Utrecht University

Master Thesis

Multimodal Analysis of Acoustic and Linguistic Features in Entrepreneurial Pitches using Deep Learning

Author:
P.W.O. van Aken
Universiteit Utrecht
p.w.o.vanaken@students.uu.nl

Supervisor:
Assist. Prof. Dr. I. Önal Ertuğrul
Universiteit Utrecht
i.onalertugrul@uu.nl

Second Examiner:
Assist. Prof. Dr. H. Kaya
Universiteit Utrecht
h.kaya@uu.nl

In Partial Fulfillment of the Requirements for the Degree of Masters of Science in Artificial Intelligence at Universiteit Utrecht

January 25, 2023

ABSTRACT

Multimodal Analysis of Acoustic and Linguistic Features in Entrepreneurial Pitches using Deep Learning

by Pepijn van Aken

Acquiring early-stage investments for the purpose of developing a business is a fundamental aspect of the entrepreneurial process, which regularly entails pitching the business proposal to potential investors. Previous research suggests that business viability data and the perception of the entrepreneur play an important role in the investment decision-making process. This perception of the entrepreneur is shaped by verbal and non-verbal behavioural cues produced in investor-entrepreneur interactions. This study explores the impact of such cues on decisions that involve investing in a startup on the basis of a pitch. A multimodal approach is developed in which acoustic and linguistic features are extracted from recordings of entrepreneurial pitches to predict the likelihood of investment. The acoustic and linguistic modalities are represented using both hand-crafted and deep features. The capabilities of deep learning models are exploited to capture the temporal dynamics of the inputs. The findings show promising results for the prediction of the likelihood of investment using a multimodal architecture consisting of acoustic and linguistic features. Models based on deep features generally outperform hand-crafted representations. Across multiple explainable models, consistent features are found to be important predictors. An early fusion multimodal model consisting of deep representations of the two modalities has been proven to be most predictive.

ACKNOWLEDGEMENTS

First of all I would like to thank my thesis supervisor, Itir, for her unwavering support and guidance during the course of this project. Her feedback and knowledge helped me immensely throughout the research and writing process. All the meetings in which we discussed the progress of the project kept me focused and motivated. I would also like to thank Heysem for his contributions to the project and for acting as the second examiner. Next I would like to thank Werner for letting me use the *Data Management Entrepreneurial Pitches* data set. Finally, I would like to thank my fellow student Francois, for the interesting discussions and fun study sessions.

Contents

Abstract **1**

List of Figures **5**

List of Tables **6**

1 Introduction **7**

1.1 Research motivation 7

1.2 Contributions to existing work 10

1.3 Research questions 11

1.4 Thesis outline 12

2 Literature Study **13**

2.1 Literature Study: Decision-making in entrepreneurial contexts 13

2.1.1 Studying social interactions 14

2.1.2 Investor decision-making based on entrepreneurial pitches 15

2.2 Literature Study: Methodology 18

2.2.1 Acoustic modality 19

2.2.2 Linguistic modality 26

2.2.3 Multimodal analysis 33

3 Methodology **40**

3.1 Data 40

3.1.1 Overview of the data set 40

3.1.2 Previous work on this data set 42

3.2 Feature extraction 42

3.2.1 Pre-processing the data 42

3.2.2 Acoustic features 44

3.2.3 Linguistic features 45

3.3 Experimental Setup 46

3.3.1 Unimodal models: single set of feature representations 46

3.3.2 Unimodal models: combining deep and hand-crafted feature sets . 47

3.3.3 Multimodal models: different fusion techniques 49

3.3.4 Explainable acoustic, linguistic and multimodal models 50
3.3.5 Generalizing the models to online settings 51
3.4 Training and evaluation 51

4 Results 54
4.1 Unimodal models: single feature representations 54
4.2 Unimodal models: combining feature representations 55
4.3 Multimodal models 56
4.4 Explainable models 58
4.5 Cross-domain experiment 64

5 Discussion & Conclusion 66
5.1 Overview of the results 66
5.2 Answering the research questions 72
5.3 Limitations and future research 73
5.4 Conclusion 75

Bibliography 77

Appendix 86

List of Figures

2.1 The convolutional layers replace hand-crafted features. Image by Trigeorgis et al. (2016) 24
2.2 Using text to make predictions. Image by Kowsari et al., (2019) 27
2.3 Early fusion versus late fusion. Image by Zhang et al. (2021b) 34
2.4 The multimodal model to estimate self disclosure by Soleymani et al. (2019) 37
2.5 The multimodal model to classify motivational interviewing by Tavabi et al. (2020) 38

3.1 Distribution of in-person pitches 44
3.2 Distribution of online pitches 44
3.3 Unimodal models consisting of either hand-crafted or deep feature sets . . 46
3.4 Unimodal acoustic models consisting of both hand-crafted and deep features 47
3.5 Unimodal linguistic model consisting of both hand-crafted and deep features 48
3.6 Multimodal model - early fusion: hand-crafted and deep feature sets for each modality 49
3.7 Multimodal model - early fusion: best performing feature set of each modality 50

4.1 SHAP summary plots for the LIWC models 60
4.2 SHAP summary plots for the openSMILE models 62
4.3 SHAP summary plots for the multimodal models 63

List of Tables

3.1 Hyper-parameters included in Grid Search for Xgboost Regressor models 52
3.2 Hyper-parameters included in Grid Search for single input GRU models . 52

4.1 Unimodal models: single feature representation 55
4.2 Unimodal models: combined feature representations 56
4.3 Multimodal models: different fusion techniques 57
4.4 Explainable acoustic, linguistic and multimodal models 58
4.5 Explainable features that appear among the top 20 of feature importance in at least 3 out of 4 models for the respective linguistic, acoustic and multimodal models 59
4.6 Cross-domain results 64
4.7 An overview of the best performing models 65

Chapter 1

Introduction

This thesis aims to contribute to the field of entrepreneurial decision-making, with a focus on investment decisions for startups based on a pitch. In Section 1.1, the motivation for the research is formulated and the potential of a multimodal machine learning model to study this context is examined. In Section 1.2, the contributions of this work to existing literature are presented. Subsequently, the research question and sub-questions are introduced, followed by a brief overview of the experiments used to answer these questions. Finally, an outline for this thesis is provided in Section 1.4.

1.1 Research motivation

Delivering a successful elevator pitch on a business proposal in front of investors is an intimidating challenge for many entrepreneurs. Convincing investors of the potential of a business plan and raising funds to realize the plan are critical parts of the entrepreneurial process. We are only beginning to understand how investors make decisions regarding investments (Clarke et al., 2019).

Decision-making, the process of determining the most appropriate course of action based on available information, in the field of entrepreneurship is characterized by high levels of uncertainty (Shepherd et al., 2015). Berner et al. (2012) claim that the main premise of entrepreneurship is to accept high levels of risk while investing or producing a good. Since factual information to base a decision is often lacking or limited, decision makers in entrepreneurial contexts rely on heuristics (Huang and Pierce, 2015). The high levels of uncertainty and the use of heuristics make it difficult to analyze the decision-making process, and due to its complexity, this process has attracted a lot of academic research (e.g. Wadeson, 2006 ; Shepherd et al., 2015). It is especially interesting to study decisions that involve social interactions, since due to the lack of factual information, this social interaction itself can influence the decision.

The interaction between a pitching entrepreneur and an evaluating investor is such an entrepreneurial setting based on social relationships that is marked by high uncertainty. The investor has to make an assessment of the feasibility of a project based on the pitch and financial data. However, research suggests that investors also rely on subtle social cues they extract from the pitch. Huang and Pierce (2015) find that investors both rely on intuition and formal analysis when making this decision. Furthermore, they find that this intuition is for a large share based on the perception of the founding entrepreneur. Multiple studies show that this perception of the entrepreneur is shaped by verbal and non-verbal cues in the pitch. Martens et al. (2007) find that the use of language and storytelling plays a key role in entrepreneur-investor interactions. Chen et al. (2009) study the influence of non-verbal behaviour cues on perceived passion of an entrepreneur. Clarke et al. (2019) study a combination of verbal cues and hand gestures and find a strong effect on funding decisions.

Processing the verbal and non-verbal cues emitted by pitchers could open up a valuable source of information for research into investor decision-making. To unlock the potential of these signals, NLP can be used to analyze the cues in the use of language in the pitch, while the Social Signal Processing domain provides the tools to automatically code nonverbal signals, resulting in a more accurate and efficient analysis (Liebregts et al., 2020). However, as noted by Clarke et al. (2019), the effect of verbal and non-verbal communication strategies are often studied in isolation. Since social interactions are the interplay of verbal and non-verbal cues, integrating them in a single analysis could provide interesting new insights. The class of vocal behaviour has been identified as a potential important driver for investments in entrepreneurial pitches (Huang and Pearce, 2015 ; Clarke et al. 2019), but has not been studied in a combined model with verbal cues yet. This is a current gap in existing literature which has been explored in this thesis.

Extracting features from the raw data is one of the main challenges when using acoustic and linguistic data. Traditionally, both these modalities have been studied using hand-crafted feature sets. For the acoustic features of speech, features such as pitch and loudness can directly be extracted from the audio signal and used to make predictions such as emotion classification (Luengo et al., 2005 ; Marchi et al., 2016). For language, a Bag-of-Words (BoW) model can be used to create a vector that represents the input text. Despite the fact that hand-crafted feature sets have been successfully applied on a number of tasks, there are some limitations regarding this approach, such as modelling the context of a text. The development of deep learning enabled the creation of models that can learn to extract feature representations themselves. Furthermore, in combination with deep feature embeddings, deep encoders can capture the temporal dynamics of the signals, leading to better performance (Zhou, 2021). Currently, these deep learning based feature extractors have become state-of-the-art in audio and language research. However, this does not imply that hand-crafted feature sets are not useful anymore. Elbanna et al. (2022) and Johnson and Marcellino (2022) argue that using an ensemble of hand-crafted and deep feature sets can lead to an increase in performance and interpretability of a model.

For a long time, vocal and verbal behavioural cues have been studied separately in the form of unimodal models. However, research suggest that verbal and non-verbal, such as vocal cues, are tightly coupled communication mechanisms that each carry complementary but distinct meanings (Clarke et al., 2019). Therefore, to properly analyze the role of vocal and verbal cues in decision-making in entrepreneurial contexts, both have to be considered in a single model. This is where multimodal models come into play. In a multimodal model, different unimodal models are combined and thus capture a wider range of behaviour. Different techniques exist to fuse unimodal models into a multimodal model, such as early and late fusion (Poria et al. 2017).

This thesis presents a multimodal model consisting of both acoustic and linguistic features to predict the likelihood of investment of entrepreneurial pitches. This study could provide insights for investors and researchers into the decision-making process and help entrepreneurs enhance their pitching abilities. In this study the *Data Management Entrepreneurial Pitches* data set is used, which is issued by Tilburg University and contains video recordings of pitchers and potential investors during a pitch competition (2020). Firstly, unimodal models consisting of either acoustic or linguistic features are developed to predict the probability of investment, using different types and combinations of feature representations. We experiment using hand-crafted representations and deep representations. It is hypothesised that integrating these two types of, potentially complementary, feature set into a single model to represent a modality could result in an improvement in model performance. Deep learning models are developed to capture the complex non-linear relationships and temporal dynamics of the acoustic and linguistic input data. Then we examine what the effect is of combining the two modalities into a single multimodal model. It is hypothesised that when analyzing different types of behavioural cues in one model, a more truthful representation of reality is created, allowing us to capture more information of the social interaction in the model. Here, we also examine the effect of applying different fusion strategies on the model performance. Finally some explainable models are considered, for which common important features are discussed, and we test how well the models generalize to a different setting in a cross-domain experiment.

1.2 Contributions to existing work

Overall, this study contributes to existing literature in four main ways. Firstly, most prior works that apply a multimodal model in an entrepreneurial context study crowdfunding campaigns (e.g. Kaminski et al., 2020 ; Cheng et al., 2019). As noted by Liebregts et al. (2020) crowdfunding decisions do not always require social interactions and for this reason the effect of verbal and non-verbal social cues on these decisions is rather limited. In this work, the direct effects of verbal and non-verbal behavioural cues during social interactions in the context of entrepreneurship are analyzed. Furthermore, we study how these cues impact the uncertainty that is inherent to these types of decisions.

Secondly, we have developed a deep-learning based model consisting of a combination of verbal and vocal features to predict the probability of investment. Research on pitching has mostly been focused on studying verbal and non-verbal behaviour in isolation (Huang and Pierce, 2015). Since social interactions are formed by the interplay of different kind of social cues, you miss out on important information when analyzing them separately. Clarke et al. (2019) do use a mixed methods approach and look at the combination of verbal features and gestures of a speaker. They find strong evidence that a combined model can predict investment decisions. Although Clarke et al. (2019) focus on gestures, they suggest that the combination of verbal and vocal behaviour may also shape the interactions of entrepreneurs and investors. We draw on this recommendation by Clarke et al. (2019).

The third contribution to the existing work is using a combination of hand-crafted and deep feature set to represent each modality within the multimodal model. Prior research has shown that combining hand-crafted and deep representations for both the acoustic and linguistic modality can lead to an increase in accuracy and explainability (Elbanna et al., 2022 ; Johnson and Marcellino, 2022). Most studies using a multimodal model, first explore several types of unimodal models, including both hand-crafted and deep based features (Soleymani et al., 2019 ; Tavabi et al., 2020). However, when fusing the modalities, only the best performing model for each modality is considered in the fusion model. This study explores whether combining the two types of feature sets in a multimodal model can improve the performance of such models.

Finally, en explainable multimodal model has been developed on the hand-crafted acoustic and linguistic feature representations. This enables us to study what features play an important role in determining the likelihood to invest score. To the best of our knowledge, this thesis research is the first to propose an explainable multimodal architecture consisting of acoustic and linguistic features to study investment decision-making based on entrepreneurial pitches.

1.3 Research questions

For this research the following research question and corresponding sub-questions are defined.

Main research question: To what extent can the likelihood of investment be predicted from acoustic and linguistic features recorded during an entrepreneurial pitch using deep and hand-crafted representations?

To answer the overarching research question, multiple unimodal and multimodal models are developed, utilizing hand-crafted and deep features to represent the acoustic and linguistic modalities. These models are evaluated by testing them on unseen data. Differences in the performance on the task to predict the likelihood of investment are compared in order to find the strongest predictors.

Sub-question 1: To what extent can an acoustic or linguistic unimodal model predict the likelihood to invest of entrepreneurial pitches while using either hand-crafted or deep feature representations?

To study the first sub-question, four feature representations are extracted from the pitch recordings. For the acoustic modality, hand-crafted features are extracted using openSMILE and deep embeddings are obtained using VGGish. For the linguistic modality, LIWC is used to create a hand-crafted representation and Longformer for the deep features. A GRU is trained to capture the temporal information in the acoustic features, while the linguistic features are modelled using a linear regression.

Sub-question 2: How does combining hand-crafted and deep feature representations in a unimodal model affect the performance of the model?

For each modality, the hand-crafted and deep representations are combined in a model. One type of combination is created by directly fusing the feature sets and training a single model to predict the likelihood. Another strategy involves fusing the output of the individual models that are developed for the previous sub-question.

Sub-question 3: What is the effect of using different multimodal fusion approaches when using both acoustic and linguistic feature representations in one model?

We evaluate the effect of using a multimodal model by combining the unimodal acoustic and linguistic models. The unimodal models are integrated by both applying early fusion and late fusion. Furthermore, models are considered where all four feature sets are used, as well as models containing only the best performing feature representation per modality.

Sub-question 4: What explainable acoustic and linguistic features play a role when predicting the likelihood of investment of entrepreneurial pitches?

To answer this sub-question, hand-crafted acoustic and linguistic features extracted using openSMILE and LIWC, are used to train a model for predicting the likelihood of investment. Then, Shapley feature importance values are obtained for each fold and compared to explain common important features.

Sub-question 5: How well do the models trained using the in-person recordings of the pitches generalize to online recordings?

For the final experiment, the generalizability of the models developed to answer the previous sub-questions is evaluated in a cross-domain experiment. For the online pitches, hand-crafted and deep features for both modalities are extracted. Then, the best performing models trained on the in-person pitches are tested using the instances of the online data set.

1.4 Thesis outline

The rest of this thesis is organized as follows. First, a literature review of related works is presented. This literature study consists of two parts: a contextual part on entrepreneurial decision making and a technical part discussing papers related to the methodology. The purpose of this literature review is to establish the potential significance of the proposed approach. In the next chapter, the data set is described and an overview of the methodologies used for the experiments is given. Chapter 4 presents the results of these experiments. Finally, in Chapter 5 the results are discussed and placed into context, some limitations are provided and the research questions are answered.

Chapter 2

Literature Study

2.1 Literature Study: Decision-making in entrepreneurial contexts

In a report called *Fostering Entrepreneurship* (1998), the OECD recognizes the central role entrepreneurship plays in driving economies and economic growth. The dynamism of entrepreneurship fosters globalisation, creates jobs and helps our economies to function as open markets. Despite this importance, there is still a lot to be understood about what entrepreneurship actually is and in what kind of processes it thrives. Shane and Venkataraman (2000), who presented a framework and definition of entrepreneurship that is widely supported among scholars in the field, describe it as “a field of business that seeks to understand how opportunities to create something new (e.g., new products or services, new markets, new production processes or raw materials, new ways of organizing existing technologies) arise and are discovered or created by specific persons, who then use various means to exploit or develop them, this producing a wide range of effects”. Following this definition, studying how individuals in entrepreneurial contexts identify and act on opportunities that are presented to them, in other words how decisions are made, is a focus point in recent research on entrepreneurship (Ucbasaran, 2008).

The Oxford Dictionary defines decision-making as “the process of acting upon the best information available in order to determine the most appropriate course of action” (Stevenson, 2010). Decision-making in entrepreneurial settings is characterized by the uncertain conditions in which it takes place, limiting the information available to determine the best course of action. According to Berner at al.(2012), the main premise of entrepreneurship is to invest or produce while accepting higher risks. For this reason, researchers are focusing on how decisions can be made under uncertainty, and decision-making has become a well-established topic (e.g. Wadeson, 2006 ; Shepherd et al., 2015). Understanding how entrepreneurs make decisions, why some are successful and some fail is crucial to the success of entrepreneurial businesses.

To account for uncertainty in decision-making, the naturalistic decision-making framework was introduced. This framework forms a means of studying how people that are experts in their respective fields can make decisions despite the uncertainty (De Winnaar and Scholtz, 2019). This framework assumes that the knowledge structures of the decision makers are formed by their cognitive structures, which in turn are developed through emotion, which means that not all possible variables involved in the decision can be known. The role of emotion makes entrepreneurial decision-making a complex process since entrepreneurs are altered by their social environment and cannot be seen as entire rational beings. Therefore, decisions in entrepreneurial settings that involve social interactions are especially interesting. When the decision involves uncertainty over the facts and consequences, this social interaction itself can play a key role (Liebregts et al., 2020).

An important entrepreneurial decision-making process involving social interactions is the hiring process of new employees. This decision is vital for a firm since human capital is related to the effectiveness of the team and to growth of the firm (Colombo and Grilli, 2005). The application process often involves multiple rounds based on which job-based performance predictions can be made. Someone's cv forms an important information source of past experience but in most cases a job interview is the deciding factor (DeGroot and Gooty, 2009). In an interview, verbal cues play an important role, for example the applicant has to be able to explain their motivation for the job. However, the social interaction does not only exist of the words being interchanged, gestures and appearances also influence the conversation. DeGroot and Gooty (2009) find that visual and vocal cues are related to the performance of applicants. Barrick et al. (2009) find that the image candidates portray in an interview plays a significant role in whether they are hired or not.

2.1.1 Studying social interactions

In social interactions people communicate with another in two main ways. Firstly, directly by speaking with another, this is called verbal communication. Developments in the field of Natural Language Processing (NLP) have enhanced the processing of verbal cues and their influence on human behaviour. In section 2.2.2 several NLP tools are examined that can be used to study the role of language in social interactions. However, when watching television in a foreign country where you do not speak the language, you can still follow the social interactions to some degree. For example, whether people are angry at each other and if the setting is tense or relaxed. This information is communicated using nonverbal cues such as vocal outbursts and facial expressions (Vinciarelli et al., 2009). These nonverbal social signals often trigger analysis of socially relevant information (Argyle and Kendon, 1967).

Traditionally, the effects of social signals have been analyzed using manual techniques to encode them. Maxwell et al. (2011) use independent raters to manually code social signals emitted by pitchers to study early stage business angel decision-making. Jimeneze Munoz (2019) studies success in non-native business pitches using independent ratings of both verbal and non-verbal interactions. The results suggest that both the use of the pitchers voice and gestures influence investment decisions. These manual methods have two main limitations. Firstly, manually encoding variables is time-consuming and inherently arbitrary. Zhang and Cueto (2017) argue that these independent raters are prone to bias. Secondly, the manual annotators may miss important information for decision-making, because it happens too fast or it is a more subtle social signal.

Given the importance of nonverbal cues on our behaviour, a more suitable approach to encode their information is necessary. This problem is addressed by the emerging domain Social Signal Processing (SSP), which aims to make social interactions understandable through analyzing nonverbal behavioural cues using machines (Vinciarelli et al., 2009). SSP focuses on human nonverbal communication and uses modern technologies such as artificial intelligence (AI) to analyze it. Five major classes of nonverbal cues are identified: physical appearance, gestures, face and eye behaviour, vocal or acoustic behaviour and the surrounding environment. In section 2.2.1, the methodologies that can be used for the analysis of acoustic features of vocal behaviour are discussed in detail. The processing of social signals and behaviours consists of two main stages. Firstly, in the pre-processing stage the recordings of social interactions are split into multimodal behavioural streams per person. Secondly, in the social interaction analysis stage the multimodal streams are mapped to social signals and behaviours (Vinciarelli et al., 2009).

2.1.2 Investor decision-making based on entrepreneurial pitches

Both verbal and nonverbal behavioural cues are used to communicate in social interactions and can thus have an influence on the decision-making process. This section reviews what role both verbal and nonverbal cues have in investment decision-making based on entrepreneurial pitches.

There are also decisions in the entrepreneurial context that are not made by the entrepreneurs themselves but by others, such as funding decisions. A vital part of the entrepreneurial process is raising funds from investors to develop your business ideas (Liebregts et al., 2020). Convincing potential investors of the viability of your business proposal is often the result of social interactions through the form of an entrepreneurial pitch, also referred to as an elevator pitch. Like other entrepreneurial decision-making processes, the setting of a pitching entrepreneur and an evaluating investor is characterized by high levels of uncertainty. The entrepreneur has to prove that the business idea is feasible, while not exactly knowing who sits in front of him. On the other hand, the investor has to make an assessment of the potential legitimacy of the project while having limited information (Clarke et al., 2019). Pitching for an investor often takes place at the earliest stages of new ventures, before a product even has been developed or produced and for that reason the risk is uncertain or unknown (Huang and Pearce, 2015).

Newell et al. (1958) coined the idea that in order to deal with high uncertainty, investment decision makers develop heuristics. This form of decision-making can be seen as expert-based intuition, compared to formal analysis when information is available. Huang and Pearce (2015) argue that when evaluating a pitch, investors rely on both expert intuition and formal analysis. The authors name this strategy of using both intuition and formal analysis, two undermining strategies, decision-making based on "gut feeling". Out of a survey among early-stage investors, this gut feeling is described as based on two main components: data on the viability of the project and perceptions of the founding entrepreneur (Huang and Pearce, 2015). As discussed before, viability data is often scarce and for that reason the second component can be more interesting for a study on decision-making. The perceptions of the founding entrepreneur are formed by personal observations of the investor based on social interactions (Huang and Pearce, 2015). As discussed in the previous section, verbal and nonverbal behavioural cues play a vital role in social interactions. Given the reliance of investors on social interactions for their decision-making, it is interesting to analyze the social cues in such an entrepreneurial pitch setting. The role of verbal and nonverbal cues of pitches in making investment decisions has attracted a lot of attention in academic research.

Martens et al. (2007) study the role of storytelling in the ability of a start-up to secure investments. They qualitatively analyse the languages used in pitches in three high-tech industries. According to the authors, the verbal content is important when entrepreneurs want to attract capital for three reasons. Firstly, it can help give investors an insight in the firm's culture and identity, providing more insights than purely factual data on the firms business (sales, revenue etc.). Secondly, using verbal cues can help the business idea to appear both original and distinctive. Thirdly, it gives an entrepreneur the opportunity to elaborate on the background and reasoning behind the business proposal. Combining these three arguments suggests that using the correct verbal cues in a pitch could reduce the uncertainty and risk as perceived by the investors. Furthermore, it can act as a call to action, by motivating and mobilizing investors to commit (Cohen and Dean, 2005). In a qualitative analysis, Martens et al. (2007) reveal that creating narratives that construct identities for firms or that create a contextual embedding have a positive influence on the amount of capital a firm can acquire. They conclude that language usage and storytelling play a key role in entrepreneur-investor interactions.

In an influential study, Chen et al. (2009) investigate the extent to which investors' perception of the passion of an entrepreneur influence investment decisions. Here, passion is defined as an affective state that is accompanied by behavioural manifestations related to the display of emotions and energy. In experiments, they found that this passion has no direct impact on investment decisions, while the substance of the business idea has a positive influence. Since then, research has aimed to replicate these findings, which led to mixed results (Mitteness et al., 2012; Murnieks et al., 2016). These conflicting findings seem to suggest that the approach of Chen et al. (2009) does not capture the entire underlying interlinkages of verbal and non-verbal behaviour in these pitch settings.

In the paper by Chen et al. (2009) non-verbal behaviour is mostly handled as secondary behaviour that can support to convey the passion of an entrepreneur. Clarke et al. (2019) challenge this view and aim to show that non-verbal cues carry meaning themselves, in combination with verbal cues they form an integral part of the interaction process. Clarke et al. (2019) study different combinations of verbal tactics (such as using figurative language) and gestures (e.g. hand gestures) on the probability of investment of entrepreneurial pitches. Their findings suggest that both verbal and the non-verbal gestures significantly influence investment judges. The effect of hand gestures to depict the business proposal had an even stronger positive effect than variations in the type of language used. Of all the non-verbal cues, Clarke et al. (2019) focus on the gestures class, however, they do deem it likely that other non-verbal modalities shape the interactions. They highlight the potential of vocal or paralinguistic acoustic elements such as pitch and loudness in these kinds of analyses.

Research on the role of vocal behaviour in entrepreneurial contexts has mostly focused on the influence on persuasion. Clarke and Healey (2022) argue that voice is an important source of information for investors. To get a better understanding how vocal cues can be used, a model is developed to investigate the relationship between a voice and investor decisions. Their findings suggest that vocal features have the ability to signal important entrepreneurial qualities such as competence and trustworthiness. Another interesting finding is that female entrepreneurs are disadvantaged when trying to persuade a investor given their naturally higher voice pitch. Wang et al. (2021a) study persuasion attempt in crowdfunding projects. The results identify three key vocal indicators that could lead to successful persuasion attempts: focus, low stress and stable emotions. The authors argue that using vocal features and audio mining should play a greater role in academic entrepreneurial research.

The studies by Clarke and Healey (2022) and Wang et al. (2021a) prove that vocal behaviour can play a significant role in the decision-making process. However, both these papers look at vocal-behaviour in isolation and not in combination with verbal behaviour. The approach in Clarke et al. (2019) suggests that looking at the combination of verbal and non-verbal is beneficial. Furthermore, Clarke and Healey (2022) and Wang et al. (2021a) mostly focus on the influence on the perceived persuasiveness of the entrepreneur and not directly on the investment decision. In some cases, it might be possible that an entrepreneur that comes across as too persuasive and thus loses their credibility. Therefore, it is interesting to analyze the direct effect of vocal behaviour on investment decisions.

Acquiring capital in the form of investments is a critical part of the entrepreneurial process for any business in the early stages of their venture. Entrepreneurial pitches form a medium where entrepreneurs and investors interact. Previous work on the assessment of entrepreneurial pitches shows that both verbal and nonverbal behavioural cues can affect the final investment decision. Processing the social signals emitted by both the pitchers and the investors opens up a valuable source of information for research into entrepreneurial decision-making. To unlock the potential of these signals, NLP can be used to analyze the verbal cues, while the SSP domain provides the tools to automatically code nonverbal signals resulting in a more accurate and efficient analysis. Since social interactions are the interplay of verbal and non-verbal cues, integrating them in a single analysis is a current gap in the existing literature and could provide interesting new insights. The class of vocal behaviour has been identified as a potential important driver in pitch settings (Huang and Pearce, 2015 ; Clarke et al. 2019), but has not been studied in a combined model with verbal cues yet.

2.2 Literature Study: Methodology

In this section, a review of related works employing a methodological approach similar to the one that has been used in this research is presented. Firstly, an overview of papers utilizing either acoustic or linguistic features is provided. Then, papers exploring the potential added value of combining these two modalities in a multimodal model are introduced. Emphasis is placed on papers that apply these methodologies in an entrepreneurial context.

2.2.1 Acoustic modality

Acoustic features of speech form the primary mode of interactions between humans. These type of acoustic or vocal features are used on a wide scale to analyse human behaviour, including decision-making. Extracting features from the actual audio signal and creating feature representations is a key part in the analysis of vocal behaviour. Generally speaking there are two different approaches to feature engineering of the acoustics of speech:

(i) Feature representations that are handcrafted using domain knowledge

(ii) Feature representations that are learnt by deep learning algorithms

In the following sections, papers utilizing either one or a combination of these approaches for the representation of the acoustic modality are be discussed.

Hand-crafted acoustic feature representations

Traditionally, feature representations of speech are generated by a hand-crafted process which requires domain knowledge. These type of feature sets are interpretable by humans. Examples of features that can be extracted by hand are pitch, which makes it possible to judge sounds as either high or low, and loudness. There is a wide range of features that are based on domain knowledge and fall under the hand-crafted category, which can be further categorized into three categories: prosodic, voice quality and spectral (Shah, 2022).

Prosodic features represent the melodic contour of the audio signals of speech and give an indication of the intonation, examples include: pitch, loudness and timing. These relatively simple features can be used to gain interesting insights on the effect of the acoustic features of human speech. Carlson (2017) used a small selection of highly interpretable prosodic features, such as loudness and speaking rate, to analyse entrepreneurial pitches from a startup competition. The aim of the study was to examine whether these vocal features capture perceived traits of entrepreneurs. Regressing perceived traits such as confidence and likeability on the prosodic features, the author finds that speech has a significant effect. The findings suggest that especially loudness plays an important role in how a speaker is perceived, as those who are louder on average are perceived to be more confident and more likable. Furthermore, the study looked at whether the effect of loudness could also be extended to funding outcomes of the entrepreneurial pitch. The findings suggest that the variance in loudness has a strong positive relationship with funding raised (Carlson, 2017).

Another application for which prosodic features are used is the detection of emotion in a speaker. Luengo et al. (2005) used an emotional speech database which included six basic emotions, namely anger, fear, surprise, disgust, joy and sadness. For each utterence, the authors extracted 12 "curves" such as pitch and power and for each of them different statiscal features were computed, resulting in 86 prosodic features. A support vector machine (SVM) was trained on the data set using this entire feature set, resulting in an accuracy of 93.5%. When only using the six best features, which include mean pitch and pitch variance, an accuracy of 92.3% is obtained. A similar approach is used by Rao et al. (2013) who specifically compare the effect of using prosodic features over segments of sentences instead of the entire sentence. Their results suggest that using local prosodic features outperforms global prosodic features. Finally, prosodic features can also be used to detect medical conditions such as dementia (Ossewaarde et al., 2019, Haulcy and James, 2021) and depression (Yang et al., 2012).

The second category of hand-crafted acoustic feature representations is voice quality. Voice quality can be defined as the characteristic auditory colouring of a person's voice (Keller, 2004). Although prosodic and voice quality features interact closely, scholars argue it is beneficial to distinguish them for explanatory and algorithmic purposes (Ibid, 2004). Examples of features that fall under this category are divergence from spectral distributions and jitter, a measure of the periodicity of the voice signal. Studies have shown that using voice quality feature representations, predictive models can be created. For example, Szekely et al. (2012) created a model using voice quality differences to detect specific regions where the speaker swaps its normal voice for a different one in audio books. Although voice quality can be a useful feature representation by itself, combining it with prosodic feature sets yields a substantial improvement in performance, compared to using a single feature set. With a combination of these two feature sets and a SVM classifier, Zhang (2008) achieved a performance of 76% in an emotion recognition task, a 10% improvement compared to single features. Similarly, Lugger and Yang (2007) found that the parameters of voice quality are a contribution in addition to prosodic features.

The third and final category of hand-crafted acoustic features describes the spectral attributes. These spectral features are treated as strong correlations of the different shapes the vocal tract makes when a person is speaking and the changes in articulator movement (Koolagudi and Rao, 2012). MFCCs (Mel frequency cepstral coefficients) and LPCCs (Linear prediction cepstral coefficients) are often used as feature sets for the spectral category. These are created by first by applying a Fourier transform on the audio signal, to transform a time-domain signal into a frequency domain signal. Spectral feature representations have been used for several tasks. Niebuhr et al. (2018) showed evidence that there is a correlation between spectral features and listener ratings of how charismatic a speaker is. Again, in many cases a combination of spectral features with one of the other types yields an increase in performance of the model, for example when creating a language identification model (Yin et al., 2006).

As discussed in the previous sections, these types of hand-crafted feature representations are in many cases more useful when combined with another. This has led to standardized feature vectors which are released as sets of feature representations. OpenSMILE is a feature extraction toolkit which extracts different types of feature representations from audio signals and creates vectors (Eyben, Wollmer, and Schuller, 2010; Eyben et al., 2013). Using openSMILE, Low Level Descriptors (LDDs) can be extracted and in addition functional statistics (such as extremes, mean etc.) of these descriptors are determined. Currently, openSMILE supports three standard feature sets: ComParE, GeMAPS and eGeMAps. The ComParE set is the baseline feature set of the INTERSPEECH Computational Paralinguistics Challenge and consists of 6373 features (Schuller et al., 2016). The openSMILE toolkit has been used in a wide range of studies.

Training (deep) models using hand-crafted acoustic features

As in this study we look at the role of social signals in entrepreneurial decision-making it is especially interesting to look at papers part of the 2013 INTERSPEECH Computational Paralinguistics Challenge. During this year, the 2012 version of openSMILE's ComParE was slightly modified in order to optimize modelling social signals and emotion (Schuller et al., 2016). As part of this challenge, Wagner et al. (2013) used phonetic patterns to detect social cues, such as laughter and fillers like "uhm", in natural conversations. Extracting features using openSMILE and training a linear kernel SVM on these features resulted in an accuracy of 83.3%.

Marchi et al. (2016) used the ComPare 2013 package from openSMILE as a baseline feature set for a model to the track the emotions and character traits of speakers on mobile platforms. Their motivation to use this feature set was twofold. Firstly, openSMILE is freely available and a well-defined standard for audio tasks. Secondly, the size of several thousands features gives a substantial amount of information for a model to be trained on. The authors extract acoustic features for speech segments of up to 120 seconds and then classify speaker's characteristics and emotional state with the help of

Support Vector Machines (SVMs) and Support Vector Regression (SVR). The openSMILE toolkit is also used in a study Galanis and Esposito (2013) that looks at detecting emotional traits in call centre interactions. By extracting features compromised of functionals of LLDs using openSMILE an SVM can be trained. The SVM learns an optimal hyperplane that separates the emotional and non-emotional utterances. This hyperplane can be used to predict the category of a new utterance. The best performing model can classify emotional speech with an accuracy of 82.11%.

The studies using openSMILE for feature extraction that have been discussed so far mostly use a SVM for classification. The SVM was a successful learning model for a number of tasks and was considered state-of-the-art. Recently, as deep learning techniques have been widely used on various tasks, researchers have also started to use them for audio classification tasks (Bae et al., 2016). Deep neural networks (DNN) are powerful learners that can learn complex non-linear relationships between the input and targeted output. A limitation of DNNs is they can only map a present input vector directly to an output, in other words the model cannot remember states in the past. Recurrent neural networks (RNNs) overcome this shortcoming. Using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, two types of RNNs, the model can learn the temporal information of consecutive input vectors. Given the importance of spatiotemporal dynamics in speech, these types of models are very practical to analyze vocal behaviour. Therefore, it is interesting to feed the feature extractions sets of openSMILE in RNN classification models (Bae et al., 2016).

Wollmer et al. (2012) propose a model for emotion recognition based on a LSTM. Previous studies on emotion recognition found that modeling long-range emotions tends to increase the performance of models for emotion recognition. These LSTM networks are able to incorporate information on how emotions develop over time and thus could increase the accuracy of this task. In this study, the audio feature extraction is based on openSMILE. Using the toolkit, a large set of LLDs and fuctionals are extracted. An LSTM is trained on this feature set to make classifications on an emotion recognition task, to discriminate between high and low levels of arousal, expectation and power. The LSTM trained on the openSMILE features is the best performing model for this task, outperforming LSTM models based on other (such as visual) feature sets. The 65% accuracy on this task using vocal features and a LSTM was the highest accuracy reported in literature when the paper was published in 2012.

These papers that successfully use feature sets extracted using openSMILE for speech based tasks show the potential power of speech as a social signal. In the chapter on entrepreneurial decision-making we have seen that perceived character traits and emotions of a pitcher in an entrepreneurial setting has an effect on how much funding is raised. Here we saw that openSMILE feature sets capture the emotion of a speaker. Therefore, it seems likely that it is interesting to use hand-crafted acoustic feature representations of speech to analyze entrepreneurial pitches.

Deep learning derived acoustic feature representations

The previous section discussed the traditional method to create feature representations of the audio signal of speech, which relies on designing hand-engineered features. After the feature engineering step, a machine learning model (such as a SVM or RNN) is trained on these features to make predictions. The performance of these models is very dependent on the quality of the data representations and for this reason feature engineering has been a vital part of the process. However, feature engineering techniques are often time-consuming and the hand-crafted features can potentially be a not so accurate representation of the actual speech (Latif et al., 2020). Recently, representation learning techniques have been started to become popular. With these techniques, machines can learn a representation of the speech input directly and automatically. These types of representations are very useful and conducive for making predictions or classifications. Data representation tools can be divided over two categories: shallow learning models and deep learning based models. In this thesis the main focus is on deep learning based models. Deep learning models, such as a CNN, can be trained to learn what the important features are of an audio signal and thus be used as a feature extraction tool (Latif et al., 2020).

Jaitly and Hinton (2011) were one of the firsts to try find a “better” way to represent speech sound waves. They argue that using low dimensional hand-crafted representations may lead to losing valuable information which makes it difficult to use it for discrimination. Jaitly and Hinton present a novel approach consisting of higher dimensional encodings with the help of a Restricted Boltzmann machine (RBM). RBMs are a type of graphical model that uses hidden variables to achieve expressive marginal distributions. Experiments using these RBMs to detect features showed that they can be used in emotion recognition task and outperform the state-of-the-art models at the time, that were based on hand-crafted features. Milde and Biemann (2015) used CNNs to create a feature extraction model for a paralinguistic task of eating condition classification. Their main idea is to train local CNN classifiers that learn feature representations on a small section of the full spectrum. The local classifications of the small segments are then aggregated to create a global classification for the entire sequence. The final model achieved a relative improvement of 15% over the baseline hand-crafted model.

Trigeorgis et al. (2016) went a step further than Jaitly and Hinton (2011) and applied an end-to-end model for an emotion recognition task. An end-to-end model refers to learning the feature representation and classifying the emotional state of the speaker in one jointly trained model. The model they use which has the raw waveform signal as input and an emotion prediction as output is illustrated in Figure 2.1. As input, the raw wave is divided into 6 seconds long segments. Subsequently, the input is fed to a convolutional layer to extract spectral information, followed by temporal pooling. Then there is another convolutional layer in order to extract more long term characteristics of speech. The convolutional layers "replace" the traditional need for hand-crafted features and the output of these layers is fed into a recurrent network. The use of this proposed model that uses the raw signal significantly outperforms models based on traditional hand-crafted feature sets such as ComParE and eGeMAPS. Furthermore, the authors studied the gate activations of the recurrent layers of the LSTM models. They find that these cells are highly correlated with prosodic features that were assumed to convey cues in speech, which suggests interpretable cells exist.

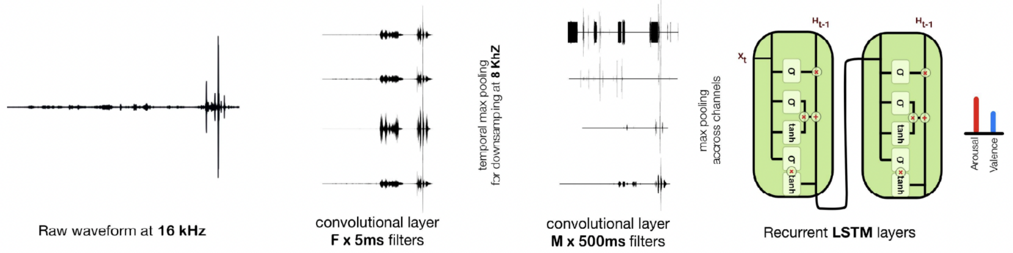

Figure 2.1: The convolutional layers replace hand-crafted features. Image by Trigeorgis et al. (2016)

Hersey et al. (2017) study various CNN architectures for large scale audio classification. Their strategy is based on the notion that an audio spectrum can be regarded as some type of image. By modifying CNN networks that are effective for image classification tasks, a CNN architecture that can perform audio classification tasks is created. In such as case, the pre-training is processed with a different task at hand and we call this kind of learning transfer learning. VGGish is a type of deep audio embedding method that works in such a manner (Hersey et al. (2017)). VGGish is based on the VGG model, which is designed for image classification. VGGish is trained to predict video tags from the Youtube-8M data set (Hersey et al., 2017). The VGGish model is composed of a series of convolutional, ReLU activation and max pooling layers, followed by three fully connected layers. Using the VGGish feature extractor a 128 dimensional audio embedding is created for every 960 ms of the audio fragment. Similarly as was the case for the hand-crafted feature vectors, a RNN can be trained on this feature set as a sequence encoder which can incorporate the temporal dynamics over the entire audio signal.

Shi et al. (2021) leveraged the VGGish architecture to extract audio features for crowdfunding campaign video's. For every 960 ms segment, a 1280-dimensional vector is outputted by the final layer. This feature vector can then be used to train a model. In this case, Shi et al. (2021) use a deep neural network model as a prediction model, the output layer being a sigmoid activation function. The model using VGGish features was the best performing model on the task of predicting crowdfunding success, outperforming models based on hand-crafted feature sets such as MFCC.

Sun et al. (2020) extract both hand-crafted and deep vocal features for an emotion recognition task. The eGeMAPS consisting of LLDs and functionals is extracted using the openSMILE toolkit. The window size is set to 5 seconds, meaning that a feature vector is created for frames of this size. Using VGGish, 128-dimensional embeddings are extracted for every 0.96 seconds of the audio signal. Then, a LSTM model is utilized in order to model the complex temporal dynamics of the audio sequences. The LSTM is trained on the feature sets and creates context-dependent hidden states. Finally, a regression layer is used to output the final emotion prediction. The VGGish based model outperforms the eGeMAPS model significantly on the arousal prediction task (0.49 versus 0.39).

Combination of acoustic hand-crafted and deep representations

The previous section highlighted the superiority of deep representations over hand-crafted features when the goal is to analyze or make predictions on acoustic features of speech. However, this does not necessarily mean that hand-crafted feature representations have to be discarded all together. In this section, we examine whether it could be beneficial to combine both deep and hand-crafted speech feature representations in a model.

Elbanna et al. (2022) studied the effect of stress on speech and aimed to create a model to detect stress in speech. Earlier works using feature representations based on deep neural networks did not succeed in accurately recognizing stress load in speech. The approach of Elbanna et al. (2022) consists of leveraging the power of the deep feature representations and the hand-crafted acoustic feature sets. Instead of training two separate models, one on deep and one on handcrafted features, and combining the outcomes of these models, the two models are incorporated into the pre-training phase. By doing so, the authors create a unique representation on which a classifier can be trained. In this study, a handcrafted feature set consisting out of 88 features is extracted using openSMILE. Deep learning representations are extracted using a variety of models, including VGGish and YAMNet (Plakal and Ellis, 2020). Comparing the performance of several models on a stress recognition task, the results suggest that a hybrid representation of speech outperforms models that solely use either handcrafted or deep feature sets. Although deep learning based models are effective for the task at hand by themselves, including hand-crafted acoustic features yields a more accurate model. The author suggests that future research should use a similar approach.

Another audio task for which feature extraction plays a important role in the performance of the model is urban sound classification. It can be valuable to classify sounds in city for numerous reasons including reducing noise pollution and improving security (Luz et al., 2021). Traditionally, urban sound classification models relied on handcrafted acoustic features but in the last years deep learning models have also successfully been used. A few studies have experimented using a combination of both deep and hand-crafted feature representations. Giannakopoulos, Spyrou and Perantonis (2019) propose a method that combines hand-crafted audio representations with a CNN based representation. Extensive experimentation on the combination of these two feature representation types showed that the combination leads to a relative increase in performance of 11%. This result suggests that the CNN model and hand-crafted method complement each other. Luz et al. (2021) use a similar strategy to classify urban sounds. By associating hand-crafted features to deep features, the performance is increased and it allows for dimension reduction of up to 62% for the combined descriptors. The combined model outperforms most of the state-of-the-art CNN models for urban sound classification.

Generally speaking, theories on ensembling techniques suggest that using a diverse set of features plays an important role in the performance of a classifier (Kuncheva et al., 2003). This theory in addition to the works discussed in this subsection show that it could be beneficial to combine hand-crafted and deep feature set in a single mode.

2.2.2 Linguistic modality

The analysis of linguistic data is commonly referred to as Natural Language Processing (NLP). The field of NLP involves developing models that allow computers to understand and process information in a natural language format. The development of AI has resulted in major breakthroughs in the development of language models. Approaches using language as input to make classifications or predictions can, generally speaking, be deconstructed into four phases: feature extraction, dimension reductions, classifier selection and evaluation (Kowsari et al., 2019). As was the case for the linguistic modality, feature extraction plays a key role when we want to analyze text. Generally speaking there are two different approaches to feature engineering of text:

(i) Feature representations that are handcrafted using domain knowledge

(ii) Feature representations that are learnt by deep learning algorithms

As text data often consists of a large number of unique words, the feature representations for a document can get very large. Large feature representation sets can cause high time and memory complexity when training a model. A solution for this problem is dimension reduction to reduce the size of the feature representations. A commonly used method for dimension reduction is Principal Component Analysis (PCA).

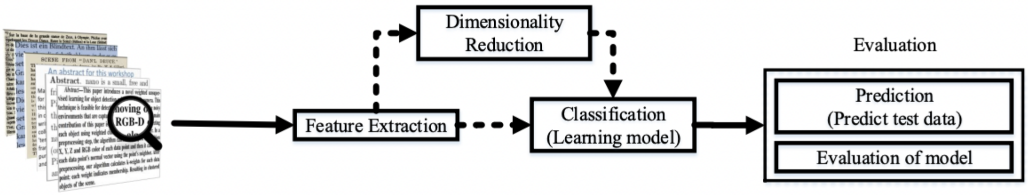

Figure 2.2: Using text to make predictions. Image by Kowsari et al., (2019)

After extracting features and reducing the dimensionality of these features, a classifier can be trained on the text data. Different types of machine learning algorithms, ranging from traditional to deep learning models, can be trained on text data, all having different advantages and disadvantages. Traditionally, simple classification algorithms such as logistic regression and Naive Bayes were very popular and have achieved good results. Currently, SVMs are used on a wide scale as baseline models to compare to more advanced models. Deep learning models outperform most of these traditional methods given their capacity to model complex and non-linear relationships in the data (Kowsari et al., 2019). The final step of the entire process is evaluating the predictions of the text model. There are many methods available to evaluate these techniques.

In the following sections, papers that employ linguistic features of text as input for machine learning models are discussed, with a particular focus on the feature extraction step. The two different types of feature representations and their combination are explored. The approaches for the other three phases are also examined in the discussion of works employing one of these methods.

Hand-crafted linguistic feature representations

Raw language data often contains text in the form of short or long documents. A documents consists of a number of sentences, where each sentence includes words that are made out of letters. If we want to train a model on this kind of data, first a structured feature set has to be extracted. The simplest form of feature representation are hand-crafted and belong to the "n-gram" category. The n-gram strategy involves a set of n-words that occur in that specific order in a text. Such a n-gram is not a representation of the text, but can be a feature to represent it (Kowsari et al., 2019).

The simplest form of n-gram technique is where n = 1, this form is also called a Bag-of-Words model (BoW). A BoW model involves two things, a vocabulary of known words and a measure to quantify the presence of these words. In this model, a vector is created in which each feature corresponds to a unique word in the text. In the simplest form, called term frequency, the value for this feature is the number of times this word is present in a text. Dirisam, Bein and Verma (2021) use a BoW model to predict whether a crowdfunding action will attract donors or not. The text of crowdfunding pages is converted into a vector using BoW. Training a Naive Bayes classifier on these feature representations resulted in an AUC score of 0.7, outperforming other types of feature sets. The authors attribute this to the fact that BoW is simple and the conditional probabilities are not overfitted. Venkata Raju and Sridhar (2020) use a BoW model to predict the score of a review based on the text of that review. The authors use a scale of one to five on a data set consisting out of hotel reviews. Using this BoW approach, 60% of the ratings can be accurately predicted with the given review. A limitation of the BoW models is that all that matters is what and how often words are used, the ordering of the words is irrelevant. Therefore, it is also common to use 2-grams and 3-grams, since here a model can detect more information and the order of words is included in the feature set (Kowsari et al., 2019).

Another limitation of BoW is that it treats every word equally. This way implicitly common words in a corpus can have a big effect on the feature representation. To lessen the effect of these common words, Sparck Jones (1972) developed the Term Frequency - Inverse Document Frequency (TF-IDF) model. This model combines term frequency with inverse document frequency, that assigns a higher weight to either words with a high or low frequency in the document. Ramos (2003) used a TF-IDF model to determine word relevance in document queries. In this experiment, the goal is to find a document relevant to a search query. The TF-IDF vector of the search query is matched with the TF-IDF vectors of the documents. The sum of the vectors is maximized, resulting in a list of documents that should match the search query. The results of Ramos (2003) show that TF-IDF is an efficient model for this task and it returns highly relevant documents. However, a limitation of the model is that it makes no relationships between words, for example when a query includes the word “priest”, documents that use the word “reverend” will not be returned.

Another BoW based feature extraction model is Linguistic Inquiry and Word Count (LIWC) which relies on pre-computed dictionaries (Pennebaker et al., 2015). LIWC is a lexical software tool that creates feature representations by matching words in a document to terms in its dictionary and creates scores along multiple dimensions. For example, LIWC scores for linguistic variables such as number of pronouns and conjunctions but also more affective aspects such as positive/negative emotion (Kahn et al., 2007). The terms in each dictionary are selected by experts and validated in different types of settings. The English LIWC Dictionary contains around 6400 words and word stems. For every term in the dictionary, one or multiple category labels are listed. These categories can be classified into five sets: linguistic processes, psychological processes, personal concerns, spoken categories and punctuation. Examples of categories in the linguistic processes category are: word count, total pronouns, personal pronouns. For a word in the input text, scores for each category are added and by doing so a feature set is created for text sequences or entire documents. For example, words in the negative emotion category are: abuse, sorry, tears. For a specific text, the LIWC analysis measures the appearance of all negative emotion words and creates a numeric value in the range 0 to 1 for the negative emotion category. (Onan, 2018).

LIWC has been used to create feature sets and analyze texts in entrepreneurial settings and achieved high model accuracies. Balachandra et al. (2021) used LIWC features to study the influence of gendered language in entrepreneurial pitching. Previous studies showed that women consistently raise less capital than men. This study explored whether gendered language might influence investment decisions. Using LIWC, variables for masculine discursive style measures (work and complexity) and feminine discursive style measures (emotions and affect) were extracted from entrepreneurial pitches presented by both males and females. The results of the analysis show that female entrepreneurs do not use a more feminine linguistic style in their pitches compared to men. However, the results do suggest that a masculine linguistic style is more effective in the pitch setting.

Zhang et al. (2021a) study what contributes to a successful crowdfunding campaign. Amongst other attributes of crowdfunding campaigns, the role of the text in the description is analyzed. For every pitch, a feature vector consisting of 92 features is extracted using LIWC. The results show that text features can contribute to 59% variance in the success of the campaign. Similarly, Babayoff and Shehory (2022) use a data set of 50.000 crowdfunding campaigns from the website *Kickstarter* and extract 120 semantic features using LIWC. A prediction model based on semantics only, can predict whether the campaign is successful or not with an accuracy of 91%. Looking more specifically at the categories of LIWC, the buzzwords and emotional features are highly correlated to funding outcome.

These different kinds of hand-crafted BoW based models are chosen as feature extraction technique for machine learning due to their simplicity and robustness (Mikolov et al., 2013). Furthermore, it is more effective to train a simple model on a large amount of data instead of training a complex model on limited data. The discussed papers indicate that feature sets extracted using LIWC can be applied successfully to train a model to make predictions in an entrepreneurial setting. However, these models do not include any semantic similarity of the words in the feature representations. As Mikolov et al., (2013) argue, in many cases a meaning of a phrase is not simple the composition of the meaning of the individual words that make up the phrase. Given the fact that the order of words is not respected and the lack of semantic information in these feature sets, a limited set of tasks can be achieved using this method.

Deep learning based linguistc feature representations

Given the shortcomings of hand-crafted BoW based models, many researchers aimed to develop a word embedding model to represent text data. Similarly as for the speech modality, models can be trained to learn feature extraction of a text. Word embeddings are feature learning models in which each word or phrase is mapped to a dimension vector of real numbers (Kowsari et al., 2019).

Mikolov et al. (2013) presented the “word to vector” (Word2Vec) representation as a word embedding model. Word2Vec is a deep learning based model which can be used to generate continuous dense vector representations of words. These embeddings are special since they capture semantic similarity. Simply said, this is a unsupervised model which can take in large amounts of text, create a vocabulary of all words and creates embeddings for every word in the vector space representing that vocabulary. Since the size of the vectors and the number of vectors can be specified, the dimensionality of these models can be lower than the BoW based models. Two model architectures exist that can be leveraged by Word2Vec to create the embeddings: the Skip-gram and the Continuous Bag of Words models. Word2Vec can be seen as both supervised and unsupervised. Using the Skip-gram and the Continuous Bag of Words models the model can derive a supervised learning task from the corpus itself. While it is unsupervised in the sense that it can learn embeddings for any corpus of your choice, without labelling.

Lilleberg et al., (2015) use Word2Vec in combination with Support Vector Machines for text classification. The main issue the authors identify in this approach is the fact that Word2Vec treats every word the same and is thus unable to distinguish between the importance of each word with respect to the classification. Therefore, they propose an approach which combines Word2Vec and TF-IDF. Since Word2Vec can only create feature vectors for words or short phrases, the representation for an entire document was created using the weighted sum of the word vectors. The result shows that the combination of the two can outperform the model using individual feature sets.

The main challenge that Lilleberg et al., (2015) ran into while using Word2Vec is the fact that it can only generate embeddings for words and not for entire sentences, which makes it difficult to incorporate the context in the embedding. This problem is solved by the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). The BERT model is based on the transformer architecture and currently is the most commonly used pre-trained model for feature extraction on text data. BERT has achieved state-of-the-art results in several NLP tasks and is very effective in representing a whole sequence of terms at once as a fixed-length vector. The BERT model is pre-trained in such a way that it is forced to learn the context and semantic information within and between sentences (Devlin et al., 2018). This is done by training the model with two learning objectives in mind: masked language modelling (MLM) and next sentence prediction (NSP). The MLM task drives the BERT model to to embed each word based on each surroundings. The NSP tasks forces the model to learn continuous semantics over sentences.

Chan et al. (2021) explore how BERT can be used to predict crowdfunding outcomes. Using word embeddings created by BERT, they analyze the writing quality of the description of crowdfunding projects. The BERT masking strategy is used to predict the probability of that word appearing in that position, given the rest of the sentence. A lower BERT score then indicates better predictability for the tokens in the sentence. Therefore, in this case a lower BERT score reflects a better use of grammar and more fluent sentences, representing better writing quality. The results of the study show that descriptions with a higher average BERT score (lower writing quality) tend to raise more donations. Zhou (2021) proposes a BERT + CNN model to classify the titles of crowdfunding campaigns. This model is used to achieve semantic information and spatial location information at the same time. The combined BERT and CNN model outperforms the individual models and traditional methods such as SVMs and Naive Bayes.

Transformer based models such as BERT are partly so successful because of their attention mechanism which enables them to capture the context of a sequence. However, a limitation of this attention mechanism is that the memory and computational requirements grow quadratically with sequence length (Beltagy et al., 2020). Therefore it is infeasible to use BERT for long sequences or entire documents. To address this limitation, Beltagy et al. (2020) propose the Longformer model, which is a transformer based model comparable to BERT with an attention mechanism that scales linearly. Longformer can be used to extract features for entire documents.

Chen and Chu (2021) propose a method to detect fake news regarding the COVID-19 vaccine in long documents. The authors use Longformer to extract a 768-dimensional vector for every document in the data set, to capture the long-range dynamics of the document. A matrix consisting of all the feature sets of each document is created and fed into a Multi Head Attention module. Finally, a softmax layer is used to get the final classification. Wang et al. (2021b) use a bimodal model consisting of speech and text for an emotion recognition task. The Longformer network is used to extract features for the text modality. An accuracy of around 70% is achieved using this method.

Combination of linguistic hand-crafted and deep representations

As discussed in the previous section, transformer based models significantly outperform BoW models that rely on pre-computed dictionaries. The same argument, based on general ensemble theory, as for the different acoustic representations applies here: since using diverse features can improve performance, the fact that one model is outperformed by another does not mean it has to be discarded. Again, we examine how it can be beneficial to combine hand-crafted and deep features, this time for linguistic features specifically. Downsides of deep transformer models include intensive computing requirements and a lack of explainability. Since BoW models are simpler models and require less computing power, some scholars tried to use these BoW models in conjunction with deep models.

Johnson and Marcellino (2022) argue that when simple BoW models are used as a supplement for deep transformer models, this can lead to both an increase in performance and improved explainability of the results. To demonstrate the validity of this argument, they perform two different experiments. Firstly, an experiment is described where implementing a BoW model drastically improves classification performance compared to only a transformer model. The authors extract features using a BERT-based model and a dictionary based model comparable to LIWC, called a stance vector. A logistic classifier is trained on either the BERT vector, the stance vector or a concatenation of these two. The two embeddings are scaled differently, however, this is not an issue since a logistic regression model fits each parameter separately. The three different implementations are trained for binary classification on three different text databases. In all three cases, the "hybrid" model where the two feature sets are concatenated performs the best. Transformer models are known to have difficulty with longer pieces of text, this also appears in the results of this experiment. The highest increase in performance over the single models of the hybrid models is on the data set consisting of long pieces of text. This finding suggest that hybrid approach are the most effective on longer text sequences. Secondly, Johnson and Marcellino (2022) show that a hybrid approach also provides insights in the classification process and thus increases explainability.

Younus and Qureshi (2020) propose a similar method as Johnson and Marcellino (2022) to combine BERT with LIWC features. Again, a logistic regression model is fed both types of features. In this case the task involves detecting propaganda in news outlets. The results show that combining the two feature sets significantly improves the performance of this task. The authors conclude that the contextual linguistics and contextual semantics play a role in text classification. El Mekki et al. (2020) use an ensemble model that applies a weighted voting techniques on one classifier based on hand-crafted N-grams and another on BERT. The ensembling model outperforms the models based on either BERT or N-grams.

The findings of these papers demonstrate the potential value of combining deep and hand-crafted feature sets for text data. Especially for longer text sequences and in computer-constrained settings a hybrid approach can be beneficial.

2.2.3 Multimodal analysis

As human beings, we often do not rely on one unimodal feature in our communication but on a combination of several unimodal features together. Using such multimodal information we get a better understanding of a speaker's intention. For example, when hearing someone's voice we are able to detect sarcasm, while this is difficult when just reading a text. The ability of multimodal systems to outperform a unimodal one is well established in literature (D'mello et al., 2015).

Multimodal fusion

So far, unimodal models consisting of either speech or text have been discussed. In this section, we look at how these unimodals can form the building blocks for a multimodal system. This step forms one of the important challenges when designing a multimodal model (Baltruvsaitis et al., 2018). The process of combining information from the different unimodal models into a multimodal one is called fusion. Different fusion strategies and techniques exist. Generally speaking, two types of fusion techniques exist: early fusion at the feature level and late fusion at the decision level. Furthermore, these two different strategies are also employed together by some researchers, as part of a hybrid approach (Poria et al., 2017).

Early or feature-level fusion refers to concatenating the feature vectors from the different modalities into a single vector. The concatenation forms a single new feature representation on which a model can be trained for classification or regression. The main benefit of fusing the feature vectors of the unimodal models is that the correlation between the different features can help when training the model to improve the performance of the task at hand. A challenge when using early fusion is the synchronization of the different modalities. If you want to concatenate the different features into a single vector, they must be synchronized and in the same format (Poria et al., 2017).

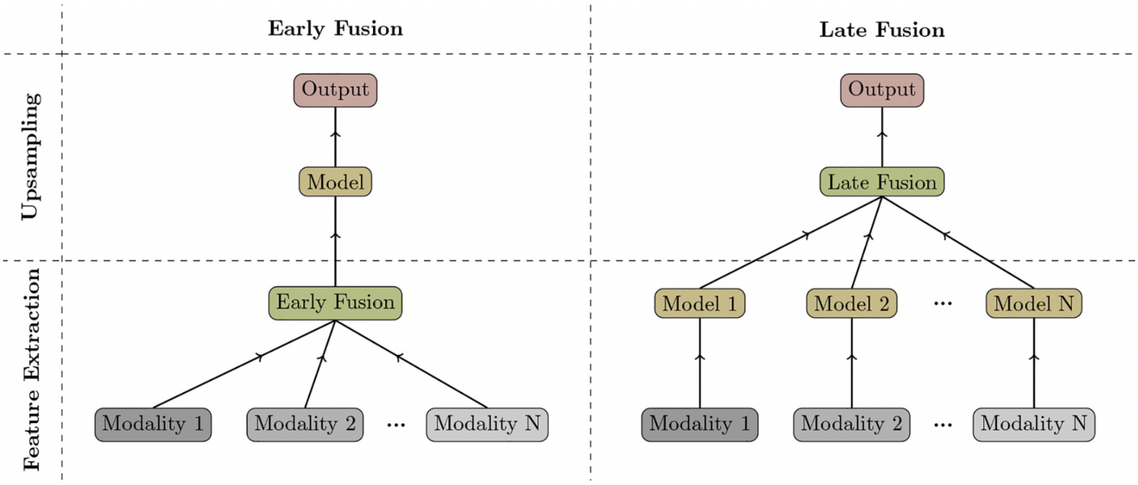

Figure 2.3: Early fusion versus late fusion. Image by Zhang et al. (2021b)

When using late fusion, first several models are trained on the feature representations of each of the individual modalities, then the results of these models are fused to form a single final decision. Aggregating the decision of the unimodal models can be done by voting, averaging, weighted sum or a trainable model (Gkoumas et al., 2021). One of the benefits of late fusion is the complement of the weakness of early fusion, it is relatively simple to fuse the outcomes of the unimodal model since these are often a similar form of data. Another advantage of late fusion is that for every modality, experiments can be done on what the most suitable classifier is and this one can be selected for the final model. A downside of training multiple models and classifiers is that it is a tedious job that consumes a substantial amount of time. A survey of different fusion methods in academic studies by Poria et al. (2017) reveals that recently researchers seem to prefer late fusion over early fusion.

Morvant et al. (2014) apply a late fusion approach based on majority voting of diverse classifiers. Over a set of classifiers, seen as voters, the lowest misclassification rate is estimated while maximizing the diversity of the voters. By maximizing the diversity of these voters, they ensure that information of different modalities are included in the final model. The proposed approach scores a good performance on the challenging PASCAL VOC'07 benchmark. Another notable late fusion method is based on the Kalman filter, a linear system based on a Markov model (Glodek et al., 2013). Dobrivsek et al. (2013) propose a relatively simple fusion model by employing weight sum and weighted product rule, where the weighted product rule outperformed the sum approach on the eNTERFACE data set. Finally, it is also possible to train a deep model to fuse the predictions of multiple classifiers. Nojavanasghari et al. (2016) train a deep model on the final prediction score of each unimodal classifier and the complementary scores (to infer the absence of the class of interest). The late fusion model based on the deep model outperformed a early fusion and a late fusion with averaging model.

In a hybrid fusion model both early and late fusion are applied. The motivation of this approach is exploiting the advantages of each strategy and overcoming the disadvantages. Different types of hybrid fusion models exist. Wollmer et al. (2013) propose a hybrid model for audio, video and text modalities. When using these three modalities the main challenge is the alignment of the audio and video versus the text. In the hybrid fusion approach of Wollmer et al. (2013) the audio and video feature representations are concatenated (early fusion) and a prediction model on this combined feature set is trained, while the textual features are fed into a separate prediction model. The decisions of the two prediction models are then fused (late fusion), resulting in a single prediction for the three modalities together.

The added value of using multiple modalities

After describing how unimodal models can be combined into a multimodal one, papers that develop multimodal models are examined in more detail. The entire methodology of feature extraction, training, and fusion are examined in order to assess the added value of utilizing multiple modalities, particularly in an entrepreneurial setting. Additionally, the evaluation of multimodal systems is discussed.

Jiménez Muñoz (2019) uses a very simple model only consisting out of handcrafted features to analyze the impact of verbal and non-verbal elements on the success rate of business pitches. The study provides statistical evidence on the impact of paralinguistic cues on the probability of success of pitches. In some cases these cues have a higher effect on the success rate of business pitches than verbal aspects. Based on the analysis of the investor reports Jiménez Muñoz (2019) concludes that both verbal and paralinguistic features play a significant role and are related to start-up valuation and probability of investment. Although this study provides some interesting evidence that both verbal and paralinguistic features are correlated with the success of pitches, it is a shallow analysis which cannot be used to make predictions.

In Section 2.2.1 the paper by Sun et al. (2020) was reviewed, who used handcrafted (eGeMAPS) and deep (VGGish) speech features to predict emotions. Apart from training a unimodal based on speech, two other unimodal models are trained in this experiment: text and vision. Several deep embeddings are extracted for these modalities, including BERT word embeddings. Sun et al. (2020) use both early and late fusion to create a multimodal model. Here the best performing feature sets for each modality are combined into the multimodal model. Here late fusion achieves consistently better performance than early fusion. The authors observe that multi-modal models can substantially improve the performance. Furthermore, the authors find that using too many unimodal models may hurt the performance.

Soleymani et al. (2019) have developed a multimodal deep neural network based on verbal and non-verbal behaviour to predict the level of self-disclosure. The data consisted of 727 responses on questions from 102 participants, where each response was rated on the willingness to disclose (on a 7 point scale). Three modalities capturing the responses of the participants were analyzed: language, speech and vision. To represent language, BERT and LIWC were used. Using BERT each instance (response on a question) was transformed into a 768-dimensional vector. With LIWC, 93 features were extracted from the text of each utterance, forming a 93-dimensional vector. For the speech (acoustic) modality two types of hand-crafted feature representations (MFCC and eGeMAP) and a deep representation are used. MFCC features are extracted using openSMILE, generating a T x 39 matrix for each sample, where T is the number of frames based on the hopsize. Using VGGish, an embedding of length 128 is extracted for each frame of the audio signal, resulting in a T x 128 matrix. Finally, two sets of vision features are extracted using openFACE and FACS.

Firstly, Soleymani et al. (2019) estimate self-disclosure using unimodal models. All the different feature representations are fed into an encoder that transforms the output into a 1 x 128 vector. For the language features this is a single fully connected (FC) layer or an instance based encoder. For speech and vision, the temporal dynamics play an important role and for that reason recurrent layers (single layer GRU) are used as sequence based encoders. Each of these encoders are then followed by a fully connected layer that outputs the predicted self-disclosure score. The models are trained and evaluated using k-fold cross-validation. Since self-disclosure is estimated using a score, the models are evaluated using Spearman correlation (r). For the unimodal models, the models based on deep representations, BERT for language, ResNet for vision and VGGish for speech. achieve the best performance. Generally speaking, the linguistic modality achieves the highest performance of the three modalities ($r = 0.58$).

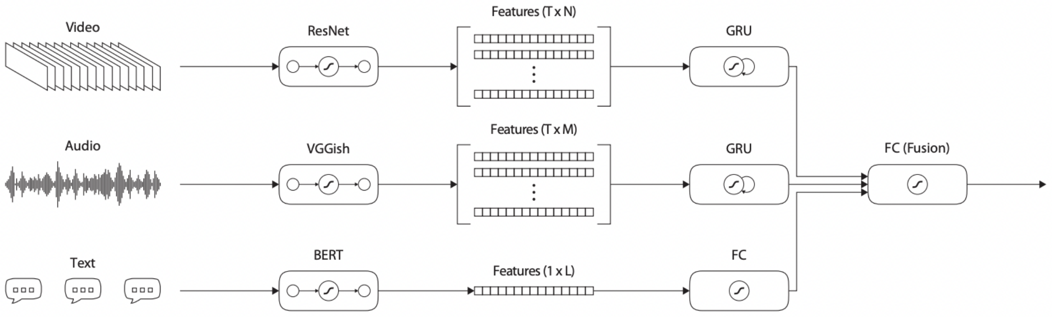

Figure 2.4: The multimodal model to estimate self disclosure by Soleymani et al. (2019)

Subsequently, a multimodal model is created by using the encoders of the best performing models for each modality (BERT, VGGish and ResNet). These three pre-trained encoders are followed by one FC layer to fuse the modalities before being fed into a final layer for prediction (see Figure 2.4). Additionally, the authors develop a late fusion model by averaging the output of the best performing models of all modalities. The results of Soleymani et al. (2019) show that for this task the performance of both fusion strategies is comparable (both $r = 0.58$). However, the fusion models are not able to improve on the performance of the best performing unimodal, BERT. On the contrary, in a cross-corpus experiment where the training is done on one data set and the evaluation on another, the late fusion model outperforms the best unimodal model. Although for this specific task in the within corpora experiment the multimodal did not improve the performance, when combining different modalities the results can be generalized beyond the data used for training.

Tavabi et al. (2020) use a multimodal model to detect whether a conversation two people have can be considered motivational, a style that evokes a persons personal intrinsic reasons to change their behaviour. As it considers intrinsic reasons to make a change in behaviour it can be seen as some sort of decision-making. The authors study behavioural cues in both speech and language features, using a database consisting of therapy sessions between therapists and clients with alcohol problems. Two feature sets for the text modality are extracted: LIWC and BERT. LIWC is chosen for its interpretability and to identify important features. For every utterance (client or therapist) a 93-dimensional feature is extracted. BERT is used to take advantages of the powerful pre-training representations, for all utterance a representation of length 768 is extracted. For the speech modality also two types of feature sets are used, namely eGeMAPS (from openSMILE) and VGGish. The eGeMAPS consists of 23 features that can be interpreted. A 128-dimensional vector is extracted using VGGish and applying PCA, using a hopsize of 0.96 seconds.

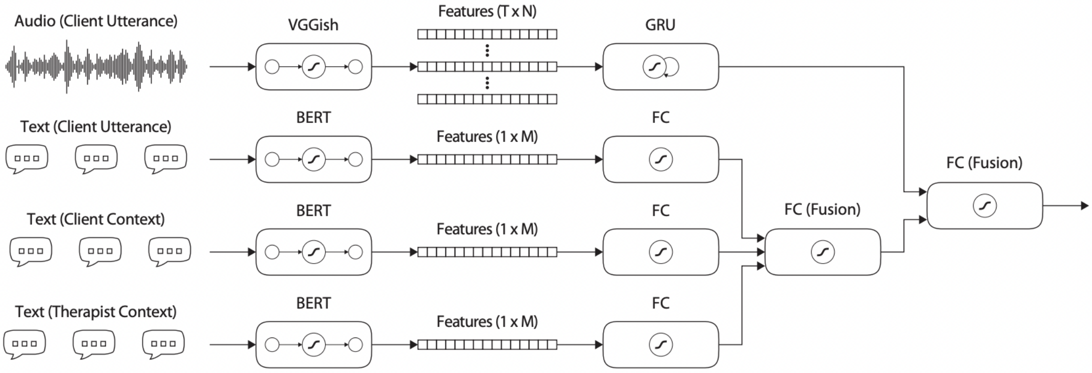

Figure 2.5: The multimodal model to classify motivational interviewing by Tavabi et al. (2020)

Tavabi et al. (2020) use a similar methodology as Soleymani et al. (2019) to train the unimodal models. For the text models, an instance based encoder is used to map the input feature space to a 256-dimensional embedding. Again, since in the speech modality the temporal dynamics are significant, a sequence based encoder (1 layer GRU) is used for the VGGish and eGeMAPS vectors. To make predictions on the unimodal models, the 256-dimensionl vector produced by the encoders are fed into a classification layer, which makes a 3-class classification (sustain talk, neutral or change talk). For the multimodal model, early fusion is used. The same 256-d embeddings from the previous models are now concatenated and fed into a final layer before classification. The authors also perform late fusion by averaging the predictions of the unimodal models.

The results of the analysis indicate that using text and speech features it is possible to estimate when interviewing can be considered "motivational", reaching an F1-score of 0.72. For this specific task, text is the strongest unimodal model. The multimodal model consisting of both text and speech obtains very similar performance scores as the unimodal text model and thus not substantially increases the performance. The authors argue that this might be explained by the fact that three different feature sets (client utterance, client context and therapist context) are extracted while only one feature set for speech (client utterance) is used. Therefore, in this case the speech modality does not add a lot of additional information for the model. When only extracting features for the utterance of the client, the (early fusion) multimodal model slightly outperforms the unimodal text model. Comparing the two fusion techniques, in this study early fusion outperformed late fusion.

Looking at some other studies that specifically combine text and speech features in a single model, we find some interesting results. Toto, Tlachac and Rundensteiner (2021) developed a deep transfer learning multimodal classification framework for depression screening. Due to privacy concerns, audio data sets consisting out of people speaking and depression labels are limited in size. To tackle this problem, the authors propose a deep learning framework based on speech and text. To overcome the problem of the small data set, pre-trained audio (VGGish) and text (BERT) feature extractors are augmented by a dual self-attention mechanism. The model, called AudiBERT, achieved state-of-the-art performance on the depression recognition task. The multimodal model demonstrated a robust improvement (ranging from 6% to 30% F1 score) in comparison to the unimodal models.

From the studies discussed in this section, we cannot conclude that either early or late fusion achieves higher accuracy. In some cases late fusion is preferred since it is simpler to implement. For all the papers discussed here, the multimodal model substantially outperformed the unimodal models, which is in line with what is expected based on theory.

Chapter 3

Methodology

In this chapter, the methodology and experimental setup employed in the research are discussed in detail. The data set utilized in this study, as well as previous work on this data set, are introduced. Then, the data pre-processing and feature extraction process are described. Subsequently, the experimental pipelines utilized to address the research question and sub-questions are outlined. Finally, the training and evaluation of the models are examined.

3.1 Data

In this thesis we used the *Data Management Entrepreneurial Pitches* data set containing video recordings from entrepreneurial pitch competitions. This data set is collected and maintained by Dr. Werner Liebregts, assistant Professor of Entrepreneurship at the Jheronimus Academy of Data Science (JADS) in Den Bosch, the Netherlands. JADS is an initiative of Tilburg University and the Eindhoven University of Technology. This data set was originally part of a study by Liebregts et al. (2020) on the potential of studying the role of social signal processing in entrepreneurial decision-making.

3.1.1 Overview of the data set

The *Data Management Entrepreneurial Pitches* data set consists of two main components: video data of entrepreneurial pitches followed by a Q&A session and survey data coming from both the pitchers and investor judges. The video data consists of students of the JADS that perform an entrepreneurial pitch on a start-up business proposition in front of judges that have experience with investing in start-up companies. Both the pitchers and the judges are video and audio recorded during the presentation. According to the guidelines, every pitcher had 3 minutes to present their business proposal, followed by a Q&A session of 10 minutes. However, in practise these guidelines were not implemented very strictly. For the sake of this study, only the actual pitch segment of the video is analyzed, the Q&A segment is discarded during the pre-processing.

The pitches of the start-up course of the JADS have been recorded for several years, starting from 2018 up to 2021, resulting in 52 videos in total. Before being recorded, all the participants (pitchers and jury members) had to fill in an informed consent form. Based on these consent forms, 42 pitches can be included in this study. All of the pitchers of these 42 pitches gave permission for their video and survey data to be used for research. For the pitch video data, only the consent forms of the pitchers are relevant since we exclusively analyze the extracted audio files of the pitch segments. As the judges do not interrupt the speaker during the pitch itself, the audio files of these segments do not include any data of the judges.

Within the 2018-2021 time frame, the COVID-19 pandemic started. Due to the pandemic, the pitch sessions of the JADS changed from in-person meetings to online settings. Therefore, one share of the pitches is documented in an offline setting and the other consists of recordings of online meetings. Kuhn and Sarfati (2021) explored whether the move to online settings affected investors' perception of social signals. Their findings suggest that since body movement is limited, acoustic features plays a more substantial role in the assessment of pitches in online settings. It is likely that the relationship between acoustic and linguistic features of speech and probability of investment is not the same in online and offline settings. For this reason, only the in-person recordings (25 in total) are used for the overarching research question and the first 4 sub-questions. To investigate how the models trained on the offline database generalize to different settings, we use the online database, consisting of 17 additional videos, as the test data for the cross-domain experiments.

Besides the video data, the data set consists of survey data on the investors and the pitchers. The data of three different surveys is included in the data set: an investor survey, a student survey and a pitch survey. The investor survey is meant for the judges and includes demographic information, character traits and investor experience. The student survey is conducted before and after giving the pitch and involves similar questions as the investor survey, except that there is no focus on investor experience but on passion for entrepreneurship. The third survey is the most interesting for the sake of this research and consists of the jury evaluation reports of the pitches. After each pitch, the investor judges evaluated the pitch on several aspects, including non-verbal behavioural cues and the business idea itself. In this research, we specifically look at the probability that the investor would invest in the pitched business idea. Every judge ranked this probability from 0 to 100. Looking at the consent forms of the judges, all of the judges who scored the 25 in-person and 17 online recorded pitches that are listed in the appendix gave permission for their survey data to be used for academic research.

3.1.2 Previous work on this data set

The *Data Management Entrepreneurial Pitches* data set has been used for previous studies. In the work of Goossens (2021) and Goossens et al. (2022) the impact of vocal behaviour on funding decisions is analyzed. In the study, the VGGish model is used to extract deep feature representations for the acoustic modality of the audio files of the pitches. This deep representation is combined with a hand-crafted feature representation and fed into a RNN in order to model the long term dependencies. Goossens (2021) uses both the online and offline videos in the analysis (42 in total) and makes a binary classification on whether a pitch would attract investment. The best performing models, the combination of a deep and hand-crafted representations trained on either a LSTM or GRU achieved an accuracy of 77.78%.

Stoitsas et al.(2022) focused on nonverbal behaviour cues and self reported characteristics from both pitchers and investors. Using different feature sets consisting of cues from several modalities such as facial expressions, head movement and vocal expressions, the investment decisions of the pitches are predicted. Their findings show promising results for the prediction of investor's evaluations of entrepreneurial pitches based on nonverbal behavioural cues. In this study the behavioural cues are stronger predictors than the models trained on the self-reported characteristics. The best performance was achieved for the models trained on head movement and vocal expression features. When predicting "the probability that you would invest", the best performing unimodal model was trained on head movement, which had an average MAE of 16.47. Here the multimodal late fusion model had an average MAE of 17.25 and thus did not result in an increase in performance.

3.2 Feature extraction

3.2.1 Pre-processing the data

Before features can be extracted from the data and a model can be trained on these features, several pre-processing steps must be completed: (i) removing non-consent pitches, (ii) extracting audio from video data, (iii) trimming the audio data, (iv) splitting the audio data in chunks, (v) converting speech to text and (vi) creating a single score for likelihood to invest.

(i) removing non-consent pitches : As discussed in the previous section, the entire data set consists of 52 videos. However, not all of these videos are used in the research, when removing non-consent pitches, 42 are available for our research. Then, the data

set is split into an in-person database (25 pitches) for the main goals of this study, and an online database (17 pitches) for the cross-domain experiment.

(ii) extracting audio from video data: In this study, only acoustic and linguistic features are considered, thus the videos are first converted into audio files. The conversion process was carried out using the MoviePy package in Python, with all files being transformed into WAV format audio files. The use of WAV format has been chosen due to its uncompressed nature, which allows for the preservation of more information for feature extraction using the openSMILE and VGGish packages.

(iii) trimming the audio data: In order to accurately analyze the pitches, it is necessary to trim the audio files to only include the pitch segment itself and exclude the Q&A session. Although the guidelines for the presentations stipulate a duration of three minutes for the pitch, in practise this was not always fulfilled. Therefore, it is necessary to manually extract the pitch portion of the audio from the full video, as the pitch may not always begin at the start of the video or end at the three-minute mark.

(iv) splitting the audio data in chunks: While using VGGish, a feature vector is extracted for every 0.96 seconds of audio data. To get similar size feature vectors for both the deep VGGish and the hand-crafted openSMILE features, chunks have to be created. Here, we split all the audio data in non-overlapping chunks of 0.96 seconds. By doing so, openSMILE can be used to extract features for the exact same frames as VGGish.

(v) converting speech to text data: Before linguistic features can be extracted, the audio files have to be transcribed to text data. This is done by using Google's Speech-to-Text API, which has state-of-the-art accuracy on automatic speech recognition tasks.

(vi) creating a single score for likelihood to invest: Every pitch in the data set is evaluated by several judges. To train the model, a single likelihood to invest score for every pitch has to be determined. Here we set the maximum score out of all judges' scores for a specific pitch as the single probability to invest score. This decision is justified for several reasons. Firstly, the goal of a pitcher is to raise money for their business. If one of the investors is very enthusiastic about a pitch but the other two are not at all, the pitch is likely to be "successful" and raise money. The average probability to invest score of this pitch would not be very high since two of the judges are not enthusiastic. For this reason the maximum score is used, since it gives a better indication of whether a pitch is successful. Another reason is that the investors come from different background and industries. Therefore, in some cases it is possible the investor has no experience in the industry a pitcher is presenting on, and would probably not invest in this pitch. Including this pitchers (low) probability to invest in the final score would not

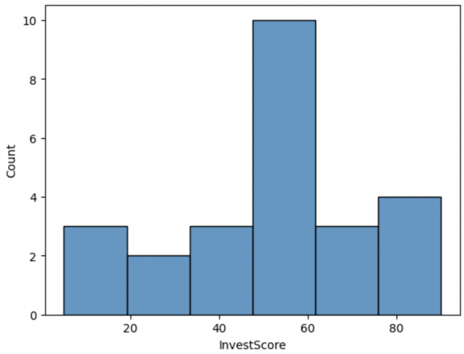

Figure 3.1: Distribution of in-person pitches

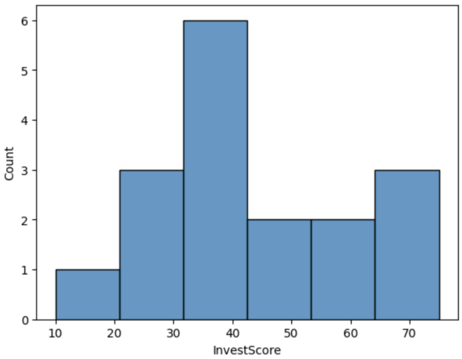

Figure 3.2: Distribution of online pitches

be an truthful reflection of the actual quality of the pitch. Following this definition, the distribution of scores in the in-person data set and the online recordings are presented in Figure 3.1 and Figure 3.2.

3.2.2 Acoustic features

For the acoustic modality, two categories of feature sets are extracted: explainable, hand-crafted features and deep representations. For the hand-crafted features the openSMILE toolkit is used, for the deep features we use VGGish:

openSMILE: Out of the openSMILE kit we specifically use *the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02)* feature set (Eyben et al. 2015). This is a basic standard acoustic parameter set intended to provide a common baseline for research in the acoustic domain. It consists of 88 features, including Low Level Descriptors (LLDs) and functionals, which are extracted from the 0.96 second chunks of audio created in the pre-processing. These features are organized into a T x 88 matrix for each pitch, where T represents the number of 0.96 second chunks that fit into the length of the pitch. In order to model the temporal information in the audio signal, a gated recurrent unit (GRU) is used. A GRU requires all inputs to be the same size, in order to ensure that all pitches have fixed-length feature vectors, we fix the length of all pitches based on the size of the second longest pitch. This allows for the incorporation of as much audio information as possible while accounting for the longest pitch (09:25 minutes), which may be considered an outlier. To create equal length feature vectors, shorter pitches are zero padded and the longer pitch is trimmed. This is a standard method to create equal length features (e.g. Han et al., 2020) in the audio modality. In addition, a "static" openSMILE representation is extracted, which does not consider the temporal information of the audio signal, by taking the mean, maximum and standard deviation of each column over all the chunks, resulting in a 264-dimensional feature vector.

VGGish: VGGish converts audio input into a semantically meaningful 128-D embedding (Hershey et al., 2017). As discussed in Section 2.2.1, for every 0.96 seconds of audio an embedding is obtained. When feeding the pitches into the VGGish framework this results in a T x 128 embedding for every pitch, where T is the number of chunks. Here, the same procedure has been applied to fix the length, by zero padding and trimming all videos to a size 330 x 128, resulting in a deep representation of the acoustic features of the text that capture the temporal information. Similarly as for the openSMILE features, we also create a "static" deep representation, by taking the mean, maximum and standard deviation over all the chunks for each column, resulting in a 384-dimensional feature vector.

3.2.3 Linguistic features

For the linguistic modality, both hand-crafted and deep representations are extracted. The hand-crafted representations consist of LIWC features, while the deep embeddings of the language used in the pitches is represented by LongFormer features.

LIWC: Linguistic Inquiry and Word Count (LIWC) is a text analysis tool that determines the percentage of words in a text that fall into one or more linguistic, psychological and topical categories. The core of the tool is a dictionary containing words that belong to these categories. The most recent version, LIWC-22, is used to extract 116 features for each pitch (Boyd et al., 2022). These features are also used in an explainable model, which enables us to draw conclusions on what word "categories" play a role in the investment decision-making process.

Longformer: As discussed in Section 2.2.2, most transformer based models cannot be used for longer text sequences. Since the pitches in the data set are many cases too long to be compatible with models such as BERT. For this reason, Longformer is used in this study, which has a linear (instead of a quadratic) attention mechanisms, allowing much longer input texts (Beltagy et al., 2020). Using the output of the pooled layer of the Longformer model, a 768-dimensional embedding is obtained for the text of every pitch. Given the large number of features in the Longformer model compared to the LIWC model, we also create another deep feature representation with a smaller number of features. With the help of principal component analysis (PCA) the number of features is reduced substantially. For every fold in the training process, principal components are obtained only using the training set, then the coefficients are used to also transform the test set. Splitting the data sets in specific folds for training and testing the model is discussed in more detail in section 3.4.

3.3 Experimental Setup

3.3.1 Unimodal models: single set of feature representations

The first part of the analysis is aimed to answer the first sub-question: To what extent can an acoustic or linguistic unimodal model predict the likelihood to invest of entrepreneurial pitches while using either hand-crafted or deep feature representations? Here we developed 7 individual models, that all predict the probability of investment. The approach for these unimodal models is inspired by the models of Soleymani et al. (2019) and Tavabi et al. (2020).

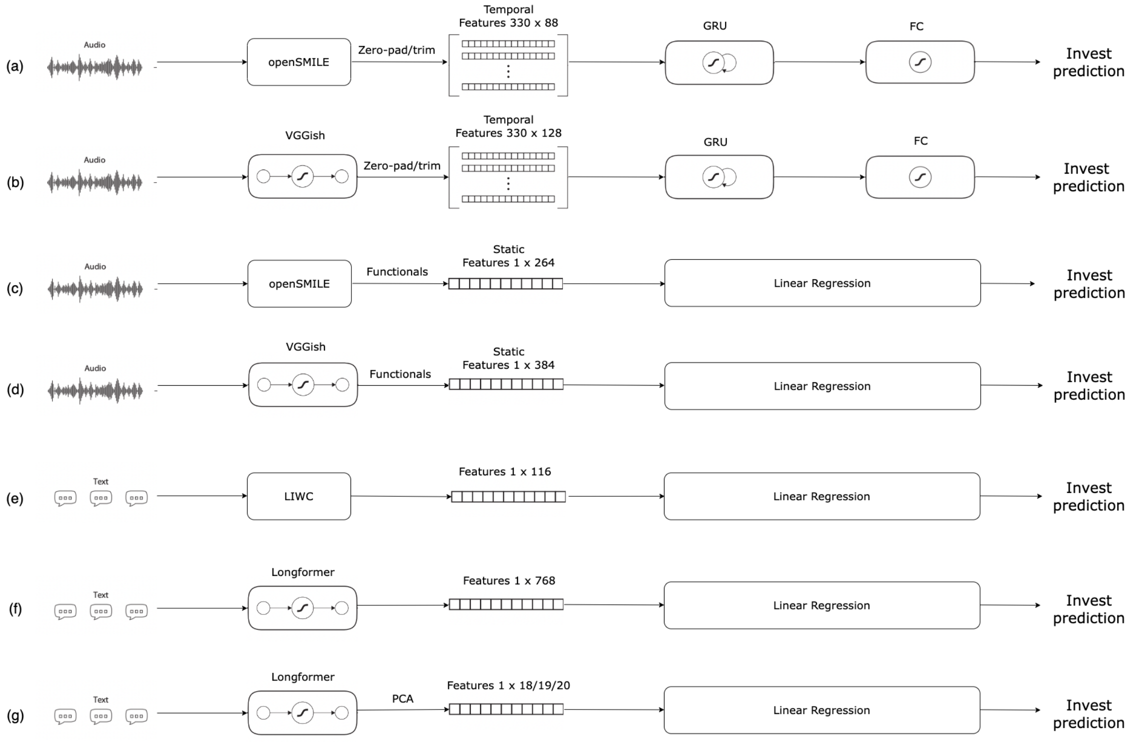

Figure 3.3: Unimodal models consisting of either hand-crafted or deep feature sets

The probability of investment of the entrepreneurial pitches is formulated as a regression problem. The first step involves extracting the feature sets. This is done following the procedure outlined in Figure 3.2. As shown in Figure 3.3, on the acoustic modality, 4 models are trained. Two of these models use a sequence of chunks to model the temporal information, a GRU is used to model this information. Since the temporal dynamics of human behaviour could be important in shaping the entrepreneur - investor interaction, we used a recurrent model as sequence-based encoder for the acoustic modality. This GRU model consists of a single GRU layer followed by a regression module that outputs a continuous value for the probability of investment of an entrepreneurial pitch. We perform hyper-parameter tuning by applying grid search over the parameters of the model, exact details on the training and evaluation of the models are discussed in Section 3.4. For the two remaining "static" acoustic models, we predict the probability of investment using a linear regression, by applying Xgboost Regressor. Here the one-dimensional feature set is used, consisting of the functionals calculated over each column.

For the linguistic modality 3 models have been developed, one on the hand-crafted feature representation and two on the deep representation. The PCA Longformer model is developed by applying PCA on every training set and subsequently transforming the test set. A smaller feature size for the Longformer feature vector is also useful when looking ahead at the model where it is combined with the LIWC feature representations. All the models in the linguistic modality are trained using a linear regression by applying Xgboost Regressor. A small grid search is performed over the number of instances, the learning rate and maximum depth of the model.

3.3.2 Unimodal models: combining deep and hand-crafted feature sets

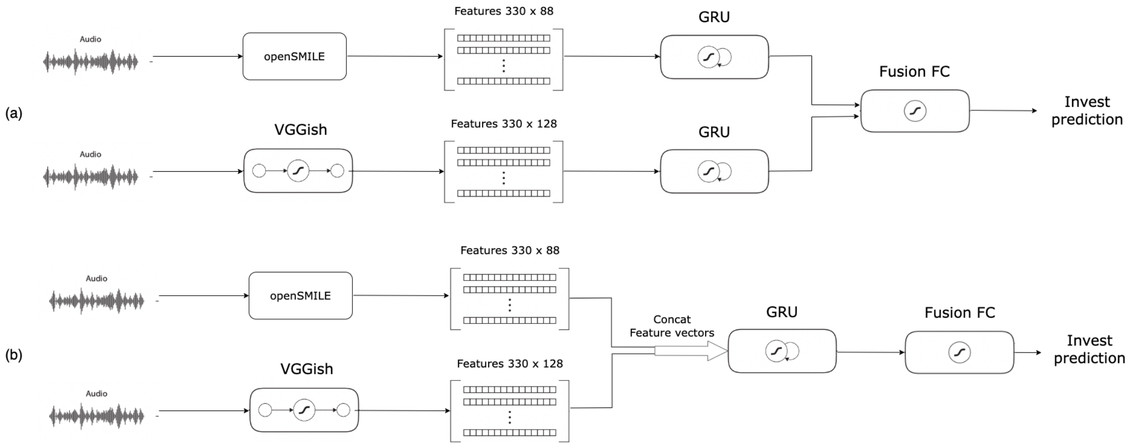

Figure 3.4: Unimodal acoustic models consisting of both hand-crafted and deep features

The second part of the analysis investigates the second sub-question of this thesis: How does combining hand-crafted and deep feature representations in a unimodal model affect the performance of the task to predict the likelihood of investment of entrepreneurial pitches? The approach is based on the suggestions raised in the papers discussed in Section 2.2.1 and Section 2.2.2. We combine the feature sets into a single model by both applying "early" fusion (a single regression model) and "late" fusion (taking the mean output of two regression models to get a single score).

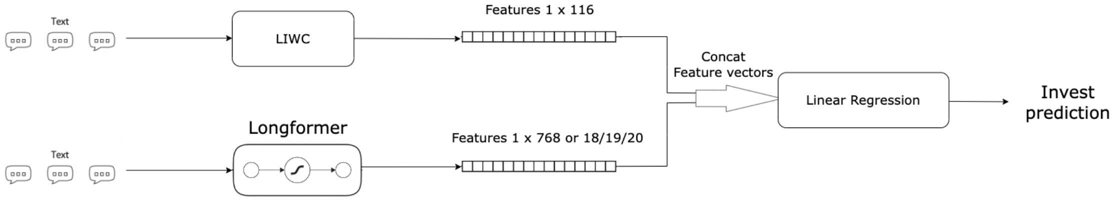

Figure 3.5: Unimodal linguistic model consisting of both hand-crafted and deep features

For all the models, the same feature extraction procedure and resulting feature vectors are used as in the previous analysis. For the early fusion acoustic models the two feature vectors are combined using two distinct strategies. The first strategy, presented in Figure 3.4 (a), involves creating a model consisting out of two GRU input (encoder) layers, one for each feature set. These two GRU layers are followed by a single fully connected layer, where the two inputs are fused, and a dense layer to output the likelihood of investment score. For the second strategy, presented in Figure 3.4 (b), we directly combine the feature vectors by merging them. The merged feature representation is then fed into a single GRU and subsequent fully connected layer to output the prediction, comparable to the models consisting out of a single feature set. For the acoustic modality, the two "static" feature sets are concatenated and modelled using a linear regression.

For the linguistic modality, two "early" fusion models have been developed. The first model consists of the merged LIWC and standard Longformer representations. The second model consists of the merged LIWC and the PCA-transformed Longformer representations. Both these features sets are modelled using a linear regression.

For both modalities we also combine the two feature sets into a single model using late fusion. For the acoustic model we select the two best performing models of each category (hand-crafted and deep): the (temporal) openSMILE and (temporal) VGGish model. For every instance in the test set, the prediction of this late fusion model is derived by taking the mean of the individual openSMILE and VGGish predictions. The performance of the model is then determined using the same method as the other models, by calculating the Mean Absolute Error between the predictions and the actual values. The same method is applied for the linguistic model, where the output of the PCA transformed Longformer model is integrated with the LIWC model output.

3.3.3 Multimodal models: different fusion techniques

To answer the third sub-question and the overarching research question, multiple models are developed in the third set of experiments. These models are designed to predict the probability of an investment score using two different multimodal models architectures. The first type employs late fusion, which involves averaging the regression output of the unimodal models to produce a single score for the entire multimodal model. Two late fusion models are developed: one using the best performing feature set of each modality (VGGish and PCA-transformed Longformer), and the other using all four feature representations.

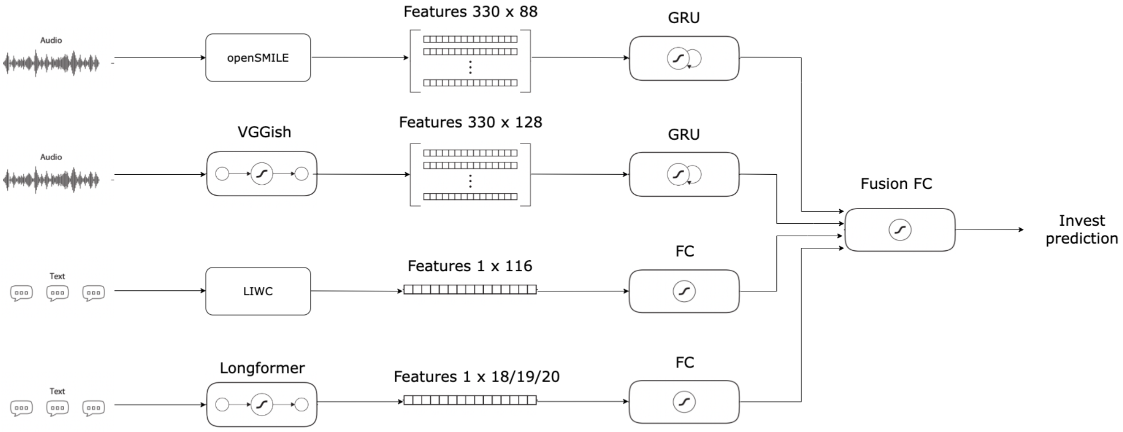

Figure 3.6: Multimodal model - early fusion: hand-crafted and deep feature sets for each modality

In the subsequent section of this research, experiments using the second type of multimodal architectures, namely early fusion, are conducted. The same feature representations and encoders as in previous steps are utilized. Similarly as was the case for the late fusion models, we develop a model consisting out of all feeature representations and a model solely consisiting of the best performing representation for both modalities. A model that can accommodate the four different inputs is created using the functional API package of Keras. As shown in Figure 3.6, temporal acoustic features are first encoded using a single layer GRU model with a similar architecture as depicted in Figure 3.4 (a). The GRU is used to capture the temporal information present in the audio signal. The hidden layers for the linguistic features are fully connected layers with ReLU activation. The four branches of the network are then concatenated using a fully connected (FC) layer. The final layers consist of two dense layers, the first of which combines all the different inputs and the second performs the actual regression and predicts an investment score. Thus, using this complete vector of information from two modalities, comprising two types of feature sets per modality, a model is trained to predict the probability of investment for the pitches. Additionally, a model composed of the best-performing feature set of each modality has been developed. The architecture for this model is shown in Figure 3.7, and consists of the components for the VGGish and Longformer features of the complete multimodal model.

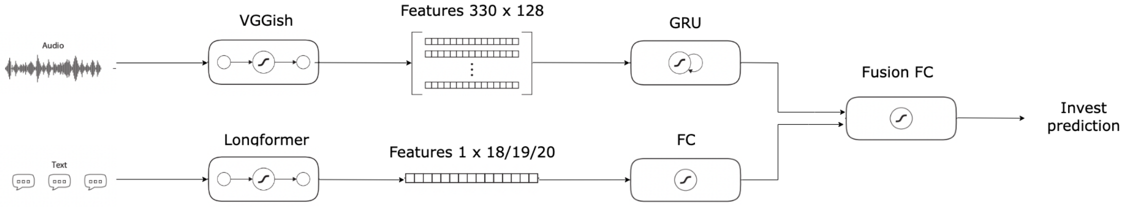

Figure 3.7: Multimodal model - early fusion: best performing feature set of each modality

3.3.4 Explainable acoustic, linguistic and multimodal models

The experimental setup for the explainable models is relatively straightforward. Three individual explainable models are developed: an acoustic, a linguistic and a multimodal model. For the acoustic modality, a slightly different feature set is used as in the previous desribed experiments. Instead of extracting features for chunks of the pitches, we extract the *eGeMAPSv02* (Eyben et al., 2015) features over the whole pitch at once. Consequently, a one-dimensional vector of length 88 is obtained for the entire pitch. For the explainable linguistic model, the same LIWC feature representation is implemented as earlier. For the multimodal model, the openSMILE and LIWC vectors are concatenated. All models are trained using Xgboost Regressor. A grid search is performed over the number of instances, the maximum depth and the learning rate.

The output of these explainable models can be analyzed using a tool such as SHAP (SHapley Additive exPlanations) (Lundberg, Scott and Lee, 2017). SHAP is an approach based on game theory, which connects optimal credit allocation with local explanations using Shapley values. With the help of SHAP, we can gain insights into which features play an important role in predicting the likelihood of investment of the pitches in our data set. To get an overview of which features are important for a model, the SHAP values of every feature of every pitch can be plotted. By doing so, both the importance of the features and the distribution of the impact each feature has on the output can be visualized. In this research this means, we can analyze whether a higher value for an acoustic or linguistic feature would increase or decrease the likelihood of investment. Since in this thesis Xgboost Regressor is applied, which is a tree based model, the TreeExplainer implementation of SHAP is used (Lundberg et al., 2020).

3.3.5 Generalizing the models to online settings

The goal of the final sub-question is to test to what extent the performances of the models outlined in Section 3.3.1 - 3.3.3 generalize to data collected in a slightly different setting. To answer this question, the best performing models for each sub-question are tested on the recordings of the pitches in the online setting. For the online videos, the exact same feature extraction procedures are followed as outlined in Section 3.2. Then, we evaluate the best performing acoustic, linguistic and multimodal models using all the instances of the online data set.

3.4 Training and evaluation

In this section, the procedures for training and evaluating the described models are detailed. Prior to training the models, the data set must be divided into a training set and a test set. Rather than using random sampling, the data is divided into folds based on the session in which a pitch was recorded. The 25 in-person recorded pitches included in this study were recorded over four distinct sessions, each with different pitches and investors. The pitches are divided into folds, with each fold consisting of all the pitches from a single session, in order to ensure that each fold or test set can be considered representative of an actual session. This results in four folds, ranging in size from 5 to 7 pitches per fold. The pitches recorded in the online setting are all tested at once and thus do not need to be split into separate folds. An overview of the pitches and the division of pitches over the folds can be found in the Appendix.

For the models based on Xgboost Regressor hyper-parameter tuning was conducted to optimize model performance. For each model and each fold, hyperparameters were explored using a grid search with 5-fold cross-validation on selected hyperparameters. The model was then evaluated by predicting pitches in the unseen test set using the best-performing parameters identified during the grid search. The absolute error was used as the objective function during training. The explored hyperparameters and values are listed in Table 3.1.

Hyper-parameter	**Explored values**
Number of instances	100, 200, 300, 400
Learning rate	0.1, 0.2, 0.3
Maximum depth	4, 5, 6, 7

Table 3.1: Hyper-parameters included in Grid Search for Xgboost Regressor models

The deep learning-based models consisting of a single input layer were trained using the Keras Sequential API. A grid search was also employed as part of the training process to optimize the hyperparameters of the model. For each model, three parameters were selected for hyper-parameter tuning: the number of units in the GRU layer, the learning rate, and the drop-out rate. Given limited computing power, the number of parameters and the number of explored values was kept relatively small. The explored parameter space is presented in Table 3.2. The Adam optimizer was used during training, and early stopping was applied by monitoring the validation error on a hold-out set of the training set (using a validation split of 0.2). The number of epochs was set to 100, although the application of early stopping means that this value is not particularly relevant in this case. A batch size of 5 was used while training all these models.

Hyper-parameter	**Explored values**
Number of units	32, 64
Learning rate	0.001, 0.1, 0.2
Drop-out rate	0, 0.1, 0.2

Table 3.2: Hyper-parameters included in Grid Search for single input GRU models

The deep learning models with multiple, diverse input layers are created using the Keras Functional API. Due to the complexity of these models, they are not compatible with grid search implementations, and therefore no hyper-parameter tuning is performed. Instead, the hyper-parameter settings for each fold of a specific model are set to be exactly the same, allowing for fair comparisons across different models and preventing overfitting on the test set. For all models, the GRU layers consist of 64 units, the dropout rate is set at 0.1, the Adam optimizer is used, and the learning rate is set to 0.001. Early stopping was applied by monitoring performance on the validation set, using a validation split of 0.2. The batch size was again set at 5.

The performance of the models is evaluated using two main metrics: performance on the prediction of the probability of investment and feature importance in explainable models. The performance of all models is measured using the same evaluation score, the

Mean Absolute Error (MAE). This allows for comparison of all models to each other and to the findings of other studies using this data set, such as Stoitsas et al. (2022). The MAE is calculated as the sum of absolute errors divided by the sample size and is a commonly used evaluation measure for regression problems. The feature importance of explainable models is evaluated using SHAP scores, as previously discussed in Section 3.3.4.

Chapter 4

Results

In this chapter, the results of the experiments discussed in the previous chapter are presented. The chapter is divided in sections based on the sub-questions of this thesis.

4.1 Unimodal models: single feature representations

The results from the unimodal models containing a single feature representation are summarized in Table 4.1. The table includes the Mean Absolute Error (MAE) for each individual fold, as well as the average MAE across all models. The best average result for each modality is highlighted in bold. The results indicate that the best performing acoustic model outperforms the linguistic models, with the acoustic VGGish model yielding the best results overall. Among the hand-crafted feature representations, LIWC outperforms openSMILE. When drawing a comparison between a hand-crafted feature set and a deep one, the results here suggest that the deep representations can form stronger predictors. For both modalities the models that use (the best performing) deep representations outperform the hand-crafted interpretable feature sets. The performance results for the non-PCA-transformed Longformer model are substantially lower than the PCA-transformed version. This could be caused by the large embedding size of Longformer in combination with a small data set and a relatively simple model, namely a linear regression. For this reason, in the remaining experiments, the PCA-transformed implementation is used to represent the deep linguistic features.

Upon analyzing the acoustic modality, it was found that the overall best performance was achieved using the deep (temporal) VGGish representation, which outperformed the openSMILE representation in all but one fold. The relatively weak performance of the "static" acoustic models, which are based on functionals of the entire pitch, suggests that it is valuable to exploit the capabilities of deep learning models to capture the complex non-linear relationships and temporal dynamics of the audio input. Using a deep model to represent the acoustic signals and a GRU layer to model the information captured in this signal results in a better performing prediction model. For this reason, the temporal implementations are used in the experiments where feature sets and modalities are combined into a single model.

Modality	**Feature Set**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Acoustic	openSMILE	18.54	15.23	13.44	20.92	17.03
Acoustic	VGGish	15.48	12.42	11.53	22.19	**15.41**
Acoustic	static openSMILE	20.90	15.30	16.35	31.09	20.91
Acoustic	static VGGish	24.07	19.76	15.14	22.34	20.32
Linguistic	LIWC	17.20	16.17	11.13	18.01	15.63
Linguistic	Longformer	16.27	21.40	15.95	27.69	20.33
Linguistic	PCA Longformer	16.03	12.25	10.94	23.21	**15.60**

Table 4.1: Unimodal models: single feature representation

The results of the linguistic modality show that the deep model outperforms the non-deep model, although this is only after reducing and transforming the number of features using PCA. It is interesting to note that the hand-crafted LIWC model has a very competitive performance across all folds. While the deep model demonstrates better average performance on the test sets, the hand-crafted models offer the benefit of explainability, which can be useful in guiding entrepreneurs to identify characteristics of an attractive pitch for investors. The performance and relevant features of these explainable models, both unimodal and multimodal, are discussed in more detail in Section 4.4.

4.2 Unimodal models: combining feature representations

In Table 4.2, the results for predicting the likelihood of investment using unimodal models that combine hand-crafted and deep feature representations are presented. For the acoustic modality, the model consisting of two separate GRU layers, one for openSMILE and one for VGGish, performs slightly worse than the model where the feature vectors are concatenated directly and fed into a single GRU. Both these models represent a significant improvement compared to the acoustic models with only one feature set. Additionally, the combined version of the static model outperforms both individual static features. However, the late fusion model, which takes the mean of the predictions of the individual models, does not improve on the single VGGish model.

The early fusion approaches for the linguistic modality are not able to improve on the average MAE of 15.60 of the PCA-transformed Longformer model. However, in this case, it was found that using late fusion increases the predictive power of the linguistic modality. Similarly as in the previous experiment, the best linguistic model performs marginally worse than the acoustic one.

The results here indicate that the method of creating a "hybrid" representation has the potential to outperform models that rely solely on hand-crafted or deep feature sets. This effect is observed for both modalities in the conducted experiment, and is consistent with findings from papers discussed in Sections 2.2.1 and 2.2.2.

Modality	**Model**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Acoustic	Concat VGGish + openSMILE, into single GRU.	14.22	17.53	10.38	17.13	**14.82**
Acoustic	VGGish + openSMILE into separate GRUs, then concat.	17.34	14.16	9.05	20.13	15.17
Acoustic	Static: VGGish + openSMILE	22.04	16.99	15.75	25.98	20.19
Acoustic	Late Fusion: VGGish + openSMILE	16.80	13.98	12.48	21.39	16.16
Linguistic	Concat: LIWC + Longformer	14.49	19.56	15.02	24.70	18.44
Linguistic	Concat: LIWC + (PCA) Longformer	17.79	17.02	11.61	23.63	17.51
Linguistic	Late Fusion: LIWC + (PCA) Longformer	16.62	12.61	9.56	20.61	**14.85**

Table 4.2: Unimodal models: combined feature representations

4.3 Multimodal models

This section reviews the results of the experiments on multimodal models and the effect of different fusion techniques on the performance of such models. The results of the multimodal explainable model are discussed in Section 4.4, which leaves four models, two for each fusion strategy, to be discussed here. For each fusion technique, a model consisting of all the four feature representations and a model containing the best performing feature set per modality (VGGish and PCA-transformed Longformer) are presented.

The results of the experiments demonstrate two main effects. Firstly, the overall results of the early fusion technique are compared to the late fusion models. We note that the best performing model is an early fusion model, namely the model in which early fusion of the best feature representations of each modality is applied. Overall, the average MAE of 13.91 of this specific model is the best performing model that has been found in the entire research. This model performs substantially better than its late fusion counterpart, consisting of the two best feature sets (average MAE of 14.95). However, it cannot be noted that early fusion is the strongest fusion approach in any scenario. From Table 4.3, we note that when using all the four feature representations, the late fusion approach slightly outperforms early fusion. We observe that the best performing early fusion model does improve on the best performing unimodal models (acoustic hand-crafted and deep features into a single GRU and the linguistic combined model of LIWC and PCA-Longformer). However, the best performing late fusion model is outperformed by both these unimodal models consisting of a combination of feature sets.

Fusion type	**Features**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Early	Explainable models	13.92	13.54	12.56	22.36	15.59
Early	Best set of each modality	17.17	13.51	5.47	19.47	**13.91**
Early	All sets	16.48	15.21	9.17	20.52	15.35
Late	Best set of each modality	15.11	11.73	10.47	22.47	**14.95**
Late	All sets	16.49	12.94	10.76	20.70	15.22

Table 4.3: Multimodal models: different fusion techniques

Secondly, for every fusion strategy, the performance of the model consisting of all the four feature sets can be compared to the model solely consisting of the two best performing representations. As discussed in Sections 1.2 and 2.2.3, most previous works using a multimodal model only use the best performing feature set to represent a modality. Since, the best features outperform all the four features, the findings here suggest that this approach is the strongest method to represent the modalities and forms the foundation for the best performing models. This effect is valid for both fusion techniques and zooming in on the individual folds, we observe that for both strategies in three out of four folds a better performance is achieved in the “best performing sets only” scenario. The increase in performance when removing the openSMILE and LIWC features is larger in the case of early fusion compared to late fusion.

4.4 Explainable models

In Table 4.4 the results of the explainable models are presented. The explainable acoustic model is a slightly different implementation of the static acoustic openSMILE model given in Table 4.1, since here the openSMILE features are obtained over the whole length of the pitch at once, instead of over chunks. We note that when looking at the average MAE across the four folds, this explainable openSMILE model outperforms the model that captures the temporal information of the audio signal presented in Table 4.1 (Average MAE of 17.03). However, when examining the individual models, we find that in 3 out of the 4 models, the temporal model outperforms the features obtained over the whole pitch at once that is given here. The better overall performance for the model here is caused by a large performance increase in performance on the first test set. The LIWC model given in the table here is exactly the same as in Table 4.1.

The average MAE of the explainable multimodal model is 15.59. This finding demonstrate that also when developing explainable models, using a multimodal model is beneficial when predicting the probability of investment of entrepreneurial pitches and outperforms the unimodal models it is made up of. Despite the fact that the average increase in performance is rather limited compared to the LIWC model, the multimodal model performs substantially better in the first two models. Comparing the explainable multimodal to the multimodal models based on deep learning frameworks in Table 4.3, we find that this approach is the worst performing model. However, we note that the improvement of this model in terms of explainability, compared to the deep learning models, is obtained at a relatively small cost in terms of model performance. The difference in performance of the best performing model, early fusion of the strongest representations, and the explainable model is rather limited (average MAE of 13.91 versus 15.59). At the cost of this slight decrease in performance, we do gain a lot of interesting insights by looking at the SHAP feature importance plots of these models.

Modality	**Feature Set**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Acoustic	openSMILE on whole pitch	12.75	16.71	14.08	23.72	16.82
Linguistic	LIWC	17.20	16.17	11.13	18.01	15.63
Multimodal	openSMILE + LIWC	13.92	13.54	12.56	22.36	**15.59**

Table 4.4: Explainable acoustic, linguistic and multimodal models

In Table 4.5 the explainable features that appear in the top 20 of feature importance for at least three out of the four folds are presented. When looking at the SHAP plots in Figures 4.1 - 4.3, some conclusions can be drawn on the effect these features have on the outcomes of the model. Here it must be noted that these values should be used carefully and should not directly be used to create guidelines for entrepreneurs to enhance their pitching skills, as is discussed in Section 5.3.

Linguistic LIWC model	**Acoustic openSMILE model**	**Multimodal model**
number	F0semitoneFrom27.5Hz_amean	Clout
achieve	loudness_amean	F0semitoneFrom27.5Hz_pctlrange0-2
Clout	F2frequency_amean	F2frequency_amean
Analytic	loudness_stddevNorm	Conversation
i	jitterLocal_amean	WC
Dic	F0semitoneFrom27.5Hz_stddevNorm	number
quantity	mfcc3_amean	
Conversation	F0semitoneFrom27.5Hz_pctlrange0-2	
	spectralFlux_stddevNorm	

Table 4.5: Explainable features that appear among the top 20 of feature importance in at least 3 out of 4 models for the respective linguistic, acoustic and multimodal models

Starting with the linguistic model, for the LIWC feature *number*, which is simply a count for the amount of numbers used in the text, the results seem to suggest that a lower value has a positive impact on the model output. This can be read from the tables by looking at the color and the side of the spectrum the majority of dots are. For *number* we generally observe more blue dots on the right half of the spectrum, this pattern is especially clear in folds 2 and 3. This would indicate that in the pitch setting studied here, using a lot of numbers during the pitch could have a decreasing effect on the probability of investment. Similarly, for the feature *quantity*, which captures words such as: all, more and some, this same effect is observed in the first model. However, in models 3 and 4, the relationship is reversed, making it difficult to draw a conclusion on the impact of this feature.

For the second category *achieve*, which contains words such as: work, better, best and working, we observe that there are mostly pink dots on the right side of the spectrum, meaning that the more these words are used during the pitch, the higher the prediction for the probability of investment. This means that in the studied setting, when the pitching entrepreneur uses more words that demonstrate a sense of achieving, this has

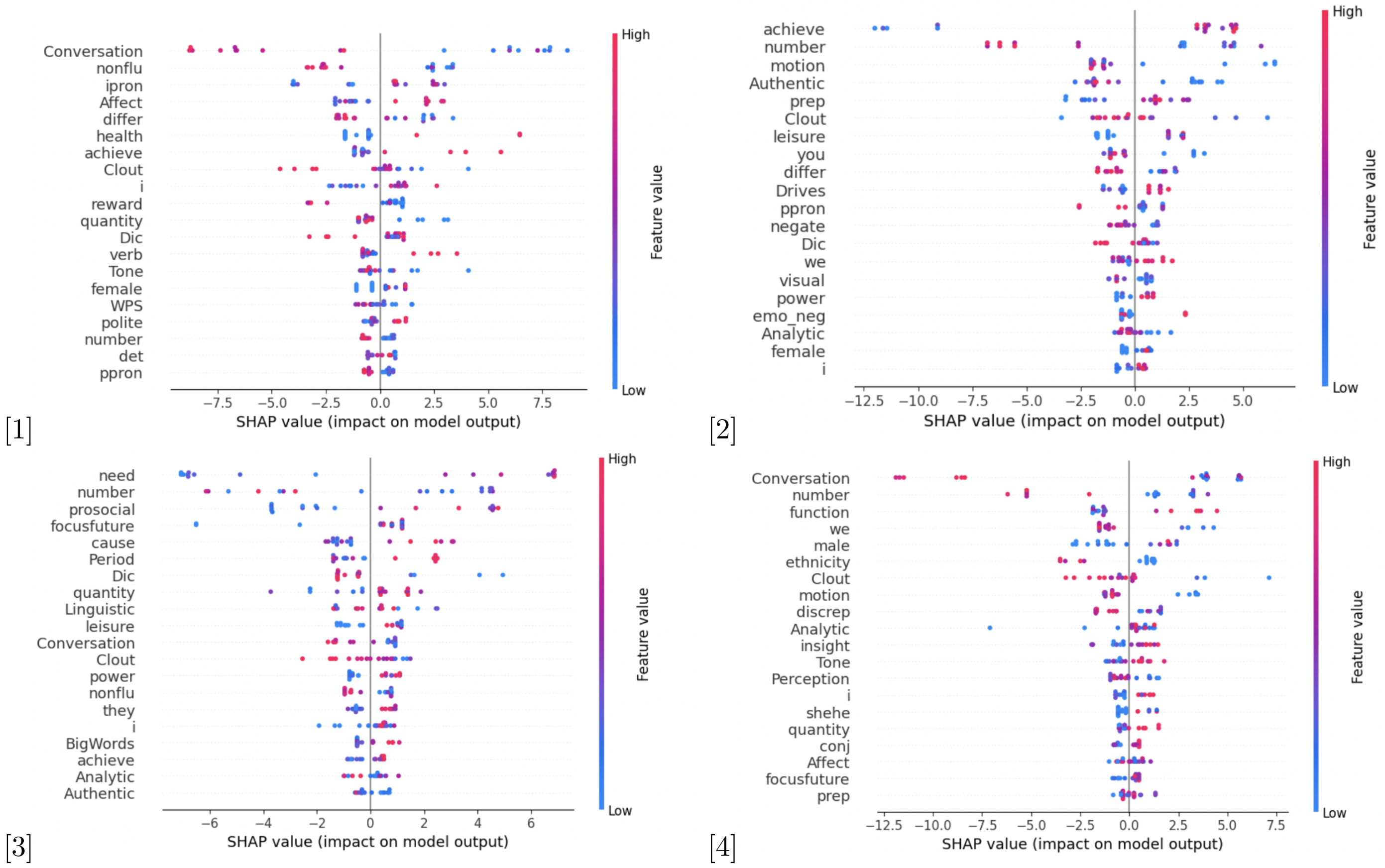

Figure 4.1: SHAP summary plots for the LIWC models

a positive impact on the likelihood that investment would be attracted.

The third common important feature is *Clout*. The LIWC feature set contains four overarching summary variables, of which *Clout* is one. This variable is described in the LIWC documentation as "language of leadership" (Boyd et al., 2022). A higher number for the *Clout* score indicates that the presenter is speaking from a perspective of high expertise and is confident, on the other hand, lower scores suggest a more tentative or humble speaking style (Pennebaker et al., 2015). For this feature, the SHAP plots for all the four models indicate that a higher value for this feature would lower the output of the model. This would mean that when the presenter seems to be highly confident, this has a negative impact on getting an investment in the pitch setting studied here.

Another potentially interesting strong feature, which is also a summary variable, is *Analytic*. A high number for the *Analytic* variable reflects formal and hierarchical thinking, while a lower number reflects informal or personal language. The effect of this feature on the model output is less straightforward compared to the previous cases. In models 2 and 3, even though the feature impact is relatively low, we observe mostly blue dots on the right and pink dots on the left, indicating that a lower value for *Analytic* would have a positive effect on the model outcome, and vice versa. However, for model 4, where the feature has a stronger impact on the model, some blue dots are spotted on the left side, while all the higher feature values (depicted by a pink/red color) are found on the right half. Therefore, we cannot draw an unambiguous conclusion on the effect the *Analytic* feature has on the likelihood of investment.

The feature *i* appears in the top features of all the four individual models. This category involves all 1st person singular pronouns, such as I, me and my. The effect of this feature is consistent across all these models, namely that using a high number of these kind of pronouns increases the probability of investment score predicted by the model (and vice versa). This finding seems to suggest that when the presenter makes the content of the pitch personal, and refers to him or herself in the story, this has a positive impact on receiving investments.

Finally, for the *Conversation* feature, which is the most important feature for two of the models, we observe that a higher score for this feature has a substantial negative impact on the outcomes of the model. It might be surprising that this feature comes up during a pitch situation where only one person is speaking. However, this category involves non-fluent speech such as: oh, um and uh, and also contains filler words. Therefore, the SHAP values for the *Conversation* feature indicate that when a pitcher presents with a lack of fluency, for example caused by stammering or usage of filler words, this has a negative impact on the likelihood of investment.

Upon analyzing the feature importance plots for the acoustic openSMILE models, 9 features are found to appear in the top 20 of at least three of the models. Since the openSMILE features are comprised of functionals over some of the descriptors of the audio signal, the features are less easy to interpret than the LIWC features. However, we can still observe some interesting results. For the common feature, *loudness amean*, which is the average loudness, a subjective auditory impression of the intensity of a sound, in model 4 we observe that a higher value for this feature reduces the model output. On the other hand, in model 2 we find a reversed relationship and observe some blue dots on the left side, indicating that a lower mean loudness reduces the likelihood of investment the model predicts. The relationship between the standard deviation of the loudness, the *loudness stddevNorm* feature and the impact on the model is more consistent across the different models. Here we find that a lower standard deviation for the loudness descriptor leads to a higher probability of investment predictions. This

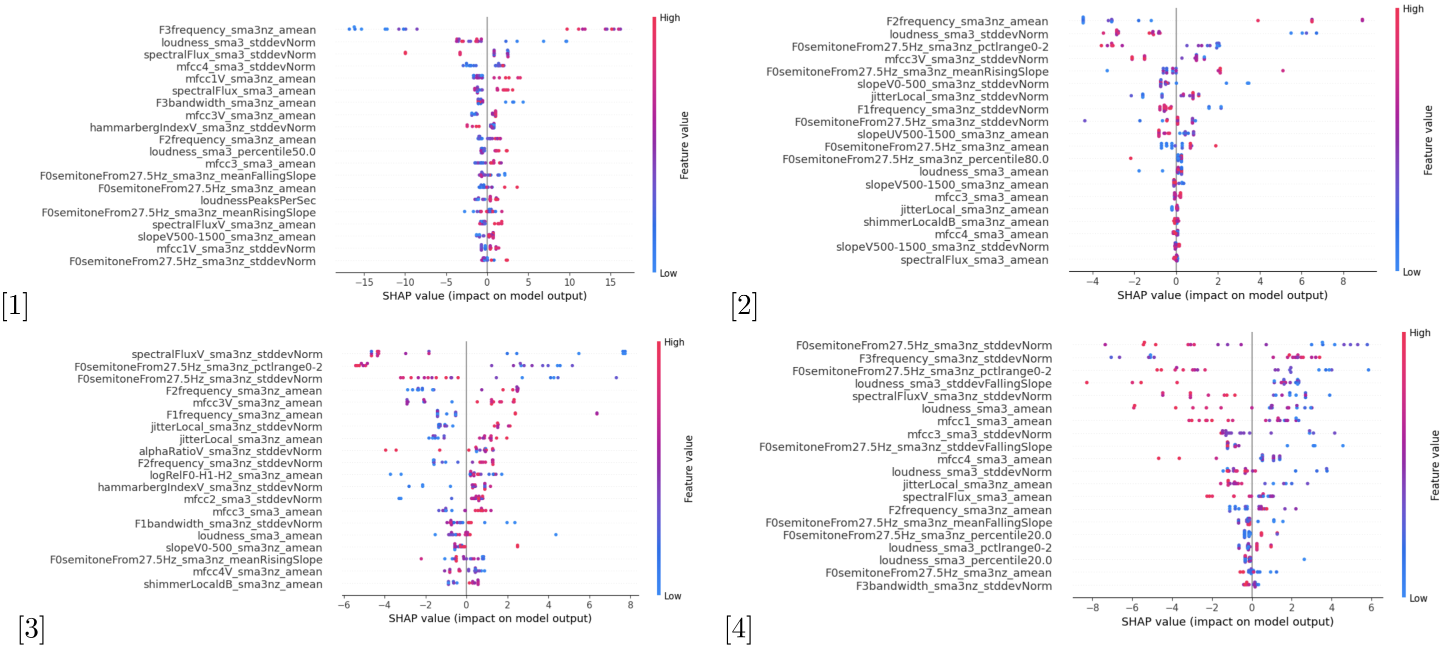

Figure 4.2: SHAP summary plots for the openSMILE models

suggest that consistently speaking on a loudness level closer to the mean increases the prediction for the likelihood of investment.

Amongst the common important features, several features representing a type of frequency are found. Starting with the *F0semitoneFrom27.5Hz amean*, this is a measure for the fundamental frequency of the audio signal, which represents the lowest frequency component of the periodic waveform and is often referred to as the pitch of a sound. For this feature we do not observe a consistent relationship. In model 4, a lower fundamental frequency increases the likelihood of investment, but in models 2 and 3 we do not find this correlation. We also find the *F2frequency amean* feature in the list of common important features. The second formant (F2) frequency is related to the vowel sounds of speech. *F0semitoneFrom27.5Hz pctlrange0-2* and *F0semitoneFrom27.5Hz stddevNorm* are both measures for the distribution of the fundamental frequency over the audio signal, where the first is useful to identify extremes and the latter to indicate the consistency of the frequency (Cooper and Sorensen, 2012). In model 4, these features are both in the top three most important features, and in both cases a lower value has a positive impact on the predictions of the model. For the *pctlrange0-2* feature this relationship is also found in models 2 and 3. The *jitterLocal amean* feature is also common and is related to the fundamental frequency, as it is defined as the variation in frequency over the signal.

Therefore, we note that for common important acoustic features there are three features related to the distribution of the fundamental frequency and two features related to the mean frequency, one for the fundamental and one for the second formant frequency.

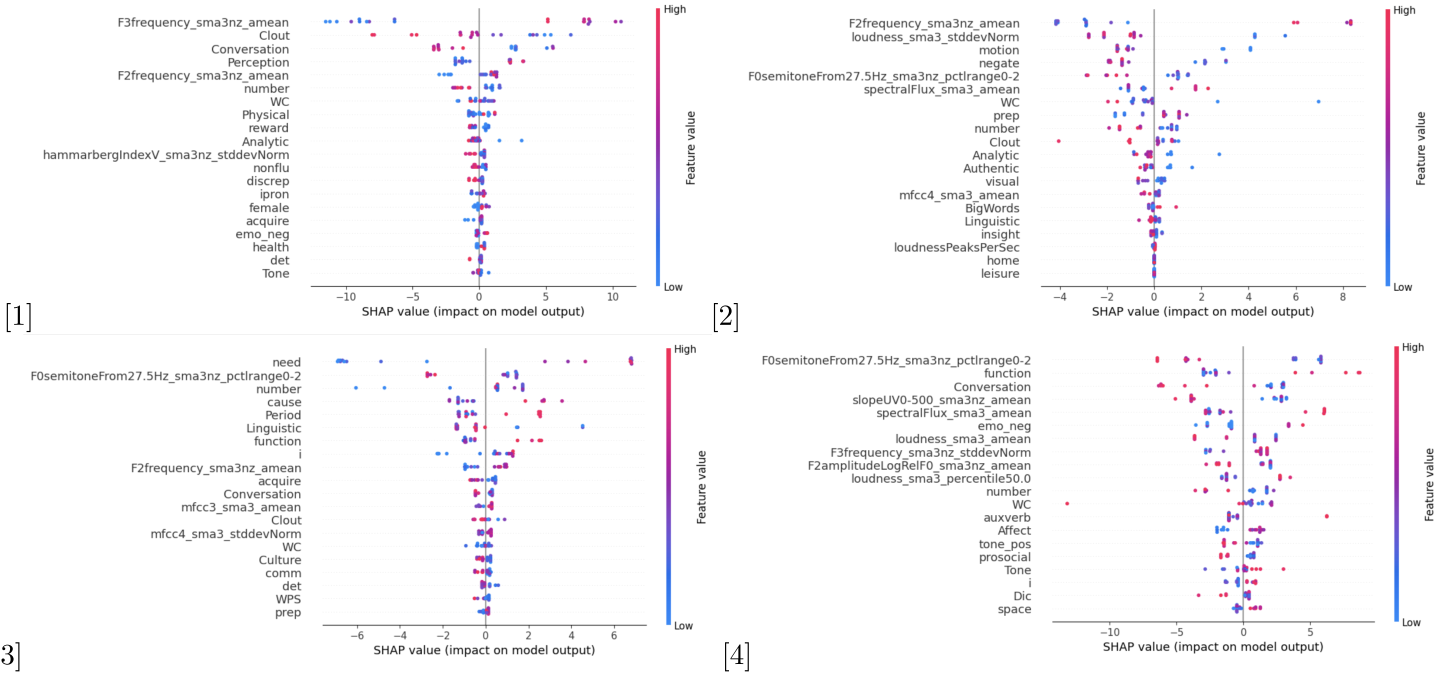

Figure 4.3: SHAP summary plots for the multimodal models

For the multimodal model, the 6 common features consist of 4 linguistic and 2 acoustic features. All of these features are also a common feature in the respective acoustic and linguistic models, except for the *WC (word count)* feature of LIWC. For the multimodal model, it is mostly interesting to look at the distributions of the feature categories (acoustic or linguistic) over the feature importance plots. Across all the 4 different models both acoustic and linguistic are found amongst the most predictive features. Looking at the overall distribution, we find this is slightly skewed to the linguistic features, since 60 out of the 80 most important features across the four models are of the linguistic category. However, when only considering the top 5 features of the 4 models, the number of acoustic and linguistic features is very similar (11 versus 9). Furthermore, in 3 models, the strongest feature is an acoustic one. This finding demonstrates that when developing a multimodal model consisting out of acoustic and linguistic features, important features used by the model to make predictions originate from both modalities . This result, in combination with the performance edge of the explainable multimodal model over its unimodal counterparts, is further evidence that studying verbal and non-verbal behaviour cues in combination is a valuable strategy when studying a social interactions. The implications of this finding are discussed in more detail in section 5.1.

Zooming in on the 6 common features and their feature importance plots, similar patterns are identified as in the unimodal models. Specifically, in the multimodal model, the linguistic feature *Clout* showed the same effect as in the unimodal linguistic scenario, with higher *Clout* scores corresponding to lower predictions for the probabilities of investment in models 1, 2, and 3. Similarly, the *Conversation* feature exhibited the same relationship as in the unimodal case, with higher scores corresponding to lower model outcomes and vice versa. The same holds for the *number* feature. The two acoustic features that are common important features in the multimodal model (*F0semitoneFrom27.5Hz pctlrange0-2* and *F2frequency amean*) are also common features in the unimodal acoustic model.

4.5 Cross-domain experiment

Finally, the results of the cross-domain experiment are presented. Here the best performing acoustic, linguistic and multimodal models are tested on the database consisting of the online recordings. The goal of the cross-domain experiment is to test to what extent the performance of the used methodology generalizes to unseen data recorded in a different context. In this experiment, we used the models trained on the four folds of the in-person database. Then, we evaluated the models with all the instances of the online dataset.

Modality	**Model**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Acoustic	Concat VGGish + openSMILE, into single GRU	16.51	15.91	19.05	21.21	18.17
Linguistic	Late fusion: LIWC + (PCA) Longformer	19.15	19.50	21.65	19.62	19.98
Multimodal	Early fusion: best set of each modality	18.59	16.22	16.91	16.84	**17.14**

Table 4.6: Cross-domain results

Comparing the result of this experiment with the results of the previous experiments suggests that the methods can generalize to an online pitch settings to certain extent. While a slight decrease in performance across all three models was observed, the size of this decrease is relatively small and the models continue to exhibit an adequate performance. We note that the some of the patterns observed in the previous experiments recur in the cross-domain experiment. Firstly, when comparing the unimodal acoustic and linguistic model, we again find that the acoustic model is superior. The difference in performance is more pronounced in this cross-domain experiment than previously (as presented in Table 4.2), where these particular acoustic and linguistic models displayed relatively similar scores. Therefore, we observe that the features representing the acoustic modality generalized better than the linguistic features. Another consistent finding in this experiment is the observation that the multimodal model again outperforms the individual unimodal models. The relative increase in performance compared to the acoustic model consisting of concatenated features fed into a single GRU is comparable as in the in-person context, namely a decrease in MAE score of around 1 point.

Model	**Features**	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	**Average MAE**
Acoustic	VGGish	15.48	12.42	11.53	22.19	**15.41**
Linguistic	PCA Longformer	16.03	12.25	10.94	23.21	**15.60**
Acoustic	VGGish + openSMILE single GRU	14.22	17.53	10.38	17.13	**14.82**
Linguistic	LIWC + PCA Longformer late fusion	16.62	12.61	9.56	20.61	**14.85**
MM: Early fusion	openSMILE + LIWC	13.92	13.54	12.56	22.36	**15.59**
MM: Early fusion	VGGish + PCA Longformer	17.17	13.51	5.47	19.47	**13.91**
MM: Late fusion	VGGish + PCA Longformer	15.11	11.73	10.47	22.47	**14.95**

Table 4.7: An overview of the best performing models

Chapter 5

Discussion & Conclusion

In the concluding chapter of this research, a comprehensive examination of the research objectives and outcomes is presented. In section 5.1, the findings reported in Chapter 4 are contextualized and compared with existing literature in the field. Subsequently, in section 5.2, the main research question and associated sub-questions are addressed and answered. The limitations of the current study and recommendations for future research are discussed in section 5.3. Lastly, an overall conclusion is drawn, summarizing the key insights of the study.

5.1 Overview of the results

The aim of this study was to examine the decision-making processes of investors during entrepreneurial pitch interactions by analyzing the acoustic and linguistic characteristics of these pitches. Previous work has shown that early-stage investors rely on two main components when making decisions: factual data on the viability of the project and perceptions of the founding entrepreneur (Huang and Pearce, 2015). Given the scarcity of factual data in this area, the resulting decisions are often characterized by a high degree of uncertainty. Therefore, the perception of the entrepreneur, for a large part based on the social interaction between the investor and entrepreneur, forms a vital part of the decision-making process. For this reason, the social interaction itself can have an impact on the outcome of the decision-making process. Social interactions are shaped by both verbal and nonverbal behavioural cues. In the pitch scenario, such cues may, for example, contribute to the level of trust the investor has in the entrepreneur's capabilities. To address this, the current study proposes a multimodal approach that combines both verbal and nonverbal behavioural cues, specifically acoustic and linguistic features, into a single model to predict the outcomes of these interactions. The underlying philosophy behind this approach is that social interactions are shaped by the interplay of different behavioural cues, and thus studying them in combination results in a more accurate representation of reality.

Acoustic and linguistic models

First the unimodal acoustic and linguistic models consisting of a single feature representation are discussed. Four models were trained on individual acoustic features, with two models utilizing a GRU to model the temporal information in the audio signal and the remaining two models utilizing static representations of openSMILE and VGGish features, respectively. These static representations were trained using Xgboost Regressor. The results of the study indicate that the temporal models performed significantly better than the static models, highlighting the importance of capturing the non-linear relationships and temporal dynamics of acoustic features in the speech signal. This finding is in line with previous research, such as the work of Bae et al. (2016), who emphasize the benefits of learning temporal information when modeling the acoustic modality. Additionally, when comparing the performance of the hand-crafted openSMILE models to the deep VGGish models, it was found that the deep representations achieved superior performance, with this difference being more pronounced in the temporal scenario than in the static scenario. This result is consistent with the study by Sun et al. (2020), who analyzed both the eGeMAPS and VGGish features using a LSTM model and found that the VGGish model outperforms the model based on the openSMILE eGeMAPS features. In the study of Soleymani et al. (2019), who use a comparable methodology as used in this research, the VGGish model also outperforms the eGeMAPS model.

In this study, three models have been trained on individual linguistic features, with one model utilizing the LIWC features and two models utilizing the Longformer representation. When comparing the performance of the hand-crafted model to the deep models, it was found that the best-performing model was based on the deep representation. Additionally, it was observed that the non-PCA-transformed Longformer model exhibited poor performance across all folds of the study. This result can be attributed to the combination of a large embedding size of 768 and a limited sample size. However, upon applying Principal Component Analysis (PCA) to the Longformer features, the results improved substantially. Furthermore, a direct comparison between the results of the PCA-transformed Longformer model and the LIWC model revealed that the Longformer model performed slightly better overall and achieved the strongest performance in 3 of the folds. This finding is consistent with previous research such as the work of Soleymani et al. (2019), who also found that deep embeddings outperform hand-crafted LIWC models. However, in that study, the difference in performance between the two models is more prominent.

In addition, experiments have been conducted to examine the effectiveness of combining hand-crafted and deep representations to create a “hybrid” representation for the unimodal acoustic or linguistic model. Specifically, for the acoustic modality, the openSMILE and VGGish features were integrated in four different models. These models were constructed in various ways, three of which were created by first integrating the feature vectors and then applying a regression module. The effectiveness of these different “hybrid” representation models on the task were then evaluated and compared to the individual models. For all these first three models, an improvement was found with respect to the individual models. The strategy of directly concatenating and feeding the features into a single GRU emerged as the most effective performer. On top of that, the model comprising of two GRUs to encode the information for each feature set also performed better than individual models. The remaining hybrid acoustic model was created by taking the mean of the individual models, comparable to a late fusion approach. However, this model performed worse than the individual VGGish model.

The findings for the first three models are in agreement with the study of Goossens et al. (2022), who use the same data set as used here, to develop a classification model which predicts whether a pitch would be invested in. When openSMILE features were added to the VGGish model, by concatenating the features and feeding this into a single GRU (comparable to the set-up of the best performing acoustic model in this study), the accuracy increased from 66 % to 78 %. Furthermore, in the study by Elbanna et al. (2022), the feature vectors of eGeMAPs and VGGish are also combined in the pre-training phase. Although deep learning based models are effective by themselves, including hand-crafted acoustic features yields a more accurate model. Therefore, in this study we find further evidence for what has been found in earlier works where acoustic feature sets are integrated.

For the linguistic modality, three “hybrid” models have been developed, of which the early fusion and late fusion of LIWC and PCA-Longformer are discussed. Our findings indicate that the combination of hand-crafted and deep features through a single regression model did not result in an improvement in performance, as measured by the mean absolute error (MAE) of 17.51. Upon closer examination of the performance on individual folds, it was found that the combined model did not achieve higher performance than either of the individual models in any fold. This outcome is in contrast to some previous studies that are discussed in Section 2.2.2. For example, both Johnson and Marcellino (2022) and Younus and Qureshi (2020) find that when a hand-crafted model is used as supplement for a deep transformer model, this leads to an increase in performance. However, looking at the results of the late fusion approach of the linguistic features we do find an increase in performance. Here it must be noted that in this case the model does not actually learn a combined feature space. We obtain a better performance using this strategy when we consistently find that one of the models under-predicts an actual score and the other over-predicts this same score.

Across all the unimodal acoustic and linguistic models, we note that the best performing acoustic model is the combined feature representation implemented using a single GRU. Similarly, the strongest linguistic model was identified as the one that employed a combined representation obtained by averaging the output of the individual models. These findings suggest that, in both cases, a combined feature set forms the strongest representation of a modality. The acoustic model slightly outperforms the linguistic model (14.82 compared to 14.85), but this difference is negligible. Furthermore, upon closer examination of the performance on individual folds, it was found that in two cases the acoustic model outperformed the linguistic model and vice versa. Therefore, it is not possible to draw a clear conclusion regarding which modality is the strongest predictor in the context of entrepreneurial pitches.

Multimodal models

The experiments on multimodal models provide three notable insights. Firstly, we studied the effect of creating a multimodal model consisting of both hand-crafted and deep features to represent the two modalities and compared this to a model where only the best performing features of the two modalities are used. As noted in Section 1.2, most previous works explore multiple unimodal models but employ the latter strategy when constructing multimodal models, utilizing a single feature set per modality. The findings presented in this study demonstrate that in both the early fusion and the late fusion architectures the strategy of only using the best performing feature is superior in terms of model performance. This finding lends credibility to the multimodal architectures proposed by authors such as Soleymani et al. (2019) and Tavabi et al. (2020). Here we note that earlier we found using both hand-crafted and deep features in a unimodal context outperforms models made up of individual features. However, this finding is not consistently replicated in the multimodal scenario.

Secondly, the performances of the early fusion models are compared to those of late fusion models. The results revealed that, when evaluating the multimodal model consisting of the best features only, the early fusion model outperforms the late fusion strategy. Conversely, in the setting where all the feature sets are used, the late fusion model demonstrated superior performance. Therefore, we cannot draw a consistent conclusion regarding the most optimal fusion strategy for creating the strongest model. Previous work also did not identify a clear superiority of one strategy over the other. For example, Nojavanasghari et al. (2016) find that late fusion outperforms early fusion, while Dong et al. (2014) find that early fusion is superior. Given these inconclusive findings, we argue that it is recommended to experiment with both early and late fusion strategies when developing a multimodal model in order to assess their impact on the performance. The early fusion approach has the benefit of allowing the model to learn a comprehensive feature space incorporating both verbal and nonverbal behavioral cues. On the other hand, the late fusion approach has the benefit of being relatively straightforward to implement and does not require the modalities to be synchronized, making it applicable in a broader range of contexts

Thirdly, the results for the multimodal models are compared to those of the unimodal models in order to test the hypothesis that multimodal models can potentially outperform unimodal models. The findings indicate that only one of the non-explainable multimodal models, specifically the early fusion of VGGish and PCA-transformed Longformer features, was superior to the best performing unimodal models. Although this result is specific to a single model, it does suggest that utilizing a multimodal approach is a viable methodology for studying investment decision-making based on pitches. Additionally, it highlights the added value of studying different modalities in conjunction, as previously demonstrated in literature discussed in Section 2.2.3. In the previous paragraph we argued that we do not find evidence for the superiority of one of the two fusion strategies. However, considering that the best performing model utilizes early fusion and it is the only multimodal model that outperforms the strongest unimodal models, it can be concluded that the early fusion strategy is the most effective in the context of predicting the likelihood of investment based on acoustic and linguistic features.

Since the early fusion model consisting of the VGGish and PCA-transformed Longformer features is the strongest model found in this study, the performance of this model is compared to the current state-of-the-art performance on the in-person recordings available in this data set. Stoitsas et al. (2022) also used the probability of investment score as a target variable and used a model consisting of different feature sets of cues from several modalities such as facial expressions, head movement and vocal expressions. Here the strongest performing model was trained on head movement and achieved an average MAE of 16.47. Here it must be noted that in this experiment, the data set is split in three folds, with the first two being the same as in this study. However, the third fold was a combination of the third and fourth folds in this experiment. Our strongest model achieved a performance of 13.91, yielding an improvement over the previous state-of-the-art performance provided by Stoitsas et al. (2022).

Explainable models

Explainable models have been developed in order to examine what features play an important role when the models predict the likelihood of investment. For these models, openSMILE and LIWC features are used. First we look at the model performance of these explainable models. Compared to the best performing non-explainable models, we observe a slight decrease in performance. The acoustic explainable model (MAE of 16.82) performs worse than the acoustic model were a GRU is used to model the openSMILE and VGGish features at once. Similarly, the explainable multimodal model is outperformed by all the four other multimodal models. However, we observe that this decrease in performance is rather limited and at the cost of this decrease we do gain a lot in terms of explainability. Furthermore, we again observe that the multimodal model outperforms the unimodal models it contains. So we can conclude that also in the context of explainable models, a multimodal architecture is suitable to predict the likelihood of investment.

We obtained the Shapley feature importance values in order to find repeated important features and to study some correlations between feature values and model predictions. Overall, we found that a relatively large number of features were consistent important features across multiple models. LIWC measures four broader "summary" variables and two of these, *Clout* and *Analytic*, were found to be common important features across the four linguistic models. In the analysis of the openSMILE features, 3 different functionals over the fundamental frequency were common important features. In addition, both the mean and the normalized standard deviation of the loudness were also frequent relevant features. For the SHAP plots of the multimodal models, our primary focus was on examining the distribution of modality categories across the feature importance plots. The results of the analysis demonstrate that both acoustic and linguistic features play an important role in determining the model output. Additionally, this finding further supports the argument that the combination of verbal and non-verbal behavioral cues captures a more comprehensive range of behavior and subsequently yields stronger predictors. The evidence that both modalities are important in determining the model output, strengthens this line of thinking.

Generalizing the models to online recordings

In a cross-domain experiment we tested how well the models that have been developed generalize to the online recorded pitches. Our findings demonstrate a degree of generalizability of the proposed methods to a slightly different context. Firstly, the results indicate that the acoustic features generalized better than the linguistic features. This is consistent with the findings of Soleymani et al. (2019), who also tested their models in a cross-domain experiment and found that non-verbal features, such as vocal features, generalized better. The paper of Kuhn and Sarfati (2021) found that in online settings acoustic features play a more critical role. Therefore, one might expect that the acoustic features have a distinct and different effect in the online setting and thus would not generalize well from an in-person to online context. However, we observe that the acoustic features generalize relatively well. One interpretation of the relatively weak generalizability of the linguistic modality might be that the verbal content changes significantly across the two settings, for example because the presenter is able to have speaker notes on his screen or is less nervous in the online setting. Secondly, we find that also in the cross-domain experiment, the multimodal model is superior to the unimodal models. These observations demonstrate the generalizability of the proposed multimodal architecture and provide further evidence that using a multimodal model leads to better performing models.

5.2 Answering the research questions

Regarding **Sub-question 1**, which stated *to what extent can an acoustic or linguistic unimodal model using either hand-crafted or deep feature representations predict the likelihood of investment*, it should be answered that it is to some extent possible to predict the investment likelihood using these models. The best performing model was trained on the VGGish features, achieving an average MAE of 15.41, followed by the PCA-transformed Longformer features. However, these models did not improve on the results of Stoitsas et al. (2022).

Sub-question 2 continues on the first sub-question and asks *whether combining the hand-crafted and deep representations in a single model affects the model performance.* The present study provides evidence that in certain contexts, the utilization of both hand-crafted and deep features can result in an improvement in model performance. Specifically, for the acoustic modality, it was observed that training a model on an integrated feature vector led to an increase in performance in two cases. Furthermore, for the linguistic modality, it was observed that only fusing the outputs of the individual LIWC and Longformer models had a positive effect on performance.

For **Sub-question 3** we examined *the effect of using different multimodal fusion approaches when using both acoustic and linguistic feature representations in one model.* For this sub-question we find that the result of applying early fusion or late fusion depends on whether all features or only the best features are used. However, only using early fusion, a higher performance was achieved compared to the unimodal models.

Sub-question 4 asked *What explainable acoustic and linguistic features play a role when predicting the likelihood of investment of entrepreneurial pitches?* Using the Shapley values, multiple frequent important features were identified across the acoustic, linguistic and multimodal models. In the multimodal model, both acoustic and linguistic features played an important role when predicting the investment likelihood.

Sub-question 5 stated *how well do the models trained using the in-person recordings of the pitches generalize to online recordings?*. Here the findings indicate that the models generalize to the online recordings to some degree. As can be expected in a cross-domain experiment, a slight decrease in performance was observed. The multimodal model was also in this setting the strongest predictor.

The **Main research question** was the following: *To what extent can the likelihood of investment be predicted from acoustic and linguistic features recorded during an entrepreneurial pitch using deep and hand-crafted representations?* The answer to the overarching question can be stated as: a state-of-the-art performance to predict the likelihood of investment of the pitches in this data set can be achieved, with an average MAE of 13.91. The strongest predictor is an early fusion model consisting of the VGGish and PCA-Longformer features. When both hand-crafted and deep representations are used, we do not observe an increase in performance over the unimodal models consisting of these two representations. The superiority of the multimodal model is further supported by experiments on generalizability and explainable models, which indicate that analyzing both verbal and nonverbal cues together is beneficial in understanding entrepreneurial decision-making.

5.3 Limitations and future research

Although the presented findings show some promising results for the prediction of investor's likelihood of investment, some limitations and context need to be considered.

Firstly, it is difficult to generalize the results, given the relatively small data set, consisting of only 25 videos on which models can be trained, and another 17 (online) pitches that can be used to test the models in a different pitch setting. For this reason, the models presented in this study are based on only 18 to 20 training instances, depending on which session is used as a test set. However, it is worth noting that the startup course where the pitches used in this experiment originate from is taught on an annual basis, thus, the dataset will continue to expand in size over time, which would enable the training of more robust models with more data. Furthermore it would be interesting to look at publicly available data sets containing entrepreneurial pitches, such as the tv shows *Shark Tank* and *Dragons' Den*. Models trained on the dataset used in this thesis could be tested on pitches originating from these sources, and vice versa.

A second limitation, and possible confounding factor, is the fact that the pitches that are analyzed in this study are part of a start-up course as part of a university program and do not take place in a "real" entrepreneurial setting. For this start-up course, the goal is mostly to come up with a business plan and convince investors that this is a well thought out plan. However,the students are not professional entrepreneurs and for the investors there is no actual money at stake. For this reason, it could be argued that in this case the effect of giving a good pitch in terms of acoustic and linguistic features on the probability of investment is larger compared to professional entrepreneurial settings. Furthermore, there was limited to no economic or factual viability data available for the evaluating investors to base their likelihood of investment score on. In reality, this is, besides the evaluation of the founder, an important factor where investors base a decision on. This is another argument for the claim that models using acoustic and linguistic features are less effective when they are applied in a more realistic entrepreneurial context.

Apart from limited viability data in the context of this study, the scope of social interactions analyzed here is narrow in comparison to more realistic entrepreneurial settings. Firstly, the likelihood of investment score that we predict here is based only on the actual pitch segment of the recordings. However, in the data set, the pitches were followed by a Q&A session between the entrepreneur and the investors. This Q&A session has the potential to significantly shape the nature of the social interaction and alter the perception of the founding entrepreneur held by investors. Therefore, future work could aim to also incorporate the Q&A session in the analysis to get a more comprehensive model to represent the reality. It is also worth noting that, in real-world scenarios, social interactions between entrepreneurs and investors are often extensive and involve multiple meetings, thus the scope of the interaction is likely to be broader than what is captured in this study.

Finally, some considerations should be noted regarding the explainable models and the effect of common important features on the outcome of the models. An interpretability tool like SHAP can make predictive machine learning models, like XGBoost Regressor, even more powerful, by uncovering informative associations between features and model outcomes. As noted in an article in the documentation of SHAP (Lundberg, Scott and Lee, 2017), it can be tempting to interpret the values as identifying specific features that can be manipulated by stakeholders in order to alter the predictions of the model. However, using predictive models to guide behaviour is often misleading, since there is an important difference between correlation and causation. While the SHAP tool provides a method to create transparency regarding correlations, it does not indicate causation. Therefore, if the goal is to create guidelines for entrepreneurs to enhance their pitching skills, it is essential to exercise caution when interpreting feature importance plots in this context and it would be necessary to conduct additional causal analysis.

In a broader context, this thesis has demonstrated that using a multimodal analysis approach is a promising direction for studying decision-making in the context of entrepreneurial pitches. Based on this, future research in this area could continue to build upon the methodology proposed in this thesis in order to address some of its limitations. An alternative direction for future research could involve expanding the methodology used in this study. For instance, it would be of interest to incorporate the visual modality, which includes features such as facial expressions, gestures, and head movement into the multimodal analysis. This is because the results of Stoitsas et al. (2022) have indicated that these features also play a role in decision-making, and thus, incorporating them along with the features analyzed in this study could potentially lead to further improvements in performance on the task at hand.

5.4 Conclusion

In this thesis, acoustic and linguistic features extracted from recordings of entrepreneurial pitches have been used to predict the likelihood of investment. Both modalities are represented using hand-crafted and deep features. Deep learning models have been used to model to the temporal dynamics of the inputs. The acoustic and linguistic models have been combined in a single multimodal by applying early and late fusion of the feature representations. Furthermore, explainable models have been trained on the hand-crafted features in order to identify common important features.

The presented findings show promising results for the prediction of investor's likelihood of investment of entrepreneurial pitches using acoustic and linguistic features. State-of-the-art performance has been achieved on this data set using a multimodal model where the best performing features of each modality are integrated using early fusion. In the experiments, deep features generally outperform hand-crafted ones. Further findings suggest that when developing a unimodal model, it is beneficial to represent this modality using both hand-crafted and deep feature sets. It was found that early fusion outperforms late fusion. Across multiple explainable models, consistent features are found to be important predictors. A cross-domain experiment demonstrated that the developed models generalize to a different context to some extent.

Bibliography

Argyle, M. and Kendon, A. (1967). The experimental analysis of social performance. In *Advances in experimental social psychology*, volume 3, pages 55–98. Elsevier.

Aytuğ, O. (2018). Sentiment analysis on twitter based on ensemble of psychological and linguistic feature sets. *Balkan Journal of Electrical and Computer Engineering*, 6(2):69–77.

Babayoff, O. and Shehory, O. (2022). The role of semantics in the success of crowdfunding projects. *Plos one*, 17(2):e0263891.

Bae, S. H., Choi, I. K., and Kim, N. S. (2016). Acoustic scene classification using parallel combination of lstm and cnn. In *DCASE*, pages 11–15.

Balachandra, L., Fischer, K., and Brush, C. (2021). Do (women's) words matter? the influence of gendered language in entrepreneurial pitching. *Journal of Business Venturing Insights*, 15:e00224.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Barrick, M. R., Shaffer, J. A., and DeGrassi, S. W. (2009). What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, 94(6):1394.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150.*

Berner, E., Gomez, G., and Knorringa, P. (2012). 'helping a large number of people become a little less poor': The logic of survival entrepreneurs. *The European Journal of Development Research*, 24(3):382–396.

Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin.*

Carlson, N. A. (2017). Simple acoustic-prosodic models of confidence and likability are associated with long-term funding outcomes for entrepreneurs. In *International Conference on Social Informatics*, pages 3–16. Springer.

Chan, C. R., Pethe, C., and Skiena, S. (2021). Natural language processing versus rule-based text analysis: Comparing bert score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights*, 16:e00276.

Chen, M., Chu, X., and Subbalakshmi, K. (2021). Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 31–38.

Chen, X.-P., Yao, X., and Kotha, S. (2009). Entrepreneur passion and preparedness in business plan presentations: a persuasion analysis of venture capitalists' funding decisions. *Academy of Management journal*, 52(1):199–214.

Cheng, C., Tan, F., Hou, X., and Wei, Z. (2019). Success prediction on crowdfunding with multimodal deep learning. In *IJCAI*, pages 2158–2164.

Clarke, J. and Healey, M. P. (2022). Giving voice to persuasion: Embodiment, the voice and cultural entrepreneurship. In *Advances in Cultural Entrepreneurship*. Emerald Publishing Limited.

Clarke, J. S., Cornelissen, J. P., and Healey, M. P. (2019). Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal*, 62(2):335–360.

Cohen, B. D. and Dean, T. J. (2005). Information asymmetry and investor valuation of ipos: Top management team legitimacy as a capital market signal. *Strategic Management Journal*, 26(7):683–690.

Colombo, M. G. and Grilli, L. (2005). Founders' human capital and the growth of new technology-based firms: A competence-based view. *Research policy*, 34(6):795–816.

Cooper, W. E. and Sorensen, J. M. (2012). *Fundamental frequency in sentence production*. Springer Science & Business Media.

De Winnaar, K. and Scholtz, F. (2019). Entrepreneurial decision-making: new conceptual perspectives. *Management Decision*.

DeGroot, T. and Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of business and psychology*, 24(2):179–192.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dirisam, J. V., Bein, D., and Verma, A. (2021). Predictive analytics of donors in crowdfunding platforms: A case study on donorschoose. org. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0834–0838. IEEE.

D'mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36.

Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., and Štruc, V. (2013). Towards efficient multi-modal emotion recognition. *International Journal of Advanced Robotic Systems*, 10(1):53.

Dong, Y., Gao, S., Tao, K., Liu, J., and Wang, H. (2014). Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications*, 17(1):37–50.

El Mekki, A., Alami, A., Alami, H., Khoumsi, A., and Berrada, I. (2020). Weighted combination of bert and n-gram features for nuanced arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274.

Elbanna, G., Biryukov, A., Scheidwasser-Clow, N., Orlandic, L., Mainar, P., Kegler, M., Beckmann, P., and Cernak, M. (2022). Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load. *arXiv preprint arXiv:2203.16637*.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Galanis, D., Karabetsos, S., Koutsombogera, M., Papageorgiou, H., Esposito, A., and Riviello, M.-T. (2013). Classification of emotional speech units in call centre interactions. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 403–406. IEEE.

Giannakopoulos, T., Spyrou, E., and Perantonis, S. J. (2019). Recognition of urban sound events using deep context-aware feature extractors and handcrafted features. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 184–195. Springer.

Gkoumas, D., Li, Q., Lioma, C., Yu, Y., and Song, D. (2021). What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197.

Glodek, M., Reuter, S., Schels, M., Dietmayer, K., and Schwenker, F. (2013). Kalman filter based classifier fusion for affective state recognition. In *International workshop on multiple classifier systems*, pages 85–94. Springer.

Goossens, I. (2021). Deep learning approach to the influence of vocal behaviour on the decision-making process in the entrepreneurial context. *Tilburg University.*

Goossens, I., J. M. M. L. W. and Onal Ertugrul, I. (2022). o invest or not to invest: Using vocal behavior to predict decisions of investors in an entrepreneurial context. In International Workshop on Human Behavior Understanding (HBU).

Han, W., Jiang, T., Li, Y., Schuller, B., and Ruan, H. (2020). Ordinal learning for emotion recognition in customer service calls. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6494–6498. IEEE.

Haulcy, R. and Glass, J. (2021). Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11:624137.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Huang, L. and Pearce, J. L. (2015). Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative Science Quarterly*, 60(4):634–670.

Jaitly, N. and Hinton, G. (2011). Learning a better representation of speech soundwaves using restricted boltzmann machines. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5887. IEEE.

Jiménez Muñoz, A. J. et al. (2019). Beyond language: a multimodal analysis of success in non-native business-english pitches. *Ibérica, 37.*

Johnson, C. and Marcellino, W. (2022). Bag-of-words algorithms can supplement transformer sequence classification & improve model interpretability.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation.*

Kahn, J. H., Tobin, R. M., Massey, A. E., and Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, 120(2):263–286.

Kaminski, J. C. and Hopp, C. (2020). Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics*, 55(3):627–649.

Keller, E. (2004). The analysis of voice quality in speech processing. *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 54–73.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

Kuhn, N. and Sarfati, G. (2021). Zoomvesting: angel investors' perception of subjective cues in online pitching. *Journal of Entrepreneurship in Emerging Economies*.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., and Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*.

Liebregts, W., Darnihamedani, P., Postma, E., and Atzmueller, M. (2020). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small business economics*, 55(3):589–605.

Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140. IEEE.

Luengo, I., Navas, E., Hernáez, I., and Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In *Ninth European conference on speech communication and technology*. Citeseer.

Lugger, M. and Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–17. IEEE.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Luz, J. S., Oliveira, M. C., Araujo, F. H., and Magalhães, D. M. (2021). Ensemble of handcrafted and deep features for urban sound classification. *Applied Acoustics*, 175:107819.

Marchi, E., Eyben, F., Hagerer, G., and Schuller, B. W. (2016). Real-time tracking of speakers' emotions, states, and traits on mobile platforms. In *INTERSPEECH*, pages 1182–1183.

Martens, M. L., Jennings, J. E., and Jennings, P. D. (2007). Do the stories they tell get them the money they need? the role of entrepreneurial narratives in resource acquisition. *Academy of management journal*, 50(5):1107–1132.

Maxwell, A. L., Jeffrey, S. A., and Lévesque, M. (2011). Business angel early stage decision making. *Journal of Business Venturing*, 26(2):212–225.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality.

Milde, B. and Biemann, C. (2015). Using representation learning and out-of-domain data for a paralinguistic speech task. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Mitteness, C., Sudek, R., and Cardon, M. S. (2012). Angel investor characteristics that determine whether perceived passion leads to higher evaluations of funding potential. *Journal of Business Venturing*, 27(5):592–606.

Morvant, E., Habrard, A., and Ayache, S. (2014). Majority vote of diverse classifiers for late fusion. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, pages 153–162. Springer.

Murnieks, C. Y., Cardon, M. S., Sudek, R., White, T. D., and Brooks, W. T. (2016). Drawn to the fire: The role of passion, tenacity and inspirational leadership in angel investing. *Journal of Business Venturing*, 31(4):468–484.

Newell, A., Shaw, J. C., and Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological review*, 65(3):151.

Niebuhr, O., Skarnitzl, R., and Tylečková, L. (2018). The acoustic fingerprint of a charismatic voice-initial evidence from correlations between long-term spectral features and listener ratings. In *Proceedings of Speech Prosody*, volume 9, pages 359–363.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.-P. (2016). Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

Ossewaarde, R., Jonkers, R., Jalvingh, F., and Bastiaanse, R. (2019). Classification of spontaneous speech of individuals with dementia based on automatic prosody analysis using support vector machines (svm). In *The Thirty-Second International Flairs Conference.*

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Plakal, M. and Ellis, D. (2020). Yamnet. *YAMNet*.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Rao, K. S., Koolagudi, S. G., and Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., Evanini, K., et al. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, pages 2001–2005.

Shah, S. M. S. (2022). *COMPUTATIONAL INFERENCE OF TRUSTWORTHINESS IN SOCIAL FIGURES THROUGH ANALYSIS OF SPEECH ACOUSTIC, TEXTUAL, AND VISUAL SIGNALS*. PhD thesis, RMIT University.

Shane, S. and Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of management review*, 25(1):217–226.

Shepherd, D. A., Williams, T. A., and Patzelt, H. (2015). Thinking about entrepreneurial decision making: Review and research agenda. *Journal of management*, 41(1):11–46.

Shi, J., Yang, K., Xu, W., and Wang, M. (2021). Leveraging deep learning with audio analytics to predict the success of crowdfunding projects. *The Journal of Supercomputing*, 77(7):7833–7853.

Soleymani, M., Stefanov, K., Kang, S.-H., Ondras, J., and Gratch, J. (2019). Multimodal analysis and estimation of intimate self-disclosure. In *2019 International Conference on Multimodal Interaction*, pages 59–68.

Stevenson, A. (2010). *Oxford dictionary of English*. Oxford University Press, USA.

Stoitsas, K., Önal Ertuğrul, I., Liebregts, W., and Jung, M. M. (2022). Predicting evaluations of entrepreneurial pitches based on multimodal nonverbal behavioral cues and self-reported characteristics. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, pages 121–126.

Strategy, T. O. J. (1998). Fostering entrepreneurship. Technical report, OECD.

Sun, L., Lian, Z., Tao, J., Liu, B., and Niu, M. (2020). Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 27–34.

Székely, E., Kane, J., Scherer, S., Gobl, C., and Carson-Berndsen, J. (2012). Detecting a targeted voice style in an audiobook using voice quality features. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4596. IEEE.

Tavabi, L., Stefanov, K., Zhang, L., Borsari, B., Woolley, J. D., Scherer, S., and Soleymani, M. (2020). Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 406–413.

Toto, E., Tlachac, M., and Rundensteiner, E. A. (2021). Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4145–4154.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Ucbasaran, D. (2008). The fine 'science'of entrepreneurial decision-making. *Journal of Management Studies*, 45(1):221–237.

Venkata Raju, K. and Sridhar, M. (2020). Based sentiment prediction of rating using natural language processing sentence-level sentiment analysis with bag-of-words approach. In *First International Conference on Sustainable Technologies for Computational Intelligence*, pages 807–821. Springer.

Vinciarelli, A., Salamin, H., and Pantic, M. (2009). Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 42–49. IEEE.

Wadeson, N. (2006). Cognitive aspects of entrepreneurship: decision-making and attitudes to risk. *The Oxford handbook of entrepreneurship.*

Wagner, J., Lingenfelser, F., and André, E. (2013). Using phonetic patterns for detecting social cues in natural conversations. In *INTERSPEECH*, pages 168–172.

Wang, X., Lu, S., Li, X., Khamitov, M., and Bendle, N. (2021a). Audio mining: the role of vocal tone in persuasion. *Journal of Consumer Research*, 48(2):189–211.

Wang, X., Zhao, S., and Wang, Y. (2021b). Bimodal emotion recognition for the patients with depression. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 40–43. IEEE.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2012). Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Yang, Y., Fairbairn, C., and Cohn, J. F. (2012). Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150.

Yin, B., Ambikairajah, E., and Chen, F. (2006). Combining cepstral and prosodic features in language identification. In *18th international conference on pattern recognition (ICPR'06)*, volume 4, pages 254–257. IEEE.

Younus, A. and Qureshi, M. A. (2020). Combining bert with contextual linguistic features for identification of propaganda spans in news articles. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5864–5866. IEEE.

Zhang, S. (2008). Emotion recognition in chinese natural speech by combining prosody and voice quality features. In *International Symposium on Neural Networks*, pages 457–464. Springer.

Zhang, S. X. and Cueto, J. (2017). The study of bias in entrepreneurship. *Entrepreneurship theory and Practice*, 41(3):419–454.

Zhang, X., Lyu, H., and Luo, J. (2021a). What contributes to a crowdfunding campaign's success? evidence and analyses from gofundme data. *Journal of Social Computing*, 2(2):183–192.

Zhang, Y., Sidibé, D., Morel, O., and Mériaudeau, F. (2021b). Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042.

Zhou, G. (2021). Donation-based crowdfunding title classification based on bert+ cnn. In *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing*, pages 291–296.

Appendix

In-person recordings

Fold 1	**Fold 2**	**Fold 3**	**Fold 4**
PREA	BubblePop	Chattern	Ar-T-ficial
SoccerAcademy	FitPoint	Choos3Wisely	HoodFood
Whitebox	FLIPR	FindIT	LockUp
YoungBoosters	HOTIDY	SmArt	Peech
Ziggurat	LittleSister	StudentFood	Recipe-Me
	RecognEyes	TAIste	Salix
	SOLON	WAIste	

Table 1: Distribution of the in-person recorded pitches over the folds

Online recordings

APlaceForNow	CommunicAid	Locify	SOLOPE
Adverlyze	CourseCompass	OutBusy	Shooze
BestCarFit	EasyTrip	Pawshake	ThriftShopBox
BookFlixDelivery	Jellefish	QTag	VintageSurprise
Calculytics			

Table 2: The online recorded pitches used for the cross-domain experiment