CORK INSTITUTE OF TECHNOLOGY

A bi-modal approach to emotion detection, using verbal and non-verbal components of speech

by

Zachary Dair

This thesis has been submitted in partial fulfillment for the degree of Bachelor of Science in Software Development

in the
Faculty of Engineering and Science
Department of Computer Science

December 2020

Declaration of Authorship

I, Zachary Dair, declare that this thesis titled, 'A bi-modal approach to emotion detection, using verbal and non-verbal components of speech' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for an undergraduate degree at Cork Institute of Technology.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Cork Institute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

CORK INSTITUTE OF TECHNOLOGY

Abstract

Faculty of Engineering and Science
Department of Computer Science

Bachelor of Science

by Zachary Dair

Human-like communication and understanding enables for a higher range of personalization from AI driven assistants, and provides a more dynamic and fluid interaction.

In this dissertation, the current state of natural language understanding is outlined in the emotion detection domain. Accomplished by exploring current work in the area of emotional and sentiment analysis, and the natural language processing methods that make the current work possible.

The inclusion of multiple modalities such as affective prosody and the text of transcribed speech, to see specifically how the increase in modalities allows for a deeper analysis of language communication in terms of emotion.

Leveraging natural language processing methodologies, machine and deep learning classification models, in order to accurately classify emotion as it is found in speech samples, in a hybrid analysis of both the verbal and non-verbal components of human communication.

The goal of this workflow is to determine emotion and contextual cues of the emotion exhibted. Which when integrated with a virtual assistant enable a deeper level of understanding of the user. Therefore providing a greater suite of personalization options, and an increased level of fluidity in human to computer interaction.

Acknowledgements

I would like to express my gratitude to **Ruairi O'Reilly** who provided guidance and excellent feedback throughout the research phase.

I would also like to thank **Ryan Donovan** who provided knowledgeable insights into the psychology of emotion, valuable writing information and excellent project feedback.

In addition a huge thank you to all of my lecturers throughout my time in Cork Institute of Technology, without whom I would not have the knowledge required for this project.

And finally, thank you to all of my family and friends for supporting me through the years.

Contents

Declaration of Authorship **i**

Abstract **ii**

Acknowledgements **iii**

List of Figures **vii**

List of Tables **viii**

Abbreviations **ix**

1 Introduction **1**
- 1.1 Research Area . . . 1
- 1.2 Problem: Creating fluid HCI . . . 2
- 1.3 Problem Abstraction . . . 3
- 1.4 High Level Solution . . . 4
- 1.5 Structure . . . 5

2 Background **6**
- 2.1 Research Fields . . . 6
  - 2.1.1 Emotion Detection . . . 6
  - 2.1.2 Speech Recognition . . . 7
  - 2.1.3 Natural Language Processing . . . 7
  - 2.1.4 Machine Learning . . . 8
- 2.2 A review of emotion detection utilising speech recognition, NLP and ML . 9
  - 2.2.1 Deep Learning Sentiment and Emotion Analysis . . . 9
  - 2.2.2 Theoretic Natural Language Processing . . . 11
  - 2.2.3 Applied NLP for sentiment analysis . . . 12

3 Methodology **14**
- 3.1 Problem Definition . . . 14
- 3.2 Objectives . . . 16
  - 3.2.1 Affective Prosody Analysis . . . 16
  - 3.2.2 Transcription . . . 16

3.2.3 Text Analysis . . . 16
3.2.4 Emotional Recognition . . . 17
3.2.5 Contextual Analysis . . . 17
3.3 Functional Requirements . . . 17
3.4 Non-Functional Requirements . . . 18

4 Implementation and Evaluation **19**
4.1 Architecture . . . 19
4.1.1 Data Gathering . . . 19
4.1.2 Affective Prosody Pre-Processing . . . 20
4.1.3 Affective Prosody Emotion Analysis . . . 20
4.1.4 Affective Prosody Semantic Analysis . . . 20
4.1.5 Transcription . . . 20
4.1.6 Text Pre-Processing . . . 21
4.1.7 Text Emotion Analysis . . . 21
4.1.8 Text Semantic Analysis . . . 21
4.1.9 Classification Convergence . . . 22
4.2 Risk Assessment . . . 22
4.2.1 Risk one - Inconsistent Transcription . . . 22
4.2.2 Risk two - Semantics lost in Transcription . . . 22
4.2.3 Risk three - Inconsistent Emotion Classification . . . 23
4.2.4 Risk four - Low Performance in Emotion Detection . . . 23
4.2.5 Risk five - Unidentifiable Semantic Relationships . . . 24
4.2.6 Risk six - Noisy Audio Samples . . . 24
4.3 Methodology . . . 25
4.3.1 Extensive research and analysis . . . 25
4.3.2 Applied research and learning . . . 25
4.3.3 Practical Evaluation . . . 25
4.4 Implementation Plan Schedule . . . 26
4.4.1 January 11th - February 1st . . . 26
4.4.2 February 1st - February 8th . . . 26
4.4.3 February 8th - February 22nd . . . 26
4.4.4 February 22nd - 8th March . . . 26
4.4.5 8th March - 29th March . . . 27
4.4.6 29th March - 5th April . . . 27
4.4.7 5th April - 5th May . . . 27
4.5 Evaluation . . . 27
4.5.1 Initial Evaluation . . . 28
4.5.2 Affective Prosody Evaluation . . . 28
4.5.3 Transcription Evaluation . . . 28
4.5.4 Semantic Analysis Evaluation . . . 29
4.5.5 Emotional Recognition Evaluation . . . 29
4.5.6 Contextual and Final Analysis Evaluation . . . 29
4.6 Prototype . . . 30
4.6.1 Pre-Processing . . . 30
4.6.2 Extracting Relevant Features . . . 30
4.6.3 Sentiment Classification . . . 31

5 Conclusions **33**
5.1 Discussion 33
5.2 Conclusion 34
5.2.1 Complications in VDA 34
5.2.2 A Bi-Modal Workflow for Emotion Detection 35
5.2.3 Implementing the Workflow 35
5.3 Future Work 36

Bibliography **37**

A Code Snippets **38**

List of Figures

3.1 A workflow for emotion detection using a bi-modal approach 15

4.1 Pre-Processing Procedure in Python 30

4.2 Extracting Relevant Features Procedure in Python 31

4.3 Likelihoods Calculation Procedure in Python 32

4.4 Naive Bayesian Classification in Python 32

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
VDA	**V**irtual **D**igital **A**ssistant
HCI	**H**uman **C**omputer **I**nteraction
BET	**B**asic **E**motion **T**heory
NLP	**N**atural **L**anguage **P**rocessing
BOW	**B**ag **O**f **W**ords
ML	**M**achine **L**earning
NLU	**N**atural **L**anguage **U**nderstanding
TF-IDF	**T**erm **F**requency **I**nverse **D**ocument **F**requency

For/Dedicated to/To my…

Chapter 1

Introduction

1.1 Research Area

Communication is when two entities express and exchange information to one another through a given medium, the medium can vary from mutually understood signs, symbols, or languages. In the context of human to human or human to virtual digital assistant (**VDA**) communication the main medium used is language. When deconstructing how language is used, verbal and non-verbal components can be identified. These components are relied upon to transfer information, from this an inherent relationship between communication and linguistics can be identified.

This relationship can be examined by analyzing the verbal and non-verbal components in a domain-specific area, and that is the area of emotion detection, where the spoken content and the affective prosody can be considered as the verbal and non-verbal features respectively.

Identifying emotion and the context to which it belongs provides deep insights into human psychology. By using emotion detection as a specific research area, it provides an information-rich and categorized domain to analyze.

Emotions can provide a wide range of information about the person. When emotion detection is used in conjunction with digital services it enables a richer suite of personalization options and more suited recommendations to the end-user.

In recent years, humans are increasingly relying on computers to manage a wide range of tasks in their lives, these artificial intelligence (**AI**) powered systems require human communication to provide their task. There are 4.2 billion digital voice assistants being used in devices around the world in 2020 [1], this shows the global scale of **VDA** usage.

VDA incorporates multiple areas of computer science developed to process inbound communication from the end-user with the aim of completing a task, the goals of these tasks can vary widely from playing a song or turning on a light, to explaining a recipe or scientific concept to the end-user.

VDA has been a highly active field of computer science, with corporations and researchers around the world striving to improve and expand capabilities. However, due to the communicative nature of **VDA** a significant problem can be identified.

Human to Computer Interaction (**HCI**) is focused on detecting, categorizing, and processing verbal communication. However, the non-verbal components are critical for deeper understanding. Therefore, by neglecting these non-verbal components current **HCI** language tends to be more restricted and static.

A **VDA** which utilizes and comprehends **HCI** that extracts features related to these non-verbal components, will be better positioned to understand communication between humans.

1.2 Problem: Creating fluid HCI

Identifying context and inferring information can be an issue in **VDA**'s requiring the user to simplify or parse their query into a format that explicitly defines everything the **VDA** needs to know. This compromise represents a challenge to fluid Human to Computer communication.

Conveying tone and emotion through succinctly written word is a prevalent issue. With our new reliance on digital communication, this issue is also occurring in human to **VDA** interactions. Conveying emotion through messages without being explicit, can often lead to misunderstandings, and if these misunderstandings can take place with humans, it makes them significantly more common to occur with computers.

There are elements such as tone, pitch, underlying emotion, etc that are significant when understanding the meaning in communication. The absence of understanding these non-verbal components in **VDA** systems can lead to an increased likelihood that the meaning will be misidentified.

1.3 Problem Abstraction

The purpose behind analyzing emotion detection in this context is to provide a domain-specific analysis of the relationship between communication and linguistics. The understanding of emotion can provide insights into both the context and the intent of an individual when they are expressing an opinion, or simply conversing.

Emotion analysis provides information on how individuals perceive a given topic, there is significant overlap between sentiment analysis and emotional analysis, the latter providing arguably a more psychological analysis, whereas the former provides a general positive or negative classification of the individual's views.

Typical emotional analysis is conducted using data in text form, which can be extracted from social media [2], or news articles, books, etc. From analyzing the language choice in text significant insights can be gained into the overall emotions exhibited. However, text analysis methods have a significant problem, which is that they lose an entire layer of information, the non-verbal components, by analyzing solely text data.

When humans converse, there is the spoken content, but also how that content is spoken, the pitch, tone, word emphasis, pauses, all of these features allow for information to be extracted. This information can provide vital insights into the overall sentiment and emotions felt by the individual about the subject matter, without taking into account the actual words spoken.

In an attempt to provide a deeper emotional analysis two modalities are analyzed. These are text and affective prosody (non-verbal components). Analysed to leverage the subtle inferences and additional information of spoken content, with an aim of extracting the contextual information from both of these modalities, providing an end model capable of emotional detection and analysis on text and also on speech.

Another problem found in typical emotion analysis is the range of emotions in question. There is significant research into what may be considered the true set of human emotions that stem from biological makeup and evolution, Basic Emotion Theory [3], defines a subset of emotions: Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise. However, it is a source of conjecture between philosophers whether this subset of emotions is truly universal. In this analysis, the universality of the emotions isn't of great importance.

Departing from the psychological, in some cases the emotion range can be too narrow, where there are too few emotions in question, this can cause the machine learning detection model to miss-classify emotions due to several being bundled together, or in the opposite case some models look at a wide range of emotions, and this leads to a decrease in the overall performance of the model.

Context is also an important part of human conversation, and this subtle but key feature is easily picked up on by humans but can be lost in the concrete processing of machine learning models. Contextual cues can provide subtle inferences that allow a greater understanding of the subject matter, in the case of emotional detection, context can be used to identify subject matter, hidden/inferred information or it can be used to determine subtle relationships between the exhibited emotion and the text content. Using a bi-modal approach, this facilitates performing contextual analysis on the affective prosody and text data, thus allowing for context between non-verbal and verbal components to be determined.

This is achieved by converting speech into its text content using transcription. The text data will then be analyzed using traditional methods, such as **NLP**, sentiment analysis, etc. The audio will then be processed using a machine learning model, trained to determine emotion from affective prosody.

1.4 High Level Solution

The data to be analyzed and utilized in the creation and training of the machine learning models is collected from an experimental study conducted at Cork Institute of Technology. This data contains both semantic and audio information about the user's reactions to emotionally provocative stimuli. Therefore, enables semantic analysis using **NLP** and analysis of affective prosody.

By using speech, the aim is to take advantage of the presence of affective prosody, which is the pitch, tone, pauses, etc, which become additional features to be analyzed in comparison to traditional text-based emotion detection methods.

This enables a two-sided approach where the subtle and indirect information cues of non-verbal data are available for analysis. Which can then be compared with the analysis of the concrete verbal data from the transcribed audio.

The speech data is first processed in order to retrieve the affective prosody data and then transcribed into text to be used with natural language processing methods.

Natural language processing methods, are a subset of machine learning that focuses on theme extraction, sentiment analysis, and much more. These methods provide an important process in adapting the data to be able to be used with machine learning and deep learning models.

Using an **NLP** method called N-grams, which allows the splitting of text data into segments. Typically used to create a summary of the data. Similarly, an additional

NLP method is text vectorization used in order to convert the data summary into a form usable by machine learning models.

There may (and will be) many more processing steps before the data is combined with the affective prosody data and labeled and inputted into a machine learning model in order to create the final classifications.

The end goal is to create a workflow employing machine learning classifiers capable of identifying emotion and context in a spoken sentence, by analysing the non-verbal and verbal components.

1.5 Structure

The rest of this dissertation is organized as follows. Chapter two outlines prior research and information pertinent to this topic, and provides an overview of systems and methods used within this bi-modal emotion detection workflow. Chapter three defines the aforementioned workflow enumerating the core components, functional and non-functional requirements. Chapter four provides an overview of the implementation for this workflow. This overview contains an implementation procedure, a risk assessment, an implementation plan, progress evaluation, and prototype information. Chapter five concludes this dissertation which a summary of the findings from the research phase and outlines potential additional features.

Chapter 2

Background

The emotion detection workflow outlined in this dissertation can be disassembled into the following fields: Emotion Detection, Speech Recognition, **NLP** and Machine Learning.

This chapter aims to outline the purpose behind using emotion detection, the relationship between linguistics and communication in speech recognition, the power of **NLP** in this domain, and finally machine learning as computational processing means.

2.1 Research Fields

2.1.1 Emotion Detection

Expressing emotion through language provides a medium to convey a psychological feeling. Therefore enabling proper understanding of the individual.

Emotion detection is the act of analyzing an individual in order to determine their emotional state. In emotion detection, there is conjecture as to what emotions should be detected.

Basic Emotion Theory, which considers the following as universal basic human emotions: Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise. **BET** [3] has dominated the affective sciences since 1972. However, the range of emotions analyzed has been criticized due to cross-cultural differences in emotion, hence disputing the universality claim.

In this emotion detection workflow, the goal is to create an emotion detection procedure implementable by **VDA** systems. Enabling a greater level of understanding of the user. The emotions to be detected follow **BET** due to its longevity and position in the affective sciences.

2.1.2 Speech Recognition

Speech Recognition is a subfield of both computer science and computation linguistics. This process typically involves converting human communication into text.

The communication between a human and a **VDA** system is the same as human to human communication. The difference however occurs not in the communication, but in the understanding process.

When analyzing the human conversation process, the understanding of spoken language can be analyzed in two parts, the verbal component, and the non-verbal component. The verbal component contains the meaning of the phrase, such as nouns, verbs, etc. And the non-verbal component contains the tone, pitch, pauses, etc. Therefore human to human communication can be seen analyzing an additional modality in comparison to **VDA** systems, that typically rely solely on the verbal components.

Further research into human communication, specifically into our ancestors shows that non-verbal components were used to communicate before words. Therefore displaying the importance of non-verbal components in communication.

Therefore, in order to provide a similar manner of communication and understanding, **VDA** systems must take into account, both verbal and non-verbal components of communication.

2.1.3 Natural Language Processing

Natural language processing is a tool to provide a bridge between the static and rigid language processing and understating found in computers, and the fluid, subtle, and dynamic nature of human speech.

NLP is typically focused on text translation, theme extraction, sentiment analysis, such methods are used in a wide range of applications, the most prominent being virtual assistants such as Amazon©Alexa.

NLP has become a significant area of study in recent years, which can in part be attributed to the vast increase in the usage of human to **AI** communication in devices such as Alexa.

A widely used **NLP** method is text vectorization, this is a general term for the process of converting a series of words into a numerical representation. This vector can then be used in a machine learning algorithm for classification, whereas before vectorization the algorithm may not have been able to process the data, or it may produce undesired results. There is a multitude of different ways to achieve text vectorization, such as the Binary Term Frequency, Bag of Words (**BOW**) Term Frequency, and more.

Another commonly used **NLP** method is to split the corpus (text) into segments, these segments can vary in size from one (mono-gram), two (bi-gram), three (tri-gram), or more. This method is known as N-grams and can be applied to both words, or individual characters. The N-gram method can be used for sentiment analysis or theme extraction.

2.1.4 Machine Learning

Machine learning is a sub-field of Artificial Intelligence and a prominent area of study.

The power of machine learning enables systems that once created, are able to learn and adapt without explicit instruction.

Machine learning models are created by inputting data into a specified algorithm this is known as training. The training data is typically a large portion of the training dataset. Once trained, the model is then evaluated by prediction classification on unseen data from the training dataset, this is known as test data. The accuracy of the model is then calculated, a model with poor accuracy may need retraining with different data or a different algorithm. If the model exhibits a high accuracy it can then be used to create accurate predictions on new unseen data.

There are two methods for training machine learning models, supervised and unsupervised.

A supervised model is trained using data that contains a classification label. An example of this can be seen in sentiment analysis of a phrase. The dataset will include the phrase and the positive or negative sentiment exhibited in that phrase.

An unsupervised model is trained using data that solely contains features. An example of this can again be seen in sentiment analysis of a phrase. The dataset will now only include the phrase, ie the features, and no sentiment label.

The features in a dataset represent the individual characteristics, or properties being examined, and the classification is a value that defines the class of a given entry in the dataset.

There are important considerations to take into account before choosing a machine learning algorithm.

Firstly the classification type must be identified. The type of classification can be considered binary or non-binary (multi-class). Binary classification is where the classification is true or false. And non-binary or multi-class is where there are more than two possible classifications.

In the case of emotion detection, an example of a binary classification would be: does the phrase contain the presence of a specific emotion such as anger? The answer to this question is True, or False, thus rendering it a binary problem. An example of a multi-classification in the same domain would be: Which emotions or traits are shown in the text? The answer to this question could be a list of multiple emotions exhibited.

There is also a further classification method, where the classification value is a continuous numerical value, these classifications are most commonly found alongside Linear Regression algorithms.

Linear Regression algorithms are based on statistical models. These models are used for predictions, they identify a classification value from the relationship between the input data and the target variable.

Secondly, the size of features is important, as certain algorithms such as Decision Trees perform poorly with a larger number of features, due to the number of branches required. If the dataset contains data-points with a large number of features an algorithm such as Linear Discriminant Analysis may provide better results.

2.2 A review of emotion detection utilising speech recognition, NLP and ML

2.2.1 Deep Learning Sentiment and Emotion Analysis

Deep learning is a method found in machine learning, where in place of typical classification algorithms we use artificial neural networks inspired by the human brain.

In [2] the author explores the usage of sentiment analysis and deep learning for emotion detection on Covid-19 related tweets.

Their study focused primarily on six emotions joy, surprise, sadness, fear, anger, and disgust with the goal of analyzing and describing the polarity of opinion shown by emotion to the global Covid-19 pandemic between different cultures.

A significant portion of the paper outlined the data gathering process utilized, and the considerations in that area. Such as only taking tweets from February to April, as they believed that post April the general population would have acclimated to the situation, thus there would be a diminished display of emotion in the area.

Some data pre-processing methods that were outlined that have relevance to the problem area. Such as tokenization of text and the removal of stop words.

In [2] the reliance on Twitter data and emotion icons (emoticons) meant their data required different levels of 'cleaning', and extraction in order to be relevant. Therefore relying on pre-processing methods such as emoticon extraction, non-ASCII character replacement, removal of mentions, and colons from the tweet text. However as speech to text transcription generates a text corpus, these issues will not be encountered.

The author describes their usage of a personally created dataset using their own data gathering and processing methods, but also the usage of three external datasets such as a Twitter dataset from Kaggle, Sentiment140 dataset from Stanford, and an Emotional Tweets dataset.

The workflow outlined in this dissertation enables the potential to leverage these datasets, and text processing methods in order to create a large training corpus. Despite the text not stemming from speech initially, it may provide the machine learning model with a comprehensive and diverse labeled dataset. Which can be used to further improve the emotion and sentiment analysis on the text modality of the data.

Sentiment analysis differs from emotional classification, as sentiment analysis classifications generally refer to positive, negative, and neutral polarities, whereas in the case of emotion detection the total range of emotions becomes the classifications.

The authors of [2], merge multiple levels of analysis in order to discern the true emotion exhibited in each tweet. Therefore in each level, a classification can be identified. The first level is sentiment analysis using a deep learning model trained on the sentiment140 dataset, this classifier is used for strict classification of a tweet as exhibiting a positive or negative sentiment.

Once a given tweet exhibits a positive sentiment, it is passed into their second level classifier, this is the positive emotion recognition classification. This second level classifier has been trained using the Emotional Tweets dataset, with the goal of classifying a given tweet as one of the following: joy or surprise.

The final level is using a classifier that classifies tweets exhibiting a negative polarity into a negative emotion category.

The structure outlined in [2] presents an interesting basis of an emotion detection classification method. By restructuring the classification problem from a multi-classification to a series of binary classifications, enables any given text to be analyzed for the presence or absence of each emotion. Therefore enabling each emotion present to be identified in the case of a text exhibiting multiple emotions.

2.2.2 Theoretic Natural Language Processing

Natural language processing as previously defined is comprised of a multitude of methods that can be used for theme extraction, sentiment analysis, pre-processing for machine learning methods, and much more.

[4] contains an outline of several aspects of **NLP** and **NLU**

The authors outline the differences between formal languages and natural languages. Formal languages such as Python or Java, have strict semantic rules and these languages have been precisely defined for computation tasks. This provides a stark contrast with natural languages such as English or Spanish. As natural languages cannot be defined in totality as a specific set of sentences, natural languages also contain a high level of ambiguity, from grammatical rules to word definitions. For a natural language model to be accurate, there is a need to use probability distributions to create approximations of meaning.

One of the first **NLP** methods explored by the authors is N-gram character models, these models are designed by splitting the text (AKA Corpus) into sequences, these sequences vary in length (n), 1-gram (uni-gram), 2-gram (bi-gram), 3-gram(tri-gram) we then calculate the probability distribution of these n-gram models.

N-gram character models are well suited to language detection, genre classification, named-entity recognition. However, an issue occurs when a sequence of characters does not appear in the training corpus. Therefore the n-gram model associates a probability of zero to this character sequence.

This issue could be seen as a design flaw by the project author. Due to the dataset in use not being suitably varied. Therefore failing to give an accurate representation of the language in question. An example of this could be the character sequence ’ht’ which is an uncommon starting character sequence for a word in the English language but it does occur, ’HTTP’ for example, in order to counter this issue we can use various smoothing methods.

Smoothing n-gram models can be done using Laplace smoothing, as done in Naive Bayesian classification. Laplace smoothing consists of adding a small non-zero probability to the rarely or non-occurring sequences in the training corpus and decreasing the probabilities of other sequences to ensure the sum of all probabilities is one.

Considering the outlined problem, character N-gram models may not produce the desired results, as the corpus is being split into segments of characters, which then loses all relationships that may have between words. Fortunately, N-gram models using words can counteract the aforementioned issue. These word models provide the benefit of immediate relationships between words, which in the problem space is ideal for identifying context or emotional cues.

When using N-gram models several factors need to be considered. Such as the greatly increased size of vocabulary (from a set of characters to a large set of words). Additionally, an issue arises when new words being invented.

As explored by an experiment of n-gram word models in [4], their findings were that using a bag of words, or uni-gram model the notion of order and relation of words was completely lost. However, using both the bi and tri-gram models it was persevered and showed a higher approximation of the overall context and meaning of the phrases.

An important finding was also that very common words, known as stop words can be removed from the training corpus which has little detriment to the accuracy and processing of the text, but increases performance greatly.

2.2.3 Applied NLP for sentiment analysis

As previously mentioned, sentiment analysis is conducted to extract an overall positive or negative sentiment from data. Typically conducted using semantic analysis.

In [5], the authors have outlined a series of processes that employ **NLP** methods in order prepare and ultimately classify text as either positive or negative polarities known as sentiment analysis.

Text data from standard datasets typically requires procedures. Such procedures involve data cleaning and processing to remove any 'noise'. Examples of 'noise' in data can be corrupted or distorted data, data containing additional symbols or spacing etc.

These **NLP** methods are bundled into a 'pipeline' which the crude text data found in the Tweet is then sent through in order to be cleaned into a use-able format, and also removes any redundant information.

In the case of the bi-modal emotion detection workflow, this 'pipeline' of **NLP** methods will be structured differently and may contain different methods. The data used in [5] is from Twitter. Therefore the text may include hyperlinks, emoticons, symbols, mentions, etc. All information that will not be found in the text that has been transcribed from speech.

However, a number of pre-processing methods will be employed in the workflow. Such as tokenization, removal of stopwords, text modeling in the form of **BOW** and **TF-IDF** Models.

Tokenization is the act of splitting a given character sequence or sequence of words into pieces, also known as tokens, during this process punctuation may also be removed.

Stop words are considered the most common words in a language. There is however no definite list of stop words, as they may vary in the problem area, and depending on the tool used. By removing stop words it also reduces redundant words that have little to no effect on the understanding of the sentence.

Bag of Word models, can also be considered as uni-gram models, where the sequence of words is split one by one, and a count is stored for the occurrence of each word, a major downside in this model type, is the loss of inferred information such as relationships or order.

Term frequency-inverse document frequency or **TF-IDF** is the statistic that is intended to reflect the importance of a given word in a corpus, it is a method of weighting words to determine the most important, and it can be used alongside a lexicon of emotional key words, if the given word in the corpus also appears in the lexicon then it can be given greater importance than other words.

These methods all aim to reduce the redundancies and highlight important information. Using these methods, along with text vectorization (the process of encoding each word of a sequence into a numerical form) enables the processing of text sequences in a machine learning algorithm to achieve a classification based on sentiment (negative or positive), or even further analysis for emotional classification.

Chapter 3

Methodology

A significant body of work has been dedicated to the understanding of written communication, using **NLP** and traditional text classification machine learning models. The bi-modal analysis of affective prosody alongside **NLP** in the area of emotional language in communication provides a workflow that could be adopted by **VDA** systems to enable a wider range of personalization and increase communication fluidity.

3.1 Problem Definition

This methodology adopted comprises several key components, each of which aims to provide a workflow that could be employed by a **VDA** to provide more fluid and personal human to computer interaction. The following can be considered as the core sections of the project: affective prosody analysis, transcription, semantic analysis, and emotional recognition.

In order to build a workflow enabling emotion to be detected from both audio and text, there must be two procedures, one for each modality. These procedures involve the pre-processing, analysis, and classification functions required to construct the overall analysis. These two procedures converge at the final stage to give an overall analysis of the input data.

Firstly the input data must be prepared, by converting the initial audio data into the two modalities, the initial audio and the transcribed text. To achieve this the software takes an audio file as an input, and transcribes the spoken text content, this transcribed content represents the text modality. This splitting enables a separate analysis of each modality.

Secondly the creation of a machine learning model that returns an emotional classification based on non-verbal components. The model requires an audio file as input, which when examined, will extract the relevant features from the audio, such as tone, pauses, etc, these components will then be used to create the classification, dictating the presence of emotion in the audio stream, this stage excludes the meaning behind the spoken content, and relies solely on non-verbal components.

Thirdly the remaining modality must be examined, the verbal components. These verbal components are created by the transcription of the audio sample. The goal of this stage is to carry out semantic analysis and emotional classification, there is a multitude of ways this may be accomplished, this approach involves using **NLP** and semantic analysis methods to extract relevant features which will then be passed through a series of binary classification machine learning models trained to detect the presence or absence of each emotion found in **BET** [3]. The end result of this stage will be a series of classifications dictating the presence of emotions from the text content.

Subsequently, the objective is to analyze the contextual information that is a bi-product of the classifications of each modality, this stage will enable the relationships between the affective prosody and the text content to be identified. These relationships will highlight the inferred/obscure contextual information present alongside the emotional language.

Finally, the aforementioned workflow will be packaged into an application that provides a start to finish approach to emotion detection. This will be achieved by encapsulating each stage of the workflow, each stage then linked into one overall system, that will allow an end-user to input an audio file or a series of files which will then pass through each process, and finally result in an emotion detection classification, and an analysis of the contextual information inferred from the analysis of both modalities.

The full workflow structure can be seen in Figure 3.1

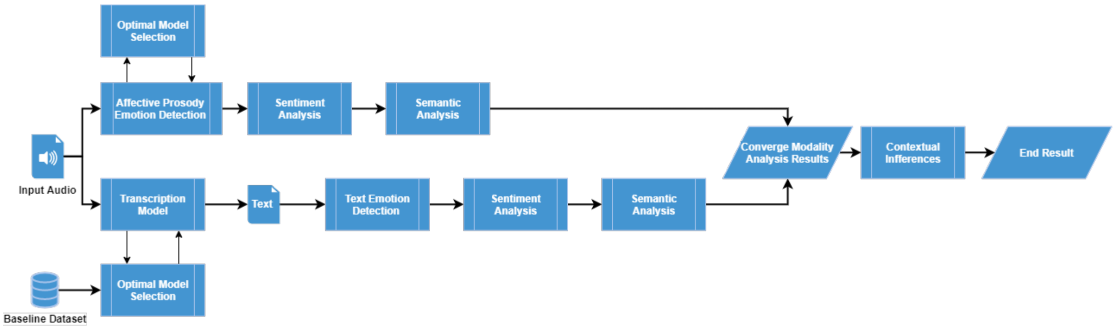


FIGURE 3.1: A workflow for emotion detection using a bi-modal approach

3.2 Objectives

The previously outlined methodology components can be enumerated into several broad objectives that provide actionable targets to evaluate the progress made in constructing this bi-modal workflow for emotion detection.

3.2.1 Affective Prosody Analysis

Affective prosody analysis is a key aspect of the workflow, achieving this objective is vital. This analysis is required to gain insights into the audio modality, these insights can be gained from various analysis methods which are as follows:

- Using semantic analysis on the audio provides insight into the structure of the data.
- Conducting sentiment analysis on the audio provides an overall positive or negative sentiment classification.

3.2.2 Transcription

Transcription is another key aspect of the workflow, the objective of transcription is to convert the initial audio sample into text, whilst keeping the structure, punctuation, content intact. Transcription is the tool that enables a second modality, text to be extracted from the initial data.

3.2.3 Text Analysis

Text analysis is a core component, similar to affective prosody analysis. This analysis is required to gain insights into the text modality, which will then be compared with the audio analysis results. These insights can be gained from various analysis methods which are as follows:

- Using semantic analysis on the text provides insight into the structure and content of the data.
- Conducting sentiment analysis on the text provides an overall positive or negative sentiment classification.
- Theme and entity extraction is a further method that can be used to gain insights into the meaning and context of the text data.

3.2.4 Emotional Recognition

Emotional recognition takes place on both modalities, utilizing the range of emotions found in **BET** [3]. This objective results in the presence or absence of the emotions being identified. This objective, emotional recognition is an important objective to achieve successfully as it provides a psychological classification for the content found in both the non-verbal and verbal components of the data.

3.2.5 Contextual Analysis

The contextual analysis objective consists of converging the results of the various analysis methods for each modality. An example of this is utilizing the sentiment and emotional classification on the audio alongside the semantic analysis to identify the context of the non-verbal components such as pauses, tone, etc. Contextual analysis can also be seen in the text modality, by utilizing the emotional classification results alongside semantic analysis, theme, and entity extraction, to provide an overall analysis that outlines the meaning of the verbal component.

3.3 Functional Requirements

The following are the functional requirements that have been outlined for this project:

1. Create or adapt an existing transcription model to allow audio to text conversions.
2. Create or adapt an existing affective prosody machine learning model to classify the emotion based on non-verbal components.
3. Create a semantic analysis workflow that aids the extraction of relevant features for the text content.
4. Create or adapt an existing emotion detection machine learning model to classify the emotion based on text data.
5. Create an analysis process, that combines the classifications and extracts relevant contextual information.
6. Create a platform that employs the emotion detection workflow, and allows an end-user to input audio and results in emotion detection on the given audio, and further analysis providing contextual information around the analysis.

3.4 Non-Functional Requirements

The following are the non-functional requirements that have been outlined for this project:

1. Each stage of project post transcription must occur synchronously, in an attempt to increase performance times.
2. Both the affective prosody and the text analysis must use the same or similar set of classifications, including an optional classification for neutral or undetected emotions.
3. The analysis models must be built into the project in a modular manner which allows future scalability and improvements, in an attempt to improve the models over time, a portion of the input data may be retained and used to re-train the models.

Chapter 4

Implementation and Evaluation

4.1 Architecture

The overall structure of this bi-modal emotion detection workflow can be seen as several key components, as outlined in the previous chapter. When considering the implementation approach there will be a number of stages that take place between and inside the key components. This chapter consists of an elaboration on the implementation of the overall architecture to provide a high-level overview of the system.

4.1.1 Data Gathering

The project focuses widely on data-driven machine learning models, therefore the initial dataset choice is vital. Utilizing a dataset constructed from an experimental study by **CIT** containing both text and audio reactions to emotionally provocative stimuli enables a bi-modal emotion analysis to be conducted.

Firstly, a dataset with both audio and text labels is required as previously stated. This dataset does not need to contain any specific emotion labels, as it's sole purpose is to act as a baseline in evaluating the chosen transcription method, which when used in conjunction with self-reporting and evaluation can allow for iterative improvements to the transcription model.

Secondly, an audio dataset containing emotional labels is required, a key benefit here would be utilizing a dataset with both text labels for the audio and emotional labels for the exhibited emotions. However, the text labels are not fully required as the transcription service can be utilized to provide them. The purpose of this dataset is to train both the affective prosody model and the text analysis model to determine the presence of emotions, similarly to the transcription model using an iterative process and self-reporting/evaluation the model can be improved.

4.1.2 Affective Prosody Pre-Processing

This stage is the pre-processing stage for the affective prosody model. The purpose of this stage is to extract the relevant features of the audio, these may be tone, pitch, pauses, etc. At this stage, 'cleaning' of the audio samples will also take place, which is the removing of noise from the audio. This aspect will be less relevant in the training and development stages but will be valuable for accentuating the key relevant features in the end user's audio samples.

4.1.3 Affective Prosody Emotion Analysis

The emotion detection of the non-verbal component takes place at this stage. Based on the relevant features extracted in the previous section, a machine learning or deep learning model will process these features and return a resulting classification that will represent the presence or absence of specific emotions.

At this stage it is likely the system will be leveraging an externally developed prosody analysis model, such as ProsoDeep [6] or Speech Emotion Analyzer [7] in order to correctly identify the emotions exhibited in the audio sample.

4.1.4 Affective Prosody Semantic Analysis

As the workflow aims to detect and define relationships between both the non-verbal and verbal components. Semantic analysis on both modalities is required, at this stage, the result of the emotion analysis stage is leveraged, alongside the semantics, auditory structure, and sentiment determined of the audio sample. All of which will enable the cross-examination of both modalities in terms of emotion, sentiment, and semantics.

4.1.5 Transcription

Transcription is used to allow for traditional text semantic, sentiment analysis and emotion detection methods to be used on the audio samples, by extracting the verbal element of the audio sample. Therefore the initial audio becomes text data.

Utilizing a labeled audio dataset alongside a commercial transcription service such as google speech to text acts as a baseline. Following this, the same dataset can be used with an open-source transcription service such as DeepSpeech [8]. The purpose of this is to determine the word error rate and overall accuracy. Once the best transcription service has been identified it can then be utilized in the transcription component of the emotion detection workflow.

4.1.6 Text Pre-Processing

Once the text content has been extracted from the audio samples, traditional **NLP** methods can be used. Therefore enabling the extraction of relevant features, this stage may use methods such as N-grams and others to determine sentiment and theme of the text prior to the emotional analysis.

This stage will also be important when it comes to removing any noise such as symbols, excess spacing, etc, and ensuring that the transcribed text is comprehensible to the machine learning model.

4.1.7 Text Emotion Analysis

The emotion detection of the verbal component takes place at this stage. Based on the relevant features extracted in the previous section, a machine learning or deep learning model will process these features and return a resulting classification. This classification will represent the presence or absence of specific emotions.

At this stage it is likely the system will be leveraging an externally developed emotional analysis model, however, this may be used in conjunction with a custom model that allows for the additional information extracted such as theme and sentiment in the prior stage to be used in the final emotional classification.

4.1.8 Text Semantic Analysis

Similarly to the semantic analysis stage of the affective prosody. The purpose of this stage is to analyze the semantics and structure of the text content. This stage is important for highlighting the relationships between both non-verbal and verbal features. A key consideration at this stage is that in transcribing the audio to text that the semantics reaming intact.

4.1.9 Classification Convergence

At this final stage of the workflow, a classification result can be identified. This classification result determines the presence of emotions for both modalities of the same sample. Therefore allowing the convergence of all the information gleaned from the initial data into one final analysis. A final analysis that highlights the emotion present, the theme, general content, the overall sentiment, and the semantics. All of which provide key insights into where inferences took place and what they may have been in the case of human interaction with this data.

4.2 Risk Assessment

The following risks have been identified during the research phase and have been classified by consequence and occurrence frequency.

4.2.1 Risk one - Inconsistent Transcription

- Consequence: Fatal
- Occurrence Frequency: Remote
- Description: Transcription is a key feature in extracting the spoken words from the audio content. In the case that the transcription service provides poor performance, this could result in the text extracted being incoherent, or containing little of the true content thus resulting in analysis on data that is irrelevant.
- Mitigation: By firstly performing evaluations using a commercial transcription service and a baseline dataset reduces the chances of this occurring. In the case that this situation occurs, alternative open-source transcription services may be utilized. If the alternative service fails to yield a sufficient result then educational licenses for a commercially tested and well-performing transcription service may be acquired.

4.2.2 Risk two - Semantics lost in Transcription

- Consequence: Critical
- Occurrence Frequency: Rare

- Description: Similarly to risk one, if the transcription service performs poorly extracting words, it may also perform poorly in extracting the overall semantics. If this is the case it may be difficult to keep the general theme of the spoken content intact. This may render the semantic analysis redundant due to the complete or partial loss of semantics.

- Mitigation: A similar mitigation as risk one, may be to find a better performing transcription service, or that to construct an additional model that can reconstruct partial phrases. The downside to this mitigation is that it may render the text quite different from the actual content. A final mitigation could be manual transcription.

4.2.3 Risk three - Inconsistent Emotion Classification

- Consequence: Major

- Occurrence Frequency: Occasional

- Description: Two separate models are required for the emotion detection. One for the non-verbal components and one for the verbal components. External emotion detection models may be used, resulting in the usage of two models that implement detection for a separate range of emotions and not the **BET** [3] emotions.

- Mitigation: In this situation, the emotional classifications need to be analyzed. As there may be the potential of drawing associations between emotions. As emotions can be different but related. If this is not sufficient, the possibility remains to explore other emotion detection models that employ the same range of emotions.

4.2.4 Risk four - Low Performance in Emotion Detection

- Consequence: Critical

- Occurrence Frequency: Remote

- Description: As with all machine learning models, their accuracy is a key factor in the overall analysis. There is a possibility that the emotional detection models for both modalities may have low accuracy. A low accuracy could stem from poor training data, or poor model construction, either case would be detrimental to the project.

- Mitigation: This can be mitigated by firstly conducting an initial evaluation of the models for each modality to ensure that they perform adequately. In the situation that they have poor performance for certain or all emotions, further evaluation can take place on the training data. Depending on the quality of this data it can either be replaced with a suitable alternative or extended with additional entries.

4.2.5 Risk five - Unidentifiable Semantic Relationships

- Consequence: Minor
- Occurrence Frequency: Rare
- Description: In analyzing both modalities, there is a possibility that the semantic relationships between the audio and the text cannot be identified. These relationships are expected to provide insights into the inferred information. If they prove to be unidentifiable this will force the analysis of semantics separately and potentially requiring manually constructing the relationships.
- Mitigation: As stated above, a possible mitigation is to observe the semantics separately in the context of the modality they are from. This may reduce the capability to identify hidden inferred information directly. In order to compensate for this, an additional deeper analysis into the modalities individually may be required. Enabling the extraction of a similar type of information that can in turn be used to construct the final analysis.

4.2.6 Risk six - Noisy Audio Samples

- Consequence: Critical
- Occurrence Frequency: Probable
- Description: As the workflow is processing audio samples, there is a probability that the system may be presented with an unclear audio sample. This issue is less likely to occur in development and training but potentially prevalent in a real-life application. Typical noise reduction methods may also impact the audio sample negatively as they may reduce the prevalence of the key features required for affective prosody analysis.
- Mitigation: Typical noise reduction methods should be sufficient in reducing excess noise, whilst ensuring the key features are still present and unaltered. However in the case that the audio sample is overly altered, the processing of both altered and unaltered versions of the audio may be required.

4.3 Methodology

For this project, a methodology was adopted that is comprised of the following methods, which has aided both research and development stages.

- Extensive research and analysis
- Applied research and learning
- Practical Evaluation
- Agile Scrum boards
- Feature Driven Development

4.3.1 Extensive research and analysis

Prior research into the current state of emotion detection, and affective prosody analysis, as well as multi-modal approaches to emotion detection, was required. These domain-specific areas provide insight into the possible structure of the project. In order to gain a high level of understanding in this area requires both theoretical and practical knowledge. Therefore research into the theoretical side of emotion detection, and also the practical aspect by examining source code of open source projects have provided key information.

4.3.2 Applied research and learning

Understanding the theoretical aspect of a given research topic provides valuable insights. However actively reconstructing and testing the methods found from research, allows for additional information to be obtained that will be vital in the development stages. An example of this would be recreating **NLP** methods and sentiment analysis systems, resulting in practical experience which makes not only understanding the theoretical portion easier but also provides code that can be adjusted to solve future problems.

4.3.3 Practical Evaluation

Similar to the applied research, analyzing working examples to determine their efficacy and performance is important when discerning which functions are best suited to this workflow. By conducting practical evaluation and comparisons of important components. Whether these components are programming functions or externally developed systems, practical evaluation enables the most educated design choices to be made.

4.4 Implementation Plan Schedule

4.4.1 January 11th - February 1st

Creation of several test projects. Each of which implements one of the externally developed systems. Therefore, resulting in a project for transcription, emotion detection from text data, and from affective prosody analysis. The purpose of these projects is to enable a deeper practical analysis of the systems. By analyzing the input and output data formats, the overall project architecture, and the project's inner components, in order to provide a comprehensive understanding of the systems. The systems can then be evaluated in terms of performance, efficacy, and suitability. At this stage any obvious issues in the choice of the external systems will be apparent, allowing for adequate time to identify replacement systems.

4.4.2 February 1st - February 8th

Implementation of the best transcription service identified at the previous stage. This transcription service is then utilized to transcribe the dataset collected. The results are then compared with the baseline transcription model to determine the overall and relative accuracy. The purpose of this stage is to ensure that the transcription service is accurate and performs adequately on the real dataset.

4.4.3 February 8th - February 22nd

At this stage, the affective prosody and text emotion detection models have been defined. Enabling a deeper analysis into the systems specifically the inputs. Therefore the pre-processing and feature extraction procedures can be defined. These procedures will be different for each modality and will result in adapting the raw input data into a format suitable for the machine learning models.

4.4.4 February 22nd - 8th March

Using the pre-processed inputs a semantic analysis algorithm for both audio and text can be defined. This function is used to extract structural information of the data, and to extract the meaning from text data, the results of which will be key for contextual analysis. At this stage, a sentiment analysis algorithm or model will be implemented to return a general positive or negative sentiment classification of the text and audio data.

These functions act as additional analysis points of the input data, the results of which will be converged into the final analysis of the data.

4.4.5 8th March - 29th March

At this stage, each separate analysis component of the workflow has been created. This stage focuses on implementing a system that allows for each of the prior functions and models to be integrated into one overarching system. This involves creating an input structure to process audio samples in batch and individually, following this the transcription model transcribes the audio, this results in the two modalities. Each modality can then be pre-processed and analyzed with its respective methods. Therefore the overarching system at this stage can take an audio input and conduct semantic and sentiment analysis and also emotion detection on each modality.

4.4.6 29th March - 5th April

Utilizing the results of the system creating at the last stage, the contextual analysis function can be defined. This function involves a level of semantic analysis to identify the meaning behind the verbal component. It also involves the convergence of the results of the analysis methods for each modality. The purpose behind this stage is to implement the final analysis function that will identify the semantic and contextual relationships between the modalities. Once the final analysis function has been defined, the outputs can be summarised and displayed to the end-user, a function will be implemented to allow for this information to be displayed.

4.4.7 5th April - 5th May

This stage will remain clear of development tasks, as it will be used for bug fixes, optimization, and any unforeseen tasks.

4.5 Evaluation

As defined in the previous chapter, this bi-modal emotion detection workflow can be broken down into several core components. As a reminder, these are affective prosody analysis, transcription, semantic analysis, and emotional recognition. The development progress can be evaluated in-line with the core components, the functional requirements, and the implementation plan.

The following outline the progress evaluation stages:

- Initial Evaluation
- Affective Prosody Evaluation
- Transcription Evaluation
- Semantic Analysis Evaluation
- Emotional Recognition Evaluation
- Contextual and Final Analysis Evaluation

4.5.1 Initial Evaluation

The initial evaluation is used to measure the progress of the analysis of the externally developed libraries and systems. As the workflow involves machine learning models this presents an opportunity. Which is to use the accuracy metrics of these models as an evaluation metric for the implementation progress itself. At the conclusion of the initial applied research stage, accuracy values for each potential model will have been calculated. Alongside this baseline accuracy values using commercial tools will also have been identified. These accuracy metrics can be used to determine the successful progress of this stage. As high accuracy demonstrates the model is performing well, and that the correct procedure for running the model has been implemented.

4.5.2 Affective Prosody Evaluation

Similarly to the initial evaluation, the accuracy metric of the chosen affective prosody model will determine the successful completion of this stage. The goal of the affective prosody model is to classify the presence or absence of emotions from an audio sample based on the non-verbal components. A high accuracy metric for this model is a key indicator that the associated functional requirement has been met.

4.5.3 Transcription Evaluation

The successful progression of this stage can be identified by the overall accuracy of the transcription model. The intended cross-examination of an open-source and a commercial transcription service provides a degree of success. This degree of success is determined by the difference in word error count between the open-source transcription model and the commercial transcription service.

Successful completion of the associated functional requirement and therefore this implementation stage is determined by the implemented transcription model having an accuracy value similar to that of the baseline accuracy value.

4.5.4 Semantic Analysis Evaluation

The evaluation of this stage is slightly different, as there is no accuracy metric to use. The semantic analysis must occur on both modalities. Therefore the evaluation must also occur on both modalities. This means that the successful progression of this stage is reliant on two algorithms. One for semantic analysis on audio that can extract the semantic information and structure of an audio sample. And a second, which conducts semantic analysis on text, extracting theme, entities, structure, and general meaning.

This stage in the implementation plan additionally includes sentiment analysis. Hence the evaluation of this stage will additionally require the accuracy of the sentiment analysis on both modalities.

The completion of this stage dictates that the functional requirements for semantic analysis have been met.

4.5.5 Emotional Recognition Evaluation

This progress evaluation stage is similar to the initial evaluation. As it is based on the accuracy of the implemented machine learning model. In this case, the emotion recognition models for both modalities. A high accuracy metric for both models dictates that the associated functional requirement has been met, and based on the accuracy value itself it's possible to determine how well this requirement has been met.

4.5.6 Contextual and Final Analysis Evaluation

The evaluation of the final stage comes in two parts. The first is the evaluation of the implementation of the contextual analysis algorithm. The second is the evaluation of the overall analysis results, and how it relates to the relationship between non-verbal and verbal components of emotional language. These two evaluations are inherently linked as a failure to identify the contextual information and conclude the correct inferred relationships will result in incorrect information being displayed to the end-user. However, there is a degree of acceptable error in this final analysis stage. Therefore providing a less concrete evaluation. To succeed in this evaluation, the end-user must be presented with the results of the prior analysis. From these results, the inferred relationships can be identified. Resulting in the associated functional requirement having been met.

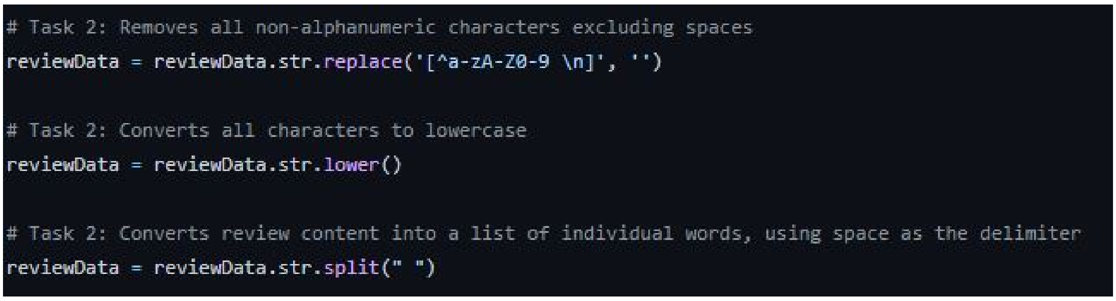
4.6 Prototype

During the research phase, a sentiment analysis project was created in python. Serving as a practical project to utilize a range of research findings. The sentiment analysis model utilized Naive Bayesian classification, to discern a positive or negative sentiment from movie reviews. The movie reviews analyzed were in text format. Therefore a significant overlap can be identified, specifically in the analysis of the text modality for emotion detection. Similar methods of pre-processing, feature extraction, and sentiment analysis can be implemented into the workflow.

4.6.1 Pre-Processing

Utilizing the Pandas [9] library, data-points were extracted from a dataset of movie reviews. The data included a string containing the review and either a positive or negative sentiment label.

As the review string initially contained symbols, numbers, and trailing spaces, pre-processing was required. Additionally, the Naive Bayesian classification method implemented considered each separate word of the review to be a feature. Therefore requiring the 'cleaned' review string to be split into words. The following Figure 4.1 is a code snippet outlining the Python code used for pre-processing.



```python
# Task 2: Removes all non-alphanumeric characters excluding spaces
reviewData = reviewData.str.replace('[^a-zA-Z0-9 \n]', '')

# Task 2: Converts all characters to lowercase
reviewData = reviewData.str.lower()

# Task 2: Converts review content into a list of individual words, using space as the delimiter
reviewData = reviewData.str.split(" ")
```

FIGURE 4.1: Pre-Processing Procedure in Python

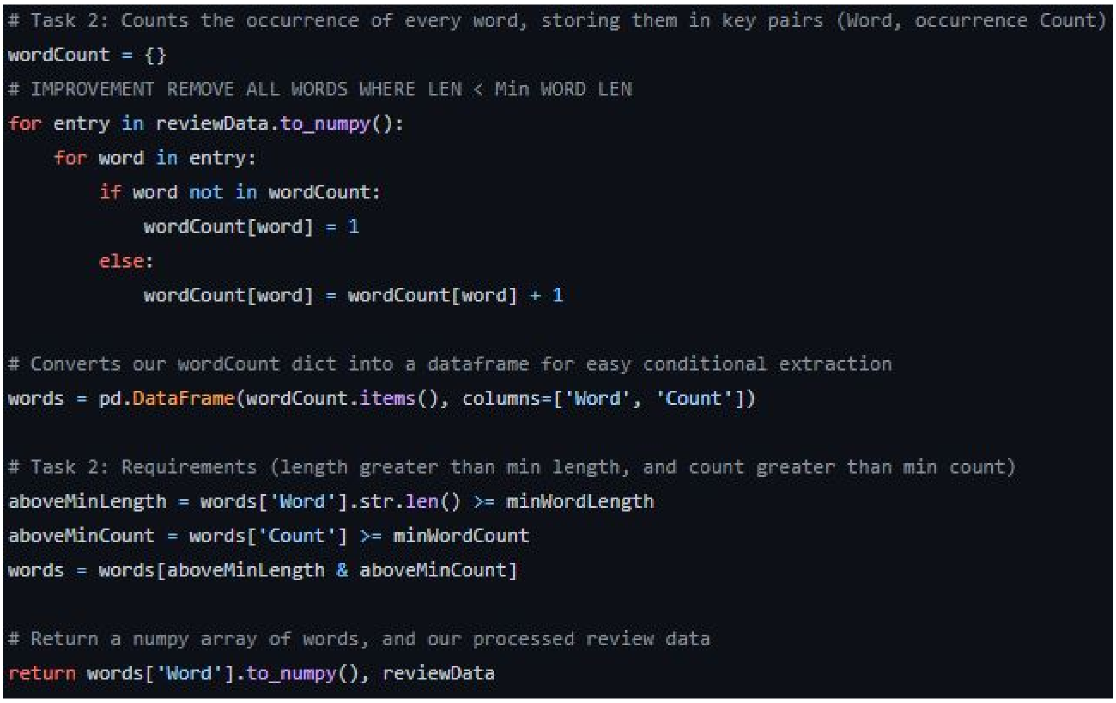
4.6.2 Extracting Relevant Features

A common procedure in machine learning models is extracting the relevant features from data. In the case of this sentiment analysis project, the relevant features were words that affected the classification in a meaningful manner. Conducting an analysis of the movie review data provides insights into the relevant features. When considering a movie review, words such as 'I, a, he, as, it' etc provide little information about the sentiment. Therefore the length of the words used as features is relevant.

Similarly, words that occur rarely have little bearing on the overall sentiment. As these words may be film-specific, or names, or even places, etc, all of which provide little insight into the sentiment exhibited in the review.

By parameterizing both the length of words used, and the occurrence frequency allows for cross-validation. Therefore by extracting a different set of relevant features with each parameter change, cross-validation can be conducted to identify the correct set of relevant features.

The following Figure 4.2 is a code snippet outlining the python code used for extracting the relevant features.



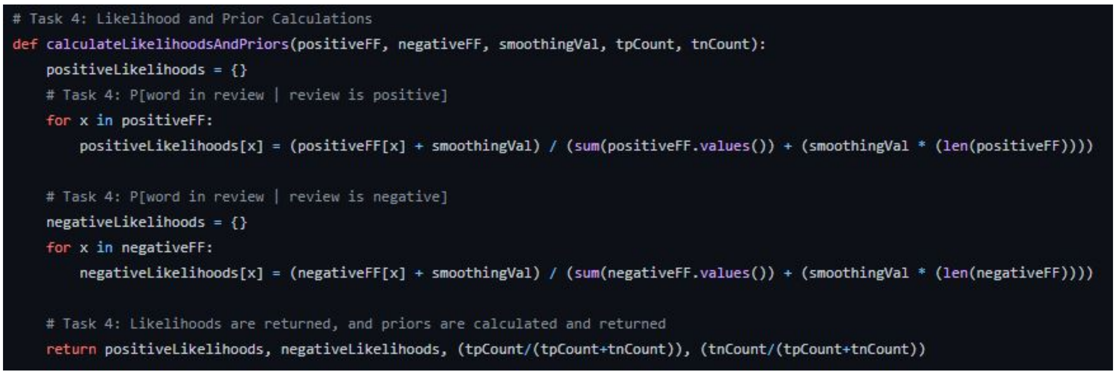
```python
# Task 2: Counts the occurrence of every word, storing them in key pairs (Word, occurrence Count)
wordCount = {}
# IMPROVEMENT REMOVE ALL WORDS WHERE LEN < Min WORD LEN
for entry in reviewData.to_numpy():
    for word in entry:
        if word not in wordCount:
            wordCount[word] = 1
        else:
            wordCount[word] = wordCount[word] + 1

# Converts our wordCount dict into a dataframe for easy conditional extraction
words = pd.DataFrame(wordCount.items(), columns=['Word', 'Count'])

# Task 2: Requirements (length greater than min length, and count greater than min count)
aboveMinLength = words['Word'].str.len() >= minWordLength
aboveMinCount = words['Count'] >= minWordCount
words = words[aboveMinLength & aboveMinCount]

# Return a numpy array of words, and our processed review data
return words['Word'].to_numpy(), reviewData
```

FIGURE 4.2: Extracting Relevant Features Procedure in Python

4.6.3 Sentiment Classification

The sentiment analysis classification was created by using Naive Bayesian classification. This requires calculating positive and negative likelihood values. These likelihood values are calculated by dividing the positive feature frequency by the total positive feature count, and similarly for the negative likelihood. The feature frequency is a value derived from counting the presence or absence of a word from a review. The positive and negative priors must also be calculated, which are the number of positive reviews divided by the total amount of reviews, and similarly for the negative prior.

The following Figure 4.3 is a code snipped outlining the Python code used for likelihood calculation.

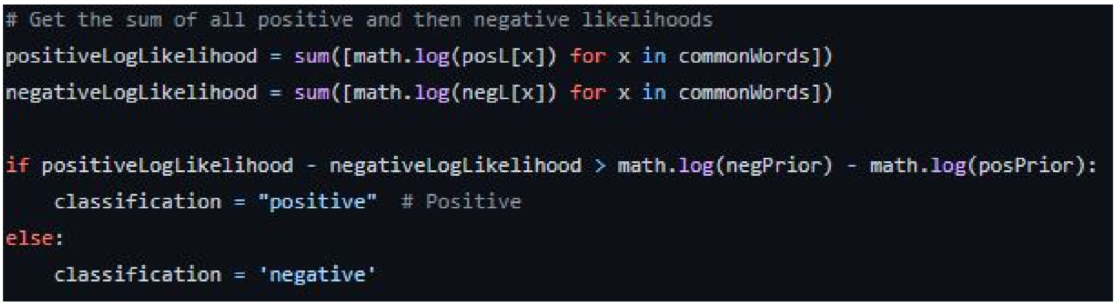


```python
# Task 4: Likelihood and Prior Calculations
def calculateLikelihoodsAndPriors(positiveFF, negativeFF, smoothingVal, tpCount, tnCount):
    positiveLikelihoods = {}
    # Task 4: P[word in review | review is positive]
    for x in positiveFF:
        positiveLikelihoods[x] = (positiveFF[x] + smoothingVal) / (sum(positiveFF.values()) + (smoothingVal * (len(positiveFF))))

    # Task 4: P[word in review | review is negative]
    negativeLikelihoods = {}
    for x in negativeFF:
        negativeLikelihoods[x] = (negativeFF[x] + smoothingVal) / (sum(negativeFF.values()) + (smoothingVal * (len(negativeFF))))

    # Task 4: Likelihoods are returned, and priors are calculated and returned
    return positiveLikelihoods, negativeLikelihoods, (tpCount/(tpCount+tnCount)), (tnCount/(tpCount+tnCount))
```

Figure 4.3: Likelihoods Calculation Procedure in Python

These likelihoods and priors are then used in the classification calculation. The classification is determined by the difference between likelihoods being greater than or less than the difference between the priors.

The following Figure 4.4 is a code snipped outlining the Python code used for the classification calculation.



```python
# Get the sum of all positive and then negative likelihoods
positiveLogLikelihood = sum([math.log(posL[x]) for x in commonWords])
negativeLogLikelihood = sum([math.log(negL[x]) for x in commonWords])

if positiveLogLikelihood - negativeLogLikelihood > math.log(negPrior) - math.log(posPrior):
    classification = "positive"  # Positive
else:
    classification = 'negative'
```

Figure 4.4: Naive Bayesian Classification in Python

Chapter 5

Conclusions

5.1 Discussion

Providing an improvement for **HCI** was the initial intention of this dissertation. Prior to the research phase the scope was much wider and potentially extravagant for the time given. As the research phase progressed it became clearer that the initial scope required refinement. In order to accomplish this, the project scope became domain-specific. The domain chosen was emotional language, as this enabled a semantic analysis and a psychological classification of the data.

An initial complication was identifying a reliable method to extract the text modality. This lead to research into both commercial and open-source transcription services. By utilizing a commercial service, it's reasonable to expect high accuracy and performance. Therefore this creates a baseline to act as a comparison against the open-source transcription services. This comparison then enables a reliable transcription service to be identified and utilized.

A significant decision was the range of emotions to be analyzed. There is a wide range of psychological theories outlining the 'basic' or 'universal' human emotions. Each of these theories supported and disputed, which resulted in a degree of uncertainty when choosing the emotions this workflow detects. After research into these theories, **BET** [3] was the theory that was chosen. As the core controversy surrounding **BET** stems from the claim that **BET** emotions are universal. In the case of this emotion detection workflow, the universality of the emotions is irrelevant, hence **BET** was the aptest choice of emotions.

5.2 Conclusion

To conclude the findings of this research phase, a domain-specific approach to a problem has been outlined. This outline includes an introduction to the issue, a specific problem definition, and finally a high-level implementation plan.

5.2.1 Complications in VDA

The issue identified is a lack of fluidity and potential for improvement in **VDA** systems. This issue is most notable due to the manner in which **HCI** occurs in **VDA** systems. By concretely processing the verbal component of the input audio, these systems lose a layer of data. Therefore reducing the system's capability to fully understand the user and increasing the chance of miscommunication.

The previously outlined issue stems from communication. Specifically the medium of communication which is used between a human and a **VDA**, language. this medium can be broken down into non-verbal and verbal components, verbal components can be processed and understood using methods such as **NLP** and semantic analysis, and are typically employed by **VDA** systems. However, this neglects the non-verbal components. And analyzing human to human communication shows the utility of these non-verbal aspects of language.

As previously stated, this issue can be addressed in a wide range of domains. However, by choosing emotional detection, a psychological classification of the data is present. Therefore enabling a narrow research area to be explored.

Emotional communication or emotionally charged language provides a wide range of insights into the subject matter and the individual expressing their views. Therefore analyzing this domain enables a cornucopia of information to be extracted from the input data.

Spoken emotional language can then be converted from audio in order to extract the verbal components. The original audio is also analyzed, the combined analysis of these modalities provides a human-like approach. This approach mimics humans, by considering both the non-verbal and verbal components in the overall classification. The classification which in this case the presence or absence of emotions.

The benefit of using this approach is that a **VDA** is able to extract the emotional information from a sentence. Therefore enabling an emotion detection method that is not limited to the analysis of verbal components. An emotionally aware **VDA** can then provide more suitable suggestions, ultimately providing a higher level of understanding of the user.

5.2.2 A Bi-Modal Workflow for Emotion Detection

Utilizing two modalities, text and audio enables the extraction of additional information from the input. The workflow defined in chapter three explores the potential behind the analysis of these modalities.

Sentiment analysis can be conducted on both modalities. Therefore providing an overall positive or negative classification of the data. This classification can then be utilized in order to identify keywords or entities that are associated with the exhibited sentiment.

Semantic analysis can also be conducted on both modalities. The semantic analysis of audio provides key insights into the structure, and context of the non-verbal components found in the audio sample. Similar semantic analysis can also be conducted on the verbal features, allowing for theme and entity extraction, overall meaning and structure to be extracted from the data. Therefore enabling a cross-examination of the results of the analysis on both modalities. This examination aids the identification of inferred relationships and contextual analysis of the data.

Emotion detection is used as a psychological classification of the data. The exhibited emotions also provide key insights that can be used alongside the sentiment and semantic analysis. The resulting information extracted from these analysis methods, enables overall insights, contextual information, and relationships between non-verbal and verbal components to be determined from the input data.

5.2.3 Implementing the Workflow

After prior research, the implementation approach has been defined for the aforementioned workflow.

The planned approach involves firstly, using a transcription service to extract text from audio. This optimal transcription method is chosen based on it's relative performance against commercial transcription services. Emotion detection can then take place on the affective prosody of the audio, and on the transcribed text content. Therefore resulting in a bi-modal emotion classification. Additionally, semantic analysis and sentiment analysis is conducted on the modalities and the results of which are combined. This combined analysis enables further contextual information to be extracted from the initial data.

This implementation approach will leverage the power of machine learning models to classify the data. And custom algorithms for semantic, sentiment, and contextual analysis. Finally, the analysis results will be combined into a final overview providing the desired insights to the end-user.

5.3 Future Work

The following features are possible improvements or additional components:

1. Custom affective prosody emotion detection model
2. Custom textual emotion detection model
3. Explore a wider range of emotions
4. Pattern Recognition (Semantics, Emotion, Context)
5. A web platform to support the workflow
6. Implementation of the workflow to a **VDA** system
7. **VDA** with decisions influence by the exhibited emotion

Bibliography

[1] [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[2] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," *IEEE Access*, vol. 8, pp. 181 074–181 090, 2020.

[3] P. Ekman, *Basic Emotions*. John Wiley and Sons, Ltd, 1999, ch. 3, pp. 45–60. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3

[4] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall, 2010.

[5] M. R. Hasan, M. Maliha, and M. Arifuzzaman, "Sentiment analysis with nlp on twitter data," in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 2019, pp. 1–4.

[6] Gerazov, "gerazov/prosodeep." [Online]. Available: https://github.com/gerazov/prosodeep

[7] Miteshputhranneu, "Miteshputhranneu/speech-emotion-analyzer." [Online]. Available: https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer

[8] Mozilla, "mozilla/deepspeech." [Online]. Available: https://github.com/mozilla/DeepSpeech

[9] [Online]. Available: https://pandas.pydata.org/

Appendix A

Code Snippets