

Re
Silence

S + T + ARTS
ReSilence

Retune the Soundscape of future cities through art and science collaboration
HORIZON- 101070278

D4.1

Techniques for soundscape experience assessment

Dissemination level:	Public
Contractual date of delivery:	Month 16, 31 December 2023
Actual date of delivery:	Month 20, 24 April 2024
Work package:	WP4: Technologies for soundscape experience assessment
Task:	T4.1: Emotional and cognitive sensing and analysis T4.2: Sentiment analysis from crawled social media data and reviews T4.3: Multimodal analysis of sound-related behaviour T4.4: Self-report generation
Type:	Report
Approval Status:	Final Draft
Version:	1.0
Number of pages:	71
Filename:	D4.1_resilience_Techniques_for_soundscape_experience_assessment_v1.0.docx
Abstract	

This deliverable will present basic techniques for emotion recognition that rely on EEG and physiological signals are going to be presented in this deliverable in the middle and finish of the project. This deliverable will also present a basic version of behavioural understanding from audio data, sentiment analysis from media content and reviews and self-report generation.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.



co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	01/02/2024	Creation of ToC	Beatrice de Gelder
0.2	11/03/2024	Input from UPF	Alexander Shvets
0.3	28/03/2024	Input from MPIEA	Melanie Wald-Fuhrmann
0.3	01/04/2024	Input from MU	Beatrice de Gelder
1.0	15/04/2024	Submission for internal review	Beatrice de Gelder

Author list

Organization	Name	Contact Information
MU	Beatrice de Gelder	b.degelder@maastrichtuniversity.nl
UPF	Alexander Shvets	alexander.shvets@upf.edu
MPIEA	Melanie Wald-Fuhrmann	melanie.wald-fuhrmann@ae.mpg.de
CERTH	Paraskevi Kritopoulou	pakrito@iti.gr
CERTH	Sotiris Diplaris	diplaris@iti.gr

Executive Summary

The ReSilence project employs diverse methodologies to understand emotional responses to aesthetic and artistic stimuli. Through the use of virtual reality (VR) environments, the project demonstrates the feasibility of eliciting and studying emotional reactions in controlled settings. This approach informs the development of VR-based methods for analysing emotional responses and underscores the importance of personalised algorithms to accommodate individual differences in emotional processing. Additionally, ReSilence focuses on sentiment analysis from social media data, multimodal analysis of sound-related experiences, and self-report generation. By integrating various techniques such as entity disambiguation (ED) and aspect-oriented sentiment analysis, the project aims to provide comprehensive insights into the relationship between artistic compositions and audience perceptions. This holistic approach facilitates transparent decision-making processes in evaluating aesthetic experiences.

Abbreviations and Acronyms

CERTH	Centre for Research & Technology Hellas
E2E	End-to-End
ECG	Electrocardiography
EEG	Electroencephalography
EL	Entity Linking
EOG	Electrooculogram
ERPs	Event-Related Potentials
fMRI	functional Magnetic Resonance Imaging
HEOG	Horizontal EOG
HS	Hate Speech
LLM	Large Language Model
MD	Mention Detection
MPIEA	Max Planck Institute for Empirical Aesthetics
NER	Named Entity Recognition
NLG	Natural Language Generation
NLI	Natural Language Inference
PPS	Peripersonal Space
RAG	Retrieval-Augmented Generation
VEOG	Vertical EOG
VPP	Vertex Positive Potential
VR	Virtual Reality

Table of Contents

1 INTRODUCTION.....	7
2 EMOTIONAL SENSING AND ANALYSIS	8
2.1 Related work.....	8
2.2 Methodology.....	9
2.3 Discussion.....	12
2.4 Relation to ReSilence	13
3 SENTIMENT ANALYSIS FROM CRAWLED SOCIAL MEDIA DATA AND REVIEWS	14
3.1 Related work.....	14
3.2 Methodology.....	15
3.2.1 Entity disambiguation	15
3.2.2 Mention detection	17
3.2.3 Neural word sense disambiguation	18
3.2.4 Sentiment analysis	20
3.2.5 Functionality test generation.....	20
3.3 Datasets	23
3.3.1 Entity disambiguation	23
3.3.2 Mention detection	25
3.3.3 Sentiment analysis	27
3.3.4 Functionality test generation.....	28
3.4 Evaluation	28
3.4.1 Entity disambiguation	28
3.4.2 Mention detection	31
3.4.3 Sentiment analysis	33
3.4.4 Functionality test generation.....	36
3.5 Future considerations	37
4 Multimodal analysis of sound-related experience and behaviour.....	39
4.1 Related work.....	39
4.2 Datasets	41
4.3 Preliminary results	41
4.4 Future considerations	42
5 SELF-REPORT GENERATION.....	43
5.1 Related work.....	43
5.2 Methodology.....	44

5.2.1 Abstract dependency template extraction	44
5.2.2 Candidate sentence retrieval	46
5.2.3 Candidate clause ranking and selection	46
5.2.4 Conversion to PredArg templates	47
5.2.5 Sample outputs of all components	47
5.2.6 Generation from fused information	48
5.3 Datasets	49
5.3.1 Abstract dependency template extraction	49
5.3.2 Generation from fused information	50
5.4 Experiments and preliminary results	50
5.4.1 Experiment setup	50
5.4.2 Results	51
5.4.3 Error analysis	53
5.4.4 Generation from fused information	53
5.5 Collaboration with Open Call Artists	55
5.6 Future considerations	63
6 CONCLUSIONS	64
7 REFERENCES	65

1 INTRODUCTION

In the rapidly evolving landscape of technology and human interaction, understanding and responding to the intricacies of human emotion and sentiment has become paramount. This is particularly evident in ReSilence, where the fusion of art, technology, and human experience is at the forefront.

Two aspects within this project underscore the significance of emotional and cognitive sensing as well as sentiment analysis in shaping user experiences and interactions. In this deliverable we delve into the development of advanced algorithms to recognize users' emotional states through EEG (Electroencephalography) and physiological signals, facilitating real-time adaptation of artistic installations. Meanwhile, we also focus on the realm of social media data and reviews, employing deep learning techniques to identify concepts, disambiguate entities, and analyse sentiment within the context of ReSilence's acoustic experiences.

Together, these tasks not only enhance our understanding of human emotion and sentiment but also pave the way for more immersive and responsive technological interactions in everyday life. The second section of this deliverable explains the emotional sensing and analysis response approach, Section 3 presents sentiment analysis from crawled social media data and reviews, Section 4 explains multimodal analysis of sound-related experience and behaviour, Section 5 frames self-report generation. The last section regards the conclusion of this deliverable.

2 EMOTIONAL SENSING AND ANALYSIS

In this activity novel physiological signal processing algorithms and signal integration frameworks will be developed to detect, extract and classify markers of emotional responses to aesthetic and/or artistic stimuli. Through this analysis ReSilence will try to understand individual and collective perceptions and preferences, leading to a more transparent decision-making process. Part of this work will be a novel algorithmic approach to investigate the correspondences between music composition and listeners' perception.

This work will be based on existing research in the context of social threat, in which the influences of various factors have been demonstrated, including agent characteristics, proximity, and social interaction, on social threat perception.

2.1 Related work

The capacity to recognize and respond to threats is a fundamental trait across many species, shaped by evolution (LeDoux and Daw, 2018). Research on defensive behaviours, both in humans and other animals, has delineated various responses to threats, primarily categorised as freezing and fleeing (Eilam, 2005). Freezing, characterised by heightened vigilance to potential threats, is a state that amplifies the processing of environmental danger signals (Blanchard et al., 1986; Mobbs and Kim, 2015; Terburg et al., 2018; Livermore et al., 2021). Investigations into freezing-like responses have explored a range of stimuli, including social cues like facial expressions and computer-based simulations such as virtual shooting tasks (Roelofs et al., 2010; Stins et al., 2011; Hagenars et al., 2014; Gladwin et al., 2016). Physiologically, freezing involves bradycardia and reduced mobility, orchestrated by subcortical connections between amygdala nuclei and brain regions regulating defensive behaviour (Roelofs et al., 2010; George et al., 2019).

The proximity of a threat is a crucial determinant of freezing reactions in humans. Studies on peripersonal space (PPS), the area surrounding the body where interactions with stimuli occur, have highlighted its role in modulating defensive behaviours during social interactions (Di Pellegrino and Làdavas, 2015; Bufacchi and Iannetti, 2018; Serino, 2019). Threats perceived within PPS elicit reduced motor cortex activity, heightened physiological responses, and enhanced neural processing in brain regions associated with defence mechanisms (Avenanti et al., 2012; Ruggiero et al., 2021; Vieira et al., 2020). Additionally, the neural network governing PPS responds to social threats invading personal space, activating regions involved in threat perception (de Borst et al., 2020; Ellena et al., 2021). EEG studies have demonstrated that threatening body expressions affect early event-related potentials (ERPs), such as N170 and vertex positive potential (VPP), with interactions observed between these electrophysiological measures and PPS (Stekelenburg and de Gelder, 2004; Van Heijnsbergen et al., 2007).

In addition to proximity, the ability to exert control over threats plays a crucial role in adaptive responses. Active control behaviours, such as escape or confrontation, can mitigate threat experiences (Iachini et al., 2016; Wendt et al., 2017). However, studying human responses to social threats presents challenges in creating realistic scenarios and measuring behaviour and physiology accurately. Virtual reality (VR) offers a promising avenue for addressing these challenges, allowing participants to experience threatening situations in a controlled environment (Parsons et al., 2017; Monti and Aglioti, 2018). VR-based designs incorporating

social threat from avatars have been successfully employed in various studies, including behavioural, fMRI and EEG investigations (Stolz et al., 2019; de Borst and de Gelder, 2022; Mello et al., 2022). Combining VR with behavioural and neural measurements provides a comprehensive understanding of how social threats are perceived and how the ability to control threats influences processing at different levels (de Borst et al., 2020; de Borst and de Gelder, 2022; Mello et al., 2022).

2.2 Methodology

i. VR Scenario

The VR setting depicts a dimly lit and narrow urban street, where an avatar displaying either anger or neutrality emerged and advanced towards the participant. This is developed using Unity 3D software (Unity Technologies, US), the VR scenario adopts similar body expressions of anger (raised arms) and neutrality (arms down) as a prior study (Mello et al., 2022), while introducing new environmental features and task parameters specific to this investigation. Prior to commencing EEG data collection, participants put on the VR headset (HTC VIVE) and have the opportunity to freely navigate the three-dimensional VR landscape by physically moving their head and walking. This allows them to explore the virtual streets and visually engage with the surroundings, fostering a sense of immersion in the VR environment.

ii. Experimental Design

Participants are briefed on a VR task where they encounter angry or neutral avatars in an urban setting. They are instructed to press a button (space bar) to halt the avatar's approach if they feel uncomfortable. Different cue conditions (five, see Figure 1) indicate the likelihood of successfully stopping the avatar, ranging from 0% to 100%. The avatar approaches from a distance of 5 metres, gradually increasing in size over time. Participants press the button upon discomfort, attempting to stop the avatar. Trials vary in whether the button press successfully halts the avatar's approach. Each trial lasts approximately 3.5 seconds. The study utilises a 5x2 design, including five cue conditions and two avatar emotions (angry, neutral), with 40 trials per condition presented randomly over five runs, totalling an hour of experiment time. The trial procedure in Figure 1 shows two screenshots corresponding to the participants' view during the VR experiment.

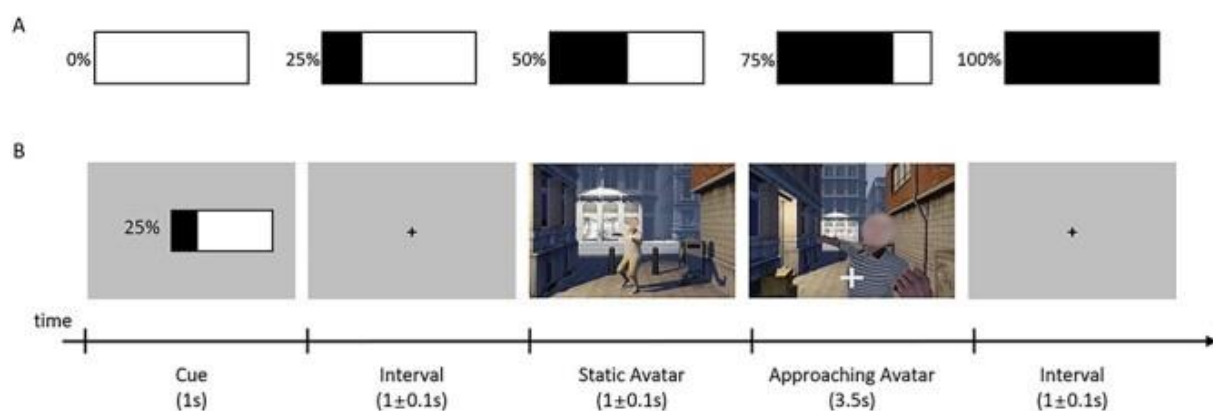


Figure 1: (A) Five kinds of controllable cues. (B) A trial procedure

iii. EEG acquisition

EEG data are collected using a scalp cap with 63 electrodes and a sampling frequency of 250 Hz. Electrodes are positioned according to the international 10–20 system, with Cz serving as the reference and FP1 as the ground electrode. Four electrodes measure electrooculogram (EOG), with two for vertical (VEOG) and two for horizontal (HEOG) eye movement.

Additionally, three electrodes are used for electrocardiography (ECG). The remaining 56 electrodes cover various scalp locations. Impedances are kept low to ensure signal quality. To minimise interference from the VR headset and potential head movements, participants use a chin rest during data collection (see Figure 2). This methodology focuses on analysing event-related potentials (ERPs) to static avatars due to their cleaner and more stable neural responses compared to dynamic stimuli. This approach aims to enhance the reliability of neural measures of emotional processing.



Figure 2: VR-chin rest-EEG setup shows a participant standing against the height-adjustable bar chair

iv. ERP analyses

The ERP analysis is conducted by aligning the timing to the appearance of the static avatar, allowing for clean ERPs in response to the image. A time window ranging from 200 ms before the avatar's onset to 1,000 ms after is extracted from each trial's pre-processed data. Baseline correction is performed by subtracting the average amplitude during the interval (-200 to 0 ms) before the avatar's onset. EEG data from each participant is averaged across experimental conditions to create ERPs for further analysis using IBM SPSS Statistics 27. The ERPs focus on responses to social threatening/non-threatening body expressions (angry/neutral avatars) and their sensitivity to threat control cues. EEG channels are divided into five spatial clusters, and prominent ERP components are identified for each region based on visual inspection and

previous research (de Gelder et al., 2004; Stekelenburg and de Gelder, 2004; Cunningham et al., 2005; Van Heijnsbergen et al., 2007; Luo et al., 2010; He et al., 2011; Chai et al., 2023). Mean amplitude within specific time windows is computed for each cluster. A repeated-measures ANOVA is applied to the mean amplitudes, separately for each ERP component, considering both threat control (0%/25%/50%/75%/100%) and avatar emotion (angry/neutral). Greenhouse–Geisser correction is used for F-ratio degrees of freedom, and significance was set at $P < 0.05$, with Bonferroni correction applied to control type I errors.

v. Time-frequency analyses

To examine the temporal fluctuations in oscillatory EEG power across the frequency range of 1 to 30 Hz, we utilise the complex Morlet wavelet transform on each trial. The parameters used for the wavelet transform include a frequency-bin size of 1 Hz, three cycles per time window, and a time-bin size of 50 ms. To mitigate edge effects, we conduct the time-frequency analysis on longer epochs corresponding to the duration of pre-processed epochs before ERP computation. Additionally, we establish a longer and earlier baseline ranging from -500 to -100 ms before the onset of the static avatar. Our focus in the statistical analysis is on the oscillatory power within the theta (4–7 Hz) band, given previous research suggesting its relevance to threat and control processing, particularly in frontal and central scalp regions. Therefore, electrodes positioned at Fz, FCz, Cz, F1, F2, FC1, FC2, C1, and C2 are chosen for this analysis. Examination of theta power reveals a peak between 100 and 200 ms after the onset time consistently across conditions in the frontal-central region. Accordingly, we extract the mean theta power during the time window of 100–200 ms at the selected electrodes and subject it to statistical analysis using the same repeated measures ANOVAs employed for the ERP analysis.

vi. ECG analysis

We extract a time frame spanning from 500 ms before the appearance of the static avatar to 4,500 ms after its onset from the continuous ECG data. Heart rate analysis is conducted using the ECGdeli toolbox (Pilia et al., 2021). Due to missing ECG data that might occur from a participant, the analysis includes data from 25 participants. The electrode positioned beneath the left collarbone, providing the strongest signal closest to the heart, is chosen for this analysis. Statistical procedures are consistent with those applied for ERP and theta power analyses, as described earlier.

vii. Behavioural analyses

The participants are instructed to press the button when they begin feeling uncomfortable with the approaching avatar. In certain trials, participants press the button once (4.81% of trials), while in others, participants either don't press the button or press it multiple times (16.42% of trials). Two behavioural metrics are captured here. Initially, we record the virtual distance between the participant and the avatar when the participants first pressed the button. To do this, we multiply the time at which the participants pressed the response button by the speed at which the avatar was approaching, then subtract the initial distance at which the avatar appeared (distance = 3.5 - response time × speed). Additionally, we tally the number of button presses. Similar to the physiological measures described earlier, each behavioural metric undergoes a 5 (controllable cue: 0%/25%/50%/75%/100%) × 2 (Avatar emotion: angry/neutral) repeated-measures ANOVA.

viii. VR questionnaire

Data regarding the participants' subjective encounter during the VR scenario is collected through a questionnaire administered after the experiment, as detailed in Figure 3 (Seinfeld et al., 2016; Seinfeld et al., 2021).

Item	Question	Mean ± SD
Realism	To what extent have you experienced the situation as if it was real?	4.69 ± 1.29
Vulnerable	Did you feel at any time that you were vulnerable?	3.96 ± 1.61
Violent	How violent do you find this scene is?	3.69 ± 1.38
Assaulted	Did you think that you could be physically assaulted during the scene?	3.19 ± 1.77
Attention to cue	How much attention did you pay to the probability cues during the experience?	5.08 ± 1.52
Fear of static angry avatar	How fearful did you feel when facing the static angry avatar?	3.23 ± 1.67
Fear of an approaching angry avatar	The sense of uncomfortable increased when the avatar got closer to me.	5.38 ± 1.55
Fear of static neutral avatar	How fearful did you feel when facing the static neutral avatar?	1.96 ± 1.34
Fear of an approaching neutral avatar	The sense of uncomfortable increased when the avatar got closer to me.	4.69 ± 1.44

Figure 3: The items and mean ± SD rating scores in the VR questionnaire are shown. Ratings were made on a 7-point scale (1 = not at all, 7 = completely)

2.3 Discussion

We examine how human participants respond behaviourally, electrically (EEG), and cardio vascularly (ECG) when encountering angry and neutral avatars in a VR setting, where they can control the interaction with the avatar to varying degrees. Behaviourally, we notice differences in participants' discomfort levels with the approaching avatar, depending on the presence of threat. This aligns with previous research indicating that imminent threat triggers defensive behaviour. The impact of personal space for social threat is initially demonstrated using full-body avatar expressions in a VR-fMRI study. By combining VR with EEG, we gain detailed insights into the timing of these processes. Questionnaire results also suggest heightened perceived threat when facing an angry avatar compared to a neutral one. Regarding controllability, we find that the controllable cue condition significantly affects two behavioural indicators: participants judged closer distances tolerable as the probability of successful control increased, and the number of button presses increased as the probability of control decreased. At the neural level, seeing a threatening body expression increases the amplitude of early ERP components (N170 and VPP) compared to non-threatening expressions. Moreover, threatening expressions elicit a smaller N3 than neutral ones, suggesting heightened emotional arousal and cognitive resource mobilisation. Full control enhances the amplitude of the late component LPP compared to partial control, indicating increased motivational engagement. Theta power in frontal central regions is modulated by avatar emotion, with greater increases for threatening expressions, consistent with higher emotional arousal. Our ECG results also reflect increased heart rate for social threat, indicating emotional arousal rather than a freezing response. In summary, our study sheds light on how humans process social threats under different levels of control, revealing distinct behavioural and neural mechanisms involved. Additionally, our novel VR-EEG-ECG setup

offers a valuable tool for future research on social interaction in naturalistic settings.

2.4 Relation to ReSilence

The findings from the above methodology on human responses to social threats in VR environments, particularly the behavioural, neural, and cardiovascular responses to angry and neutral avatars, can contribute significantly to the goal of developing novel physiological signal processing algorithms and integration frameworks.

Firstly, our study demonstrates the feasibility and effectiveness of using VR technology to elicit emotional responses in a controlled laboratory setting. This could inform the development of VR-based stimuli presentation methods for studying emotional responses to aesthetic and artistic stimuli. Secondly, our investigation into the temporal dynamics of behavioural control in response to threat perception provides insights into how individuals modulate their responses based on perceived control. This understanding can be applied to develop algorithms that track and analyse changes in emotional responses over time, helping to identify patterns and markers of emotional engagement with aesthetic and artistic stimuli.

Furthermore, the neural and cardiovascular measures obtained in the above study offer potential biomarkers for emotional responses. By developing algorithms to detect and classify these markers, ReSilence can gain deeper insights into individual and collective perceptions and preferences towards aesthetic and artistic stimuli. This could lead to the development of more transparent decision-making processes in evaluating the emotional impact of various artistic compositions and experiences.

Lastly, this methodology highlights the importance of considering individual differences in emotional processing, which can be addressed through personalised algorithms that account for variations in emotional responsiveness to different aesthetic stimuli. This personalised approach aligns with the goal of understanding the correspondence between music composition and listeners' perception, as it allows for a more nuanced analysis of how individuals interpret and emotionally engage with artistic content.

3 SENTIMENT ANALYSIS FROM CRAWLED SOCIAL MEDIA DATA AND REVIEWS

This part of WP4 is dedicated to language understanding. To understand feedback on a sonic experience that a user has provided in a written form as a report, review or a message on a social platform on the Web, it is necessary to recognize expressed emotions and identify mentions they were attributed to, i.e., the “concepts” or “entities”. Entities play a central role in formal knowledge structures which represent information in a format allowing for comprehensive automated data analysis. Detection of entity mentions in a free-form text, i.e., mention detection (MD), and matching them against existing entries in a knowledge structure known as entity disambiguation (ED) are two common subtasks of a task of entity linking (EL). In the following, we present our progress in these subtasks as well as in the task of sentiment analysis in the first part of the Project.

3.1 Related work

There are three major approaches to address entity linking: (i) solve the entire task in an end-to-end mode (E2E) by training a model to extract and disambiguate within one architecture, (ii) apply a promising model for ED in a combination with a basic MD (MD+ED) or (iii) consider EL as a retrieval task and perform a search as an extra stage of (i) or (ii).

The autoregressive model GENRE (De Cao et al., 2021) is an example of the E2E model. It produces outcomes by outputting the same input text while generating entry names and markup annotation that highlights mentions, left to right, token-by-token in an autoregressive fashion, conditioned on the context, leveraging the trie structure with possible entry names to constrain decoding. Having MD as its strong side it outperforms many modern models, however, when models provided with ground truth mentions, they have evaluation scores higher with a significant margin, as, e.g., Bootleg (Orr et al., 2021), revealing that GENRE performs poorly on the part related to ED. Another limitation consists in the inability to disambiguate entities unseen during the training, i.e., when global knowledge is updated.

Bootleg (Orr et al., 2021) is an MD+ED model for proper names that copes with rear entities using a transformer for encoding special reasoning patterns. It extracts candidate profiles from a knowledge repository and ranks their representations created by concatenating entity, relation, and type embeddings. The authors provide evidence that the learned representations from Bootleg successfully transfer to other tasks that require entity-based knowledge.

Another practice is to apply a retrieval stage at the ED step. Optionally, re-ranking or pairwise comparison (validation) can be further applied, e.g., using cross-encoders (Wu et al., 2020). TABi (Leszczynski et al., 2022) has a bi-encoder trained within contrastive learning to account for entity types. It retrieves entries by the similarity of their textual descriptions to the query text. It deals with both proper names and open discourse concepts but returns only one entity per text.

Proper dataset compilation is an accompanying problem met by every approach that includes the training of the models, especially in specialised domains. Generating synthetic examples using knowledgeable large language models (LLMs) became a valid option due to their superior generative capacity (Meng et al., 2022; Ye et al., 2022). However, the generated text

usually struggles with low-quality issues and may exhibit domain shifts with test data (Gao et al., 2023). Recently, a framework that utilises an LLM for labelling of examples rather than for generation of them was proposed (Zhang et al., 2023). It was shown that a specific model for named entity recognition and relation extraction trained on *silver*¹ data can outperform their teacher LLMs. In the Project, we carry out the design of some training datasets respecting these findings and develop our models being inspired by the state-of-the-art techniques discussed in this section.

3.2 Methodology

3.2.1 Entity disambiguation

Entity disambiguation as one of the major subtasks of entity linking plays an important role within T4.2. Recognition of sound-related concepts and music-related attributes like genre or theme is a necessary step in entity-oriented emotion analysis of opinion-rich textual material for understanding how people perceive acoustic experiences.

In the first part of the project, we developed an end-to-end approach for detecting entities of a specific category in a supervised learning paradigm. In particular, we focus on recognition of names of music genres since many of them possess a number of alternative senses causing a large amount of ambiguity. As we design our approach in a generic way and the chosen category makes a substantial challenge in entity disambiguation, success with it would signify the applicability of our solution to other categories as well. In the following, we detail individual steps.

First, we turn the task of linking the entities located in a text into a sequence labelling problem so that every token in a given text shall receive one of the following “*BIO*” tags: “*O*” - token is *outside* of a genre name, “*B-Genre*” - token corresponds to a single-word genre or to the *beginning* of a genre name, “*I-Genre*” - token is *inside* the multi-word genre; see Figure 4. Once a token(s) is classified as being a genre in a given context (i.e., the sense is disambiguated), linking to a knowledge entry becomes rather straightforward because, in general, concepts do not share the same names within a given category and we need only account for the fact that a concept that can be mentioned in a variety of ways (e.g., “rock and roll”, “rock and roll music”, “Rock 'n' Roll”, “rock&roll”, etc.) shall be assigned to the same knowledge entry to avoid semantic duplication. This formulation provides us with several advantages: (i) it is possible to detect several genres with a single model inference that makes it computational efficient; (ii) it is possible to detect single- and multi-word concepts among which there can be concepts whose tokens are embedded into mentions of other concepts including named entities (e.g., genre “jazz” is embedded into a concept “Barcelona International Jazz Festival”); (iii) it is possible to identify candidates for creating additional (missing or novel) entries in a knowledge structure when a model classifies some tokens as of a given category but they cannot be mapped to any of the existing entries.

¹ We refer *gold* data to ground-truth/human-labelled samples, and *silver* data to LLM-labelled samples.

Goral music in the first sentence in Figure 5) and it requires a lot of individual inferences for a single sentence with an excessive number of them if there are entries embedded into others (like *blues*, *dance*, and *swing* in the third sentence Figure 5). Therefore, we use it only for one-time annotation.

He wrote in his article "About Goral Music": "My discovery of the essential beauty of Goral music, **dance** and architecture is a very personal one; much of this beauty I have absorbed into my innermost **soul**".

dance → No
soul → No

Described by Carey as an homage to **Motown soul music**, Carey employed the help of Walter Afanasieff, who only had a small role on her debut, as well as Robert Clivillés and David Cole, from the **dance** group C+C Music Factory.

Motown → Yes
soul music → Yes
dance → No

Jackson explored genres including **pop**, **soul**, **rhythm and blues**, **funk**, **rock**, **disco**, **post-disco**, **dance-pop** and **new jack swing**.

pop → Yes
soul → Yes
rhythm and blues → Yes
blues → Yes
funk → Yes
rock → Yes
disco → Yes
post-disco → Yes
dance → Yes

Figure 5: Yes/No labels produced by the Mistral model for provided mentions and sentences

Finally, we create training, development, and test sets using annotated sentences and fine-tune and evaluate a transformer-based model (Vaswani et al., 2017). It has a smaller number of parameters than LLMs like Mistral and therefore requires less computational resources, performs classification faster, and may outperform them in generalisation in a specific task due to a smaller number of training examples needed to tune the model. At inference time, mentions labelled with positive labels are mapped to knowledge entries using traditional non-contextual character-based embedding models like FastText (Bojanowski et al., 2017).

3.2.2 Mention detection

The stage of mention detection is substantial for the purpose of data analysis and exploration. It is especially prominent for open personalised categories that cannot be described with a straightforward definition and are rather defined via a seed list like, for example, lists of musical tags in the specialised music datasets such as MSD². Due to the category openness, LLMs cannot be applied for the verification in the manner described above to create a dataset for end-to-end disambiguation. Instead, MD stage can be performed using generic concept

² <http://millionsongdataset.com/lastfm/>

extraction models like UPF’s pointer-generator-based model that we successfully applied in the MindSpaces project³ for public discourse analysis in the domain of interior and urban design (Stathopoulos et al., 2022).

In ReSilence, we design a more specific extraction technique that narrows the scope of the concepts to be identified by a seed list. Our aim is to train a generative model that generalises well over a predefined seed list and outcomes concepts that are more likely to be in a possible extension of the list. With this model, we would also like to account for the overlapping and discontinuous multi-word concepts that create a problem in applying traditional sequence labelling-based extractors. In this respect, we opt for deep neural generative models like T5 (Raffel et al., 2020) that are capable of translating a sentence into an arbitrary sequence of words. In our case, the model output contains concept mentions in the form appearing in a sentence one after another with a separation symbol in between.

For English, we work with two versions of the pretrained transformer T5, *T5-small* and *T5-large*. For languages other than English, mT5 shall be used as it allows for cross-lingual transfer learning. We also fine-tune UPF’s pointer-generator-based concept extraction model CE-PGN (Shvets and Wanner, 2020) as an alternative. Originally, this model was designed to extract concepts mainly in a form of nominal phrases which is not an exclusive form for the concepts we consider in the Project (as, e.g., with music genres that might be named only with an adjective). Still, we assume that tuning it on relevant data could extend its behaviour.

3.2.3 Neural word sense disambiguation

The entities extracted with a mention detection technique, even if limited by a seed list, can be of various categories. This may lead to not well-distinguishable senses, in contrast to senses of entities detected by our entity disambiguation model for a single category (like music genre). In this case, non-contextual word-embedding models are going to work poorly for a sense-matching step, especially for out-of-vocabulary unmapped entities. Therefore, we opt for using contextual word-embedding models as a more general approach.

Within the first period of the Project, we were experimenting with a contextual Transformer-based model BERT (Devlin et al., 2019). In particular, we attempted to study how much cluster-ability of embeddings is changed after a model is fine-tuned on a task in lexical item labelling. We explored the model that we trained to detect lexical collocations of lexical functions representing various semantic classes (Shvets and Wanner, 2022).

We considered embeddings of collocation bases, i.e., items less bounded by the semantics of the lexical function they are considered within. Two lexical functions that correspond to the “intense” and “minor” relations between bases and collocates were selected for general comparison of vectors before and after fine-tuning of the model. We visualise vectors in 2D space using the t-SNE algorithm (van der Maaten and Hinton, 2008) as shown in Figures 6 and 7 (“C-0” and “C-1” signify “intense” and “minor” relations correspondingly).

³ <https://cordis.europa.eu/project/id/825079>

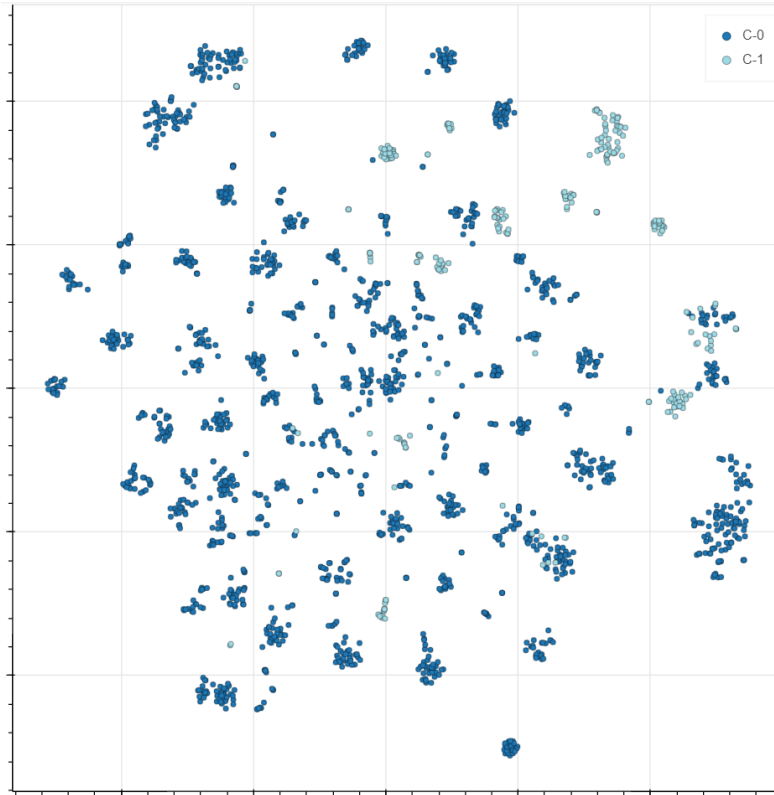


Figure 6: Contextualised embeddings of BERT

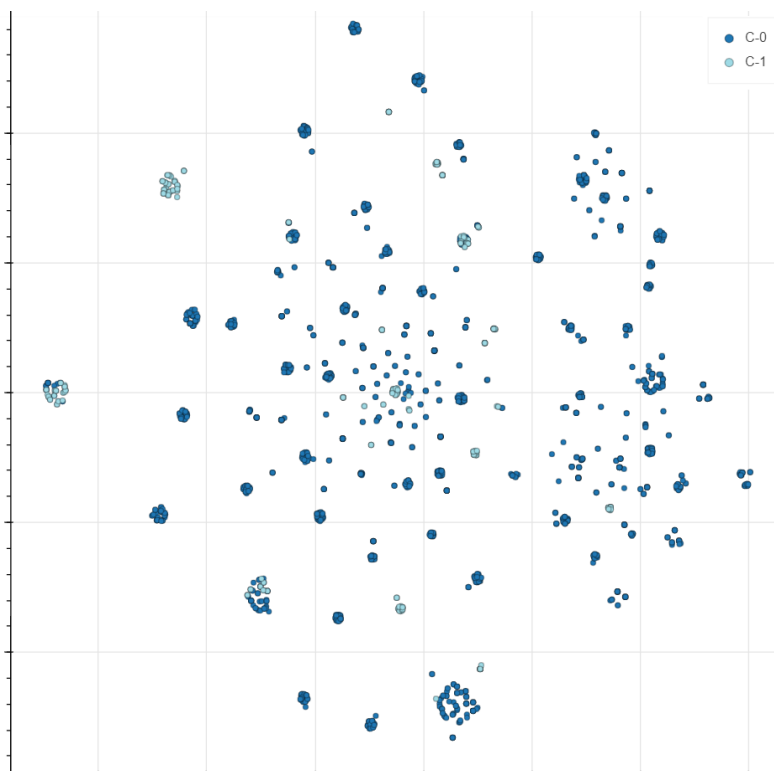


Figure 7: Contextualised embeddings of a fine-tuned BERT

We observe that with the original BERT (Figure 6), bases of the same semantics are located nearby and, depending on the context, situated in slightly different positions forming rather fuzzy clusters. Depending on the relation, a base of the same sense could appear in two

clusters. In contrast to this, embeddings of bases encoded by a BERT fine-tuned on lexical function identification task, became more distinct and do not depend on the relation (Figure 7), i.e. they appear in a single cluster while clusters become very fine-grained, dedicated to a specific base. Senses for bases of similar semantics are still located nearby but are not mixed together anymore. This means that the overall “coarse-grained” semantics were not changed after the fine-tuning but new embeddings distribute senses in well-separated discrete clusters. This property may ease the task of word sense disambiguation for open categories. We will continue exploring the approach of detecting senses by the similarity of their vector representations and make proper evaluations within the course of the Project.

3.2.4 Sentiment analysis

Entity-oriented sentiment analysis requires recognition of emotion and detection of an aspect this emotion is expressed about. An aspect can be detected as an individual entity in a text or a more general class of entities. Should a plethora of individual entities be opinionated within a given corpus, their grouping will provide more observability to data and allow for a more comprehensive analysis. In ReSilence, we explore an unsupervised way of maintaining open classes for both emotions and aspects, so that any not pre-modelled emotion-aspect pair could be automatically caught without a need for retraining.

We conducted a pilot study using 20,000 comments on music experiences provided by the users of YouTube. The comments were collected from pages of videos of a concert of the Chicago Symphony Orchestra⁴ and a concert of the group Coldplay⁵ by CERTH with their crawling techniques. We further indexed crawled messages making the search capabilities available within the Solr framework⁶ hosted on a UPF’s server dedicated to ReSilence. In order to come up with an idea of how diverse the discussed aspects are, we performed data clustering with further automatic naming of the resulting clusters. For this, we encoded every comment using Transformer-based sentence embeddings⁷, compared them pairwise using cosine distance as a metric (from 0 to 1), and created a graph where each node corresponded to a single comment and an edge was established only between nodes being in a small cosine distance under a defined threshold (we used a threshold of 0.3 to have a sparser graph). We then applied a graph clustering algorithm to decompose the graph (Blondel et al., 2008).

3.2.5 Functionality test generation

Tasks on fine-grained classification such as sentiment analysis suffer from biases incurred in data sampling, annotation, and model pre-training. Therefore, measuring the averaged performance over all examples in held-out test data is inadequate. Instead, we must identify specific model weaknesses and be informed when it is more likely to fail. For this, we explored a way to generate diverse and realistic functional tests by instructing large language models. Our study was carried out within a task of hate speech (HS) detection which is similar by nature to sentiment detection facing similar challenges. The results were discussed in a full paper that was accepted and will be presented at LREC-COLING 2024 (Jin et al., 2024). We

⁴ <https://www.youtube.com/watch?v=rOjHhS5MtvA>

⁵ <https://www.youtube.com/watch?v=Fpn1imb9qZg>

⁶ <https://solr.apache.org/>

⁷ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

believe that conclusions drawn can be applied to the task of sentiment analysis.

High accuracy on available benchmark datasets does not warrant that a classification model can detect representatives of a concerned class successfully in real-world data, especially when applied to challenging functionalities (such as “*implicit HS/emotions*”, “*hate/emotion phrased as a question*”, “*hate/emotion expressed using negated positive statements*”). To address the issue, Röttger et al. (2021) introduced HateCheck, a comprehensive suite of functional tests with handcrafted examples that cover various model “functionalities”. Each functionality captures a specific kind of hate speech, e.g., “hate expressed using slurs” (a corresponding example is “You are just a [slur] to me.”).⁸ While HateCheck provides important diagnostic insights, it suffers from simplistic sentence structures in examples, which deviate from the style of online language and may make models struggle when encountering novel aspects not covered in the training data.

To address these limitations, we propose GPT-HateCheck, a framework to generate functionality tests using large language models. We handcraft prompts to instruct an LLM to generate test cases corresponding to the functionalities in HateCheck. Furthermore, we employ a natural language inference (NLI) model (Williams et al., 2018; Yin et al., 2019) to validate that the generated test cases correspond to the gold-standard labels and the intended functionalities to be tested.

We use GPT-3.5 (Ouyang et al., 2022) as the LLM to generate test cases by invoking OpenAI’s API.⁹ To set the context, we provide the following system prompt, mentioning that the generated messages are for research purposes so that GPT will not reject the request due to violation of the usage policies¹⁰:

“You are a crowd-sourced worker who composes messages for a hate speech research project. Write messages with diverse styles and lengths.”

We use a prompt factory method with templates (shown in Figure 8) and alter only the instruction fragments specific to each functionality in order to speed up the prompt creation process.

```

1 def prompt_factory(instruction, n=10, demonstration=None):
2     if demonstration:
3         demo_str = f"Example: ``{demonstration}``\n"
4     else:
5         demo_str = ""
6     return f"""
7         Write {n} messages that {instruction}.
8         {demo_str}
9         Return the result in JSON list format with
10        the following key: messages.
11        Result:
12        """

```

Figure 8: The prompt factory method that instantiates prompt templates w/ or w/o demonstrations

⁸ We don’t list here the entire list of functionalities in HS as this goes beyond the scope of D4.1.

⁹ gpt-3.5-turbo-0613 checkpoint; <https://platform.openai.com/docs/models/gpt-3-5>.

¹⁰ <https://openai.com/policies/usage-policies>

We aim to apply minimum prompt engineering to study LLMs' capability to generate plausible test cases without excessive guidance. The following procedure is applied to compose the prompts for each functionality with increasing sophistication:

- a) Use only the functionality definition provided in HateCheck.
- b) Rephrase the instruction so that the LLM understands it better.
- c) Use the instruction and one demonstration randomly drawn from HateCheck dataset.
- d) For the functionalities involving slur words, additionally provide the list of slur words from HateCheck in the prompt.

The following list shows example instruction segments:

- 1) "express strong negative emotions against [identity]";
- 2) "reclaim one of the following slurs referring to [identity] to show endearment in a non-hateful context. Slurs: '[slur], [slur], [slur]'";
- 3) "1. each contain two clauses 2. the second clause contains hate expressions against [identity]";
- 4) "contain factual statements about [identity] with neither positive or negative sentiment".

These segments serve as input to the prompt factory method in Figure 8 to form the complete prompt templates. The first instruction is taken directly from the definition of the corresponding functionality "*Expression of strong negative emotions (explicit)*". The second provides slur words from HateCheck in addition. The other two examples require rephrasing to clarify the discourse structure ("*Hate expressed through reference in subsequent clauses*") or the target sentiment ("*Neutral statements using protected group identifiers*").

LLMs may not always follow the instructions strictly (Ouyang et al., 2022). For the generated test cases to be usable, we must ensure that the examples (i) agree with the gold labels, and (ii) are targeted to the specified functionalities. We were inspired by Goldzycher and Schneider (2022)'s proposal to use an NLI model (Yin et al., 2019) to improve zero-shot HS detection. They compose supporting hypotheses $[h_1, \dots, h_n]$ for four strategies: (1) filtering by target group identity, (2) filtering by counter speech, (3) filtering reclaimed slurs, and (4) catching dehumanisation comparisons. Then, they test each hypothesis independently using an NLI model (Yin et al., 2019). The predictions are combined using predefined rules to assign a label $y = \{1, 0\}$ to each input text, where '1' denotes *hateful* and '0' denotes *non-hateful*. Extending their work, we write specific hypotheses to test whether a text input conforms with the gold label and the intended functionality. We use the *bart-large-mnli* model from the transformers library¹¹, which predicts one of the following labels: *entailment*, *neutral*, and *contradiction*. We follow Goldzycher and Schneider (2022) to perform a softmax over the logits of *entailment* and *contradiction*. Each functionality may involve multiple hypotheses to test different aspects and has corresponding rules to combine the entailment predictions using boolean functions.

¹¹ <https://huggingface.co/facebook/bart-large-mnli>

3.3 Datasets

3.3.1 Entity disambiguation

In order to create examples of how music genres can be mentioned in various contexts and to cover as many possible genres as possible we went through music-related lists of Wikipedia articles and picked up for our corpus those that include many items and provide a diversity of topics as a whole. The selected lists with the number of articles they contain are provided in Table 1.

Name	# articles	Name	# articles	Name	# articles
Lead vocalists	3,102	Greek musical artists ¹²	129	Music festivals in the United Kingdom	239
Singer-songwriters	2,897	Greek musical groups	14	Music festivals in Canada	133
Composers by name	4,956	Spanish musicians	202	Music festivals in the United States	211
Principal conductors by orchestra	256	Bands from Spain	196	Music festivals in Australia	100
Symphony orchestras	227	People from Italy (Musicians)	406	Canadian musicians	1,498
Pops orchestra	24	Italian musical groups	87	Canadian composers	446
Symphony orchestras in Europe	429	French musical groups	61	Bands from Canada	2,087
German musicians	216	Music festivals	418	Symphony orchestras in the United States	285
German musical groups	175	Music festivals in Italy	28	TOTAL¹³	17381

Table 1: Selected music-related lists provided by Wikipedia

We collected and processed texts of the articles to split them into sentences and individual tokens using spacy-udpipe¹⁴. We further selected sentences that include the same tokens as positioned in names of musical genres. The extensive lists of music genre names that we used

¹² Country-related lists include some articles that are missed in generic lists of the same type.

¹³ The number of articles 17381 is less than a sum of list sizes 18822 as some lists include the same articles.

¹⁴ <https://spacy.io/universe/project/spacy-udpipe>

are maintained on Wikipedia^{15,16,17,18}. We found more than 1000 out of 1400 names in our corpus that confirms the diversity we aimed to reach.

We further verified genre names found in the sentences with the LLM (see details in Section 5.2) that gave us a silver annotation for the dataset. In total, we resulted in 188,854 positive mentions and 95,181 negative mentions (the ratio is 66.5% and 33.5% correspondingly) located in 216,912 sentences. Sentences with positive mentions contain 1.28 ($\sigma = 0.76$) positive mentions on average. The outlier examples contain up to 49 mentions, e.g.:

*“Across their career the band has also been said to play within the genres [alternative metal](#), [alternative rock](#), [pop rock](#), [electronic rock](#), [hard rock](#), [heavy metal](#), [post-hardcore](#), [pop metal](#), [pop](#), [nu metal](#), [electropop](#), [hip hop](#), [EDM](#), [arena rock](#), [melodic metalcore](#), [electronicore](#), [electronica](#), [screamo](#), [hardcore punk](#), [technical metal](#), and [emo](#).”*¹⁹

As already mentioned in Section 5.2, in order to train the model in a contrastive manner, we include in our dataset examples with occurrences of both positive and negative classes for the same entry (“B”/“I” vs. “O”). However, we remove examples that exclusively contain mentions of entries for which we observe less than 1% of positive cases in the dataset (among them such genre mentions as *march*, *were*, *church*, *early*²⁰). Erroneously, this may lead to an incorrect balance for such mentions in case their tokens still appear in examples with other entries kept. Therefore, we assign negative labels to tokens corresponding to excluded entry mentions regardless of the correctness of the class. This prevents the model from observing positive occurrences for which some sentences with corresponding negative occurrences were removed from the dataset. Instead, these words will be treated as neutral as any other word always tagged with “O”. In total, the resulting dataset contains 406 out of 1027 genres mentioned with a negative class.

We split our dataset 80/10/10 to create training, validation, and test subsets correspondingly. The statistics over the dataset is provided in Table 2.

	Size	# pos mentions	# neg mentions	# true genres	# false genres
Training	173,528	151,088	76,394	999	386
Validation	21,691	18,774	9,376	655	229
Test	21,693	18,992	9,411	642	224
TOTAL	216,912	188,854	95,181	1027	406

Table 2: Statistics for the genre-disambiguation dataset

For clarity, we provide Figures 9 and 10 that show the number of positive mentions for genres

¹⁵ https://en.wikipedia.org/wiki/List_of_styles_of_music:_A-F

¹⁶ https://en.wikipedia.org/wiki/List_of_styles_of_music:_G-M

¹⁷ https://en.wikipedia.org/wiki/List_of_styles_of_music:_N-R

¹⁸ https://en.wikipedia.org/wiki/List_of_styles_of_music:_S-Z

¹⁹ The sentence is from https://en.wikipedia.org/wiki/Bring_Me_the_Horizon.

²⁰ Other mentions for these genres like “church music” and “early music” remain labelled in the dataset.

in the training set.

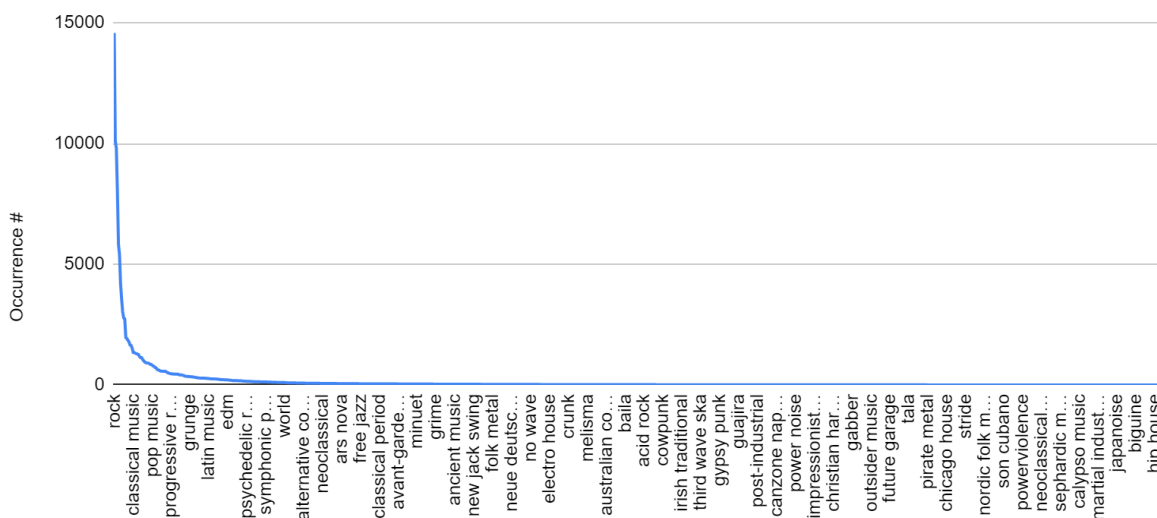


Figure 9: Number of positive mentions for genres in the training set

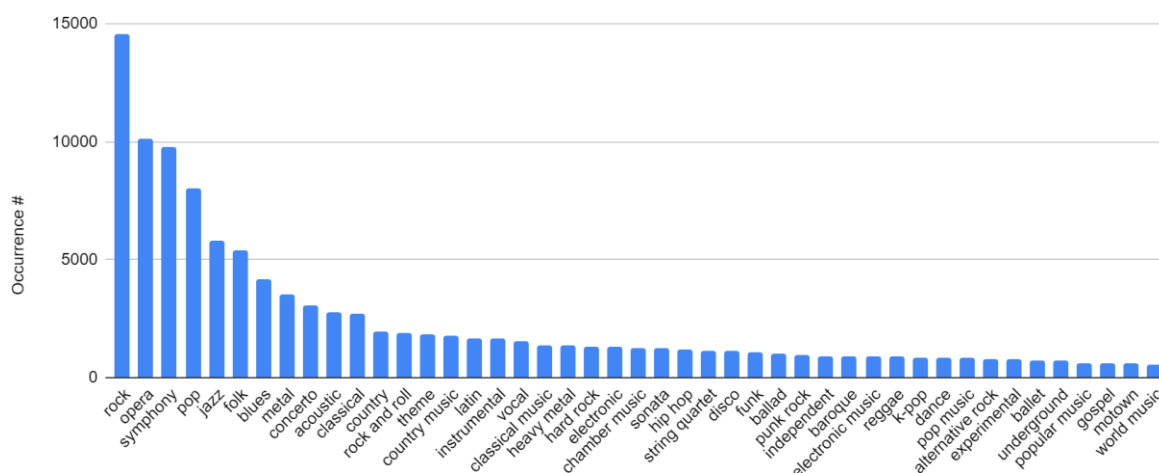


Figure 10: Number of positive mentions for top-frequent genres in the training set

3.3.2 Mention detection

As this task requires analysis at a lexical level, we first carried out a pilot study with an expert in lexicography, in particular, an expert in the domain of green chemistry. This drift in the domain does not affect our conclusions about the appropriateness of the approach, and, therefore, further we will take advantage of its portability to other domains and apply to the data collected within ReSilence. The results of this study were published and presented by us at the *2nd International Conference on Multilingual Digital Terminology Today, MDTT 2023* (Gotkova and Shvets, 2023).

As a result of continuous sampling of environmental lexical material²¹, we compiled the initial seed list of 268 unique environment-related keywords. These keywords were further divided

²¹ The process included both manual and automatic selection and was partly done in collaboration with an expert in green chemistry and an expert in lexicology (Gotkova and Chepurnykh, 2022).

into two categories. We selected 104 keywords which we see as core-candidates, i.e., similar to (Drouin et al., 2018), keywords which may potentially be validated as core environmental terms which cut across the entire field of the environment (such as *carbon*, *climate*, *global warming*, *greenhouse gas*); further in the text called *CKeywords*. The supervised models are expected to expand this list. The remaining 164 keywords were categorised as supplementary; further in the text called *SKeywords*. We considered the following keywords as supplementary: complex keywords built with Ckeywords (*air pollution*, *anthropogenic carbon dioxide*, *atmospheric warming*) and keywords which do not refer to the most essential domain concepts but still exhibit strong semantic connections with other domain-related terms and therefore are nevertheless important for domain discourse (*ice*, *Earth*). None of the lists contains a lexicon very specific to specialised environmental discourse.

Our specialised corpus is a monolingual English domain-specific corpus composed of 44 reports issued by international environmental organisations such as European Environmental Agency, Intergovernmental Panel on Climate Change, United Nations Environmental Program and World Meteorological Organization. These reports give a comprehensive overview of the current and emerging environmental issues. We converted documents to plain text excluding figures and tables and manually cleaned artefacts remained after the conversion. Consider an example of a sentence with Ckeywords given in bold and Skeywords underlined:

Moreover, the degradation of wetlands releases stored carbon, fuelling **climate change**.

We designed a simple procedure to annotate the entire corpus of about 30K sentences to have enough data samples for training a neural network. The first step is to parse the corpus using UDPipe²², while the second is to consider all the sequences of tokens of lengths from one to the maximum number of words in keywords in our lists taking the normal forms of lexical items using their lemmas, and looking them up (with conditions on part-of-speech tags) in the lists of keywords which we automatically expanded with alternatives beforehand (e.g., for *biodiversity conservation* we added *conservation of the biodiversity*, for *bio-based – biobased*, etc.). Finally, each sentence with the corresponding found items made a single data sample. The obtained samples cover 103 out of 104 Ckeywords (*carbon-free* was not found in this corpus) and, in total, 255 out of 268 keywords. The search procedure took into account many possible occurrences including the cases of overlapping and discontinuous keywords such as *soil pollution* and *air pollution* in a phrase “*soil and air pollution*”.

Resulting samples were shuffled and split into the training, development (dev), and test subsets in the proportion 80/10/10. We performed shuffling several times until the examples were distributed among the subsets in such a way that only 80% of the keywords are used for training (they also appear in two other subsets), while other 10% and 10% are used exclusively in the dev and test subsets without intersections.²³ We preserve these 20% of keywords to assess the ability of the model to extract “new” keywords unseen during the training. We leverage the dev set to select the most prominent intermediate states of the model obtained during the training, and the test set – for the final evaluation. Sentences without keywords were also added proportionally to the subsets to guide the model when it should not extract anything.

²² <https://ufal.mff.cuni.cz/udpipe>

²³ A couple of thousands of samples were removed from the dataset to meet the condition of exclusiveness.

In addition to our simple automatic annotation, we manually selected and examined 200 sentences from the corpus (excluded from the subsets) and created fully annotated samples (with some keywords beyond the existing lists) that we refer to as a gold standard. The size of the overall dataset is shown in Table 3.

	Ckey + Skey	Ckey	Ckey new	# pos	# neg
Training	24,565 (206)	17,618 (80)	-	10,301	8,803
Dev	3,711 (184)	2,723 (83)	183 (9)	1,449	1,149
Test	3,703 (172)	2,737 (80)	183 (9)	1,505	1,117
Gold	592 (238)	192 (37)	35 (9)	100	100

Table 3: Statistics over the dataset. Ckey+Skey/Ckey/Ckey new – number of occurrences in a subset (number of unique keywords in parenthesis); # pos, # neg – number of samples w/ and w/o keywords

3.3.3 Sentiment analysis

We conducted a pilot study using comments on music experiences provided by the users of YouTube on pages of videos of a concert of the Chicago Symphony Orchestra²⁴ and a concert of the group Coldplay²⁵. The data are presented by the CERTH’s crawler in the following JSON format (Figure 11):

```
[
  {
    "video_metadata": {
      "video_id": "",
      "video_title": "",
      "video_meta": {
        "publishedAt": "",
        "channelId": "",
        "title": "",
        "description": "",
        "channelTitle": "",
        "tags": [],
        "categoryId": "",
        "liveBroadcastContent": ""
      }
    },
    "video_comments": [
      {
        "comment_text": "",
        "timestamp": "",
        "likes": 0,
        "replies": []
      },
      ...,
      {
        "comment_text": "",
        "timestamp": "",
        "likes": 0,
        "replies": []
      }
    ]
  }
]
```

Figure 11: The structure of the crawled data

²⁴ <https://www.youtube.com/watch?v=rOjHhS5MtvA>

²⁵ <https://www.youtube.com/watch?v=Fpn1imb9qZg>

We extracted messages from the “comment_text” fields and encoded them with sentence embeddings²⁶ resulting in 10,932 vectors for the first video and 9,333 vectors for the second video. We used all the messages across these two subsets for graph construction and further decomposition to come up with topics and compared them to each other to study similarities and differences in opinions on two different types of music.

3.3.4 Functionality test generation

The generated dataset GPT-HateCheck that we evaluate consists of 40 examples for each (target group, functionality) pair. We use the same list of target groups as HateCheck (Röttger et al., 2021). Targets are not applied to the domain of ReSilence, therefore, in the future, the dataset for sentiment analysis will be either target-independent or various performance aspects will play a role of targets, depending on the priorities within the Project. Table 4 presents the number of examples for HateCheck, GPT-HateCheck and the candidates generated by GPT before filtering.

Target Group ²⁷	HC	GPT-HC	-Filter
Women	509	606	824
Trans ppl.	463	611	795
Gay ppl.	551	646	822
Black ppl.	482	741	812
Disabled ppl.	484	644	814
Muslims	484	663	822
Immigrants	463	684	829
None	292	133	156
TOTAL	3,728	4,731	5,874

Table 4: Number of examples for each target group in different datasets (HC: HateCheck; GPT-HC: GPT-HateCheck). “-Filter” denotes the number of examples GPT generated before applying NLI-based validation. “None” denotes functionalities that do not target any protected group

3.4 Evaluation

3.4.1 Entity disambiguation

We trained two transformer-based models on a genre-annotated dataset, specifically the

²⁶ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

²⁷ As noted above, targets in this pilot study do not apply to the ReSilence domain. They may be substituted with aspects of music performances without a need for adjustment of the overall approach.

base version of RoBERTa²⁸ and a large version of BERT²⁹, both for five epochs. We compare them to a baseline model which is a greedy dictionary-based labeller that was used to create initial data annotation before it was refined with Mistral³⁰.

Table 5 provides evaluation scores: *micro average scores* demonstrate the ability of the models to discover a music genre in the text in general, regardless of a particular genre, while *weighted average scores* are calculated by us by going through individual genres and averaging across them taking their frequencies as weights. Due to the fact that there is a pool of frequent genres that are easy to detect, the latter scores turn higher. However, there is a long tail of other genres that appear rare individually but very frequent as a whole (see Figure 9), therefore *micro average scores* provide a more realistic picture of what users will experience in practice.

	Micro average			Weighted average		
	<i>Precision</i>	<i>Recall</i>	<i>F₁-score</i>	<i>Precision</i>	<i>Recall</i>	<i>F₁-score</i>
Baseline	0.67	1.00	0.80	0.88	1.00	0.92
RoBERTa-base	0.92	0.96	0.94	0.93	0.96	0.94
BERT-large-uncased	0.92	0.96	0.94	0.93	0.96	0.94

Table 5: Evaluation scores for music genre identification problem

According to the scores, fine-tuned transformer models perform equally well, both outperforming the challenging baseline in F_1 -score by 15% on micro average and by 2% on weighted average. This reduces the error rate by 70% and 25% respectively.

Table 6 contains statistics over genres that are the most problematic for the dictionary-based approach, i.e., those that gain more ambiguity due to the variety of senses their naming words might possess. We selected only those entries that appear in the test dataset more than ten times. Fine-tuned transformer model achieves precision of 0.64 on average for these cases which is 2.2 times higher than of the baseline (35 percentage points higher). This results in an average 1.57 times rise of the F_1 -score.³¹

<i>Genre</i>	Baseline			BERT-large-uncased			#
	<i>Precision</i>	<i>Recall</i>	<i>F₁-score</i>	<i>Precision</i>	<i>Recall</i>	<i>F₁-score</i>	
country	0.28	1.00	0.44	0.68	0.76	0.72	246
theme	0.50	0.93	0.65	0.76	0.92	0.83	225
vocal	0.34	1.00	0.51	0.69	0.75	0.72	177
independent	0.31	0.81	0.44	0.66	0.73	0.70	116
dance	0.17	1.00	0.29	0.71	0.68	0.69	108

²⁸ <https://huggingface.co/FacebookAI/roberta-base>

²⁹ <https://huggingface.co/google-bert/bert-large-uncased>

³⁰ <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>

³¹ Since both models RoBERTa and BERT performed similarly, we provide scores only for one of them.

ballet	0.43	1.00	0.60	0.68	0.84	0.75	93
beat	0.42	1.00	0.59	0.68	0.76	0.72	70
suite	0.38	1.00	0.55	0.65	0.66	0.65	53
soul	0.12	1.00	0.21	0.48	0.63	0.55	46
melodic	0.44	1.00	0.61	0.59	0.64	0.62	45
chamber	0.13	1.00	0.23	0.64	0.55	0.59	38
christian	0.13	1.00	0.23	0.77	0.71	0.74	34
progressive	0.46	1.00	0.63	0.71	0.86	0.78	29
mass	0.19	1.00	0.32	0.57	0.62	0.59	26
string	0.08	1.00	0.15	0.39	0.50	0.44	22
world	0.01	1.00	0.02	0.56	0.47	0.51	19
noise	0.17	1.00	0.30	0.54	0.39	0.45	18
cabaret	0.45	1.00	0.62	0.59	0.76	0.67	17
medieval	0.40	1.00	0.58	0.69	0.65	0.67	17
surf	0.36	1.00	0.53	0.75	0.75	0.75	12
chant	0.46	1.00	0.63	0.57	0.73	0.64	11
children's	0.07	1.00	0.14	0.75	0.82	0.78	11
AVERAGE	0.29	0.99	0.42	0.64	0.69	0.66	65

Table 6: Evaluation scores for the most ambiguous music genre names

Table 7 contains the most frequent genres in the test set. It allows us to check that the model not only overcomes the baseline in general including complex cases but also does not decrease performance for the most popular music genres. In fact, we observe the increase in performance even though the baseline is tough-to-beat for these very frequent genres most of which are unambiguous.

Genre	Baseline			BERT-large-uncased			#
	Precision	Recall	F_1 -score	Precision	Recall	F_1 -score	
rock	0.91	1.00	0.95	0.94	0.96	0.95	1893
opera	0.98	1.00	0.99	0.98	0.99	0.99	1286
symphony	0.89	1.00	0.94	0.92	0.97	0.94	1252
pop	0.84	1.00	0.91	0.90	0.96	0.93	987
jazz	0.99	1.00	1.00	0.99	1.00	0.99	702
folk	0.99	1.00	0.99	0.99	0.99	0.99	678
blues	0.98	1.00	0.99	0.98	1.00	0.99	516

metal	0.94	1.00	0.97	0.96	0.98	0.97	438
concerto	0.90	1.00	0.95	0.93	0.99	0.96	354
acoustic	0.96	1.00	0.98	0.97	0.99	0.98	340
classical	0.97	1.00	0.99	0.98	0.98	0.98	310
country	0.28	1.00	0.44	0.68	0.76	0.72	246
country music	0.99	1.00	0.99	0.99	1.00	0.99	236
theme	0.50	0.93	0.65	0.76	0.92	0.83	225
rock and roll	1.00	1.00	1.00	1.00	1.00	1.00	221
latin	0.65	1.00	0.79	0.84	0.94	0.89	221
instrumental	0.73	1.00	0.84	0.86	0.93	0.89	216
hard rock	0.98	1.00	0.99	0.99	1.00	0.99	179
vocal	0.34	1.00	0.51	0.69	0.75	0.72	177
heavy metal	0.99	1.00	1.00	0.99	0.99	0.99	177
disco	0.93	1.00	0.96	0.93	0.98	0.96	168
chamber music	1.00	1.00	1.00	1.00	1.00	1.00	166
electronic	0.83	0.99	0.90	0.92	0.96	0.94	164
classical music	0.98	1.00	0.99	0.98	0.99	0.99	157
AVERAGE	0.85	1.00	0.90	0.92	0.96	0.94	370

Table 7: Evaluation scores for the most frequent music genre names

In summary, we can see that *silver* examples of the training set taken from the same domain in a contrastive way allowed models to learn differences in rather similar contexts for the ambiguous music genres while keeping at a high level the performance for less ambiguous names. Our model overcomes the strong but not scalable baseline by 15% in F_1 -score on this dataset. For future experiments, we will create a *gold* annotation and evaluate more models to choose the one that better fits the needs of ReSilence.

3.4.2 Mention detection

We report on the precision $P = TP/(TP + FP)$ and recall $R = TP/N_P$ scores for different types of keywords (*Skeywords*, *Ckeywords*, and *Ckeywords new* – *Ckeywords* unseen during the training) in Tables 8-10 where TP is the number of correctly extracted mentions of the scored type, FP – the number of extracted mentions out of all ground-truth mentions (FP does not depend on the type under scoring), and N_P – the number of ground-truth mentions of the scored type.

	Development set					
	Ckeywords + Skeywords		Ckeywords		Ckeywords new	
	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>
CE-PGN	5.96	31.66	4.27	30.41	0.36	36.61
CE-PGN tuned	94.94	87.47	93.18	86.82	0.57	0.54
T5-small tuned	84.64	85.85	80.73	88.91	12.95	46.99
T5-large tuned	88.13	91.05	84.90	93.94	22.09	70.49
T5-large c-tuned ³²	93.45	69.23	93.41	93.65	29.96	42.08

Table 8: Results breakdown for mention detection (development set)

	Test set					
	Ckeywords + Skeywords		Ckeywords		Ckeywords new	
	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>
CE-PGN	6.06	32.40	4.55	32.37	0.50	50.82
CE-PGN tuned	95.11	85.53	93.41	84.40	2.40	2.19
T5-small tuned	89.60	83.53	86.89	86.92	22.46	56.83
T5-large tuned	94.50	92.28	92.80	93.68	32.08	51.36
T5-large c-tuned	93.02	67.65	93.02	91.49	26.27	36.61

Table 9: Results breakdown for mention detection (test set)

	Gold set					
	Ckeywords + Skeywords		Ckeywords		Ckeywords new	
	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>
CE-PGN	25.73	36.99	9.71	35.42	2.62	48.57
CE-PGN tuned	81.23	51.18	67.29	75.00	4.11	8.57

³² The model was tuned on the same training set but annotated only with Ckeywords.

T5-small tuned	82.45	52.36	71.79	87.50	20.48	48.57
T5-large tuned	82.82	59.46	70.68	91.67	24.74	68.57
T5-large c-tuned	85.31	46.11	78.64	90.10	32.86	65.71

Table 10: Results breakdown for mention detection (gold set)

As expected, the original UPF’s CE-PGN model extracts a small number of keywords with a very low precision as it tends to find all the concepts independently of the domain. The fine-tuning re-oriented it towards the environmental domain – both scores were significantly improved. However, it performed poorly on extracting unseen keywords.

T5-large performed better than other models apart from when the small version gained a slightly higher recall score on the unseen Ckeywords of the test set. Interestingly, the annotation with Skeywords helped to detect Ckeywords better. The model that was trained only to extract Ckeywords (T5-large c-tuned) generalised poorer and missed many more unseen Ckeywords.

For the quality check of the extraction results, we manually checked 171 non-annotated keywords extracted from dev set using T5-large. As a result, 70 novel keywords (41%) were obtained, other 32 (19%) corresponded to existing keywords missing in automatic annotation due to mistakes of the parser, and only the rest 69 (40%) were false negatives, i.e., not keywords. 45 keywords out of the 70 novels were combinations of already existing keywords in our lists, the remaining 25 keywords were new to us. This result is linguistically valuable for us: all 25 new keywords are pertinent to the environmental topic. Although some keywords are too specific, all of them are considered as an important addition.

Results of our experiments provided several valuable insights as regards both linguistics and information extraction areas. First, tests performed with T5-large demonstrated that supplementary lexical material (Skeywords) enhanced the model’s ability to detect Ckeywords. Therefore, as the list of Ckeywords used to train the model grows, it is necessary to add to the list of Skeywords as well. Second, we consider it now important to increase the number of manually annotated samples to improve the gold standard dataset and this would allow us to train the model on annotated data of high quality in addition to automatically annotated sets. We will take this into account when we develop the dataset in the domain of the Project. Third, T5-large model proved efficient for extracting unseen keywords: it detected 50-70% of them in a set (62% across all the evaluation sets). Finally, the models extracted many novel keywords pertinent to the chosen topic which were not present in the preselected keyword list, i.e., they allow for dynamic in-domain knowledge enhancement.

3.4.3 Sentiment analysis

Following the methodology described in Section 5.2 applied to comments of YouTube users, we obtained about 400 subgraphs among which about 100 subgraphs included at least 5 nodes (comments). In order to name the subgraph that represents an aspect, we applied the GPT-3.5 LLM³³ with the following prompt: *“Characterize shortly (in several words) the topic of the following group of comments. The total number of comments in this group: <number>”*:

³³ OpenAI: Gpt-4 technical report (2023)

<comments>”, where <number> and <comments> are placeholders for the size of a subgraph and up to 10 random comments from the subgraph correspondingly. Examples of generated outcomes are provided below:

- “Observations and comments on the audience’s behaviour during the performance of Beethoven’s Symphony No. 9, specifically regarding the absence of applause between movements, humorous remarks about continuous clapping, and expressions of appreciation for the uninterrupted flow of the music.”
- “Praising the concert experience and expressing admiration for Coldplay’s performance, particularly in Argentina. Commenters emphasise the energy, enthusiasm, and passion of the Argentine crowd and highlight the visual spectacle of the concert. Some comments also acknowledge the economic challenges faced by the audience but still express appreciation for the music and the overall experience.”
- “Appreciation and admiration for the drone shots and the drone operator during Coldplay’s performance. Comments highlight the impressive filming, the skill of the drone pilot, and the excitement generated by the footage captured from the drone.”
- “Curiosity and fascination regarding the synchronised wristbands worn during Coldplay’s performance. Commenters express amazement at the coordination and synchronisation of the light bracelets with the music, and inquire about the technology and mechanics behind their operation.”
- “Emotional reactions and appreciation for the music and performances, including comments on specific movements and moments.”
- “Specific timestamps, moments, and reactions related to the music and performances.”
- “Expressing intense pleasure and sensory satisfaction while listening to Beethoven’s Symphony No. 9, likening the experience to a “medicine for the ears”.”
- “Curiosity and inquiries regarding the female performers wearing colourful dresses (particularly pink and red) in the orchestra during the performance of Beethoven’s Symphony No. 9. Commenters express fascination with their presence and inquire about their identities and vocal abilities, while acknowledging their appreciation for the overall musical composition.”

We observe a wide coverage of various aspects that include the music itself, the audience, the technical support at performances, multimodal inputs (e.g., sound with visuals), specific moments, sensory experience, vocal abilities, etc.

Figures 12-14 contain examples of comments and a corresponding aspect/topic as described by the model (we manually highlighted in bold phrases that summarise the description). Figure 12 shows comments on the physical impact of the music for both concerts. It is possible to see cultural differences between audiences of Coldplay and Beethoven by the average length of messages on the same aspect. Beethoven’s public tend to express themselves in a more descriptive way, more open to sharing their thoughts. Figure 13 demonstrates two groups of comments similar to each other but still clustered by graph decomposition algorithm separately as they cover slightly different aspects. We can see that the LLM captured these details and generated distinct descriptions. We also include Figure 14 to show the advantage of using sentence embeddings for semantic representation of messages (instead of vocabulary overlap): the comments grouped not necessarily share the same lexicon but opinions, indeed, refer to the same aspect (in this case, recognising the genius of Beethoven as a composer and expressing gratitude for the ability of his music to inspire, uplift, and even provide solace in difficult times).

Topic: Expressing intense emotions and the physical reaction of getting goosebumps while listening to Coldplay and Beethoven's music. Commenters emphasize the **powerful impact of the music**, describing it as awe-inspiring, emotional, and **capable of evoking a strong physical response**. The **repeated mention of goosebumps** highlights the intensity and immersive nature of the musical experience.

{'Coldplay': 82, 'Beethoven 9': 32}

Coldplay: my goosebump has goosebumps and that goosebump has goosebumps aswell!	Beethoven 9: Goosebumps.
Coldplay: goosebumps	Beethoven 9: I have goosebumps all over my body and i have a teardrop in my eye. This is something truly unbelievable!
Coldplay: i was having goosebumps when i was watching this for the 1st time and i still get it everytime i watch this.	Beethoven 9: Goosebumps
Coldplay: If you don't get goosebumps from this get help !	Beethoven 9: Couple seconds in and i'm getting goosebumps from the video, i can only imagine how astonishing it sounds in person.
Coldplay: goosebumps	Beethoven 9: Gt goosebumps just by watching on phone wonder what its like to actually be there
Coldplay: I get goosebumps watching this, then what about in person 😭😭	Beethoven 9: Goosebumps head to toe.
Coldplay: Awesome emotions	Beethoven 9: Call me crazy, but I actually had goosebumps when I jumped right in at around 3 minutes into it
Coldplay: goosebumpsssss	Beethoven 9: At 2:33 , I get goosebumps everyone I listen to this
Coldplay: Goosebumps	Beethoven 9: I got goosebumps and amazed by the instrument, this instrument for me can calm the mind.
Coldplay: Wow Goosebumps	Beethoven 9: I listen to this every day while I work. It brings me such joy. Thank you for this amazing work of art! and for all the goosebumps! <3

Figure 12: Description of a cluster of comments grouped by semantic similarity

Topic: **Acknowledging the collaboration between Coldplay and Avicii** in a song. Expressing emotions and **appreciation for Avicii's musical contributions**. Mentioning the absence of Avicii's recognition in the performance.

{'Coldplay': 181}

Coldplay: A real masterpiece by Avicii and Coldplay!! ❤️
 Coldplay: Avicii would be happy to see this from the other side ☐
 Coldplay: I wish Avicii was there with them...
 Coldplay: that drop hits me with so many emotions rest in peace Avicii your music will live with us forever
 Coldplay: Sky full of fans (stars) 🌧️❤️
 Coldplay: Tim bergling's master piece ❤️
 Coldplay: AVICII X COLDPLAY = MASTERPIECE
 Coldplay: When I see this my heart is like 😭😭😭
 Coldplay: The musician is avicii ❤️,,and this lyrics and coldplay voice ❤️
 Coldplay: Sad to see he didn't even bother mentioning Avicii even though he wrote the song

Topic: Expressions of admiration and awe for the beauty and impact of Coldplay's performances. **Appreciation for the emotions evoked and the overall experience**.

{'Coldplay': 326}

Coldplay: this is just incredible
 Coldplay: This is beautiful, damn ... ❤️
 Coldplay: unreal
 Coldplay: i love you;))
 Coldplay: Thank you for being with me through my bests and worst. I love you forever.
 Coldplay: Wowwwww that was great to watch!!
 Coldplay: That just looks absolutely amazing ❤️
 Coldplay: Awesome as always
 Coldplay: Best of the best love you
 Coldplay: The most beautiful thing I have ever seen. Out of this world ❤️🎉

Figure 13: Descriptions of two clusters of comments grouped by semantic similarity

Topic: **Deep appreciation and admiration for Beethoven's 9th Symphony**, recognizing its significance, emotional impact, and the **genius of Beethoven as a composer**. Reflections on the symphony's themes, structure, and the **profound effect it has on listeners**. Expressing gratitude for Beethoven's artistry and the **ability of his music to inspire, uplift, and even provide solace in difficult times**.

{'Beethoven 9': 946}

Beethoven 9: Hearing the 9th for the first time caused a massive paradigm shift in my life, as it does for most people who hear it and truly appreciate it's significance. There's no going back once you hear this symphony in it's full glory. I'd argue that it's saved me from being engulfed entirely by the void of depression and for that I owe Beethoven more than I could ever offer as gratitude for his art. Everyone should be able to hear this work.
 Beethoven 9: "To play a wrong note is insignificant ; to play without passion is inexplicable " -- Ludwig van Beethoven
 Beethoven 9: very nicely done and Beethoven would be proud of all you and your hard working putting this all together
 Beethoven 9: Lord how much effort and time you needed to give us Beethoven
 Beethoven 9: people nowadays can't create a simple, 5 minute overture or lied, and these classical composers composed 60+ minutes long symphonies at the age of 7 while being deaf. People, 300-400 years ago were both blessed and cursed to break the barriers of civilization and innovation. I pity and admire them equally...
 Beethoven 9: Is the conductor always look like Beethoven, or he is just cosplaying?
 Beethoven 9: As a music lover, I cannot imagine anything more heartbreaking than being unable to hear it anymore. I do hope Beethoven got some joy out of knowing that his audiences loved (and still love) his music.
 Beethoven 9: When I was in college, our music department did a Beethoven marathon on his birthday doing all nine symphonies. It was quite an experience and we were exhausted but filled with joy at the end of the ninth. A night I remember 40 years later.
 Beethoven 9: The fact that Beethoven composed this after going deaf is beyond amazing 🤩

Figure 14: Description of a cluster of comments grouped by semantic similarity

The generated descriptions already include emotional feedback specifying rather fine-grained emotions such as surprise, curiosity, appreciation/admiration/excitement, boredom, nostalgia, frustration, disappointment and others. Thus, the identification of sentiments using a generic LLM can be a valid alternative to training models on specialised sentiment datasets

that may suffer from underrepresented data for rare emotions and contexts not related to the ReSilence domains. As a result of this pilot study, we conclude that our approach which includes a combination of neural text similarity assessment, graph clustering, and an LLM-based description generation for aspect and emotion identification satisfies the needs of the Project.

3.4.4 Functionality test generation

We addressed three research questions to prove the quality of the proposed approach to automated dataset generation: (RQ1) Which functionalities does GPT struggle to generate examples for? (RQ2) Can GPT generate diverse and natural test cases? (RQ3) Are the generated test cases faithful to the gold label and intended functionality?

To tackle the RQ1 question, we calculate each functionality’s NLI test passing rates. The result is presented in Figure 15. This result is in line with previous research on creating adversarial examples for HS detection (Hartvigsen et al., 2022; Ocampo et al., 2023), showing that some nuanced functionalities pose challenges to both HS detectors and text generators. We will pay attention to these differences in functionalities when we port them to the sentiment analysis task.

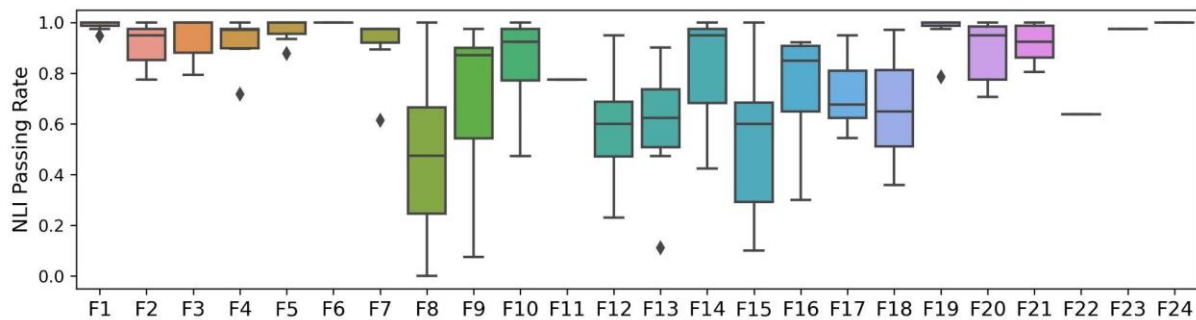


Figure 15: Functionality-wise NLI test passing rates

Addressing the RQ2 question, we conduct automatic evaluations to measure intra-example lexical diversity using self-BLEU (Zhu et al., 2018)³⁴ and naturalness using perplexity³⁵.

Since GPT-HateCheck contains more examples than HateCheck, we calculate the metrics for HateCheck using the entire dataset while drawing ten random subsamples from GPT-HateCheck with the same number of examples as HateCheck. Table 11 shows the averaged result.

³⁴ The Self-BLEU score of a dataset is calculated as the average BLEU score of each generated example using the rest of the examples as references. We report BLEU-2/3/4 scores.

³⁵ We use the *gpt2-large* model from HuggingFace library to calculate perplexity.

Dataset	self-BLEU			PPL
	n=2	n=3	n=4	
HC	0.937	0.863	0.761	67.47
GPT-HC	0.864 (1.2e3)	0.735 (2.2e3)	0.594 (2.6e3)	21.52 (.088)

Table 11: Result of self-BLEU scores to measure intra-example diversity (the lower the better) and perplexity to measure naturalness (the lower the better). The best results are highlighted in bold; the standard deviations are shown in brackets. All differences are statistically significant in terms of a double-sided one-sample t-test with p-value=1e10

We observe that the examples in GPTHateCheck have a higher lexical diversity than in HateCheck, the gap being larger for longer n-grams. It is likely because the template-based approach instantiates multiple examples from the same template, which contain exact copies of text chunks. Qualitatively, we can also observe from the samples that GPT-HateCheck contains novel aspects/arguments that are neither in HateCheck nor in the prompts. Furthermore, the examples in HateCheck have a much higher perplexity score, confirming that the template-based generation method is prone to producing rigid and unnatural examples.

To answer the RQ3 question, we select all 795 GPT-generated messages targeting trans people and 156 messages that do not target any protected group (cf. Section 5.3) to conduct human evaluation. We ran two separate annotation tasks indicating whether a message is hateful and consistent with the indicated functionality. We present the hateful and functionality consistency evaluation in Table 12.

Setting	Hateful	Func
GPT-HC	92.65%	88.57%
GPT-HC -filter	91.48%	83.28%

Table 12: Hatefulness and functionality consistency scores. The best scores are highlighted in bold

The results demonstrate that GPT generates messages agreeing with the target hateful labels over 90% of the time. However, the generations are more likely not to follow the intended functionalities, still reaching a high score over 88%. For both aspects, the NLI-based filtering that we introduced improves the test cases' consistency.

3.5 Future considerations

We plan to expand entity disambiguation and mention detection modules towards other music-related categories and cover more languages. For this we are going to look into the direction of transferability of the obtained extractive models to other domains and multilingual neural models such as mBERT (Devlin et al., 2019), XLM-Roberta (Conneau et al.,

2020), and mT5 (Xue et al., 2021). We will also explore the use of multilingual LLMs such as Mixtral (Jiang et al., 2024) for the creation of silver annotation for other ReSilence languages.

The proposed technique for neural word sense disambiguation will be consolidated and tested on the materials pertinent to the ReSilence domain.

As for aspect-oriented sentiment analysis, in the course of the Project, we will experiment with publicly shared LLMs, develop techniques for extracting and categorising sentiments from the generated descriptions and perform an evaluation of the proposed approach on a large corpus.

Furthermore, an extension of our experiments in functionality test generation towards sentiment analysis may bring us to more robust models in emotion detection and consequently to more reliable insights we can identify in the social media data.

4 MULTIMODAL ANALYSIS OF SOUND-RELATED EXPERIENCE AND BEHAVIOUR

To understand audience experience and behaviour in a comprehensive way, it is imperative to collect and analyse multimodal data that combines the outcomes of different components, e.g. video recordings of audience behaviour, self-reports or physiological signals. The goal of this task is to provide flexible and mobile multimodal measurement setups as well as custom-made analysis pipelines that the individual Use cases and artistic projects can use according to their needs.

4.1 Related work

A starting point is the multimodal measurement setup that already exists in the ArtLab of the MPIEA. The ArtLab Auditorium is a multifunctional event space for psychophysiological and behavioural reception research in the context of speech, sound, music and multimedia performances. In combination with the adjacent control room for scientists and a reference sound control room, it serves as a laboratory for studies and experiments on experience and behaviour in a naturalistic reception situation.

Here, multimodal time series data such as physiological responses (electrodermal activity, blood volume pulse, electrocardiogram, respiration, EEG, facial electromyography), self-reports, and behavioural data (motion capture, movement and facial expression from video recordings) can be recorded from participants while they take part in sound and music performances. To obtain videos of audience behaviour, two PTZ dome cameras are permanently installed at the back of the stage. For the continuous recording of physiological responses, a mobile system by the company Biosignals Plux with eight channels and a sampling rate of up to 1000 hz is used in combination with Raspberry PI 5 single-board computers. Centrally controllable tablet computers are used to prompt self-reports from the audience members at any point in time. In addition, audiovisual recordings of the performance itself can be made. All signals that are recorded are time-synched via the system clock that is distributed to all connected devices via the network. (See figures 16–19.)



Figure 16: Participants with physiological sensors on their hand



Figure 17: Participants with physiological sensors and questionnaires in the ArtLab auditorium



Figure 18: An audience with tablet computers for self-reports



Figure 19: Motion capture of a performer

(all © MPIEA)

The data can be analysed with event-related and time-series oriented methods to assess the intensity and qualities of experiences on the individual and group level and to relate it to the features of the respective stimulus.

To adapt this system also to other ReSilence use scenarios outside the ArtLab, we have run tests with individual components. This way, we have developed prototype solutions for the flexible and mobile use of questionnaires (via tablet computers), video-based motion capture and behaviour analysis, and the recording of physiological responses that could be requested by individual ReSilence projects if needed.

For the artistic project “Theatre of Memory” by the ReSilence artist-in-residence Tim Otto

Roth that was displayed at the MPIEA, it was agreed with the artist to capture a measure of time spent in the exhibition for each visitor as well as self-reports on their behaviour, perception, appreciation and experience together with sociodemographic and psychometric information. To obtain this data, we adapted the system in the following way: Upon entering the exhibition space, all visitors received an admission ticket with an individual QR code. This was scanned by an assistant on a tablet next to the entrance immediately before being handed out. On leaving the exhibition, the same code was scanned again by an assistant on the tablet. The times of the scanning process were saved so that the duration of stay could be calculated. The time stamps can also be used to assign the stay to the pieces heard in the sound installation, for which time stamps were created.

After their visit, participants had the opportunity to voluntarily take part in a questionnaire study. Completing the questionnaire took about 10 minutes. Five laptops were provided in a separate room in a quiet environment for this purpose. The laptops were in kiosk mode so that the participants could not make any changes to the settings. The questionnaires were implemented on the LimeSurvey survey platform and made available on the MPIEA's internal server. To start the questionnaire, participants had to present their individual QR code from the admission ticket to the laptop's camera. A sub-ID was then created in LimeSurvey, in which the code was stored in order to link the answers from the questionnaire with the duration of stay.

For the display of the work on a different site (Tieranatomisches Theater Berlin), we migrated the questionnaire to a public server and made it accessible via people's own smartphones.

4.2 Datasets

In addition to several already existing multimodal datasets from earlier MPIEA studies and performances in the ArtLab that could be used to develop algorithms for data preprocessing and analysis, as well as for machine learning procedures, a novel dataset was collected in the context of the "Theatre of Memory" project. In Frankfurt/M., data for time spent at the exhibition were collected from $N = 192$ participants. Of those, $n = 91$ also provided self-reports. In Berlin, self-reports were collected from another $N = 79$ participants.

4.3 Preliminary results

The behavioural and self-report data obtained in the context of the project "Theatre of Memory" offers first insights into some of the main questions of ReSilence as far as user experiences are concerned.

Overall, visitors responded positively to this novel form of cross-disciplinary conceptual artwork: From the questionnaire items about their experience, the highest mean values were found for "interesting" ($M = 5.3$), "entirely new work" ($M = 4.9$), "liking" ($M = 4.6$), and "thought provoking" ($M = 4.3$, all based on 7-point Likert scales). Negative experiences had much smaller means, e.g. "I couldn't really get to grips with the work" ($M = 3.4$) or "I was overwhelmed by the work" ($M = 3.9$). However, the positive experiences that were strongest were almost all related to an experience dimension we called "intellectual stimulation" in earlier studies (e.g., Wald-Fuhrmann et al. 2021), whereas dimensions such as absorption or being immersed were only rarely activated (M immersion = 3.3, M moved = 3.1).

Correlations between duration of visit and aesthetic behaviour and experience were not

found for how people approached the work, for only two items from the battery of what they perceived the work to be (as a piece of musical theatre: $r = 0.25$, as an artistic representation of neural processes: $r = 0.32$), but for several of the experience items (in particular, “I forgot the time”: $r = 0.44$, “thought provoking”: $r = 0.36$, “interesting”: $r = 0.33$).

Since ReSilence aims at using the arts to address important societal questions and to make novel technologies more accessible, we had also included some questionnaire items that were supposed to measure such associations. The items “The work made aesthetically convincing use of novel technologies” had a mean rating of $M = 4.1$ and was positively correlated with “I found the work to be something entirely new” ($r = 0.32$) and “I perceived the work as a forward-looking form of art” ($r = 0.59$). These correlations show that a positive aesthetic experience of a piece of art that uses up-to-date technologies can also lead to a positive evaluation of these novel technologies.

Although ratings of these three items were partly related with people's general attitudes towards contemporary arts and technologies, they were not found to be significant predictors of ratings of “The work made aesthetically convincing use of novel technologies” in a linear regression model that included experience items, attitudes and personality traits. The only significant predictor in the final model was the experience item “I found the work captivating” ($R^2 = 0.31$, coefficient of predictor “captivating” = 0.59), indicating that a genuine aesthetic experience can outweigh also negative and sceptical attitudes.

4.4 Future considerations

We are prepared to adapt and develop the multimodal measurement system further according to the needs of other ReSilence projects. For example, the existing components can be used individually or in any combination. Additional components can also be integrated if required, e.g. mobile eye tracking. Further, it is planned to construct a display tool that presents the content of all data channels at the same time and in their own mode and relative to the time series. This way, it would be possible to listen to and/or watch the performance while seeing individual or group responses as they evolved during the performance.

Currently, an adaptation of the system according to the still emerging needs of the artist Lea Luka Siekau is being discussed.

5 SELF-REPORT GENERATION

The task of self-report generation T4.4 concludes WP4. Its aim is to give clarity to the results of different types of multimodal analyses performed in WP4 by joining the outcomes of individual techniques developed within T4.1-T4.3 and producing descriptive easy-to-follow reports in a natural language. The final goal is to make these reports personalised and self-sufficient so that individual users obtain necessary information tightly related to their objectives in soundscape experience assessment.

In the first period of the Project, we focused on research and development in automatic domain-independent crafting of the semantic structures needed for the enhancement of the coverage of grammar-based generators. This general direction opens a perspective in adaptation of generators to any specific domain by conditioning the inputs of the crafting. Further limiting inputs to the project domain would result in specialised structures to match preferences of the users in ReSilence.

To address the coverage issue faced by rule-based text generators, we designed a pipeline for extracting abstract dependency templates (predicate-argument structures) from Wikipedia text to be used as input for generating text from structured data with the UPF's FORGe system³⁶. The pipeline comprises three main components: (i) candidate sentence retrieval starting from a property of a given semantic triple, (ii) clause extraction, ranking and selection, and (iii) conversion to predicate-argument form. In the following, we present related work, our approach, and preliminary evaluation for the ranking and selection module.

5.1 Related work

A number of papers have tackled the extraction of templates from text to be used as input for natural language generation (NLG). Duma and Klein (2013) mine and prune sentence templates from Wikipedia articles, but (i) extract templates given an entity (instead of a property as in our case), and (ii) manage to obtain a template for about 20% of the target entities. Ell and Harth (2014) achieve impressive coverage with their (multi-property) sentence templates, but also suffer accuracy problems, with the text faithfully representing the input in only about half the cases. Perera and Nand (2015) use an open Information Extraction (IE) tool to identify candidate sentence spans that verbalise a given property, and then acquire lexicalisation information via VerbNet, resorting to default strategies when a predicate is not covered by VerbNet. Hoang et al. (2022) suggest several general approaches to align triple components and textual elements, namely string, substring, hypernym and synonym matching; for property matching, they also use a pre-trained vector model to calculate the distance between words. Other recent work on this topic uses keyword matching (Kaffee et al., 2022) or cosine similarity (Abhishek et al., 2022) for aligning triples and text in under-resourced languages. In order to assess the strength of the alignment, Abhishek et al. (2022) apply a natural language inference (NLI) model to detect (lack of) entailment between the triples and the candidate sentences.

One difference between our approach and most of the related work on template mining for NLG is that we want to extract predicate-argument templates, and not full-sentence templates. However, the approaches have a lot in common, since we extract the predicate-

argument structures from sentences. The main issue with most of the approaches above is the lack of accuracy. Recently, Transformers have been shown to improve accuracy for Question-Answering (Karpukhin et al., 2020), including for the specific task of aligning text and structured data (Oguz et al., 2022) and also for fact checking, for instance for comparing tables and text (Zhang et al., 2020). In our approach, we therefore explore another way of aligning linguistic predicates and properties via Transformer-based meaning similarity scoring.

5.2 Methodology

5.2.1 Abstract dependency template extraction

Semantic triples we operate with consist of three related elements called a Property, a Subject and an Object in Semantic Web terminology. DBpedia is a comprehensive semantic resource containing a large variety of such triples (Lehmann et al., 2015). As it covers knowledge in many domains, including sound- and music-related contents pertinent to ReSilence, it is chosen as a base resource for our work. A Subject (denoted here by DB-Subj) is usually an entity that has a Property and a value for this Property, which is the Object (DB-Obj). For example, in Figure 20, the entity `Sanremo_Music_Festival` is associated with 5 properties: `Name`, `Years_active`, `Dates`, `Location` and `Genre`. The semantics of each property is defined by DBpedia editors³⁷, but in most cases, “*the Property of the DB-Subj is DB-Obj*” makes it clear (e.g., “*the Location of the Sanremo_Music_Festival is Sanremo, Liguria, Italy*”).

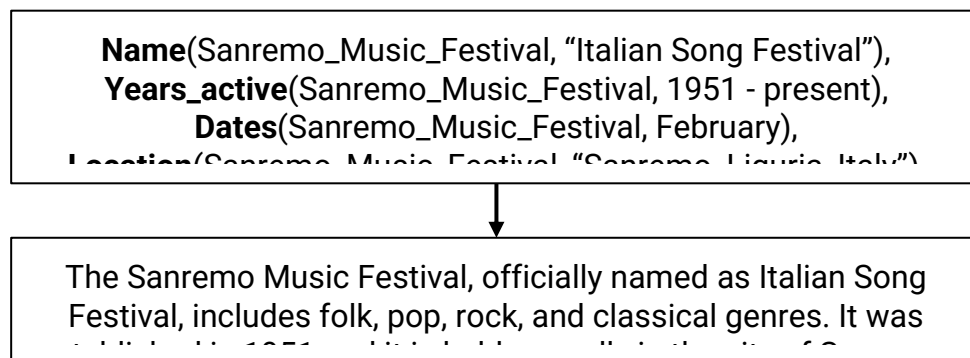
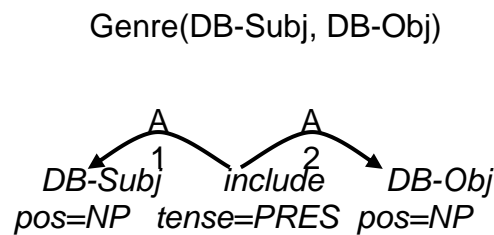


Figure 20: A DBpedia triple set and a corresponding generated text. Triple = Property(DB-Subj, DB-Obj), where the DB-Subj is an entity, and the DB-Obj another entity, a numeric, a date, etc.

FORGe operates on dependency structures at several levels of representation (syntax, semantics), and needs partially lexicalised predicate-argument (*PredArg*) structures in the PropBank style (Kingsbury and Palmer, 2002) to use as input for generation (see Figure 21b). In other words, if a mapping between property and *PredArg* structure as shown in Figure 21a-b does not exist, the generator cannot introduce the appropriate words and, unless a backup mechanism is in place, it fails to generate a text.

³⁷ http://mappings.dbpedia.org/index.php/How_to_edit_the_DBpedia_Ontology



The DB-Subj includes DB-Obj.
DB-Obj is included by DB-Subj.

Figure 21: (a) The Property Genre, (b) a corresponding PredArg template (graph with no linear order information), and (c) two possible verbalisations of the property via the template. A1/2 = first/second argument

Thus, the overall problem that we are tackling is the following: given (i) the rule-based FORGe generator that covers all properties in the training data, (ii) the mappings between these properties and their respective PredArg template, (iii) an input triple set that contains one or more properties not currently covered by the generator, automatically extend the mapping in ii with new unseen property/template pairs that will enable FORGe to generate a text that verbalises all input properties.

We carry out this automatic extraction of property/template pairs via a pipeline that retrieves and ranks candidate clauses from Wikipedia that correspond to a given DBpedia instantiated property (i.e., a triple), and converts them to predicate-argument representations. The pipeline and components have been defined and connected, and we have identified and addressed two primary challenges: one is candidate clause extraction, i.e., find a sentence or a clause that likely matches the input triple, the other is the candidate classification for the selection of correct clauses, i.e., identify clauses that contain an exact match. Figure 22 shows a complete view of the components that allow us to extract PredArg template(s) given one input DBpedia triple³⁸. The three main components of the pipeline (indicated across the bottom in Figure 22) are: (i) Candidate sentence retrieval, (ii) Candidate clause ranking and selection, (iii) conversion to predicate-argument template.

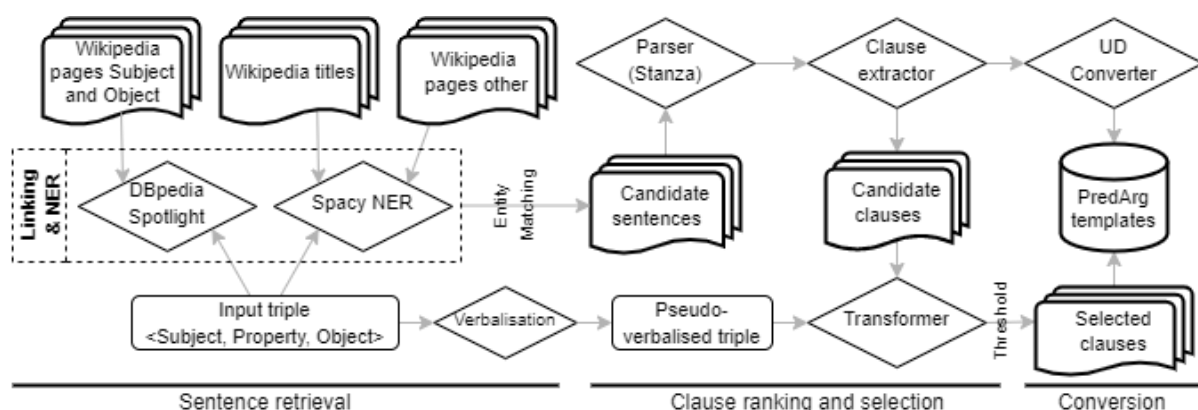


Figure 22: Overview of the pipeline for PredArg template extraction

³⁸ Since FORGe performs triple aggregation during the generation, we don't need to extract PredArg templates that correspond to multiple triples.

5.2.2 Candidate sentence retrieval

The first step is to find candidate sentences for a given input triple; since DBpedia triples are often verbalised in Wikipedia texts, we use the Wikipedia contents as a candidate source. Via the HuggingFace dataset³⁹, we have access to the title and the cleaned (plain) text of each article. We first find the Wikipedia articles of both the DB-Subj and the DB-Obj (if any), and then run the entity linking tool DBpedia Spotlight (Mendes et al., 2011) on the input triple’s DB-Subj and DB-Obj and on the article text to find sentences that mention both the DB-Subj and the DB-Obj.

In order to find more candidate sentences and possibly get better candidates, we also perform a relaxed search. We get a named entity type for the DB-Subj using Spacy NER⁴⁰, and parse Wikipedia article titles until we find an article about an entity of the same type as the DB-Subj. We then proceed to run Spacy NER on the DB-Obj and the found article so as to find sentences that contain two entities of the type of the DB-Subj and DB-Obj, and replace these entities with the ones from the original input for the ranking phase.

5.2.3 Candidate clause ranking and selection

To extract minimal clauses and calculate their semantic similarity with the input triple, we use a Sentence Transformer bi-encoder model⁴¹ (Reimers and Gurevych, 2019), so that candidates scored above a given threshold are kept while others are discarded. Existing sentence similarity approaches return a score for a pair of sentences; in our case, we need a similarity score between a triple and a clause. To account for this peculiarity, we fine-tune the model using a dataset we created for this purpose. The fine-tuning dataset consists of pseudo-verbalisations of input triples aligned with reference sentences⁴² as follows. For each triple T, we compiled 4 sets of sentences that correspond to 4 levels of similarity with T (on a scale from 0 to 1): 1.00 (sentences that verbalise exactly T), 0.66 (sentences that verbalise a triple that has 2 elements in common with T, either DB-Subj and Property, DB-Subj and DB-Obj, or Property and DB-Obj), 0.33 (1 element in common with T), and 0.00 (no element in common with T); see Table 13.

Triple: Location (Agra Airport, India)	
1.00	'Agra Airport is in India.', 'Agra airport is located in India.'
0.66	'Agra Airport is located in Uttar Pradesh.', 'The Taj Mahal is in India.', etc.
0.33	'AGR is the ATA Location Identifier for Agra Airport.', 'AC Hotel Bella Sky Copenhagen is in Copenhagen.', 'Mother Theresa is from India', etc.

³⁹ <https://huggingface.co/datasets/wikipedia>

⁴⁰ <https://spacy.io/api/entityrecognizer>

⁴¹ <https://huggingface.co/sentence-transformers/nli-distilroberta-base-v2>

⁴² See Section 5.3 for the origin of reference sentences and statistics over the dataset in our experiments.

0.00	'Agnes Kant is a national of the Netherlands.', 'FC Köln played the 2014-15 season in the Bundesliga.', 'Ampara Hospital has 476 beds.', etc.
------	---

Table 13: Sentences with different similarity levels; in **bold**, the elements in common with the triple

To make the representation of a triple closer to a sequential format of inputs that the pre-trained sentence bi-encoder model operates with, we converted each triple to a typed pseudo-verbalised form (Pasricha et al., 2020): *Location(Agra Airport, India) → <AIRPORT> Agra Airport <PROP> location <PLACE> India*. When an unknown property is detected in the input, we do not have at hand a verbalisation of the triple that contains it since the objective of our pipeline is to discover such verbalisations. Therefore, the pseudo-verbalisation here is an adequate strategy: the pseudo-verbalised input triple will be compared to the candidate clauses. We adapted a couple of aspects in introducing types in the outcome in (Pasricha et al., 2020) with respect to the specifics of DBpedia data format: (i) we implemented a check if a DB-Obj is of type <NUMBER> or <DATE>, and (ii) we took the first ontology type (starting with dbo:) in the rdf:type section of the DBpedia page for the other types.

The sentences retrieved from Wikipedia are usually long, in the encyclopaedic style; we thus reduce each sentence to the minimal subtree that contains a finite verb and two elements of the same types as the DB-Subj and the DB-Obj respectively. Each candidate sentence is parsed with the Stanza Universal Dependency parser (Qi et al., 2020); the output syntactic structures are then processed to extract the minimal subtree via graph-transduction grammars implemented in FORGE. The original sentence span that corresponds to this clause subtree is selected (for illustration, see sample output below).

5.2.4 Conversion to PredArg templates

The predicate-argument structures of the selected clauses from the previous step are created. For this, we use the grammar-based UD Converter released for the Surface Realisation Shared Tasks (Mille et al., 2019), which returns a predicate-argument structure for a given UD parse. The specific DB-Subj and DB-Obj are replaced by generic [DB-Subj] and [DB-Obj] placeholders resulting in a structure shown in Figure 21b.

5.2.5 Sample outputs of all components

In this subsection, each step of the pipeline with sample inputs and outputs is illustrated.

- **Input triple:** Alan_Bean || birthDate || "1932-03-15"
- **Entity linking (DBpedia Spotlight):**
 - DB-Subj: Alan Bean
 - kb_{id}: 11139903761698166211
 - dbpedia link: http://dbpedia.org/resource/Alan_Bean
 - DB-Obj: "1932-03-15"
 - kb_{id}: 0 (No dbpedia entity found)
- **Entity type assignment (Spacy):**
 - DB-Subj: Alan Bean
 - Entity label: 380 (PERSON)

- DB-Obj: "1932-03-15"
- Entity label: 391 (DATE)
- **Typed pseudo-verbalisation:** <PERSON> Alan Bean <PROP> birth date <TIMEPERIOD> "1932-03-15"
- **Sentence extraction (Entity matching) and parsing (Stanza):** See Figure 23.

1	The	DT	Definite=Def PronType=Art	2	det
2	seat	NN	Number=Sing	11	nsubj
3	of	IN	5	case	
4	Wheeler	NNP	Number=Sing	5	compound
5	County	NNP	Number=Sing	2	nmod
6	,	,	2	punct	
7	in	IN	8	case	
8	Texas	NNP	Number=Sing	5	nmod
9	,	,	11	punct	
10	is	VBZ	Mood=Ind ...	11	cop
11	Wheeler	NNP	Number=Sing	0	root
12	,	,	11	punct	
13	where	WRB	PronType=Rel	16	mark
14	Alan_Bean	NNP	subject=true	16	nsubj:pass
15	was	VBD	Mood=Ind ...	16	aux:pass
16	born	VRN	Tense=Past ...	11	acl:relcl
17	on	IN	18	case	
18	1932-03-15	CD	NumForm=Digit ...	16	obl
19	.	.	11	punct	

Figure 23: Sample UD structure (selected columns)

- **Clause Extraction (graph transduction grammars):** 'Alan_Bean was born on "1932-03-15"' (*the minimal subtree that contains both the DB-Subj and the DB-Obj, with additional trimming (e.g., a relative pronoun "where" before the DB-Subj is removed)*)
- **Clause ranking (Transformer):** 'Alan_Bean was born on "1932-03-15"' -> 0.8853' (*If the clause is above the defined threshold, it is selected for the template*)
- **Conversion to PredArg (UD Converter):** See Figure 24 for the delexicalised predicate-argument template extracted from the selected clause.

1	bear	VERB	Tense=Past ...	0	ROOT
2	[Subject]	PROPN	subject=true ...	1	A2
3	[Object]	NUM	NumForm=Digit ...	1	Time

Figure 24: Sample PredArg template (selected columns)

5.2.6 Generation from fused information

As the final purpose of the T4.4 is to generate descriptive reports from fused information, we envision this to be realised in a retrieval-augmented generation fashion using LLMs which combine pre-trained parametric and non-parametric memory for language generation (Lewis et al., 2020). The fused multimodal information may be presented in two formats acceptable by two types of generators considered in the scope of the Project, i.e. (i) formal knowledge structures that can be handled by the grammar-based generator as knowledge-to-text generation, and (ii) relevant textual information from crawled social media data and reviews for text-to-text generation. In order to produce personalised reports that would benefit from outcomes of both generators and sound coherently as a whole, we propose to incorporate these outcomes into an instruction-oriented LLM that would perform aggregating the information and presenting it in a single style with an emphasis on specific aspects as provided in instructions.

We envision retrieval component to operate as a query-based summarizer based on encoder-decoder architecture with dual encoder similar to the approach presented in (Hasselqvist et al., 2017): for a given query (in a form of a semantic RDF triple) and a document (or a set of candidate sentences) separately encoded by two jointly trained encoders, the model returns sentences that would expand the information provided in a query decoding one sentence at a time and taking into account already decoded sequence to enhance coherence and avoid repetitions; in case there are no relevant sentences, the model returns a statement “no related information”. Depending on the complexity of the model, we may consider its replacement with a faster approach implemented in the “candidate clause ranking and selection” module presented above, although training a bi-encoder with a similarity metric special for summarisation purposes may present a challenge.

The outcome of the retrieval component and knowledge-to-text verbaliser are taken as a joint input to the decoder-based LLM like GPT (Ouyang et al., 2022). The overall architecture is presented in Figure 25.

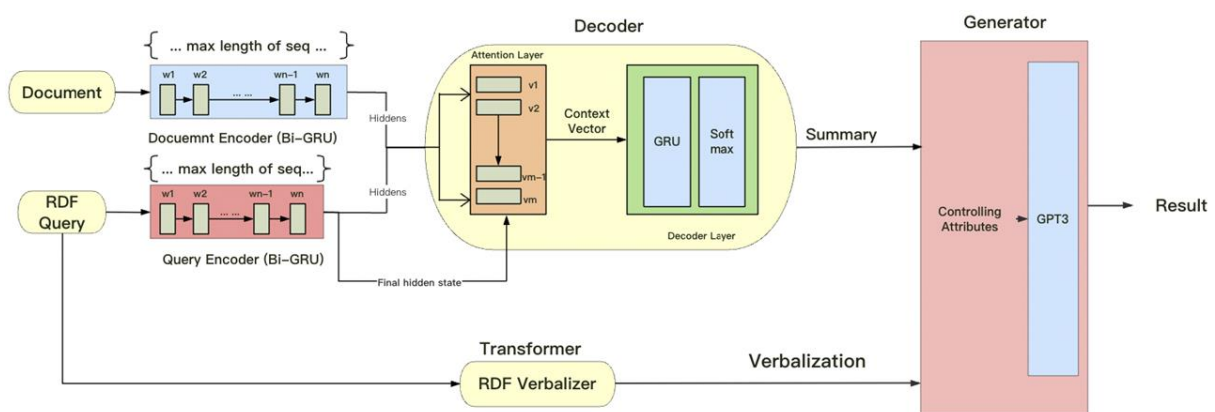


Figure 25: Overview of the proposed architecture, includes the RDF verbaliser, encoder-decoder summariser, and the generator

The structure of the final report depends on the instruction/context provided to the LLM. We consider three possible prompting approaches, i.e. ways of setting the context: PM_b – basic instruction prompting “write a short description based on: {text}. {knowledge}.”, where we include all the input retrieved in a field {text} and verbalised from the knowledge structures in a field {knowledge}; PM_k – more knowledge-specific prompting explicitly differentiating input from the summarizer and knowledge verbalizer, “Your task is to write a short description based on the text and the knowledge provided. <text>: {text}. <knowledge>: {knowledge}.”; PM_f – few-shot prompting that shows examples of how the outcome is expected to look like, “Your task is to write a short description based on the text and the knowledge provided. <text>: {text}. <knowledge>: {knowledge}. <response>: {response}. <text>: {text}. <knowledge>: {knowledge}. <response>: {response}. <text>: {text}. <knowledge>: {knowledge}. <response>:” . We explore which approach tends to generate fewer hallucinations when parametric memory is used to consistently join the pieces of provided information.

5.3 Datasets

5.3.1 Abstract dependency template extraction

Dataset compilation is done over the data of WebNLG+ shared task (Ferreira et al., 2020)

which is a benchmark for data-to-text NLG consisting of aligned DBpedia triple sets and texts. A part of its test set inputs contains features not seen in the training or development data.

To get Wikipedia pages, we retrieve (i) the page of the DB-Subj, (ii) the page of the DB-Obj if any, and (iii) 1,000 random article about an entity that has the same type as the DB-Subj (matching the Spacy tag of the title with that of the DB-Subj). We then look for candidates on the pages, based on the type predicted by DBpedia Spotlight (pages of DB-Subj and DB-Obj) or by Spacy (other pages). We detokenise the DB-Subj and the DB-Obj for them to be parsed as one single named entity.

The fine-tuning dataset is built from the one-triple items in the test set of the WebNLG+. Following the procedure of assigning four levels of similarity for sentences for a given triple (as described in Section 5.2), we obtained 7,645 triple/sentence pairs in total for the set of similarity 1.00, 24K pairs for 0.66, 399K for 0.33, and 23M for 0.00. To balance the dataset, we randomly picked 7,645 pairs from the sets of 0.66, 0.33, and 0.00. The sample is divided 70/15/15 for training, development and test sets, respectively.

We also compiled two evaluation subsets, (a) the WebNLG+ development subset of single-property inputs (401 triples), and (b) the subset of the WebNLG+ test set comprising all and only items with properties not seen in the WebNLG+ training data (113 triples). For each input triple, there are 1 to 3 exactly matching sentences (the corresponding reference sentences in the WebNLG+ dataset), which are the target sentences that we want the model to prefer (rank highest) for the input triple. For use as the non-matching candidates, which should be dispreferred (ranked lower) by the model, we select all other sentences that verbalise one-triple inputs, and all sentences that verbalise two-triple inputs; the total $Eval_{Dev}$ and $Eval_{Test}$ candidate pools contain 1,834 and 2,887 sentences respectively. This way, we ensure that we have candidates with a significant meaning overlap with the target sentences (one-triple inputs can share elements with one another and two-triple inputs can include elements or even full triples of the one-triple inputs).

5.3.2 Generation from fused information

The data to train the query-based summarizer is an expanded dataset sourced from WebNLG+ training and validation subsets. To collect data that suits the summarization task, we retrieve Wikipedia articles using the subject in the triple as the search keyword and split the text into sentences. To determine the relevance between the split sentences and the triple, we calculate Euclidean distances for a subject and an object of a triple to sentence words in the GloVe embedding space (Pennington et al., 2014). To accelerate the comparison, we mainly focus on words of specific part-of-speech tags such as PROP (proper nouns), NOUN (common nouns), and NUM (numbers). If the distance exceeds a predefined threshold, it indicates a potential relationship between the compared words. If both the subject and object are in a potential relationship with the words of the sentence, we denote it as a closely related sentence.

5.4 Experiments and preliminary results

5.4.1 Experiment setup

Parameters for the fine-tuning of the sentence bi-encoder model for triple-to-clause similarity prediction were chosen as follows. The train batch size is 16, and the train loss is Cosine

Similarity Loss. The training stage is performed for 4 epochs with warm-up steps (10% of the training data) applying the Embedding Similarity Evaluator (which uses the development set) with evaluation at every 1000 steps.

For the final evaluation, we compare two models, the off-the-shelf Transformer (Reimers and Gurevych, 2019) and our fine-tuned version of it using $Eval_{Dev}$ and $Eval_{Test}$ subsets. The objective is to obtain performance upper and lower bounds for the fine-tuned model by examining how accurate it is at selecting the right candidate (a) for properties seen during fine-tuning, and (b) for unseen properties, which is the most realistic scenario for PredArg template extraction.

5.4.2 Results

On the development data (top half of Table 14), the fine-tuned model ranks all the target sentences at the top in 98.5% of the cases, and at least one of the target sentences at the top in 99.5% of the cases. The average similarity score of the correctly top-ranked sentences is 0.963, and the first non-target sentence is on average scored 0.346 points below. The off-the-shelf model is effective at placing one, but not all, target sentences at the top, and the difference in scores between the target and non-target sentences is half of what it is for the fine-tuned model (0.170 and 0.346 respectively).

	All properties (401 triples)		Unseen properties (113 triples)	
	<i>Off-the-shelf</i>	<i>Fine-tuned</i>	<i>Off-the-shelf</i>	<i>Fine-tuned</i>
$Accuracy_{All}$ (%)	91.02	98.50	56.64	73.45
$Accuracy_{One}$ (%)	98.25	99.50	87.61	96.46
$Avg. top P_{OK}$	0.785	0.963	0.787	0.929
$Margin$	0.170	0.346	0.110	0.212
$Avg. top P_{Mod}$	0.684	0.754	0.702	0.776
$Avg. top P_{InvSO}$	0.803	0.971	0.815	0.952
$F_1 (Eval_{Dev} thresh.)$	0.798 (0.73)	0.955 (0.87)	0.429 (0.73)	0.694 (0.87)
$F_1 (best thresh.)$	0.798 (0.73)	0.955 (0.87)	0.537 (0.78)	0.745 (0.93)

Table 14: Evaluation of the ranking module (WebNLG+). **Accuracy_{All/One}** = % of cases with all/one good candidate(s) ranked at the top; **Avg. top P_{OK}** = Average score (0 to 1) of correctly top-ranked n candidates for a given input triple; **Margin** = difference in % between top-ranked candidates and first non-correct candidate; **Avg. top P_{Mod/InvSO}** = Average score (0 to 1) of the top-ranked candidate for a given input triple in which the property name was randomly changed / the DB-Subj and DB-Obj were inverted; **F₁** = best F₁-score for candidate selection obtained via the indicated threshold

To assess to what extent the models capture the semantics of the properties, we repeated the experiment above but modifying the input triples in two ways: replacing the property name by another randomly selected property (Avg. top P_{Mod}), and inverting the DB-Subj and DB-Obj (Avg. top P_{InvSO}). The off-the-shelf model has a harder time discriminating between correct and wrong properties than the fine-tuned model (similarity scores of 0.785 and 0.684, respectively, for the off-the-shelf model, 0.963 and 0.754 for the fine-tuned model). However, neither of the models is able to discriminate cases where the DB-Subj and DB-Obj are switched, yielding even higher scores on average than with the original triple (Avg. top P_{InvSO}).

We then looked for the threshold at which a model gets the best F_1 -score when selecting a candidate sentence. We tested all thresholds (in steps of 0.01 from 0 to 1) for each model on the $Eval_{Dev}$ set and obtained values of 0.73 and 0.87 for the off-the-shelf and fine-tuned models respectively, which yield a F_1 of 0.798 and 0.955 respectively. On the unseen test set $Eval_{Test}$, these thresholds yield a significantly lower F_1 -score, the fine-tuned model reaching an F_1 of only 0.694 and the off-the-shelf model 0.429. A better F_1 can be achieved on these unseen triples by selecting different thresholds (both higher, at 0.93 and 0.78 respectively). Figures 26 and 27 show a plot of the F_1 -score in function of the selection threshold for candidate sentences.

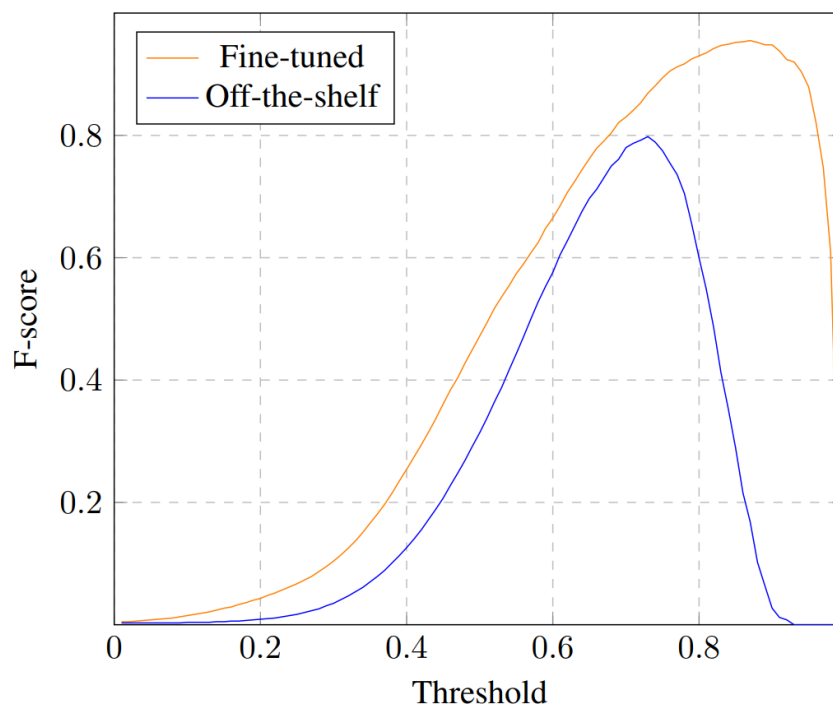


Figure 26: Threshold definition for clause selection ($Eval_{Dev}$ set)

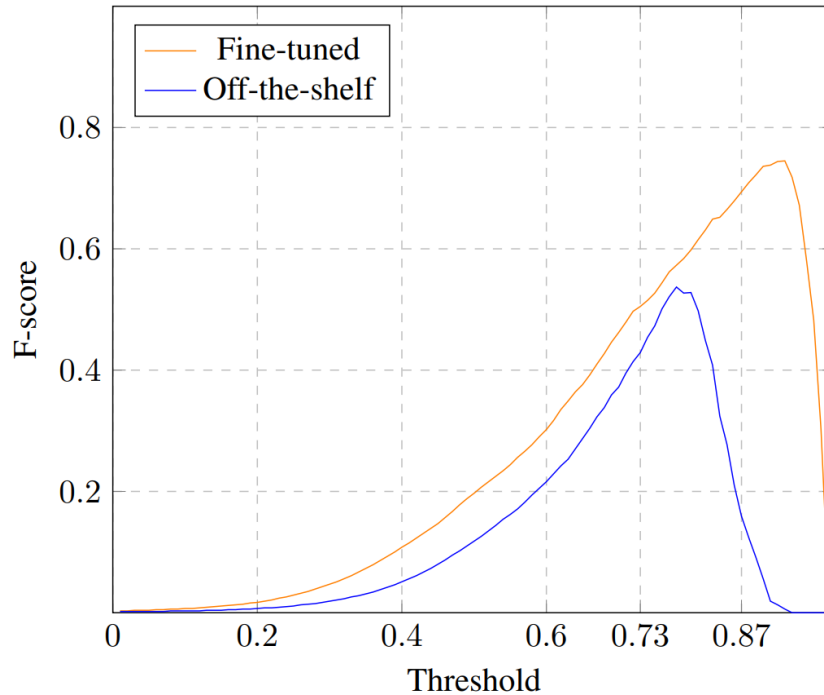


Figure 27: Threshold definition for clause selection (Eval_{Test} set)

5.4.3 Error analysis

We examined all the false positives and false negatives for the best threshold on the Eval_{Dev} set (0.87), and found the following errors.

False positives (53 errors): (i) a sentence that corresponds to 2 triples was selected, because one or more elements of the second triple are very similar with the input triple’s DB-Subj, DB-Obj or Property (75% of errors); (ii) the selected sentence verbalises a triple that is almost identical to the input triple (25%).

False negatives (35 errors): (i) mismatch between a DB-Subj, Property or DB-Obj and their corresponding verbalisation due to an accent, a comma in a number, quotation marks, parentheses, casing (57%); (ii) a triple element is verbalised with a word judged semantically distant (29%); (iii) a reference sentence is wrong (14%). Only false negatives (i) and (iii) stem from errors or lack of normalisation in the data; the other errors are due to the model.

The decrease in the score between the Eval_{Dev} and the Eval_{Test} sets is rather surprising, especially for the off-the-shelf Transformer, for which we would expect no difference between seen and unseen properties. We hypothesise that the Eval_{Test} set is more challenging: (i) the reference sentences seem less similar (0.910 on Eval_{Test} VS 0.932 on Eval_{Dev} when running the off-the-shelf Transformer on the gold sentences for triples of size 1); (ii) some problematic cases are more frequent (e.g., the DB-Subj or DB-Obj has content in parentheses in 34% of the Eval_{Test} triples, VS 12% in the Eval_{Dev} set); (iii) there are more candidate sentences for the Eval_{Test} set (see Section 5.3).

5.4.4 Generation from fused information

We performed two types of comparisons between different prompting techniques. In particular, we first checked how far the generated text deviates from the input that the model received in the *{text}* and *{knowledge}* fields using METEOR and ROUGE scores (Banerjee and

Lavie, 2005; Lin, 2014). The results are shown in Table 15.

	METEOR	ROUGE_1	ROUGE_2	ROUGE_L
PM_b	0.35	0.39	0.30	0.36
PM_k	0.55	0.60	0.50	0.57
PM_f	0.78	0.70	0.64	0.69

Table 15: Evaluation of the prompting techniques (**generated text against input text**). PM_b – basic prompting, PM_k – knowledge-specific prompting, PM_f – few-shot prompting

We observe that few-shot prompting adds less information from the parametric memory of the model, filling the outcome to a major extent with the text based on the original input.

Second, we carried out evaluation using reference texts from the validation set considering various domains separately. The results are presented in Table 16. Consistently across the categories, the PM_f approach leads to better coverage of reference text which occupies roughly 40-50% of the text produced. The remaining part corresponds to the retrieved data and additional information from parametric memory of the model.

Category	Method	METEOR	ROUGE_1	ROUGE_2	ROUGE_L
Airport	PM_e	0.21	0.25	0.17	0.22
	PM_k	0.31	0.44	0.30	0.35
	PM_f	0.42	0.51	0.40	0.45
Astronaut	PM_e	0.24	0.25	0.19	0.24
	PM_k	0.36	0.41	0.35	0.41
	PM_f	0.44	0.51	0.47	0.50
Building	PM_e	0.27	0.26	0.17	0.23
	PM_k	0.30	0.39	0.31	0.36
	PM_f	0.43	0.50	0.43	0.47
City	PM_e	0.22	0.27	0.17	0.22
	PM_k	0.36	0.45	0.36	0.38
	PM_f	0.48	0.53	0.41	0.50
Comics	PM_e	0.21	0.28	0.19	0.23
	PM_k	0.35	0.39	0.35	0.38

	PM_f	0.44	0.51	0.40	0.46
Food	PM_e	0.20	0.24	0.16	0.21
	PM_k	0.37	0.42	0.33	0.40
	PM_f	0.42	0.50	0.43	0.47
Monument	PM_e	0.28	0.25	0.17	0.25
	PM_k	0.38	0.45	0.38	0.40
	PM_f	0.41	0.50	0.46	0.47
SportsTeam	PM_e	0.27	0.34	0.21	0.36
	PM_k	0.32	0.45	0.39	0.42
	PM_f	0.43	0.52	0.43	0.47
University	PM_e	0.16	0.30	0.19	0.23
	PM_k	0.38	0.44	0.37	0.41
	PM_f	0.47	0.52	0.49	0.52
WrittenWork	PM_e	0.29	0.32	0.27	0.27
	PM_k	0.38	0.43	0.36	0.39
	PM_f	0.48	0.53	0.49	0.52

Table 16: Evaluation of the prompting techniques (**generated text against reference text**). PM_b – basic prompting, PM_k – knowledge-specific prompting, PM_f – few-shot prompting

We expect that the few-shot prompting technique will be of the most relevance to the project as it keeps the reports concise (not much extra information related to the results of the multimodal analysis is added), and allows for varying the style and placing emphasis on various parts of contents depending on the custom user preferences by enhancement of the instructions with corresponding specifications.

5.5 Collaboration with Open Call Artists

In the first part of the Project, we carried out preliminary experiments within the artistic residency of Lea Luka Sikau. We collaboratively explore how the lexicon which is specific to her work can be placed in various discourses related to the sound experiences, perhaps, being used metaphorically, and what types of sentiments it can be accompanied with.

Lea compiled the following list of concepts we would like to be found in real examples on topics deviating from the common domains pertinent to these concepts: *microbiome, irritable bowel syndrome, gastric signals, bowel movements, bowel sound, gastroenterology, gastric,*

spill the gut, stomach, gut punch, gutted, gut feeling, gut, to gut, probiotics, gut bacteria.

After an unsuccessful attempt to use traditional search techniques to find resources with an abundance of relevant lexicon which would not be leveraged in a direct meaning (first of all, we were investigating social question-answering platforms such as Reddit⁴³ and Quora⁴⁴), we tried to turn the task into an LLM-based text generation task. Instead of having linguistic analysis over a single resource, we would rely on the parametric memory of a model trained on nearly all the topics and sources of information available on the Web. We also compared this approach to generation with retrieval-augmented generative models (RAGs) which apart from producing a text in a requested discourse return the references to real sources used as a base for the response. For the pure LLM-based generation we used instruction-oriented model ChatGPT⁴⁵, while GPT-based Bing Copilot⁴⁶ and Perplexity.ai⁴⁷ were used as RAGs.

We tried various prompts exploring the possibilities of the models, whereas, the main prompt was the following (*{vocabulary}* placeholder is replaced with the list of concepts):

Give me some examples of metaphors for the provided vocabulary related to emotional and sonic (musical, sound) experiences: {vocabulary}

Examples of the statements generated by the ChatGPT model are provided in Figure 28. As stated by the model itself “these examples incorporate both emotional and sonic elements, illustrating how metaphors can bridge the gap between visceral experiences and the auditory dimensions of life”.

⁴³ <https://www.reddit.com/>

⁴⁴ <https://www.quora.com/>

⁴⁵ <https://chat.openai.com/>

⁴⁶ <https://www.bing.com/copilot>

⁴⁷ <https://www.perplexity.ai/>

Bowel Movements; Bowel Sound:

The rhythm of life's bowel movements echoed through the routine, a daily cadence of existence.

Silence enveloped the room, interrupted only by the soft bowel sounds of nature.

Gastroenterology; Stomach; Gut:

In the symphony of bodily functions, the stomach played the role of a bass, its notes resonating through the digestive composition.

The gut, a conductor of internal orchestration, directed the movements of digestion with precision.

Spill the Gut; Stomach:

As the music played, emotions spilled from her like a melody, each note representing a chapter of her inner symphony.

He decided to spill the gut of creativity onto the canvas, letting the colors dance like stomach butterflies.

Figure 28: Metaphors generated by the ChatGPT

We checked that adding details to the instruction can affect the style of the generated text. Thus, by asking to give examples from literary works, the statements became rather brief (Figure 29).

Gut; Gut Feeling:

The decision churned in her gut like a stormy sea.

His gut feeling whispered warnings that couldn't be ignored.

Gutted; to Gut:

The revelation gutted her, leaving emotions scattered like broken glass.

To gut the truth from the lies, he had to navigate a maze of deception.

Gut Punch:

The news hit him like a gut punch, knocking the wind out of his sails.

The betrayal delivered a gut punch that echoed through the corridors of trust.

Punch:

Her words packed a punch, leaving an imprint on his soul.

The realization landed with a punch, shattering illusions he had held.

Stomach:

In the pit of her stomach, anxiety danced like a flickering flame.

His dreams settled in the depths of his stomach, aching for realization.

Figure 29: Metaphors generated by the ChatGPT as if inspired by the literary works

Looking at Bing's outputs we observe that depending on the prompt, its model can explain a keyword in detail (Figure 30), find related literature, and generate metaphors (Figure 31) pointing to sources to look for further information.

Gut Feeling:

This phrase refers to an intuitive sense or instinctive reaction. Interestingly, the gut is not just a digestive organ; it also houses a complex neural network that communicates with the brain. The gut microbiome—trillions of bacteria residing in our intestines—plays a role in influencing both physical and psychological well-being. [These gut-brain connections give rise to our “gut feelings” 12.](#)

Irritable Bowel Syndrome (IBS):

IBS is a common gastrointestinal disorder characterized by abdominal pain, bloating, and altered bowel habits. [Stress and emotional factors often exacerbate IBS symptoms 1.](#)

Learn more

<https://www.apa.org/monitor/2012/09/gut-feeling>

<https://neuroscience.ubc.ca/our-second-brain-more-than-a-gut-feeling/>

Figure 30: Bing Copilot’s explanations of concepts with references to the sources

Gutted:

Emotional: The loss left her heartstrings frayed, like a violin gutted of its sweet melodies.

Sonic: The room echoed with gutted laughter—a haunting sound devoid of joy.

Gut Punch:

Emotional: The breakup hit her like a gut punch, leaving emotional bruises that reverberated for months.

Sonic: The bass drop in the song delivered a gut punch, shaking the dance floor.

Gastroenterology:

Emotional: The study of emotions is like gastroenterology for the soul—exploring hidden depths.

Sonic: The doctor's voice, clinical and precise, dissected emotions like a sonogram.

Microbiome:

Emotional: Our inner microbiome—the orchestra of emotions, playing intricate compositions.

Sonic: The gut's microbial chorus hummed, shaping our well-being.

[Punch to the gut - Idioms by The Free Dictionary](#)

[GUT PUNCH | English meaning - Cambridge Dictionary](#)

[idioms - What's the meaning of "punch to the gut"? - English Language Learners Stack Exchange](#)

Figure 31: Metaphors generated using Bing Copilot's

Alternatively, Perplexity.ai generated explanations of how concepts can be used metaphorically with an emphasis on various aspects without providing examples (Figure 32). This gave us an intuition about the model's beliefs of how the lexicon would be incorporated in the discourse. Refining the prompt led to expected outcome with examples, however, many of them didn't include requested concepts but only captured the intended meaning (Figure 33).

Emotional Metaphors:

"Gut feeling" can be metaphorically linked to "intuitive resonance," conveying a sense of deep inner conviction or understanding.

"Gut punch" may metaphorically evoke "emotional impact," likening it to a powerful and jolting emotional experience.

"Gutted" could be metaphorically associated with "emotional devastation," reflecting a profound sense of loss or emptiness.

Sonic Metaphors:

"Gut feeling" might be metaphorically expressed as "musical intuition," akin to the way a musician senses the right notes or phrases without conscious thought.

"Gut punch" could be likened to a "sonic impact," reflecting the powerful and visceral nature of the sound or music.

"Gutted" may be metaphorically related to "sonic emptiness," conveying a sense of hollowness or absence in the auditory experience.

Figure 32: Explanations for possible metaphors generated by Perplexity.ai

Gutted:

Metaphor: "The news hit him like a wrecking ball, leaving behind a landscape of shattered hopes."

Explanation: This metaphor likens the emotional impact of the news to the destructive force of a wrecking ball, emphasizing the devastation and loss.

Gut Punch:

Metaphor: "The unexpected betrayal was a sledgehammer to his emotions, leaving him breathless in its wake."

Explanation: This metaphor equates the emotional blow of betrayal to the force of a sledgehammer, emphasizing its overwhelming and staggering effect.

Gut Microbiome:

Metaphor: "The *gut microbiome* is a bustling metropolis, with diverse communities working in harmony to maintain the city's balance."

Explanation: This metaphor compares the complex microbial communities in the gut to the diverse and interconnected population of a city, highlighting their collective role in maintaining equilibrium.

Gut Bacteria:

Metaphor: "The *gut bacteria* are like diligent gardeners, tending to the soil of our intestines and nurturing the seeds of health."

Explanation: This metaphor likens the role of gut bacteria to that of dedicated gardeners, emphasizing their nurturing and essential contribution to overall well-being.

Figure 33: Metaphors generated by Perplexity.ai

In addition, we used ChatGPT to look for candidate resources where metaphors can be found within human-produced texts. The outcome includes social platforms we already considered and some additional sources/databases that potentially can be useful to us:

Literary Databases:

- Project Gutenberg (<https://www.gutenberg.org/>)
- Google Books (<https://books.google.com/>)

Idioms and Proverbs:

- The Free Dictionary - Idioms (<https://idioms.thefreedictionary.com/>)
- Proverb Hunter (<https://www.proverbhunter.com/>)

Medical Literature:

- PubMed (<https://pubmed.ncbi.nlm.nih.gov/>)
- ScienceDirect (<https://www.sciencedirect.com/>)

Online Forums and Communities:

- Reddit (<https://www.reddit.com/>) - Explore subreddits related to health, emotions, and personal experiences.

- Quora (<https://www.quora.com/>)
- Creative Writing Platforms:
- Wattpad (<https://www.wattpad.com/>)
 - Writing.com (<https://www.writing.com/>)

Overall, our impression is that the models give quite generic answers relating to all these metaphors. However, the databases (such as Project Gutenberg) might be a good starting point. Thus, generated metaphors on their own are less applied to Lea's work and what is more relevant is how prominent these metaphors are within the mentioned databases. We will continue exploring various databases looking for discourse relevant to Lea's project. Metaphors generated by the LLMs so far will be helpful examples to narrow the scope of our search.

5.6 Future considerations

Considering the pipeline for extracting predicate-argument structures, we will investigate how to increase the F_1 -score for candidate selection on unseen data, for instance by using cross-encoders for the final ranking of the top candidates or NLI to filter out non-matching candidates (Abhishek et al., 2022). To find more and better candidates, we will apply coreference resolution on the Wikipedia pages, test open information extraction approaches to identify text spans (Perera and Nand, 2015), and explore the use of Simple Wikipedia (Duma and Klein, 2013) and WEXEA (Strobl et al., 2020). We will further develop our prototype clause extractor and apply our approach to other languages to test its portability.

As for the generation from multimodal fused information, we will perform experiments with real data produced within the Project and perform human evaluation to assess the relevance of the retrieved data, the accuracy of verbalisation of knowledge provided in abstract representation structures, and the degree of personalization and alignment with user requirements.

We will also explore the ways in which UPF can expand its collaborative effort with Artists who work within the ReSilence residencies.

6 CONCLUSIONS

In conclusion, the ReSilence project encompasses diverse methodologies aimed at understanding emotional responses to aesthetic and artistic stimuli. The utilisation of virtual reality (VR) environments in studying human reactions to social threats has demonstrated the feasibility of eliciting emotional responses in controlled settings. This approach not only informs the development of VR-based stimuli presentation methods for studying emotional responses but also emphasises the importance of personalised algorithms to account for individual differences in emotional processing. Additionally, the project's focus on sentiment analysis from social media data and reviews, multimodal analysis of sound-related experiences and behaviours, and self-report generation highlights the interdisciplinary nature of the research. By integrating various techniques such as entity disambiguation, aspect-oriented sentiment analysis, and multimodal measurement systems, ReSilence aims to provide robust insights into the correspondence between artistic compositions and listeners' perceptions, ultimately leading to more transparent decision-making processes in evaluating aesthetic experiences.

7 REFERENCES

- Abhishek, T., Sagare, S., Singh, B., Sharma, A., Gupta, M., & Varma, V. 2022, April. "Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages", In Companion Proceedings of the Web Conference 2022, p. 171-175.
- Avenanti, A., Annala, L., & Serino, A. 2012. Suppression of premotor cortex disrupts motor coding of peripersonal space. *Neuroimage*, 63(1), 281-288.
- Banerjee, S., & Lavie, A. 2005, June. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments", In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, p. 65-72.
- Blanchard, R. J., Flannelly, K. J., & Blanchard, D. C. 1986. "Defensive behaviors of laboratory and wild *Rattus norvegicus*", *Journal of comparative psychology*, 100(2), 101.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. 2008. "Fast unfolding of communities in large networks", *Journal of statistical mechanics: theory and experiment*, 2008 (10), P10008.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. 2017. "Enriching word vectors with subword information", *Transactions of the association for computational linguistics*, 5, p. 135-146.
- Bufacchi, R. J., & Iannetti, G. D. 2018. "An action field theory of peripersonal space", *Trends in cognitive sciences*, 22(12), p. 1076-1090.
- Chai, X., Liu, M., Huang, T., Wu, M., Li, J., Zhao, X., ... and Zhang, Y. X. (2023). "Neurophysiological evidence for goal-oriented modulation of speech perception", *Cerebral Cortex*, 33 (7), p. 3910-3921.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L. and Stoyanov, V., 2020, July. "Unsupervised Cross-lingual Representation Learning at Scale", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 8440-8451.
- Cunningham, W. A., Espinet, S. D., DeYoung, C. G., & Zelazo, P. D. 2005. "Attitudes to the right- and left: frontal ERP asymmetries associated with stimulus valence and processing goals"; *NeuroImage*, 28(4), 827-834.
- Daniel Duma and Ewan Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers, pages 83–94.
- de Borst, A. W., and de Gelder, B. 2022. "Threat detection in nearby space mobilizes human ventral premotor cortex, intraparietal sulcus, and amygdala", *Brain Sciences*, 12(3), 391.
- de Borst, A. W., Sanchez-Vives, M. V., Slater, M., and de Gelder, B. 2020. "First-person virtual embodiment modulates the cortical network that encodes the bodily self and its surrounding space during the experience of domestic violence", *Eneuro*, 7(3).
- De Gelder, B., Snyder, J., Greve, D., Gerard, G., and Hadjikhani, N. 2004. "Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body", *Proceedings of the National Academy of Sciences*, 101(47), 16701-16706.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. "BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding”, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171-4186.

Di Pellegrino G, and Làdavas E. 2015. *“Peripersonal space in the brain”*, *Neuropsychologia*. 66, p.126–133.

Duma, D., and Klein, E. 2013. *“Generating natural language from linked data: Unsupervised template extraction”*, In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)--Long Papers (pp. 83-94). ASSOC COMPUTATIONAL LINGUISTICS-ACL.

Eilam, D. 2005. *“Die hard: a blend of freezing and fleeing as a dynamic defense—implications for the control of defensive behavior”*, *Neuroscience & Biobehavioral Reviews*, 29 (8), p. 1181-1191.

Ell, B. and Harth, A. 2014, June. *“A language-independent method for the extraction of RDF verbalization templates”*, In Proceedings of the 8th international natural language generation conference (INLG), p. 26-34.

Ellena, G., Starita, F., Haggard, P., Romei, V., and Làdavas, E. 2021. *“Fearful faces modulate spatial processing in peripersonal space: An ERP study”*, *Neuropsychologia*, 156, 107827.

Ferreira, T. C., Gardent, C., Ilinykh, N., Van Der Lee, C., Mille, S., Moussallem, D., and Shimorina, A. 2020, December. *“The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020)”*, In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+).

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... and Neubig, G. 2023, July. *“Pal: Program-aided language models”*, In International Conference on Machine Learning. p. 10764-10799,. PMLR.

George, D. T., Ameli, R., & Koob, G. F. 2019. *“Periaqueductal gray sheds light on dark areas of psychopathology”*, *Trends in neurosciences*, 42 (5), p. 349-360.

Gladwin, T. E., Hashemi, M. M., van Ast, V., & Roelofs, K. 2016. *“Ready and waiting: Freezing as active action preparation under threat”*, *Neuroscience Letters*, 619, p. 182-188.

Goldzycher, J., & Schneider, G. 2022. *“Hypothesis engineering for zero-shot hate speech detection”*, arXiv preprint arXiv:2210.00910.

Gotkova, T., & Chepurnykh, N. 2022. *“Public perception and usage of the term: Linguistic analysis in an environmental social media corpus”*, *Psychology of Language and Communication*, 26 (1), p. 297-312.

Gotkova, T., and Shvets, A. 2023. *“Key Environmental Lexicon Extraction Using Generative Transformer”* (Short Paper).

Hagenaars, M. A., Roelofs, K., and Stins, J. F. 2014. *“Human freezing in response to affective films”*, In *Anxiety, Stress & Coping*, 27(1), p. 27-37.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. 2022. *“Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection”*, arXiv preprint arXiv:2203.09509.

Hasselqvist, J., Helmertz, N., and Kågebäck, M. 2017. *“Query-based abstractive summarization*

using neural networks”, arXiv preprint arXiv:1712.06100.

He, W. Q., Luo, W. B., He, H. M., Chen, X., and Zhang, D. J. 2011. “N170 effects during exact and approximate calculation tasks: an ERP study”, *Neuroreport*, 22 (9), p. 437-441.

Hoang, T. T., Gelbukh, A., & Sidorov, G. 2022. “Mapping Process for the Task: Wikidata Statements to Text as Wikipedia Sentences”, arXiv e-prints, arXiv-2210.

Iachini, T., Coello, Y., Frassinetti, F., Senese, V. P., Galante, F., and Ruggiero, G. 2016. “Peripersonal and interpersonal space in virtual and real environments: Effects of gender and age”, *Journal of Environmental Psychology*, 45, p. 154-164.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... and Sayed, W. E. 2023. “Mistral 7B”, arXiv preprint arXiv:2310.06825.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... and Sayed, W. E. 2024. “Mixtral of experts”, arXiv preprint arXiv:2401.04088.

Jin, Y., Wanner, L., and Shvets, A. 2024. “GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection?”, arXiv preprint arXiv:2402.15238.

Joyce C, Rossion B. The face-sensitive N170 and VPP components manifest the same brain processes: the effect of reference electrode site. *Clin Neurophysiol*. 2005;116(11):2613–2631.

Kaffee, L. A., Vougiouklis, P., & Simperl, E. 2022. “Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective”, *Semantic Web*, 13(2), p. 163-194.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... and Yih, W. T. 2020. “Dense passage retrieval for open-domain question answering”, arXiv preprint arXiv:2004.04906.

Kingsbury, P. R., and Palmer, M. 2002, May. “From TreeBank to PropBank”, In LREC, pp. 1989-1993.

LeDoux, J., and Daw, N. D. 2018. “Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour”, *Nature Reviews Neuroscience*, 19 (5), p. 269-282.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... and Bizer, C. 2015. “Dbpedia—a large-scale, multilingual knowledge base extracted from Wikipedia”, *Semantic web*, 6(2), p. 167-195.

Leszczynski, M., Fu, D. Y., Chen, M. F., and Ré, C. 2022. “Tabi: Type-aware bi-encoders for open-domain entity retrieval”, arXiv preprint arXiv:2204.08173.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... and Kiela, D. 2020. “Retrieval-augmented generation for knowledge-intensive nlp tasks”, *Advances in Neural Information Processing Systems*, 33, p. 9459-9474.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... and Zitnick, C. L. 2014. “Microsoft coco: Common objects in context”, In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, p. 740-755, Springer International Publishing.

Livermore, J. J., Klaassen, F. H., Bramson, B., Hulsman, A. M., Meijer, S. W., Held, L., ... and Roelofs, K. 2021. “Approach-avoidance decisions under threat: the role of autonomic

psychophysiological states”, *Frontiers in Neuroscience*, 15, 621517.

Luo, W., Feng, W., He, W., Wang, N. Y., and Luo, Y. J. 2010. “*Three stages of facial expression processing: ERP study with rapid serial visual presentation*”, *Neuroimage*, 49(2), p. 1857-1867.

Mello, M., Dupont, L., Engelen, T., Acciarino, A., de Borst, A. W., and de Gelder, B. 2022. “*The influence of body expression, group affiliation and threat proximity on interactions in virtual reality*”, *Current Research in Behavioral Sciences*, 3, 100075.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. 2011, September. “*DBpedia spotlight: shedding light on the web of documents*”, In *Proceedings of the 7th international conference on semantic systems*, p. 1-8).

Meng, Y., Huang, J., Zhang, Y., and Han, J. 2022. “*Generating training data with language models: Towards zero-shot language understanding*”, *Advances in Neural Information Processing Systems*, 35, p. 462-477.

Mille, S., Dasiopoulou, S., Fisas, B., and Wanner, L. 2019. “*Teaching FORGe to verbalize DBpedia properties in Spanish*”, In *Proceedings of the 12th International Conference on Natural Language Generation*, p. 473-483.

Mobbs, D., and Kim, J. J. 2015. “*Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans*”, *Current opinion in behavioral sciences*, 5, p. 8-15.

Monti, A., & Aglioti, S. M. 2018. “*Flesh and bone digital sociality: On how humans may go virtual*”, *British Journal of Psychology*, 109 (3), p. 418-420.

Ocampo, N. B., Cabrio, E., & Villata, S. 2023, July. “*Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection*”, In *Findings of the Association for Computational Linguistics: ACL 2023*, p. 2758-2772.

Oguz B., Chen X., Karpukhin V., Peshterliev S., Okhonko D., Michael Schlichtkrull M., Gupta S., Mehdad Y., and Yih S. 2022. “*UniK-QA: Unified representations of structured and unstructured knowledge for open-domain question answering*”, In *Findings of the Association for Computational Linguistics: NAACL 2022*, p. 1535–1546.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... and Lowe, R. 2022. “*Training language models to follow instructions with human feedback*”, *Advances in neural information processing systems*, 35, p. 27730-27744.

Parsons, T. D., Gaggioli, A., & Riva, G. 2017. “*Virtual reality for research in social neuroscience*”, *Brain sciences*, 7(4), p. 42.

Pasricha, N., Arcan, M., and Buitelaar, P. 2020. “*NUIG-DSI at the WebNLG+ challenge: Leveraging transfer learning for RDF-to-text generation*”, In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, p. 137-143.

Pennington, J., Socher, R., and Manning, C. D. 2014, October. “*Glove: Global vectors for word representation*”, In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543.

Perera, R., and Nand, P. 2015. “*A multi-strategy approach for lexicalizing linked open data*”, In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II* 16, p. 348-363, Springer

International Publishing.

Pilia, N., Nagel, C., Lenis, G., Becker, S., Dössel, O., and Loewe, A. 2021. *“ECGdeli-An open source ECG delineation toolbox for MATLAB”*, SoftwareX, 13, 100639.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. 2020. *“Stanza: A Python natural language processing toolkit for many human languages”*, arXiv preprint arXiv:2003.07082.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... and Liu, P. J. 2020. *“Exploring the limits of transfer learning with a unified text-to-text transformer”*, Journal of machine learning research, 21 (140), p. 1-67.

Reimers, N., and Gurevych, I. 2019. *“Sentence-bert: Sentence embeddings using siamese bert-networks”*, arXiv preprint arXiv:1908.10084.

Roelofs, K., Hagenars, M. A., and Stins, J. 2010. *“Facing freeze: social threat induces bodily freeze in humans”*, Psychological science, 21(11), p. 1575-1581.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. 2020. *“HateCheck: Functional tests for hate speech detection models”*, arXiv preprint arXiv:2012.15606.

Ruggiero, G., Rapuano, M., Cartaud, A., Coello, Y., and Iachini, T. 2021. *“Defensive functions provoke similar psychophysiological reactions in reaching and comfort spaces”*, Scientific reports, 11(1), 5170.

Seinfeld, S., Bergstrom, I., Pomes, A., Arroyo-Palacios, J., Vico, F., Slater, M., and Sanchez-Vives, M. V. 2016. *“Influence of music on anxiety induced by fear of heights in virtual reality”*, Frontiers in psychology, 6, 169495.

Seinfeld, S., Zhan, M., Poyo-Solanas, M., Barsuola, G., Vaessen, M., Slater, M., ... and de Gelder, B. 2021. *“Being the victim of virtual abuse changes default mode network responses to emotional expressions”*, cortex, 135, p. 268-284.

Serino, A. 2019. *“Peripersonal space (PPS) as a multisensory interface between the individual and the environment, defining the space of the self”*, Neuroscience & Biobehavioral Reviews, 99, p. 138-159.

Shvets, A., & Wanner, L. 2020. *“Concept extraction using pointer-generator networks and distant supervision for data augmentation”*, In International Conference on Knowledge Engineering and Knowledge Management, p. 120-135.

Shvets, A., and Wanner, L. 2022. *“The relation dimension in the identification and classification of lexically restricted word co-occurrences in text corpora”*, Mathematics, 10 (20), 3831.

Stathopoulos, E. A., Shvets, A., Carlini, R., Diplaris, S., Vrochidis, S., Wanner, L., & Kompatsiaris, I. 2022, June. *“Social Media and Web Sensing on Interior and Urban Design”*, In 2022 IEEE Symposium on Computers and Communications (ISCC), p. 1-6.

Stekelenburg, J. J., and de Gelder, B. 2004. *“The neural correlates of perceiving human bodies: an ERP study on the body-inversion effect”*, Neuroreport, 15 (5), p. 777-780.

Stins, J. F., Roelofs, K., Villan, J., Kooijman, K., Hagenars, M. A., and Beek, P. J. 2011. *“Walk to me when I smile, step back when I’m angry: emotional faces modulate whole-body approach-avoidance behaviors”*, Experimental brain research, 212, p. 603-611.

- Stolz C, Endres D, and Mueller E. M. 2019. *“Threat-conditioned contexts modulate the late positive potential to faces—a mobile EEG/virtual reality study”*, *Psychophysiology*, 56 (4), e13308.
- Strobl, M., Trabelsi, A., and Zaiane, O. R. 2020, May. *“WEXEA: Wikipedia exhaustive entity annotation”*, In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1951-1958.
- Terburg, D., Scheggia, D., Del Rio, R. T., Klumpers, F., Ciobanu, A. C., Morgan, B., ... and van Honk, J. 2018. *“The basolateral amygdala is essential for rapid escape: a human and rodent study”*, *Cell*, 175 (3), p. 723-735.
- Van der Maaten, L., and Hinton, G. 2008. *“Visualizing data using t-SNE”*, *Journal of machine learning research*, 9(11).
- Van Heijnsbergen, C. C. R. J., Meeren, H. K. M., Grezes, J., and De Gelder, B. 2007. *“Rapid detection of fear in body expressions, an ERP study”*, *Brain research*, 1186, p. 233-241.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. 2017. *“Attention is all you need”*, *Advances in neural information processing systems*, 30.
- Vieira, J. B., Pierzchajlo, S. R., & Mitchell, D. G. 2020. *“Neural correlates of social and non-social personal space intrusions: Role of defensive and peripersonal space systems in interpersonal distance regulation”*, *Social Neuroscience*, 15 (1), p. 36-51.
- Wendt J., Löw A., Weymar M., Lotze M., and Hamm A. O. 2017. *“Active avoidance and attentive freezing in the face of approaching threat”*, *NeuroImage*, 158, p.196–204.
- Williams, A., Nangia, N., & Bowman, S. R. 2017. *“A broad-coverage challenge corpus for sentence understanding through inference”*, arXiv preprint arXiv:1704.05426.
- Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., and Ma, H. 2020. *“Clear: Contrastive learning for sentence representation”*, arXiv preprint arXiv:2012.15466.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... and Raffel, C. 2020. *“mT5: A massively multilingual pre-trained text-to-text transformer”*, arXiv preprint arXiv:2010.11934.
- Ye, X., and Durrett, G. 2022. *“The unreliability of explanations in few-shot prompting for textual reasoning”*, *Advances in neural information processing systems*, 35, p. 30378-30392.
- Yin, W., Hay, J., and Roth, D. 2019. *“Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach”*, arXiv preprint arXiv:1909.00161.
- Zhang, C., Ding, W., Peng, G., Fu, F., and Wang, W. 2020. *“Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems”*, *IEEE Transactions on Intelligent Transportation Systems*, 22 (7), p. 4727-4743.
- Zhang, R., Li, Y., Ma, Y., Zhou, M., and Zou, L. 2023. *“Llmeta: Making large language models as active annotators”*, arXiv preprint arXiv:2310.19596.
- Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. 2018, June. *“Texygen: A benchmarking platform for text generation models”*, In *The 41st international ACM SIGIR*

conference on research & development in information retrieval, p. 1097-1100.