Annotating social data with speaker/user engagement. Illustration on online hate characterization in French

Delphine Battistelli, Valentina Dragos, Jade Mekki

Annotating social data with speaker/user engagement. Illustration on online hate characterization in French

Delphine Battistelli[1], Valentina Dragos[2], and Jade Mekki[1]

[1] Paris Nanterre University, 200 Av. de la République, Nanterre, France
[2] ONERA-The French Aerospace Lab, 6 chemin de la Vauve aux Granges, Palaiseau, France

Abstract. This paper presents an annotation framework relying on the linguistic notion of speaker/user engagement. This notion is suitable for a finer characterization of online hateful discourse as it allows addressing the following question: does the speaker engage himself to the truthfulness of hate content? Two resources were built to support the annotation framework: a taxonomy of speaker/user engagement degrees and a rich semantic resource of pictograms/emoticons. The paper describes the resources used and the annotation process. Preliminary experiments and results on sexism characterization in French are also presented.

Keywords: opinion mining · social computing · extremist content

1 Introduction

Online hate includes abusive, insulting, intimidating, and harassing expressions that support violence, hatred, or discrimination against specific targets related to race, ethnic origin, religion, gender, age, physical condition, disability, and sexual orientation of persons [20]. Linguistically, online hate speech is manifold. If hatred content is obviously highlighted by offensive words and explicit incitation to violence [5], it can also be less explicit. For example, [9] showed that for online data, stereotypes propagate implicit sexism. Another relevant linguistic aspect specific to hatred contents is that one can also immediately distinguish different ways of marking the user engagement to the truthfulness of what is said.

This study is motivated by several limitations of current approaches developed for online hate detection. Those limitations are illustrated by the examples (1) to (4).

- (1) Tous ces migrants envahissent nos hôpitaux!!! (All these migrants are invading our hospitals!!!)
- (2) Il a raison de dire que tous ces migrants envahissent nos hôpitaux! (He is right to say that all these migrants are invading our hospitals!")
- (3) Entendre dire que tous ces migrants envahissent nos hôpitaux me révolte! (To hear that all these migrants are invading our hospitals revolts me!)
- (4) Il a dit que tous ces migrants envahissent nos hôpitaux. (He said that all these migrants are invading our hospitals).

(1) is a direct and explicit hateful content expressed by the speaker/user with an implicit engagement to the truthfulness of what is said in the statement; (2) is an indirect

hateful content as it indicates an opinion expressing support to hateful content, and thus there is a clear engagement to the truthfulness of what is said ; (3) is similar to (2), but it indicates an opinion expressing no support to hateful content and there is a clear disengagement with respect to the truthfulness of what is said ; (4) is "just" reporting a hateful content, thus there is just an engagement to the truthfulness of a speech act and not of the hate content it contains. Currently, automatic approaches address the detection of direct online hate speech like (1) by using training data composed of explicitly hatred content, or by creating lexicons of hate-relevant words [3]. However, they fail to distinguish cases similar to the examples illustrated in (2), (3), and (4). The analysis of those cases requires creating datasets labeled by considering not only intrinsic features of abusive language but also features of what can be analyzed as different positions or attitudes against the truthfulness of hate content. This paper investigates the following research question: how useful could be the linguistic notion of a speaker's engagement in a better characterization of what is a hateful attitude, especially among these four types of attitudes denoted by the examples (1) to (4) mentioned before? The paper has three main contributions: first, it explores the notion of user engagement as a feature to be further used to better characterize and detect hatred content; then, it describes the construction of a resource of pictograms integrating this notion of user engagement and finally the paper contributes to the analysis of online content in French. Moreover, this study provides a new perspective on online hate in French social media, and brings new in-depths linguistically grounded analysis that benefit society. The case of hate speech detection provides valuable insight into the promises and limitations of automated online content characterization for sociological analyses.

The outline of the paper is the following: section 2 presents a selection of related approaches. The construction of resources is discussed in section 3 while section 4 illustrates the use of resources for the annotation of sexist tweets in French. Conclusion and directions for future work end the paper section 5.

2 Related work

Engagement is related to speakers' own perspective on informational content and indicates whether they assert with confidence what they perceive in the environment, know from their experience, interact with, or attend to. In some particular cases, for example, when indirectly reporting, the authors can only assume, to varying degrees of certainty, and this aspect is also covered by engagement [6]. Although engagement is a quite well-established notion in linguistics and has been explored at the intersection of modality, evidentiality, and commitment categories [10], there has been less work on considering speaker/user engagement for social data analysis. However, previous research confirms that, for open-dialogue systems, taking into account user engagement as real-time feedback benefits the analysis of social interactions [19]. More specifically, online users use emojis to enrich the context and convey additional emotions, and using emojis increases user engagement [22].

Emoticons are extremely popular on social platforms and there is a growing literature investigating how they affect the interpretation of messages online [11]. The main topics investigate whether emoticons capture emotional states [23] and shed light on this

connection by compiling data sets [13] which quantify people's reported association between emoticons and emotions. Several experimental studies illustrate the impact of considering emoticons for online content analysis. First, for visual sentiment analysis, [2] demonstrate that, while emoticons carry a sentiment signal by themselves, they also act as sentiment modifiers of surrounding contents. Then, for opinion detection, implementing a multi-modal fusion by combining emoticons with text improves the performance of opinion detection in terms of recall and precision [1]. Recent works also investigate the intrinsic semantics of emoticons [15] and their links with semantic representations of general, abstract, or concrete, concepts. More specifically, emoticons prove to be good indicators for non-direct hate speech detection, as statistical links mapping certain emoticons and key terms of hatred discourses can be used to assess differences in sentiments conveyed by those contents [7].

Taking a step further, several resources such as the Emoji Sentiment Ranking [18] or Emojipedia [3], have been created to describe the semantics of emoticons and make explicit relations between emoticons and emotions. Whilst resources provide different perspectives about emoticons and their meanings, they do not take into account their context of usage [12]. However, emoticons are generally considered as variant artifacts, and their meaning change due to time [21]. From a different perspective, studies also consider the evaluation of detection models based on emojis, and [17] presents HatemojiCheck, a test suite of 3,930 short-form statements that allows to evaluate performance on hateful language expressed with emoji and highlights weaknesses in existing hate detection models.

As shown above, using emoticons for social data analysis was tackled by several research studies, but there are not numerous approaches considering both emoticons and user engagement. The core insight of this work is to provide an annotation framework aggregating labels based on the analysis of speaker/user engagement, which is expressed, by both strictly linguistic markers and pictograms. Unlike previous work that mainly develop training data or pre-trained models, we focus on the development of linguistic resources that can be integrated into various classification approaches. Keeping apart the construction of resources and the implementation of algorithms can also be helpful when considering how annotation procedures, sampling data and training models can be merged together. In addition, the use of annotation categories provides a finer characterization of online hate and allows us to go beyond the limitations of the binary classification. This is a trendy research topic and the need for a finer characterization of online hate is also highlighted by the tasks recently addressed at SemEval-2023 challenge [16], which included a four-class classification for sexist posts into threats, derogation, animosity, and prejudiced discussions.

3 Building resources to describe user engagement

This section presents the construction of two resources capturing user engagement under two headlines. The first one is a taxonomy for labeling different degrees of user engagement and the second one is a semantic resource of pictograms/emoticons.

[3] https://emojipedia.org/

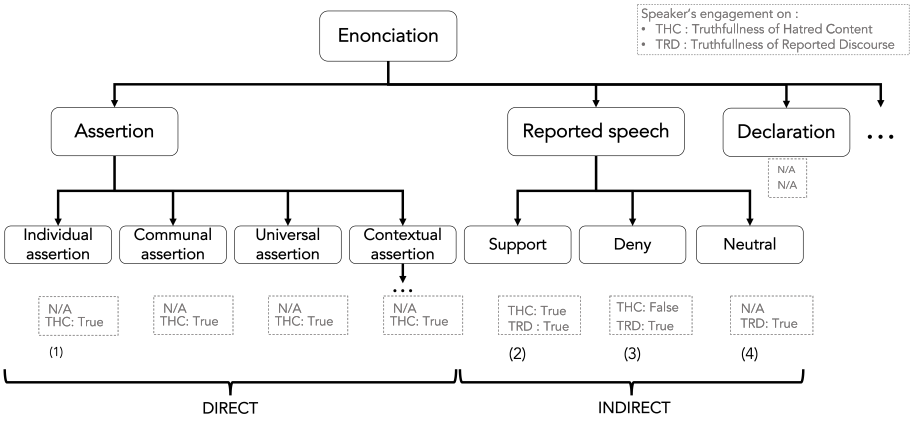

Fig. 1. Taxonomy of speaker's engagement

3.1 A taxonomy for labelling user engagement

As mentioned before (cf. examples 1 to 4), it is important to take into account to what extent the user supports or denies the truth of hateful content. In order to do so, we developed a taxonomy for user engagement. It was created by analyzing existing linguistics research on user engagement [10] and knowledge claims [4]. Fig. 1 presents a partial view of this taxonomy, that is relevant for experiments described in section 4. It highlights three main categories of enunciative acts: assertion, reporting, and declaration. Each category is described by using two types of boolean variables expressing the engagement of the speaker on two kinds of truthfulness: the truthfulness of a hatred content (noted THC) and the truthfulness of a reported discourse (noted TRD).

When a category is not concerned with one of those two variables, it is noticed with N/A. According to this taxonomy:

- (1) can be analysed as an Individual Assertion with $THC = True$;
- (2) is then a Reported Speech having $TRD = True$ and $THC = True$;
- (3) is a Reported Speech having $TRD = True$ and $THC = False$;
- (4) is a Reported Speech with $TRD = True$ and THC has the value N/A.

3.2 A semantic resource of pictograms

A pictogram is a symbol used to represent a variety of images: facial mimicry, (e.g. ,), physical postures (e.g. ,), food (e.g. ,), tools (e.g. ,), vehicles (e.g. ,), flags (e.g. ,), plants or animals (e.g. ,), punctuation marks (e.g. ,) or abstract concepts (e.g. ,), etc.

From the set of pictograms, emoticons can be distinguished as a specific subset According to [14], what makes a pictogram an emoticon is the fact that it becomes

a conventional index of an attitude, an emotion, of any element experienced by the speaking subject. Therefore, an emoticon is a pictogram imitating a facial expression (e.g. 😭 , 😺), or a pictogram giving additional information on what the speaker feels or experiences (e.g. ❤️ , 🔥). Moreover, [15] also identifies four types of emoticons according to the facial expressions they imitate:

- positive: raising corners of the mouth and/or squinting eyes (e.g. 😆 , 😊 , 😺);
- negative: with droopy and/or twisted mouth (e.g. 😓 , 😖 , 😐);
- surprised: with rounded mouth and/or eyes (e.g. 😲 , 😳 , 🤯 , 😵);
- address: when it imitates gestures intended to address an interlocutor (e.g. 😉 , 🤫).

It is also possible to highlight the ability of pictograms to convey user engagement with any textual content. That's what we did. In the end, a rich semantic resource was created by classifying a collection of 1 877 pictograms (among which 180 emoticons). Each pictogram/emoticon is described by three semantic attributes:

1. the type whose values can be: *pictogram*, *emoticon*;
2. the mimicry whose values can be: *absence, positive, negative, surprised, address*, a value, when different of absence, is represented with a rank from 1 to 4;
3. the user engagement whose values can be: $true$, $false$, the value is represented with a rank ranging from 1 to 2.

The protocol of giving ranks allows us to describe the semantics of pictograms even when they are ambiguous or with a heterogeneous use, especially with generational differences. For example, the item 😺 will be interpreted non-ambiguously as positive whereas the item ☠️ will be interpreted as negative but also as positive for the younger generation for the expression "dead funny".

Table 1. Examples of pictograms described with three semantic criteria.

	Type	**Mimicry**				**User Engagement**	
		Positive	**Negative**	**Surprised**	**Adress**	**True**	**False**
😜	emoticon	2			1	1	
❤️	emoticon	1				1	
😏	emoticon	1			2	1	1
🥑	pictogram						1
🇫🇷	pictogram					1	

Table 1 shows five examples of pictograms and their attributes. The first three cases of emoticons illustrate different types of facial expressions and user engagement: the item 😜 is an emoticon addressing someone in a positive way and fully engages the user; the item ❤️ is a non-ambiguous emoticon which fully engages the user in a positive

posture; the item 😏 is more ambiguous about user engagement with an ironic posture. Some pictograms, such as 🥑, don't engage a user. However, others like the last item, 🇫🇷, even if it doesn't refer to a facial mimic, is a good example of a pictogram that is used to engage a user through his nationality. The rank values are converted into percentages needed for machine learning algorithms. For instance, the obtained percentages for the item 😏 are therefore 66% ($\frac{2}{3}$) in terms of positive meaning, and 33% ($\frac{1}{3}$) in terms of address meaning. The next section illustrates how those resources are used for sexism annotation with online data.

4 Illustration on sexism annotation

Objective of experiments. The goal of this experimental phase is to evaluate how useful the annotation on user engagement is in order to improve the characterization of online hateful data. The application context is sexism classification in online data in French. More specifically, the paper addresses one of the limitations identified by [8], which gathers under the label sexist two types of tweets: a tweet where a user's sexist experience is reported, a tweet where a user writes sexist content directly addressed to another user(s). In the first case, the user is the victim, while in the second, the user is the aggressor. To overcome this limitation, we propose a method able to distinguish the sexist content and the user's engagement with respect to the truthfulness of this sexist content.

Data sets. Annotation was carried out by using a set of data created by the previous cited study dedicated to sexism detection in online data [8]. The initial data set comprises around 12 000 tweets gleaned online and manually annotated with two labels indicating whether a tweet is sexist or non sexist. 2.

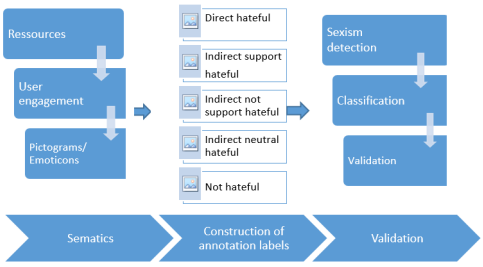

Fig. 2. General architecture for experiments

General methodology The experimental protocol involves a chain of methodological decisions illustrated in Fig. 2. This chain is structured by three main steps: a semantic

analysis and representation of the notion of speaker/user engagement in the form of a taxonomy, the determination of a set of useful labels for annotations from this linguistic taxonomy, and a validation phase of the entire protocol. The first step has been detailed in the previous section, while the last two ones are presented in this section. We begin by presenting our annotation process of a corpus of French tweets, then we detail first experiments conducted in order to valid our methodology.

4.1 Annotation of tweets

Annotation of tweets The purpose of the annotation is to refine the description of the content in a tweet by 1) characterizing it as hateful or not, 2) identifying whether the user states it directly or reports it, 3) and finally whether the user supports the content or not. In the end, each tweet is described with those three features and gets one of the five given labels shown with the leaves Fig. 3.

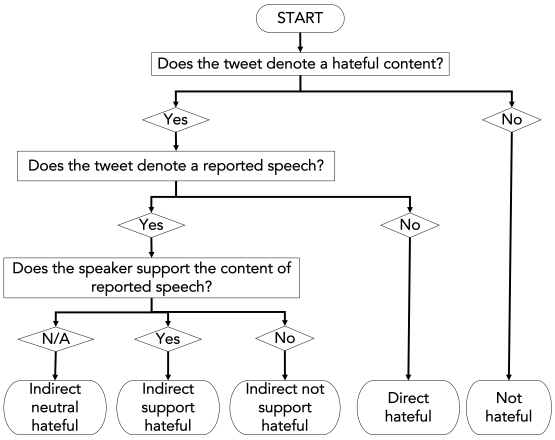

Fig. 3. Annotation procedure: ways for labels identification

Set of labels As said, five labels are considered:

- **Not hateful**, noted L_1: the tweet does not contain hateful content:
  example: *Les bleues font un très, très, très bon match !* 🔥 *#FRANOR @X*
  (The blue ones make a very, very, very good game! 🔥 #FRANOR @X);
- **Direct hateful**, noted L_2: the tweet contains hateful content directly said by the user:
  example: *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);

Table 2. Example of descriptors and their values to describe a tweet

	Labels	**Hateful content**	**Emoticons/Pictograms**						**Enonciation types**			
			Type	**Mimicry**				**User engagement**				
				Positive	**Negative**	**Surprised**	**Address**		**Assertions**	**Reported speech**	**Declaration**	**...**
Values	L_1		Emoticon	*N/A*	*N/A*	*N/A*	*N/A*					
	L_2	N/A		*1*	*1*	*1*	*1*	N/A	N/A	N/A	N/A	
	L_3	True		*2*	*2*	*2*	*2*	True	True	True	True	...
	L_4	False	Pictogram	*3*	*3*	*3*	*3*	False	False	False	False	
	L_5			*4*	*4*	*4*	*4*					

- **Indirect not support hateful**, noted L_3: the tweet contains hateful content reported by a user who does not support it:
  example: *"Hé mademoiselle, est-ce que je peux me permettre de vous dire que t'es bonne ?" Agen – place Jasmin #payetashnek*
  ("Hey miss, can I tell you that you're hot?" Agen – Place Jasmin #payetashnek);
- **Indirect support hateful**, noted L_4: the user reports hateful content in the tweet and supports it:
  example: *Ouais. T'as raison. Le viol après 15 ans ça n'existe pas. Et si on est violées, on l'a cherché.*
  (Yeah. You're right. There's no such thing as rape after 15. And if we are raped, we asked for it.);
- **Indirect neutral hateful**, noted L_5: the user remains neutral with respect to reported content:
  example: *"C'est une grosse tarlouze": une association saisit le CSA après le dérapage de JoeyStarr à ONPC http://url.com"*
  ("He's a big faggot": an association refers to the CSA after JoeyStarr's slip at ONPC http://url.com).

The annotator does not assign directly these labels to tweets but annotates a set of linguistic descriptors for each tweet, based on which the labels are then obtained by following the procedure for deciding the labels Fig. 3. These descriptors are precisely detailed in the following section.

Description of our annotated tweets data set Annotation of tweets is carried out by considering three types of descriptors indicating: 1) whether the tweet's content is hateful or not, i.e. sexist[4] or not in this use case, 2) pictograms/emoticons with the three semantic criteria exposed section 3.2, and 3) types of enunciations according to the taxonomy Fig. 1. These three main categories of descriptors are shown in Table 2 (categories are underlined) which details each descriptor and its possible values. The entity to be annotated is the entire tweet. By following this annotation procedure, 300 tweets randomly selected from the initial collection of 12 000 tweets were annotated. Inter-annotator agreement has not been calculated since, to our knowledge, no measure has

[4] The definition of sexism used it's the one given by the European Council.

Table 3. Results for all classifiers and the camemBERT model

Classifiers	Precision	Rappel	F-measure
Naive Bayes	0.62	0.63	0.62
Random Forest	0.65	0.70	**0.67**
Ridge	0.60	0.63	0.60
SVM	0.49	0.68	0.55

Model	Precision	Rappel	F-measure
CamemBERT	0.35	0.59	**0.45**

been proposed to measure agreement between two annotations in terms of proportions (like in Table 1).

4.2 Experiments and first results

In order to validate our methodology for describing hate content (sexist content in our use case), we compared it to a generic approach. The experiments demonstrate the consistency of our annotation protocol. Although this protocol is complex and time-consuming, we show that a very small annotated dataset produces satisfactory results. Therefore, this section presents the first experiments to compare the results of supervised classifiers using user engagement with results provided by a BERT-type model fine-tuned on the same data. We make this comparison to test our hypothesis that few expert features give better results than a fine-tuned generic language model.

The goal of the multiclass classification task is to distinguish the five categories shown with the leaves Fig. 3: *Not hateful, Direct hateful, Indirect not support hateful, Indirect support hateful,* and *Indirect neutral hateful*.

Experimental protocol Each tweet is considered as a document for which a label is predicted. All features are boolean: 1 is assigned for their presence, 0 for their absence. To annotate descriptors as present when their annotations are proportions, the proportion must be greater than 50%. To evaluate the relevance of these features, several supervised classifiers (Naive Bayes, SVM, Random Forest, and Ridge) were compared to a pre-trained model (camemBERT-base). To train these different classifiers, we separated the set of 300 manually annotated tweets into 3 subsets: the test set (10%), the development set (10%), and the training set (80%). Traditional evaluation metrics in automatic classification were used: precision, recall, and f-measure. All experiments were implemented in Python with scikit-learn[5] for the classifiers and ktrain[6] for camemBERT.

Analysis of results The results for camemBERT presented in this section are those obtained with the base version[7], a learning rate of 0.02, and 8 epochs. Table 3 shows the results of each classifierGiven the limited size of corpora used for experiments and the

[5] Supervised learning models from scikit-learn
[6] Ktrain's GitHub
[7] camemBERT-base version on Hugging Face site

difficulty of the task, those results are still acceptable. Moreover, they confirm that taking into account user engagement improves the overall performance: Random Forest has a better f-measure (0.67) than camemBERT (0.45). These results confirm our hypothesis: with a small training dataset, few expert features lead to better results than a fine-tuned generic language model trained on large corpora. In addition, experiments show that our annotation procedure facilitates the corpus inspection and illustrates how features of user engagement and emoticons can help to identify otherwise inaccessible examples of hate speech like those in [8]. Table 4 shows five examples of predictions obtained with our Random Forest model.

Table 4. Examples of predictions with the Random Forest model

ID	fr/en	example	predicted label
1	fr	Tom Villa sur Aurore Bergé : "quand y a buffet à volonté, elle goute à tous les plats" 😭 #SLT	Not hateful
	en	*Tom Villa on Aurore Bergé:* *"When there's an all-you-can-eat buffet,* *she tastes every dish"* 😭 *#SLT*	
2	fr	Les filles encore pucelle vous etes comme des cartes bleues pour moi	Direct hateful
	en	*The girls still virgin* *you are like blue cards for me*	
3	fr	Heyy mignonne, je te trouve même carrément bonne. J'allais te demander une cigarette, mais je préfère que tu me suces la bite. #payetashnek	Indirect not support hateful
	en	*Heyy cutie,* *I even think you're downright hot.* *I was going to ask you for a cigarette,* *but I'd rather you sucked my dick. #payetashnek*	
4	fr	Une femme qui prie est plus dangereuse qu'une femme qui se venge	Indirect support hateful
	en	*A woman who prays is more dangerous* *than a woman who takes revenge*	
5	fr	Et les commentateurs PSG Toulouse : j'ai jamais vu ça un tir de femme enceinte.	Indirect neutral hateful
	en	*And the PSG Toulouse commentators:* *I've never seen a pregnant woman shoot like that.*	

The model proposed by [8] would have labeled examples 2 and 3 with the same "sexist" label, whereas our model distinguishes between them according to the speaker's engagement with the truth of the sexist content. At last, because sexism is a sensitive subject, the performance of detection models should be evaluated with the use of qualitative methods, and using human-in-the-loop procedures.

There are several limitations affecting the overall approach described in this study. First, the resources were created from scratch and thus further experiments are required for their empirical validation. Secondly, another annotation campaign could be carried out to test our annotation schema and models. Finally, we could refine types of reported speech.

5 Conclusion and perspectives

The main contribution of the paper is the development of linguistic resources for French. This work also provides a basis for future research on the usage of pictograms/emoticons and user engagement for social data analysis. More specifically, the pictogram resource augments from a semantic point of view the classical description of pictograms by adding new attributes useful for the automatic interpretation of online hate discourses. Further work could be done to study pictograms/emoticons by looking at their syntactic integration to refine their role in representing user engagement. Another direction for future work is the setup of classification experiments, in order to investigate the impact of user engagement and pictograms when used as features for automatic online hate detection. The annotated dataset is available for research purposes upon request.

References

1. Al-Azani, S., El-Alfy, E.S.M.: Early and late fusion of emojis and text to enhance opinion mining. IEEE Access **9**, 121031–121045 (2021)
2. Al-Halah, Z., Aitken, A., Shi, W., Caballero, J.: Smile, be happy:) emoji embedding for visual sentiment analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
3. Bassignana, E., Basile, V., Patti, V., et al.: Hurtlex: A multilingual lexicon of words to hurt. In: CEUR Workshop Proceedings. vol. 2253, pp. 1–6. CEUR-WS (2018)
4. Battistelli, D., Amardeilh, F.: Knowledge claims in scientific literature, uncertainty and semantic annotation: A case study in the biological domain. In: Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM 2009) (2009)
5. Battistelli, D., Bruneau, C., Dragos, V.: Building a formal model for hate detection in french corpora. Procedia Computer Science **176**, 2358–2365 (2020)
6. Bergqvist, H., Knuchel, D.: Explorations of engagement: Introduction. Open linguistics **5**(1), 650–665 (2019)
7. Bick, E.: Annotating emoticons and emojis in a german-danish social media corpus for hate speech research. RASK–International Journal of Language and Communication **52**, 1–20 (2020)
8. Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., Coulomb-Gully, M.: An annotated corpus for sexism detection in french tweets. In: Proceedings of the 12th language resources and evaluation conference. pp. 1397–1403 (2020)
9. Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., Coulomb-Gully, M.: He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4055–4066 (2020)
10. Desclés, J.P.: Prise en charge, engagement et désengagement. Langue française (2), 29–53 (2009)

11. Evans, V.: The emoji code: How smiley faces, love hearts and thumbs up are changing the way we communicate. Michael O'Mara Books (2017)
12. Fernández-Gavilanes, M., Costa-Montenegro, E., García-Méndez, S., González-Castaño, F.J., Juncal-Martínez, J.: Evaluation of online emoji description resources for sentiment analysis purposes. Expert Systems with Applications **184**, 115279 (2021)
13. Godard, R., Holtzman, S.: The multidimensional lexicon of emojis: A new tool to assess the emotional content of emojis. Frontiers in Psychology **13** (2022)
14. Halté, P.: Les marques modales dans les chats: étude sémiotique et pragmatique des interjections et des émoticônes dans un corpus de conversations synchrones en ligne. Ph.D. thesis, University of Luxembourg, Luxembourg, Luxembourg (2013)
15. Halté, P.: Les émoticônes: de la signification des affects aux stratégies conversationnelles. Communiquer (28) (2020)
16. Hemati, H.H., Alavian, S.H., Beigy, H., Sameti, H.: Sutnlp at semeval-2023 task 10: Rlat-transformer for explainable online sexism detection. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 347–356 (2023)
17. Kirk, H.R., Vidgen, B., Röttger, P., Thrush, T., Hale, S.A.: Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv preprint arXiv:2108.05921 (2021)
18. Kralj Novak, P., Smailović, J., Sluban, B., Mozetič, I.: Emoji sentiment ranking 1.0 (2015)
19. Liang, W., Liang, K.H., Yu, Z.: Herald: an annotation efficient method to detect user disengagement in social conversations. arXiv preprint arXiv:2106.00162 (2021)
20. Malecki, W., Kowal, M., Dobrowolska, M., Sorokowski, P.: Defining online hating and online haters. Frontiers in Psychology **12**, 744614 (2021)
21. Robertson, A., Liza, F.F., Nguyen, D., McGillivray, B., Hale, S.A.: Semantic journeys: quantifying change in emoji meaning from 2012-2018. arXiv preprint arXiv:2105.00846 (2021)
22. Rong, S., Wang, W., Mannan, U.A., de Almeida, E.S., Zhou, S., Ahmed, I.: An empirical study of emoji use in software development communication. Information and Software Technology **148**, 106912 (2022)
23. Shoeb, A.A.M., De Melo, G.: Emotag1200: Understanding the association between emojis and emotions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8957–8967 (2020)