

Tunisian Dialect Speech Corpus: Construction and Emotion Annotation

Latifa Iben Nasr^a, Abir Masmoudi^b and Lamia Hadrich Belguith^c

MIRACL Laboratory, University of Sfax, Tunisia

Keywords: Emotion, Spontaneous Speech, Tunisian Dialect, Multi-Domain.

Abstract: Speech Emotion Recognition (SER) using Natural Language Processing (NLP) for underrepresented dialects faces significant challenges due to the lack of annotated corpora. This research addresses this issue by constructing and annotating SERTUS (Speech Emotion Recognition in Tunisian Spontaneous speech), a novel corpus of spontaneous speech in the Tunisian Dialect (TD), collected from various domains such as sports, politics, and culture. SERTUS includes both registers of TD: the popular (familiar) register and the intellectual register, capturing a diverse range of emotions in spontaneous settings and natural interactions across different regions of Tunisia. Our methodology uses a categorical approach to emotion annotation and employs inter-annotator agreement measures to ensure the reliability and consistency of the annotations. The results demonstrate a high level of agreement among annotators, indicating the robustness of the annotation process. The study's core contribution lies in its comprehensive and rigorous approach to the development of a dataset of spontaneous emotional speech in this dialect. The constructed corpus has significant potential applications in various fields, such as human-computer interaction, mental health monitoring, call center analytics, and social robotics. It also facilitates the development of more accurate and culturally nuanced SER systems. This work contributes to existing research by providing a high-quality annotated corpus while emphasizing the importance of including underrepresented dialects in NLP research.

1 INTRODUCTION


Human emotions arise in response to objects or events in their surroundings, influencing various aspects of human life, including attention, memory retention, goal achievement, recognition of priorities, knowledge-based motivation, interpersonal communication, cognitive development, emotional regulation, and effort motivation (Gannouni et al., 2020). Speech Emotion Recognition (SER) has emerged as a crucial area of research with diverse applications, such as mental health assessment, customer service optimization, and human-computer interaction (Goncalves et al., 2024).


As for languages, most speech emotion datasets are implemented in German, English, and Spanish. Several SER studies have also been conducted in Dutch, Danish, Mandarin, and other European and Asian languages (Aljuhani et al., 2021). Arabic SER is an emerging research field (Besdouri et al., 2024). The developmental lag in Arabic SER is due to the


lack of available resources compared to other languages (Alamri et al., 2023).

Expanding upon the linguistic complexities, the Arabic language, with its rich diversity of dialects and linguistic nuances, presents a formidable hurdle in SER tasks. Among these dialects, Tunisian Dialect (TD) stands out due to its unique characteristics and complexities. Spoken by approximately 12 million people across various regions (Zribi et al., 2014), TD encompasses numerous regional varieties, including the Tunis dialect (Capital), Sahil dialect, Sfax dialect, Northwestern TD, Southwestern TD, and Southeastern TD (Gibson, 1999).

Moreover, TD exhibits a distinctive linguistic fusion, incorporating vocabulary from multiple languages, such as French, Turkish, and Berber (Masmoudi et al., 2018). For instance, Tunisians often borrow expressions from French, incorporating them into everyday conversations with phrases like 'ça va,' 'désolé,' 'rendez-vous,' and 'merci.' Indeed, the linguistic situation in Tunisia is described as 'polyglossic,' where multiple languages and language varieties coexist. This linguistic diversity is a testament to Tunisia's rich historical heritage, influenced by var-

^a  <https://orcid.org/0009-0008-0677-7458>

^b  <https://orcid.org/0000-0002-5987-8876>

^c  <https://orcid.org/0000-0002-4868-657X>

ious powers, including French colonization, the Ottoman Empire, and earlier influences.

Therefore, analyzing emotional expressions in TD requires a nuanced understanding of the linguistic intricacies of the dialect, including the dual registers of intellectualized and popular dialects (Boukadida, 2008). The intellectualized dialect, used by scholars and in media broadcasts, features a formal lexicon and borrowings while retaining dialectal structures. In contrast, the popular (familiar) dialect is used for everyday communication. For example, the sentence "Disappointing situation" is rendered as [مَقْصِيَّةُ الْحَالَةِ / mVaynaT AlHAAlaT] in the familiar register and [الْوَضْعِيَّةُ حَايِيَّةُ / Alwa.diyaT xAybT] in the intellectualized register.

Furthermore, emotional expressions in Tunisian Arabic are heavily influenced by cultural nuances. For instance, a speaker from Tunis might express joy by saying [فَرِحْتُ بَرَشًا / fraHit bar\$A] ("I'm very happy"), using [بَرَشًا / bar\$A] to intensify the emotion, while a southern speaker might say [هِيَا نَحْسُ رُوحي تَقُولُ فِي الْجَنَّةِ / hika ni-His rowHy tVowl fiy AljanaT] ("I feel like I'm in paradise") to convey a deeper sense of joy through metaphor. Similarly, frustration might be expressed as [تَعَبْتُ مِنَ الْحَالَةِ / tibit mn Al.hAlaT] ("I'm tired of the situation") in the north, whereas a southern speaker might use [دَاوْ بِيَا الْحَالِ / dAV biyA AlHAL] ("I'm overwhelmed by the situation") to indicate a higher level of emotional intensity.

These variations illustrate how cultural and regional nuances shape emotional expressions in Tunisian Arabic. Despite this linguistic diversity, the field of SER in TD remains limited, with only a few studies, such as those conducted by (Nasr et al., 2023), (Meddeb et al., 2016), and (Messaoudi et al., 2022), which are annotated with distinct emotion classes but have a limited size distribution.

The primary objective of this research is to address this gap by providing emotional speech annotations in Tunisian Arabic based on a categorical approach, which will be made publicly available to the scientific community. To achieve this, we aim to develop a newly collected corpus of spontaneous TD across different domains and capture multiple registers of TD, some of which have already been utilized in previous studies (Nasr et al., 2023), to ensure the richness of our dataset. Figure 1 elucidates the meticulous construction process of our SERTUS corpus.

The major contributions of this research are summarized as follows:

- Collecting spontaneous data from various domains in TD.
- Annotating the corpus using a categorical approach.
- Assessing the effectiveness of this annotation using an agreement measure.

The remainder of this paper is structured as follows: In Section 2, we review the related works presented in the literature. In Section 3, we outline the construction of SERTUS (Speech Emotion Recognition in TUnisian Spontaneous Speech). In Section 4, we establish a discussion. Finally, in Section 5, we draw conclusions and discuss future research directions.

2 RELATED WORK

The critical assessment of SER systems depends on the quality of the databases used and the performance metrics achieved. To select an appropriate dataset, several criteria must be taken into account, including the naturalness of emotions (natural, acted, or induced), the size of the database, the diversity of available emotions, and the annotation approach (dimensional or categorical).

Arabic SER is an emerging field. However, there is a notable scarcity of available Arabic speech emotion datasets. Between 2000 and 2021, emotional databases in Arabic speech accounted for only 4.76% of the total number of databases across all languages (Iben Nasr et al., 2024). Additionally, acted and elicited emotion databases are primarily used for Arabic SER (Alamri et al., 2023). For example, KSUE-motions is an acted database for Standard Arabic that comprises 5 hours and 5 minutes of recorded data. Another example is EYASE, which represents a semi-natural Egyptian Arabic dataset consisting of 579 utterances ranging in length from 1 to 6 seconds (El Seknedy and Fawzi, 2022). To the best of our knowledge, only two datasets are available in TD: the REGIM_TES dataset (Meddeb et al., 2016), which includes 720 acted emotional speech samples with lengths of up to 5 seconds, and TuniSER (Messaoudi et al., 2022), which contains 2,771 induced speech utterances with durations varying from 0.41 to 15.31 seconds, averaging around 1 second.

In comparing Arabic with other languages such as English and French, we notice a notable lack of both the quality and quantity of available datasets for SER. For instance, English language datasets, such as the IEMOCAP dataset (Busso et al., 2008), include natural and spontaneous conversations between ac-

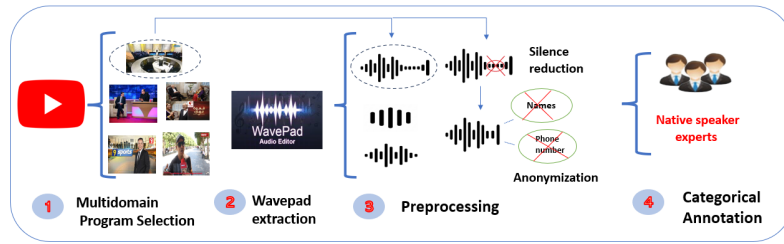


Figure 1: Process of SERTUS Construction.

tors, totaling approximately 12 hours of audio recordings. Similarly, the SAVEE dataset (Jackson and Haq, 2014) comprises natural emotional speech recorded from four English male speakers, with a total duration of 4 hours. Furthermore, French language datasets, like the RECOLA dataset (Ringeval et al., 2013), contain natural emotional speech recordings in various contexts, with a total of around 8 hours of audio data.

However, the Arabic language suffers from a dearth of comparable resources in terms of both quality and quantity compared to these well-established English and French datasets. As previously mentioned, the available Arabic datasets, predominantly consisting of acted or elicited speech, lack the naturalness and spontaneity of real-world emotional expressions. Moreover, compared to other datasets, the sizes of the existing Arabic datasets are significantly smaller, exacerbating the challenges faced in developing robust Arabic SER systems. This disparity underscores the urgent need for additional efforts to address the shortcomings in Arabic SER, particularly regarding the quality and size of the datasets. Hence, it is vital to apply this approach in diverse linguistic contexts and identify its gaps in Arabic SER corpora.

3 SERTUS CONSTRUCTION

3.1 Data Collection

In Tunisia, it is possible to make fair use of copyrighted materials according to Article 10 of Law No. 94-36 of February 24, 1994, related to literary and artistic property, as amended by Law No. 2009-33 of June 23, 2009. In our case, within a research framework, and as mentioned above, there are two oral linguistic registers (Boukadida, 2008). For the first register, we drew inspiration from Tunisian television programs, particularly talk shows on YouTube. We aimed to analyze the range of emotions expressed by guests who shared their views on various topics related to social, political, cultural, and sporting matters. As for the second register, we extracted data

from the 'Tunisian Reality' YouTube series, which consists of interviews conducted on the streets using sidewalk microphones in Tunis, the capital city of Tunisia. These series capture spontaneous emotional responses and comments from individuals in public spaces, addressing a wide range of topics such as politics and the economy. We collected data from different regions, including southern and northern Tunisia, to capture dialectical differences.

Based on a variety of talk shows and 'Tunisian Reality' series, with an emphasis on the various domains and the two registers of TD, and by including diverse public regions, we were able to build a comprehensive representation of spontaneous emotional reactions within the Tunisian population, ultimately leading to a better understanding of the SER process. These resources offer a wide range of emotional expressions and regional linguistic variations to better capture the main TD characteristics, including multi-lingual aspects, various word types, as well as morphological, syntactical, and lexical differences (Masmoudi et al., 2018) (Nasr et al., 2023).

Our SERTUS database contains over 23.85 hours of spontaneous TD speech by 1,259 speakers. We used the 'WavePad' software, thanks to its user-friendly interface and advanced features, to extract voiced segments from the audio files. We extracted the speakers' responses and eliminated not only the accompanying questions raised by TV show presenters or journalists after conducting street interviews but also extraneous noise sources. Each speaker's response was treated as a distinct audio segment to ensure clarity and coherence. Our dataset includes 3,793 recordings with durations ranging from 3.15 seconds to 10.97 minutes, providing a wide range of speech scenarios. In fact, each segment spans approximately 20 seconds to capture a broad spectrum of linguistic expressions and emotional nuances. A uniform sampling rate of 44.1 kHz was applied to enhance consistency, and the corpus was gathered over a five-month period.

Furthermore, to provide new insights into the distribution and composition of our dataset, Table 1 categorizes the collected data into different domains and

provides a detailed breakdown of the size and composition of our dataset in each domain.

Table 1: Statistics on the number of minutes collected per domain.

Domain	Program	Number of minutes
Culture	Kahwa Arbi Labess Show	180
Policy	Wahech Checha	60
Sport	Attesia Sport	60
Society	Safi Kalbek Maa Ala	180
Multi-domain	Tunisian Reality	951

3.2 Pre-Processing Steps

To ensure accurate annotations, we have carried out some pre-processing steps, namely reducing periods of silence and anonymizing personal information.

3.2.1 Silence Reduction

Our study has primarily focused on selected speeches delivered by speakers or guests, rather than presentations made by TV program presenters. Therefore, we have excluded presenters' statements from our corpus. It is worth noting that our corpus is made up of the voices of TV speakers or guests without any overlapping speech. Given the spontaneous nature of our corpus, periods of silence used for reflection before giving a specific response are common. To ensure consistency, any silence lasting more than two seconds has been standardized to a duration of exactly two seconds.

3.2.2 Anonymization

Our speech data, collected from authentic conversations, may include personal information such as names, professions, locations, and other details. To ensure confidentiality and compliance with data protection regulations, it is crucial to closely follow and oversee the legal framework surrounding the management of personal data. By adhering to the Good Practice Guide Principles (2006), we prioritize the anonymization of our data to prevent the identification of individuals.

Our approach involves identifying sensitive elements in recordings that may reveal personal information. To ensure anonymity, it is essential to identify and replace cues that may disclose this information with periods of silence, thereby reducing the frequency of personal speech information. We acknowledge that specific combinations of cues may still re-

veal individuals' identities. For example, a speaker may mention only their first name, which is not considered identifiable personal information. Also, professional roles (e.g., journalists) are not generally considered identifiers unless specific details are disclosed.

After analyzing the corpus, uncommon instances, such as specific job titles and locations, require careful handling to avoid the direct identification of personal information. Anonymization is a manual process that involves thoroughly examining all audio segments to protect privacy and adhere to legal standards.

3.3 Guidelines

The annotators were provided with clear and detailed guidelines to ensure a thorough understanding of the study's objectives and the precise ways their contributions would be utilized. These guidelines emphasized the importance of objectivity throughout the annotation process to minimize any potential biases stemming from individual perceptions, emotions, or subjective interpretations.

To achieve objectivity, annotators were instructed to approach the emotional labeling task with a neutral perspective, focusing solely on the linguistic and acoustic cues present in the audio data rather than their personal emotional responses. This objectivity was critical, given the diverse range of spontaneous emotions expressed in the TD and the need for consistent and reliable annotations across the dataset.

To facilitate the annotation process, a structured and user-friendly method was employed. Annotators were provided with a pre-designed table that they were to complete using Microsoft Excel. After listening to each audio segment, they were required to associate the identified emotions with the corresponding audio file name. This structured approach ensured a systematic and organized method of capturing emotional labels, reducing the likelihood of errors or inconsistencies.

The recruitment of the annotation team was handled with care and precision. A public call for tender was launched to develop a robust and transparent selection process, ensuring that only the most qualified individuals were chosen for the task. The selection criteria were stringent, focusing on expertise in sentiment annotation, fluency in TD, and a strong background in linguistic analysis. We specifically targeted individuals with academic qualifications from the Faculty of Arts and Humanities, particularly those with advanced degrees in linguistics and a solid understanding of the emotional nuances in the TD.

The final annotation team consisted of three na-

tive speakers of TD: two females and one male, all of whom were experts in the field of sentiment analysis. Inspired by previous works such as (Messaoudi et al., 2022) and (Macary et al., 2020), which also utilized three annotators, each annotator independently reviewed and labeled the audio files to ensure a diverse yet consistent perspective on the emotional content of the corpus. This independent annotation process, combined with their linguistic expertise, provided a high degree of reliability and validity in the emotional labeling of our dataset. Additionally, the diversity in the team, both in terms of gender and analytical perspectives, further strengthened the robustness of the annotations, contributing to the overall quality of the study's outcomes.

3.4 Annotation

At this stage, annotators must take into account the definitions of emotions identified in each audio segment of our dataset. Each segment receives a specific emotion. Manual annotation is extremely important as it helps annotators build a dedicated learning corpus for emotion analysis. The final designation for each clip was determined by a majority vote. Our categorical annotation includes six different emotions as follows:

- Neutrality: الفيلم اليوم يعرضه / Alfylm Alywm y'r.dwh / The movie is showing today.
- Joy: حاجة تفرح علخر / hAjT tfr.h 'lxr / Its makes me so happy
- Disgust: الوضع مهواش جملاً / Alw.d' mhwA\$ jmlA / The situation is not going well at all.
- Satisfaction: بصراحة أنا راضية و توا بشوة بشوة / b.srA.hT 'nA rA.dyT w tW b\$wT b\$wT t\$'hsn / Honestly, I'm satisfied with this situation, and little by little, it will get better.
- Sadness: زاني تعبت مل الفقر و الإحتياجية / rAniy t'bt ml Alfqar w AlA.htyAjyT/ I suffer from poverty and low living standards.
- Anger: سراق كيما هكا ميخفوش ري / srAaq kymA hkA myxfw.S rby/ Thieves like these people do not fear God.

The selection of emotions in our corpus—neutral, joy, sadness, disgust, anger, and satisfaction—was carefully determined through discussions with annotators and based on the specific characteristics of the data. This choice was influenced by several considerations, including the inclusion of "neutral" to capture

moments devoid of intense emotions and "satisfaction," which reflects a blend of joy and calm (Plutchik, 1980). The presence of these categories in our corpus was essential to accurately represent the observed emotional diversity. While this study does not directly compare these emotions with those in datasets from other languages, the selection aligns with established models such as (Plutchik, 1980) and (Ekman, 1992), while being tailored to the unique features of our dataset.

During the data collection phase, we encountered a significant challenge: the overwhelming prevalence of negative emotional discourse among Tunisian speakers. This pattern appeared to reflect the broader socio-economic situation in Tunisia, where ongoing challenges related to political instability, economic difficulties, and social unrest have shaped public sentiment. Despite our concerted efforts to balance the dataset by actively seeking out episodes that focused on more positive topics—such as love, personal achievements, and national celebrations—our corpus remained heavily skewed toward negative emotional expressions.

The emotional landscape captured in our data leaned heavily towards emotions like anger and sadness, while more positive emotions such as joy and contentment were notably underrepresented. As shown in Figure 2, the "joy" class accounted for a mere 4.69% and the "neutral" class accounted for only 3.02% of the emotion distribution, whereas the "anger" class represented 28.76%, indicating a stark contrast in the emotional categories present in the corpus.

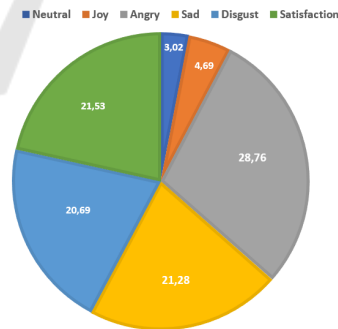


Figure 2: Class distribution in the corpus.

The domains represented in the data collected from Tunisian talk shows and street interviews have a direct impact on the types of emotions expressed by speakers. For example, political discussions, such as those from the program "Wahech Checha", often evoke negative emotions like anger. In one instance, a speaker expressed their anger over the political instability,

saying, [ينهبو في مخازن الدول بطريقة غير مباشرة سراق] /ynhbw fy mxAzn aldwl b.tryqT .gyr mbAsr srAq /They are looting the state's warehouses indirectly, thieves], reflecting the socio-political challenges faced by the country. Conversely, cultural topics, featured in programs like "Kahwa Arbi" and "Labess Show," tend to elicit more neutral or positive emotions. For example, a guest on "Labess Show" expressed satisfaction, stating, [فرحانة أي اليوم في برنامجك] /fr.hAnT 'ny alywm fy brnAmjk/ I am happy to be in your program today]. Sports-related discussions, such as those from "Attesia Sport," also display a range of emotions, from joy during team victories to disappointment during losses. For instance, a fan celebrating a win said, [ترجي ديمنا مفرحتنا] /trjy dymA mfr.htnA/ Attaraji always bring us joy]. The societal topics addressed in the programs "Safi Kalbek" and "Maa Ala", as well as some episodes in the "Tunisian Reality" YouTube series, evoke emotions of sadness and disgust that reflect the economic and social situation in Tunisia, stating, [حالنا يعلم به كان ري] /hAlnA ylm bh kAn rby/ Only God knows our situation]. By analyzing these varied domains, we can observe distinct emotional patterns associated with each field, enriching our understanding of how emotions are tied to the topics being discussed. This domain-emotion correlation is vital for a nuanced analysis of emotional discourse in TD, especially when considering the regional and contextual factors that influence emotional expression.

3.5 Calculation of the Inter-Annotator Agreement

Evaluating agreement between annotators in corpus annotation tasks is a critical step in assessing the consistency of the annotations. Inter-annotator reliability ensures that the annotated data reflect consistent judgments, minimizing individual biases. One of the most widely used measures to assess this agreement is Fleiss' kappa coefficient (Gwet, 2021). It measures the degree of concordance between multiple annotators, regardless of the number of categories, and accounts for chance agreement, making it a robust tool for such analyses. Several Python libraries, such as scikit-learn and NLTK, provide convenient tools for researchers to compute this coefficient and validate the reliability of their annotations.

The formula for Fleiss' kappa is as follows:

$$K = (Po - Pe) / (1 - Pe) \quad (1)$$

Po: represents the proportion of observed agreement between annotators.

Pe: denotes the proportion of agreement expected by chance.

In our case, the annotators demonstrated a high level of agreement with with $K = 0.89$, indicating strong consistency in annotating the emotions expressed in spontaneous speech. This result demonstrates the reliability of the annotations; however, some disagreements persisted due to the inherent challenges posed by the nature of the data..

The first challenge is the spontaneous nature of our data, which includes background noise and disfluency phenomena (Boughariou et al., 2021). This issue affects the performance of the annotation process. Annotating emotions in a spontaneous corpus presents additional challenges due to the complexity and overlap of emotional states. Emotions are not always expressed unequivocally and can shift within the same discourse, making it difficult to achieve perfect agreement between annotators. For example, disagreements can arise based on how each annotator perceives subtle nuances within the same sentence.

Consider the following sentence in TD:

ياخي هذا كولو صارلي ومازلت نتسمّاح / yAxy h*A klw .sArly w mAzelt ntsemAe.h/ After everything that happened to me, I continue to hurt myself]. This sentence could be interpreted as expressing sadness, due to the sense of resignation it conveys. However, it could also be perceived as reflecting disgust towards oneself or the situation, illustrating the difficulty of assigning a single emotional label to nuanced expressions.

Another example that illustrates emotional overlap is: [كنت مغموم و الدنيا صاقت بيا، أما الحمد لله] /konit ma.gmwm w AldenyA .dAqt biyA, amA Al.hmd llh t.hasnt w fra.het bar\$A /I was sad and felt overwhelmed, but thank God, things improved, and I am very happy now.] Here, we see a clear transition from sadness to joy, demonstrating an emotional shift marked by the improvement in the situation. This kind of emotional complexity makes it difficult to standardize annotations, as each annotator may focus on different aspects of the discourse.

Thus, while the high kappa coefficient reflects strong overall agreement between the annotators, these examples show that some divergence can persist, especially in situations where emotions are ambiguous or overlapping.

Table 2: Comparative table of Arabic SER corpora.

Work	Name	Langue	Nature	Size	Number of emotions
(Meddeb et al., 2016)	REGIM.TES	TD	Acted	720 samples of up to 5s each	5
(Messaoudi et al., 2022)	TuniSER	TD	Semi-natural	2771 utterances from 0.41 to 15.31s lengths	4
(Meftah et al., 2020)	KSUEmotions	MSA	Acted	5 hours and 5 minutes of data	5
(El Seknedi and Fawzi, 2022)	EYASE	Egyptian dialect	Semi-natural	579 utterances from 1s to 6s	4
Our work	SERTUS	TD	Spontaneous	23.85h	6

4 DISCUSSION

In the realm of Arabic SER, our study provides a comprehensive analysis of the SERTUS dataset. As shown in Table 2, our work stands out across several key dimensions, with the SERTUS dataset serving as the cornerstone of our research. Notably, our dataset features a substantial size, consisting of approximately 23.85 hours of audio recordings, with a total of 3793 samples. Furthermore, the data collected for SERTUS comes from a variety of programs across multiple domains, enhancing its applicability to a broad range of research areas and real-world applications. One of the main strengths of our dataset lies in the spontaneous nature of the speech data. Unlike datasets composed of acted or semi-natural speech, the natural interactions captured in SERTUS offer a more authentic representation of human emotions in everyday environments. This characteristic makes our dataset especially valuable for developing emotion recognition models that are better suited to real-life applications, surpassing the limitations posed by artificially constructed datasets. Additionally, SERTUS covers six distinct emotion categories, exceeding the four categories commonly found in comparable datasets, such as TuniSER and EYASE. This broader emotional range allows for the capture of more nuanced and varied emotional expressions, thereby contributing to the development of more accurate and robust emotion recognition models. The diversity and scale of spontaneous recordings in SERTUS provide fertile ground for advancing research in this area, facilitating improvements in systems designed for SER across various applications, such as human-computer interaction, clinical emotion detection, and behavioral analysis in complex environments.

The TD holds potential for generalization to other Arabic dialects, such as Algerian and Libyan, due to

their shared linguistic roots and phonological, lexical, and syntactic similarities. This facilitates transfer learning between these dialects. Our Tunisian emotional corpus, rich in annotated audio data, provides a solid foundation for developing models that can be adapted to neighboring dialects. By leveraging the linguistic and emotional nuances in this corpus, it is possible to refine emotion recognition algorithms, thereby enhancing the understanding of emotions in Algerian and Libyan dialects. This approach enables significant advancements in computational linguistics and in the emotion recognition community, benefiting not only the TD but also other Arabic dialects.

5 CONCLUSION AND FUTURE PERSPECTIVES

In the current research work, we introduced our spontaneous emotional speech corpus in TD, named SERTUS, with categorical annotation. We labeled our corpus with six emotions: neutral, disgust, anger, joy, satisfaction, and sadness, following established guidelines. This corpus includes more than 23.85 hours of recordings in different domains, such as sports and politics, and from various sources to capture regional differences in TD. The ultimate objective of this work was to ensure the consistency of this new corpus, and we achieved good agreement between annotators, measured by the kappa coefficient.

In future directions, we will address the limitations of our research, particularly the imbalance in our corpus, by applying data augmentation techniques and measuring their performance. Voici une version mise à jour du paragraphe avec l'ajout de la granularité plus fine :

Additionally, our focus will pivot towards refining

the annotation process to better capture the evolving nature of human emotions, especially within spontaneous speech. This will involve exploring alternative annotation methods, such as dimensional annotation, which can provide a more nuanced understanding of emotional subtleties. Specifically, we aim to adopt finer granularity in our annotation to detect overlapping emotions, such as disgust and anger. We also plan to extend the applicability of our dataset beyond its current scope. By collaborating with experts from various fields, we aim to explore how our annotated corpus can be utilized to advance research in areas such as NLP, affective computing, and cultural studies.

Furthermore, we will continue to investigate potential real-world applications for our dataset, including its use in developing emotion recognition systems, improving human-computer interaction interfaces, and facilitating cross-cultural communication.

REFERENCES

- Alamri, H. et al. (2023). Emotion recognition in arabic speech from saudi dialect corpus using machine learning and deep learning algorithms.
- Aljuhani, R. H., Alshutayri, A., and Alahdal, S. (2021). Arabic speech emotion recognition from saudi dialect corpus. *IEEE Access*, 9:127081–127085.
- Besdouri, F. Z., Zribi, I., and Belguith, L. H. (2024). Challenges and progress in developing speech recognition systems for dialectal arabic. *Speech Communication*, page 103110.
- Boughariou, E., Bahou, Y., and Belguith, L. H. (2021). Classification based method for disfluencies detection in spontaneous spoken tunisian dialect. In *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 182–195. Springer.
- Boukadida, N. (2008). *Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale)*. PhD thesis, Université Rennes 2; Université de Tunis.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- Ekman, P. (1992). Are there basic emotions?
- El Seknedy, M. and Fawzi, S. A. (2022). Emotion recognition system for arabic speech: Case study egyptian accent. In *International conference on model and data engineering*, pages 102–115. Springer.
- Gannouni, S., Aledaily, A., Belwafi, K., and Aboalsamh, H. (2020). Adaptive emotion detection using the valence-arousal-dominance model and eeg brain rhythmic activity changes in relevant brain lobes. *IEEE Access*, 8:67444–67455.
- Gibson, M. L. (1999). *Dialect contact in Tunisian Arabic: sociolinguistic and structural aspects*. PhD thesis, University of Reading.
- Goncalves, L., Salman, A. N., Naini, A. R., Velazquez, L. M., Thebaud, T., Garcia, L. P., Dehak, N., Sisman, B., and Busso, C. (2024). Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results. *Development*, 10(9,290):4–54.
- Gwet, K. L. (2021). Large-sample variance of fleiss generalized kappa. *Educational and Psychological Measurement*, 81(4):781–790.
- Iben Nasr, L., Masmoudi, A., and Hadrich Belguith, L. (2024). Survey on arabic speech emotion recognition. *International Journal of Speech Technology*, 27(1):53–68.
- Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.
- Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2020). Allosat: A new call center french corpus for satisfaction and frustration analysis. In *Language Resources and Evaluation Conference, LREC 2020*.
- Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., and Belguith, L. (2018). Automatic speech recognition system for tunisian dialect. *Language Resources and Evaluation*, 52:249–267.
- Meddeb, M., Karray, H., and Alimi, A. M. (2016). Automated extraction of features from arabic emotional speech corpus. *International Journal of Computer Information Systems and Industrial Management Applications*, 8:11–11.
- Meftah, A., Qamhan, M., Alotaibi, Y. A., and Zakariah, M. (2020). Arabic speech emotion recognition using knn and ksuemotions corpus. *International Journal of Simulation-Systems, Science & Technology*, 21(2):1–5.
- Messaoudi, A., Haddad, H., Hmida, M. B., and Graiet, M. (2022). Tuniser: Toward a tunisian speech emotion recognition system. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 234–241.
- Nasr, L. I., Masmoudi, A., and Belguith, L. H. (2023). Natural tunisian speech preprocessing for features extraction. In *2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS)*, pages 73–78. IEEE.
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1.
- Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.
- Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., and Habash, N. (2014). A conventional orthography for tunisian arabic. In *LREC*, pages 2355–2361.