

# What makes the audience engaged?

## Engagement prediction exploiting multimodal features

Daniele Borghesi<sup>1</sup>, Andrea Amelio Ravelli<sup>2</sup> and Felice Dell’Orletta<sup>2</sup>

<sup>1</sup>Università di Pisa, Lungarno Pacinotti 43, 56126, Pisa, Italy

<sup>2</sup>Istituto di Linguistica Computazionale “A. Zampolli” (ILC-CNR), Via Giuseppe Moruzzi 1, 56124, Pisa, Italy

### Abstract

This paper reports a series of experiments and analyses aimed at understanding which, among numerous linguistic and acoustic aspects of the spoken language, are distinctive in the detection of an engagement potential within speech. Starting from a dataset consisting of numerous sentences, pronounced during guided sightseeing tours, and characterised by a set of multimodal features, various classification algorithms were tested and optimised in different scenarios and configurations. Thanks to the implementation of a recursive feature elimination algorithm, it has been possible to select and identify which characteristics of the language play a key role in the presence of an engagement potential, and which can thus differentiate an engaging sentence or speech from a non-engaging one. The analyses on the selected features showed that, among the strictly linguistic aspects, only basic features (i.e. sentence or word length) proved to be relevant in the classification process. In contrast, aspects of acoustic nature showed to play a considerably important role, in particular aspects related to sound spectrum and prosody. Overall, a feature selection led to appreciable increases in the performance of all implemented classification models.

### Keywords

multimodal dataset, feature selection, engagement prediction, audience engagement

## 1. Introduction and motivation

In recent years we have witnessed to major advances in Artificial Intelligence and Natural Language Processing, to the point that we now have models capable to write complete (and most of all, sounding) pieces of text out of a simple prompt [2, 3]. The ability to generate content is impressive, but the scope of a text is often beyond the pure information conveyed with it. Nevertheless, the effectiveness of information transfer is often due to the willingness of the receiver to accept it. This is particularly evident if we move our focus from the written page to more interactive communication media and channels, such as face-to-face interactions. In fact, the average (human) speaker is generally very good at estimating the interlocutor’s

---

NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy [1]

✉ d.borghesi@studenti.unipi.it (D. Borghesi); andreaamelio.ravelli@ilc.cnr.it (A. A. Ravelli);


felice.dellorletta@ilc.cnr.it (F. Dell’Orletta)

🌐 <http://www.ilc.cnr.it/content/andrea-amelio-ravelli> (A. A. Ravelli);

<http://www.ilc.cnr.it/content/felice-dellorletta> (F. Dell’Orletta)

🆔 0000-0002-0979-0585 (D. Borghesi); 0000-0002-0232-8881 (A. A. Ravelli); 0000-0003-3454-9387 (F. Dell’Orletta)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

level of involvement from visually accessible signals (e.g. body postures and movements, facial expressions, eye-gazes), and at refining his/her communication strategy, in order to keep the communication channel open and the attention high in the audience. Such visible cues are mostly signals of attention, which is considered as a perceivable proxy to broader and more complex inner processes of engagement [4]. Moreover, recent studies have shown that the processing of *emotionality* in the human brain is performed on modality-specific basis [5]: prosody, facial expressions and speech content (i.e. the semantic information) are processed in the listener's brain with the selective activation of the auditory cortex, the fusiform gyri and the middle temporal gyri, respectively.

Understanding of non-verbal feedback is not easy to achieve for virtual agents and robots, but this ability is strategic for enabling more natural interfaces capable of adapting to users. Indeed, perceiving signals of loss of attention (and thus, of engagement) is of paramount importance to design naturally behaving virtual agents, enabled to adjust the communication strategy to keep high the interest of their addressees. That information is also a general sign of the quality of the interaction and, more broadly, of the communication experience. At the same time, the ability to generate engaging behaviours in an agent can be beneficial in terms of social awareness [6].

The objective of the present work is to understand the phenomena correlated to the increase or decrease of perceivable engagement in the audience of a speech, specifically in the domain of guided tours. We are interested in highlighting which features, from which specific modality, have a key role in driving the attention in the listener(s), in order to exploit a reduced set of features as dense but highly informative representations.

## 1.1. Related Work

With the word engagement we refer to the level of involvement reached during a social interaction, which assumes the shape of a process through the whole communication exchange. More specifically, [7] defines the process of social engagement as the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing the interaction. Another definition, adopted by many studies in Human-Robot Interaction (HRI),<sup>1</sup> describes engagement as the process by which interactors start, maintain, and end their perceived connections to each other during an interaction [8]. The majority of the studies are often conducted on a dyadic base (i.e. one-to-one) in context where one of the participants is often an agent/robot [9, 10, 11]. Nevertheless, engagement can be measured in groups of people as the average of the degree to which individuals are involved [12, 13, 14].

## 2. Dataset

Data for the experiments described herein derive from a subset of the data collected for the CHROME Project<sup>2</sup> [15, 16]. The domain of the data is Cultural Heritage; more specifically, the project has been focused on guided tours in 3 Charterhouses in Campania (Italy), where an expert historian led groups of 4 persons. Tours are organised in 6 Points of Interest (POI),

---

<sup>1</sup>For a broad and complete overview of works on engagement in HRI studies, see [6].

<sup>2</sup>Cultural Heritage Resources Orienting Multimodal Experience. <http://www.chrome.unina.it/>

i.e. rooms or areas inside the Charterhouses where the visits stop and the guide describes the place with its furnishings, history and anecdotes. The communication event type is *quasi-unidirectional* (one-to-many), i.e. one of the participants is the holder of the knowledge (the expert guide) and talks to the others (the audience), with a few moments of dialogue (e.g. when the guide asks something to the audience).

The original data collection campaign led to a multimodal corpus with aligned transcriptions, audios and videos. From this, we selected a subset composed of 3 visits (i.e. 3 different groups of 4 persons) led by the same expert guide inside one of the three Charterhouses (San Martino Charterhouse, Naples). Given the exploratory objective of the present study, we made this selection in order to leverage differences such as voice features and discourse style, which are speaker-specific.

The final set of data on which we run our experiments is composed of 1,114 sentences, enriched with the annotation of the perceivable engagement of the audience, and characterised with a total of 452 features extracted from multiple modalities (127 linguistic, 325 acoustic) used to model the speech of the guide. The process through which we obtained our dataset is described in the following.

## 2.1. Human engagement annotation

We considered the attention of the audience as a perceivable proxy to model and highlight participants' engagement, in line with the assumption that engagement is a complex process, a multidimensional meta-construct [17] composed of behavioural, emotional and cognitive aspects. Behavioural aspects (and some externalisation of emotional states) can be tracked by observing the subject, while for the others it is necessary to exploit specific equipment to record biomarkers such as heart rate or neural activity. Nevertheless, all aspects of engagement are highly interrelated and do not occur in isolation, thus attention plays a crucial role in defining if the audience is engaged or not [18].<sup>3</sup>

To annotate audience engagement, we exploited the visual part of the original CHROME dataset, consisting of 2 parallel video recordings for each visit: one focused on the speaker, the other on the audience. We asked 2 annotators to watch at the same time the audience and the guide videos, with the guide video in a small window superimposed on the audience one, and to annotate the level of attention and its variation among the attendee. We recorded this information by means of PAGAN Annotation Tool [19], that enables the annotator to easily track the observed phenomenon with a simple press of two keys on the keyboard: arrow-up if a rise is perceived, arrow-down otherwise. Our annotators reached a high agreement on this task, with an average Spearman's rho of 0.87.<sup>4</sup> The resulting annotation is a continuous series of values, indicating rise or fall of engagement along the whole visit, for all the visits in our dataset.

---

<sup>3</sup>We will continue to use the term *engagement* referencing to the perceivable attention of the audience.

<sup>4</sup>For a more detailed description of the annotation process, see [20].

## 2.2. Sentence segmentation

We acquired the textual data in the form of ELAN annotation files [21], containing the orthographic transcription of single words tagged with their start- and end-time (in milliseconds), aligned on the timeline of the whole speech. In order to obtain more exploitable units of text, we split the flow of the speech in *sentence-like* segments, by concatenating together all the words that can represent a finite unit of language.<sup>5</sup> We asked two annotators to segment our texts, relying on a pure perceptual principle: mark the end of a sentence whenever conceptual completeness is perceived. We relied on the capability of mothertongue speakers of Italian to mentally segment the flow of the speech, the same as we normally do during everyday conversations. In other words, we asked the annotators to identify terminal breaks and mark them with a full stop. Given that punctuation is a convention of the written medium the annotators were asked to minimise the use of it, but beside the full stop we allowed for the use of commas to signal short pauses or listings, and question marks when a questioning intonation was identified.

A limitation to this methodology is that it is often possible that the speech rate makes difficult to finely segment, especially taking into account the necessity to propagate the segmentation from the text to the audio files. In fact, we projected the start-end spans of sentences onto the audio files in order to obtain the audio objects from where we extracted acoustic features. We kept together in the same text/audio object multiple sentences if uttered at high rate and difficult to cleanly separate on the audio level, in order to avoid noise that would have altered the computing of acoustic features.

We measured the accuracy of the segmentation on a portion of the data (about the 40% of the total) by adapting an IOB (Inside-Outside-Begin) tagging framework. We labelled all the tokens, according to each annotator, on the basis of their position at the beginning (B), the inside (I), the end (E) or the outside (O) of a constructed sentence. By applying this annotation, we registered an agreement of 91.53% in terms of accuracy on the basis of the two series of labelled tokens, thus the obtained segments can be considered reliable and consistent.

## 2.3. Engagement projection on sentences

As anticipated in 2.1, the engagement annotation consists of a continuous series of values along the timeline of each video/visit: we dispose of a numerical value indicating the level of engagement for each instant in which the latter has changed. Our aim was to use these values to extract the level of engagement for each individual sentence; for this, we aggregated all values within the span of the segmented sentences, in order to adapt the continuous annotation of the engagement to discrete units (i.e. the sentences), by translating those values into finite classes: *engaging* (associated with class 1) vs. *non-engaging* (associated with class 0). In this regard, two different aggregation methods were designed and implemented: by subtraction and by summation.

---

<sup>5</sup>Speech segmentation is not a trivial task, and many researchers debated (and they are still debating) on the problem. A recent special issue on the topic has been collected in [22].

### 2.3.1. Aggregation by subtraction

By using the subtraction method, we considered the delta between the first and the last value of engagement annotated in the time span of a sentence. Considering the time interval of a sentence  $S$ , where  $n$  values of engagement were annotated (one for each variation), to obtain the engagement level  $E_S$  of an entire sentence we subtracted the first engagement value ( $e_0$ ) from the last one ( $e_n$ ), as illustrated by equation 1:

$$E_S = e_n - e_0 \quad (1)$$

### 2.3.2. Aggregation by summation

By using the summation method, all the values and variations in the series of engagement values, within the time span of a sentence, are taken into account. Considering a series of  $n$  values of engagement (one for each variation) annotated within the time interval of a sentence  $S$ , a cumulative sum was calculated, to which 1 was added in the case where an increase in the level of engagement ( $e_n > e_{n-1}$ ) occurred, while  $-1$  was added in the case where, on the other hand, a decrease in the level of engagement ( $e_n < e_{n-1}$ ) occurred. The final result of the sum allows us to obtain the level of engagement  $E_S$  of an entire sentence, as illustrated by the equation 2, based on the system of equations 3:

$$E_S = \sum_{i=1}^n a_i \quad (2)$$

$$a_i = \begin{cases} 1, & \text{if } e_i > e_{i-1} \\ -1, & \text{if } e_i < e_{i-1} \end{cases} \quad (3)$$

### 2.3.3. Engagement thresholds

After computing the engagement level for each sentence, we further converted these values to Boolean classes ( $C_s$ ): 1 if resulting *engaging*, 0 if *non-engaging*. We considered 3 thresholds as different degrees of *inclusiveness*:

- $-1$ , to generate a more *generous* classification;
- $0$ , to generate a more *balanced* classification;
- $+1$ , to generate a more *sceptical* classification.

Every sentence with an engagement level  $E_s$  above the threshold  $t$  was considered engaging, while the others were considered non-engaging, as illustrated by the system of equations 4:

$$C_S = \begin{cases} 1, & \text{if } E_S > t \\ 0, & \text{if } E_S \leq t \end{cases} \quad (4)$$

In conclusion, we obtain six different sentence classification series: three series (one for each engagement threshold) for each of the two aggregation methodology. The selection of the most suitable series is specified within the section 3.4.

## 2.4. Features

In this section we describe the methodology and the tools used to extract features for both the textual and acoustic modality. We relied on explicit feature extraction systems in order to explore which specific features, and to which extend, convey the most of the information that create an engagement status in the audience.<sup>6</sup>

### 2.4.1. Linguistic Features

**Table 1**

Set of linguistic features extracted with Profiling-UD.

Linguistic features	n
Raw text properties	2
Morpho-syntactic information	52
Verbal predicate structure	10
Parsed tree structures	15
Syntactic relations	38
Subordination phenomena	10
<b>Total</b>	<b>127</b>

The textual modality has been encoded by using Profiling-UD [26], a publicly available web-based application<sup>7</sup> inspired to the methodology initially presented in [27], that performs linguistic profiling of a text, or a large collection of texts, for multiple languages. The system, based on an intermediate step of linguistic annotation with UDPipe [28], extracts a total of 129 features per each analysed document. In this case, Profiling-UD analysis has been performed per sentence, thus the output has been considered as the linguistic feature set of each segment of the dataset. Table 1 reports the 127 features extracted with Profiling-UD and used as textual modality features for the classifier.<sup>8</sup>

### 2.4.2. Acoustic Features

The acoustic modality has been encoded using OpenSmile<sup>9</sup> [29], a complete and open-source toolkit for analysis, processing and classification of audio data, especially targeted at speech and music applications such as automatic speech recognition, speaker identification, emotion recognition, or beat tracking and chord detection. The acoustic features set used in this case is the Computational Paralinguistics Challenge<sup>10</sup> (ComParE), which comprises 65 Low-Level Descriptors (LLDs), computed per frame.

---

<sup>6</sup>The current state of the art in both linguistic and acoustic feature extraction make use of recent Deep Learning methods and technique [23, 24, 25], but those systems extract features that are by nature not explainable.

<sup>7</sup>Profiling-UD can be accessed at the following link: <http://linguistic-profiling.italianlp.it>

<sup>8</sup>Out of the 129 Profiling-UD features, *n\_sentences* and *tokens\_per\_sent* (raw text properties) have not been considered, given that the analysis has been performed per sentence.

<sup>9</sup><https://www.audeering.com/research/opensmile/>

<sup>10</sup><http://www.compare.openaudio.eu>

**Table 2**

Set of acoustic features extracted with OpenSmile.

<b>Acoustic features</b>	<b>n</b>
<b><i>Prosodic</i></b>	
F0 (SHS and viterbi smoothing)	1
Sum of auditory spectrum (loudness)	1
Sum of RASTA-style filtered auditory spectrum	1
RMS energy, zero-crossing rate	2
<b><i>Spectral</i></b>	
RASTA-style auditory spectrum, bands 1–26 (0–8 kHz)	26
MFCC 1–14	14
Spectral energy 250–650 Hz, 1 k–4 kHz	2
Spectral roll off point 0.25, 0.50, 0.75, 0.90	4
Spectral flux, centroid, entropy, slope	4
Psychoacoustic sharpness, harmonicity	2
Spectral variance, skewness, kurtosis	3
<b><i>Sound quality</i></b>	
Voicing probability	1
Log. HNR, Jitter (local, delta), Shimmer (local)	4
<b>Total</b>	<b>65</b>

Table 2 reports a summary of the ComParE LLDs extracted with OpenSmile, grouped by type: prosody-related, spectrum-related and quality-related. Given that the duration (and number of frames, consequently) of audio segments varies, common transformations (min, max, mean, median, std) have been applied on the set of per-frame features of each segment, leading to a total of 325 acoustic features (65 LLDs x 5 transformations).

### 3. Experimental setting

In order to explore multiple methodologies and techniques to study the task of engagement potential prediction, we set our experiments in different classification scenarios, exploiting two Machine Learning models, applying alternative feature normalisation and engagement class assignment methods, and executing a selection of the most representative features to predict the engagement potential of a sentence.

#### 3.1. Classification scenarios and baseline

Dealing with a few data, as in this case (1,114 total items), may lead to an overestimation of the classification performances, making the predictions unreliable, especially if relying on a simple train-validation split of the dataset [30, 31, 32]. To avoid this, we opted for a Cross-Validation approach [33, 34, 35], declining our experimentation in 3 classification scenarios:

- By *stratified Random Partitioning (RaP)*: the dataset is divided into 10 equally sized parts, composed of randomly extracted elements. The stratified approach makes it possible to



maintain the same proportion between classes in the dataset even in individual subdivisions; this is possible exploiting a Stratified Cross-Validation technique;<sup>11</sup>

- By *Visits (Vis)*: the dataset is divided on the basis of tourist visits, thus obtaining three partitions, related to the three visits considered;
- By *Points Of Interest (POI)*: the dataset is partitioned on the basis of Points of Interest, thus obtaining six partitions, based on the POIs taken into consideration.

It is important to specify that, in each classification scenario, an unseen part of the dataset (a *test-set*) has been kept aside until the conclusion of the study, in order to ultimately test the performance of the fully optimised system on unknown data. For the *RaP* scenario we excluded from the Cross-Validation a portion of 20% of the data, which is also stratified. In the case of the *Vis* scenario, the test-set is represented by the data related to the first visit (V01), while for the *POI* scenario, the test-set is represented by the data related to the the first point of interest (P01).

For each classifier, and in each scenario, we trained 3 different models, namely *Multimodal*, *Linguistic* and *Acoustic*, on the basis of the type (or the combination of types) of features used as training. We decided also to calculate and use a baseline for each validation-set and each test-set: each sentence in the set was assigned the *Most Frequent Class* within the respective training-set. The individual baselines can be found in the appendices, where we report tables with details of every experiment we run in this work, with the figures of each baseline.

## 3.2. Classifiers

One of the primary objectives of the study is to obtain a model capable of classifying a sentence as either engaging or not engaging. To achieve this goal, as anticipated, we selected two Machine Learning models: *Linear Support Vector Classifier* [36, 37] (Linear-SVC) and *Random Forest Classifier* [38, 39] (Random-Forest). Choosing two radically different classifiers, rather than using a single one, allows us to perform an accurate comparison between two different classification processes, in terms of behaviour and performances. Most important, we relied on fully explainable classification models, where it is possible to work with explicit features, thus focusing on the phenomena behind a decision.

More precisely, at the feature selection stage, it will be possible to highlight which feature categories were deemed important by both classifiers, i.e. could be considered relevant for detecting an engagement potential in language. Indeed, both classifiers are able to sort the training features on the basis of their influence in the classification process, assigning them a rank [40, 41, 42, 39] that can be used for performing feature selection and subsequent in-depth analysis.

### 3.2.1. Hyperparameters tuning

A very important aspect in setting up the classifiers is the optimisation of the hyperparameters: the Machine Learning models, in fact, have several hyperparameters that can be modified to improve classification performance, allowing of more accurate results [43, 44, 45, 46].

---

<sup>11</sup>[https://scikit-learn.org/stable/modules/cross\\_validation.html#stratified-k-fold](https://scikit-learn.org/stable/modules/cross_validation.html#stratified-k-fold)



A complete engineering of the chosen models would have been outside the objectives of the study, thus we choose to optimise exclusively the most relevant hyperparameter in each of the two chosen classifiers:

- For Linear-SVC, the regularisation parameter (commonly referred to as parameter C) was optimised by testing a range of values (0.001, 0.01, 0.10, and 1.00) [47];
- For the Random-Forest, the number of decision trees (Decision-Trees) that make up the "forest" was optimised. In this case, a number of trees equal to 10, to 100, and to 1000 was tested.

The hyperparameters tuning results showed that the Linear-SVC achieved the best performance by using the regularisation parameter of 0.001, while the Random-Forest scores best with a Decision-Tree number of 1000. Detailed results relative to hyperparameters tuning, on a cross-comparison with the aggregation methods explained in section 2.3, can be found in Appendix A (tables 3 and 4).

### 3.3. Feature normalisation

Standardising and normalising data (e.g., scaling within a common numerical range) can benefit the training and performance of Machine Learning models [48]. In this regard, we tested many normalisation methods, that we can divide in two main groups:

- Linear normalisation methods: Standard-Scaler (StaS), Max-Abs-Scaler (MAS), Min-Max-Scaler (MiMaS) with two different numerical ranges (0 to 1, and -1 to 1), and Robust-Scaler (RoS);
- Nonlinear normalisation methods: Power-Transformer (PoT) and Quantile-Transformer (QuT).

In our experimentation, no appreciable differences emerged in terms of accuracy between all the normalisation methods. However, the Quantile-Transformer (QuT) provided slightly best overall results, thus it has been selected as default for the subsequent experiments. All the results relative to the comparison between data normalisation methods, for both the classifiers, can be found in Appendix A (Table 5).

### 3.4. Engagement class assignment

As anticipated, we considered 2 alternative methodologies (i.e. summation and subtraction) with 3 thresholds to determine whether a sentence could be classified as engaging or not. From our experimentation it resulted that the summation extraction method led to the best results; therefore, we applied this in our configuration. As for the engagement thresholds, however, a further test was performed: by comparing the three devised thresholds (-1, 0 and 1), it was found that threshold 0 (considered the most neutral) allowed for the best accuracy. Accuracy results relative to the comparison between engagement thresholds, for each classifier, can be found in Appendix A (Table 6).

### 3.5. Feature selection algorithm

The performance of a Machine Learning model can be improved by reducing the number of features it is trained with, based on their influence in the classification process [49]. For this reason, we implemented a recursive feature elimination algorithm to identify which features are most relevant for the prediction of engagement potential in a sentence, and consequently to improve the performance of the models. The process of the feature selection algorithm is structured in four steps:

1. Using the total set of features, the value of Accuracy in Cross-Validation is calculated;
2. The Accuracy value is compared with the best result obtained so far (0, if we are at the first iteration):
  - If the value obtained is greater, a ranking of features is made (based on the degrees of importance provided by the classification model), which will be considered the new optimal feature combination;
  - If the value obtained turns out to be lower, the previous optimal feature combination (obtained from the model that provided the higher Accuracy result) is retained;
3. Steps 1 and 2 are repeated, recursively eliminating a predefined number of features (recursively deleting a predefined number of features, starting with the least important based on the ranking), until it is reached the minimum threshold of about 10% of the total feature set;
4. The algorithm provides the selection of the most important features with which the best Accuracy result was obtained.

Given the long calculation times required for training Random-Forest (with 1000 estimators), it was decided to set the number of features to be eliminated at each iteration as follows:

- 10, for experiments performed with all features (*Multimodal*);
- 3, for experiments performed with linguistic features only (*Linguistic*);
- 7, for experiments performed with acoustic features only (*Acoustic*).

The choice was made taking into account the approximate proportion of linguistic features, and acoustic features, to the number of total features. In this way, it is possible to significantly reduce the number of iterations, and therefore speed up the experiment, while maintaining excellent Accuracy. With Linear-SVC, given its speed of execution, the number of features to be deleted at each iteration was set to 1 for all tests.

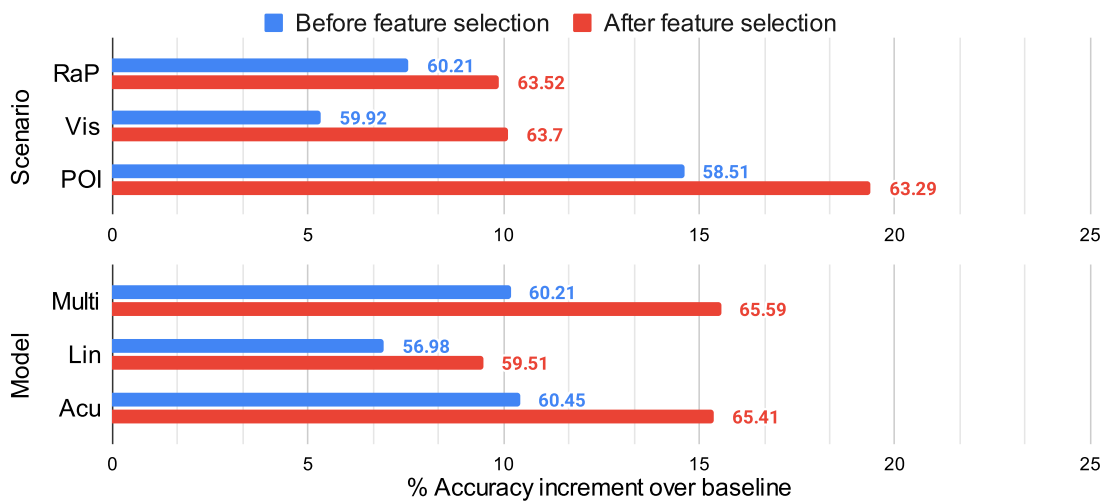
Another aspect that needs to be illustrated is the minimum number of features that the algorithm is forced to select: about 10% of the total set, for each type of feature:

- 45, for experiments performed with all features (*Multimodal*);
- 13, for experiments performed with linguistic features only (*Linguistic*);
- 32, for experiments performed with acoustic features only (*Acoustic*).

In this case, the proportion was maintained both for the experiments performed with Linear-SVC and for those performed with Random-Forest. Maintaining a minimum percentage of features of about 10% allows us to have a sufficient number of features in the analysis phase, to draw more precise and in-depth conclusions on the behaviour and choices of the models.

## 4. Experimental results

This section discusses the results for the most relevant experiments in the study, with a focus on the effects of feature selection. All the figures reporting the results contain absolute scores for accuracy values and bar extension for increments over baseline. Simultaneously, the percentages for each classification scenario represent the average of those obtained for the three models (*Multimodal*, *Linguistic* and *Acoustic*); conversely, each model represents the average of those obtained for the three scenarios (*RaP*, *Vis*, *POI*). The reported results are meant as an average between Linear-SVC and Random-Forest models. As stated in the introduction, the focus of this paper is not the state-of-the-art performance, but the analysis of the most salient features in both modalities. That said, we are not considering Linear-SVC and Random-Forest the same, but we are looking at an average performance based on a specific subset of features. Detailed tables on all the experiments can be found in the Appendix B, where we report single classifier results.



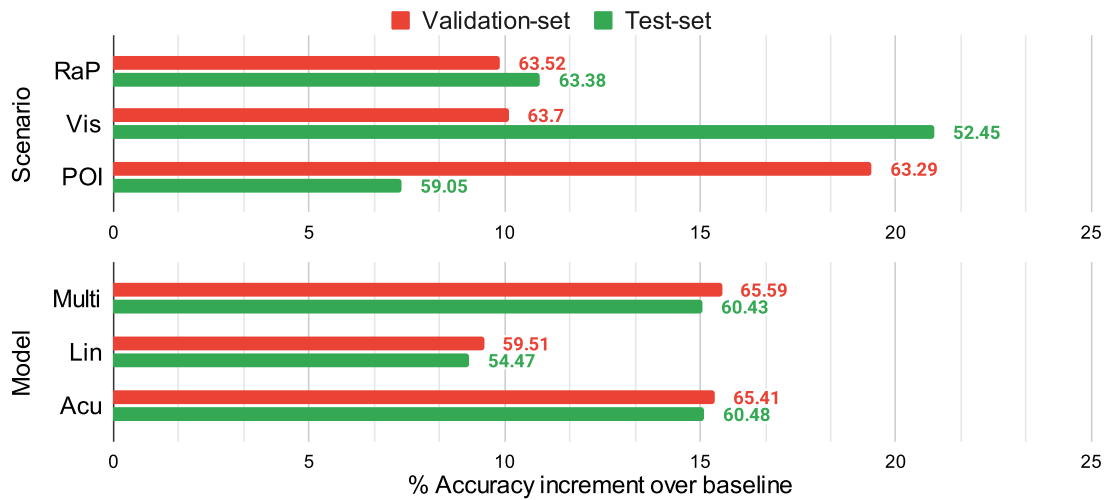
**Figure 1:** Comparison between average accuracy increments over baseline, before and after feature-selection (bar percentages refer to absolute Accuracy values, not to increments). All the values represent a mean between Linear-SVC and Random-Forest (*Multi* = Multimodal; *Lin* = Linguistic; *Aco* = Acoustic).

### 4.1. Feature selection effectiveness

The feature selection algorithm was run on both classifiers in use, testing all classification scenarios and all models. As illustrated in Figure 1, feature selection leads to increases in mean accuracy percentages in all the cases, by reducing the feature space in a range between 10 and 32% (average range between Linear-SVC and Random-Forest) of the original features set.

The increases in accuracy percentages are a signal of the presence of features that are particularly relevant to predict an engagement potential and, conversely, that many of the original features are redundant and noisy for both scenario and model variation.

## 4.2. Results on test-set



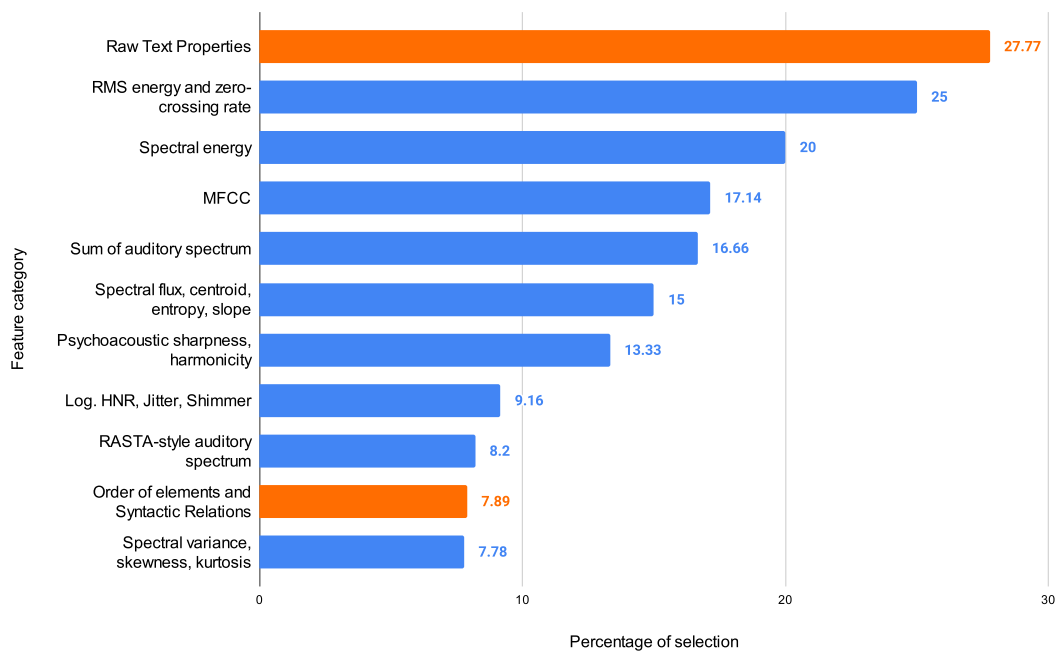
**Figure 2:** Comparison between average accuracy increments over baseline on validation-set and test-set (bar percentages refer to absolute Accuracy values, not to increments). All the values represent a mean between Linear-SVC and Random-Forest (*Multi* = Multimodal; *Lin* = Linguistic; *Aco* = Acoustic).

We run the final step of experimentation on an unknown portion of the dataset (i.e. test-set) for each classification scenario: as illustrated in Figure 2, the classifiers achieved Accuracy increments over baseline on test-set that are extremely similar to those achieved on the validation-set, despite the fact that these were unknown data. A single exception is detectable in the case of POI classification scenario: the gaps between the percentages can be attributed to the large differences between the baseline of the validation-set and the test-set (Detailed accuracy results on test-set can be found in Appendix C, for each classification model).

The good results achieved on the test-set indicate that classifiers trained with a restricted set of features are able to effectively detect an engagement potential in unseen data. This provides us with a definitive confirmation of how the most important features, selected by the feature selection algorithm, can indeed constitute a set of fundamental aspects to detect the engagement of a sentence.

## 5. Feature analysis

In order to understand which linguistic and acoustic features are the most relevant to detect an engagement potential, it is necessary to analyse the subset of features with which the classifiers performed best. Specifically, it is possible to define what percentage of each feature category (e.g. linguistic:morpho-syntactic, acoustic:spectral) was included by the feature selection algorithm among the most relevant. It is important to specify that only those categories of features selected by both classifiers were considered, i.e. only those that resulted to be highly relevant to the classification process, regardless of the exploited classifier.



**Figure 3:** Average percentage (between the two classifiers) in which each feature category was included in the top 10% of the features, based on the ranking developed by the classifier models (only the feature categories selected by both the classifiers are considered).

Considering the most important 10% of all features (on the basis of the algorithm ranking), we can observe that acoustic features seem to be the most important for the classification of engaging and non-engaging sentences: the average percentage of acoustic features (9.34%), included in the total set of most important features selected, is about 1.57 times higher than the average percentage of linguistic features (5.94%). We can derive that acoustic features play a significantly more important role, compared to linguistic features.

A closer look at the selected feature categories shows us what percentage of them were included among the most important ones by the classifiers. As shown in Figure 3, *Raw Text Properties* (i.e. sentence and word length) are the most relevant group of features. Other linguistic features included in the selected features regard syntactic relations and the order of elements, but those are selected only for the 7.89% of the total. Nevertheless, the rest of the selected features are all coming from the acoustic modality, specifically related to the sound spectrum. In this regard, it is possible to observe that the timbre of the speech, its amplitude and richness in the frequency range are decisive factors in the maintenance of attention. However, it is also necessary to note that the first group of acoustic features (*RMS energy and zero-crossing rate*), turns out to be a prosodic feature; the rhythmic features of the voice, therefore, which highlight traits such as irony and sarcasm, still play a strong role. Voice quality aspects, on the other hand, do not seem to be particularly implicated in the classification process. Detailed percentages of included features per modality can be found in Appendix D.

## 6. Conclusions

The implemented Machine Learning models were able to detect an engagement potential in language in multiple scenarios and on unknown data. It emerged that certain phenomena and features of language, mainly acoustic in nature (like prosodic or spectral), play a key role in the classification process, and thus in assessing an engagement potential of an uttered sentence.

Ultimately, it is possible to observe that all the results were achieved by fully exploiting the potential of a restricted set of features (between 10 and 32% of the total sets). This study, therefore, also aims to show to what extent optimised Machine Learning models, combined with a selected and optimised data representation (i.e. relevant features), can succeed in achieving better accuracy results. The stringent feature selection, moreover, proved to be crucial in understanding which aspects, among the various linguistic and acoustic ones considered, play a critical role in making a sentence engaging or not. On the acoustic level, prosodic and spectrum related features play a major role in discriminating engaging and non-engaging sentences, while on the linguistic level raw text properties give the main contribution. We can conclude that the attention of the listener(s), and thus the perceivable engagement, can be driven by acoustic and linguistic features, and for this reason we studied the phenomenon of engagement by means of fully explainable classification models.

### 6.1. Future developments

One of the critical issues of this study undoubtedly concerns the size of the dataset, that can be considered relatively small (1,114 sentences) and not very varied (the 3 visits are lead by the same guide, thus all the data regard one person). To conduct an even more precise and accurate study, and to generalise the results, it would be necessary to increase the size of the dataset by including data coming from more guides and groups of visitors.

Another important enrichment of the dataset could involve visual data. Currently, we exploited the visual part of the dataset exclusively for the annotation of the variation of attention/engagement, but it would be interesting to explore visual features in the classification process, and to measure the performance of a model that considers linguistic, acoustic, and visual data to predict the engagement potential of a communication act.

## References

- [1] D. Nozza, L. Passaro, M. Polignano, Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI\*IA 2022), November 30, 2022, CEUR-WS.org, 2022.
- [2] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, *Minds and Machines* 30 (2020) 681–694.
- [3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer., *J. Mach. Learn. Res.* 21 (2020) 1–67.

- [4] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, U. Trautwein, Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction, *Educational Psychology Review* 35 (2019) 463–23.
- [5] C. Regenbogen, D. A. Schneider, R. E. Gur, F. Schneider, U. Habel, T. Kellermann, Multi-modal human communication – Targeting facial expressions, speech content and prosody, *NeuroImage* 60 (2012) 2346–2356.
- [6] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, C. Peters, Engagement in human-agent interaction: An overview, *Frontiers in Robotics and AI* 7 (2020) 92.
- [7] I. Poggi, Mind, hands, face and body: a goal and belief view of multimodal communication, Weidler, 2007.
- [8] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, C. Rich, Explorations in engagement for humans and robots, *Artificial Intelligence* 166 (2005) 140–164.
- [9] G. Castellano, A. Pereira, I. Leite, A. Paiva, P. W. McOwan, Detecting user engagement with a robot companion using task and social interaction-based features, in: *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 119–126.
- [10] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, in: *Proceedings of the 6th International Conference on Human-Robot Interaction, HRI '11*, Association for Computing Machinery, New York, NY, USA, 2011, p. 305–312. URL: <https://doi.org/10.1145/1957656.1957781>. doi:10.1145/1957656.1957781.
- [11] A. Ben-Youssef, C. Clavel, S. Essid, Early detection of user engagement breakdown in spontaneous human-humanoid interaction, *IEEE Transactions on Affective Computing* 12 (2021) 776–787. doi:10.1109/TAFFC.2019.2898399.
- [12] D. Gatica-Perez, L. McCowan, D. Zhang, S. Bengio, Detecting group interest-level in meetings, in: *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, IEEE, 2005, pp. I–489.
- [13] C. Oertel, S. Scherer, N. Campbell, On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation., in: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 1541–1544.
- [14] J. A. Fredricks, P. C. Blumenfeld, A. H. Paris, School engagement: Potential of the concept, state of the evidence, *Review of educational research* 74 (2004) 59–109.
- [15] F. Cutugno, F. Dell'Orletta, I. Poggi, R. Savy, A. Sorgente, The chrome manifesto: integrating multimodal data into cultural heritage resources, *Computational Linguistics CLiC-it 2018* (2018) 155.
- [16] A. Origlia, R. Savy, I. Poggi, F. Cutugno, I. Alfano, F. D'Errico, L. Vincze, V. Cataldo, An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the chrome project, in: *2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*, volume 2091, 2018, pp. 1–4.
- [17] J. A. Fredricks, P. C. Blumenfeld, A. H. Paris, School engagement: Potential of the concept, state of the evidence, *Review of educational research* 74 (2004) 59–109. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.



- [18] P. Goldberg, O. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, U. Trautwein, Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction, *Educational Psychology Review* 35 (2019) 463–23. URL: <http://link.springer.com/10.1007/s10648-019-09514-z>. doi:10.1007/s10648-019-09514-z, publisher: Springer US.
- [19] D. Melhart, A. Liapis, G. N. Yannakakis, Pagan: Video affect annotation made easy, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 130–136.
- [20] A. A. Ravelli, A. Origlia, F. Dell'Orletta, Exploring attention in a multimodal corpus of guided tours, in: *Computational Linguistics CLiC-it 2020*, 2020, p. 353.
- [21] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, Elan: a professional framework for multimodality research, in: *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 1556–1559.
- [22] S. Izre'el, H. Mello, A. Panunzi, T. Raso, In Search of Basic Units of Spoken Language, volume 94 of *A corpus-driven approach*, John Benjamins Publishing Company, Amsterdam, 2020. doi:10.1075/sc1.94, iSSN: 1388-0373.
- [23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, et al., Speechbrain: A general-purpose speech toolkit, arXiv preprint arXiv:2106.04624 (2021).
- [25] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in Neural Information Processing Systems* 33 (2020) 12449–12460.
- [26] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-ud: a tool for linguistic profiling of texts, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 7145–7151.
- [27] S. Montemagni, Tecnologie linguistico-computazionali e monitoraggio della lingua italiana, *Studi Italiani di Linguistica Teorica e Applicata (SILTA) XLII* (2013) 145–172.
- [28] M. Straka, J. Hajič, J. Straková, UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4290–4297. URL: <https://aclanthology.org/L16-1680>.
- [29] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [30] J. Gareth, W. Daniela, H. Trevor, T. Robert, *An introduction to statistical learning: with applications in R*, Springer, 2013.
- [31] G. C. Cawley, N. L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, *The Journal of Machine Learning Research* 11 (2010) 2079–2107.
- [32] G. Seni, J. F. Elder, Ensemble methods in data mining: improving accuracy through combining predictions, *Synthesis lectures on data mining and knowledge discovery* 2

(2010) 1–126.

- [33] D. M. Allen, The relationship between variable selection and data augmentation and a method for prediction, *technometrics* 16 (1974) 125–127.
- [34] M. Stone, Cross-validatory choice and assessment of statistical predictions, *Journal of the royal statistical society: Series B (Methodological)* 36 (1974) 111–133.
- [35] M. Stone, An asymptotic equivalence of choice of model by cross-validation and akaike's criterion, *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1977) 44–47.
- [36] N. Cristianini, J. Shawe-Taylor, et al., *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.
- [37] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.
- [38] T. K. Ho, Random decision forests, in: *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, IEEE, 1995, pp. 278–282.
- [39] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
- [40] A. L. Blum, P. Langley, Selection of relevant features and examples in machine learning, *Artificial intelligence* 97 (1997) 245–271.
- [41] P. S. Bradley, O. L. Mangasarian, Feature selection via concave minimization and support vector machines., in: *ICML*, volume 98, Citeseer, 1998, pp. 82–90.
- [42] P. S. Bradley, O. L. Mangasarian, W. N. Street, Feature selection via mathematical programming, *INFORMS Journal on Computing* 10 (1998) 209–217.
- [43] F. Hutter, L. Kotthoff, J. Vanschoren, *Automated machine learning: methods, systems, challenges*, Springer Nature, 2019.
- [44] M. Claesen, B. De Moor, Hyperparameter search in machine learning, *arXiv preprint arXiv:1502.02127* (2015).
- [45] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, *Advances in neural information processing systems* 24 (2011).
- [46] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization., *Journal of machine learning research* 13 (2012).
- [47] F. Cucker, S. Smale, et al., Best choices for regularization parameters in learning theory: on the bias-variance problem, *Foundations of computational Mathematics* 2 (2002) 413–428.
- [48] J. Han, J. Pei, H. Tong, *Data mining: concepts and techniques*, Morgan kaufmann, 2022.
- [49] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *Journal of machine learning research* 3 (2003) 1157–1182.

## A. Setting of classifiers and data preprocessing

This appendix shows the detailed tables containing the Accuracy results obtained during the first phase of experimentation, where the classifier models and dataset were configured and optimised. Specifically, the percentages obtained from the cross-comparison of various hyperparameter configurations (Section 3.2.1), for both classifier models, and of the two engagement aggregation techniques implemented in the study (Section 2.3) are illustrated. The results of the comparison between the various data normalisation techniques tested (Section 3.3), and between the engagement thresholds used (Section 2.3.3) are also shown. Each table shows the baseline percentages (based on the most frequent class of engagement, i.e. 0 for non-engaging or 1 for engaging) for each classification scenario, for each engagement threshold, and for each engagement aggregation technique (experiments shown in these Tables were all performed in the *RaP* classification scenario).

**Table 3**

Accuracies with different regularisation parameter values used by the Linear Support Vector Classifier (*RaP* scenario).

	<i>Subtraction</i>	<i>Summation</i>	MEAN
<i>0.001</i>	58.59 (+5.84)	<b>59.27</b> <b>(+6.63)</b>	58.93 (+6.23)
<i>0.01</i>	58.23 (+5.48)	<b>58.79</b> <b>(+6.15)</b>	58.51 (+5.82)
<i>0.10</i>	57.15 (+4.40)	<b>57.38</b> <b>(+4.74)</b>	57.27 (+4.57)
<i>1.00</i>	55.35 (+2.60)	<b>56.68</b> <b>(+4.04)</b>	56.01 (+3.32)
MEAN	57.23 (+4.58)	<b>58.03</b> <b>(+5.39)</b>	57.63 (+4.94)
BASELINE	52.75	52.64	52.69

**Table 4**

Accuracies with different Decision-Tree quantities used by the Random Forest Classifier (*RaP* scenario).

	<i>Subtraction</i>	<i>Summation</i>	MEAN
<i>10</i>	<b>60.36</b> <b>(+7.61)</b>	58.66 (+6.03)	59.51 (+6.82)
<i>100</i>	60.14 (+7.39)	<b>61.63</b> <b>(+8.99)</b>	60.89 (+8.19)
<i>1000</i>	61.93 (+9.18)	<b>63.14</b> <b>(+10.50)</b>	62.53 (+9.84)
MEAN	60.81 (+8.06)	<b>61.14</b> <b>(+8.50)</b>	60.97 (+8.28)
BASELINE	52.75	52.64	52.69

**Table 5**

Comparison of normalization techniques: accuracies of classifiers on *RaP* scenario depending on data normalization (using only aggregation by summation technique, and engagement threshold 0).

	<i>StaS</i>	<i>MAS</i>	<i>MiMaS</i> (-1,1)	<i>MiMaS</i> (0,1)	<i>RoS</i>	<i>PoT</i>	<i>QuT</i>
<i>Linear-SVC</i>	59.93	59.26	59.03	59.03	58.13	58.92	<b>60.60</b>
<i>Random-Forest</i>	63.19	<b>63.41</b>	<b>63.41</b>	<b>63.41</b>	63.30	61.95	63.30
MEAN	61.55	61.33	61.22	61.22	60.71	60.44	<b>61.95</b>
BASELINE	52.64						

**Table 6**

Comparison of engagement thresholds on *RaP* scenario (using only aggregation by summation technique).

	-1	0	1
<i>Linear-SVC</i>	57.35 (-23.12)	<b>60.60</b> <b>(+7.96)</b>	58.47 (-4.04)
<i>Random-Forest</i>	80.47 (+0.00)	<b>63.30</b> <b>(+10.66)</b>	64.64 (+2.13)
MEAN	68.91 (-11.56)	<b>61.95</b> <b>(+9.31)</b>	61.55 (-0.96)
BASELINE	80.47	52.64	62.51

## B. Accuracies in Cross-Validation

This appendix details the averages of all Accuracy results obtained during Cross-Validation, with a cross-comparison between the feature combinations used and the classification scenarios. The tables are divided by individual classifier model, and show the accuracies obtained both before and after feature selection (see section 3.5). Each table also shows the baseline percentages (based on the most frequent class) for each classification scenario and for each modality, used for comparison with the results obtained.

**Table 7**

Accuracy values obtained by comparison between models and classification scenarios, before feature selection, with *Linear-SVC*.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	60.60 (+7.96)	59.55 (+5.96)	<b>58.02</b> <b>(+14.12)</b>	59.39 (+9.35)
<i>Linguistic</i>	57.46 (+4.82)	58.08 (+4.49)	<b>58.44</b> <b>(+14.54)</b>	57.99 (+7.95)
<i>Acoustic</i>	60.27 (+7.63)	59.59 (+6.00)	<b>59.58</b> <b>(+15.68)</b>	59.81 (+9.77)
MEAN	59.44 (+6.80)	59.07 (+5.48)	<b>58.68</b> <b>(+14.78)</b>	59.06 (+9.02)
BASELINE	52.64	53.59	43.90	50.04

**Table 8**

Accuracy values obtained by comparison between models and classification scenarios, before feature selection, with Random-Forest.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	63.19 (+10.55)	59.45 (+5.86)	<b>60.46</b> <b>(+16.56)</b>	61.03 (+10.99)
<i>Linguistic</i>	56.22 (+3.58)	56.96 (+3.37)	<b>54.76</b> <b>(+10.86)</b>	55.98 (+5.94)
<i>Acoustic</i>	65.53 (+12.89)	59.89 (+6.30)	<b>59.86</b> <b>(+15.96)</b>	61.09 (+11.05)
MEAN	60.98 (+8.34)	58.77 (+5.18)	<b>58.36</b> <b>(+14.46)</b>	59.37 (+9.33)
BASELINE	52.64	53.59	43.90	50.04

**Table 9**

Accuracy values obtained by comparison between models and classification scenarios, after feature selection, with Linear-SVC.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	64.42 (+11.78)	64.80 (+11.21)	<b>65.21</b> <b>(+21.31)</b>	64.81 (+14.77)
<i>Linguistic</i>	59.92 (+7.28)	63.21 (+9.62)	<b>61.04</b> <b>(+17.14)</b>	61.39 (+11.35)
<i>Acoustic</i>	64.20 (+11.56)	64.66 (+11.07)	<b>65.42</b> <b>(+21.52)</b>	64.76 (+14.72)
MEAN	62.85 (+10.21)	64.22 (+10.63)	<b>63.89</b> <b>(+19.99)</b>	63.65 (+13.61)
BASELINE	52.64	53.59	43.90	50.04

**Table 10**

Accuracy values obtained by comparison between models and classification scenarios, after feature selection, with Random-Forest.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	67.46 (+14.82)	65.51 (+11.92)	<b>66.17</b> <b>(+22.24)</b>	66.38 (+16.34)
<i>Linguistic</i>	57.46 (+4.82)	59.45 (+5.86)	<b>55.98</b> <b>(+12.08)</b>	57.63 (+7.59)
<i>Acoustic</i>	67.68 (+15.04)	64.59 (+11.00)	<b>65.91</b> <b>(+22.01)</b>	66.06 (+16.02)
MEAN	64.20 (+9.56)	63.18 (+9.59)	<b>62.69</b> <b>(+18.79)</b>	63.36 (+13.32)
BASELINE	52.64	53.59	43.90	50.04

## C. Accuracies on test-set

This appendix shows the Accuracy percentages obtained on the test-set, in the final phase of testing the classifier models. In particular, there is a table of accuracies for each of the two classification models, with a cross-comparison between the feature combinations used and the classification scenarios. Each table also shows the baseline percentages (based on the most frequent class) for each classification scenario and for each modality.

**Table 11**

Accuracy values on test-set, obtained by comparison between models and classification scenarios, with Linear-SVC.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	64.57 (+12.10)	<b>54.75</b> <b>(+22.79)</b>	57.68 (+6.00)	59.00 (+13.63)
<i>Linguistic</i>	59.64 (+7.17)	<b>49.68</b> <b>(+17.72)</b>	57.68 (+6.00)	55.67 (+10.30)
<i>Acoustic</i>	65.92 (+13.45)	<b>53.48</b> <b>(+21.52)</b>	58.05 (+6.37)	59.15 (+13.78)
MEAN	63.38 (+10.91)	<b>52.64</b> <b>(+20.68)</b>	57.80 (+6.12)	57.94 (+12.57)
BASELINE	52.47	31.96	51.68	45.37

**Table 12**

Accuracy values on test-set, obtained by comparison between models and classification scenarios, with Random-Forest.

	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Multimodal</i>	65.47 (+13.00)	<b>55.70</b> <b>(+23.73)</b>	64.42 (+12.73)	61.86 (+16.49)
<i>Linguistic</i>	57.85 (+5.38)	<b>48.42</b> <b>(+16.46)</b>	53.56 (+1.87)	53.28 (+7.91)
<i>Acoustic</i>	66.82 (+14.35)	<b>55.70</b> <b>(+23.27)</b>	62.92 (+11.24)	61.81 (+16.44)
MEAN	63.38 (+10.91)	<b>52.27</b> <b>(+21.31)</b>	60.30 (+8.62)	58.98 (+13.61)
BASELINE	52.47	31.96	51.68	45.37

## D. Feature selection results

This appendix shows the results of the analysis of the features selected by the classifier models through the feature selection algorithm. The tables show the percentage by which each feature category and subcategory was included in the top 10% of the features, based on the ranking processed by the classification models. It's important to notice that the indicated percentages are an average value between the percentages found with Linear-SVC and Random-Forest.

**Table 13**

Percentage of inclusion of linguistic features categories among the top 10% of the features, based on the ranking processed by the classification models (average between Linear-SVC and Random-Forest).

Linguistic feature category	<i>RaP</i>	<i>Vis</i>	<i>POI</i>	MEAN
<i>Raw Text Properties</i>	16.66	33.33	33.33	27.77
<i>Morphosyntactic information</i>	2.25	3.82	3.82	3.30
<i>Verbal Predicate Structure</i>	13.63	13.63	0.00	9.09
<i>Parsed Tree Structures</i>	12.50	12.50	25	16.67
<i>Syntactic Relations</i>	6.58	6.58	10.52	7.89
<i>Use of Subordination</i>	15.00	15.00	5.00	11.67

**Table 14**

Percentage of inclusion of acoustic features categories among the top 10% of the features, based on the ranking processed by the classification models (average between Linear-SVC and Random-Forest).

ACOUSTIC FEATURE CATEGORIES	<i>PaC</i>	<i>Vis</i>	<i>POI</i>	MEAN
PROSODICS	13.75	10.00	15.00	12.91
<i>F0 (SHS and Viterbi smoothing)</i>	10.00	10.00	0.00	6.66
<i>Sum of auditory spectrum (loudness)</i>	20.00	10.00	20.00	16.66
<i>Sum of RASTA-style filtered auditory spectrum</i>	0.00	0.00	10.00	3.33
<i>RMS energy and zero-crossing rate</i>	25.00	20.00	30.00	25.00
SPECTRAL	13.23	10.44	11.59	11.77
<i>RASTA-style auditory spectrum, bands 1-26</i>	6.92	9.61	8.07	8.20
<i>MFCC 1-14</i>	20.71	14.28	16.43	17.14
<i>Spectral energy 250-650 Hz, 1 k-4 kHz</i>	20.00	20.00	20.00	20.00
<i>Spectral roll off point 0.25, 0.50, 0.75, 0.90</i>	0,00	2.50	0,00	0.83
<i>Spectral flux, centroid, entropy, slope</i>	15.00	15.00	15.00	15.00
<i>Psychoacoustic sharpness, harmonicity</i>	20.00	5.00	15.00	13.33
<i>Spectral variance, skewness, kurtosis</i>	10.00	6.66	6.67	7.78
VOICE QUALITY	7.75	7,50	8.75	7.91
<i>Voicing probability</i>	10.00	0,00	10.00	6.67
<i>Log. HNR, Jitter (local, delta), Shimmer (local)</i>	5,00	15,00	7.50	9.16