A Corpus-Based Comparison of two Approaches for Emotion Annotation in French Texts

Valentina Dragos, Delphine Battistelli
ONERA-The French Aerospace Lab, Palaiseau, France
Paris-Nanterre University, Nanterre, France
valentina.dragos@onera.fr, delphine.battistelli@parisnanterre.fr

Abstract

Emotion annotation in texts remains a challenging task in the field of Natural Language Processing (NLP), as, unlike voice or images, texts might not only contain peculiar cues to express emotions. Methods for emotion annotation are based on lexicons or on machine learning techniques which are based on the use of manually annotated corpora. This paper aims to explore if and how the combination of these two types of methods might be useful for the annotation of emotions in texts. Four data sets are used for comparison of the two approaches, and then to investigate to what extent the results are distinct or complementary on three aspects: (i) identification of emotional sentences; (ii) identification of emotion categories; (iii) identification of one specific mode of expression of emotions called "behavioral emotions" (e.g. shout, cry). Findings show that not all emotions are equally easy to annotate, and, most specifically, the learning-based approach tends to over detect *Admiration*.

Keywords: detection of emotions, comparison of approaches, French corpus

1. Introduction

Detecting human emotions from the diversity of social networks' data is a key part of gaining deeper insight into the nature of online content. Emotion detection in textual data is a challenging and tedious task because emotions can be expressed in a variety of subtle and complex ways (Al Maruf et al., 2024), (Gladys and Vetriselvi, 2023). Some emotions are explicitly indicated by specific words such as *sadness* or *joy*, while others are implicitly derived from events (*birthday*) or behaviors (*he cried)*. Consider the following examples:

- Je dénonce avec fermeté la colonisation de notre territoire. (I firmly denounce the colonization of our territory) (1).
- Nous nous battons pour bien plus que cela. (We are fighting for much more that that) (2).
- Ebola encore plus proche de français ! Merci qui ? Merci les migrants ! (Ebola ever closer to French! Who do you have to thank for this? Thank you, migrants!) (3)
- CA SUFFIT! (Enough!) (4)

The first example (1) shows a direct and explicit *Anger* emotion expressed by the speaker. The second sentence (2) also illustrates *Anger*, but the emotion is not explicitly expressed, but rather suggested by the willingness to engage in the fight. Moreover, the third example (3) shows a sarcastic sentence conveying the same emotion. *Anger* is also present in the last sentence (4), and in this case the emotion is illustrated by a characteristic of the utterance and the exclamation mark.

Detection of emotions in text has attracted numerous research efforts in the field of Natural Language Processing (NLP). Computational approaches for emotion detection are based on various paradigms, ranging from sentence-level analysis with semantic rules (Seal et al., 2020) to building sophisticated learning models (Chavan et al., 2023). Deep learning techniques (Chutia and Baruah, 2024) and Large Language Models (LLMs) have been used to perform various tasks related to emotion analysis (Zou and Markov, 2024). Despite the prevalence of many techniques, the complexity of emotion expressions in texts has not been fully addressed.

This paper describes the comparison of two parsers developed for emotion annotation in French texts. As part of this study, we examine to what extent the tools provide complementary or divergent results on various data sets collected on the Internet.The experimental protocol defines several emotion annotation tasks, such as detecting emotional sentences and categories of emotions, and then compares the results of both tools on four data sets. The rest of the paper is organized as follows: section 2 discusses related work. Approaches for emotion annotation are discussed in section 3 and corpora used for the comparison are presented in Section 4. The experimental setup and comparison method are introduced in section 5. Section 6 concludes the paper and presents directions for future research.

2. Related approaches

This section gives an overview of the different methods used to detect emotions in texts. It also dis-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 38–46
16 May 2026.

cusses research work focusing on annotation disagreement.

2.1. Machine learning and lexicon-based approaches

Emotion annotation is a multi-classification problem (Kusal et al., 2023), and research studies have proposed a variety of lexicon-based techniques and machine learning methods that are commonly used for emotion classification (Chutia and Baruah, 2024).

Lexicon-based techniques are grounded in emotion dictionaries that are built to represent prior knowledge of how emotions are expressed. The literature presents several emotion lexicons, such as SentiWordNet (Baccianella et al., 2010), WordNet-Affect (Poria et al., 2012) and FEEL (Abdaoui et al., 2017), a lexicon built to detect emotion in French. Those techniques rely on recognizing lexicon keywords in a given text and assigning an emotion category based on those keywords and their frequency. Although lexicon-based techniques perform well in detecting explicit emotions, they fail to detect emotions that are not explicitly mentioned in text but rather derived from other clues, such as descriptions of behaviors.

Machine learning models that are trained on sample data sets and make predictions on unseen data afterward are more suitable to detect implicit emotions. Support Vector Machines (SVM) (Tzacheva and Chatterjee, 2024), XGBoost (Bandara et al., 2024) and Naïve Bayes (NB) (Kumar and Kumar, 2022) are some prominent learning techniques adopted to perform emotion detection in texts. The accuracies of supervised learning methods are in the range of 65–80% on benchmark data sets (Al Maruf et al., 2024), although the scarcity of large emotion-oriented labeled data sets is a major limitation of their use. Methods based on transfer learning address these limitations by leveraging the ability of pre-trained language models to detect emotions (Feng and Chaspari, 2020). These methods can be fine-tuned with smaller labeled data sets, and several studies have used adaptations of BERT (Acheampong et al., 2021) or GPT (Lian et al., 2024) to recognize emotions, with detection accuracy reported in the range 75 to 99%. In spite of their high accuracy values, these methods require constrained domain adaptation and demonstrate limited capacities to detect emotions from unforeseen expressions.

2.2. Hybrid approaches

From a different perspective, hybrid methods have also been developed recently, combining heuristics with learning techniques in an effort to improve the accuracy of emotion identification and refine the categories of emotions detected in texts. Among them, (Li et al., 2021) describes a dual graph convolutional network (DualGCN) model that takes into account the complementarity of syntax structures and semantic correlations simultaneously to capture general semantics. The model detects words that are semantically related, and the evaluation of emotion classification tasks shows levels of accuracy higher than 80%.

A meta-learning approach where predictions from transformers are combined with lexicon features for emotion detection is proposed in (De Bruyne et al., 2021) and the authors demonstrate an improvement on emotion detection compared to a basic BERT model. In (Meškelė and Frasincar, 2020), the authors describe an approach for emotion detection utilizing jointly a lexicalized domain ontology to model field-specific knowledge and BERT for word embeddings. More specifically, the joint utilization provided a mechanism for measuring the influence of words detected in a given sentence on emotion expression and the integration of knowledge improved the overall accuracy of emotion detection. In (Aljedaani et al., 2022), the authors describe a novel hybrid method for emotion detection that combines machine learning and deep learning techniques with a lexicon-based model called TextBlob (Loria et al., 2018). The lexicon is used for data annotation and training and fine-tuning of learning models. The authors investigated the impact of TextBlob on the performance of classification models and showed that the accuracy values of both machine learning and deep learning models are considerably improved when trained using TextBlob annotations, achieving the highest F1 score of 0.96.

Having remarkable capabilities in various natural language understanding and generation tasks, Large Language Models (LLMs) have also been used as annotators (Sun et al., 2023), (Tan et al., 2024) and they have demonstrated good potential for sample annotation (Gilardi et al., 2023) and excellent abilities to annotate data with a limited and well-defined set of labels (Ding et al., 2022).

2.3. Analysis of annotation disagreement

Although detection methods have specific detection capabilities, the annotation schemes they rely upon have an impact on the quality of results, and sometimes induce discordant emotion annotations (Beck, 2023). Recently, several studies started to investigate the factors that potentially cause disagreements when performing emotion annotation from texts. Among them, (Labat et al., 2022) designed an experiment intended to scrutinize the variation in the expression and annotation of emotions retrieved in human-computer conversations. The authors show that using annotation schema

having fine-grained emotion categories leads to low values of annotator agreement. Following a similar approach, (Barz et al., 2025) analyzed the factors of disagreement when annotating emotions in communications about the environment and climate change. Their findings show that two prominent factors of disagreement are the lexical biases of annotators and the association of an emotion to several domain-specific topics.

The work presented in this paper contributes to this last research direction and the goal of our study is to understand, thanks to a qualitative and quantitative comparison, the disagreements of annotations provided by two automatic approaches on four data sets collected on the Internet. In comparison to related works investigating disagreements on emotion annotation, this study is novel in using automated emotion annotation and considering the detection of emotion categories.

3. Automatic approaches for emotion annotation

This section presents EMOTYC and EMOTIONS, and indicates the emotion detection paradigm, the underlying emotion annotation schema and annotation examples.

3.1. EMOTYC: a learning-based parser for emotion annotation

EMOTYC (Étienne et al., 2024) is an emotion detection parser based on a classifier previously trained to detect emotions in texts. EMOTYC was developed using the CamemBERT model (Martin et al., 2019). The annotation schema used by EMOTYC is based on linguistic and psycholinguistic approaches that describe emotions by considering their type, mode, and category, as shown in Fig. 1.

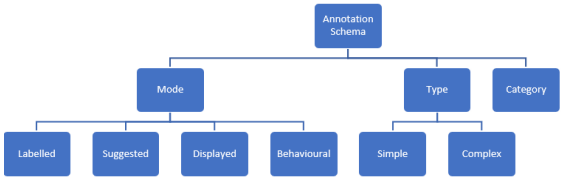

Figure 1: EMOTYC anotation schema

EMOTYC performs the following tasks for emotion analysis: (i) detection of sentences containing emotions; (ii) identification of the emotion mode; (iii) recognition of basic or complex emotions; and (iv) identification of the category of emotion.

The detection approach used by EMOTYC categorizes linguistic markers into four emotion expression modes:

- labeled : explicit emotion expressed by lexical, syntactical or typographical clues clearly showing that the utterance conveys an emotion;
- displayed: description of situations that are usually associated with emotions;
- behavioral : description of behaviors that are usually associated with emotions;
- suggested : emotions that are indirectly indicated or implied, often from the context.

EMOTYC identifies the *Type* (*Basic* or *Complex*) and distinguishes eleven categories of emotions, comprising the 6 basic emotions introduced by (Ekman, 1992) (*Anger, Fear, Happiness, Surprise, Disgust* and *Sadness*) and 4 complex emotions taken from (Blanc and Quenette, 2017) and (Davidson, 2006) (*Guilt, Embarrassment, Pride* and *Jealousy*). A fifth complex category (*Admiration*) was also added in order to better balance basic and complex emotions. The category *Autre* (*Other*) is used to capture emotions that are not related to any of the previous categories. Fig. 2 shows an example of annotation provided by EMOTYC, indicating an emotional sentence, the type (basic) and the category of emotion (Anger).

Sentence : il ne profite qu'aux multinationales qui importent de la main d'œuvre bon marche et délocalisent à l'intérieur comme à l'extérieur du pays. (It only benefits multinationals that import cheap labor and relocate it both within and outside the country.)
Emotional sentence: YES
Basic emotion: YES
Emotion category: Anger

Figure 2: Example of EMOTYC annotation

Although EMOTYC detects emotional sentences, the tool is not able to identify the tokens triggering the emotions within the sentence.

3.2. EMOTIONS: a lexicon-based parser for emotion annotation

EMOTIONS (Etienne, 2023) is an emotion annotation parser using an augmented version of the EMOTAIX emotion lexicon (Piolat and Bannour, 2009). The lexicon serves as a basis for recognizing terms from different emotional categories.

EMOTAIX models 24 categories of emotions, including the 11 categories of EMOTYC, 12 additional categories (*love, boldness, resentment*, etc.) and a behavior category that indicates emotional behavior, without a specific emotion category, such as *cry*, that can refer to both *sadness* or *happiness*.

The annotation schema used by EMOTIONS is based on the EMOTAIX lexicon, see Fig. 3.

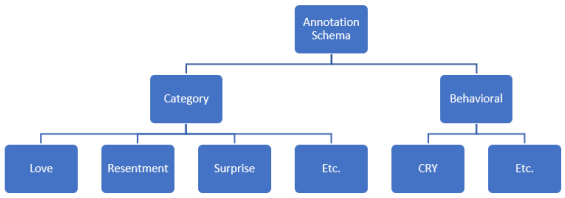

Figure 3: EMOTIONS annotation schema

In order to detect emotions, EMOTIONS performs tokenization and lemmatization of the sentences. Whenever a sentence contains a lexicon lemma, EMOTIONS identifies the emotion associated to this lemma in the EMOTAIX lexicon and thus detects the emotions and its trigger in the initial text. By following this lexicon-based approach, EMOTIONS performs the following tasks for emotion analysis: (i) detection of sentences containing emotions; (ii) identification of behavioral emotions; (iii) identification of the category of an emotion, and (iv) detection of tokens triggering emotions.

Fig. 4 shows an example of an EMOTIONS annotation.

Sentence : Le contexte social, économique et politique actuel, bien plus oppressant dans nos villes que dans nos campagnes, laisse clairement percevoir toute la fragilité du monde moderne. (The current social, economic and political context, much more oppressive in our cities that in the countryside, clearly reveals the fragility of the modern world.)
Tokens having emotions: oppressant (oppressive)
Category of emotions: deplaisir (discontent)

Figure 4: Example of EMOTIONS annotation

4. Overview of corpora

Data sets used for this study have been collected online and have different topics and volumes.

Corpus C1 (right-wing extremism). This data set was created by a project investigating the nature of extremism online [1]. The corpus is composed of Tweets and messages collected on social platforms and online 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 the sentenced conveying extremist attitudes have been identified and the corpus was annotated with *extremist* and *nonextremist* labels. The data set comprises 638 extremist sentences and 792 neutral sentences.

Sentences (E1) and (E2) show examples of sentences that convey extremist attitudes.

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 C2 (hateful content). This data set 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:

E3: *Quand il s'agit d'entretenir les etrangers 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.)

E4: *J'espère que tu vas avoir un cancer et mourir!* (I hope you will get cancer and die!)

Corpus C3 (TextToKids). This data set comprises 2506 sentences randomly extracted from the corpus used to train the EMOTYC parser. The corpus is composed of articles from the childrens' newspaper Le P'tit Libé [2]. The following examples illustrate emotional sentences from this data set.

E5: *Pourquoi je les aime autant ? Ils me donnent l'impression de jouer toute la journée.* (Why do I love them so much? They make me feel like I'm playing all day.)

E6: *On est très loin de voir les profs être remplacés par des androides !* (We are very far from seeing teachers being replaced by androids!)

Corpus C4 (Reddit) comprises 20.000 posts collected on the Reddit platform. The corpus was created by the Online European Hate Lab [3], using specific words and terms to collect toxic content in French subreddits. For the purposes of this study 2,006 sentences have been randomly selected. Two toxic sentences are shown below:

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

[2] https://ptitlibe.liberation.fr/

[3] https://europeanonlinehatelab.com/

E7: *Les post-fascistes sont un symptôme, pas l'origine du problème. (Post-fascists are a symptom, not the origin of the problem.)*

E8: *Il est meilleur mathématicien que politique. (He is a better mathematician than politician.)*

Remarks: All data sets contain data that was collected online in the framework of different projects and by different research teams.

Corpus	Nature	Balanced
C1	Extremist vs. NonExtremist	NON
C2	Hateful vs.NonHateful	YES
C3	No classes	-
C4	No classes	-

Table 1: Characteristics of data sets

C1, C2 corpora have contrasted classes, highlighting extremist vs. non-extremist content and hateful vs. non-hateful content, respectively. C3 and C4 data set are homogeneous, non-contrasted collections of texts, see tab. 1. Those data sets have been selected for their potential to convey emotions, and also for their differences in nature, size and structure.

5. Comparison of EMOTYC and EMOTIONS

This section assesses the potential for EMOTYC and EMOTIONS to agree on emotion annotation and highlights the differences between their results.

5.1. Experimental design

The experimental protocol adopted to compare EMOTYC and EMOTIONS consists of defining three tasks to be performed by the two parsers and then analyzing the set of emotion annotations they provided. The tasks are intended to:

- Detecting sentences conveying emotions (T1);
- Detecting behavioral emotions (T2);
- Detecting categories of emotions (T3).

As some emotion categories detected by EMOTYC and EMOTIONS do not overlap, a mapping was established in order to define their correspondences, as shown in tab. 2.

For all tasks of the experimental protocol, annotations provided by EMOTYC and EMOTIONS are compared with a focus on their disagreements. Fig. 5 presents an overview of the general architecture of the experiment.

If we consider a1 and a2 as the sets of anchors annotated by EMOTYC (as the A1 annotator) and EMOTIONS (as the A2 annotator), then the disagreement of annotators can be captured by the

EMOTYC	EMOTIONS
Colère (Anger)	Colère (Anger), Déplaisir (Displeasure)
Autre (Other)	Amour (Love), Apaisement (Alleviation), Audace (Temerity), Désir (Desire), Empathie (Empathy), Impassibilité (Impassivity), Humanité (Humanity), Mépris (Contempt), Ressentiment (Resentment)
Fierté (Pride)	Fierté (Pride), Orgueil (Vanity)

Table 2: Mapping of categories detected by EMOTYC and EMOTIONS

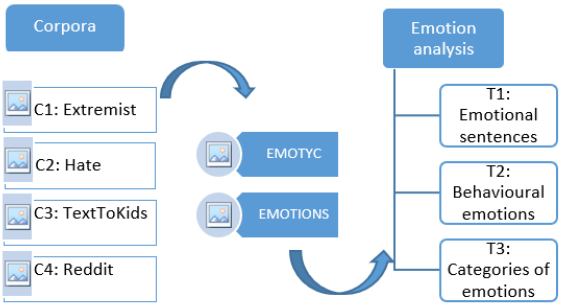

Figure 5: General architecture for comparison

Recall measure, where the Recall of A2 with respect to A1 is given by:

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

Recall values range from 0, when the intersection of the two sets of anchors is void, to 1, when the two ensembles are identical.

The results of experiments are described hereafter.

5.2. Experiments and analysis of results

Experiments were carried out using the data sets presented in section 4.

Corpus of right-wing extremism (C1):(tab. 3 and 4).

Label	Recall(A2||A1)	Recall(A1||A2)
Emotional	49.72	46.63
Non Emotional	87.52	89.88
Behavior	60.00	9.37

Table 3: Corpus C1: Results of tasks T1 and T2

Task T1: the two parsers identify a similar number of emotional sentences, and the results are

identical for half of them, as the values of Recall(A1||A2) and Recall(A2||A1) are very close for both the detection of emotional and non-emotional sentences.

Task T2: compared to EMOTYC, EMOTIONS identifies a limited number of behavioral emotions. This outcome is aligned with previous findings described in (Etienne, 2023), where the authors also point out that the lexicon used by EMOTIONS has a limited number of terms related to behavioral emotions.

Label	Recall(A2||A1)	Recall (A1||A2)
Anger	72.72	12.69
Joy	80.00	28.57
Admiration	0.00	0.00
Sadness	40.00	25.00
Shame	100	50.00
Fear	55.17	55.17
Pride	40.00	100
Surprise	50.00	25.00
Other	1.00	100

Table 4: Corpus C1: Results of task T3

Task 3: Comparison of results for the identification of emotions' categories shows that the two processors agree well when it comes to the detection of *Fear* category. However, there is no agreement between EMOTYC and EMOTIONS on the detection of *Admiration*. In-depth analysis of results indicates that the category is correctly detected by EMOTIONS. For example, the sentence *Gloire à ceux qui luttent à l'Est ! (Glory to those who fight in the East!)* is correctly annotated as Admiration by EMOTIONS, although the processor is wrong when assigning the same annotation to the sentence *Nous ne sommes pas un parti politique et nous estimons que ce sont aux gens qui nous gouvernent d'assurer notre sécurité, travail pour lequel ils sont grassement payés. (We are not a political party and we believe that it is up to the people who govern us to ensure that we are safe, a job for which they are well paid.)*. For the *Admiration* category, EMOTYC performs rather poorly.

Corpus of hate content (C2): (tab. 5 and 6).

Task T1: Values of the Recall indicate a strong disagreement between EMOTYC and EMOTIONS on the detection of emotional and non-emotional sentences. This can be explained by the nature of the corpus, this data set containing many implicit emotions.

Task T2: as for the previous corpus, EMOTIONS detects only a limited number of behavioral emotions, and the agreement with EMOTYC is low.

Task T3: The disagreement levels are significant also for the detection of emotion categories. More specifically, the disagreement on the *Anger* category is strong (for example, 186 sentences conveying *Anger* are detected by EMOTYC while EMOTIONS detects only 7 ones). This can be explained by the fact that *Anger* is mostly associated with the displayed and suggested modes that are not detected by EMOTIONS. In addition, many sentences annotated as *Anger* by EMOTYC contain swearing and insulting words. As those words are not included in the lexicon used by EMOTIONS, the sentences are not annotated by the parser, which also explains the disparity of *Anger* labels detected by the two parsers. More specifically, *Other* is the emotion category most detected by EMOTIONS within this corpus, indicating that lexicon-based tool encounters difficulties in detecting accurately the categories of emotions in this data set.

Label	Recall(A2||A1)	Recall (A1||A2)
Emotion	77.43	33.15
NonEmotion	59.81	91.52
Behavior	20.00	5.55

Table 5: Corpus C2: Results of tasks T1 and T2

Label	Recall(A2||A1)	Recall (A1||A2)
Anger	100	3.76
Joy	50.00	34.61
Admiration	28.57	1.76
Sadness	37.5	6.97
Shame	66.66	33.33
Fear	78.57	50.00
Pride	100	3.33
Surprise	100	11.62
Other	0.90	25.00

Table 6: Corpus C2: Results of task T3

Corpus TextToKids (C3): (tab. 7 and 8).

Task T1: For this data set, EMOTYC and EMOTIONS demonstrate good agreement on the detection of non-emotional sentence. In contrast, the tools show divergent results on the detection of emotional sentences.

Label	Recall(A2||A1)	Recall (A1||A2)
Emotional	55.67	36.94
Non Emotional	86.75	94.85
Behavior	53.84	5.88

Table 7: Corpus C3: Results of tasks T1 and T2

Task T2: For this data set, EMOTYC annotates a large number of behavioral emotions, as the corpus contains numerous behavior-related expressions, such as *criticizing*, *having fun* and *smiling*. Those expressions are specific to journal articles written for kids, that often contain descriptions and explanations. The behavioral emotions are still difficult to

Label	Recall(A2||A1)	Recall (A1||A2)
Anger	73.68	13.72
Joy	87.87	52.00
Admiration	28.00	55.76
Sadness	58.82	13.51
Shame	33.33	6.25
Fear	92.30	41.73
Pride	66.66	25.00
Surprise	100	12.82
Other	0.48	9.09

Table 8: Corpus C3: Results of tasks T3

identify for EMOTIONS on this data set as well, and thus the tools have a strong disagreement when performing the task T2.

Task T3: The parsers achieve high agreement regarding the detection of *Joy*, as most of the instances detected by EMOTIONS are also detected by EMOTYC. Similarly, the agreement is also high for the identification of *Fear*. As the data set contains stories for kids, *Fear* and *Joy* are the most prevalent emotions in the texts and they are expressed in a variety of direct and implicit ways, which can explain the agreement of tools on their detection.

Corpus C4 (Reddit) (tab. 9 and 10).

Task T1: For this data set, the tools demonstrate good agreement on detecting non-emotional sentences, although they have a low agreement on the detection of emotional sentences.

Label	Recall(A2||A1)	Recall (A1||A2)
Emotional	29.74	20.02
Non Emotional	92.98	91.69
Behavior	33.33	4.10

Table 9: Corpus C4: Results of tasks T1 and T2

Task T2: the results of this task are perfectly aligned with the results obtained for the previous corpora: the number of behavioral emotions detected by EMOTIONS is limited, and thus the agreement with EMOTYC is low.

Task T3: The two parsers demonstrate good agreement on the detection of two categories: *Joy* and *Fear*. They strongly disagree when it comes to detecting *Sadness* and *Pride*. Also, there is a significant asymmetry in the detection of *Anger* (for example as EMOTYC detects 204 sentences conveying *Anger* while EMOTIONS identifies only 30 ones). *Anger* is highly represented in this corpus, and the analysis shows that sentences annotated by EMOTYC with this label show reactions to previous events, such as *It is shocking!*, or *How dare you ?*.

Other is the most detected emotion category for

Label	Recall(A2||A1)	Recall (A1||A2)
Anger	50.00	7.35
Joy	75.71	53.00
Admiration	6.66	1.48
Disgust	0.00	0.00
Sadness	16.66	16.47
Shame	57.14	33.33
Fear	55.55	51.47
Pride	7.14	12.5
Surprise	61.90	22.03
Other	0.00	0.00

Table 10: Corpus C4: Results of task T3

EMOTIONS, and the agreement of tools on the detection of this category is low.

Overall, the experiments show that the nature of data sets has an impact on the quality of emotion annotation, as EMOTYC performs better in detecting implicit emotions, while EMOTIONS is accurate in detecting explicit ones. Moreover, the results also show that some categories of emotions, such as the basic ones *Fear* and *Joy*, are easier to detect for both parsers, while complex categories are both difficult to identify.

6. Conclusion and perspectives

This paper compares EMOTYC and EMOTIONS, two parsers performing emotion annotation from texts by using distinct approaches and annotation schema. The results show that data sets have an impact on the accuracy of the techniques used by the tools, and thus affect their agreement. In general, implicit and, most specifically, behavioral emotions are difficult to detect for EMOTIONS, while EMOTYC performs poorly in detecting *Admiration*. In addition, for all data sets, the tools have a good agreement in detecting the non-emotional sentences, while the agreement is lower when it comes to the detection of emotional ones. Similarly, there is good agreement on detecting basic emotions, while for complex emotions, the results show consistent disagreement.

Future research aims at improving the annotation of emotions from texts by tuning the learning model used by EMOTYC and enriching the lexicon used by EMOTIONS. Expanding the data sets to include more types of corpora and studying the impact of annotation schema on the results provided by the tools and their agreement will also be investigated in the future.

Data availability statement

data sets presented in this article are available for research purposes upon request.

Ethical considerations and limitations

For ethical concerns, the results are presented in a way that avoids reidentification of any URIs and hashtags that have been used to collect data.

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.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: a survey. *IEEE access*, 12:18416–18450.

Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. 2022. Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry. *Knowledge-Based Systems*, 255:109780.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta.

Sudesh Bandara, Tharindu Dasath, Senula Nanayakkara, Samadhi Rathnayake, and Thusithanjana Thilakarthna. 2024. Enhancing typing dynamics emotion recognition: A multiclass xgboost approach for accurate sentiment detection. In *2024 6th International Conference on Advancements in Computing (ICAC)*, pages 354–359. IEEE.

Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. Analyzing the online communication of environmental movement organizations: Nlp approaches to topics, sentiment, and emotions. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76.

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

Jacob Beck. 2023. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17(3):331–353.

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

Disha Chavan, Esha Anvekar, Megha Dandapat, Vaibhav Bichave, and Jayashree Jagdale. 2023. Machine learning applied in emotion classification: a survey on dataset, techniques, and trends for text based documents. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1697–1703. IEEE.

Tulika Chutia and Nomi Baruah. 2024. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57(8):203.

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(3):232–242.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Emotional robbert and insensitive bertje: combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 257–263. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

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

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*. Ph.D. thesis, Université de Nanterre-Paris X.

Aline Étienne, Delphine Battistelli, and Gwénolé Lecorvé. 2024. Emotion identification for French in written texts: Considering modes of emotion expression as a step towards text complexity

analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 168–185, Bangkok, Thailand. Association for Computational Linguistics.

Kexin Feng and Theodora Chaspari. 2020. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science*, 2:9.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

A Aruna Gladys and V Vetriselvi. 2023. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556:126693.

Akhilesh Kumar and Awadhesh Kumar. 2022. Human sentiment analysis on social media through naïve bayes classifier. *J. Sci. Res*, 66(1).

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215.

Sofie Labat, Naomi Ackaert, Thomas Demeester, and Véronique Hoste. 2022. Variation in the expression and annotation of emotions: A wizard of oz pilot study. In *Proceedings of the 1st workshop on perspectivist approaches to NLP@ LREC2022*, pages 66–72.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.

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*.

Donatas Meškelė and Flavius Frasincar. 2020. Aldonar: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, 57(3):102211.

Annie Piolat and Rachid Bannour. 2009. Emotaix: un scénario de tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année psychologique*, 109(4):655–698.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Peipei Yang, Amir Hussain, and Tariq Durrani. 2012. Merging senticnet and wordnet-affect emotion lists for sentiment analysis. In *2012 IEEE 11th international conference on signal processing*, volume 2, pages 1251–1255. IEEE.

Dibyendu Seal, Uttam K Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018*, pages 423–430. Springer.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Angelina Tzacheva and Sanchari Chatterjee. 2024. Actionable pattern discovery for emotion detection in bigdata in education and business. In *International Conference on AI, Machine Learning and Data Science (AIMDS 2024)*, volume 13.

Xinhao Zou and Konstantin Markov. 2024. Combining graph nn and llm for improved text-based emotion recognition. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 143–154. Springer.