ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

UNIQUE

Semester Thesis

Sentiment Analysis in Video Calls

Autumn Term 2022

Supervised by:
Prof. Dr. Perez-Cruz Fernando
Dr. Hauri Pascal

Author:
Alexis Tabin

Declaration of Originality

I hereby declare that the written work I have submitted entitled

Sentiment Analysis in Video Calls

is original work which I alone have authored and which is written in my own words.

Author(s)

Alexis	Tabin

Student supervisor(s)

Fernando	Perez Cruz
Pascal	Hauri

With the signature, I declare that I have been informed regarding standard academic citation rules and that I have read and understood the information on 'Citation etiquette' (`https://www.ethz.ch/content/dam/ethz/main/education/rechtliches-abschluesse/leistungskontrollen/plagiarism-citationetiquette.pdf`). The citation conventions usual to the discipline in question here have been respected.

The above-written work may be tested electronically for plagiarism.

Place and date	Signature

Contents

Abstract

This report presents the results of a semester project at Unique. Unique is a fast-growing startup that aims to implement *Artificial Intelligence (AI)* in online meeting platforms like Microsoft Teams or Zoom. The project focused on developing a *Sentiment Analysis (SA)* model using recent *Natural Language Processing (NLP)* techniques. SA is used to understand a person's underlying feelings and emotions regarding certain topics they discussed.

Throughout this report, we will be providing an overview of Unique's product and the different algorithms they use. The report discusses the transcription, important moments, participation, summary, insights, and checklist features. After that, this report's main focus is SA and further developing Unique's SA model. The different strategies used to design efficient SA models will be explained, and we will see in this report how they can be useful for a company like Unique.

This report will provide valuable insights into using AI in online meeting platforms and how it can enhance sales performance. It will be helpful to anyone interested in understanding the potential of AI, more precisely NLP, applied to video calls and sales.

Chapter 1

Introduction

During the pandemic of the COVID-19, when everyone was confined at home, the number of online meetings grew significantly. Many employers were forced to adapt and demanded that workers be in a full or hybrid remote setting.

The rise of remote work and online meetings has increased the use of platforms like Google Meet or Discord, making it crucial for companies to optimize and improve these interactions. Unique's idea is to take advantage of this trend and bring AI into online video calls. With their product, Unique aims to improve sales performance and provide valuable insights during and after sales calls.

This report provides an overview of Unique's product and a comprehensive summary of the project, focusing on the *Sentiment Analysis (SA)* [1], [2], [3] aspect. SA is crucial for Unique as it enables insights into public opinion, helping to sell projects or ideas [4]. The project aimed to enhance the current rudimentary SA model by improving accuracy and efficiency with transformer models [5], [6]. Results, methodology and implications of the findings are included.

Chapter 2

Unique

By providing organisational tools and AI in online meeting platforms, Unique's main objective is to increase the productivity and performance of the sales team. Their product reports important call metrics, such as the time spent talking, the number of interruptions, the sentiment of the conversation, the transcription of the call, and a summary. This information can be used to improve call preparation and to gain insights into the effectiveness of the call.

Unique's product is not limited to sales calls. It can be used in other types of calls, such as interviews or online courses, to gain insights into participant interactions. The company aims to make online communication more efficient and effective by providing valuable insights through AI.

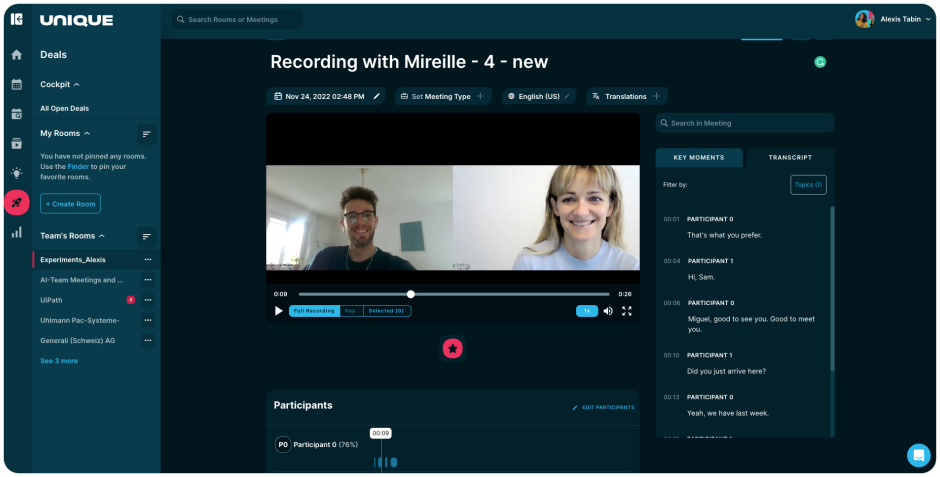

Figure 2.1: View of one recording in the Unique app. On the left panel, users can navigate through the different deals (meetings) and corresponding recordings. In the center is displayed a recording with some pieces of information such as title, date, etc. On the right, there is the transcription panel and the key moments.

Unique's product is built on a suite of AI algorithms, including speech recognition, natural language processing, and machine learning. These algorithms work together to analyse the audio and transcriptions of a call, providing a detailed analysis of the conversation. The key features include identifying key topics discussed, detecting sentiment, and identifying important moments in the call.

In figure 2.1, an overview of Unique's platform is presented. The following sections describe in more detail the different tools used by the startup.

2.1 Transcription

Transcription is the process of converting audio speech into written text, and it is a key element of Unique's product (see fig. 2.2). Transcription technology has come a long way since the early days of manual transcription, and today's speech recognition techniques can transcribe speech with high accuracy and in real-time.

Unique uses speech recognition technology to transcribe the audio of a call into text. The transcription happens nearly instantaneously and is available during the call. The transcription accuracy is important as most of the NLP models used at Unique are based on it. Searching in the transcription allows the user to recover a specific piece of information discussed during the call faster than going through the entire video.

With the advent of *Deep Learning (DL)*, the performance of speech recognition systems has improved dramatically. Models such as Deep Speech 2 [7] and Listen, Attend and Spell [8] have achieved great performance on various speech recognition benchmarks and have been shown to be highly accurate. These models can transcribe speech in real time and handle a large vocabulary.

Transcription technology is useful in various applications, including speech-to-text dictation, call centre automation, and online meetings. In the case of Unique, transcription technology is used to analyze the audio of a call and to provide a written record of the conversation. This allows for detailed conversation analysis, including identifying key topics discussed, detecting sentiment, and identifying important moments in the call.

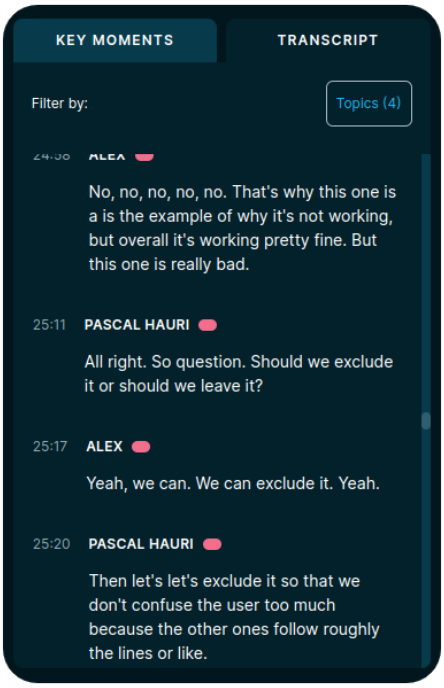

Figure 2.2: Details of the transcript panel

Unique uses a forked version of Whisper [9], a model developed by OpenAI. Whisper is an automatic speech recognition (ASR) system that uses a vast amount of data (680,000 hours of audio files data) collected from the web. This data is multilingual and multitasking, which allows the system to be robust in the face of different accents, background noise, and technical language. The developers of Whisper are open-sourcing the models and inference code to enable researchers and developers to build practical applications and to advance the field of speech processing further.

The transcription process is essential for this project, as the NLP done for the sentiment analysis task will be based on it.

2.2 Important Moments

Currently, users can mark important moments during calls by clicking on a small red star on the top left of the transcription (see fig. 2.3). After the call, the insights panel displays the important moments that happened, and users can edit the important moments (see fig. 2.4). With a glance, the user can rewatch the most important parts of the recording.

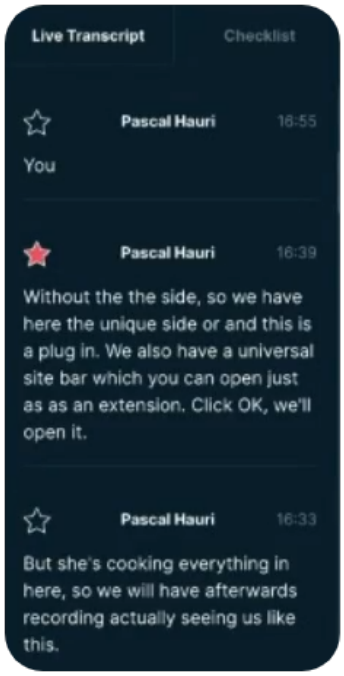

Figure 2.3: Red stars indicate important moments. Users can click on the star to add important moments during the call

Detecting and predicting important moments in a conversation will be another key feature of Unique's product, as it allows for a deeper understanding of the conversation and the ability to identify key points that led to a sale or other important outcome. However, the implementation of automatic detection of important moments is not yet fully developed. During this last semester, a Bachelor thesis was carried out to develop this feature.

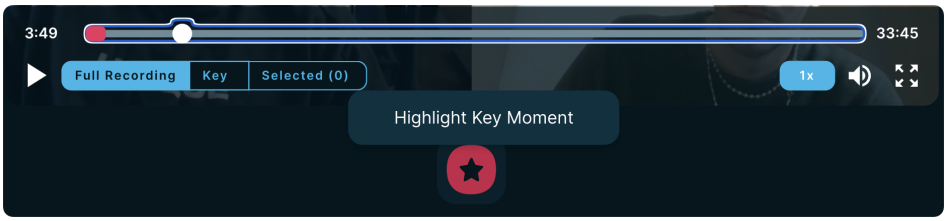

Figure 2.4: Users can add or delete important moments after the call

2.3 Participation

With its algorithm, Unique can also track each participant's participation in the call (see fig. 2.5), measuring the amount of talking time and the number of interruptions. This can identify potential issues in the call, such as one person dominating the conversation or a lack of engagement from the buyer or other participants.

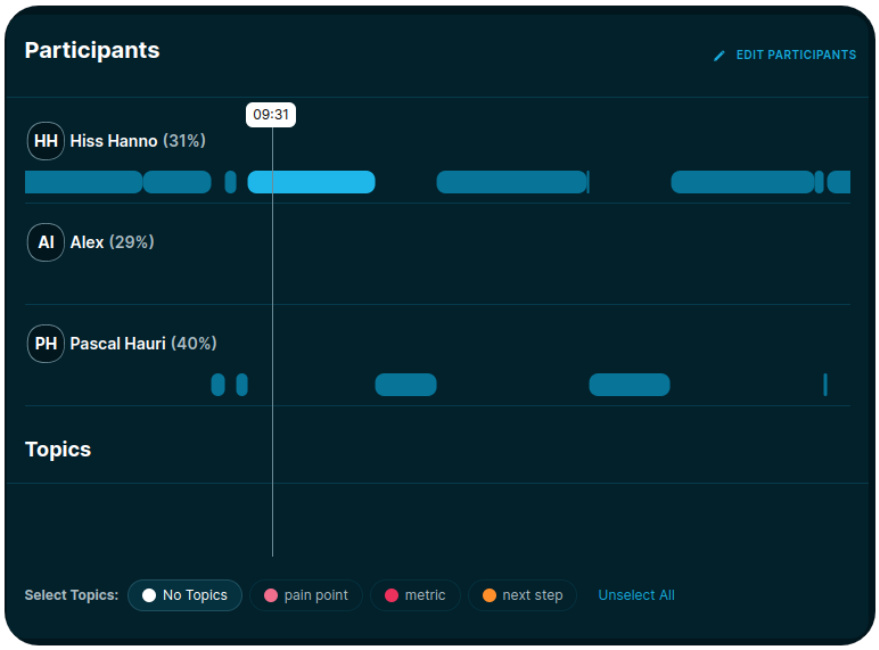

Figure 2.5: Details of the participation panel

The participation analysis process works by transcribing the audio of the call into text with the transcription model, then training a classifier that will learn the voice features of each participant and classify each chunk of text to the corresponding participant.

For this purpose, *pyannote.audio* [10] is a pretrained speaker segmentation model. Its usage simplicity makes it one of the most popular audio models available on Huggingface.

For conversational speech recognition, various convolutional and LSTM acoustic model architectures can be used to achieve performance on par with human performance [11].

More recently, *A Framework for Self-Supervised Learning of Speech Representations* [12] demonstrate the feasibility of speech recognition with limited amounts of labelled data. The authors used a technique called *Contrastive Learning*, a concept in which the input is mapped in two ways (e.g. audio and text). Afterwards, the model is trained to identify whether two input transformations are still the same object. The features of the audio modality can then be used to perform a classification task between the different speakers of a conversation.

2.4 Summary

Unique's product summarises the call, highlighting the main subjects discussed. This allows instantaneous understanding of the key takeaways from the call, which can be useful for follow-up and post-call analysis.

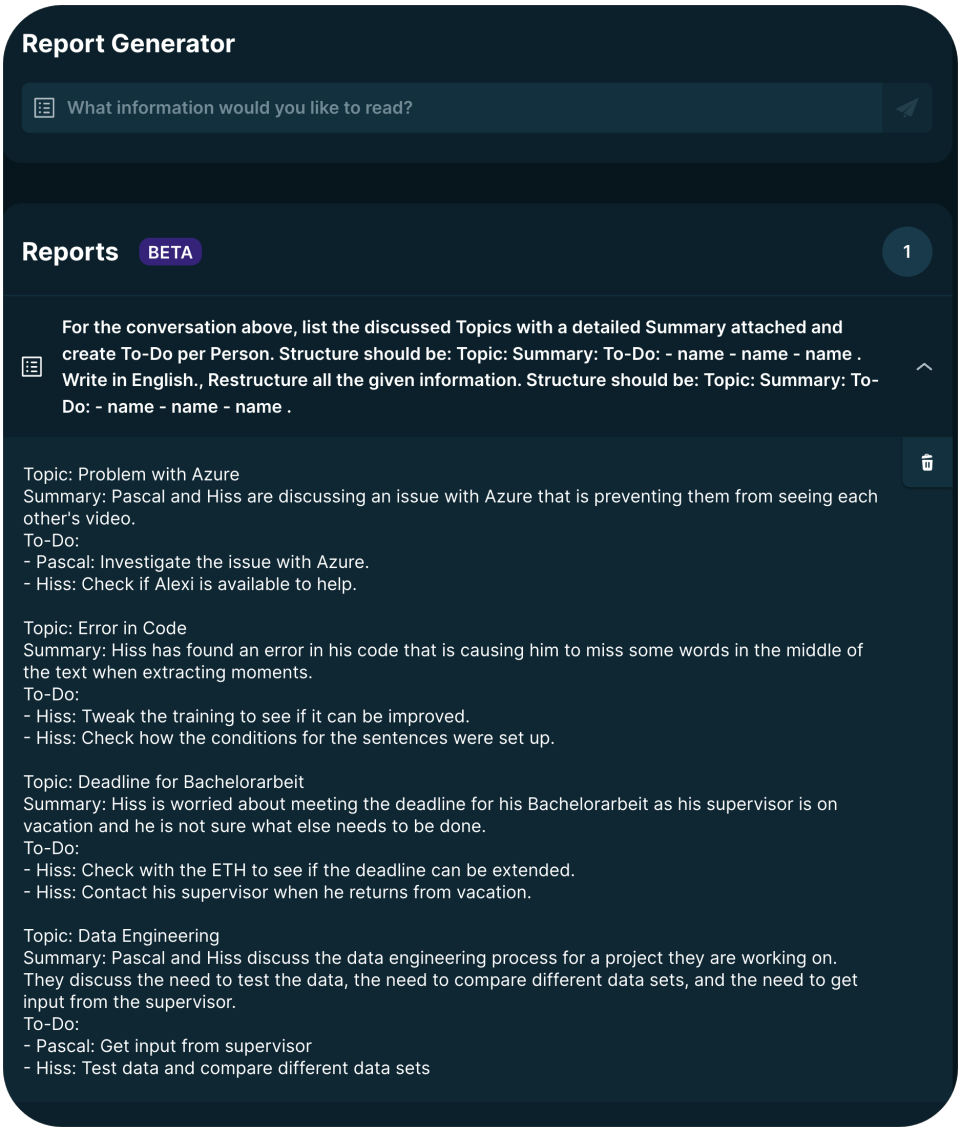

Figure 2.6: Details of the summary panel

The algorithm used for the summary is the famous GPT-3 [13]. Using *Large Language Models (LLM)* allows summarising complete conversations, such as the example illustrated in fig. 2.6. When using GPT-3, one of the keys to a great output is the prompt that is fed. For example, the prompt allows the user to choose a specific output format.

In *Large Language Models are Zero-Shot Reasoners* [14], the authors demonstrated the importance of the prompt design. They found out that adding the sentence **Let's think step by step** in the prompt can facilitate step-by-step answers across various reasoning tasks, for instance, arithmetic [15], [16].

More recently, *Chain of Thought Prompting Elicits Reasoning in Large Language Models* [17] took prompt designing one step further by explaining explicitly to LLMs how they should reason about a problem.

2.5 Checklist

Unique offers a checklist of items to be covered before, during and after the call to assist the user to prepare and to increase the chances of success for the call (see fig. 2.7). This checklist can be customized for calls, like sales, interviews or online courses.

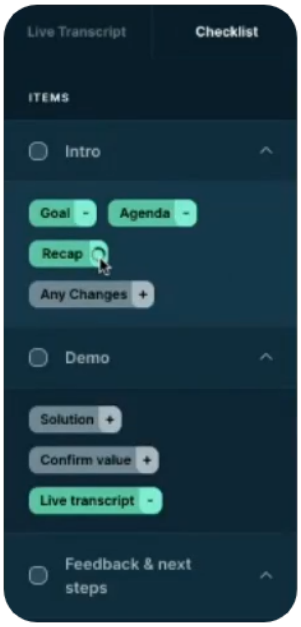

Figure 2.7: Details of the insights panel

Unique automatically detects when a subject has been talked about and then checks it in the list. It is a naive word-matching algorithm that detects when the user mentions a specific topic and updates the checklist accordingly.

2.6 Insights

The insights panel is available after the call. It provides a detailed understanding of the call, including sentiment analysis, participation analysis, and time spent on each topic (see fig. 2.8). These insights can improve call preparation, identify areas for improvement, and gain a deeper understanding of the conversation.

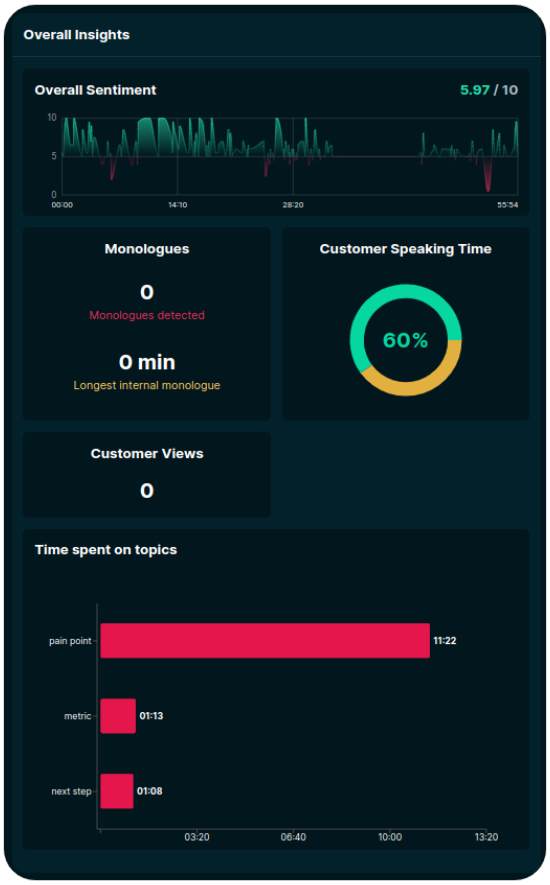

Figure 2.8: Details of the insights panel

On top of the panel, the user can see the sentiment analysis. This part will be covered thoroughly during chapter 3. Below the sentiment analysis subpanel, a counter indicates the number of monologues during the call. In this panel, the participant's participation appears too. Finally, at the bottom of the panel, the user can see how much time was spent on which topic.

Chapter 3

Sentiment Analysis in Video Calls

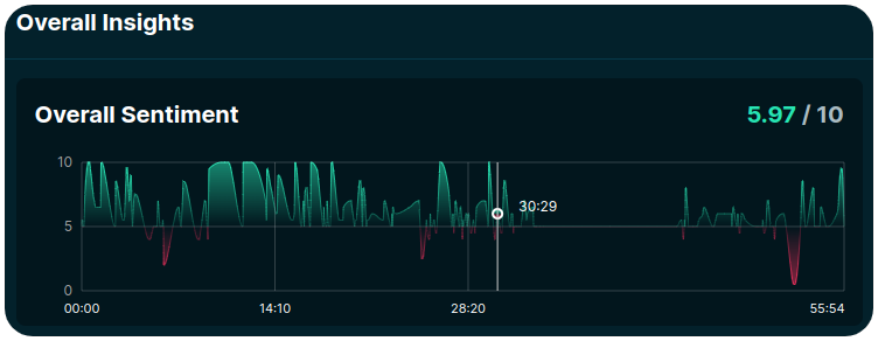

Figure 3.1: Details of the SA panel

Sentiment Analysis (SA) is a key component of the insights feature, as it allows for a deeper understanding of the feelings and beliefs of the participants in the conversation.

At Unique, we can see in figure 3.1 that two important metrics are reported. The first one is the evolution of the sentiment in function of the time. We can see that the conversation lasted approximately 55 minutes and had three negative moments, at time $t_1 \approx 7$ min, $t_2 \approx 23$ min and $t_3 \approx 52$ min. The second one is the overall sentiment in the conversation reported in the upper right corner. In this example, the score reported is 5.97.

The following sections describe how SA is performed in general and at Unique.

3.1 Multi-modal Sentiment Analysis

When analysing video calls, we have access to the video, audio and text transcription of the speaker. In the video, face and emotion detection could be used to gather precious information on the speaker's sentiments. With the audio modality, one can analyse the change in the pitches of the voice to detect different emotions.

Regarding the text modality, it provides a precise understanding of the semantics of the message pronounced by the speaker. The field that combines all the modalities to perform sentiment analysis is called *Multi-Modal Sentiment Analysis*. To this purpose, multiple datasets [18], [19] have been recently developed, allowing the emergence of new multi-modal models [20],[21],[22],[23].

In this project, we will solely focus on the text modality, mainly because of time and complexity constraints.

3.2 Sentiment Analysis in NLP

Sentiment Analysis (SA) is one of the fastest-growing study areas in AI. The availability of subjective texts on the Web created an outbreak of computer-based SA. After 2004, more than six thousand papers have been written on the subject. This makes SA an established field of research and a growing industry [24], [1].

However, language resources for SA are developed by distinct enterprises or research organisations. They are usually not shared, besides for a few publicly available resources such as WordNet-Affect [25] and SentiWordNet [26].

The process of determining the sentiment of a piece of text, such as a message in a conversation transcript, is typically performed using machine learning models, such as sentiment classification models like *Latent Dirichlet Allocation (LDA)* [27] and *Bidirectional Encoder Representations from Transformers (BERT)* [28].

3.3 Sentiment Analysis in Conversations

SA in multi-turn conversation is a subtask of SA in NLP where the context differs from a piece of text. The main goal is to understand the sentimental change of each participant in a conversation. Currently, there are not many approaches to tackling this problem. As we will see through this report, the research of new methods is critically limited by the lack of labelled interactive sentiment datasets [29].

The following chapter presents the methods implemented during this project to tackle this problem.

Chapter 4

Methods

This chapter contains the different techniques and experiences tried during the project. It is organised in chronological order of the discoveries.

4.1 Data exploration

During the first week, much time was spent understanding the inner workings of Unique, the data that is stored, and how it is stored. A meeting was held with Andreas Hauri, the CTO of Unique, where the data was explained.

A quick summary of the data organisation was provided. Users can create events, and each event can have multiple recordings, with a transcript of the conversation and a list of cues for each recording. A cue is a message in the conversation. It can be composed of multiple sentences. The data's most important features can be visualised in figure 4.1.
For each cue in the recording, some metadata was provided. It gives information like who is speaking, at what time, for how long, etc. More importantly, this is where we can find the vanilla sentiment analysis. In this chapter, the vanilla sentiment analysis refers to the model implemented at Unique. Indeed, the data contains no ground truth value but only the evaluation per cue of a vanilla model. Each cue has a field *sentiment*, which indicates the sentiment polarity of the corresponding cue. Here, it is important to note that this output is the only access to the vanilla model. The model itself was not provided. For each recording, the overall sentiment is also reported.
At Unique, the sentiment polarity ranges continuously from zero to ten. But most of the existing models output a discrete sentiment polarity taking the values -1 for a negative sentiment, 0 for a neutral sentiment and 1 for a positive sentiment.

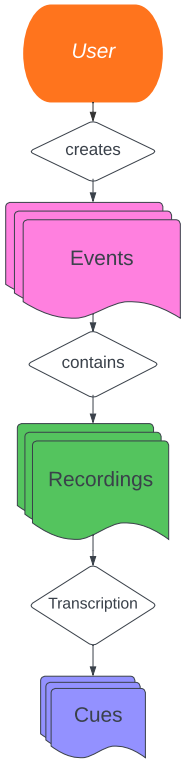

Figure 4.1: Visualisation of Unique's data structure

4.2 Model comparison at a cue level

After the data's structure was understood, a BERT model [30] that had been pre-trained on German text data was used to

analyse the sentiment at a cue level. The model was developed
to add sentiment analysis in the context of a conversational bot. In their paper [31], the authors report that BERT can perform better than models like FastText [32] on unseen data.
As Unique's data is not labelled, it would be hard to train a model from scratch to perform sentiment analysis, and therefore it makes sense to lean toward a model that performs well on unseen data.
The first question was how to compare the performance of the vanilla model (the one used currently at Unique) with newly developed models. At this point, it was decided to compare the models visually.
To compare both models, the first method was trivial. The continuous output of the vanilla model was mapped to discrete values in the following way:

$$f : [0, 10] \to \{-1, 0, 1\}; \quad f(x) = \begin{cases} 1 & \text{if } x > 0.66, \\ 0 & \text{if } 0.33 < x \leq 0.66, \\ -1 & \text{if } x \leq 0.33. \end{cases} \tag{4.1}$$

where x is the cue's sentiment polarity evaluated by the vanilla model.

The motivation behind the values 0.33 and 0.66 are to divide the range $[0, 10]$ into three distinct intervals, each representing a different sentiment polarity.

0.33 is the lower threshold for the neutral category, indicating that any value less than or equal to 0.33 will be classified as negative. Similarly, 0.66 is used as the upper threshold for the neutral category, indicating that any value greater than 0.66 will be classified as positive. This means that the interval between 0.33 and 0.66 is the neutral category, with any value in this interval classified as neutral.

The data was filtered to analyse only the german cues, and then, a counter was incremented each time both models had predicted the same polarity classification. It turned out that both models agreed only on 32% of the cases. One of the reasons is that the vanilla model outputs mostly polarities that are mapped to neutral. In contrast, the pre-trained BERT model favours rather negative and positive polarities.

As this mapping may introduce a bias towards the neutral class, another mapping was designed. The pre-trained BERT model outputs the probability of each class (Negative, Neutral or Positive), and the sentiment polarity classification is obtained by taking the $softmax$ over the probabilities. The idea is to remove the $softmax$ layer and map the probabilities to continuous output. The mapping is defined in equation 4.2.

Let x, y, z be the probability of the classes Negative, Neutral and Positive, respectively, i.e. $x + y + z = 1$. The following equation defines the mapping:

$$\begin{aligned} g : [0, 1] \times [0, 1] \times [0, 1] \to [0, 10]; \\ g(x, y, z) = 0x + 5y + 10z \end{aligned} \tag{4.2}$$

0, 5 and 10 are weights attributed to each probability so that the output will range between $[0, 10]$.

This approach assumes that the sentiment probability is correlated with the sentiment strength. It is a strong simplification done principally for two reasons; It allows us to compare models together, but more importantly, Unique's front-end expects

a continuous sentiment that ranges continuously between 0 and 10. Better techniques to learn the sentiment direction and strength are discussed in this paper [33].

Another suggestion is to include more information, such as towards which entity, subject or characteristic the sentiment is directed within the text. This research field is called *Aspect-Based Sentiment Analysis (ABSA)* and is discussed in more detail in this paper [34]

With this mapping on the pre-trained BERT model, both models rank the sentiment polarity on a scale of 0 to 10. On figure 4.2, we can observe how the BERT model (in blue) offers a wide output range. In contrast, the vanilla model (in orange) is always close to 5, or a corresponding neutral sentiment.

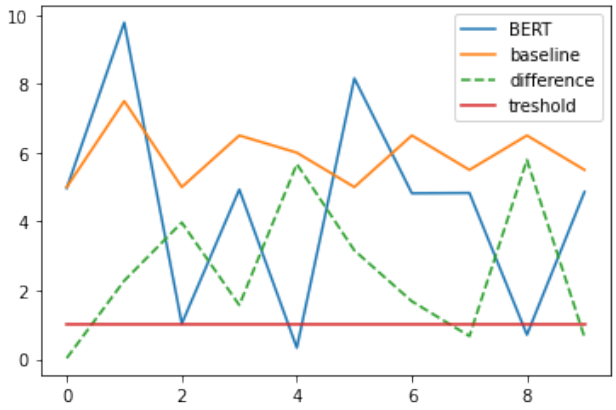

Figure 4.2: Comparison between BERT fine-tuned and baseline at a cue level. The baseline is the model currently used at Unique. BERT is a transformer-based model. The green dotted line plots the differences between the baseline and BERT. A difference below the threshold indicates that both models agree on sentiment polarity.

The issue with the vanilla model is that at the end of the day, as it always outputs values around 5 (Neutral polarity), it does not provide great insights into the overall sentiment of the conversation. For this reason, the outputs obtained by the pre-trained BERT model are promising.

4.3 Model comparison in a Conversation

Instead of working on german data, English data will be used from now on. This choice was motivated mainly because most of the existing datasets for sentiment analysis are in English. Therefore, pre-trained models on English text data may perform better than german on unseen data.

When analysing the sentiment in a conversation setup, the main difference with a single cue setup is that the previous conversation cues also play a role. The vanilla model obtains the overall sentiment by averaging the sentiment polarity over all the individual cues. The central problem with this approach is that it does not consider the order of the cues and their impact on the conversation. An example of a conversation is provided in table 4.1. The number at the end of each cue indicates the participant's sentiment. The overall sentiment is the average over the sentiment

Table 4.1: Example of a conversation at Unique.

A: I'm really excited about the new project we're working on. Do you have any ideas for how we can improve it? 7.2 **B**: I think we should focus on making the user interface more user-friendly. Maybe we can conduct some user testing to gather feedback. 5.1 **A**: That's a great idea! I'll reach out to our design team and see if they can come up with some mockups. I also think it would be beneficial to gather feedback from our target market. 6.8 **B**: Sounds good; I'll start conducting some market research to see what our target audience is looking for in a product like this. 4.4 **A**: Great! I think by working together, we can really make this project a success. What are your thoughts on the project deadline? 5.6 **B**: I think we should aim to have a minimum viable product ready in 6 months. That way, we can gather feedback and make improvements before releasing the final product. 4.8 **A**: I agree, 6 months sounds like a good deadline. Let's make sure we're on track and make adjustments if needed. 6.7 **B**: Absolutely, I'm looking forward to working on this project with you. 5.9 **Overall Sentiment :** 6.1

of all cues.

To fix that, the new approach was the following:

Rather than extracting sentiment polarity at a cue level, multiple following cues of the same user are merged so the model can retrieve more context. This is useful when a user speaks for a long time or numerous times in a row. The merged cues were then analysed with the large version of a model named RoBERTa [6], a model based on BERT that is trained longer, with bigger batches and more data. Using this technique, two problems quickly emerged.

The first one was that with a larger model, the maximum sequence length remains the same even if the number of parameters increases. So the merged cues could still be longer than the maximum number of tokens accepted by the model.

The second issue is that this approach loses too much context information. For example, in the conversation in table 4.2, the overall sentiment is interpreted as negative, whereas it should be positive because the participant is not injured.

To address the second issue, RoBERTa was used to analyse the entire conversation. The first issue remains; The maximum number of tokens accepted by RoBERTa is 512, and most discussions are longer. This paper [35] suggests six different strategies to deal with long text: 1) head-only, 2) tail-only, 3) head+tail, 4) hierarchical mean, 5) hierarchical max, and 6) hierarchical self-attention. Of those six strategies, only one was implemented because the discovery of ScenarioSA (see section 4.4) allowed a new promising solution. This solution is further explored in the following section.

Table 4.2: Importance of context in a conversation

A: Did you hurt yourself ? 3.1
B: No, I did not. 3.5
Overall Sentiment : 3.3

4.4 Conversational Dataset : ScenarioSA

The technique developed in the previous section still lacks one significant thing, which is that we cannot compare the models' performances qualitatively. As mentioned, the models are not learning from Unique's conversational data. They are pre-trained on other datasets and then used for inference purposes. Based on this observation, some literature research was done to find a dataset containing ground truth labels.

A promising dataset, ScenarioSA [29], was found. ScenarioSA is a dyadic conversational database for interactive sentiment analysis. In this paper, they introduced a manually labelled dataset of conversations like the one presented in table 4.3.

Table 4.3: ScenarioSA example

A: You voted, right? 0
B: You know I did. 1
A: Who did you vote for? 0
B: I voted for NAME, of course! 1
A: Can you believe that he actually won? 0
B: I knew he would win. 1
A: I didn't think he would. 0
B: He was the top candidate. 1
A: I figured people wouldn't vote for him because he's African American. 0
B: That just goes to show that America is finally turning over a new leaf. 1
A: You're absolutely right. 1
B: I'm excited that NAME NAME is our President. 1
A Overall Sentiment : 1
B Overall Sentiment : 1

The number at the end of each message represents the corresponding sentiment polarity (-1 for negative, 0 for neutral, 1 for positive). It is independent of the context. It only depends on the message itself. The two numbers at the end of the conversation represent the overall sentiment of each participant. The dataset contains only conversations with two participants.

Using this dataset, a new model could be developed and analysed. The model designed worked in two steps, and the first step was predicting the sentiment of each message independently. BERTweet [5], Texblob [36] and, as in the previous section,

RoBERTa were used for that. The latter obtained significantly better results and performed well on unseen data [6].

The results obtained by RoBERTa are depicted in section 4.4. We can see that the model has an accuracy of 81%. However, it does not predict very well the negative sentiments.

	precision	recall	f1-score	support
Positive	0.81	0.75	0.78	16365
Neutral	0.84	0.86	0.85	43670
Negative	0.67	0.68	0.67	11043
accuracy			**0.81**	71078
macro avg	**0.77**	**0.76**	**0.77**	71078
weighted avg	**0.81**		**0.81**	71078

Table 4.4: Performance of the RoBERTa on ScenarioSA.

The second step is the aggregation of the independently analysed sentiments to the overall sentiment. In their paper [29], the authors present different techniques. In the following subsection, we will see the results of some of them because they provide great insights on achieving a suitable performing merge.

For the following merge definitions, let $\{a_i\}_{i=1}^{n}$ be the ordered list of sentiment polarity associated with each message of participant A, with $a_i \in \{-1, 0, 1\}$ and n the number of messages.

4.4.1 Vanilla Merge

The idea behind the vanilla merge is straightforward; The overall sentiment participant is fully reflected in his last message. So we have :

$$A_{overall} = a_n \tag{4.3}$$

The results of the vanilla merge are reported in table 4.5. We can also analyse the model's behaviour with the confusion matrix (fig. 4.3).

Table 4.5: Performance of the Vanilla merge

	precision	recall	f1-score	support
Positive	0.76	0.78	0.77	2894
Neutral	0.60	0.67	0.63	2308
Negative	0.61	0.47	0.53	1317
accuracy			**0.68**	6519
macro avg	**0.66**	**0.64**	**0.64**	6519
weighted avg	**0.68**	**0.68**	**0.67**	6519

4.4.2 Mean Merge

The mean is defined in the following way :

$$A_{overall} = \frac{\sum_{i=1}^{n} a_i}{n} \tag{4.4}$$

The results of the mean merge are reported in table 4.6. We can also analyse the model's behaviour with the confusion matrix (fig. 4.4).

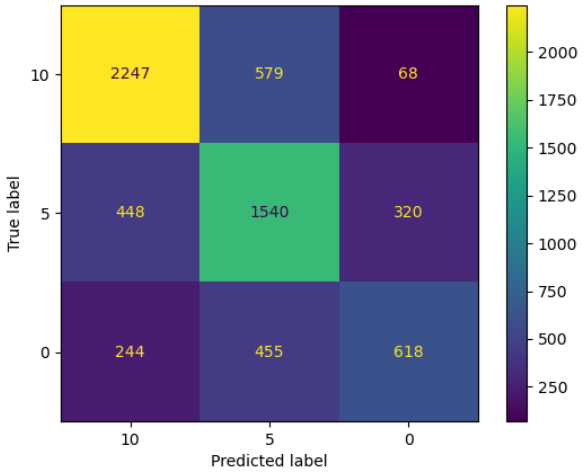

Figure 4.3: Confusion matrix of the vanilla merge, 10 stands for a positive classification, 5 neutral, 0 negative

Table 4.6: Performance of the mean merge

	precision	recall	f1-score	support
Positive	0.90	0.46	0.61	2894
Neutral	0.47	0.91	0.62	2308
Negative	0.86	0.40	0.55	1317
accuracy			**0.61**	6519
macro avg	**0.74**	**0.59**	**0.59**	6519
weighted avg	**0.74**	**0.61**	**0.60**	6519

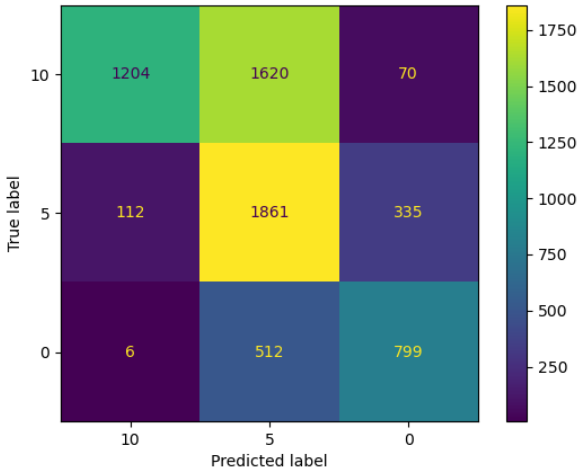

Figure 4.4: Confusion matrix of the mean merge, 10 stands for a positive classification, 5 neutral, 0 negative

4.4.3 Exponentially Weigthed Merge

The exponentially weighted merge is defined in equation 4.5.

$$A_{overall} = \frac{\sum_{i=1}^{n}(a_i \times 2^i)}{2^n} \tag{4.5}$$

The results of the exponentially weighted merge are reported in section 4.4.3. We can also analyse the model's behaviour with the confusion matrix (fig. 4.5).

	precision	recall	f1-score	support
Positive	0.87	0.72	0.79	2894
Neutral	0.57	0.83	0.68	2308
Negative	0.78	0.46	0.58	1317
accuracy			**0.71**	6519
macro avg	**0.74**	**0.67**	**0.68**	6519
weighted avg	**0.75**	**0.71**	**0.71**	6519

Table 4.7: Performance of the exponentially weighted merge

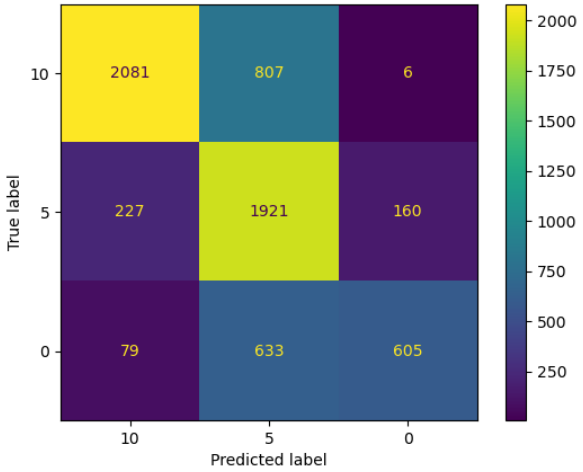

Figure 4.5: Confusion matrix of the exponentially weighted merge, 10 stands for a positive classification, 5 neutral, 0 negative

We can compare the different merge techniques with the tables of results and the confusion matrices. The fact that the vanilla merge performs better than the mean merge tells us that the last messages may carry useful information. The exponentially weighted merge tries to take advantage of this observation but tends to predict too much towards the neutral sentiment.

Now that we have a working model that shows good results, it is time to compare it again with the vanilla model used at Unique. The same issue remains; The model developed on ScenarioSA outputs discrete polarities when analysing the sentiment at a cue level. To compare it with Unique, continuous polarities are needed.

To fix that, the mapping defined in equation 4.2 is used to obtain, for each message, a sentiment polarity that ranges continuously between 0 and 10. Then, the

exponentially weighted sum is adapted in the following way, as shown in equation 4.6.

Let $a_i \in [0, 10]$,

$$A_{overall} = \frac{\sum_{i=1}^{n}(a_i \times 2^i)}{10 \times 2^n} \tag{4.6}$$

A comparison of the two models' behaviour, when used to analyse a conversation in table 4.8, is shown in fig. 4.6. Note that the overall sentiment of the model using RoBERTA and the exponentially weighted merge is obtained by taking the mean of the overall sentiment of each participant.
As this is still a quantitative and subjective way of comparing models, we did a survey among the sales team at Unique to get a better quantitative measure of the models' performances. The creation of the survey and its results are depicted in chapter 5.

Table 4.8: Conversation 5

A: I'm worried about this meeting. 3.5, 1.7
B: Why? 5, 5
A: I'm meeting with a really tough negotiator. 5, 3.7
B: Be confident, you can do it. 6, 9.8
A: I'm just worried because our company really needs this contract. 2.5, 1.8
B: I know we do, but don't let him know how much we need. 5, 3.1
A: I won't. I will look tough. 5, 2.9
B: Don't ever let him see you sweat. That will make him think you're nervous. 4, 2.8
A: I am nervous ! 5, 2.7
B: Yeah, but he doesn't need to know that 4.8, 2.7
Overall Sentiment : 4.55, 2.94

Table 4.9: *
In red is the output of Unique's model, and in green is the output of RoBERTa

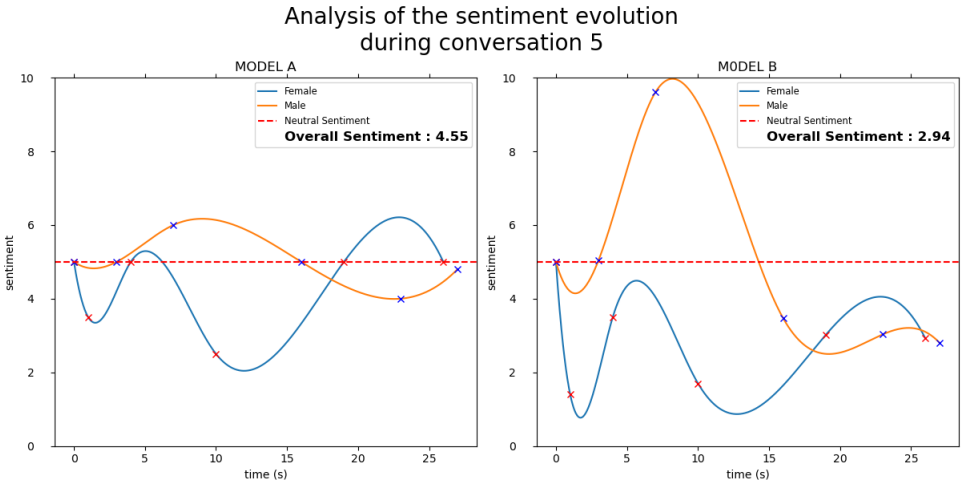

Figure 4.6: Comparison between Unique's model (left) with the one composed of RoBERTa + exponentially weighted merge (right).

Chapter 5

Results

As pointed out in the previous chapter, comparing the performance of the models is challenging, as we do not have access to any ground truth label. It was then decided to compare models quantitatively by plotting their outputs and analysing the differences in their behaviour. The survey design and corresponding results are discussed in the next section.

5.1 Survey Design

The goal of the survey is for the team at Unique to evaluate which model has the best output out of ten discussions. Thirteen people from Unique's team replied to the survey between the 20 of December and the 17 of January. These discussions were taken randomly in the ScenarioSA dataset and then analysed with both models.

At first, the video with the audio was included in the question, and the survey participant would have watched the video to rate the models. But, as it appeared that the actors interpreting the conversations might bias the raters, it was chosen to include only a link to the audio. Examples of the survey questions are depicted in figures 5.1 and 5.2. The complete survey can be found here.

5.2 Survey Discussion

As shown in table 5.1, the survey participants preferred Unique's model in most cases. Unique's model obtained a majority of the vote in 5 cases out of the ten conversations, the newly developed model earned it only twice, and the rest were draws.
As disappointing as they appear, these poor results can be explained by multiple reasons.

First, the design of the survey could be at fault. The actors' intonation may have biased the survey respondents. Both models perform on text data, and the models do not have access to audio information to perform the sentiment analysis. So it would have made more sense to include only the conversation transcript in the survey, not the audio.

Another reason could be the interpolation function. As we can see on figure 5.1, an interpolation was made between the points where the sentiments are analysed.

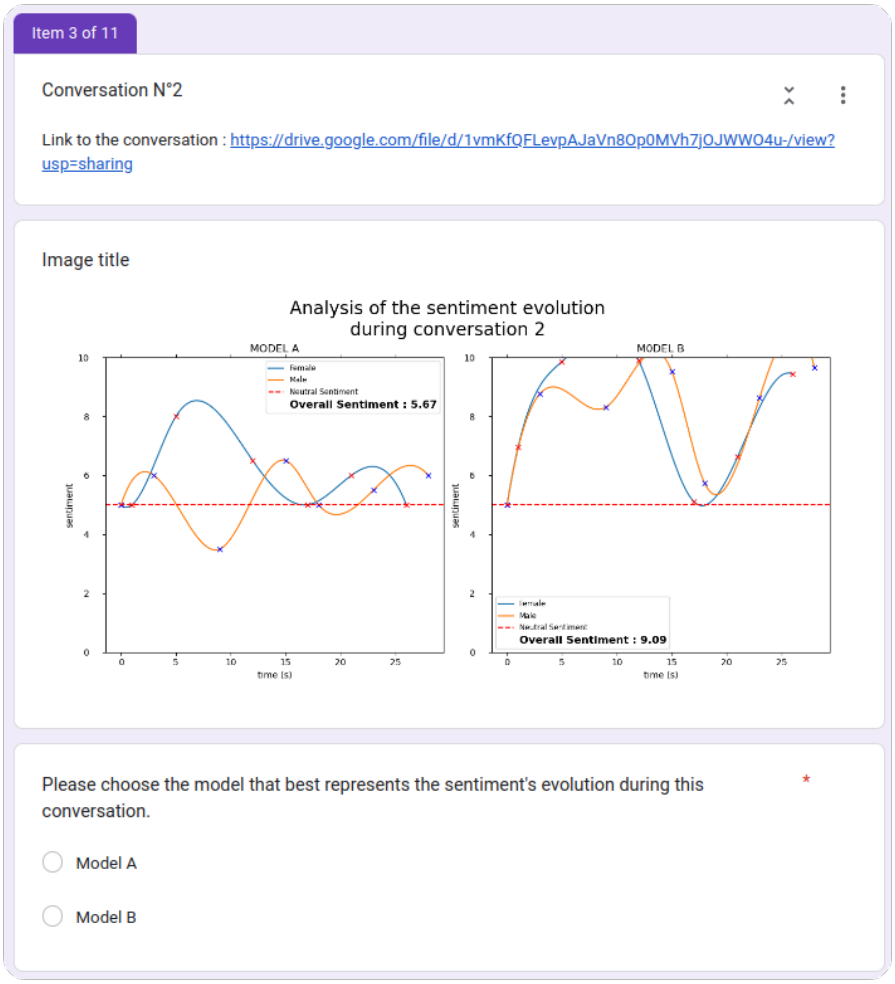

Figure 5.1: Example of one question in the survey

The reason behind this interpolation is merely to make the plot nicer. The issue with this interpolation is that it does not represent the model behaviour perfectly, and the interpolated values may be out of the plot's scope. The survey respondents may have misinterpreted this.

For the first conversation, the votes are the most unbalanced (83% for Unique's model, only 17% for the new model (see table 5.1)). Looking at the model outputs (see fig. 5.2), we can see that it is a perfect representation of the models' behaviour differences. On the left, Unique's model is flat, predicting mostly values around a neutral sentiment. On the right, the variance of the output value is much bigger.

Indeed, it seems like variations in the output of the new model can not fully follow as much interest as expected. In the chapter 6, we will consider other approaches that could achieve better results.
.

Table 5.1: Results of the survey

	Unique	New Model
Conversation 1	0.83	0.17
Conversation 2	0.5	0.5
Conversation 3	0.75	0.25
Conversation 4	0.83	0.17
Conversation 5	0.58	0.42
Conversation 6	0.42	0.58
Conversation 7	0.5	0.5
Conversation 8	0.33	0.67
Conversation 9	0.5	0.5
Conversation 10	0.67	0.33

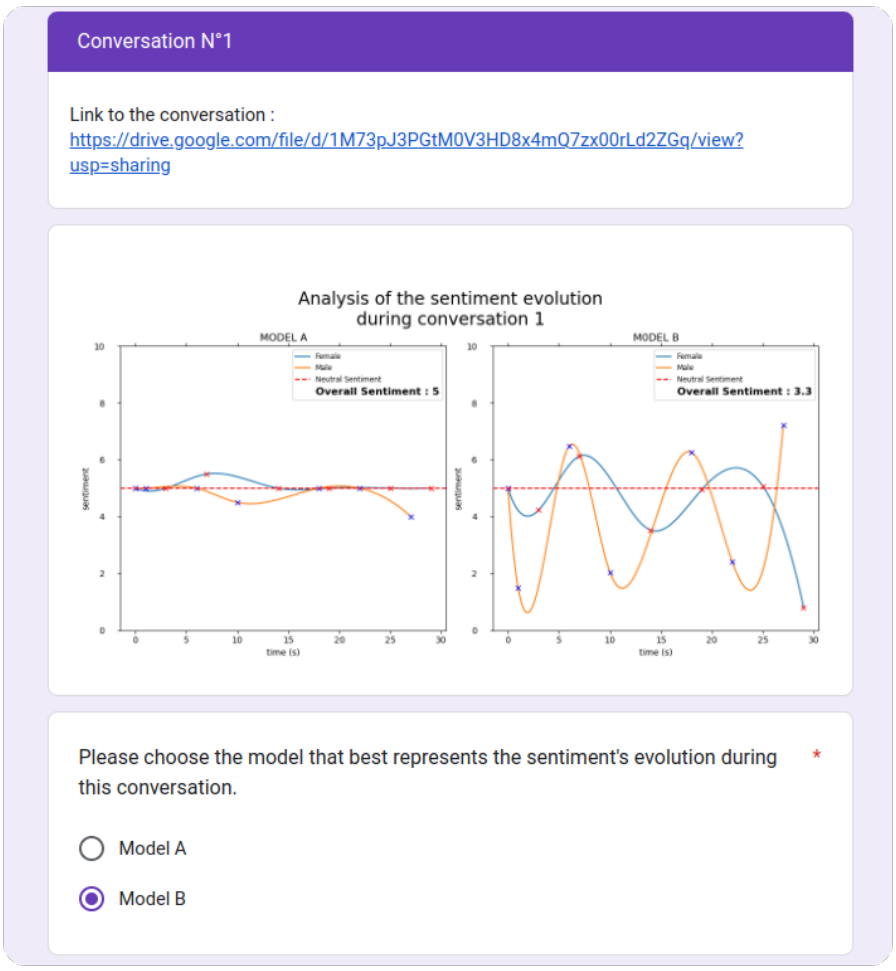

Figure 5.2: Example of one question in the survey

Chapter 6

Discussion

SA at Unique This report provides an example of what can be achieved by doing SA on the transcription of video calls. It highlights some improvements that could be made in the sentiment analysis panel of Unique's front-end—for instance, displaying the evolution of the sentiment of each participant in the conversation independently. Moreover, each participant should have their overall sentiment presented.

SA in text messages As with most of the current research [37], [38], the model developed during this project is focused mainly on classifying user sentiments into sentiment polarities (i.e., positive, negative, or neutral), but not sentiment strength. The mapping done to get a continuous output works conceptually, but it was mainly performed to adapt to Unique's front-end. Still, there are a few existing techniques [39], [40], [41] that directly output the sentiment polarity with the sentiment strength. Only rare studies have considered the sentiment strength, but this aspect is crucial for multiple applications, Unique included.

SA in conversations SA in a conversational context differs from typical SA. The surrounding utterances are considered context, which assimilates rich relations between multiple speakers, entangling the exchange of information and how one speaker influences another [42]. There exists limited research that focuses on sentiment in multi-turn conversation [43], [44], and even fewer available datasets [29] to train high-performing models. The exponentially-weighted merge (see eq. (4.6)) used in this project does not take full advantage of the context given by the conversation. A better technique could use longformers [45]. Longformer can attend to longer sequences and, thus, overcome one major drawback of models like BERT or RoBERTa. Regardless, the use of LLMs could be helpful for this task. For instance, GPT-3 with a chain of thoughts [17] prompt could perform well.

SA in multimodal conversations Restricting ourselves to only using text data (mainly because of the simplicity and time constraint) is a tremendous loss in the quantity of information available in video calls. A significant number of multimodal records of communications between people have been produced by the recent growth of social platforms like TikTok, Twitter, Instagram or Reddit [46], [47], [48]. Such conversational data intrinsically combines NLP, facial expression detection, motions and speech modulations. Those are rich sources of information that could be used together to analyse the sentiment of speakers in conversations better.

Chapter 7

Conclusion

This semester project at Unique was an excellent opportunity to better understand the organisation behind a fast-growing startup, like the importance of the weekly check-ins that provide remarkable productivity boosts throughout the semester. Furthermore, it provided great insight into how AI could improve daily used tools.

This experience also demonstrates the difficulties and constraints that AI developers need to consider when developing new models. The survey results show that the model should not only perform excellently, but the output should also be of high quality.

On the technical side, multiple possibilities exist to improve the current Unique SA model. The NLP model could be improved by using models capable of interpreting sentiment intensities using new transformer models or LLMs.

Moreover, Audio and Video modalities could be added to gain a more profound knowledge of the underlying participant's sentiment in online meetings.

Acknowledgement

I want to extend my gratitude to Nathalie and Mireille, who recorded the conversation examples with me.
I would also like to thank Hanno for his collaboration throughout the semester and Adam for his detailed and helpful corrections on this report.
My thanks also go to Pascal, who consistently made time to check in weekly with me and offer valuable guidance.
Most importantly, I am thankful to Prof. Perez-Cruz for his overall supervision and the autonomy he allowed me to have during this project.

Bibliography

[1] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[5] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[7] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. H. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," *CoRR*, vol. abs/1512.02595, 2015.

[8] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *CoRR*, vol. abs/1508.01211, 2015.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[10] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, 2021.

[11] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *CoRR*, vol. abs/1610.05256, 2016.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[14] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2022.

[15] S. Roy and D. Roth, "Solving general arithmetic word problems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1743–1752.

[16] K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *CoRR*, vol. abs/2110.14168, 2021.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *CoRR*, vol. abs/2201.11903, 2022.

[18] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, 2016.

[19] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246.

[20] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 873–883.

[21] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Information Fusion*, vol. 76, pp. 204–226, 2021.

[22] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis," *Information Fusion*, vol. 66, pp. 184–197, 2021.

[23] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *CoRR*, vol. abs/1705.09406, 2017.

[24] C. Iglesias, J. Sánchez-Rada, G. Vulcu, and P. Buitelaar, "Chapter 4 - linked data models for sentiment and emotion analysis in social networks," in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Boston: Morgan Kaufmann, 2017, pp. 49–69.

[25] C. Strapparava, A. Valitutti *et al.*, "Wordnet affect: an affective extension of wordnet." in *Lrec*, vol. 4, no. 1083-1086. Lisbon, Portugal, 2004, p. 40.

[26] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[27] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[29] Y. Zhang, Z. Zhao, P. Wang, X. Li, L. Rong, and D. Song, "Scenariosa: A dyadic conversational database for interactive sentiment analysis," *IEEE Access*, vol. 8, pp. 90 652–90 664, 2020.

[30] "German sentiment BERT finetuned on news data," https://huggingface.co/mdraw/german-news-sentiment-bert?text=I+like+you.+I+love+you, accessed: 2023-01-21.

[31] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Training a broad-coverage german sentiment classification model for dialog systems," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1620–1625.

[32] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016.

[33] T. T. Thet, J.-C. Na, and C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of information science*, vol. 36, no. 6, pp. 823–848, 2010.

[34] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 187–196.

[35] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" *CoRR*, vol. abs/1905.05583, 2019.

[36] S. Loria and al., "textblob documentation," *Release 0.15*, vol. 2, no. 8, 2018.

[37] J. Du, L. Gui, Y. He, R. Xu, and X. Wang, "Convolution-based neural attention with applications to sentiment classification," *IEEE Access*, vol. 7, pp. 27 983–27 992, 2019.

[38] B. Bansal and S. Srivastava, "Hybrid attribute based sentiment classification of online reviews for consumer intelligence," *Applied Intelligence*, vol. 49, no. 1, pp. 137–149, 2019.

[39] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual lstm model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, 2018.

[40] Y. Lu, X. Kong, X. Quan, W. Liu, and Y. Xu, "Exploring the sentiment strength of user reviews," in *International Conference on Web-Age Information Management*. Springer, 2010, pp. 471–482.

[41] H. Chen, X. Li, Y. Rao, H. Xie, F. L. Wang, and T.-L. Wong, "Sentiment strength prediction using auxiliary features," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 5–14.

[42] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, and J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Information Fusion*, vol. 93, pp. 282–301, 2023.

[43] F. Cui, H. Di, L. Shen, K. Ouchi, Z. Liu, and J. Xu, "Modeling semantic and emotional relationship in multi-turn emotional conversations using multi-task learning," *Applied Intelligence*, vol. 52, no. 4, p. 4663–4673, mar 2022.

[44] C.-H. Kao, C.-C. Chen, and Y.-T. Tsai, "Model of multi-turn dialogue in emotional chatbot," in *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, 2019, pp. 1–5.

[45] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *CoRR*, vol. abs/2004.05150, 2020.

[46] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in *A practical guide to sentiment analysis*. Springer, 2017, pp. 1–10.

[47] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, and B. Wang, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Information Fusion*, vol. 62, pp. 14–31, 2020.

[48] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: a transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.