Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Automatic Emotion and Sentiment Extraction from Speech

James Gorman

15319250

Supervisor: Professor Khurshid Ahmad

April 30, 2020

A Final Year Project submitted in partial fulfilment
Of the requirements for the degree of
ICS (Computer Science)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University calendar for the current year, found at – http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: James Gorman   Date: 30/04/2020

Abstract

This project aims to extract emotion and sentiment from a speaker's voice. This entails the use of an automatic speech recognition system to analyse the acoustic features of the human voice, as well as using a sentiment analysis technique to analyse a speaker's word choice. This system will be applied to a carefully selected dataset consisting of politicians for analysis. Each politician's acoustic speech signals will be analysed as well as their word choice. These modalities of communication will be fused together, with the aim of validating the extent to which an emotion is present. The automatic feature extraction software OpenSMILE was used to analyse speech signals. A Lexicon-Based sentiment analysis approach was used to analyse the speaker's word choice.

Acknowledgments

I would first like to thank my supervisor, Professor Khurshid Ahmad, for his invaluable guidance, support and help throughout the course of this project.

I would also like to thank Yunis Lone for his help over the course of this project, particularly with the configuration of OpenSMILE.

Finally, I wish to thank my family, friends and classmates for their encouragement and support over the 4 fantastic years I’ve spent at Trinity College Dublin.

Contents

List of Tables

List of Figures

1. Introduction

This section will begin with a short introduction in relation to the motivation for undertaking this specific research project. Subsequently, the methods throughout the project will be outlined, followed by a summary of the dataset gathered and used throughout the project. The key results from this project will be briefly discussed and finally this section will conclude with a summary of the structure of the remainder of the report.

1.1 Motivation

Humans communicate with each other by both spoken language and non-verbal cues. “Emotion modulates almost all modes of human communication – word choice, tone of voice, facial expression, gestural behaviours, and many more” (Picard et al., 2001). These emotions can be categorised into six main sections – “happiness”, “anger”, “sadness”, “surprise”, “fear” and “disgust” – it is the way in which humans deliver modes of communication that these emotions can be expressed.

The tone of voice plays a key role in expressing emotion. Recognizing emotion from speech can be done by analysing the distortion and contortion of the tone of voice from specific parts of the human throat. Contortions are observed by doing high frequency time series analysis and learning about the statistical properties of signals from a very specific part of the throat. These are said to be the physical correlates of an emotional state. No human can perfectly recognize your innermost emotions and sometimes people cannot even recognise their own emotions (Picard et al., 2001). The question is how well can a machine monitor the emotions of a human being?

My aim for this project is to investigate if a human beings dominant emotional state can be detected by an automatic speech recognition system from the physical correlate of their speech and by cross-checking one mode of communication (word choice) with another if this detection can be validated.

1.2 Methods Employed

As outlined in section 1-1, the aims of this project are to a) detect the dominant emotional state and b) to cross-check one mode of communication against another to accept or reject the emotional correlate of the voice and detect emotional leakage.

These objectives contributed to my decision to examine the project in four sections. These sections include:

(a) the obtaining and pre-processing of data;
(b) an analysis of the acoustics of speech;
(c) an analysis of word choice, and time series; and
(d) a fusion of results from both speech and word choice.

Obtaining of suitable data to be used throughout this project was very important. All audio had to be clear, concise and of enough length to be tested. This dataset was comprised of political figures. I will speak about this data in more detail in section 1-3. Pre-processing was also necessary in order to make this data compatible with automatic speech recognition systems.

The next stage of this project was to process each speech/interview through an automatic speech recognition system. This system uses the physical features of speech signals in order to determine what emotion is being expressed. The software I used was OpenSMILE.

Extraction of the audios transcript then followed. This was to prepare each speech/interview for sentiment analysis. I used a Lexicon-based approach to examine the word choice from each speech/interview. I measured the percentage of negative bearing words used as well as positive bearing words.

Once all data had been processed, an analysis was performed on each audio clip to find the dominant emotion state. Data fusion was performed between results from the automatic speech recognition system and sentiment analysis to detect leakage of emotion and establish if a relationship exists between the physical correlate of emotion and choice of words.

1.3 Summary of Dataset

The dataset chosen for this project is based around the 2019 UK elections. The dataset is comprised of interviews and speeches of 9 politicians. These politicians are geographically dispersed across the UK and Ireland, ranging between 40 and 70 years of age and evenly balanced with respect to gender (4 males, 5 females).

The politicians chosen for analysis are established politicians and have fantastic public speaking skills, often used to hide their true emotions to remain rational. This is what we can call “semi-wild data” as these politicians are intending to express specific emotions, but they are not in a professional recording environment. This dataset is used in both the automatic speech recognition stage and for sentiment analysis.

1.4 Key Results

Results were processed and represented using statistical formulae.

Emotional states that are of low valence (sadness, boredom, etc.) were highly dominant across all politicians. There was no distinct difference in emotions expressed depending on gender, age or geographical location.

Data fusion between the physical correlate of speech and a speaker’s word choice showed an overall anti-correlation between emotions expected to have an association with either a negative or positive valence, which implies the detection of emotional leakage.

1.5 Overview of Report

The first section of the report introduces the research motivation, methods used throughout the project, data used in those methods and the key results.

The second section focuses on the existing work in this research are. Some of the main research of both emotion recognition and sentiment analysis is discussed in this section.

The third section describes the methods, step by step, carried out in this project. In four main stages – gathering and pre-processing of data, extraction of emotion by use of an automatic speech recognition system, extraction of sentiment by analysis of speaker word choice, assessment and validation of obtained results.

The fourth section discusses the results obtained from section 3 and their significance.

In the fifth section I will deliver my conclusion derived from section four and talk about possible and probable routes for this project to go to in the future.

2. Motivation and Literature Review

This section will discuss the process of automatic speech recognition systems and sentiment analysis. Automatic speech recognition systems are related to the extraction of emotions from physical speech and sentiment analysis is related to the speaker's choice of words.

2.1 Motivation

Emotion recognition is a very exciting and rapidly advancing field of study. Emotion recognition can be done by analysing many forms of human communication, both verbal and non-verbal. Understanding how communication is being expressed is most accurately predicted when there are multiple communication modalities present as well as context about the message being delivered. Picard et al., (2001) states that "A combination of low-level pattern recognition, high-level reasoning and natural language processing is likely to provide the best emotion inference."

There are many applications in which the analysis of different modalities can be used. Research in the detection of emotion through the analysis of acoustics in speech has been applied to several sectors of everyday life. Medicine has been influenced the ability to use characteristics of speech for the early detection of diseases in the human vocal cords (Teixeira, 2013). Businesses can use this analysis to improve their customer support systems by recognising what emotion a caller is showing to help approach proposing a solution in a suitable manner (Narayanan, 2005). Studies have also been done to show that predictions can be made from analysing the acoustics of speech. Dietrich et al., (2019) carried out an investigation if the acoustic parameters within the speech of judges can be used to predict the verdict of judicial cases. Similarly, the analysis of a person's word choice can be influential in many different sectors. Sentiment analysis has been used successfully validate changes in the stock market with respect to investor sentiment (Ahmad et al., 2016), (Tetlock, 2007). Businesses often also use this to retrieve feedback about products from customers by applying sentiment analysis to product reviews on website platforms such as Twitter (Ravi, 2014).

I aim to examine if a machine can accurately detect emotions being expressed by humans through speech over rapidly changing events (UK elections) and attempt to validate this with the analysis of their word choice.

2.2 Emotion Recognition

This section will deal with the modality of communication: speech.

2.2.1 Speech Recognition Background

Speech is perhaps the primary means of communication between human beings. Many steps have been taken to develop technical systems to interpret the physical correlate of the human voice and apply this to emotions. Juang (2005) speaks of the main milestones in the development of automatic speech recognition systems. In the early 1950's, Bell Laboratories built a system for isolated digit recognition for a single speaker using phonetic frequencies (frequencies of speech) measured during vowel regions of each digit. Developments in this project over the next decade would lead to the creation of a machine that could recognize 16 words. This machine was called "Shoebox" and was developed by IBM in 1962. Shortly after this, there was a lull in the development of speech recognition technologies due to the lack of funding at Bell Laboratories. In the early 1970's, DARPA invested heavily in research behind speech recognition, aiming to create a machine that could recognize 1000 words. Around the same time, the Hidden Markov Model began to be used in speech recognition. This propelled the field's development and over the next 20 years speech recognition systems with vocabularies consisting of thousands of words. Since the 1990's, businesses have made use of such technology. AT&T were the first company to use automated telephone calls. In today's world, speech recognition systems can be seen everywhere – from smart speaking virtual assistants such as Amazon's Alexa, to smart cars – it is a rapidly advancing field of technology.

2.2.2 Properties of Speech

The word "speech" is defined by Oxford dictionary as "the way in which a particular person speaks." We as human beings, communicate through speech by moving air through a specific part of our throat to produce vibrations with different frequencies and amplitudes. How we adjust these frequencies and amplitudes determines the way in which our communication is expressed. These frequencies and amplitudes can be called "acoustic signals".

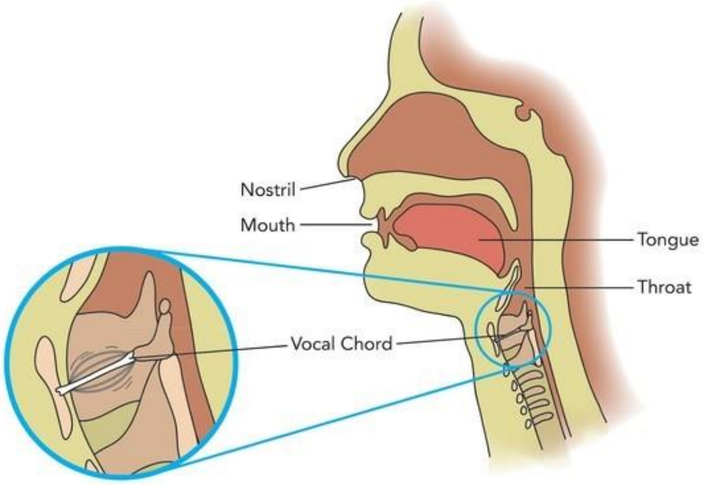


Figure 1: Location of vocal cords in the human throat

From these acoustic signals, it is possible to extract a set of parameters to analyse. The main parameters that are assessed with respect to acoustic signals are: the fundamental frequency (f0), jitter and shimmer (Teixeira et al., 2013).

Fundamental frequency, in terms of speech, is defined as the number of glottal pulses in a specific time frame. Glottal pulses refer to the opening and closing the vocal cords. These frequencies differ in range depending on if the person speaking is male or female and what age they are. The value for f0 is measured in Hertz and typically ranges between 200-300Hz for females, 80-200Hz for males and 400-500Hz for children. F0 has been said to depend on many factors other than age and gender, such as "the state of mind of the person, the time of day that fit the lifestyle and professional use of voice" (Teixeira et al., 2013). The term “pitch” is often used to describe the was in which we as humans perceive the fundamental frequency.

Jitter and *shimmer* are also very important parameters within acoustic signals. Jitter denotes the variation of frequency over a specific time period while shimmer denotes the variation of amplitude over a specific time period (Farrus et al., 2007). Small variations in tracheal pressures will be picked up by jitter and shimmer. The formulae for jitter (representing the absolute difference between two consecutive time periods) and shimmer (representing the absolute difference between the amplitudes of two consecutive time periods divided by the average amplitude) are as follows:

$$jitt_{abs,l} = \frac{1}{N-1}\sum_{1=1}^{N-1}|\overline{T_\imath} - T_{i-1}|$$

Where T_i = period duration in seconds, N = number of periods

$$shim_{Local} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|\,A_i - A_{i+1}|}{\sum_{i=1}^{N-1}|\,A_i|}$$

Where A_i = amplitude value, N = number of periods

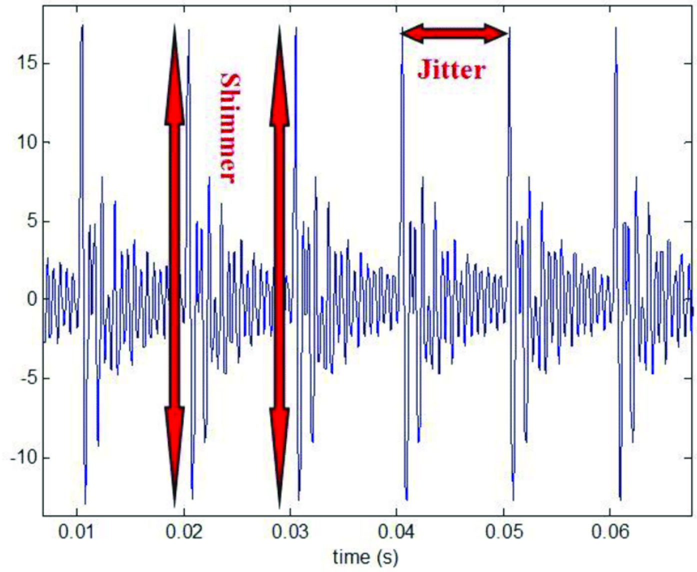


Figure 2: Visualization of Jitter and Shimmer in wave frequency

Rachman et al., (2018) discusses consequences of what happens when the values of f0, jitter and shimmer are changed. By increasing the value of the pitch (f0) by a constant gradient, this will cause for the communication to be expressed with a state of high arousal such as happiness and if this decreased it will generate a state of emotion with low valence e.g. sadness. If jitters value is altered, this can put emphasis on certain words and even as low as parts of words. Shimmer (also known as inflection) can affect the output emotion by a sudden change in the pitch at the beginning of each utterance, returning to the normal pitch value quickly. When this is increased it results in the expression of an emotion with high intensity such as happiness, anger, etc. If it is decreased, it will result on an emotion of low valence being expressed.

Using characteristics of f0, jitter, shimmer and many more from these “acoustic signals”, it is possible to extract emotions from a human being’s speech. This is referred to the physical correlate of emotion. Different algorithms such as the Hidden Markov Model (Schuller, 2003) and Deep Neural Networks (Han, 2014) have been used to approach the extraction of emotion from speech, bearing similar accuracies in their predictions.

2.2.3 Speech Recognition Process

A system architecture of how emotion can be detected through speech recognition can be seen in Figure 3.

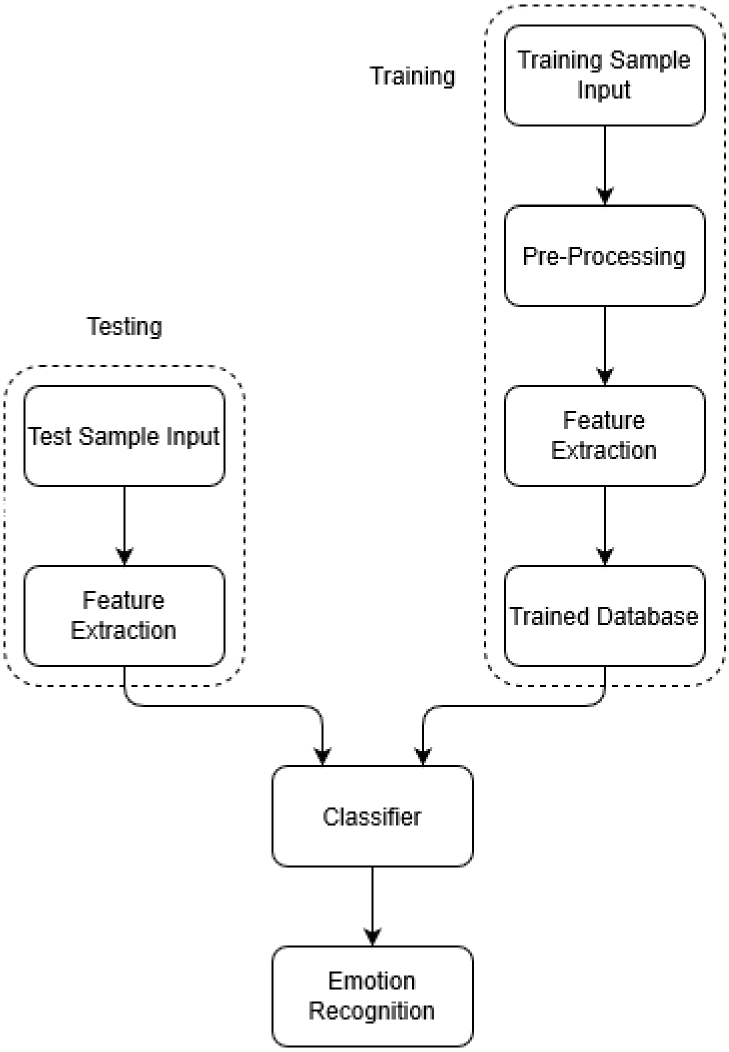


Figure 3: Architecture of speech recognition system

2.2.3.1 Training the System

Emotion recognition systems consist of many stages. Systems can be “trained” to recognise and categorise parameters from acoustic signals to specific emotions. This training is done by employing actors to read a script while expressing a certain emotion. This is done for several emotions and each script reading is analysed, with features of the acoustic signals being extracted (section 2-3-3-2). These extracted features are used to distinguish between emotions. The accumulation of information about each emotion regarding its associated acoustic signal properties is recorded and stored in a database (Ayadi et al., 2011). These databases can be used to test a given speech input. By comparing the given speech's acoustic signal parameters with the characteristics of each emotion in the database, the given speech can be assigned an emotion.

2.2.3.2 Extracting Features and Classification

As outlined in section 2-3-2, the human voice is comprised of many parameters. How these parameters change determines the emotion which is being expressed. Extraction of these features is the most important part of the speech recognition system, both for testing and training. The acoustic features focused on are the pitch (f0) and the energy values (jitter and shimmer), deriving descriptive statistics (mean, standard deviation, skewness, kurtosis, etc. (explained in section 2-5-1)). These features are extracted in segments of the provided audio. This can be visualised below in Figure 4.

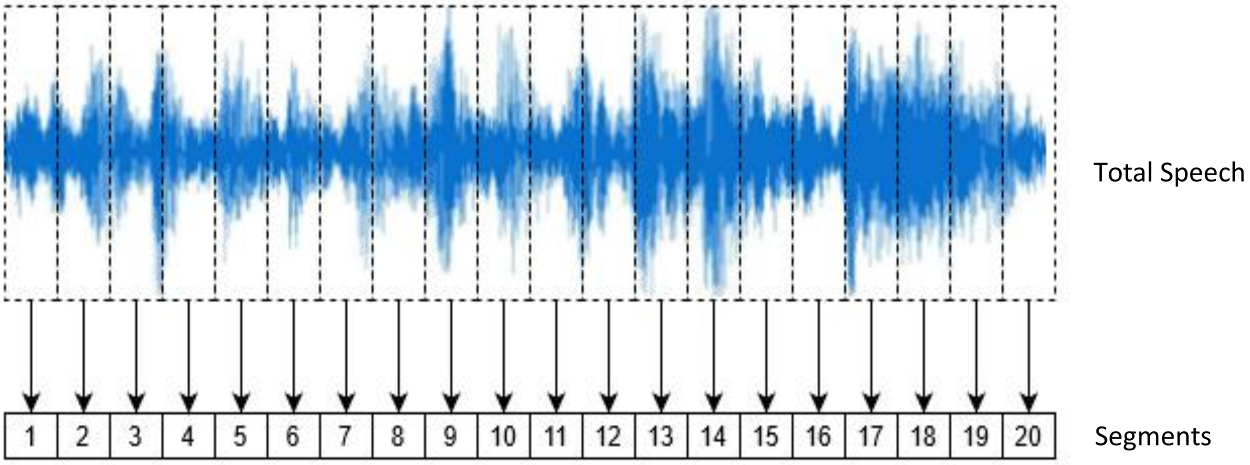


Figure 4: Visualization of an audio clip being segmented

These segments that are created from the complete speech (Figure 4), have different descriptive statistic values. I have shown in Tables 1 and 2 examples of a small portion of parameteres within first two segments of a speech from Mary Lou McDonald on January 30th, 2019.

0-1999	Mean	Std Dev	Skewness	Kurtosis	Int-Q Range
F0final	159.92	44.79	0.389	3.99	23.98
jitterLocal	0.14	0.186	2.56	10.23	0.035
shimmerLocal	0.25	0.2	1.57	5.54	0.08

Table 1: Acoustic features in first 2000ms of speech (segment one)

2000-3999	Mean	Std Dev	Skewness	Kurtosis	Int-Q Range
F0final	170.3	43.55	2.13	6.31	3.61
jitterLocal	0.05	0.09	3.03	13.67	0.01
shimmerLocal	0.16	0.16	2.88	13.56	0.057

Table 2: Acoustic features in second 2000ms of speech (segment two)

Due to the vast number of parameters within these signals, it is not essential to include every one of them. A feature selection stage is implemented to eliminate irrelevant features and maximise the performance of the emotion recognition system. A forward selection system can be implemented to combine the relevant values and removed unnecessary data. This has been shown to improve the accuracy of classification by a great deal (Narayanan, 2005). Once these features have been determined, they are passed onto a classifier. Swain et al., (2018) investigates and reviews many proposed and their pros and cons as classifiers, including the Hidden Markov Model, Neural Networks, Gaussian Mixture Models and many more. This classifier then determines and outputs what emotion is being expressed in the speech segment.

2.2.3.3 Existing Software

There are many existing toolkits which can carry out a process like this pipeline that would apply to the scope of this project.

- OpenSMILE toolkit is an open-source software which performs extraction of features from audio signals. This software has been used in many emotion-recognition projects and very well documented. This project began in Munich in 2008 (Eyben, 2015).
- A similar software is DAVID. This is also an open-source software which performs extraction of features from audio signals. DAVID works with both audio files and can run in real time (Rachman, 2018).
- Praat is a software package created for the analysis of speech and phonetics with its latest version released in 2015. Praat was started at the University of Amsterdam and was developed by Paul Boersma and Weenink (Boersma, 2018).
- The Hidden Markov Toolkit is a widely used speech and pattern recognition system. Its development began in Cambridge University and was first released in 2004.

2.3 Sentiment Analysis

This section will deal with the modality of communication: word choice.

2.3.1 Introduction

“Opinions and related concepts such as sentiments, evaluations, attitudes and emotions are the subject of study of sentiment analysis” (Liu, 2012). Sentiment analysis is the process of identifying and categorizing opinion from text to determine whether this opinion is positive or negative towards a specific topic or product. This natural language process has been used in many fields including business and research. (Ahmad et al., 2016) and (Tetlock, 2007) have used this successfully to validate changes in the stock market with respect to investor sentiment.

Sentiment analysis can be applied on multiple levels. (Kolkur et al., 2015) explains that sentiment analysis on a document level is the simplest form of classification. This refers to a document of text having only one subject throughout meaning that the output can be classified as one core opinion. If multiple subjects appear in a body of text, it is not possible to say only one opinion exists – this requires breaking the body of text down into sentences, if there are multiple opinions within a sentence this can be broken down into individual phrases for analysis.

The application of sentiment analysis in this project is to assess word choice as a modality of communication. Detection of negative and positive valence can be used to cross check against the physical correlate of emotion for validation and to detect leakage of emotion.

2.3.2 Textual Analysis

There are two main ways in which sentiment analysis can be approached: A Lexicon based approach and a machine learning approach. The Lexicon based approach is a rule-based system that uses lists of polarized words (positive and negative) to detect sentiment in a body of text. The machine learning approach is an automatic method, usually modelled as a classification problem whereby a classifier is fed a text and returns a category of sentiment. There also exists hybrid approaches and techniques. (Appel, 2016).

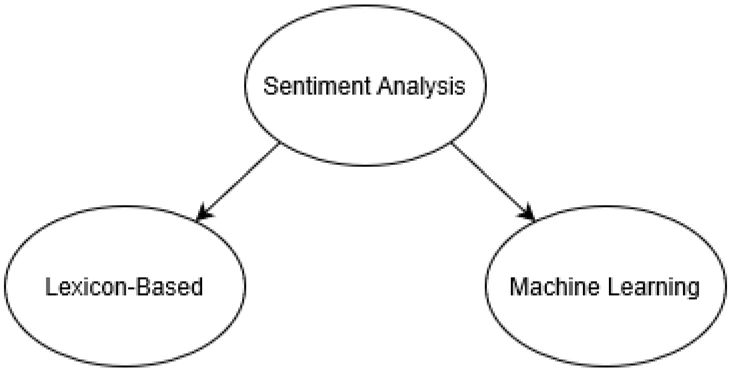


Figure 5: Sentiment analysis approaches

2.3.2.1 Lexicon Based Approach

Lexicon-based (dictionary based) approaches to sentiment analysis are straightforward and accurate. The way in which this approach can be applied is to break down a body of text into single word tokens and perform calculations of sentiment bearing words. Predefined dictionaries of words with positive and negative valence are used to cross check with against the body of text to identify to what extend the document is positive and negative. This approach is called a “bag-of-words” method (Da Silva, 2016). Existing dictionaries can be used such as Harvard IV-4 dictionaries (Ahmad et al., 2016) or can be created from a corpus related words to the given dataset. As in politics speeches/interviews, some words may be construed in a different way to everyday situations. This means that creating dictionaries containing words of negative and positive valence relating to politics may be necessary to detect sentiment accurately in my dataset (section 1-3).

2.3.2.2 Machine Learning Approach

Working with large Lexicon-Based systems can sometimes be quite labour intensive. For this reason, quite a lot of research has gone into the implementation of machine learning in sentiment analysis (Hutto, 2014). The machine learning approach to sentiment analysis entails many different techniques. Some of these techniques are Naïve Bayes, Support Vector Machines, Maximum Entropy, Decision Trees, K-Nearest Neighbour Classifiers, Window Classifiers, Adaboost Classifiers. (Sharma, 2012).

Pang and Lee (2002) investigate and discuss the performance of some of these machine learning approaches. Naïve Bayes classifier is a supervised learning algorithm that is based on Bayes Theorem. This assumes that each word in a sentence is independent of each other and because of this assumption it reduces a large amount computation. Despite its level of simplicity, Naïve Bayes method has been found to be very accurate (Lewis, 1998). Support Vector Machines (SVM) are also a supervised machine learning algorithm. The approach of SVMs is to take the plot of a dataset that is not linearly separable and map it to a new area where it can be separated linearly. SVMs provide lenience in classification (Mullen, 2004). The Maximum Entropy technique is done by assigning context to classes and checking if documents belong to these classes. No biases are present in the system upon maximising of entropy (Mehra, 2002). Each of these techniques are very accurate, with some outperforming others. Overall, the performance of these techniques is similar (Pang, 2002).

Having said that, machine learning techniques does present itself with certain shortcomings. These approaches rely on being trained extensively and possessing specific training data. This training can be quite intensive with respect to both labour and time and training data may be difficult to obtain. In comparison to Lexicon-Based approaches, these machine learning techniques are also more expensive in terms of processing and computing (Hutto, 2014).

2.3.3 Transcript Extraction

To perform textual analysis correctly, it is necessary to have a completely accurate transcript of each speech/interview within the dataset (referring to section 1-3). There are many potential ways in which these transcripts can be retrieved.

Brute force extraction of transcripts can be done by listening to audio clips and writing down what is being said. Although this is perhaps the most accurate method, it is also extremely labour intensive. Brute force extraction is the ideal approach for transcribing short audio clips but for a large dataset, like in this project, it is not a feasible approach.

There exist online transcribing tools such as “Trint.com”, “transcribe.wreally.com”, “happyscribe.com”, AWS Transcribing tool and many more. These are highly reliable and accurate software systems, and for this reason they can be quite expensive for large scale extraction. As the content of my dataset for this project is substantial, the approach of using one of these transcribing tools may not be possible.

A plausible solution is to write a python script to extract each transcript. “SpeechRecognition” is a python library which makes use of Google’s Cloud Speech API. It is free to use and can be executed through the terminal command line. This may be the most suitable approach for my project with respect to the size of my dataset.

2.4 Data Comparison and Fusion

2.4.1 Statistical Comparison

When assessing and comparing statistics derived from both automatic speech recognition and sentiment analysis there are certain formulae of descriptive statistics that can be used. These can be referred to as “moments” of data.

Mean: The mean, commonly denoted as μ, is the average value of a set of numbers. This is the first moment of data. This can be calculated by adding all numbers within a set to find the sum and dividing this sum value by the total number of numbers within the set. The formula to find μ can be described as:

$$\mu = \frac{\{x_1 + x_2 + x_3 + \cdots + x_n\}}{n}$$

Standard Deviation: The standard deviation, denoted by σ, is used to measure how far a group is spread out from the mean. This is the second moment of data. If the standard deviation is low, it means that the set is generally close to the mean. If the standard deviation is high, it means that there is a large range within the set of numbers. The formula for standard deviation is:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$$

Skewness: Skewness, denoted by S, is used to measure the asymmetry of a probability distribution. This is the third moment of data. If there is a positive skew in the distribution, this implies that tail is on the left-hand side. If there is a negative skew in the distribution, this implies that the tail is on the right-hand side. Skewness can be described by the formula:

$$S = \frac{1}{n}\sum_{1=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

Kurtosis: Kurtosis, denoted by k, is used to measure extreme values in either tail of the distribution. This is the fourth moment of data. This helps to explain if the curve of a distribution is either normal or abnormal. It can be described with the following formula:

$$k = \frac{1}{n}\sum_{1=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4$$

Z-Score: A z-score measure the number of standard deviations that a value within the distribution is away from the mean. This can be expressed by the formula:

$$z\ score = \frac{x - \mu}{\sigma}$$

Correlation: Correlation refers to two values to change or fluctuate together, whether that be in a positive or negative way. This distinguishes to what two sets of data share a relationship. This can be calculated with the formula:

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

Where cov is the covariance and can be calculated with:

$$Cov_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)}$$

2.4.2 Leakage of Emotion

"*Sometimes it is not what is said that was most important, but how it is said*" Picard (2001).

One can ascertain that when a person is speaking of a joyous event or topic, they will usually express a positive emotion (happiness, etc.), as well as use words of positive valence. Correspondingly, when a person is speaking of an event or topic that is distressing, they will express negative emotion (anger, etc.) as well as using words of negative disposition. If this is true between two modes of communication (speech and word choice), it can be said that the speaker is being transparent. However, this is not always the case. If a speaker is not expressing the same emotion across all modalities simultaneously, this can imply that the speaker is concealing something from the listener. For example, if someone is attempting to remain positive while feeling nervous, there may be a combination of positive word choice with a negative expression of emotion. This is regarded as a 'leakage of emotion' and may provide an argument that the speaker is trying to deceive the listener (Eckman, 1969). A machine is said to have an 'emotional intelligence' if it can automatically detect the emotion being expressed.

As mentioned briefly in section 2-1, Dietrich et al., (2019) carried out a study of the process US judges take to decide the outcome of judicial cases. This was done by examining the way which the judge addressed the defendants and prosecution over the course of the trial. By observing parameters of the judge's voice, different levels of emotional arousals were detected in the vocal pitch. It was found that the vocal pitch could be correlated to the way in which the case was settled. This is a prime example of emotional leakage as judges are supposed to remain neutral throughout the course of the trial.

3. Method

This chapter will show the methods used to carry out this project. As can be seen from the below system architecture diagram (Figure 6), this project was done in four main stages, gathering and pre-processing of data, emotional classification of speech acoustics, textual analysis of this data and the fusion of the two modalities.

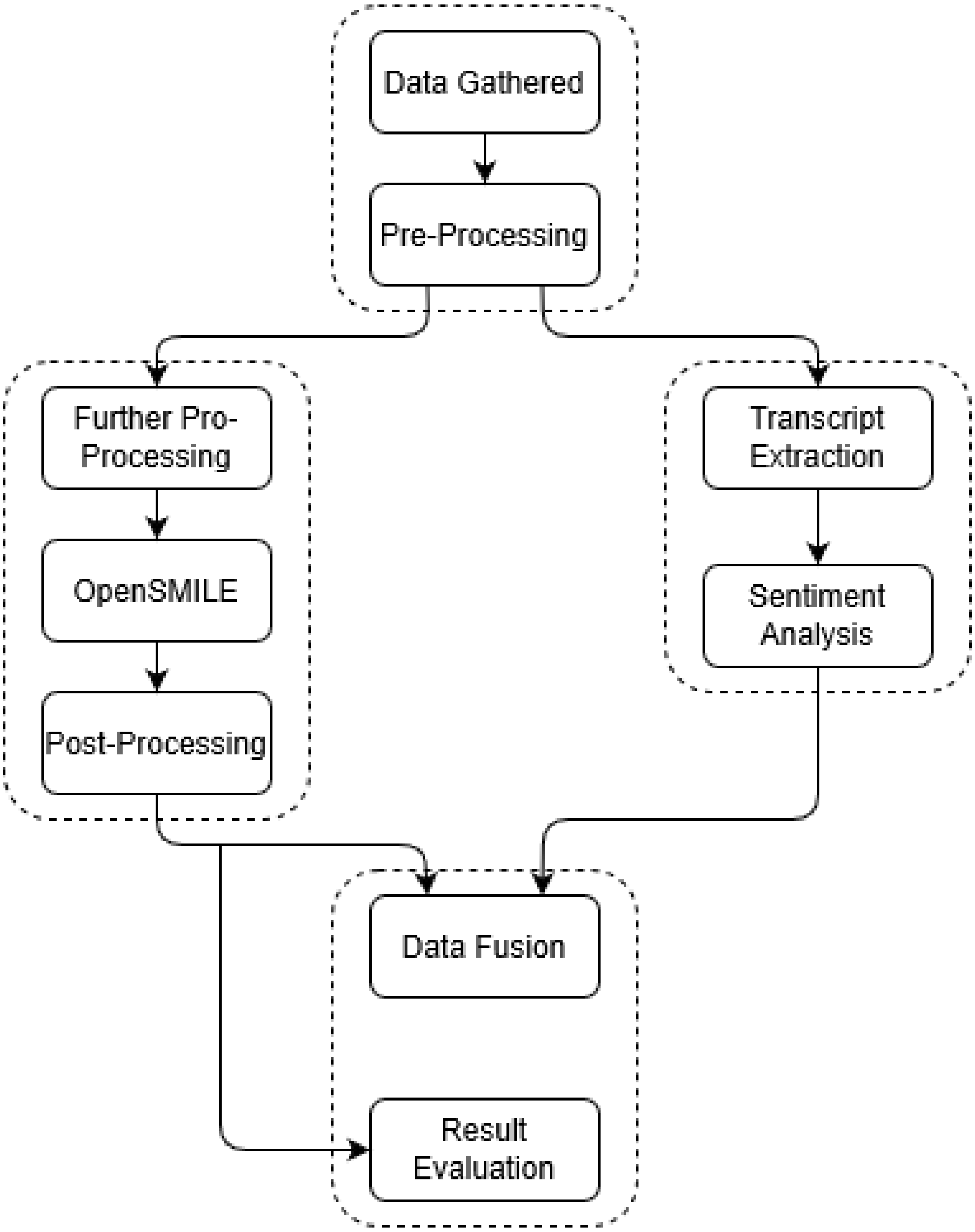


Figure 6: Full System Architecture

3.1 Data Chosen

3.1.1 Data Description

The first step to proceed to our methods was to collect the correct data. This dataset I chose consisted of politicians who are highly trained at speaking in a rational way. The same data is used for both acoustic feature extraction as well as analysis of word choice.

I based my data acquisition around the 2019 UK political elections. This dataset consists of 9 politicians (5 females, 4 males) spread out across the United Kingdom and Ireland with ages ranging between 40 and 70.

The dataset has a total of 91 speeches/interviews and an overall runtime of 4 hours, 38 minutes and 18 seconds (after pre-processing). All speeches/interviews were gathered ethically from public sources, mainly YouTube as well as other media outlets. (All links and dates can be found on Page 58).

Politician	Age	Gender	Nationality	Data Samples	Total Duration	% of Database
Arlene Foster	49	Female	Northern Irish	11	33m 33s	12.05%
Boris Johnson	55	Male	English	10	31m 35s	11.34%
Jeremy Corbyn	70	Male	English	9	25m 37s	9.2%
Leo Varadkar	41	Male	Irish	11	33m 00s	11.86%
Mary Lou McDonald	50	Female	Irish	13	41m 21s	14.86%
Nicola Sturgeon	49	Female	Scottish	10	31m 51s	11.44%
Nigel Farage	56	Male	England	10	32m 28s	11.66%
Rebecca Long Bailey	40	Female	England	7	21m 52s	7.86%
Theresa May	63	Female	England	9	27m 01s	9.71%

Table 3: Complete description of dataset

3.1.2 Data Gathering and Pre-Processing

The way in which I acquired this data from publicly available sources was by searching for each politician with a month of the year (e.g. Boris Johnson March 2019) and finding a speech or interview with a length of approximately 3 minutes. After listening to the clip and ensuring it was of high quality, I used the macOS application QuickTime Player to record the full screen. I then exported this recording to iMovie where I could pre-process each speech/interview.

In the analysis of both speech and word choice, we are only interested in the politician's voice. Often in speeches, there are prolonged pauses and extended applause which are unnecessary for examination. Interruptions during interviews are also very common by interviewers and/or other interview participants. Data trimming could easily be done using iMovie to remove unwanted sections of each speech/interview.

The image below shows how the precision of iMovie allows us to split the speech before and after a long portion of applause. The split portion can then be deleted.

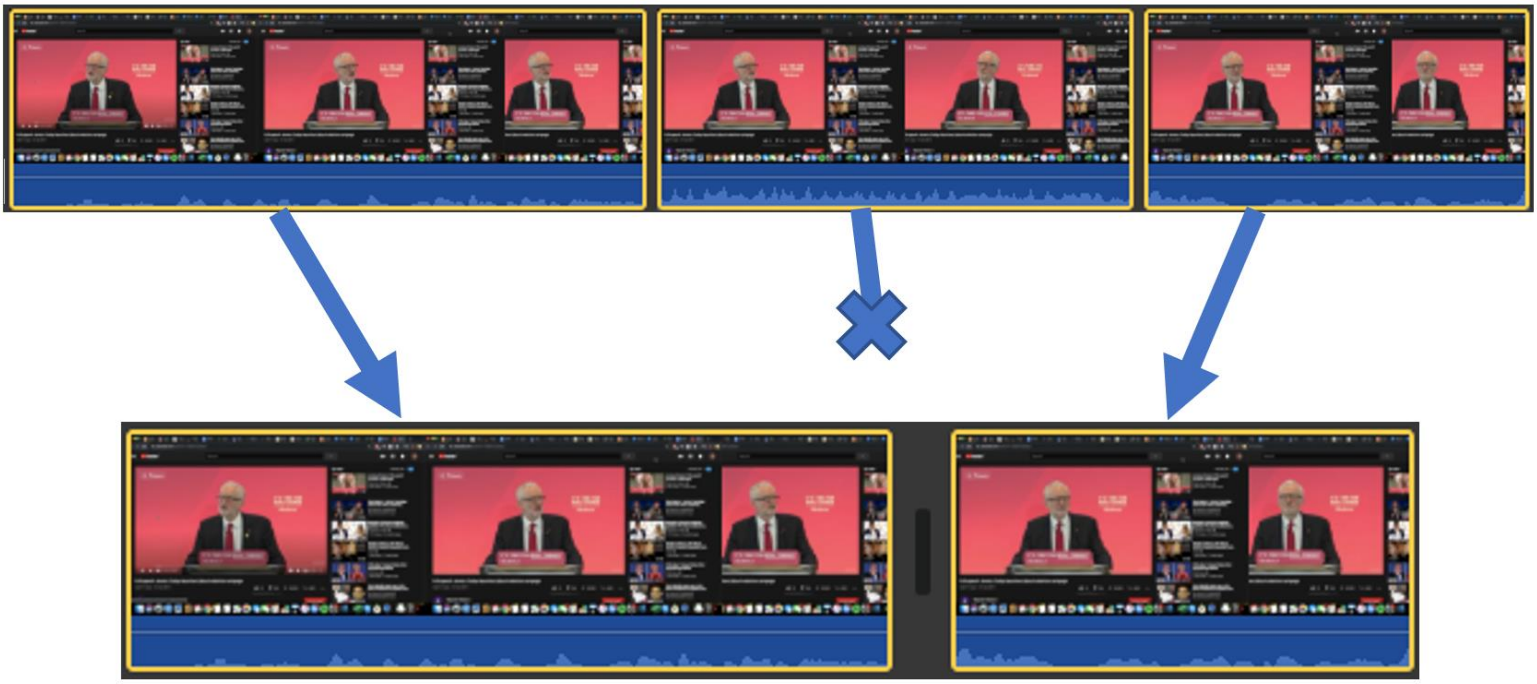


Figure 7: Trimming audio clip

Once these unwanted sections are deleted, the clip can be refused and saved. As iMovie saves in .mp4 format, I had to convert these to .wav files. This could easily be done by the following command line terminal:

```
ffmpeg -i <infile> -ac 2 -f wav <outfile>
```

This saves the file in the desired format and is now ready to be processed.

3.2 Design

3.2.1 Overview

Once the dataset was successfully pre-processed, I used the OpenSMILE feature extraction toolkit to extract the acoustics of speech and used Lexicon based method to perform sentiment analysis on each politicians' transcripts.

3.2.2 Speech Recognition System

As discussed in section 2-3-6, there are many high-quality software systems related to the extraction of emotion in speech. I decided to use OpenSMILE as it is very well documented and has been cited in over 1000 scientific papers. This toolkit is called openSMILE as it is an open-source software freely available for research purposes and SMILE is an acronym of Speech and Music Interpretation by Large feature Extraction. The development of the OpenSMILE toolkit started at Technische Universität München (TUM) in 2008 by Florian Eyben**,** Martin Wöllmer and Björn W. Schuller. OpemSMILE provides a simple, scriptable console application where modular feature extraction components can be freely configured and connected via a single configuration file. (Eyben, 2015).

3.2.2.1 OpenSMILE Architecture

The architecture of OpenSMILE has many modules and is designed for an incremental flow of data. This flow of data is handled by a central component, CDataMemory (Central Data Memory). cDataMemory manages multiple data memory ‘levels’ internally. A single component can write data to these dataMemory ‘levels’ and multiple components can read this data. There are data-source components which can read data from files or other external sources and introduce them to the central dataMemory. Along with these data-source components, there are also data-processor components. These data-processor components read data, make modifications and save to a new dataMemory ‘level’. These new dataMemory levels serve as feature extraction components. Data-sink components read the final data and save them to files. A visualization of this system can be seen in the figure below. (Eyben, 2015).

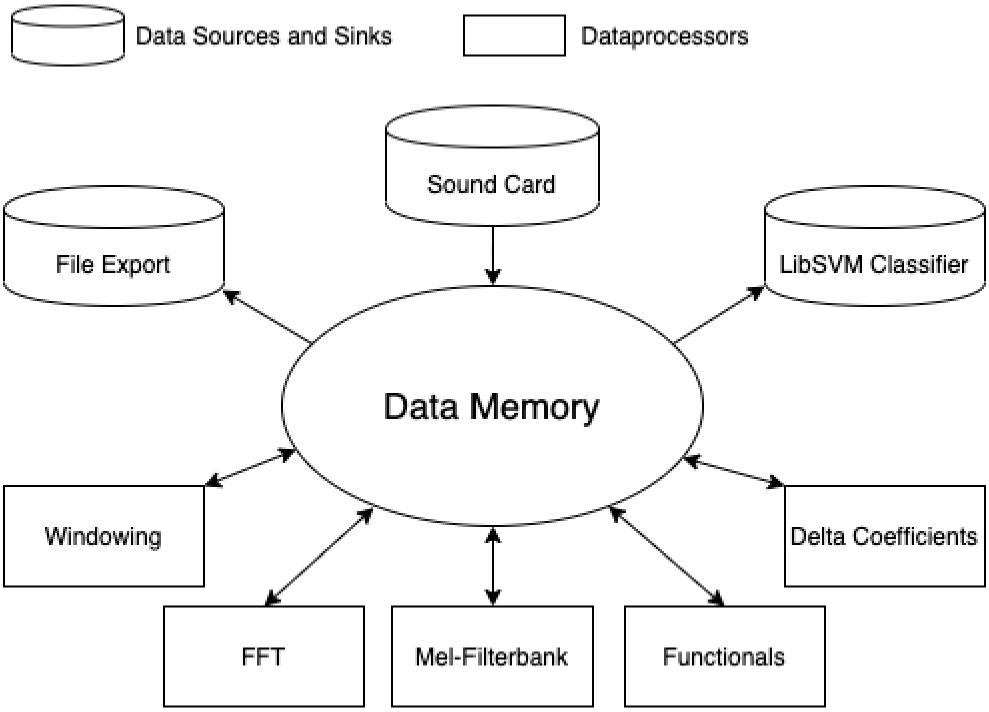


Figure 8: Complete OpenSMILE Architecture

There is a common functionality between all components which process the data and connect to dataMemory, they are all derived from a single base class cSmileComponent. The figure below shows the hierarchy in which these class is ordered and the connected between the cDataWriter and cDataReader components to the dataMemory which is displayed with dotted lines. (Eyben, 2015).

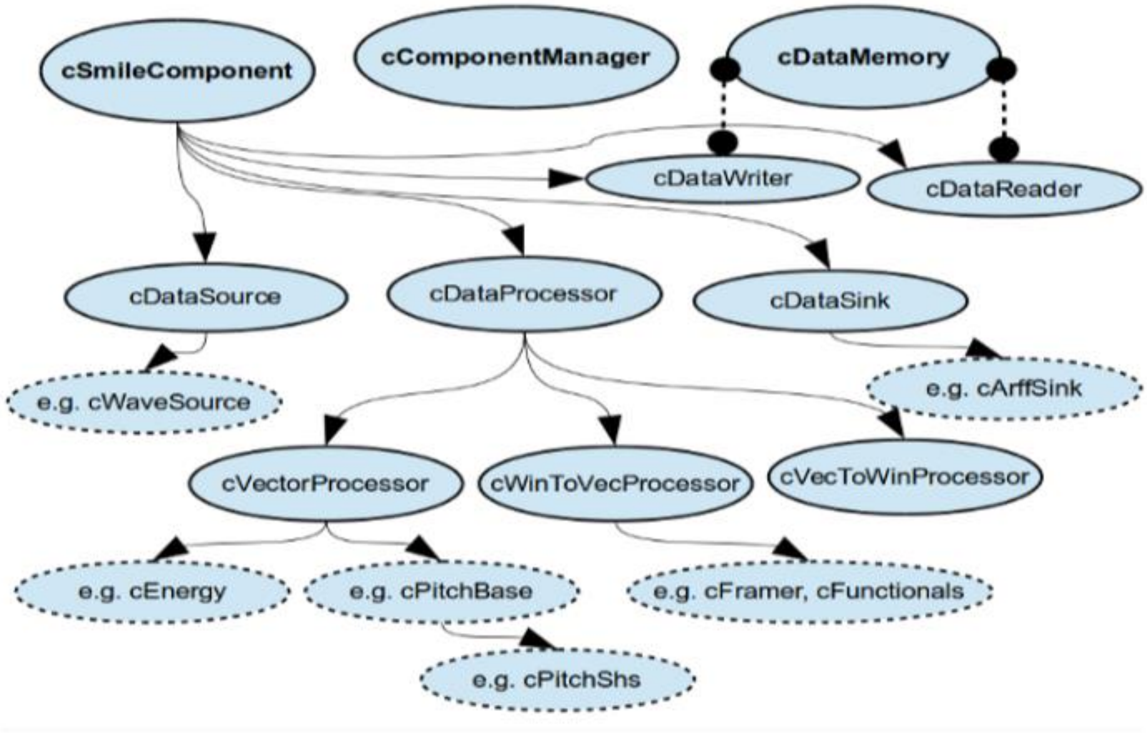
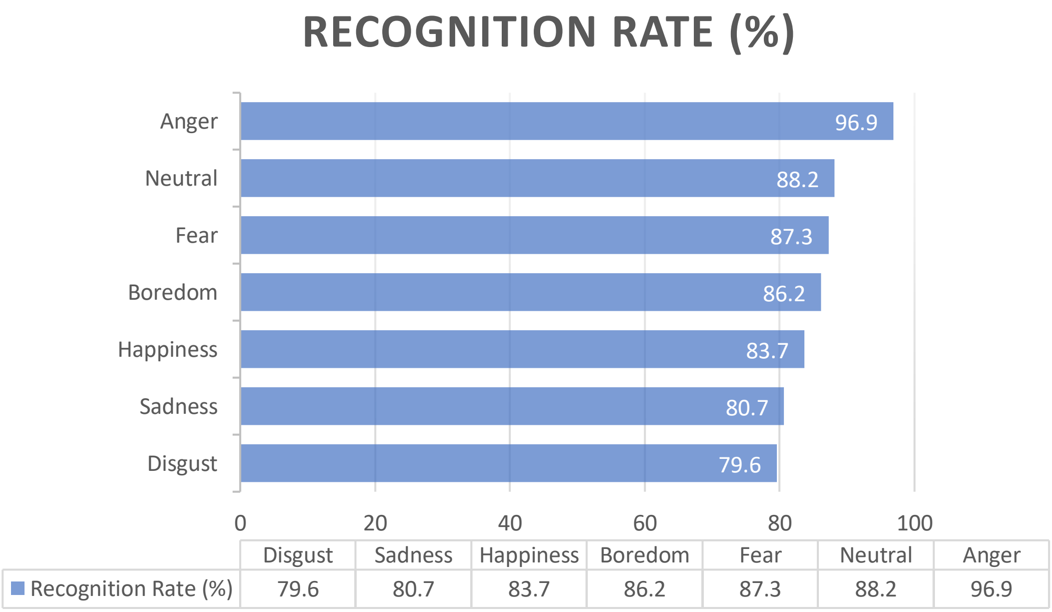


Figure 9: OpenSMILE flow of data

3.2.2.2 OpenSMILE Database

As described briefly in section 2-3-6, OpenSmile uses a database that was developed in Technische Universität Berlin (TUM) to detect emotion. This database is comprised of ten actors (5 male and 5 females) who simulated the emotions “anger”, “fear”, “neutral”, “joy”, “sadness”, “disgust” and “boredom” across 10 utterances. These utterances were common everyday sayings and are interpretable in all emotions which were applied. To achieve high audio quality the recordings took place in an anechoic at TUM. Each recording was under the supervision of three phoneticians, two of them giving instructions and feedback and one monitoring the recording equipment.

A perception test was carried out to ensure the emotional quality of the database. 20 people took part in this test, where they listened to each recording once and decided which emotional state the actor was trying to convey. The recognition rate is displayed in the figure below. (Burkhardt. F, et al., 2005).

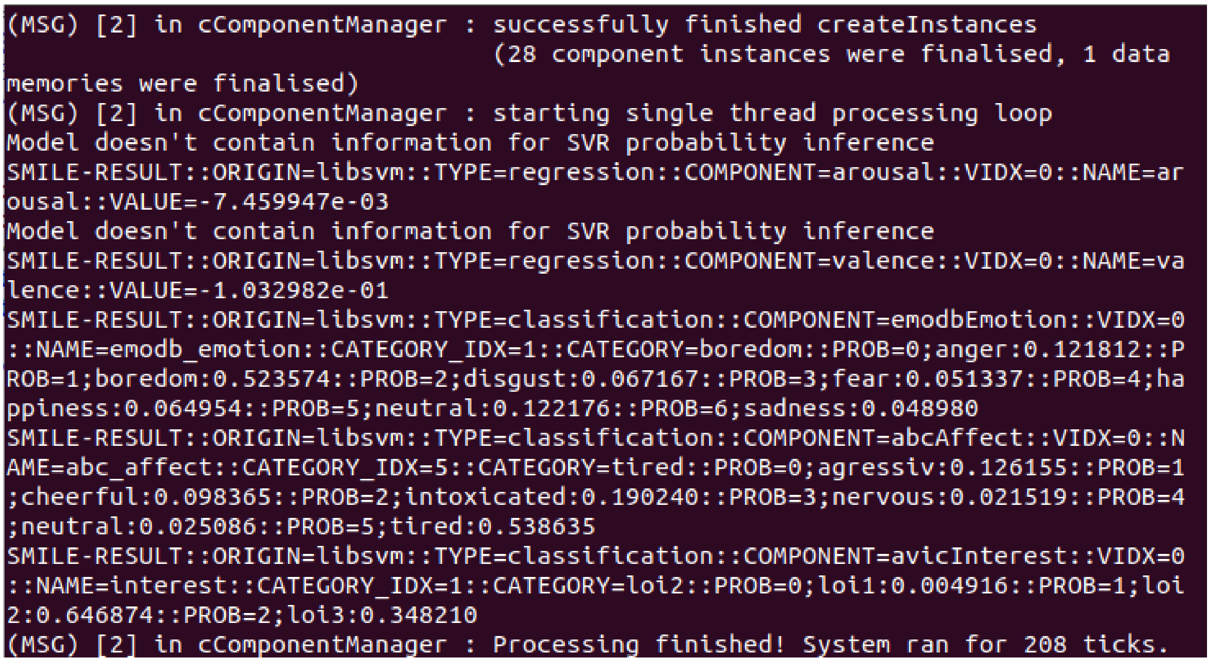


	Disgust	Sadness	Happiness	Boredom	Fear	Neutral	Anger
■ Recognition Rate (%)	79.6	80.7	83.7	86.2	87.3	88.2	96.9

Figure 10: OpenSMILE database (Emo-DB) recognition rate

3.2.2.3 OpenSMILE Process and Output

To successfully process each audio clip through OpenSMILE, they must be split into frames of 2000ms. I used a python script to split each interview/speech into these 2000ms segments. Once this was achieved, I passed each segment through the OpenSMILE system to extract its features. An example of what raw OpenSMILE output looks like can be seen below in Figure 11. Using a batch script through the Linux terminal, I processed each frame, saving the output for each speech/interview as a text file.



```
(MSG) [2] in cComponentManager : successfully finished createInstances
                                 (28 component instances were finalised, 1 data
memories were finalised)
(MSG) [2] in cComponentManager : starting single thread processing loop
Model doesn't contain information for SVR probability inference
SMILE-RESULT::ORIGIN=libsvm::TYPE=regression::COMPONENT=arousal::VIDX=0::NAME=ar
ousal::VALUE=-7.459947e-03
Model doesn't contain information for SVR probability inference
SMILE-RESULT::ORIGIN=libsvm::TYPE=regression::COMPONENT=valence::VIDX=0::NAME=va
lence::VALUE=-1.032982e-01
SMILE-RESULT::ORIGIN=libsvm::TYPE=classification::COMPONENT=emodbEmotion::VIDX=0
::NAME=emodb_emotion::CATEGORY_IDX=1::CATEGORY=boredom::PROB=0;anger:0.121812::P
ROB=1;boredom:0.523574::PROB=2;disgust:0.067167::PROB=3;fear:0.051337::PROB=4;ha
ppiness:0.064954::PROB=5;neutral:0.122176::PROB=6;sadness:0.048980
SMILE-RESULT::ORIGIN=libsvm::TYPE=classification::COMPONENT=abcAffect::VIDX=0::N
AME=abc_affect::CATEGORY_IDX=5::CATEGORY=tired::PROB=0;agressiv:0.126155::PROB=1
;cheerful:0.098365::PROB=2;intoxicated:0.190240::PROB=3;nervous:0.021519::PROB=4
;neutral:0.025086::PROB=5;tired:0.538635
SMILE-RESULT::ORIGIN=libsvm::TYPE=classification::COMPONENT=avicInterest::VIDX=0
::NAME=interest::CATEGORY_IDX=1::CATEGORY=loi2::PROB=0;loi1:0.004916::PROB=1;loi
2:0.646874::PROB=2;loi3:0.348210
(MSG) [2] in cComponentManager : Processing finished! System ran for 208 ticks.
```

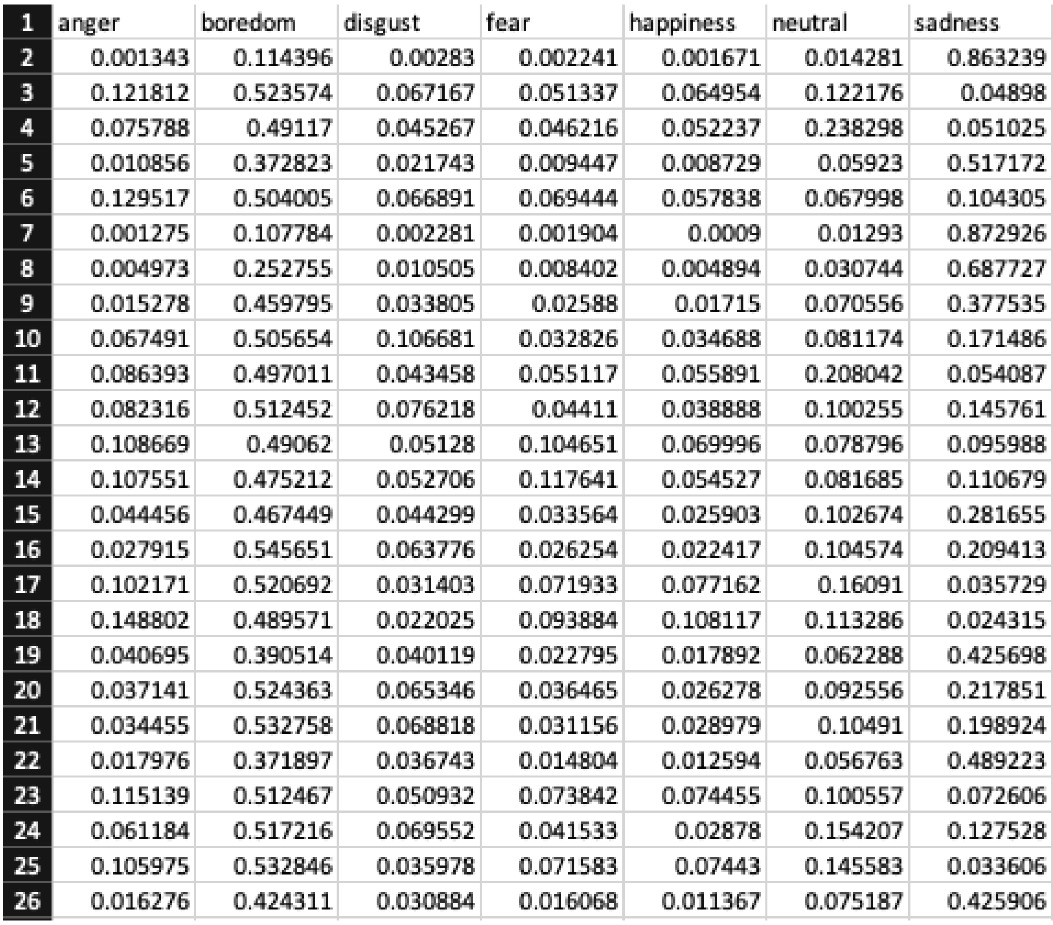
Figure 11: OpenSMILE raw output

As seen in Figure 11 raw data from OpenSMILE is quite hard to interpret and requires post-processing. The emotions output from OpenSMILE are “anger”, “boredom”, “disgust”, “fear”, “happiness”, “neutral” and “sadness”. These emotion labels are followed by their corresponding value for that specific frame. To retrieve the desired values, I opened these text files in Sublime Text (text editor platform) which has the option of searching by regex commands.

By searching for the regex string:

(anger\:[0-9]\.[0-9])|(boredom\:[0-9]\.[0-9]*)|(disgust\:[0-9]\.[0-9]*)|(fear\:[0-9]\.[0-9]*)|(happiness\:[0-9]\.[0-9]*)|(neutral\:[0-9]\.[0-9]*::PROB=6)|(sadness\:[0-9]\.[0-9]*)*

This selected all desired information from each text file and could easily be copied and exported to an Excel spreadsheet, where values could be sorted by their respective emotions. OpenSMILE output in post-processed form can be seen in Figure 12.



1	anger	boredom	disgust	fear	happiness	neutral	sadness
2	0.001343	0.114396	0.00283	0.002241	0.001671	0.014281	0.863239
3	0.121812	0.523574	0.067167	0.051337	0.064954	0.122176	0.04898
4	0.075788	0.49117	0.045267	0.046216	0.052237	0.238298	0.051025
5	0.010856	0.372823	0.021743	0.009447	0.008729	0.05923	0.517172
6	0.129517	0.504005	0.066891	0.069444	0.057838	0.067998	0.104305
7	0.001275	0.107784	0.002281	0.001904	0.0009	0.01293	0.872926
8	0.004973	0.252755	0.010505	0.008402	0.004894	0.030744	0.687727
9	0.015278	0.459795	0.033805	0.02588	0.01715	0.070556	0.377535
10	0.067491	0.505654	0.106681	0.032826	0.034688	0.081174	0.171486
11	0.086393	0.497011	0.043458	0.055117	0.055891	0.208042	0.054087
12	0.082316	0.512452	0.076218	0.04411	0.038888	0.100255	0.145761
13	0.108669	0.49062	0.05128	0.104651	0.069996	0.078796	0.095988
14	0.107551	0.475212	0.052706	0.117641	0.054527	0.081685	0.110679
15	0.044456	0.467449	0.044299	0.033564	0.025903	0.102674	0.281655
16	0.027915	0.545651	0.063776	0.026254	0.022417	0.104574	0.209413
17	0.102171	0.520692	0.031403	0.071933	0.077162	0.16091	0.035729
18	0.148802	0.489571	0.022025	0.093884	0.108117	0.113286	0.024315
19	0.040695	0.390514	0.040119	0.022795	0.017892	0.062288	0.425698
20	0.037141	0.524363	0.065346	0.036465	0.026278	0.092556	0.217851
21	0.034455	0.532758	0.068818	0.031156	0.028979	0.10491	0.198924
22	0.017976	0.371897	0.036743	0.014804	0.012594	0.056763	0.489223
23	0.115139	0.512467	0.050932	0.073842	0.074455	0.100557	0.072606
24	0.061184	0.517216	0.069552	0.041533	0.02878	0.154207	0.127528
25	0.105975	0.532846	0.035978	0.071583	0.07443	0.145583	0.033606
26	0.016276	0.424311	0.030884	0.016068	0.011367	0.075187	0.425906

Figure 12: OpenSMILE output post-processed

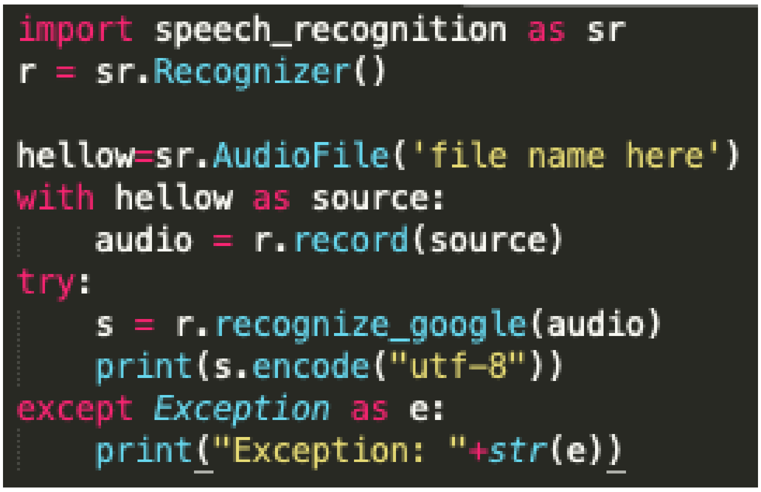
This shows the first 25 frames of a speech from Mary Lou McDonald on January 30th, 2019, with each column representing 2000ms and the values representing the percentage of each emotion within these frames of their respective columns.

3.2.3 Sentiment Analysis

The aim of this section is to identify any extract any sentiment present from each politician's speeches/interviews. The first step of this section was to retrieve the transcript of each speech/interview in full. Once these transcripts were obtained, I performed a textual analysis. This study opted for the use of a Lexicon-Based approach.

3.2.3.1 Transcript Extraction

As discussed in section 2-3-3, there are many ways to extract the transcript of an audio file. I decided to write a python script to process each audio file as this is an accurate, free and efficient approach. By watching YouTube documentaries and reading discussions on Stack Overflow I was able to write this program as shown below in Figure 13.



```
import speech_recognition as sr
r = sr.Recognizer()

hellow=sr.AudioFile('file name here')
with hellow as source:
    audio = r.record(source)
try:
    s = r.recognize_google(audio)
    print(s.encode("utf-8"))
except Exception as e:
    print("Exception: "+str(e))
```

Figure 13: Transcript extraction python script

This relatively short python program had a short processing time (approximately 1 minute for a 3-minute speech/interview on my local machine). The accuracy of this transcription method was also very impressive. I validated that the transcripts were correct by listening to each speech/interview while reading through its corresponding transcript to check for errors on the programs end.

3.2.3.2 Textual Analysis

As outlined in section 2-3-2, there are many ways to approach sentiment analysis. I opted for the use of a Lexicon-Based approach rather than a machine learning system as this has been successfully used in many projects, including the validation of changes in stock market changes with investor sentiment (Ahmad et al., 2016) (Tetlock, 2007). The two main sentiment proxies measured with this method will be negative sentiment and positive sentiment.

The way in which this method was carried out throughout the course of this project was to build both ‘negative’ and ‘positive’ dictionaries of words and find the occurrence of these ‘positive’ and ‘negative’ words.

I constructed my ‘negative’ and ‘positive’ dictionaries by breaking individual speeches/interviews down into single work tokens. From this I could differentiate between words being used with negative or positive implications. I carried out this method for 27 speeches/interviews (3 per politician) to create negative dictionary of 512 words and a positive dictionary of 799 words.

I then used a “Bag-of-Words” approach to calculate the percentage of negative and positive words being used in the remaining speeches/interviews. A “Bag-of-Words” approach once again breaks down each speech/interview into single words tokens. This approach maintains the frequency of each word and these words can be matched against my constructed dictionaries to produce a sentiment score of either positive or negative. This can be visualised in Table 4.

The following sentence has been taken from the beginning of a speech from Mary Lou McDonald on January 30th, 2019:

"The reason why nurses and midwives are today on picket lines is because of your failures"

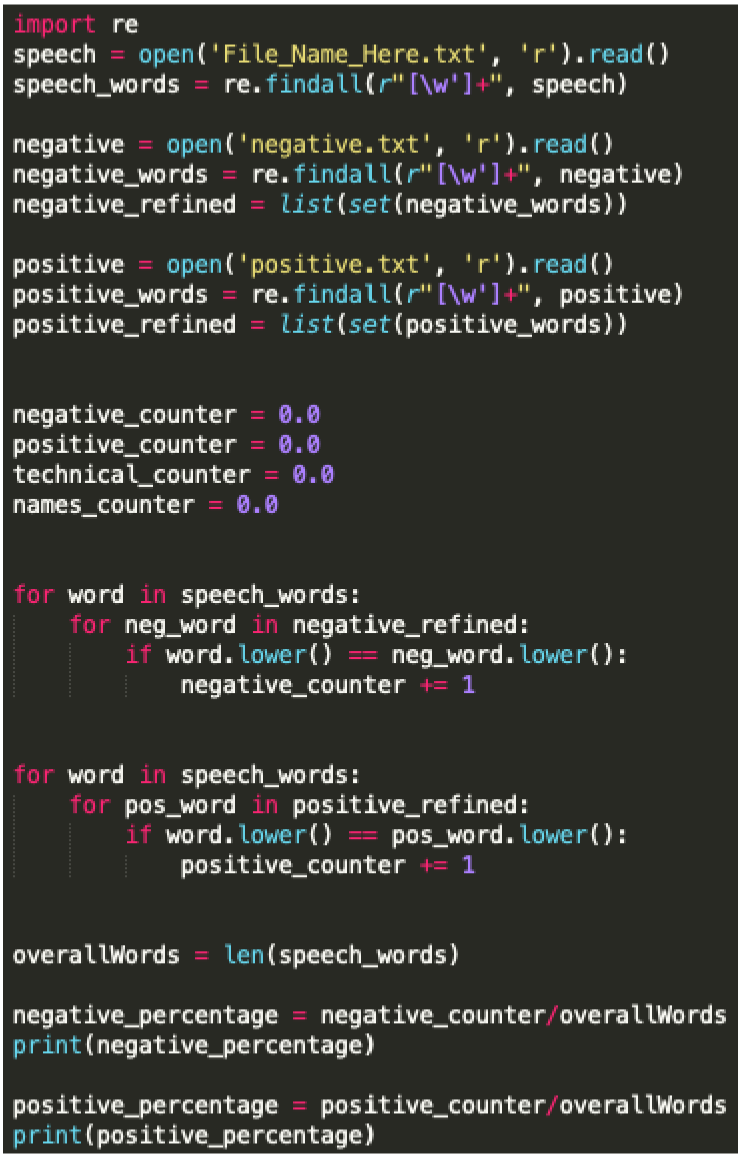
Term	Frequency
The	1
Reason	1
why	1
nurses	1
and	1
midwives	1
are	1
today	1
on	1
picket	1
lines	1
is	1
because	1
of	1
your	1
failures	1

Table 4: Bag-of-words demonstration

In this example there are a total of 15 words and 2 words of negative implication. The negative sentiment score can be calculated by dividing the negative number word count by the total number of words:

Negative Sentiment Score = 2/15 = 13.33%

The same process can be carried out to calculate the positive sentiment score. To carry out this process for each speech/interview I wrote this python script.



```
import re
speech = open('File_Name_Here.txt', 'r').read()
speech_words = re.findall(r"[\w']+", speech)

negative = open('negative.txt', 'r').read()
negative_words = re.findall(r"[\w']+", negative)
negative_refined = list(set(negative_words))

positive = open('positive.txt', 'r').read()
positive_words = re.findall(r"[\w']+", positive)
positive_refined = list(set(positive_words))


negative_counter = 0.0
positive_counter = 0.0
technical_counter = 0.0
names_counter = 0.0


for word in speech_words:
    for neg_word in negative_refined:
        if word.lower() == neg_word.lower():
            negative_counter += 1


for word in speech_words:
    for pos_word in positive_refined:
        if word.lower() == pos_word.lower():
            positive_counter += 1


overallWords = len(speech_words)

negative_percentage = negative_counter/overallWords
print(negative_percentage)

positive_percentage = positive_counter/overallWords
print(positive_percentage)
```

Figure 14: Sentiment analysis calculation python script

This script reads in the desired transcription, negative dictionary and positive dictionary and splits them into single word tokens. Using two nested for loops the script then records the number of matches between the transcription with the negative and positive dictionaries, printing out the negative and positive sentiment scores.

3.2.4 Output Assessment

This section discusses the output of the systems for both modalities: speech and word choice. As mentioned in both the section 1 and 2, the aims of this project are to investigate if the dominant emotional state can be detected from automatic speech recognition and by cross-checking one mode of communication with another if leakage of emotion can be detected.

3.2.4.1 Emotional States

As shown in section 3-2-2-3, we have retrieved the emotion shown in individual 2000ms frames from each speech/interview for each of our politicians. Finding the average of each emotion per speech/interview will result in the emotion with the highest percentage being the dominant emotion.

This can be calculated by simply using the mean formula (section 2-5-1):

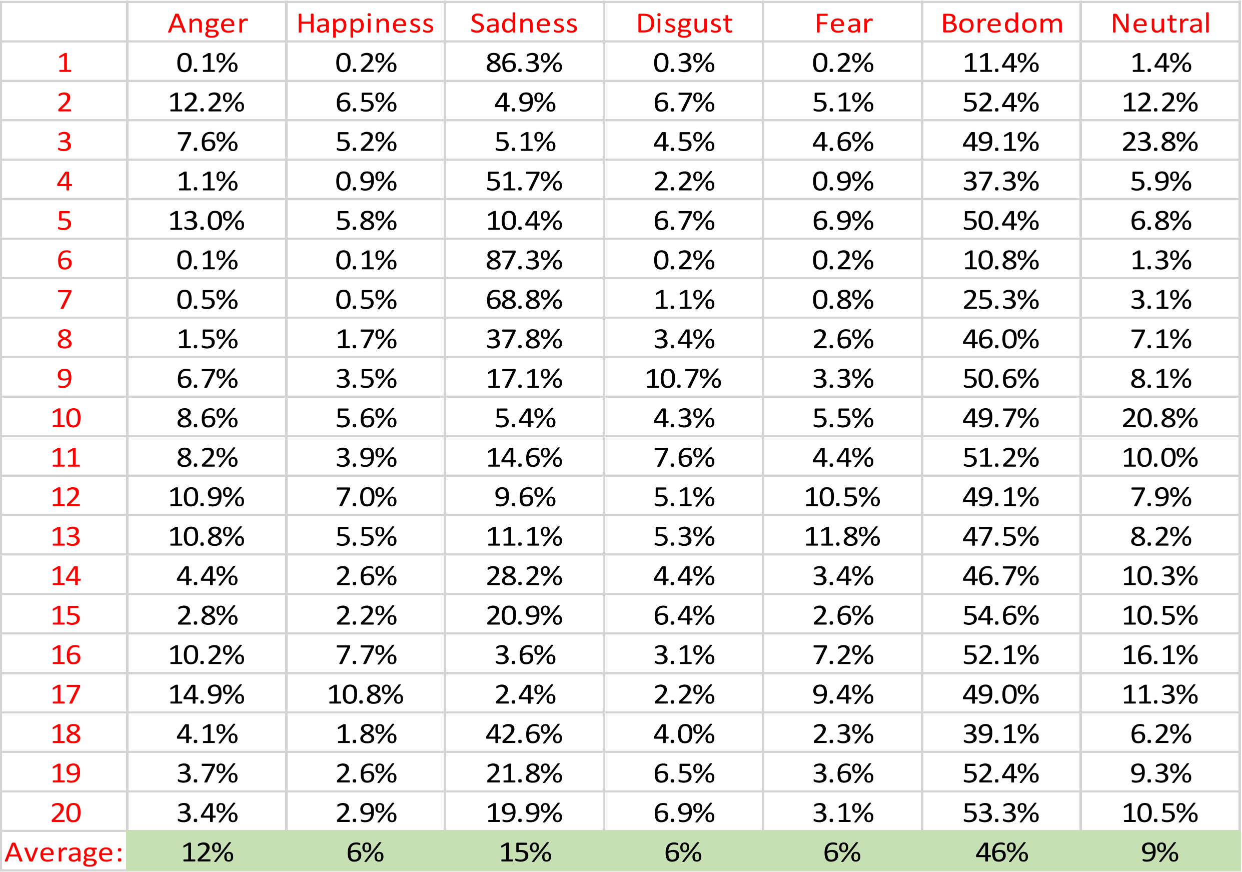
$$\mu = \frac{\{x_1 + x_2 + x_3 + \cdots + x_n\}}{n}$$

Where, m = mean, n = number of frames, x_1= first emotion value, x_n= nth emotion value

These values can be recorded as the mean value of each emotion per speech/interview. The same process can be carried out to find the dominant emotion per politician. This is done by using the same formula and changing,

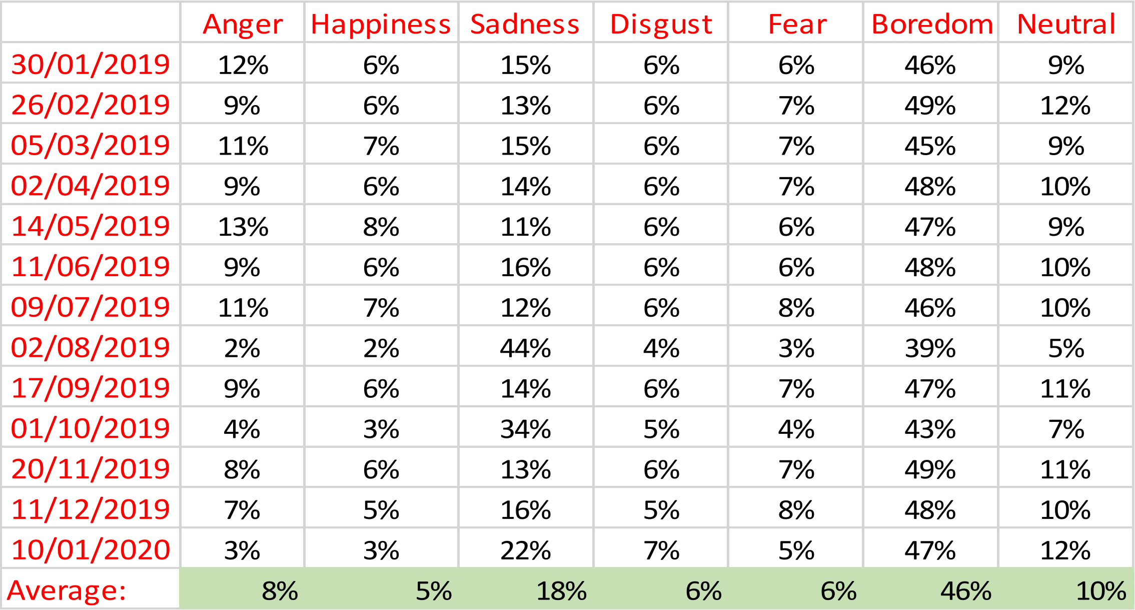
m = mean, n = number of speeches/interviews, x_1= first emotion mean value, x_n= nth emotion mean value

A demonstration of finding the dominant emotion in speech/interview can be seen in Table 5 below and the dominant emotion for a politician in Table 6.



	Anger	Happiness	Sadness	Disgust	Fear	Boredom	Neutral
1	0.1%	0.2%	86.3%	0.3%	0.2%	11.4%	1.4%
2	12.2%	6.5%	4.9%	6.7%	5.1%	52.4%	12.2%
3	7.6%	5.2%	5.1%	4.5%	4.6%	49.1%	23.8%
4	1.1%	0.9%	51.7%	2.2%	0.9%	37.3%	5.9%
5	13.0%	5.8%	10.4%	6.7%	6.9%	50.4%	6.8%
6	0.1%	0.1%	87.3%	0.2%	0.2%	10.8%	1.3%
7	0.5%	0.5%	68.8%	1.1%	0.8%	25.3%	3.1%
8	1.5%	1.7%	37.8%	3.4%	2.6%	46.0%	7.1%
9	6.7%	3.5%	17.1%	10.7%	3.3%	50.6%	8.1%
10	8.6%	5.6%	5.4%	4.3%	5.5%	49.7%	20.8%
11	8.2%	3.9%	14.6%	7.6%	4.4%	51.2%	10.0%
12	10.9%	7.0%	9.6%	5.1%	10.5%	49.1%	7.9%
13	10.8%	5.5%	11.1%	5.3%	11.8%	47.5%	8.2%
14	4.4%	2.6%	28.2%	4.4%	3.4%	46.7%	10.3%
15	2.8%	2.2%	20.9%	6.4%	2.6%	54.6%	10.5%
16	10.2%	7.7%	3.6%	3.1%	7.2%	52.1%	16.1%
17	14.9%	10.8%	2.4%	2.2%	9.4%	49.0%	11.3%
18	4.1%	1.8%	42.6%	4.0%	2.3%	39.1%	6.2%
19	3.7%	2.6%	21.8%	6.5%	3.6%	52.4%	9.3%
20	3.4%	2.9%	19.9%	6.9%	3.1%	53.3%	10.5%
Average:	12%	6%	15%	6%	6%	46%	9%

Table 5: Calculating average emotion for a single speech/interview

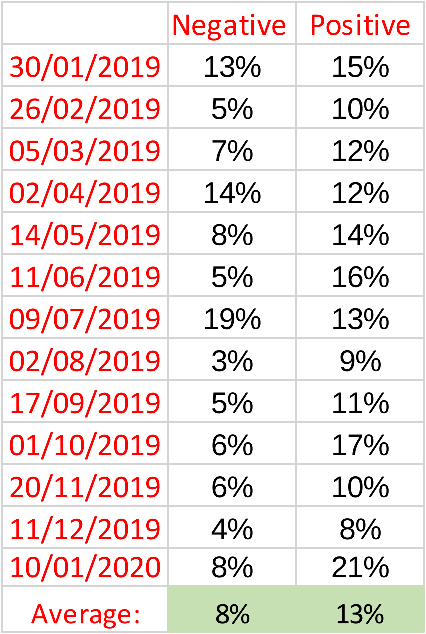


	Anger	Happiness	Sadness	Disgust	Fear	Boredom	Neutral
30/01/2019	12%	6%	15%	6%	6%	46%	9%
26/02/2019	9%	6%	13%	6%	7%	49%	12%
05/03/2019	11%	7%	15%	6%	7%	45%	9%
02/04/2019	9%	6%	14%	6%	7%	48%	10%
14/05/2019	13%	8%	11%	6%	6%	47%	9%
11/06/2019	9%	6%	16%	6%	6%	48%	10%
09/07/2019	11%	7%	12%	6%	8%	46%	10%
02/08/2019	2%	2%	44%	4%	3%	39%	5%
17/09/2019	9%	6%	14%	6%	7%	47%	11%
01/10/2019	4%	3%	34%	5%	4%	43%	7%
20/11/2019	8%	6%	13%	6%	7%	49%	11%
11/12/2019	7%	5%	16%	5%	8%	48%	10%
10/01/2020	3%	3%	22%	7%	5%	47%	12%
Average:	8%	5%	18%	6%	6%	46%	10%

Table 6: Calculating total average emotion for Mary Lou McDonald

3.2.4.2 Sentiment Scores

As discussed in 3-2-2-2, the output of this sentiment analysis method is two values: negative sentiment score and positive sentiment score. This is represented as a percentage of each speech. An overall sentiment scores for each politician can be calculated be calculated by finding the mean sentiment score from all speeches/interviews. An example of this being done for politician Mary Lou McDonald can be seen in Table 7.



	Negative	Positive
30/01/2019	13%	15%
26/02/2019	5%	10%
05/03/2019	7%	12%
02/04/2019	14%	12%
14/05/2019	8%	14%
11/06/2019	5%	16%
09/07/2019	19%	13%
02/08/2019	3%	9%
17/09/2019	5%	11%
01/10/2019	6%	17%
20/11/2019	6%	10%
11/12/2019	4%	8%
10/01/2020	8%	21%
Average:	8%	13%

Table 7: Calculating average sentiment scores

3.2.4.3 Data Fusion

Data fusion is the combination and integration of datasets to derive a better understanding of the individual datasets. This can be performed between the sets of emotional data derived from the acoustics of speech and the sets of sentiment derived from the speaker's word choice, to determine what the relationship is between the two steps. The way in which I did this was to use formulae outlined in section 2-4-1. I calculated the z-score value for each element in both sets being fused and used Excels "=correl" function to calculate the correlation between two sets. An example of such fusion can be seen in Figure 15, where the emotion "anger" is being fused with the sentiment class "negative" for all speeches of Mary Lou McDonald. This can be considering validation at an aggregate level.

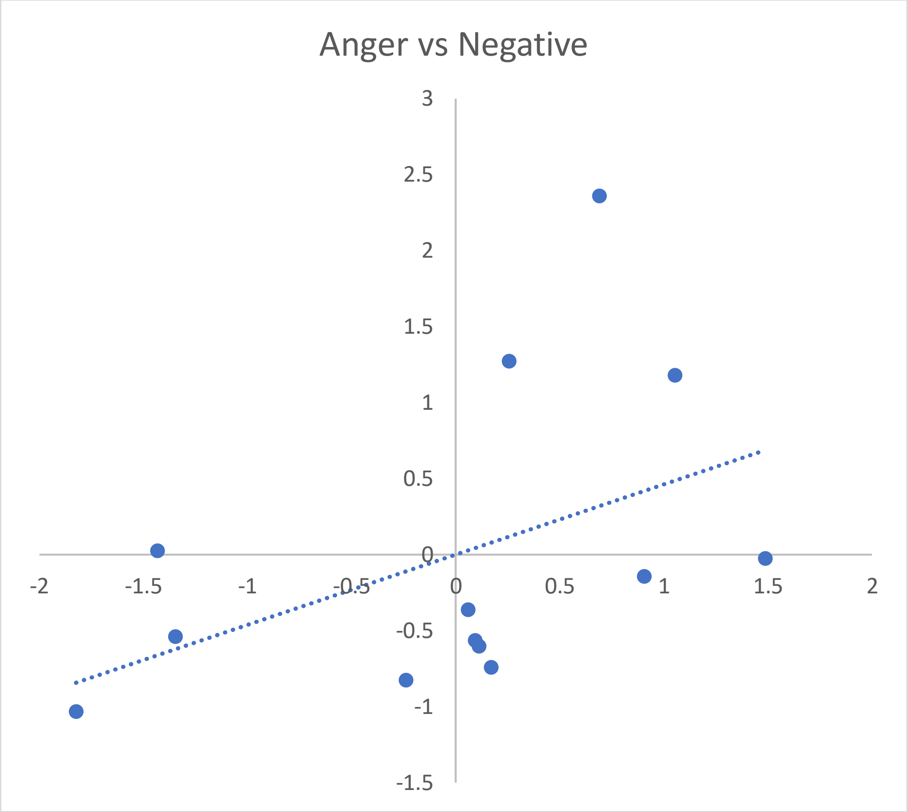


Figure 15: Showing correlation between anger and negative sentiment

A positive correlation of 46% is shown in Figure 15 between “anger” and words which are of “negative” sentiment. This is to be expected considering that anger is associated with negativity. If there is no correlation (0) or a negative correlation (>0), this may seem counter intuitive, but this is the detection of **emotional leakage**. As discussed in section 2-4-2, leakage of emotion is present when speaker is expressing more than one emotion over multiple modalities.

4. Results and Evaluation

This section will discuss results I have obtained from section 3 and discuss different ways in which these results can be interoperated. I will display the results obtained for each politician from both automatic extraction of emotion by OpenSMILE and sentiment analysis.

4.1 Emotions by Politician

In this section I will discuss the derived emotions from OpenSMILE for each politician.

Arlene Foster: Dominant emotional state is boredom (46%).

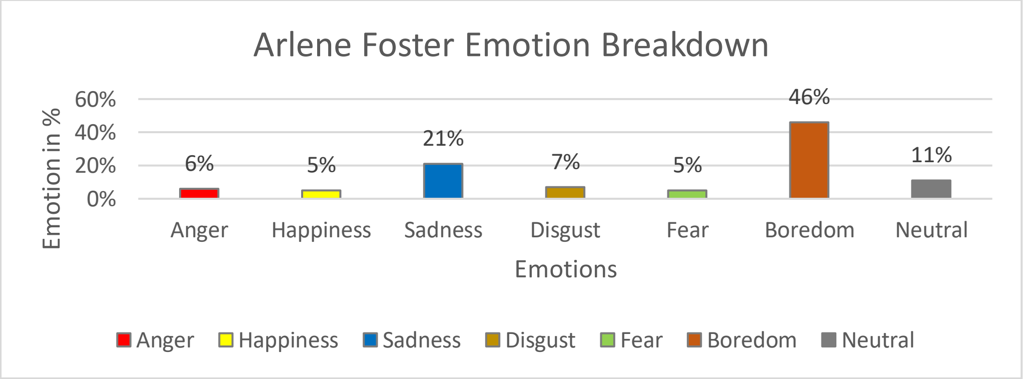


Figure 16: Arlene Foster Emotion Breakdown

Boris Johnson: Dominant emotional state is shared equally between sadness and boredom (39% each).

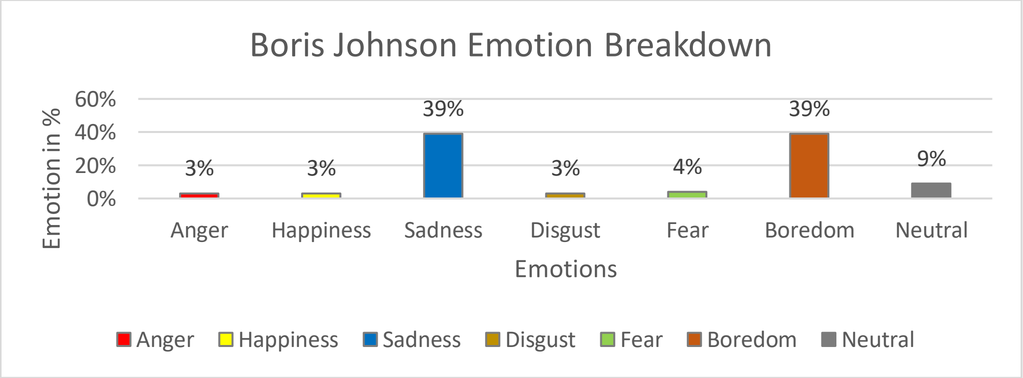


Figure 17: Boris Johnson Emotion Breakdown

Jeremy Corbyn: Dominant emotional state is sadness (50%).

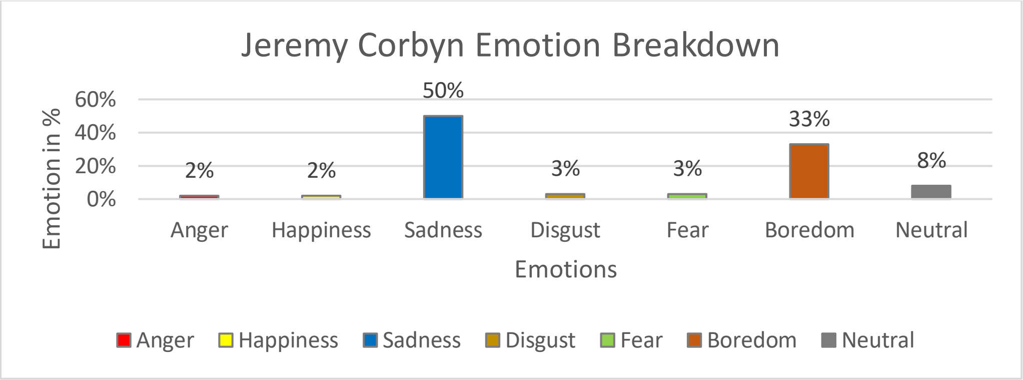


Figure 18: Jeremy Corbyn Emotion Breakdown

Leo Varadkar: Dominant emotional state is boredom (48%).

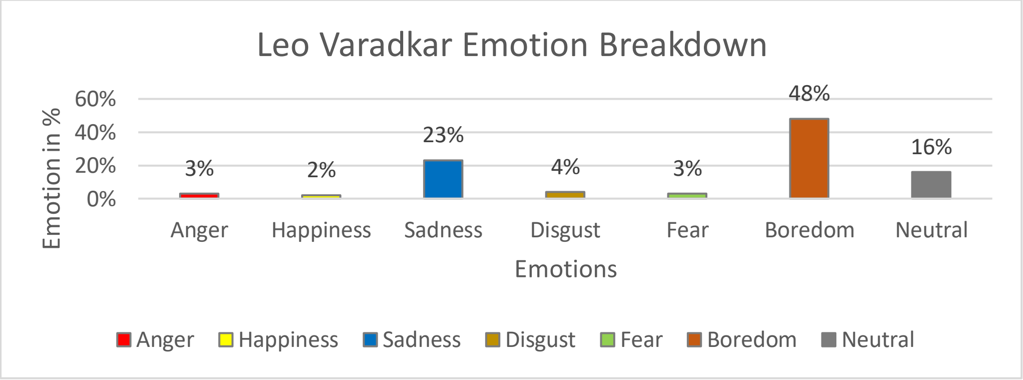


Figure 19: Leo Varadkar Emotion Breakdown

Mary Lou McDonald: Dominant emotional state is boredom (46%).

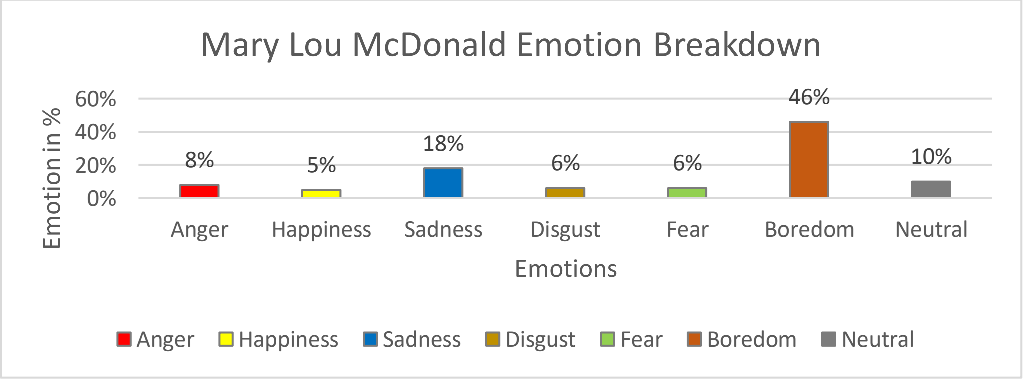


Figure 20: Mary Lou McDonald Emotion Breakdown

Nicola Sturgeon: Dominant emotional state is boredom (44%).

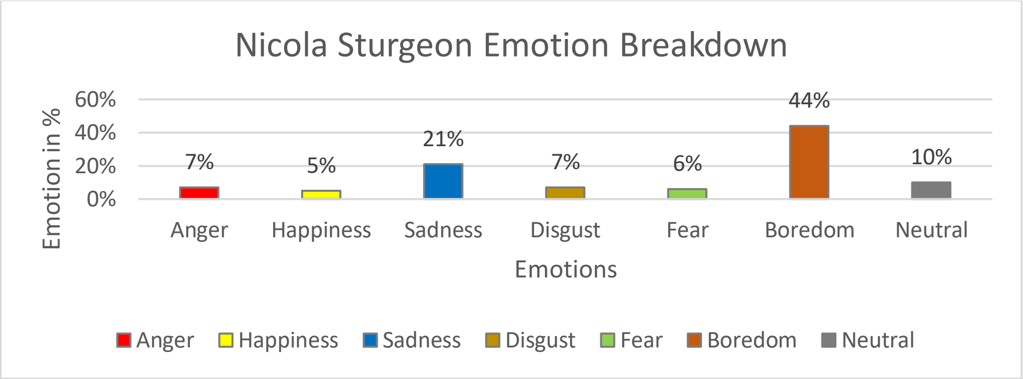


Figure 21: Nicola Sturgeon Emotion Breakdown

Nigel Farage: Dominant emotional state is boredom (47%).

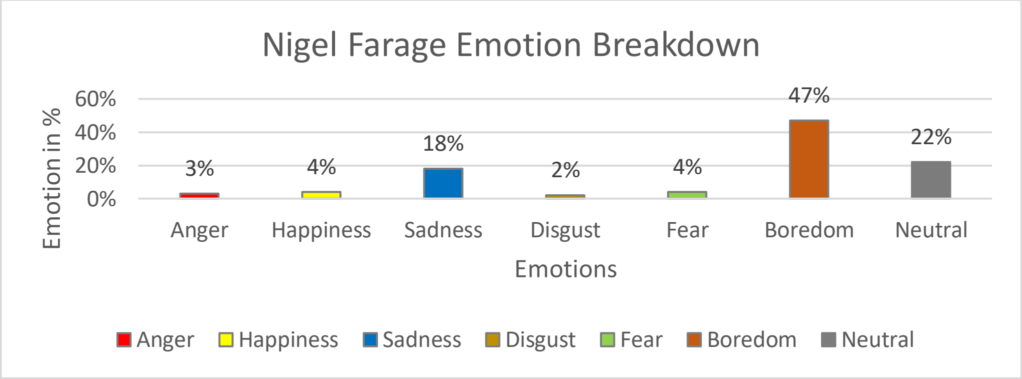


Figure 22: Nigel Farage Emotion Breakdown

Rebecca Long Bailey: Dominant emotional state is boredom (44%).

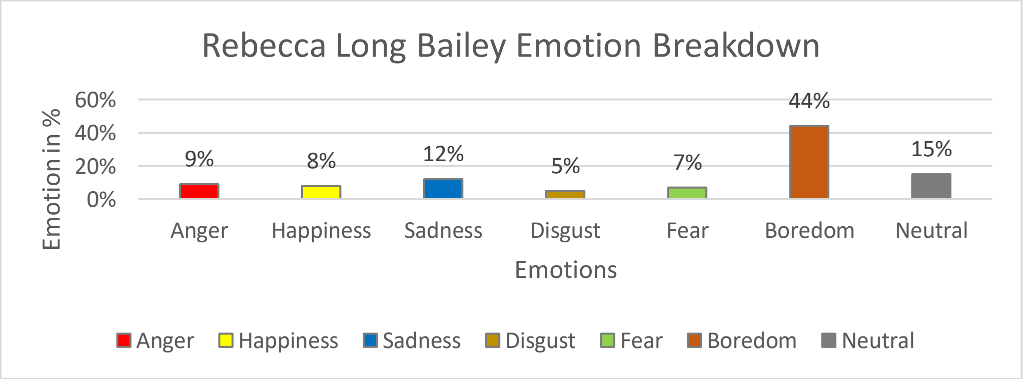


Figure 23: Rebecca Long Bailey Emotion Breakdown

Theresa May: Dominant emotional state is sadness (39%).

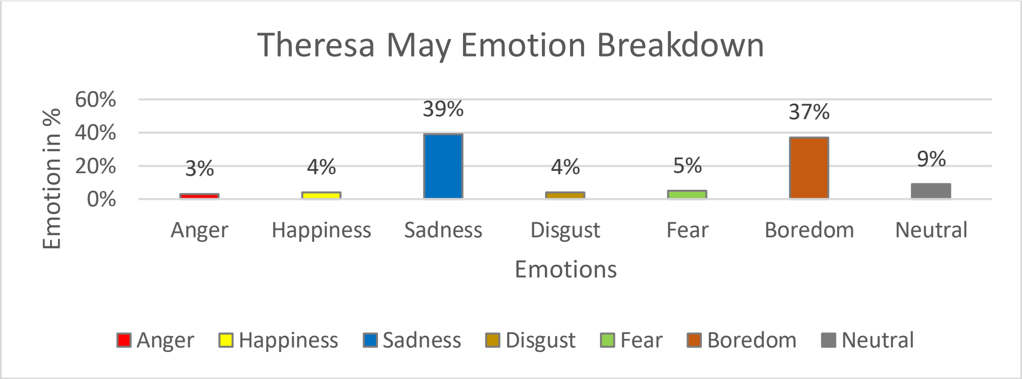


Figure 24: Theresa May Emotion Breakdown

From Figures 16 to 24 it can be said that the values for each emotional state are quite similar when comparing these politicians. Table 8 below shows the weighted mean and range of each emotion calculated across all politicians.

Emotion	Mean	Range
Anger	5.4%	2-9%
Happiness	4.6%	2-8%
Sadness	26.1%	12-50%
Disgust	4.6%	2-7%
Fear	4.8%	3-7%
Boredom	43%	37-48%
Neutral	12%	8-22%

Table 8: Overall emotion statistics

As outlined in section 3-1-1, this database consists of politicians who are highly skilled at public speaking. When speaking to an audience through a speech or answering questions in an interview, it is important to stay calm and composed to avoid causing worry or distress to the listeners. This calm way of speaking is slow and soft which are characteristics of low valence states of emotion (sadness, boredom, etc.). This is evident in the results retrieved from OpenSMILE as sadness and boredom account for a combined average total of 69.5% (see Table 8) across all 9 politicians. In comparison, emotional states that are associated with high arousal (anger, happiness) are seldom seen. It is important to note that from these results, we can say there are no distinct differences in emotions dependent on age, gender or geographical location.

4.2 Data Fusion

In this section I will display all results retrieved from the sentiment analysis section of this project as well as results from the fusion of the two modalities of communication.

4.2.1 Sentiment Analysis Results

As shown in section 3-2-4-2, there are two sentiment scores for each speech/interview for every politician: negative sentiment score and positive sentiment score. The average for each of these values across all politicians can be seen below in Table 9.

Politician	Average Negative Score	Average Positive Score
Arlene Foster	6%	14%
Boris Johnson	6%	15%
Jeremy Corbyn	7%	16%
Leo Varadkar	6%	13%
Mary Lou McDonald	8%	13%
Nicola Sturgeon	7%	13%
Nigel Farage	7%	11%
Rebecca Long Bailey	8%	14%
Theresa May	6%	15%

Table 9: Overall word choice scores

Like the emotional state results in section 4-1, the results for sentiment scores are very consistent across all politicians. The weighted average ratio between negative sentiment scores and positive sentiment scores is **1:2**. For every negative bearing word there is just over two positive bearing words.

4.2.2 Validation

Validation of the emotional states derived from the acoustics of speech by OpenSMILE (results in section 4-1) can be performed by fusion with sentiment analysis results from each speech/interview. This is validation at an aggregate level for each speech/interview. By using the method of data fusion described in section 3-2-2-3, I have obtained the following results for each politician. As highly intense emotional states "anger" and "happiness" and be categorized as negative valence and positive valence respectively, I have chosen to fuse these two emotional states with my sentiment analysis results. These results can be visualised below in Table 10.

	Negative/Anger Correlation	Positive/Happiness Correlation
Mary Lou McDondald	46%	-18%
Jeremy Corbyn	40%	37%
Nicola Sturgeon	-17%	-20%
Arlene Foster	-27%	-26%
Rebecca Long Bailey	-27%	-21%
Theresa May	-30%	-56%
Nigel Farage	-55%	-36%
Boris Jonshon	-57%	-33%
Leo VaradKar	-57%	-4%
Weighted Average:	-19%	-20%

Table 10: Data fusion table

This validation at an aggregate level. As can be seen from the derived results in Table 10, a large portion of these comparisons display an anti-correlation. This however does not necessarily mean that the OpenSMILE or Lexicon-based sentiment analysis are inaccurate, it may imply that there is a large emotional leakage present. For example, we can see from

Table 10 that politician Leo Varadkar has an anti-correlation of 57% between negative bearing word choice and the emotional state "anger". This can be interpreted as "57% of the time that Leo Varadkar is using negative bearing words, he is not in an emotional state of anger".

It is also noticeable that the weighted average of both columns in Table 11 are very similar. This led me to check what the correlation between "anger" and "happiness" is and between negative and positive words. I found there is a very strong positive correlate between "anger" and "happiness" which implies that they these emotions can coexist within speech.

	Happiness/Anger Correlation	Positive/Negative Correlation
Mary Lou McDondald	98%	20%
Jeremy Corbyn	92%	93%
Nicola Sturgeon	95%	81%
Arlene Foster	86%	35%
Rebecca Long Bailey	92%	70%
Theresa May	15%	41%
Nigel Farage	57%	67%
Boris Jonshon	87%	56%
Leo VaradKar	96%	48%
Weighted Average:	81%	54%

Table 11: Happiness vs Anger

5. Afterword

5.1 Conclusion

This project was associated with the physical correlate of human speech, a person's word choice and how emotion/sentiment can be extracted from them and cross-checked against each other. I believe that the dataset gathered for this project was perfect – high quality audio and evident emotions expressed throughout. In my opinion, OpenSMILE was very accurate in showing what percentage of each emotion was being expressed throughout the dataset. This opinion is derived from listening to audio files and analysing the OpenSMILE output. Sentiment analysis is a very interesting topic and an effective method. I believe that this processed well throughout the course of this project, however, it may have been more effective to perform sentiment analysis at a very low level for validation rather than at an aggregate level.

On a personal note, I really enjoyed working on this project. Emotion recognition is a very interesting field and has many applications across different sectors of everyday life. Working with an automatic speech recognition system of such accuracy in OpenSMILE was a brilliant experience and I look forward using it again in the future. Gaining an in depth understanding of sentiment analysis is an aspect of the which I am very happy with also.

5.2 Future Work

This project has many possible ways in which it can be developed in the future. There are some aspects I would like to analyse in further detail.

The first aspect of this project that I would consider investigating in the future is use of an alternative automatic speech recognition system like OpenSMILE. As outlined in my literature review, there are many fantastic automatic speech recognition systems that can extract the features of speech necessary to detect emotion from the human voice, particularly DAVID. DAVID is an exciting open-source software and very well documented. I would like to carry out the same project with the use of this software to compare it to my findings from OpenSMILE.

Another aspect which I would like to investigate is improving the validation procedure. In this project, I carried out validation at an aggregate level for each speech/interview. Performing sentiment analysis at a lower level (each 2000ms segment) may improve the accuracy of validation. Also, in terms of sentiment analysis, using alternative dictionaries (Harvard IV-4 dictionaries) to search for negative and positive bearing words would be interesting to see if/how overall results could change.

Bibliography

Picard, Rosalind W., Elias Vyzas, and Jennifer Healey. (2001). "Toward machine emotional intelligence: Analysis of affective physiological state." *IEEE transactions on pattern analysis and machine intelligence* 23.10 (2001): 1175-1191.

Teixeira, João Paulo, Carla Oliveira, and Carla Lopes. (2013). "Vocal acoustic analysis-jitter, shimmer and HNR parameters."

Lee, Chul Min, and Shrikanth S. Narayanan. (2005). "Toward detecting emotions in spoken dialogs." *IEEE transactions on speech and audio processing* 13.2 (2005): 293-303.

Dietrich, Bryce J., Ryan D. Enos, and Maya Sen. (2019). "Emotional arousal predicts voting on the US supreme court." *Political Analysis* 27.2 (2019): 237-243.

Ahmad, Khurshid, et al. "Media-expressed negative tone and firm-level stock returns." *Journal of Corporate Finance* 37 (2016): 152-172.

Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of finance* 62.3 (2007): 1139-1168.

Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." *Knowledge-Based Systems* 89 (2015): 14-46.

Juang, Biing-Hwang, and Lawrence R. Rabiner. "Automatic speech recognition–a brief history of the technology development." *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1 (2005): 67.

Farrús, Mireia, Javier Hernando, and Pascual Ejarque. "Jitter and shimmer measurements for speaker recognition." *Eighth annual conference of the international speech communication association*. 2007.

Rachman, Laura, et al. "DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech." *Behavior research methods* 50.1 (2018): 323-343.

Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*. Vol. 2. IEEE, 2003.

Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." *Fifteenth annual conference of the international speech communication association*. 2014.

El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* 44.3 (2011): 572-587.

Swain, Monorama, Aurobinda Routray, and Prithviraj Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." *International Journal of Speech Technology* 21.1 (2018): 93-120.

Eyben, Florian, and Björn Schuller. "openSMILE:) The Munich open-source large-scale multimedia feature extractor." *ACM SIGMultimedia Records* 6.4 (2015): 4-13.

Boersma, Paul, and David Weenink. "Praat: Doing phonetics by computer [Computer program]. Version 6.0. 37." *RetrievedFebruary* 3 (2018): 2018.

Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

Kolkur, Seema, Gayatri Dantal, and Reena Mahe. "Study of different levels for sentiment analysis." *International Journal of Current Engineering and Technology* 5.2 (2015): 768-770.

Appel, Orestes, et al. "A hybrid approach to the sentiment analysis problem at the sentence level." *Knowledge-Based Systems* 108 (2016): 110-124.

Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." *Decision Support Systems* 66 (2014): 170-179.

Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Eighth international AAAI conference on weblogs and social media*. 2014.

Sharma, Anuj, and Shubhamoy Dey. "A comparative study of feature selection and machine learning techniques for sentiment analysis." *Proceedings of the 2012 ACM research in applied computation symposium*. 2012.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

Lewis, David D. "Naive (Bayes) at forty: The independence assumption in information retrieval." *European conference on machine learning*. Springer, Berlin, Heidelberg, 1998.

Mullen, Tony, and Nigel Collier. "Sentiment analysis using support vector machines with diverse information sources." *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

Mehra, Nipun, Shashikant Khandelwal, and Priyank Patel. "Sentiment identification using maximum entropy analysis of movie reviews." *St anford Univer sity, USA in* (2002).

Ekman, Paul, and Wallace V. Friesen. "Nonverbal leakage and clues to deception." *Psychiatry* 32.1 (1969): 88-106.

Burkhardt, Felix, et al. "A database of German emotional speech." *Ninth European Conference on Speech Communication and Technology*. 2005.

Dataset Links

Mary Lou McDonald

30.01.2019 - https://www.youtube.com/watch?v=CfKYFNlZSSY
26.02.2019 - https://www.youtube.com/watch?v=vQzHU9KegYc
05.03.2019 - https://www.youtube.com/watch?v=RW0f469siaA
02.04.2019 - https://www.youtube.com/watch?v=W3c1o1fOvOE
14.05.2019 - https://www.youtube.com/watch?v=wZHBI_H8FPs
11.06.2019 - https://www.youtube.com/watch?v=eAh0utqzJUM
09.07.2019 - https://www.youtube.com/watch?v=loZLFwlWHj4
02.08.2019 - https://www.youtube.com/watch?v=q1OHPQtEJ1c
17.09.2019 - https://www.youtube.com/watch?v=S-XCfkiu5QA
01.10.2019 - https://www.youtube.com/watch?v=4tLdmLUvBI4
20.11.2019 - https://www.youtube.com/watch?v=ozlkQc2dwBw
11.12.2019 - https://www.youtube.com/watch?v=xQmv0PWOqcU
11.01.2020 - https://www.youtube.com/watch?v=pZMrZhzvGsA

Leo Varadkar

29.01.2019 - https://www.irishtimes.com/news/politics/oireachtas/varadkar-i-have-to-be-taoiseach-for-the-whole-of-the-country-1.3774804
15.02.2019 - https://www.youtube.com/watch?v=so_GKeam9X8
05.03.2019 - https://www.youtube.com/watch?v=4EfuO9-S0WY
16.04.2019 - https://www.youtube.com/watch?v=fDjnKM99AbU
16.05.2019 - https://www.youtube.com/watch?v=GHVET3YEgdA
11.06.2019 - https://www.youtube.com/watch?v=eAh0utqzJUM
10.07.2019 - https://www.youtube.com/watch?v=00gUGrlPyds
09.09.2019 - https://www.youtube.com/watch?v=IjUpdkdqlkg
13.11.2019 - https://www.youtube.com/watch?v=ZXg1yrfWw60
04.12.2019 - https://www.facebook.com/rtenews/videos/802331270227974/
14.01.2020 - https://www.youtube.com/watch?v=DicCV8IP1Og

Nigel Farage

20.01.2019 - https://www.youtube.com/watch?v=HZxJ5J9Jt1Q
19.02.2019 - https://www.youtube.com/watch?v=uNUDVXFrzvo
26.03.2019 - https://www.youtube.com/watch?v=z1uZl6leA00
08.04.2019 - https://www.youtube.com/watch?v=YGO_C6G35w4
29.05.2019 - https://www.youtube.com/watch?v=P25XWwNoTS4
23.06.2019 - https://www.youtube.com/watch?v=VoM0_Kf9ASo
28.07.2019 - https://www.youtube.com/watch?v=IY5lMVvO18k
29.08.2019 - https://www.youtube.com/watch?v=I-xrngX1vvI
11.09.2019 - https://www.youtube.com/watch?v=S1Iafez9gg0
09.10.2019 - https://www.youtube.com/watch?v=X3gNeLqaJDM
01.11.2019 - https://www.youtube.com/watch?v=vlNLTxSNuhc
29.01.2020 - https://www.youtube.com/watch?v=4xk2-Ol8tLk

Boris Johnson

04.03.2019 - https://www.youtube.com/watch?v=E5lYRdmCVoE
24.07.2019 - https://www.youtube.com/watch?v=8QABH1lMoVU
27.07.2019 - https://www.youtube.com/watch?v=_jbLzMEZNrw
21.08.2019 - https://www.youtube.com/watch?v=BqxVl-KjZi4
02.09.2019 -
https://www.youtube.com/watch?time_continue=168&v=zKlirT5shzQ&feature=emb_title
15.10.2019 - https://www.youtube.com/watch?v=70kfn35nLM4
29.11.2019 - https://www.youtube.com/watch?v=HLLgbSEHC8Q
13.12.2019 - https://www.youtube.com/watch?v=1zmBEZnII0Q
31.12.2019 - https://www.youtube.com/watch?v=fwYTklNu0xQ
31.01.2020 - https://www.youtube.com/watch?v=cvpH51Qzq-A

Theresa May

16.01.2019 - https://www.youtube.com/watch?v=u8gZ8wlBmwo
26.02.2019 - https://www.youtube.com/watch?v=TMa1aBTzrWA
20.03.2019 - https://www.youtube.com/watch?v=0zC6c9gZ-is
21.04.2019 - https://www.youtube.com/watch?v=Fh6WjR6DC8Y
24.05.2019 - https://www.youtube.com/watch?v=_t25xAp270o
26.06.2019 - https://www.youtube.com/watch?v=gx8MOHr_DQI
04.07.2019 - https://www.youtube.com/watch?v=p27fY0hsS-c
24.07.2019 - https://www.youtube.com/watch?v=EIt_qW4ebZQ
10.09.2019 - https://www.thesun.ie/news/4509786/where-is-theresa-may-now-and-is-she-still-a-conservative-mp/
12.12.2019 - https://www.youtube.com/watch?v=1Om0bDwGaKU

Arlene Foster

15.01.2019 - https://www.youtube.com/watch?v=uz-sWPefcoM
14.03.2019 - https://www.youtube.com/watch?v=ao0cbwjU6ns
29.03.2019 - https://www.youtube.com/watch?v=-s2duLPnFFk
26.04.2019 - https://www.youtube.com/watch?v=N1-6XbW3CQ4
28.05.2019 - https://www.youtube.com/watch?v=_syB6j7d3KI
12.09.2019 - https://www.youtube.com/watch?v=nCZeeT_S-OE
03.10.2019 - https://www.youtube.com/watch?v=y35uQFg-j-c
26.10.2019 - https://www.youtube.com/watch?v=p_n5VYw3I3A
28.11.2019 - https://www.youtube.com/watch?v=77fNQacVKDg
09.12.2019 - https://www.youtube.com/watch?v=vG9L84ct4sk ,
11.01.2020 - https://www.youtube.com/watch?v=TFY14ySiOGs

Nicola Sturgeon

24.01.2019 - https://www.youtube.com/watch?v=mAI2uUCjJdw
21.02.2019 - https://www.youtube.com/watch?v=InUt8rD8Xuk
28.03.2019 - https://www.youtube.com/watch?v=jLH1x3ZaLQA
28.04.2019 - https://www.youtube.com/watch?v=PEsj2jO9lmM
22.05.2019 - https://www.youtube.com/watch?v=cBtnSnyxGyM
28.06.2019 - https://www.youtube.com/watch?v=rToL6b9wScc
04.08.2019 - https://www.youtube.com/watch?v=ioSn6J3m_k4
06.09.2019 - https://www.youtube.com/watch?v=gJzSWacrkKo
16.10.2019 - https://www.youtube.com/watch?v=1CEvljetRdQ
03.11.2019 - https://www.youtube.com/watch?v=dGwO5XWVGOo
13.12.2019 - https://www.youtube.com/watch?v=KCKiUy5A6LM

Jeremy Corbyn

01.21.2019 - https://www.youtube.com/watch?v=F2N17v3250M
08.03.2019 - https://www.youtube.com/watch?v=tGyWlo1EPHg
20.04.2019 - https://www.youtube.com/watch?v=2muT_Vo1HPI
15.07.2019 - https://www.youtube.com/watch?v=w8qfBCoThrE
10.09.2019 - https://www.youtube.com/watch?v=EF-yiDqJePI
31.10.2019 - https://www.youtube.com/watch?v=062iHu1pwpU
30.11.2019 - https://www.youtube.com/watch?v=Em32dFJIe5o
06.12.2019 - https://www.youtube.com/watch?v=LlkYPDbt5mk
24.12.2019 - https://www.youtube.com/watch?v=7o8Qu4PC_hw

Rebecca Long-Bailey

04.02.2019 - https://www.youtube.com/watch?v=qLBDDTPsYfY
28.04.2019 - https://www.youtube.com/watch?v=ixfc-ZrVEI4
02.05.2019 - https://www.youtube.com/watch?v=qneMhrsCMvU
24.08.2019 - https://www.youtube.com/watch?v=yIEH2iR20PM
24.09.2019 - https://www.youtube.com/watch?v=FAgL6PLWeJs
07.11.2019 - https://www.youtube.com/watch?v=yno_MkS9pFE
12.02.2020 - https://www.youtube.com/watch?v=Gp1TKeSLE50