limitations

For ethical concerns, keywords are not provided in this paper and the results are presented in a way that avoids reidentification of any URIs and hashtags that have been used to collect online data. All URIs where deleted after preparing this work. Although carefully selected, some examples may still be considered harmful.

7. Bibliographical References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. 2020. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.

Delphine Battistelli, Aline Etienne, Rashedur Rahman, Charles Teissèdre, and Gwénolé Lecorvé. 2022. Une chaîne de traitements pour appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycholinguistique.

Delphinel Battistelli, Valentina Dragos, and Jade Mekki. 2023. Annotating social data with speaker/user engagement. illustration on online hate characterization in french. In *International Conference on Computing and Communication Networks 2023: ICCCN 2023*.

Nicole Baumgarten, Eckhard Bick, Klaus Geyer, Ditte Aakær Iversen, Andrea Kleene, Anna Vibeke Lindø, Jana Neitsch, Oliver Niebuhr, Rasmus Nielsen, and Esben Nedenskov Petersen. 2019. Towards balance and boundaries in public discourse: expressing and perceiving online hate speech (xperohs). *RASK: International Journal of Language and Communication*, 50(Autumn 2019):87–108.

Jonah Berger. 2016. *Contagious: Why things catch on*. Simon and Schuster.

Nathalie Blanc and Guy Quénette. 2017. Identifier les émotions du protagoniste du récit. In *58ème Congrès Annuel de la Société Française de Psychologie" Diversité, Connaissances, Emotions", Université Nice Sophia Antipolis, Université Côte d'Azur, Nice, 30 août-1er septembre 2017*.

Lea Canales, Walter Daelemans, Ester Boldrini, and Patricio Martínez-Barco. 2019. Emolabel: Semi-automatic methodology for emotion annotation of social media text. *IEEE Transactions on Affective Computing*, 13(2):579–591.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in french tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. *arXiv preprint arXiv:2305.10446*.

Sara Creissen and Nathalie Blanc. 2017. Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans? apports d'une étude multimédia.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Caldevilla Domínguez, José Rodríguez Terceño, and Almudena Barrientos Báez. 2019. Social unrest through new technologies: Twitter as a political tool. *Revista Latina de Comunicación Social*, (74):1264–1290.

Valentina Dragos, Delphine Battistelli, Aline Etienne, and Yolène Constable. 2022. Angry or sad? emotion annotation for extremist content characterisation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 193–201.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Aline Etienne. 2023. *Analyse automatique des émotions dans les textes : contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants*. Theses, Université de Nanterre - Paris X.

Aline Etienne, Delphine Battistelli, and Gwénolé Lecorvé. 2022. A (psycho-) linguistically motivated scheme for annotating and exploring emotions in a genre-diverse corpus. In *13th Conference on Language Resources and Evaluation (LREC 2022)*.

Amel Fraisse and Patrick Paroubek. 2015. Utiliser les interjections pour détecter les émotions. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 279–290.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Lida Ketsbaia, Biju Issac, and Xiaomin Chen. 2020. Detection of hate tweets using machine learning and deep learning. In *2020 IEEE*

19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pages 751–758. IEEE.

Roman Klinger. 2023. Bridging emotion role labeling and appraisal-based emotion analysis. *arXiv preprint arXiv:2309.02092*.

Ritesh Kumar, Soumya Pal, and Rajendra Pamula. 2021. Sexism detection in english and spanish tweets. In *IberLEF@ SEPLN*, pages 500–505.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, pages 1–36.

WP Malecki, Marta Kowal, Małgorzata Dobrowolska, and Piotr Sorokowski. 2021. Defining online hating and online haters. *Frontiers in Psychology*, 12:744614.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Raphaël Micheli. 2014. Les émotions dans les discours. modèle d'analyse, perspectives empiriques.

Piotr Miłkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocoń. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th international conference on computational linguistics*, pages 2104–2119.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN postproceedings*, pages 134–144.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020a. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6237–6246.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020b. Misogyny detection in twitter: a multilingual and cross-domain study. *Information processing & management*, 57(6):102360.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.