EDAR: A pipeline for Emotion and Dialogue Act Recognition

Elie Dina
Happiso
IDMC, U.Lorraine
dina.happiso@gmail.com

Rania Ayachi Kibech
Happiso
ayachi.happiso@gmail.com

Miguel Couceiro
U.Lorraine, CNRS, LORIA
miguel.couceiro@loria.fr
U.Lisbon, IST, INESC-ID
miguel.couceiro@inesc-id.pt

Abstract

Individuals facing financial difficulties often make decisions driven by emotions rather than rational analysis. EDAR, a pipeline for Emotion and Dialogue Act Recognition, is designed specifically for the debt collection process in France. By integrating EDAR into decision-making systems, debt collection outcomes could be improved. The pipeline employs Machine Learning and Deep Learning models, demonstrating that smaller models with fewer parameters can achieve high performance, offering an efficient alternative to large language models.

1 Introduction and Motivation

Debt collection is a challenging field that demands persistence, diligence, and a high degree of empathy, as financial decisions are often driven by emotions rather than logic (Lucey and Dowling, 2005). Traditional methods, such as Sentiment Analysis (SA), often overlook the emotional complexities of debtors, leading to increased stress for both parties. Previous papers lack a clear distinction between SA and Emotion Recognition (ER). While SA refers to the classification of sentiment as positive, neutral, or negative; ER classifies a person's emotional state, such as happiness, sadness, worry, and anger.

This paper focuses on ER, which offers a promising solution by identifying the emotional state of a debtor and enabling empathetic responses, potentially improving repayment outcomes (Bachman et al., 2000; Wang et al., 2022). EDAR improves this process by recognizing nuanced emotional states, helping the bailiff tailor their responses accordingly. Unlike conventional practices, EDAR balances efficiency with empathy, improving both debt collection outcomes and debtor satisfaction, positioning it as a novel solution in the industry[1].

Emotions are reactions that human beings experience in response to events or situations[2]; and they are able to determine how we function socially, make decisions, and more (Suhaimi et al., 2020). Understanding emotions is a major challenge for both humans and machines (Shaheen et al., 2014). People find it challenging in the context of textual messages, due to the lack of non-verbal emotional cues, such as facial expression and tonality (Derks et al., 2008; Baron-Cohen and Wheelwright, 2004). Furthermore, machines need an accurate ground truth for emotion modeling. Achieving such truth is difficult, as emotions are very subjective (Barrett et al., 2007).

Despite extensive research, there is no consensus on the definition of emotions. Paul Ekman (1972) argued that emotions are universal, identifying six basic emotions: fear, disgust, anger, surprise, joy, and sadness, which are biologically hardwired and consistent across cultures (Ekman et al., 1999). In contrast, some researchers claim that emotions are culturally specific and vary depending on social context and geography (Mesquita and Frijda, 1992). Furthermore, researchers such as Robert Plutchik (1980) introduced the "wheel of emotions," suggesting that emotions are interconnected and evolve through complex interactions, rather than being distinct, unrelated states (Plutchik, 1980).

Given the complexity of emotions and the difficulty in pinpointing what constitutes one, this article adopts the term "emotional state" to limit the ambiguity regarding the definition of emotions. An emotional state is perceived as a prolonged and less intense experience that reflects a person's overall mood or affect condition over time for a specific situation.

This work will be used as a baseline for a decision-support process for debt collection, help-

[1] https://www.metcredit.com/blog/the-role-of-emotional-intelligence-in-debt-collection/

[2] https://www.verywellmind.com/what-are-emotions-2795178/

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 175–186
January 19–24, 2025.

ing to categorize the debtor's profile based on multiple criteria. Debt collection is important for the economy, as it helps lower lending interest rates, improves individual credit scores, and strengthens the overall economy. Consequently, this work contributes to the United Nations (UN) eighth Sustainable Development Goal[3] (SDG), which focuses on promoting decent work and economic growth.

The main contributions of the paper are twofold. Firstly, it provides a method to recognize five main emotional states and five dialogue acts recursively present in textual messages through Machine Learning (ML) and Deep Learning (DL) models. Secondly, it demonstrates that even with a low number of parameters, the latter ML and DL models can achieve good performance with low energy and resource consumption, thus avoiding the use of Large Language Models (LLMs) that entail a negative environmental impact.

2 Related Work

Interest in the field of ER has increased significantly in the last decade (Han et al., 2023). This section will examine the key factors shaping this field, including the modalities used to detect emotions, the various emotion models employed alongside the dataset used, as well as the evolution of methodologies.

Modalities in this field can be divided into four main different categories: *textual*, which involves determining the emotions embedded within a textual message (Yohanes et al., 2023); *vocal*, which focuses on extracting vocal features such as tone, pitch, etc. (Luthman, 2022); *visual* through facial expression and body gestures (Wei et al., 2024), and *multimodal* taking into account multiple modalities simultaneously (Castellano et al., 2008).

One of the challenges that we address in this paper is to determine the emotion embedded in textual messages exchanged between the debtor and the file administrator. To design models with high performance and good generalization capabilities, well-annotated datasets with good Inter-Annotator Agreement (IAA) are required (Bobicev and Sokolova, 2017). Previous studies presented different datasets that vary in the emotion model followed, language support, domain application, label count, and labels used. In terms of emotion models, many datasets focus on Ekman's basic emotions (Ekman, 1992), such as Emobank (Buechel and Hahn, 2017) and Aman (Aman and Szpakowicz). Other datasets extended Plutchik's wheel of emotions (Plutchik, 1980), such as DENS (Liu et al., 2019). Finally, some datasets included a broader nuanced emotional states, such as GoEmotion, considering 27 different emotions (Demszky et al., 2020). In debt collection, the emotions identified during interactions between the debtor and the debt administrator revealed five distinct emotions, some of which were not observed in datasets from previous studies. On the one hand, this is partially because the definitions of emotions are concise, and the annotators can confuse and/or combine two or more different emotions. On the other hand, it is partly due to the lack of interest in these emotions, such as "suspicion".

Methodologies are evolving significantly in ER, ranging from simple rule-based systems to advanced DL models. Recent studies have shown five main approaches with promising results in their respective datasets. Earlier methods focused on a keyword-based approach that classified a text based on emotion-related keywords (Shivhare et al., 2015); also, the use of rule-based approaches, which used predefined rules and lexicons to identify emotions, was applied (Udochukwu and He, 2015). After the development of AI, learning approaches took the lead with different ML models such as Naive Bayes (NB) (Sharupa et al., 2020), Decision Tree (DT) (Lee et al., 2011), Logistic Regression (LR) (Basile et al., 2019) and more. In addition, DL models were developed and significantly improved ER with the use of Convolutional Neural Networks (CNN) (Cahyani et al., 2022), Recurrent Neural Network (RNN) (Li et al., 2021), and Attention Layers (Han et al., 2023). Today, interest is peaking towards LLMs that further enhance ER capabilities by understanding context and subtle nuances in the text (Pico et al., 2024).

Table 1 shows a sample of the best model performance in some research papers dealing with Textual ER (TER) using ML, DL, or by leveraging LLMs. The significant performance gap is mainly attributed to differences in the dataset rather than to the model used. These studies utilize different

Research Paper	Model	F1-Score
(Sharupa et al., 2020)	NB	0.956
(Han et al., 2023)	XLNet-BiGRU-Att	0.825
(Pico et al., 2024)	GPT-3.5	0.479
(Demszky et al., 2020)	BERT	0.460

Table 1: SOTA models' performance

[3] https://sdgs.un.org/goals

datasets, each with varying labels and label counts, making direct performance comparisons unfair and potentially misleading. This paper specifically addresses the classification of emotional states and dialogue acts within the context of debt collection, focusing on a specialized lexicon tailored to this domain.

3 Data Preparation

In the field of AI, understanding data is crucial to enhance the explainability and performance of the model. This section describes the data used in both the pre-processing and processing stages.

3.1 Data Acquisition

The data was given by a justice commissioner located in France. The latter, with the approval of debtors and in strict accordance with the ethical guidelines set by the GDPR, continues to collect the needed data from the debtors, for further analysis, and possibly to develop a decision-support system for debt collection.

The extracted messages, predominantly written in French, were primarily sent via email. Although email communication is generally formal, some messages exhibit informal language or contain significant grammatical errors. In fact, many debtors are non-native French speakers, even if having a primary residence in metropolitan France.

Non-native french speakers, make up 10.7% of the population in France, often express emotions differently due to cultural and linguistic factors. Recognizing this in our model is essential for accurately capturing the varied emotional cues present in debtor communications. The prevalence of grammatical errors among non-native speakers further underscores the importance of designing a model that can handle linguistic diversity, thus enhancing its robustness.

A total of approximately 5,130 messages were collected. No specific selection criteria were applied, except for a defined date range to ensure the relevance of the data.

3.2 Cleaning Process

The cleaning process followed for this work consists of three main parts.

The first step in the cleaning process ensured consistency and readability.

- Address encoding errors, remove irrelevant content, and ensure text uniformity;
- Remove formalities, salutations, and irrelevant references that might be present in emails. For example: *Bonjour* (Good morning), *Cordialement* (Cordially), references of images in the text, and so on;
- Divide messages into segments, based on punctuation, for more precise annotation. In fact, long messages present multiple emotions and dialogue acts.

The second step ensured anonymization, as the data contains personal and private information.

- Anonymize and standardize personal and sensitive information, by tagging the debtor's name, credit card numbers, etc.;
- Tag and categorize digits, dates, time, and monetary values, to ensure consistency in the text, and no bias towards specific values.

The final step in the cleaning process was achieved to proceed with building the models, this step was achieved to remove unrequited data.

- Remove emojis and emoticons. Although emoticons and emojis present emotional cues, they were disregarded, as only two were found in the entire dataset. This might be due to the fact, that the incoming messages are emails, thus requesting formality;
- Remove extraneous information, such as information between brackets and square brackets;
- Remove punctuation marks, except "?" and "!"; and stopwords, except those showing negation. The retained information might show emotional tones or dialogue acts cues.

3.3 Annotation Process

As mentioned, this task was developed to analyse textual messages received from the debtors. Administrators do not annotate these messages; therefore, a manual annotation was made in an attempt to determine the emotions presented in the messages automatically. Given the sensitivity and privacy of the data, the annotation process was performed locally using EZCAT (Guibon et al., 2022), a user-friendly tool for annotating conversations.

To facilitate the annotation process, a guideline was created to address the context of debt collection and the various scenarios that may arise. The guidelines in Appendix A were frequently updated

when new cases emerged. An annotation guideline is crucial to ensure consistency in labeling criteria across different annotators, provide clear instructions for annotators on classifying different types of textual data.

Humans are prone to errors. Since models are trained on human classifications, they inherit the same errors made by annotators, which results in misleading evaluations. An IAA assessment was performed to ensure the validity and reliability of the annotated data. The π coefficient was used to assess IAA due to its suitability in handling multi-class categorization in highly specific and nuanced emotional datasets. This metric offered a practical alternative for evaluating consistency across emotional states and dialogue act categories, aligning well with the needs of this study's custom annotation scheme. The latter assessment was performed on a subset of 100 messages. Two annotators independently labeled each of the 100 messages according to the annotation guideline in the Appendix A, followed by reconciliation.

For example, the segment: "*Je viens de faire un paiement, pourriez-vous confirmer sa reception*" (I have just made a payment, could you confirm receipt); can be considered both informative ("*je viens de faire un paiement*") and interrogative ("*pourriez-vous confirmer sa reception*"). However, since the manual segmentation process was not performed for this step, the annotators mentioned the most relevant discourse acts, which is in that case *Informative*.

Figures 5 and 6 in Appendix B illustrate the frequency of agreement between both annotators, with respect to emotional states and discourse acts, respectively. To be able to calculate the IAA and determine the reliability of the annotation, the coefficient π was taken into account. The latter gives a probability for each category. Equations (1) and (2) in Appendix B show how the coefficients IAA and π were calculated.

The IAA results presented in Figures 5 and 6 (see Appendix B), 0.866 and 0.857 respectively, demonstrate a high level of consistency among annotators. These values reflect excellent reliability in the annotation process. Furthermore, it suggests that both annotators consistently understood the categorization criteria. The 100 most confusing messages were selected and, by ensuring consistency in these segments, the reliability of the annotation process can be inferred.

3.4 Trials Done

Three different trials of annotation were conducted successively, until satisfactory results were obtained. These changes were discussed with file administrators to ensure their need and validity.

1. The choice of the labels was based on a quick overview of the actual data. Six different labels were identified: *Collaborative*, *Neutral*, *Preoccupied*, *Angry*, *Surprised*, and *Uninterested*.

2. Eight different labels were defined: *Neutral*, *Collaborative*, *Informative*, *Preoccupied*, *Angry*, *Surprised*, *Mistrust*, and *Uninterested*.

3. Definition of two different annotation sets. The first Emotional Tones focusing on: *Neutral*, *Worry*, *Anger*, *Mistrust* and *Surprise*. The second subset would focus on dialogue acts: *Collaborative*, *Informative*, *Interrogative*, *Uninterested*, and *Other*.

3.5 Exploratory Data Analysis

The dataset contains approximately 5,130 messages. Following automatic and manual segmentation, a total of 14,853 segments were identified. Among these, 1,810 segments were found to be duplicates. These duplicates often arose from repeated emails in response to the bailiff, showing anger or mistrust from the debtor, or recurring short phrases such as "*un virement a été fait*" (a transfer has been made). After removing the duplicates, roughly 13,000 unique segments remained.

The annotation process was conducted on a subset of the dataset due to its time-consuming nature and the necessity to evaluate the model's performance on previously unseen data. Various debt case files were selected for annotation, whether active or closed. The cases varied as for example some individuals had filed for over-indebtedness[4] (*dossier de surendettement*); others were deceased or experiencing financial difficulties. Additionally, some cases demonstrated debtor cooperation and willingness to make payments, while others involved rebuttals and denials of the debt. In total, 1,960 segments were annotated.

Tables 2 and 3 summarize the distribution of emotional states and discourse types (or discourse acts) in the annotated segments. Most segments express a neutral emotion, indicating that neutrality

[4]A procedure in France that cancels all previous debts.

Dialogue Acts	Frequency
Collaborative	814
Informative	738
Interrogative	290
Uninterest	80
Other	38

Table 2: Frequency Distribution of Discourse Types in Annotated Segments

Emotional Tone	Frequency
Neutral	1316
Worry	296
Anger	100
Mistrust	194
Surprise	54

Table 3: Frequency Distribution of Emotions in Annotated Segments

is the predominant emotional tone in the dataset. With respect to discourse types, *collaborative* discourse is the most frequent, closely followed by *informative* discourse. This suggests a substantial prevalence of collaborative and informative discourse acts in the data. The overall frequency distribution underscores the diverse range of emotional and discursive expressions captured, emphasizing neutral and cooperative interactions.

Figures 7 and 8 (see Appendix C) illustrate the word clouds for the mistrust emotion and the collaborative discourse type, respectively. As shown in the mistrust word cloud (Fig. 7), words such as "*arnaque*" (scam), "*escroquerie*" (swindle) and "*fraude*" (fraud) are primarily present. These terms suggest that the debtor perceives the communication as a scam and believes that the bailiff is attempting to defraud them financially. This perception is plausible, particularly for debts over a year old, as some debtors may have forgotten or assumed the debt was already settled. With respect to the type of collaborative discourse, words such as "*virement*" (transfer), "*échéancier*" (payment schedule) and "PT"[5] are frequently observed. These terms indicate that the debtor is cooperating by proposing or requesting a payment plan or promising to make a payment on a specific date.

In conclusion, the previous analysis reveals that *neutral* emotions and *collaborative* discourse are the most prevalent in the dataset, with significant *mistrust* associated with perceived fraud.

[5]Tag used for reference to monetary value (Price Tag)

3.6 Tasks Developed

In an initial attempt (Task 0), the models were built to compare all the different categories simultaneously for both trials 1 and 2. This first attempt, yielded in overfitting and resulted in unsatisfactory results. This is mainly due to high imbalance between the different categories, especially in earlier trials.

To address the challenge of data imbalance in multiclass classification of emotional tones, we implemented a three-task strategy (see Figure 1), which were applied to the latter two trials (Trial 2 and 3):

1. Combining Emotional Tones: Emotional tones such as *Worry*, *Anger*, *Mistrust*, and *Surprise* were grouped into a single class labeled as *Others*, thereby allowing for the comparison between the more frequent class *Neutral* and *Others* using the first classification model.

2. Differentiating Emotional Tones within *Others*: Messages classified as *Others* by the first model were further analyzed using a second classification model to distinguish among the individual emotional tones.

3. Classifying Dialogue Acts: A third classification model was developed to differentiate between various dialogue acts, providing additional contextual understanding.

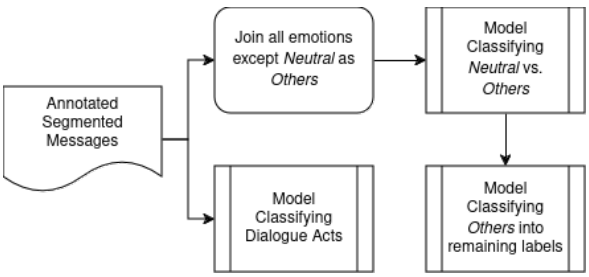

Figure 1: Tasks for trials 2 and 3

4 Model Building

Eleven ML and DL models were implemented and developed to identify the one with the highest performance. These models are LR, Multinomial NB (MNB), Support Vector Machine (SVM), eXtreme Gradient Boost (XGB), Adaptive Boosting (AdaBoost), DT, Random Forest (RF), Gradient Boosting Classifier (GBC), K-Nearest Neighbor Classifier (KNC), LightGBM, and a DL model based on a Bi-LSTM.

The latter models were chosen due to their diverse strengths in handling classification tasks, enabling a comprehensive comparison to determine the model with the highest performance in accurately identifying emotions. The Bi-LSTM model was built using PyTorch, a Python library. The model architecture consists of several key components designed for multi-task classification, which is developped in Appendix D.

A testing size of 20% was taken into account for each emotion. To ensure that the models do not overfit, hyper-parameter tuning was achieved, considering a wide range of hyper-parameters. Therefore, Grid Search Cross-Validation (GridCV) was used.

5 Results and Discussions

This section presents the outcomes of the experimental trials, highlighting the best-performing models for each task and discussing their implications for emotion recognition in the debt collection domain.

Table 4 presents the best performing models for the different trials carried out and the tasks developed. Each task considered different annotation guidelines, sets of emotions and dialogue acts, and datasets. The last round of annotation presents the most promising results, except in the second task, where the second trial outperforms the third. This might be due to chance or to the fact that the dataset was much smaller. The difference between both trials is insignificant and therefore can be ignored.

Task #	Trial #	Model	Vectorizer	F1-Score
0	1	MNB	CV	0.335
0	2	GBC	CV	0.507
1	2	RF	TF-IDF	0.829
1	3	Bi-LSTM	TF-IDF	0.901
2	2	MNB	CV	0.932
2	3	MNB	TF-IDF	0.926
3	2	MNB	TF-IDF	0.746
3	3	Bi-LSTM	TF-IDF	0.922

Table 4: Models performance over the different trials and tasks.

The macro F1-score was used instead of the weighted F1-score to ensure that the evaluation equally reflects the performance across all classes, regardless of their frequency. This approach addresses the issue of class imbalance, where certain emotional tones and discourse types classes may be underrepresented, by giving each class equal importance. Consequently, the macro F1-score provides a more balanced assessment of the model's ability to accurately classify less frequent emotions.

Figures 2, 3, and 4 present the confusion matrices (CM) for each of the tasks in the third trial, presenting the performance of the models that perform the best.

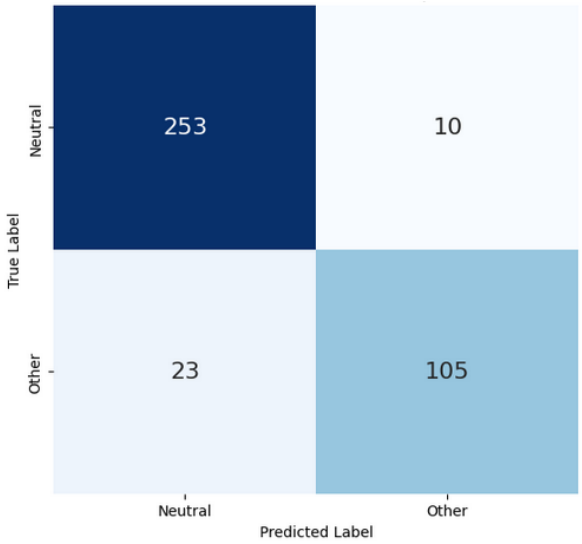

Figure 2: CM Task 1

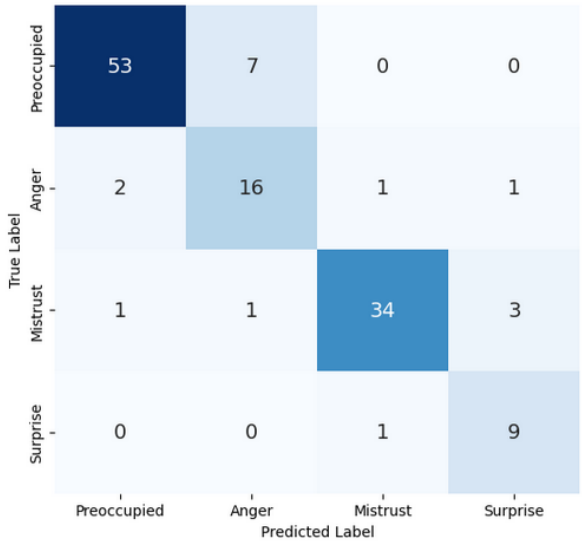

Figure 3: CM Task 2

The discrepancies in the first task (Figure 2) could be due to errors in the annotation or mainly confusion between the *preoccupied* and *neutral* class. Regarding the second task (Figure 3), misclassifications are mainly present in sentences where confusion was present as well for the annotators, as some segments might present more than one emotion, as the segmentation process is not the most effective and accurate, as it is only based on

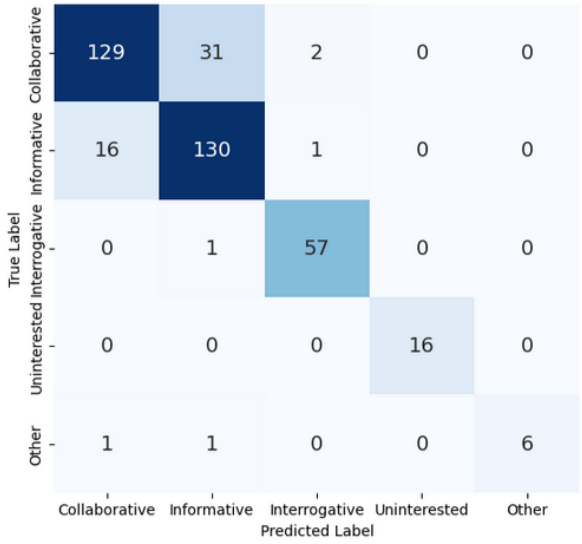

Figure 4: CM Task 3

punctuation. Finally, in regards to the third task, misclassifications appear the most between *collaborative* and *informative* discourse types. These discrepancies could also be due to inaccurate segmentation or annotation errors for some segments.

To test models' generalization capability on similar unseen data, we used 150 additional newly collected segments. The model demonstrated its ability to correctly identify emotions with an accuracy of 87% and discourse acts with an accuracy of 91%, suggesting promising results.

No LLMs were employed to achieve ER and discourse type classification due to their computational expense and time requirements. Instead, traditional ML and DL models were developed, which achieved satisfactory performance. These results demonstrated that conventional models can produce excellent outcomes in such tasks. Additionally, these models are more eco-friendly, as they involve significantly fewer parameters compared to the large number needed to train and fine-tune LLMs, thereby reducing the environmental impact associated with computational resources.

6 Conclusion and Future Work

Debt collection is a delicate but critical professional field, as administrators deal with private financial information. With the increasing number of scams nowadays, people tend to be more suspicious of incoming communications that ask for money for any reason. Thus understanding human behavior is essential in debt collection as trust plays a pivotal role in successful outcomes. By accurately assessing and dealing with debtors, bailiffs can build a cooperative base fostering trust, ultimately leading to effective debt recovery. This underscores the importance of ER in debt recovery, as it helps to interpret emotional signals and respond accordingly. The work done on ER in this work showed promising results without the need for extensive annotation or the usage of LLMs, confirming that traditional models, such as ML and DL models, can be very effective while remaining eco-conscious compared to LLMs.

The models developed in this application classify emotions after automatic segmentation based on punctuation. The drawback of such a method is the inaccurate segmentation, as some debtors might overuse or even underuse punctuation, thus leading to confusion in the model. To mitigate this limitation, a DL model with attention mechanisms could be developed to identify specific segments of the text that convey different emotions.

Additionally, multi-label models could be developed to capture the complexity of textual messages, where multiple emotions or dialogue acts might coexist within the same segment. This approach would address the limitations of automatic segmentation by allowing the model to assign more than one label per segment, thus providing a more nuanced understanding of the message's emotional and communicative intent. Such models could improve overall performance by accounting for the overlapping nature of emotions and dialogue acts often present in human communication.

Although concrete metrics have not been gathered at this stage, future work will focus on evaluating EDAR's effectiveness through key performance indicators. These will include metrics such as the overall emotional feedback from debtors and response rates to specific intervention templates. By comparing emotional response patterns before and after EDAR implementation, we aim to quantify its impact on debt recovery outcomes. Tracking de-escalation in emotionally charged interactions will also provide insights into its potential for reducing debtor stress and improving collection rates.

Finally, while the dataset was sourced from a French-speaking justice commissioner, future work will prioritize expanding the dataset to include data from different regions, linguistic backgrounds, and diverse debt collection contexts. This will contribute to more robust and generalizable findings, enabling the pipeline to adapt to a wider variety of communication styles and legal frameworks.

Ethical Considerations

Working in the field of debt collection involves handling personal and private data, which are protected by the National Commission on Informatics and Liberty (CNIL[6]) and the European General Data Protection Regulation (GDPR[7]).

According to the CNIL, personal data[8] are considered to be "any information relating to an identified or identifiable individual; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number (e.g., social security number) or one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity (e.g., name and first name, date of birth, biometrics data, fingerprints, DNA, etc.):"

As aforementioned, debt collectors attempt gathering personal information regarding the debtor for different reasons. These data collected fall under the category of personal information, thereby necessitate adherence to the CNIL and the GDPR.

To ensure the confidentiality and data security of these sensitive data, all employees within both the justice commissioner and our company have signed a Non-Disclosure Agreement (NDA) prohibiting them from sharing any of the data accessed or processed. Furthermore, the GDPR imposes regulations on the collection, processing, and storage of personal data, ensuring the protection of individuals' privacy rights.

The three main articles that should be taken into consideration, while applying our work are:

- Article 7, mentioning the importance of a free given, informed and unambiguous consent regarding the data storage and processing.
- Article 17, granting the right to have the personal data erased under certain circumstances, when the data is no longer necessary.
- Article 24, necessitating the implementation of robust security measures to safeguard personal data.

While the use of emotion recognition in debt collection offers benefits, it raises ethical concerns around the potential for manipulation or the exacerbation of debtor stress. To mitigate these risks, EDAR ensures that sensitive interactions are flagged for human review, allowing administrators to handle them with empathy and care. Furthermore, strict adherence to GDPR ensures that personal data is handled securely, with clear consent obtained from debtors. As part of future work, we will explore additional safeguards to ensure that the emotional data is used to empower rather than exploit debtors.

A final consideration should be explicitly stated about our work. Indeed, although, the pipeline we proposed achieves SOTA results, these are to be taken with *a grain of salt*, especially, when deploying it in real-world, legal domains. For instance, the fact that training was performed on some benchmark datasets that are prone to biases could have undesirable ethical implications or generalization issues.

Limitations

Several limitations of our study should be acknowledged. Firstly, the study was conducted using a single annotated dataset that might raise questions on the model's generalization capability. Also, the data was sourced from a single justice commissioner in France, which may introduce potential geographic, cultural, as well as other social biases such as political or religious orientations, which have not been accounted for in the current analysis.

Secondly, while many existing models are trained on datasets from platforms like Twitter or other social media, this paper focuses uniquely on the debt collection domain. This is the first model to incorporate the specialized vocabulary and context of debt recovery, making it directly relevant to this field. When tested on approximately 200 unseen messages, the model achieved an accuracy of 87%, demonstrating its capacity to generalize effectively within this specific domain. However, further research is needed to confirm performance across even larger and more diverse debt-related datasets.

Thirdly, we did not investigate possible proxies or biases within this dataset. Addressing these biases in future work could lead to more robust conclusions. Additionally, voice data contains a wealth of information, which may mitigate some of the aforementioned biases. Exploring the use of automated speech emotion recognition to infer characteristics such as gender, nationality, and other demographic factors could enhance the pipeline's performance and provide further insights, while al-

[6]https://www.cnil.fr/en
[7]https://gdpr-info.eu/
[8]https://www.cnil.fr/en/personal-data-definition

ways taking into consideration GDPR regulations.

Finally, positive emotions are rarely encountered in debt collection communications, as debtors typically express negative or neutral sentiments. While some debtors experience relief when reaching an agreeable payment solution, the occurrence of positive emotions is minimal (0.5 per thousand) and does not significantly enhance analysis. Thus, we chose to classify these instances within the "Collaborative" dialogue act and "Neutral" emotional state category, ensuring focus on more prevalent and analytically valuable emotions.

Acknowledgments

We would like to thank the different debt file administrators for their time and collaboration, especially while developing the different annotation guideline, and clustering the different emotional states and dialogue acts of debtors.

References

Saima Aman and Stan Szpakowicz. *Identifying Expressions of Emotion in Text*, page 196–205. Springer Berlin Heidelberg.

John Bachman, Steven Stein, K. Campbell, and Gill Sitarenios. 2000. Emotional intelligence in the collection of debt. *International Journal of Selection and Assessment*, 8:176 – 182.

Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34(2):163–175.

Lisa Barrett, Batja Mesquita, Kevin Ochsner, and James Gross. 2007. The experience of emotion. *Annual Review of Psychology*, 58:373–403.

Angelo Basile, Marc Franco-Salvador, Neha Pawar, Sanja Štajner, Mara Chinea Rios, and Yassine Benajiba. 2019. SymantoResearch at SemEval-2019 task 3: Combined neural models for emotion classification in human-chatbot conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 330–334, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 97–102. INCOMA Ltd.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Denis Eka Cahyani, Aji Prasetya Wibawa, Didik Dwi Prasetya, Langlang Gumilar, Fadhilah Akhbar, and Egi Rehani Triyulinar. 2022. Emotion detection in text using convolutional neural network. *2022 International Conference on Electrical and Information Technology (IEIT)*, pages 372–376.

Ginevra Castellano, Loic Kessous, and George Caridakis. 2008. *Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech*, pages 92–103. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Preprint*, arXiv:2005.00547.

Daantje Derks, Agneta H. Fischer, and Arjan E.R. Bos. 2008. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785. Instructional Support for Enhancing Students' Information Problem Solving Ability.

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

Paul Ekman, Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, and Terrence J. Sejnowski. 1999. Classifying facial actions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(10):974–989.

Gaël Guibon, Luce Lefeuvre, Matthieu Labeau, and Chloé Clavel. 2022. EZCAT: an easy conversation annotation tool. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1788–1797. European Language Resources Association.

Tian Han, Zhu Zhang, Mingyuan Ren, Changchun Dong, Xiaolin Jiang, and Quansheng Zhuang. 2023. Text emotion recognition based on xlnet-bigru-att. *Electronics*, 12(12).

Chi-Chun Lee, Emily Mower Provost, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53:1162–1171.

Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang. 2021. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683.

Chen Liu, Muhammad Osama, and Anderson de Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6292–6297. Association for Computational Linguistics.

Brian Lucey and Michael Dowling. 2005. The role of feelings in investor decision-making. *Journal of Economic Surveys*, 19:211–237.

Felix Luthman. 2022. Multilingual speech emotion recognition using pretrained models powered by self-supervised learning. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS).

B. Mesquita and N. H. Frijda. 1992. Cultural variations in emotions: a review. *Psychological Bulletin*, 112(2):179–204.

Aaron Pico, Emilio Vivancos, Ana Garcia-Fornes, and Vicente Botti. 2024. Exploring text-generating large language models (llms) for emotion recognition in affective intelligent agents. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 1: EAA*, pages 491–498. INSTICC, SciTePress.

Robert Plutchik. 1980. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, volume 6, page 383–392. IEEE.

Nazia Anjum Sharupa, Minhaz Rahman, Nasif Alvi, M. Raihan, Afsana Islam, and Tanzil Raihan. 2020. Emotion detection of twitter post using multinomial naive bayes. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Shiv Naresh Shivhare, Shakun Garg, and Anitesh Mishra. 2015. Emotionfinder: Detecting emotion from blogs and textual documents. pages 52–57.

Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. 2020. Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*, 2020(1):8875426.

Orizu Udochukwu and Yulan He. 2015. A rule-based approach to implicit emotion detection in text. In *Natural Language Processing and Information Systems*, pages 197–203, Cham. Springer International Publishing.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Preprint*, arXiv:2203.06935.

Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. 2024. Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, 237:121419.

Daniel Yohanes, Jessen Surya Putra, Kenneth Filbert, Kristien Margi Suryaningrum, and Hanis Amalia Saputri. 2023. Emotion detection in textual data using deep learning. *Procedia Computer Science*, 227:464–473. 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).

A Annotation Guidelines

This appendix shows the guideline followed for the process of annotation of both the emotional states and discourse acts.

A.1 Emotions Annotation Guideline

Neutral

- Greetings or polite expressions, e.g. "merci"
- File or Debtor's reference
- Giving general information about the debt procedure or themselves

Preoccupied

- Mention of financial difficulties, through informing allocations reception.
- Health problems, such as hospitalization, dealing with cancer, and more.
- Informing about breaking the law, and being imprisoned.
- Family difficulties, death in the family, recent divorce, and such.

Anger

- Using curse words
- Throwing blame on the bailiff or creditor
- Refusing to pay the debt
- Considering that the messages are harassment

Mistrust

- Surprised by the legal proceeding or the pursuit from the administrator
- Does not remember the debt
- Interpret that the message is a scam

Surprise

- Surprised by a reminder, while a message was already sent explaining the situation
- Surprised by the amount, as they remember a different amount

A.2 Discourse Acts Annotation Guideline

Collaborative

- Giving personal information, such as matrimonial situation or number of dependents
- Accepting to pay the debt, or to a payment plan
- Proposing or requesting a payment plan
- Requesting a phone call

Informative

- Informing that the payment was made
- Informing about a call attempt
- Repeating information that were previously mentioned in a phone call or in a previous email

Interrogative

- Requesting more information regarding the debt
- Requesting a payment confirmation
- Requesting information about the study
- Asking questions about the functionality of the debtor's secure space

Uninterested

- Does not want to pay the debt
- When the whole message consist of curse words
- Warning the bailiff about legal procedure for harassment

Other

- When the act of dialogue does not fit any of the previous categories.

B Inter-Annotator Agreement

$$IAA = \frac{A_0 - A_e^\pi}{1 - A_e^\pi} \tag{1}$$

where:

- A_0 represents the observed agreement among annotators.
- A_e^π denotes the expected agreement by chance.

The expected agreement by chance A_e^π is given by:

$$A_e^\pi = \frac{1}{(2N)^2} \sum_{q \in Q} (n_q^2) \tag{2}$$

where:

- Q is the set of categories.
- n_q is the total number of items categorized as q by all annotators.
- $2N$ accounts for the total number of annotations, considering that each item is annotated by multiple annotators.

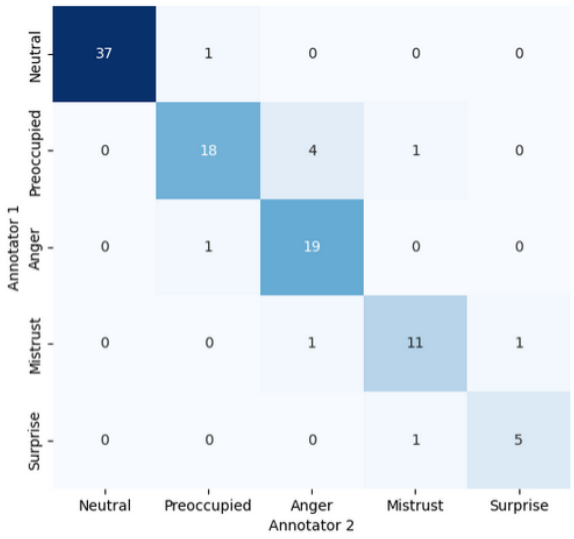

Figure 5: IAA heatmap for emotions - 0.866

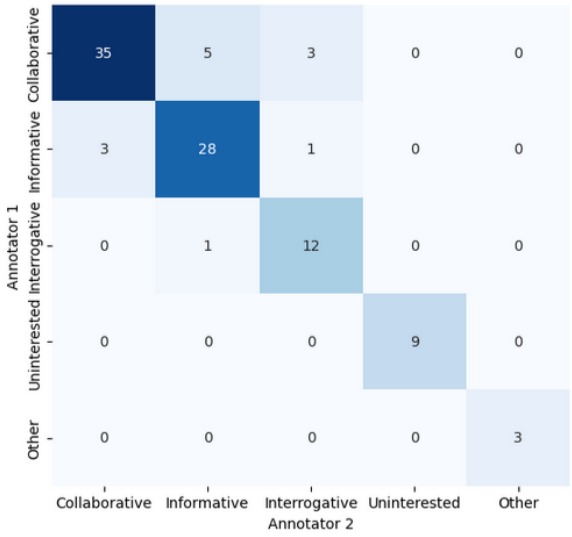

Figure 6: IAA heatmap for discourse acts - 0.857

C Exploratory Data Analysis

Figure 7: Wordcloud for Mistrust Emotion

Figure 8: Wordcloud for Collaborative discourse acts

D Bi-LSTM Model

- Input layer: The model accepts TF-IDF encoded fixed-length vectors as input.
- Embedding layer: This layers converts high-dimensional TF-IDF vectors into a lower-dimensional embedding space, using linear transformation.
- Bi-LSTM layer: This layer takes different parameters such as:
  - Input size: Equivalent to the number of features in TF-IDF vectors, in other words, the dimension of the embedding spave
  - Hidden size: Number of hidden units in the LSTM, taken to 256 in the first layer, and 128 in the second layer
  - Number of LSTM layers: two layers
  - Bidirectionality: It enables to capture contextual information from both forward and backward directions, making the LSTM model bidirectional
  - Dropout: Regularization applied to reduce the possibility of overfitting, set to 0.4
- Fully Connected (FC) layer: This layer uses the Rectified Linear Unit (ReLU), it processes the output from the Bi-LSTM layer, through linear transformation, introducing non-linearity and enabling complex feature combinations. The number of units set are 128.
- Output layer: This final layer is a FC layer, having different number of units, based on the classification made. Three tasks are achieved in the classification process:
  - Task 1: Joining all actual emotions into *Other* and comparing them to *Neutral*. Number of unit: 2 (having two classes: *Neutral* and *Other*).
  - Task 2: Classifying the emotions that were joined (*Preoccupied*, *Anger*, *Mistrust*, and *Surprise*). Number of unit: 4 (having four different classes).
  - Task 3: Classifying the acts of dialogue. Number of unit: 5 (having five different discourse types: *Collaborative*, *Informative*, *Interrogative*, *Uninterested*, *Other*).
- Loss function and Metrics: The loss function used for both tasks is categorical cross-entropy, as it is the most suitable for multi-class classification. In addition, as an optimizer, Adam was employed to updayes model parameters during training.