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

Multi-modal Sentiment Analysis with Data Fusion

by Basit Hamid Sofi

Supervisor: Dr. Khurshid Ahmad

M.Sc Thesis Report
in completion of the M.Sc Taught programme
School of Computer Science & Statistics
Trinity College Dublin, Ireland

Declaration

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

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Basit Hamid Sofi

September 11, 2020

Acknowledgments

I would like to express gratitude towards my thesis supervisor, Professor Khurshid Ahmad for his constant guidance and advice throughout the course of the dissertation, and his Research Assistant, Pranav Jain, who was extremely helpful in updating me with the systems and their configurations. I would like to thank Trinity College Dublin for giving me the opportunity to learn and apply my learning throughout the duration of the course.

I would also like to take this opportunity to thank my parents and friends who have been extremely supportive and encouraging during my time at Trinity College Dublin.

Abstract

The aim of this project is to examine emotions in three different modalities of communication and validate the presence of similar emotions in each of those modalities at any given time, and detect potential leakage of emotions. The three modalities in question include two non-verbal modalities, i.e. Facial Expressions and Voice, and one verbal modality, i.e. Speech/Text.

The extraction of emotional information from facial expressions was performed using an intelligent system, called Emotient FACET. This system makes use of Computer Vision and Machine Learning to generate emotional information based on the movement of facial muscles. Similarly, for extracting emotional information from voice, a system called OpenSMILE was used which makes use of Machine Learning and Speech Analytics to analyze speech signals. A Bag-of-Words approach with pre-categorized lexical databases has been used for analyzing the word choices of the speakers.

The dataset consists of 27 audio and videotapes of 4 politicians during public appearances such as speeches, press conferences, and interviews.

The main contribution of this project was to perform Data Fusion using the three modalities and analyzing the data to draw conclusions regarding the similarity of emotions in the three modalities. Data Fusion is the main problem and refers to the process of combining the data from the three sources that have different rates of transmission and are thereby perceived at different rates by humans.

A separate pipeline of programs was created for processing the raw output of each of the three modes of communication before fusing the processed outputs. This included carefully selecting the data, writing custom programs for processing the raw output obtained by Emotient and OpenSMILE and performing sentiment analysis on verbal data before fusing the outputs into a single file.

We can also detect potential leakage of emotions by identifying the presence, or lack thereof, of similar emotions at a given time frame from the fused data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Human beings' behavior in any interpersonal situation can be categorized into two types. These are non-verbal and verbal. These can be distinguished based on the medium in which humans choose to express their emotions [1].

Non-verbal communication can be defined as the use of facial expressions, tone of voice, gestures, body language, posture and other ways to communicate without the use of spoken word or language. These are normally perceived using visual sense organs, sometimes supplemented by auditory senses in case of vocal communication. The non-verbal communication perceived by the auditory senses can convey emotions using the three basic characteristics of sound i.e. pitch, intensity and timbre. This is also called vocal communication.

Verbal communication pertains to speech, choice of words and the content of the statements of vocal communication. Both vocal and verbal behaviors are related as they originate in the pharynx which is the part of the throat behind the mouth and nasal cavity, and above the esophagus and larynx, and as both of these behaviors are perceived by the auditory senses.

Humans can express a wide range of emotions. A number of psychologists have postulated that humans can express a limited number of basic emotions that are considered to be universal across cultures and human ethncities [2] [3]. Izard suggested that emotions have evolved due to their value in adapting and performing fundamental life tasks [4]. The basic emotions are used to aide in performing the basic or fundamental tasks e.g., fight or flight responses can be triggered by understanding of the basic emotions of fear and anger. Although the notion of Basic Emotion Theory has been widely discussed and accepted, there is no consensus regarding the actual number of basic emotions. Ekman proposed that there are six basic emotions. These are Joy, Sadness, Fear, Disgust, Surprise and Anger. Sauter et al. [5] investigated the identification of nonverbal and emotional vocalizations,

across two significantly different cultures. The vocalizations of “basic” emotions were recognized to be similar in case of both the cultural groups. These emotions can be conveyed by humans through speech, facial expressions, gestures, movement of head, voice modulation, or choice of words.

Facial expressions are generated by the contractions or relaxation of muscles in the face, which leads to temporary movement or deformity of the facial features such as eye brows, eye lids, lips, wrinkling of the skin near the eyes, etc. Typically, these changes in the movement of the facial muscles are brief, lasting not more than 5 seconds, or less than 250 ms [6]. There are different kinds of expressions, such as spontaneous facial expressions which are expressed without any prior thoughts and tend to betray emotions. They are difficult to measure. The other kind of expressions are Posed Facial Expressions which are measured after being acted out by actors as part of a script in controlled conditions. They are easier to measure because everything in the script is defined including start and finish times along with intensity. The Facial Action Coding System (FACS) developed by Ekman and Friesen [7] is used for measuring facial movement. Each independent movement of the facial muscles is defined in terms of Action Units, or AUs. Ekman and Friesen defined 46 AUs. A linear combination of the movement of one or more of these AUs is able to detect a wide range of emotions.

Analyzing sentiments from voice is different from using facial expressions as it depends on the modulation of the voice and not on the movement of the facial muscles. The modulation of voice can be done by increasing or decreasing the pitch or frequency of the voice, or increasing or decreasing the loudness or amplitude.

Another form of non-verbal communication is gestures. It is defined as the movement of part of the body, especially the hands and head, to convey a meaning or an idea. Gestures form part of day-to-day lives. There are about 700,000 different gestures, which is 150,000 more than the number of words in the largest dictionary [8]. Gestures have no vocabulary or grammar and it enriches the communication between people.

Humans with emotional intelligence can easily convey and understand emotions due to our ability to process speech and non-verbal cues simultaneously. This can sometimes mean using multiple modalities at the same time such as facial expressions, gestures, voice modulation, etc. while listening to another human being to understand emotions that the speaker is trying to convey. It is claimed that there may be some leakage of emotions in case of strained situations that the speaker may be trying to hide but can be detected by the listener(s) due to their emotional intelligence.

Leakage of emotion in unimodal communication can be detected if two or more conflicting emotions occur at the same time such as Joy and Anger, or if there is a sudden change in the emotions for a very short duration such as change from Joy to Anger for a fraction of a second and then a switch back to Joy. Emotional Leakage in multimodal communication may be detected if there are inconsistencies

in the emotion displayed by the different modes at any given time whereas the occurrence of the same emotion across different modalities reinforces the presence of an emotion.

A major technical challenge is the computation of emotional intelligence, which can also be referred to as Artificial Emotional Intelligence. There are a number of techniques that use machine learning, computer vision and speech analysis, together with inexpensive hardware that can allow us to develop computer programs and/or machines that possess emotional intelligence. These systems use an extensive dataset of emotions recorded by actors who act out a number of emotions in an exaggerated manner.

This study will focus on multi-modal sentiment analysis using facial expressions, speech/voice and text.

1.2 Methodology Outline

The aim of this study is to:

I. Create a dataset of videos, audio and transcripts of audio/speech, representing an application domain, and curate it in a way that the focuses on the subject.

II. Extract and analyze facial expressions from Videos by passing it to the Facial Expression Recognition software (Emotient).

III. Post process the raw output of Emotient.

IV. Extract and analyze audio features and expressions by passing audio to the Speech Recognition software (OpenSMILE).

V. Extract transcript from audio and add timestamps to the transcript using a software called Descript.

VI. Analyze the emotions in text of speech using known affect dictionaries such as General Inquirer, dictionary of Affect words using Rojet's thesaurus and a dictionary of political Affect words and phrases.

VII. Perform Data Fusion for the three modalities by performing aggregation on Facial Expression data and text data, and combine the outputs from all three sources.

VIII. Analyze expressions in all three modalities to validate whether similar emotions exist in all of them.

IX. Perform the analysis in step VI across the dataset.

X. Detect leakage of emotion across modalities.

The first phase of the study involved data collection and preprocessing. This includes downloading videotapes of famous politicians and public figures such as Donald Trump and Boris Johnson from publicly available sources over the Internet such as YouTube and BBC, preprocessing the videos using an Audio and Video Editing Software called Adobe Premiere Pro to make sure that only the respective candidate's face appears in the video, extracting the audio from the video and storing it on the system in a format that is suitable for the Speech Recognition software (openSMILE).

The next phase of the study involved extracting emotions from facial expressions. For this purpose, we used the software Emotient by iMotions. This software gives the raw output of facial expressions in the form of Action Units. The presence of an emotion is calculated using linear combination of the Action Units which is based on the Facial Action Coding System [7].

The third phase of the study involved post processing the raw output obtained from facial expression analysis by Emotient. Here, frames with low evidence of an emotion were ignored and frames with acceptable evidence value (i.e., value above a certain threshold) were accepted and used in further analysis.

The fourth phase of the study involved extracting emotions from speech using Speech Recognition software called OpenSMILE. This software extracts the physical features of the voice from the speech to determine the emotions being expressed.

The fifth phase of the study dealt with transcribing the audio from the speech signals. This is performed using a tool called Descript that automatically transcribes the speeches and then adds timestamps to identify the times at which different words were spoken.

The sixth phase of the study involved analyzing the transcribed text. This was performed writing a Python program that imported a number of well known and accepted dictionaries of emotional words. The purpose of writing the script was to identify the emotions associated with each word of the speech and find the distribution of emotions in speech, i.e., number of positive and words, number of closed class words, and words associated with the basic emotions. The process will be elaborated in the methodology section.

The seventh phase of the study involved Data Fusion. The process of Data Fusion is important for our study because the signals for each modality arrive at different frequencies. To achieve data fusion, aggregation of facial expression data and transcribed data was performed to bring it to the same frequency as that of the audio frequency i.e, 2000 milliseconds.

After performing data fusion, the next phase of the study focused on main analysis of the fused data. The data was used to find correlation between the different emotions in each of the modalities.

1.3 Summary of Dataset Used

The dataset chosen for this study consists of audio and videotapes of 4 well known public figures and famous politicians. There are total of 27 taped appearances of these figures. The audio and videotapes are taken from a wide variety of political speeches, interviews, press conferences and University convocation addresses, mostly recorded by large media outlets and as such the quality of videos is high with good lighting and resolution, while the focus remains on the speaker.

The candidates chosen for the study are highly influential politicians and/or public figures who are trained public speakers and have good public speaking ability. They can hide their actual emotions in order to convey the message in the context of the speech or event. This is called “semi-wild” data as these people are intending to express a particular emotion, while not being present in a recording environment. The data is used in the three modalities of facial expression recognition, automatic speech recognition and text analysis.

1.4 Structure of the Report

The first section of the report talks about the problem statement, a summary of the methodology applied and a summary of the data used.

The second section of the report talks about the motivations for this project, some of the applications of sentiment analysis, followed by the existing literature available on sentiment analysis through facial expressions and speech acoustics.

The third section of the report focuses on the methodology and design of the project. It explores the technologies behind Emotient, OpenSMILE and the ways to perform Sentiment Analysis on text with the help of existing affect dictionaries.

The fourth section of the report discusses some case studies and the results obtained.

The fifth and sixth sections of the report discusses the conclusions and the scope of future work in the field of multi-modal sentiment analysis.

Chapter 2

Motivation and Literature Review

This section covers the existing literature on the analysis of non-verbal communication such as facial expressions and speech acoustics as the physical correlates of emotion and verbal communication such as content of speech.

2.1 Motivation and Applications

The aim of the study is to observe the emotional behavior of public figures, with a keen focus on political leaders, and observe the distribution of emotions in their speeches, while also identifying emotional patterns expressed by them in public. From existing studies of the effect of emotions displayed by leaders in various situations, it is clear that positive emotions displayed by politicians are more desirable than negative emotions. For example, there is strong correlation between positive emotion and a leader's charisma [9] [10]. It has been observed that in case of an emergency or disaster, a fear inducing and negative appeal tends to reduce the overall persuasiveness of communication [10].

The human brain has a number of functions that enable us to recognize and grasp objects. Another significant brain function is our ability to be skillful in social interactions [11], such as recognizing nonverbal cues in a conversation.

Human beings possess knowledge which can be defined as facts, information and skills that have been acquired over time from education or experience. We possess cognitive abilities such as language, control, creativity, decision making, problem solving, reasoning, and perceptive capabilities such as sense of sight, smell, touch, hearing and taste which help us translate different environmental inputs such as visual, acoustic, chemical, etc. into bio-chemical sensory inputs that can be processed by the body's nervous system. With the help of these abilities, human beings are capable of making decisions, performing actions, conveying thoughts, ideas and feelings.

Detection of emotions using a machine is a problem that belongs to the field of Artificial Intelligence (AI) because a general solution to such a problem is not known in computing and it is a problem that takes advantage of certain level of human intelligence. The motivation for this study is to create an Artificially Emotionally Intelligent expert system that uses a set of programs which applies the knowledge of experts in field of sentiment analysis, so that data from such a system can be used to compute and analyze emotions and make inferences using the data obtained. The data is used to identify patterns in the way a person expresses emotions using the different modalities and the distribution of emotions displayed by different personalities.

There are numerous applications of emotion recognition and analysis in various domains such as healthcare, psychology and technology. The applications of facial expression analysis include identifying differences between true and simulated pain [12], identifying when someone is lying versus when someone is telling the truth [13], and identifying differences between the facial expressions of non-suicidal and suicidally depressed patients [14]. One of the applications of the analysis of acoustics in speech and text is identification of mild and severe mental disorders, while also being helpful in determining psychotherapy treatments [15].

2.2 Emotion Recognition

2.2.1 Overview

Emotion recognition can be defined as the process of identifying human emotion in the form of verbal and non-verbal communication. In case of non-verbal communication, emotions can be expressed through facial expressions and voice and in case of verbal communication, emotions can be expressed through speech. In the following section, we will discuss Automatic Facial Expression Analysis, Speech Analysis and Text Sentiment Analysis.

2.2.2 Automatic Facial Expression Recognition Background

Research in the field of facial emotion recognition has gained huge momentum over the last two decades. Automatic Facial Expression Recognition uses the techniques of Computer Vision and Machine Learning with the help of cheap computational power to detect emotions from facial expressions [16]. In the following subsection, we will discuss Facial Action Coding System (FACS) which forms the basis of facial expression recognition and the development of Automatic Facial Expression Analysis.

2.2.2.1 Facial Action Coding System

The Facial Action Coding System (FACS) developed by Ekman and Friesen in 1978 is the leading objective and systematic method for quantifying and measuring facial movement in terms of component actions in a sequence of images. FACS does this by identifying changes in facial expressions as a result of contractions and relaxations of the facial muscles. These muscular changes are quantified in terms of measurement units called Action Units (AUs). Each independent motion of the face corresponds to an AU. Ekman and Friesen defined 46 AUs and each AU is a combination of one or more facial muscles [7]. A linear combination of AUs and the score assigned to each AU helps in categorizing the facial expressions into one of the basic emotions. For example, AU6 corresponds to the raising of the cheek which contributes to the emotion of happiness; AU12 corresponds to the pulling of the corner of the lip, which also contributes to the emotion of happiness and contempt when the action appears unilaterally. The presence of different AUs gives a score to the different facial expressions and in the case of both AU6 and AU12 being present, the system categorizes the person as being happy at a given time.

FACS coding is performed by video and it provides the exact specifications of the facial movement dynamics such as duration, onset and offset times along with the AUs that are morphed. Human observers perform FACS using stop-motion video. FACS defines facial actions according to the changes in image in a sequence of facial images present in a video.

Although FACS provides an encouraging approach in the analysis of facial expressions, one of the major limitations of the system that hinders its widespread application is the amount of time it takes to train human experts as well as to score the videotapes. It takes over 100 hours to train a human in FACS and each minute of video takes about an hour to score to achieve minimal competency. Automating FACS makes the system more reliable and widely accessible.

2.2.2.2 Dataset for Facial Action Coding System

With the increasing growth in research in the field of Automatic Facial Expression Analysis, a lot of researchers felt that the existing datasets at the time were quite limited because of the unknown generalizability of these datasets. In 2000, Kanade et al. [17] presented the CMU-Pittsburgh AU-Coded Face Expression Image Dataset, which consisted of 2105 sequences of digitized images from 182 adult subjects of different ethnicities. They described that the problem space for facial expression analysis consisted of level of description, validity and reliability of training and testing data, orientation of the head of the subject, differences in subjects, characteristics of images, and transition of expressions. This was the most comprehensive dataset for facial expression analysis at the time and was used for classification of Action Units and emotions. However, some limitations of this dataset

were identified such as: 1) Action Units were classified reliably while emotions were not necessarily correctly classified due to the fact that emotions were not validated like the Action Units; 2) There was no availability of a standardized performance metric to evaluate the performance of algorithms; and 3) There were no standard protocols for common databases. To address these limitations, the Extended Cohn-Kanade (CK+) database was released [18]. This extended dataset increased the number of sequences by 22% and the number of subjects increased by 27%. Each sequence was fully coded with FACS and emotion labels were revised and validated. In addition, a number of sequences for spontaneous or non-posed expression such as different types of smiles and the metadata associated with them were also added to the database. The Computer Expression Recognition Toolbox (CERT), which is a tool for real-time automatic facial expression recognition achieved an average performance of 90.1% accuracy on a database of posed facial expression (CK+), and accuracy of about 80% on a database of spontaneous expressions [19].

2.2.2.3 Automated Facial Expression Analysis

Computer vision together with Machine learning techniques has the potential to observe and calculate the changes in the measurements of the facial muscles [16]. Automated Facial Expression Analysis is quite complex as it consists of a number of intricate steps such as face acquisition, facial feature extraction and finally facial expression classification. This can be seen in Figure 2.1 below which represents the multiple steps of the for generic facial expression analysis framework [6].

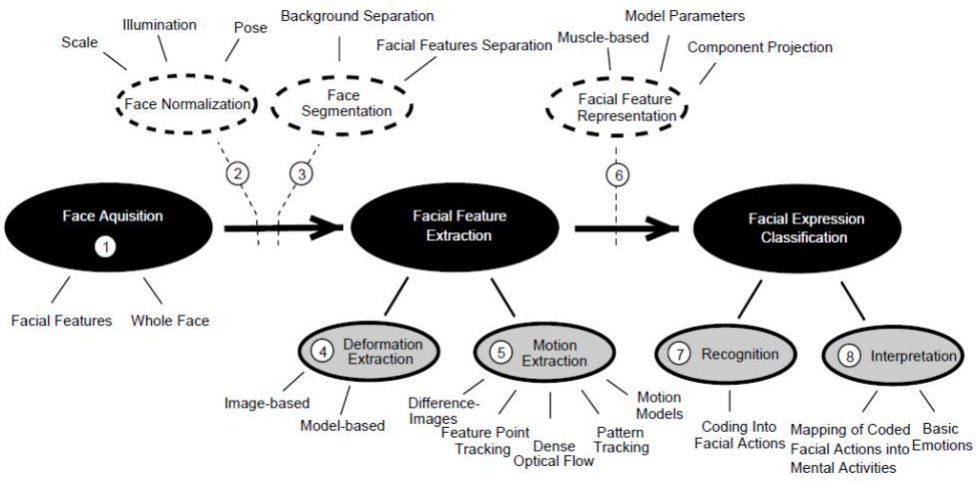

Figure 2.1: Generic Facial Expression Analysis Framework

A patent was filed based on the work of Paul Ekman in the development of the Facial Action Coding System (FACS) and the pipeline for generic facial expression analysis framework [20]. The patent is the basis for Automated Facial Expression Analysis and provides a pipeline starting with

identifying a face in an image, aligning one or more facial features, defining windows in the image, feature selection, and classifying one or more facial actions present in the image. The pipeline of the system is described in Figure 2.2 below.

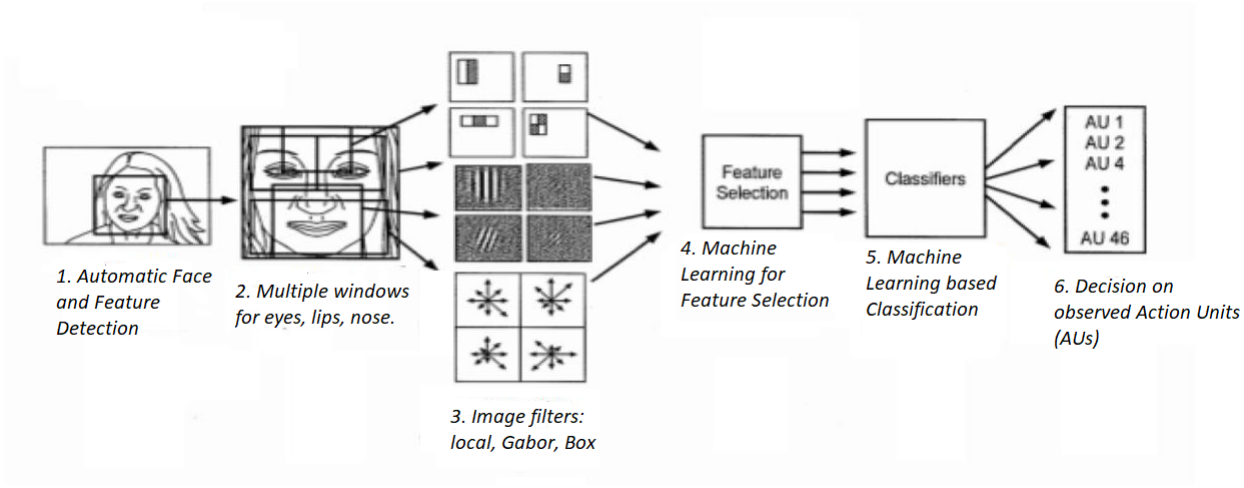

Figure 2.2: Automated Facial Expression Classification

2.2.3 Speech Recognition Background

Humans primarily communicate with each other using spoken words or speech. Analysis of emotion through speech refers to the use of methods for analyzing the changing vocal characteristics. Speech recognition dates back to the early 1950s and there have been a number of key events in the development of this technology. In 1952, three researchers at Bell Laboratories created a speech recognition system called “Audrey” which could recognize digits spoken by a single speaker [21]. Ten years later, IBM introduced “Shoebox”, a system that could understand 16 English words. Shortly after that, in 1969, the funding at Bell Laboratories dried up for a few years after John Pierce, who was an influential Engineer and author, wrote a letter critical of speech recognition and defunded the research in that area. In 1970, Itakura and Saito proposed a speech coding method called Linear predictive coding (LPC) was proposed [22]. The next decade saw significant advancements in the field of speech recognition, mainly due to the US Department of Defense and the funding provided by DARPA. The Speech Understanding Research (SUR) that they funded helped in the creation of Carnegie Mellon’s “Harpy” speech system that was able to understand 1,011 words which is almost the same as the vocabulary of a 3-year-old child [23]. Around the same time, IBM and ATT Bell Laboratories were looking at two different areas of application for speech recognition. By the mid-80s, IBM had created a voice-activated typewriter called Tangora, which had the ability to recognize 20,000 words. The system made use of statistical techniques such as the Hidden Markov Model (HMM) for speech recognition over the approach of trying to mimic the way humans process and recognize

speech [24]. The application of HMM was pioneering in the field of speech recognition and laid the foundation of systems which could recognize thousands of words for the next two decades. ATT Bell Laboratories' goal was to create a speaker-independent system that would provide automatic telecommunication services to the public such as voice dialing, routing of phone calls, etc. By early 2000s, the speech recognition systems had achieved around 80% accuracy. Later in the decade, Google arrived with Google Voice Search which could recognize over 230 billion words from user searches. In the last decade, we have seen significantly smarter systems making use of accurate speech recognition systems, such as Apple's Siri, Amazon's Alex and Google Home.

2.2.3.1 Speech Properties

The means of communication that involves use of the vocal cords to produce vibrations of varying amplitude and frequency is called speech. Emotions in voice can be explored at three levels i.e., (1) the physiological level that deals with the nerves and the muscle patterns of the different parts present in the process of voice-production; (2) the phonatory-articulatory level such as the movement of the vocal folds; and (3) the acoustic level such as the waves that characterize the wave form of the speech that comes out of the mouth [25].

Extracting emotions at the physiological and phonatory-articulatory can be quite difficult because it requires special equipment and high expertise. In contrast, the acoustic cues of voice can easily be extracted from the recordings of the conversations, speeches, interview, etc.

There are characteristics or parameters in the vocal cues or acoustic signals that can be extracted to analyze sentiment. These characteristics are: (1) Fundamental frequency (f0) which is physical correlate of the perceived 'pitch'; (2) Quality of voice which is correlate of the perceived 'timbre'; (3) Intensity (a physical correlate of the perceived loudness); (4) Voice perturbation which is the short-term variation in the sound; and (5) Temporal aspects of speech such as rate of speech. Shimmer and jitter are important parameters of vocal acoustics and are related to the temporal aspect of speech. The most common parameters that are used in the analysis of acoustic signals include fundamental frequency, shimmer and jitter.

For voice, the Fundamental frequency (f0) is defined as the number of times the vocal cords produces a sound wave. It is also defined as the number of times the glottis open and closes. The glottis is the part of the larynx consisting of the vocal cords and the opening between them. The fundamental frequency is measured in terms of Hertz (Hz). It depends on a number of factors such as gender, age and lifestyle [26]. The fundamental frequency lies between 85Hz – 180Hz in males, 165Hz – 255Hz in females, and 260Hz - 280Hz in children.

Shimmer and jitter correspond to the disturbance in the fundamental frequency. Jitter refers to the

variation of frequency between cycles. It is the average absolute difference between consecutive time periods. Shimmer is defined as the variation in the amplitude of the sound wave. It is the average absolute difference of amplitudes between two consecutive time periods [27].

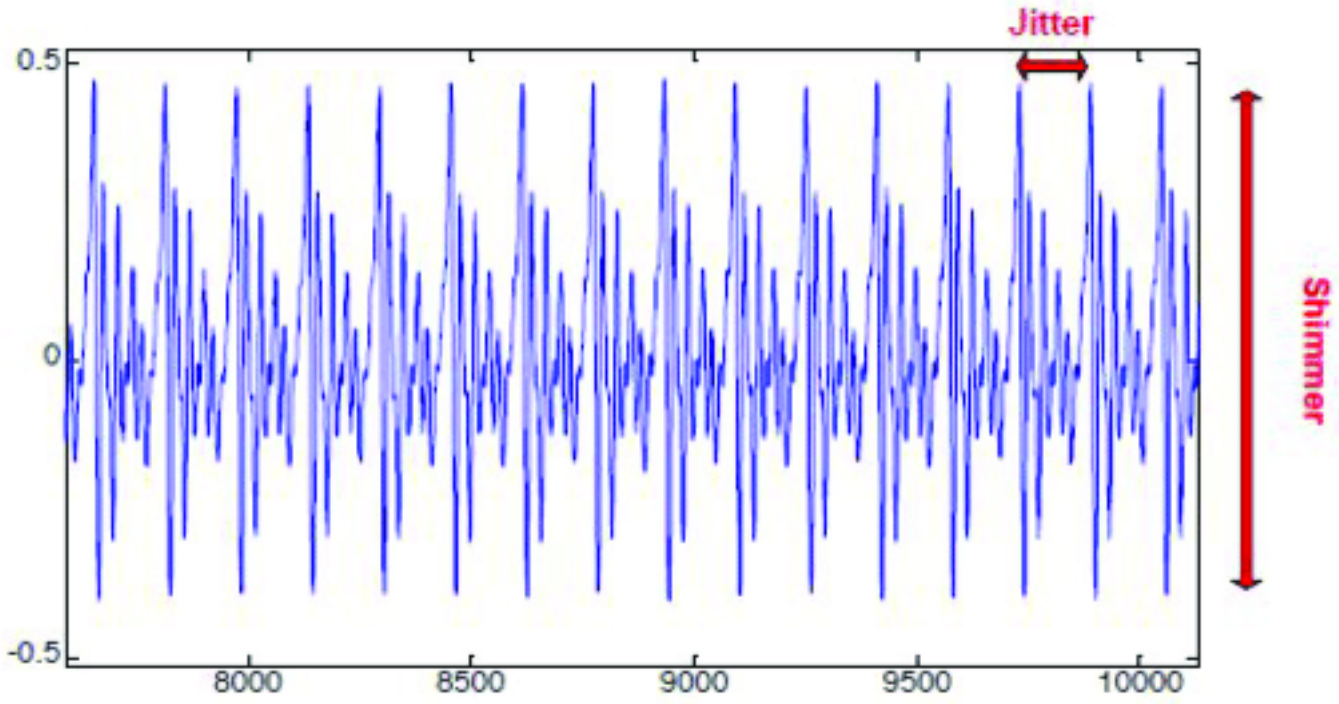

Figure 2.3: Jitter and Shimmer in Speech

Pichora-Fuller et al. [28] discussed the significance of F0 for predicting vocal emotion categorization. A wide variety of stimuli are used by researchers while examining the vocal cues that are important for perceiving emotions in voice. The study analyzed various stimuli for determining which acoustic cues contributed the most in categorizing the basic emotions. Mean F0 was found to be the most important acoustic feature contributing to categorizing the stimuli into various emotions. Li et al. [29] evaluated the jitter and shimmer features of speech for classification and analysis of speaking styles and arousal levels of humans and animals. The addition of shimmer and jitter resulted in increased classification accuracy when applied to an HMM-based classification model. Rachman et al. [30] discussed the effects that varying the values of f0, shimmer and jitter has on the emotion. Increasing the pitch of the fundamental frequency f0 by a constant factor was correlated with states of high arousal such as happiness and decreased value of f0 was correlated with states of low valence such as sadness [31] [32] [33]. Jitter, or Vibrato, is correlated with high arousal and is considered to be a significant marker of emotion. Change in the value of Shimmer (or Inflection) at the beginning of each utterance also has an effect on the emotion. Increasing the inflection corresponds to emotions of high intensity and positive valence whereas decreasing inflection usually corresponds to emotions

of low valence [34].

These features of f0, shimmer and jitter are referred to as the physical correlates of emotion and using these features, it is possible to identify emotions of a person.

2.2.3.2 Automatic Speech Recognition Process

Emotions from speech signals can be extracted as given in Figure 2.4:

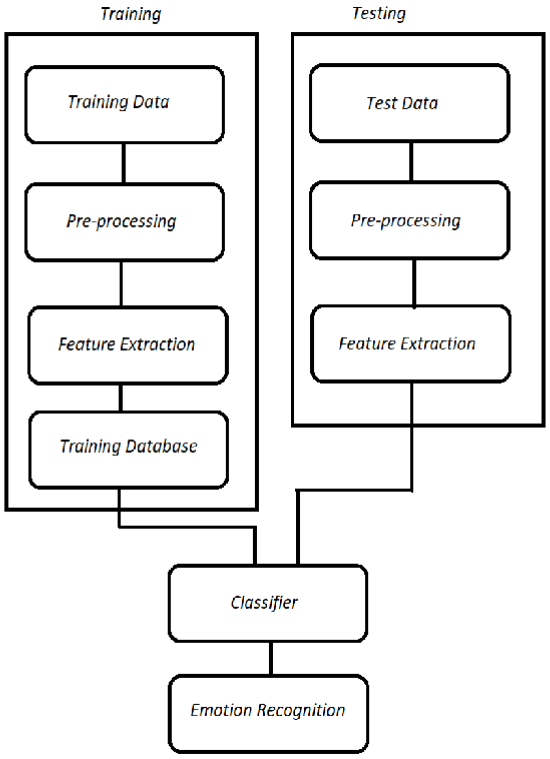

Figure 2.4: Automatic Speech Recognition

The steps for automatic speech recognition are as follows:

1. ***Training the System:*** The first stage is training the system to recognize the various features of the acoustic signal to specific emotions. In most systems, trained actors are employed to act out a script with pre-determined emotions. However, in some other systems semi-trained or semi-professional actors are invited to perform the script to avoid exaggeration of emotions. It is considered more realistic to use data from real-life events as they tend to convey emotions naturally. Using acted emotions over real ones has been criticized at times as acted emotions can be exaggerated [35]. Nonetheless, there is no contradiction between relationship of acoustic features and real emotions and the relationship of acoustic correlates and acted ones. This stage of recording actors reading a script with fixed emotions is done for several emotions and

the features from the acoustic signals are extracted. These extracted features are used to differentiate between the various emotions. The acoustic information along with the emotion's label is stored in a database [36]. A given speech input can be classified into an emotion by extracting the features from the acoustic signals and comparing the extracted features to the database.

2. ***Extracting features and Classification:*** This is one of the most important steps in the process of emotion recognition in speech. Some issues need to be addressed while selecting these features. One such issue is deciding the region of analysis for feature extraction. Some researchers prefer dividing the speech into regular intervals called frames which is used to extract local feature vectors for the analysis, while many other prefer extracting the global statistic for analysis over the entire duration of speech utterance. Most researchers believe that global features are better in terms of classification accuracy and classification time. Global features are also much less in number as compared to local features and therefore, applying cross-validation and feature selection algorithms is quicker on global features. However, global features are only efficient in differentiating between emotions of high-arousal like anger and joy against low-arousal emotions e.g., sadness. One more disadvantage of using global features is that temporal information is completely lost. Therefore, it might not be prudent to use such features on classifiers such as Hidden Markov Model (HMM) and Support Vector Machine (SVM) [36]. On the contrary, the large number of local features can be applied on these classifiers reliably and the parameters can be estimated much more accurately. Another important aspect is deciding which features are best for the task e.g., pitch, energy, zero crossing, etc. For this study, the focus is on the local features of fundamental frequency (f0), the energy values (jitter and shimmer) and the descriptive statistics (mean, median, standard deviation, etc.). Variation in these features determines which emotion is being expressed. Not all the features that are present in the speech signal need to be used. There is a feature selection stage that selects the features that are of significance and removes unnecessary features to improve the performance of the speech emotion recognition system. On identifying the features, the data is passed on to a classifier such as the Hidden Markov Model (HMM), deep neural networks and Gaussian Mixture Model to classify the speech intervals into various emotions.

There are a number of tools and softwares for speech emotion recognition available that follow this architecture. Some of them include OpenSMILE which is an open-source software that extracts features from speech intervals and classifies them into the specific emotions based on the probability values of emotions [37]. OpenSMILE is the system used for this study. There are other systems like DAVID [30] and Praat which work in similar ways.

2.2.4 Text Sentiment Analysis

Sentiment Analysis is the process of applying text analysis, natural language processing and computational linguistics to interpret and categorize words in a text into various categories of emotions (such as positive, negative, neutral, etc.), and study subjective information in text. Sentiment Analysis has many applications in a wide range of domains such as using customer reviews and responses to surveys, social media posts, etc. in domains ranging from customer service to healthcare and financial markets. Neri and Aliprandi [38] performed a study comprising of over 1000 Facebook posts that highlights the significance of Social Media websites such as Twitter and Facebook as a platform for online marketing. Ahmad et al. [39] built a corpus of over 5 ½ million new articles to textually analyze media content and its relationship with market outcomes. Most researchers have concluded that negative content expressed by the media lead to notably negative returns on the next day [40]. Devitt and Ahmad [41] studied the market data with a human gold standard for sentiment judgment whose preliminary results suggested that there is a correlation between stock prices with positive and negative text in financial news. Another application of Sentiment Analysis was used to classify millions of tweets from Twitter during the 2016 Presidential Elections into positive, negative and neutral emotions.

In this study Sentiment Analysis techniques have been used to categorize words into the various emotions. Identification of the words and their emotion value can be used to examine whether all modalities display similar emotions at the same time and also for detecting leakage of emotion.

2.2.4.1 Text Analysis

Sentiment Analysis can be performed using lexicon based approaches and machine learning techniques. Lexicon based approach uses a bag-of-words approach with pre-categorized lexical database and character matching to classify words into sentiments [42]. This approach involves extracting the sentiment from the semantic orientation of words and assigning them positive or negative values by using well known existing affect dictionaries, or by creating domain specific Affect dictionaries.

Different machine learning techniques can be used for analyzing sentiments. Lexicon based approaches can be labor intensive due to construction and collection of a corpus of related dictionaries. Rathi et al. [43] proposed a solution of that uses ensemble machine learning techniques by merging a Support Vector Machine with Decision Trees. The proposed approach provided better classification results for analysis of tweets. Wu et al. [44] integrated support vector machines with popular sentiment analysis techniques to predict the correlation between stock forums sentiment and stock price trends. The results showed that statistical machine learning approach had higher accuracy than the semantic approach.

Although machine learning techniques have certain advantages over the lexicon based approaches such as being quite expensive in term of computing power and time required for training and testing the data.

2.2.5 Data Fusion

Data Fusion is the process of integrating and combing data from multiple sources for deriving more useful and consistent results to get a better understanding of the data. This data can be used to derive important insights. Researchers in different areas, from natural language processing and artificial intelligence to psychology and social sciences, have come together with the growth of research in the field of Affect Computing and its emergence as an interdisciplinary research field. With the rapid growth in the number of videos posted online for customer reviews, movie reviews and political views, affect computing has evolved from traditional uni-modal analysis to complex multi-modal analysis. In the field of emotion recognition, multi-modality can be expressed by the presence of multiple modes of communication such as facial expression, speech emotion recognition, gestures, head movement and textual content. In 2010, a study on multi-modal emotion recognition in speech-based interaction with acoustical analysis accompanied by body gestures and facial expression was performed [45]. Features relevant to the emotion were extracted for each of the three modalities. A Bayesian classifier for automatic classification was used and results of uni-modal and multi-modal approaches were combined. It was observed that the recognition rate was increased by 10% by the multi-modal approach as compared to the most recent uni-modal approach. Poria et al. [46] carried out an extensive study of various state-of-the-art data fusion techniques. The study confirmed previous researchers' conclusions that multi-modal classifiers outperform uni-modal classifiers, and the audio-visual affect detection performance is improved with the help of text.

Chapter 3

Proposed Design and Methodology

In this section, we will discuss the process of creating the dataset, pre-processing the raw input, post-processing the raw output, sentiment analysis of text, Data Fusion and statistical analysis. Figure 3.1 shows the Methodology flowchart.

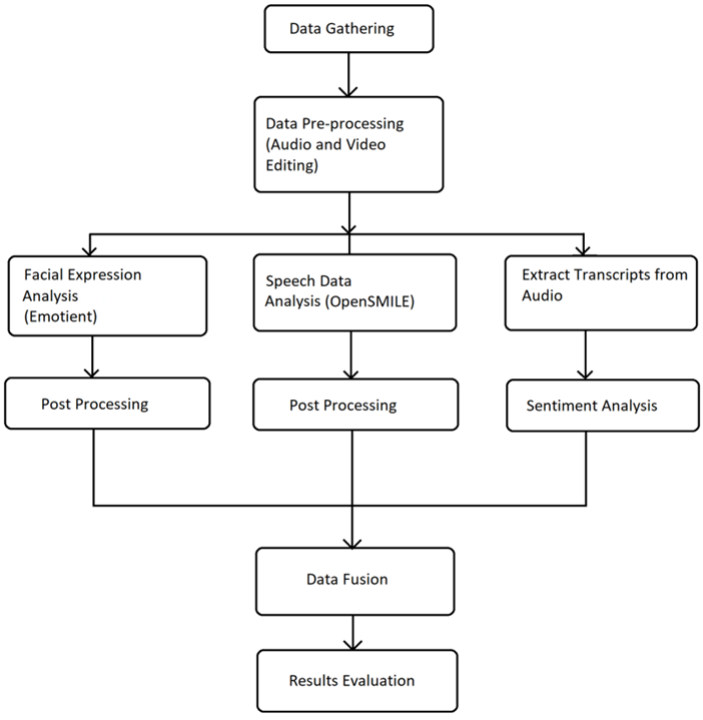

Figure 3.1: System Architecture

3.1 Data Gathering

For this study, a dataset was created consisting of the speeches, press conferences, public addresses and interviews of 4 politicians with Donald Trump being the main subject of the study. The dataset is gender balanced with 2 male and 2 female politicians. The other candidates in focus are Boris Johnson (Prime Minister of United Kingdom), Nancy Pelosi (Speaker of the United States House of Representatives) and Gina Raimondo (Governor of Rhode Island). The dataset consists of a total of 27 audio and videotapes. The recordings contain speeches in various positive and negative contexts such as press briefing regarding the daily updates and restrictions of Covid19, election winning speeches of both Boris Johnson and Donald Trump, addresses to University students, topics including Syria and the Charlottesville shooting. Most of these tapes were recorded by media outlets such as BBC, NBC News, ABC News, The Guardian, etc. and therefore majority of the videos have adequate lighting and are of high quality with a resolution between 720p and 1080p and the audios are clear and intelligible. The age of the candidates lies between 49 (min) and 80 (max), with a mean of 64.75 years. The length of the tapes ranges from 90 seconds up-to 20 minutes. The candidates chosen for this study are trained public speakers who can hide their actual emotions.

As these videos are uploaded to public platform by Media Outlets and consist of videos of public personalities who are aware of the recording, there are no ethical issues that arise in the process of data collection. Table 3.1 shows the breakdown of the dataset by politicians:

Politician	Age	Total Tapes	Percentage of Total Dataset
Donald Trump	74	10	37.03%
Boris Johnson	56	7	25.92%
Nancy Pelosi	80	5	18.51%
Gina Raimondo	49	5	18.51%
Total		**27**	**100%**

Table 3.1: Breakdown of Dataset by Politicians

3.1.1 Data Gathering Steps

The data for this study was gathered manually. The algorithm for gathering the data is described in the steps below:

1. For each candidate, lookup video streaming sites and websites of various media outlets to find suitable videos of speeches and interviews posted in different contexts (such as Press Confer-

ences related to Covid19, acceptance speech by Donald Trump and Boris Johnson, university addresses, etc.). The suitability of the videos is dependent on the resolution (between 720p and 1080p) and sufficient amount of lighting.

2. Download the video using media outlet websites or YouTube downloader in mp4 format.

3. Perform pre-processing (discussed in the next subsection) on the videos and save the pre-processed video in mp4 format, and extract audio from the pre-processed video saving it in WAV format.

4. Now, we have the video as well as the audio of the recording in the required formats.

3.2 Data Pre-processing

Before the tapes can be provided to the Emotion Recognition softwares as inputs, some pre-processing needs to be performed. There are two noteworthy problems with the original videos due to which pre-processing needs to be performed.

1. There are multiple people in the video and therefore, the main focus of the video may shift from the candidate if the camera moves back and forth between people. The change in facial expressions between two or more people would cause inconsistencies in the output generated by the Facial Expression Analysis Software (Emotient).

2. In case of interviews when there are multiple people speaking, there are instances when even though the camera is focused towards a candidate's face, a different person's voice (e.g., interviewer) is heard. Therefore, it is important to ensure that only the candidate's voice is heard. Otherwise the Speech Recognition Software (OpenSMILE) would result in inconsistent output.

To avoid these issues, the videos were trimmed manually using Adobe Premiere Pro CC 2019. The videos were trimmed using the following algorithm:

1. For each video in the dataset, find and delete frames which consist of the face of anyone who is not the candidate.

2. Find frames which consist of the face of the candidate but where the voice output belongs to a different person. Delete the frames.

3. Save the video in mp4 format and extract the audio in WAV format.

Section 3.3 and 3.4 will discuss the processing of video and audio for facial and speech expression analysis by their respective Emotion Recognition softwares.

3.3 Facial Expression Analysis – Emotient

The system used for the analysis of Facial Expression Analysis is Emotient by iMotions. This software was built on top of the pioneering work done by Paul Ekman in the field of emotions and facial expressions. Emotient, also known as FACET, uses the Facial Action Coding System (FACS) to automatically calculate the intensity of 19 facial actions and classify them into one of the 6 basic emotions. It also approximates the locations of 10 facial features along with the 3-D orientation of the head (pitch, yaw and roll) [19]. Each frame of the sequence is compared against the Extended Cohn-Kanade (CK+) database to give a log likelihood result. Emotient, previously known as Computer Expression Recognition Toolbox (CERT) follows the following pipeline of steps for analyzing emotions:

A. *Face Detection*

The CERT facial detection algorithm was trained using an extension of the Viola-Jones approach [47]. It uses a set of Haar-like features that involves summing of the image pixels in rectangular areas. These features are calculated by finding the difference between the sums of pixel values of two rectangular regions. This approach helps in identifying faces in a frame. This algorithm is applied on each frame of the sequence and only the largest face in each frame is segmented for further processing.

B. *Facial Feature Detection*

After the face has been segmented, 10 facial features consisting of inner and outer mouth corners, center of the mouth, eye centers, inner and outer eye corners, eye centers, and tip of the nose are detected using feature-specific detectors. The facial feature detectors output the log-likelihood ratio of the presence of a feature at location (x, y) of the face [19]. The likelihood term along with feature-specific priors calculates the posterior probability of a feature's presence at location (x, y).

C. *Face Registration*

After the set of 10 facial features are located, the size of the rectangular region is re-calculated at a size of 96x96 pixels with the help of an affine warp.

D. *Face Extraction*

The re-estimated 96x96 pixel rectangular region is converted into a single feature vector using complex Gabor filters.

E. *Action Unit Recognition*

The feature vector that is computed at the Face Extraction stage is then input to separate Sup-

port Vector Machines (SVM) to check for the presence of each AU. The output of the SVM is interpreted as an estimate of intensities for each AU.

F. *Expression Intensity and Dynamics*
For every single AU, the output of CERT is a continuous value that signifies the distance between the feature vector and the hyperplane of the SVM. The output was highly correlated to the intensities of the Action Units coded by FACS experts.

3.3.1 Facial Expression Data post-processing

The output from Emotient is a CSV file that contains evidence for the presence of Actions Units (AUs) and the basic emotions (Joy, Anger, Fear, Sadness, Surprise, Contempt and Disgust) as well as the evidence of presence of head movements like yaw, pitch and roll. The evidence is in the form of base 10 log-likelihood. This means that if the evidence value for a certain emotion is 1, then we say that there is a 10 times greater likelihood of that emotion being present. Similarly, if the evidence for the emotion is -1, then we can say that there is a 10 times greater likelihood of the emotion not being present. If the evidence value is 0, then there is a 50 percent chance that the emotion is present.

A python program was developed to automatically post-process the raw output of Emotient. The program converted negative evidence values for each emotion to 0 as negative values indicate absence of emotions and positive values indicate significant evidence for presence of emotion.

3.4 Speech Emotion Recognition – OpenSMILE

The system used to perform this part of the study is an open source software that extracts the features from the acoustic signals and applies Machine Learning and Speech Analytics techniques for analyzing emotions of each candidate. The SMILE in openSMILE stands for Speech and Music Interpretation by Large-space Extraction. OpenSMILE combines the techniques of speech processing and Music Information Retrieval, thereby enabling researchers in both fields to take advantage of the two domains. OpenSMILE allows easy configuration of modular feature extraction components that can be connected via a single configuration file with the help of a simple and scriptable console application [37]. The next subsection will explore the architecture of openSMILE.

3.4.1 OpenSMILE Architecture

This subsection discusses the limitations that were considered during the design of the openSMILE's architecture and the summary of the resulting architecture. The following requirements had to bet met

to overcome the limitations in the existing systems such as lack of extensibility and flexibility, lack of cross-domain feature-set, and lack of incremental processing:

1. **Incremental processing:** Frame by frame pushing data from an input stream through the processing chain.
2. **Ring-buffer memory:** For reusability of data, i.e., to eliminate re-computation of data that is used by multiple feature extractors, and for features that require temporal buffer.
3. Carefully developed algorithms in C/C++ for fast and lightweight computation.
4. **Modular architecture** that employs a run-time plug-in interface and a well-structured API that allows easy addition of feature components and feature combination.
5. Combining multiple feature extraction components into a single configuration file.

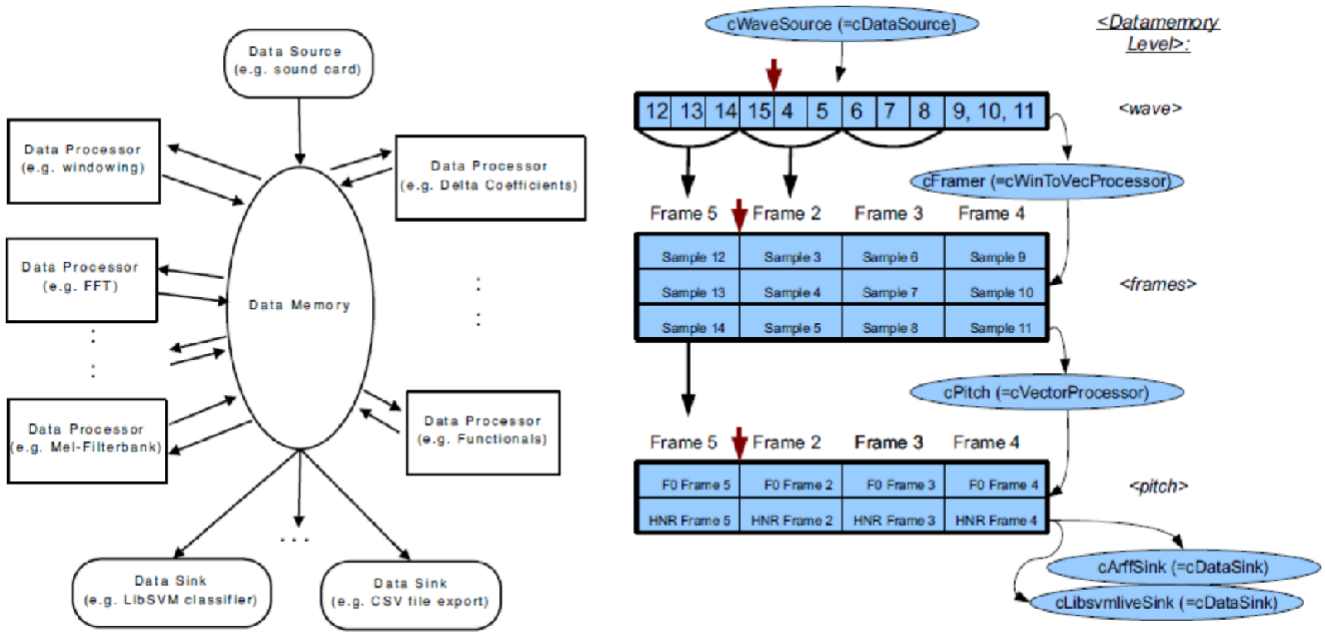

Figure 3.2: Data Flow architecture of openSMILE (left); Ring-buffer based incremental processing (right)

The overall data-flow architecture of openSMILE can be seen in the Figure 3.2 (left). It consists of a central Data Memory that acts as a link between all the Data Sources, Data Processors and Data Sinks. Data Sources are the components that write data from the external sources to the Data Memory. Data Processors are components that read data from the Data Memory, update it and write it

back to the Data Memory. Data Sinks read data from Data Memory and write it to external destinations like files.

The incremental processing using the ring-buffer memory is illustrated in Figure 3.2 (right). It consists of three levels i.e., wave, frames and pitch. The cWaveSource component writes samples to the 'wave' level. These sample are picked up by the cFramer that produces non-overlapping frames and written to the 'frames' level. Pitch features are extracted from these frames and writes them to the 'pitch' level by the cPitch component. All these components have similar functionalities of processing the data and writing it to the Data Memory.

OpenSMILE is capable of extracting Low Level Descriptors (LLD) and then applying various filters and transformations to them. Statistical functions are applied to the low level features for recognizing emotions and Music Information Retrieval.

3.4.2 OpenSMILE Data Processing and Output

The process of emotion recognition is performed by dividing the audio file into chunks of 2 seconds or 2000ms each using a python script. For example, an audio file of 2 minutes will be divided into 60 chunks. After splitting the main audio input file into chunks, a shell script is run from the Windows terminal which takes one chunk at a time and analyzes it for emotion recognition. The raw output is saved in a text file. The outputs for all the segmented audio inputs are read and concatenated into a single CSV file by running a python script. The raw output file that is generated as a result of this script is illustrated in Figure 3.3.

	0	1	2	3	4	5	6	7	8	
0	SMILE-RESULT	ORIGIN=libsvm	TYPE=regression	COMPONENT=arousal	VIDX=0	NAME=(null)	VALUE=-8.158220e-003	NaN	NaN	
1	SMILE-RESULT	ORIGIN=libsvm	TYPE=regression	COMPONENT=valence	VIDX=0	NAME=(null)	VALUE=-2.201908e-001	NaN	NaN	
2	SMILE-RESULT	ORIGIN=libsvm	TYPE=classification	COMPONENT=emodbEmotion	VIDX=0	NAME=(null)	CATEGORY_IDX=1	CATEGORY=boredom	PROB=0	angei
3	SMILE-RESULT	ORIGIN=libsvm	TYPE=classification	COMPONENT=abcAffect	VIDX=0	NAME=(null)	CATEGORY_IDX=1	CATEGORY=cheerful	PROB=0	agressiv
4	SMILE-RESULT	ORIGIN=libsvm	TYPE=classification	COMPONENT=avicInterest	VIDX=0	NAME=(null)	CATEGORY_IDX=1	CATEGORY=loi2	PROB=0	loi1
5	SMILE-RESULT	ORIGIN=libsvm	TYPE=regression	COMPONENT=arousal	VIDX=0	NAME=(null)	VALUE=1.143066e-001	NaN	NaN	
6	SMILE-RESULT	ORIGIN=libsvm	TYPE=regression	COMPONENT=valence	VIDX=0	NAME=(null)	VALUE=-2.442830e-001	NaN	NaN	
7	SMILE-RESULT	ORIGIN=libsvm	TYPE=classification	COMPONENT=emodbEmotion	VIDX=0	NAME=(null)	CATEGORY_IDX=1	CATEGORY=boredom	PROB=0	angei

Figure 3.3: OpenSMILE Raw output

The raw output of openSMILE is hard to understand and requires post processing for easy interpretation and for making the data compatible for the Data Fusion stage. A python script was written

to extract the probability values of emotions (boredom, happiness, sadness, fear, disgust, and anger) into a readable format. The post-processed openSMILE output is illustrated in Figure 3.4.

The figure shows the post-processed output of Donald Trump's speech on the topic of Syria from July 4th, 2017. Each column in the image represents an emotion and each row represents the probability value of that emotion being present at an interval of 2000ms. As it can be seen from the post-processed output, some values of emotions are calculated to be zero. This happens when openSMILE is not able to process a particular chunk of the audio input. In that case, the original output for that audio segment is an empty file and the python script that performs the post-processing automatically sets the value for each emotion to 0. These values could lead to inconsistent results during the analysis phase and therefore, such records are discarded while performing statistical analysis.

	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.023646	0.546082	0.051608	0.088662	0.041894	0.139271	0.108838
3	0.021063	0.566311	0.062623	0.056324	0.029471	0.113971	0.150236
4	0.016738	0.147499	0.012752	0.273520	0.035078	0.479843	0.034569
5	0.020294	0.539383	0.045548	0.022287	0.016551	0.109933	0.246005
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.020894	0.523407	0.041637	0.066267	0.025361	0.227810	0.094624
8	0.012114	0.416806	0.028847	0.009652	0.009580	0.443518	0.079483
9	0.024441	0.581812	0.039549	0.059194	0.032327	0.143888	0.118790
10	0.052838	0.533801	0.032942	0.087896	0.044771	0.145304	0.102447
11	0.029358	0.456058	0.025215	0.162320	0.034952	0.210350	0.081747
12	0.028057	0.519635	0.017710	0.148430	0.119639	0.036784	0.129744
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.023209	0.496243	0.038204	0.092901	0.071549	0.118927	0.158968
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3.4: OpenSMILE Post-processed output

3.5 Sentiment Analysis

This subsection aims at extracting words from the speeches/interviews and using a Lexicon-based approach for categorizing the words into sentiments. The first stage of this process is extracting the transcript of each individual audio file. The extracted text is tokenized to find the emotion in every single word present in the transcript with the help of affect dictionaries.

3.5.1 Transcript Extraction

Initially, the transcript for each audio recording was extracted using a python program written by James Gorman. The program used Google's speech recognition API to process each audio file by converting its audio to text and then saving it in a text file. The program was easily able to process audio files of up to 4 minutes. For files longer than 4-5 minutes, the program could not process them and threw an error. In such cases, a new method was created. Using the script that split audio files into chunks of 2000 ms each, the audio files were divided into segments of 1 minute each or 60000 milliseconds by modifying the parameters of the program. Each segment of the original audio file was processed by the program to generate separate transcripts and a separate text file was created for each one. The output of each single-minute transcript was manually concatenated into a new text file. However, this method had a drawback. By cropping the audio at the one-minute mark, some words may be lost if the words begin right before the minute mark and continue into the next minute. To deal with this problem, a tool called Descript was used that performs automatic transcription. It is easily able to transcribe the entire file irrespective of the length. It was observed that Descript was able to transcribe the text more accurately than the speech recognition API.

Once the final transcript of the audio is ready, it is important to add timestamps to it in order to identify the exact time at which a certain word was spoken. This was also done using Descript. It takes the audio file and its original transcript as inputs and then synchronizes them by time. It adds timestamps for every 1000ms and the words for each interval are saved in a word file.

3.5.2 Text Analysis

Analyzing the sentiment in text was done using a Lexicon-based approach. This was done with the help of Python's Natural Language Processing library along with some existing affect dictionaries for the classification of words into categories like 'positive', 'negative', 'political', 'names', 'places', 'active', 'passive', etc. and counting the occurrence of words in each category for each frame. A Python program was developed to tokenize the words to find the sentiment of each word, find percentages of each category of emotion that is present in the text, create a Data Frame and save it in an excel file for persisting the data. The steps followed for sentiment analysis are enumerated as follows:

1. The first step is creating a Data Frame (or table) by reading the time stamped Word file and extracting the words spoken at an interval of 1000 milliseconds. Every record at 1000 ms represents the words spoken during that time. The output of this step is displayed in Figure 3.5.

MediaTime	Study	Words
0	01-05-20 davos day two	
1000	01-05-20 davos day two	it's a total
2000	01-05-20 davos day two	hoax it's a
3000	01-05-20 davos day two	disgrace
4000	01-05-20 davos day two	they talked about
5000	01-05-20 davos day two	their tremendous
6000	01-05-20 davos day two	case
7000	01-05-20 davos day two	and it's
8000	01-05-20 davos day two	all done that had no
9000	01-05-20 davos day two	case

Figure 3.5: Words spoken at 1000 ms interval during Donald Trump's speech at Davos.

2. The next step is to assign an emotive value and metadata to the word in the table. New columns that represent such emotions and metadata are added to the table such as 'positive', 'negative', 'place', 'political', etc. To achieve this, the entire table is iterated one record at a time and the words in the 'Words' column are tokenized. Each word is then looked up in various dictionaries to classify the word into the various emotional and metadata categories. The different dictionaries that are looked up for this purpose are: (a) General Inquirer; (b) A dictionary of emotive words created using Rojet's Thesaurus; (c) List of Closed Class words; (d) List of political words and phrases. The importance of each dictionary is explained as follows:

   A. *General Inquirer*: It is a system for content analysis that considers sentence as a unit of information [48]. The General Inquirer consists of categories from two main sources: (1) the Harvard IV-4 dictionary; and (2) Lasswell value dictionary. The two dictionaries along with a few more make up a total of 182 categories. It consists of various valence categories such as 'Positive', 'Negative', 'Pain' and 'Pleasure', etc. Currently, the 'negative' category is the largest and contains 2291 entries. A word could have multiple categories associated with it. The General Inquirer consists of 11789 words which includes duplicates.

   B. *Rojet's Thesaurus*:This is a collection of words, their synonyms and antonyms. This thesaurus was used to create a new dictionary for emotive words such as 'happy', 'sad', 'anger', 'fear', 'disgust' and 'surprise' using their synonyms and antonyms. Rojet's thesaurus is different from a traditional dictionary as it is not arranged in an alphabetical order, but rather it is arranged based on the ideas they convey. All words with similar meaning are grouped under one column. For example, the words 'glad', 'delighted', and 'cheerful' are classified under the 'Happy' category.

   C. *Closed-class words*: The words in a language can be divided based on certain features that

they possess in common. The most familiar category of words is the Open word-classes that contain the major parts of speech such as nouns, verbs, adjective and adverbs. These words give information about the content of the communication and tend to give emotional information. Another category of words is the Closed class words that contain the minor parts of speech such as determiners ('the', 'a', 'this', 'that', 'my', 'your', 'one', 'four', 'some', 'many', etc.), pronouns, prepositions, conjunctions, etc. These words act as markers for the structure of a sentence and generally possess no emotive value.

D. *List of Political words and phrases*: A new dictionary created out of politically related terminologies, words, and phrases was created [49]. The words in the dictionary were categorized into positive and negative emotions as well as some political categories such as 'Political movement', 'Constitution', 'Election', etc.

After tokenizing the words, each word is looked up in one or more of the pre-defined dictionaries. First, the word is looked up in the list of closed class words. If the list consists of the word, it is ignored with no emotive value. If the word is not present in the list of closed class words, then it is looked up in the combination of two dictionaries i.e., General Inquirer and the dictionary of emotive words derived from Rojet's thesaurus. The two dictionaries are combined using a full outer join. Full outer join is a concept used in database management which returns all record when there is a match in either of the two tables. Therefore, by performing the full outer join on the two tables, a larger dictionary of words is created where every word occurs only once but it has attributes of both the dictionaries. Now, if the word is found in the combined dictionary, the count of all the relevant categories that are associated with that word is increased by one and saved in the table. If the word is not found in the joined dictionary, it is looked up in the list of political words.

It is important to note that if a word or phrase is found in one dictionary and that word or phrase has an emotive value in that particular dictionary then it is not looked up in any other dictionary otherwise the total count of the emotion will be inconsistent.

Another important factor to consider is the inflection of words. Inflection is the change in the form of word, generally the ending, to express grammatical function such as tense, person and case. The words present in the transcript could take any form based on the way a person speaks. Therefore, it is important to ensure that the emotive information is not lost while performing look ups in such dictionaries. This problem is tackled with the help of Python' Wordnet library. The library contains a function that can reduce the words to their word stem, base or root form. If the word present in the transcript is not present in the dictionaries in its original form, then the dictionaries are looked up with the root form of the word, and if the root form of the word is

present in the dictionaries then the emotive information is updated in the table. The final result of the sentiment analysis of words is displayed in Figure 3.6.

MediaTime	Words	Study	Place	Positive	Negative	Hostile	Disgust	Strong	Weak	Active	...	Ethnicity	Movement	Constitution	Institution
0		01-05-20 davos day two	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
1000	it's a total	01-05-20 davos day two	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	...	0.0	0.0	0.0	0.0
2000	hoax it's a	01-05-20 davos day two	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3000	disgrace	01-05-20 davos day two	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0
4000	they talked about	01-05-20 davos day two	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0
5000	their tremendous	01-05-20 davos day two	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0
6000	case	01-05-20 davos day two	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

Figure 3.6: Sentiment Analysis of Donald Trump' Speech at Davos

It can be observed that during Media time 2000, there is 1 negative word represented by the word 'hoax' in the text. Similarly, in Media time 3000, there is 1 negative word represented by the word 'disgrace'. During Media time 5000, 1 positive word can be observed which is represented by the word 'tremendous'. Other metadata in the table is represented by columns 'Strong', 'Weak', 'Active', etc.

The program also performs basic analysis that includes counting occurrence of each word in the text as well finding percentage of words that belong to a specific emotional or metadata category. Table 3.2 shows part of sorted list of words that occur in Trump's speech at Davos on 05/01/2020. This sorted list of words is inclusive of the closed class words.

Words	the	to	and	I	a	of	it's that	they	is	going	have	this	...
Count	19	10	9	8	7	6	6	5	5	5	5	5	...

Table 3.2: Sorted list of occurrence of words in a transcript

On further observation, it was seen that 177 words out of the total 316 present in the transcript were closed class words which makes up 56% of the total words of the text. Percentages of some of the categories of words are calculated with respect to the non-closed-class words, as displayed in Table 3.3. The total number of remaining words is 316 – 177 = 139.

	Positive Words	**Negative Words**	**Anger Words**	**Strong Words**	**Weak Words**
Total	19	8	0	34	5
Percentage	13.66%	5.75%	0.0%	24.46%	3.57%

Table 3.3: Percentage of words with respect to non-closed-class words

A word may belong to multiple categories. For example, the word 'tremendous' at Media time 5000 belongs to 'Positive' as well 'Strong' category.

The tables created from the transcripts and the basic sentiment analysis results are stored in an Excel file which is used while performing Data Fusion and further analysis.

3.6 Data Fusion

As discussed in section 2.2.5, data fusion is the process of combining and integrating the data from multiple sources to find more useful and consistent information. Data Fusion is a key computational challenge in this study.

Humans can see or hear due to the visual and sound signals that are emanated from a source respectively. Humans can see things around them due to the light that falls on the retina after being reflected from a surface. It is important to note that the speed of light is $3x10^8$ m/s. In contrast, the speed of sound is quite low i.e., 340 m/s. Due to the difference in their speeds, the rates at which these two signals are perceived by humans are also different. Each frame from the output of the facial expression analysis tool i.e., Emotient, can be seen at a rate of 33 milliseconds while each frame of the output of speech emotion recognition tools i.e., openSMILE, can be observed at 2000 milliseconds. For every frame of audio data, there would be approximately 60 frames of facial expression data and therefore, the facial expression output needs to be aggregated for every 2000 milliseconds. Similarly, the output of text sentiment analysis which is present at every interval of 1000 milliseconds also needs to be aggregated to be brought to the same rate as the audio output. For example, the clip for Donald Trump's speech at Davos on 05/01/2020 is 106 seconds or 106000 milliseconds in duration. Since, each frame of the audio output is present at a 2000 milliseconds' interval, there would be a total of 53 frames in the output of openSMILE for Trump's speech. For 53 frames of audio data, we have 3196 frames of facial expression data which is approximately 60 frames for every 2000 milliseconds. In a

similar way, there are 106 records of text analysis data and therefore, 2 records of text data need to be aggregated/combined to bring it to the same rate as that of the audio data. The aggregation is performed using a python program and it is explained in the following subsections.

3.6.1 Aggregation of Sentiment Analysis Data

Aggregating sentiment analysis data is a relatively simple task. As discussed, two records of text sentiment analysis table need to be aggregated to bring it to the same rate as audio output, as the number of frames in the audio output is half of that of text sentiment analysis. This is achieved by concatenating the text for every two consecutive rows and then performing summation of the values for each category for those two rows. Doing this for every two rows of the sentiment analysis table gives the aggregated table of half the size of the original table. The table is illustrated in Figure 3.7.

MediaTime	Words	Study	Place	Positive	Negative	Hostile	Disgust	Strong	Weak	...	Ethnicity	Movement	Constitution	Institution	Disease
0	it's a total	01-05-20 davos day two	0	0	0	0	0	1	0	...	0	0	0	0	0
2000	hoax it's a disgrace	01-05-20 davos day two	0	0	2	0	0	0	1	...	0	0	0	0	0
4000	they talked about their tremendous	01-05-20 davos day two	0	1	0	0	0	1	0	...	0	0	0	0	0
6000	case and it's	01-05-20 davos day two	0	0	0	0	0	0	0	...	0	0	0	0	0
8000	all done that had no case	01-05-20 davos day two	0	0	0	0	0	0	0	...	0	0	0	0	0

Figure 3.7: Aggregated table for Sentiment Analysis

It can be observed in the figure that after aggregating two consecutive rows, the Media Time column consists of values for every 2000 milliseconds instead of the previous interval of 1000 milliseconds. On further observation, it can be seen that the text for Media Time 2000 has the words “hoax it’s a disgrace”, which is the result of the concatenation of the text “hoax it’s a” and “disgrace” at Media Times 2000ms and 3000ms in Figure 3.6 respectively. Figure 3.6 represents the text sentiment analysis table before aggregation. It can also be seen that the value of ‘Negative’ column in the aggregated table is 2, which is the summation of the ‘Negative’ values in Figure 3.6 for rows with Media Time 2000 ms and 3000 milliseconds. Therefore, the aggregation is successfully performed on sentiment analysis data.

3.6.2 Aggregation of Facial Expression Data

The aggregation for facial expression data is also done for every 2000 milliseconds. But facial expression data consists of approximately 60 frames of information or 60 rows in the Emotient Output for every 2000 milliseconds. Therefore, 60 frames have to be aggregated into one single record. This is achieved by finding the descriptive statistics i.e., mean, median, maximum, minimum and standard deviation for every 60 frames. These statistics are calculated for all the basic emotions and the Action Units. The speech by Donald Trump at Davos on January 5th, 2020 has a total of 3196 frames of emotional information for a 106 second video. These 3196 frames have to be aggregated to 53 frames. However, out of 3196 frames, only 3180 frames are required to be aggregated to 53 frames. The remaining 16 are discarded because the media time for these frames exceeds the overall media time for the audio output.

	Joy Score Min	Joy Score Max	Joy Score Mean	Joy Score Std	Anger Score Min	Anger Score Max	Anger Score Mean	Anger Score Std
0	0.018971	0.518739	0.284057	0.238383	0.224912	0.625950	0.407675	0.140301
1	0.000000	0.000000	0.000000	0.000000	0.003445	0.880916	0.435359	0.202433
2	0.000000	0.000000	0.000000	0.000000	0.008110	1.225302	0.308634	0.308033
3	0.210143	0.210143	0.210143	0.000000	0.014959	1.247510	0.305235	0.353502
4	0.210443	0.513362	0.357501	0.141483	0.017038	1.362829	0.460069	0.370748
5	0.000000	0.000000	0.000000	0.000000	0.045110	1.251315	0.468308	0.244633
6	0.000000	0.000000	0.000000	0.000000	0.007538	1.081823	0.524906	0.259814
7	0.000000	0.000000	0.000000	0.000000	0.000204	0.276555	0.109834	0.094246
8	0.000000	0.000000	0.000000	0.000000	0.062372	1.169212	0.722865	0.225424
9	0.156540	0.572776	0.395499	0.094747	0.179598	1.023355	0.567470	0.171421
10	0.000000	0.000000	0.000000	0.000000	0.001647	0.527692	0.225647	0.132919
11	0.252316	0.729610	0.455453	0.150587	0.036074	0.584931	0.313098	0.164035
12	0.000000	0.000000	0.000000	0.000000	0.014774	0.124809	0.075874	0.051058
13	0.000000	0.000000	0.000000	0.000000	0.005392	1.096260	0.287394	0.247502
14	0.000000	0.000000	0.000000	0.000000	0.099858	1.168300	0.483975	0.306303
15	0.000000	0.000000	0.000000	0.000000	0.137489	1.258402	0.555368	0.283946

Figure 3.8: Aggregated table for Joy and Anger - Trump at Davos

The aggregation is performed by iterating through 60 frames at a time and calculating the descriptive statistics for each emotion and Action Unit. For simplicity, the focus is on Joy values. While iterating through the rows of Joy column, the evidence value is used to calculate the mean, median, maximum, minimum and standard deviation. It is important to note that only positive evidence values are considered while finding these statistics. Evidence is represented as base 10 logarithmic values, which implies that values below 0 suggest that the likelihood of an emotion being present is less than 50%. As evidence less than 0 is considered as the likelihood of absence of an emotion, these statistics are calculated only by using evidence values greater than 0. These statistics are used as the aggregated values and converted into one single row of emotional information for 60 frames. This

process is repeated till all the frames of the Emotient output are processed for all the emotions and Action Units. Figure 3.8 illustrates the first 16 frames of the aggregated information for Joy and Anger for Donald Trump's speech at Davos.

3.6.3 Fusion

After aggregating the data for facial expressions and text sentiment analysis, they are brought to the same number of frames as that of audio data. The tables for each of these sources are then concatenated using a script. The concatenated table includes emotional information from all three modalities which is used for further analysis. After concatenating the tables, new columns are derived by performing logical and summation operations on columns for all these modalities.

One of the aims of performing Data Fusion is to identify whether similar emotions are present at the same interval of time. This can be achieved by creating new columns by performing Logical operations such as OR and AND. The data to be used for these operations contains emotion scores for facial expression, probability of emotion in speech, and the presence of emotion in text. The logical operations are performed on columns of similar emotions of the integrated data. The fusion using these operations makes use of threshold values for each modality. If the value for an emotion is greater than or equal to the threshold value during a particular time interval, then it is concluded that the emotion is present. For facial expressions, the threshold is fixed at 0 which implies that if the evidence value is greater than or equal to 0, then the emotion is considered to be present in the face. In case of speech audio, the threshold is fixed at 0.3 which is a probabilistic value of an emotion being present. An emotion is present in text if the value of the emotion at a time interval is greater than 0. The number in the emotion column for text denotes the number of words that belong to a particular emotion. For example, if the number of positive words in the text at a particular time interval is 2, then the value in the 'Positive' column for that interval will be 2. Therefore, the threshold used for text data is 1 as a means to identify the presence of an emotion.

The logical operations provide results in the form of True or False, or 0 and 1. 0 denotes absence of emotion and 1 denotes presence of emotion. The OR operation results in 1 if the emotion is present in at least one of the modalities. On the other hand, the AND operation results in 1 if the emotion is present in all modalities at once. Another way of representing the presence of emotions is the use of the summation method. The summation method assigns the value of 1, 2 or 3 depending on the number of modalities in which the emotion is present.

Using the new columns that are created by the logical and summation operations, multi-modal sentiment analysis is performed for all three modalities.

Chapter 4

Case Studies and Results

In this section, we will carry out statistical analysis of post-processed as well as fused data. A comparison of the distribution of emotions in each mode will be calculated and contrasted with other candidates in the study.

We will also analyze the data to find similarity in emotions across different different modalities and find correlation between the overall emotion to its dependence on each modality.

4.1 Individual Distribution of Emotions

This is the first step in the analysis. It involves calculating the distributions of the emotions of happiness, anger and sadness. The average and standard deviations are calculated for 10 videos in which Donald Trump had appeared. This can be displayed in Table 4.1.

The data used for this analysis consists of percentages of words spoken for text, and the post-processed output for facial expressions and speech i.e., data before the data fusion stage. This means that the percentage of facial expressions is calculated corresponding to the number of frames in the output of Emotient.

From a cursory look at Table 4.1, it is apparent that Donald Trump's facial expression tend to display emotions of Anger and Sadness more than that Happiness or Joy. Although the presence of negative emotions is larger than happiness, they show more variation as is clear from the standard deviation of 16.54% and 18.97% for Anger and Sadness respectively, whereas it is only 5.36% for happiness. It is also clear that the evidence of these three emotions in speech is extremely low. Another noteworthy observation is that the distribution of emotions in text. The average of positive words in Trump's speeches exceed negative words by almost 2 times.

Date	Happiness			Anger			Sadness		
	Facial	**Speech**	**Text**	**Facial**	**Speech**	**Text**	**Facial**	**Speech**	**Text**
05-01-2020	19.36%	0.00%	6.65%	29.21%	0.00%	0.00%	20.55%	0.08%	2.84%
17-02-2017	3.66%	0.01%	9.78%	1.22%	0.00%	0.00%	14.22%	0.02%	5.43%
11-01-2017	6.85%	0.01%	7.07%	1.02%	0.00%	0.00%	7.03%	0.01%	3.19%
12-03-2020	0.10%	0.00%	5.99%	3.81%	0.00%	0.00%	24.89%	0.02%	5.22%
30-03-2018	1.36%	0.03%	17.50%	28.40%	0.00%	0.00%	11.30%	0.00%	2.12%
07-04-2017	9.85%	0.02%	8.45%	2.16%	0.00%	0.44%	35.66%	0.11%	8.88%
15-08-2017	2.22%	0.01%	5.91%	48.10%	0.02%	0.00%	10.93%	0.01%	6.67%
18-04-2017	5.08%	0.00%	12.24%	4.32%	0.00%	0.00%	15.23%	0.00%	4.34%
09-11-2016	2.65%	0.00%	9.53%	11.30%	0.00%	0.00%	12.08%	0.07%	4.06%
20-01-2017	3.06%	0.00%	9.66%	35.89%	0.01%	0.00%	37.78%	0.03%	4.35%
Average	**5.42**%	**0.01**%	**9.28**%	**16.54**%	**0.00**%	**0.04**%	**18.97**%	**0.03**%	**4.71**%
Std Dev	**5.36**%	**0.01**%	**3.34**%	**16.41**%	**0.00**%	**0.13**%	**10.10**%	**0.04**%	**1.88**%

Table 4.1: Donald Trump Emotion Distribution

These calculations are done for all 4 politicians. The averages and standard deviations are calculated for each politician. Table 4.2 shows the values computed. The table clearly shows that all the politicians in the study display similar values for emotions in speech which lies between 0.01% and 0.17%.

The table shows that Donald Trump and Boris Johnson display a similar distribution of emotions in every category except for sadness in facial expressions. Trump tends to show more sadness in his speeches than Boris Johnson by almost 3 times.

On average, positive words account for 8.32% of the total words spoken by a politician, while negative words account for 4.36% of the total words. Positive words occur almost twice the total number of negative words on average for each politician.

Out of the two female politicians, Nancy Pelosi tends to be more joyful in her facial expressions with an average of 33.05%, while Gina Raimondo expresses the least happiness using facial expressions i.e., 2.09%. Subsequently, Nancy Pelosi expresses the least sadness in facial expression with 1.11%, while Gina Raimondo expresses the most sadness in her speeches using facial expressions with an average of 32.12%.

Politician	Happiness			Anger			Sadness		
	Facial	**Speech**	**Text**	**Facial**	**Speech**	**Text**	**Facial**	**Speech**	**Text**
Donald Trump	5.42%	0.01%	9.28%	16.54%	0.00%	0.04%	18.97%	0.03%	4.71%
Boris Johnson	4.03%	0.02%	8.04%	16.62%	0.00%	0.06%	6.73%	0.17%	4.27%
Gina Raimondo	2.09%	0.01%	7.76%	0.06%	0.02%	0.11%	32.12%	0.01%	5.12%
Nancy Pelosi	33.05%	0.01%	8.20%	0.16%	0.00%	0.03%	1.11%	0.05%	3.39%
Average	**11.14**%	**0.01**%	**8.32**%	**8.33**%	**0.005**%	**0.06**%	**14.73**%	**0.22**%	**4.36**%
Std Dev	**14.66**	**0.005**	**0.665**	**9.5**	**0.01**	**0.035**	**13.78**	**0.07**	**0.74**

Table 4.2: Politician's Average Emotional Distribution

4.2 Multi-modal analysis

This step of the analysis focuses on identifying similar emotions in all three modes of communication at a given time. One way of representing the presence of emotions in multiple modalities is by adding new columns to the table for each emotion representing the presence of the emotion in one, two or three modalities. The columns that denote this are suffixed by '1', '2' or '3'. For example, 'Happiness_1', 'Happiness_2', and 'Happiness_3'. This means that if only one modality contains happiness for a given time interval, then 'Happiness_1' will contain the value '1', else it will be '0'. Similarly, if two or more modalities contain the emotion then the columns 'Happiness_2' or 'Happiness_3' will be assigned the corresponding value. As discussed in section 3.6.3, logical operation of OR indicates the presence of an emotion in at least one modality, and the opeartion of AND indicates the presence of emotions in all modalities at once. Figure 4.1 shows the presence of *Happiness* across modalities for Trump's speech at Davos. Using the threshold values of 0 for Facial Expressions, 0.3 for speech and 0, the presence of emotions in the different modalities can be validated from the table. On calculating further, it was observed that there were 28 instances of Happiness or *Joy* being present in just one modality, 8 instances of happiness in 2 modalities and 0 instances of happiness in all three modalities at once.

Table 4.3 displays the information of the number of modalities that are present for the emotions of happiness, sadness and anger. The table shows the number of frames in which the emotions of happiness, anger and sadness are observed for Donald Trump. It is clear that the presence of only a single modality is much higher than its presence across more than one modality for each emotion. The figures in the table understandably indicate that the presence of all the modalities in expressing

	Joy Score Max	Positive_Text	Happiness_Audio	Happiness_by_AND	Happiness_by_OR	Happiness_1	Happiness_2	Happiness_3	Happiness_Sum
0	3.053530	0	0.057057	0	1	1	0	0	1
1	1.827157	0	0.076841	0	1	1	0	0	1
2	2.482759	0	0.124858	0	1	1	0	0	1
3	2.782357	0	0.021486	0	1	1	0	0	1
4	2.856571	1	0.059213	0	1	0	1	0	2
5	2.603873	0	0.050416	0	1	1	0	0	1
6	2.651033	0	0.026399	0	1	1	0	0	1
7	2.141612	1	0.048955	0	1	0	1	0	2
8	0.263518	0	0.109739	0	1	1	0	0	1
9	2.251266	0	0.001182	0	1	1	0	0	1
10	3.168813	0	0.053402	0	1	1	0	0	1
11	3.200344	0	0.087033	0	1	1	0	0	1
12	2.393666	0	0.073112	0	1	1	0	0	1
13	3.666274	1	0.063379	0	1	0	1	0	2
14	1.926570	1	0.082004	0	1	0	1	0	2
15	1.757967	0	0.065183	0	1	1	0	0	1

Figure 4.1: Presence of Happiness across modalities

a sentiment at the same time is negligible. However, there is some evidence that suggests that two modalities are slightly more likely to be present for emotions of happiness and sadness, but not for anger.

Date	**Happiness**			**Anger**			**Sadness**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
05-01-2020 in Davos	28	8	0	36	0	0	42	3	1
11-01-2017 First press	118	39	0	28	0	0	145	1	0
17-02-2017 Press conf.	172	3	0	55	0	0	341	8	0
12-03-2020 Covid address	33	0	0	133	0	0	266	7	0
30-03-2018 on Easter	20	6	0	38	0	0	36	0	0
07-04-2017 On Syria	29	3	1	20	0	0	44	6	0
18-04-2017 Speech	22	6	0	14	0	0	31	0	0
15-08-2017 Charlottesville	47	8	0	116	2	0	81	5	0
09-11-2016 Victory Speech	70	14	0	93	0	0	137	11	0
20-01-2017 Inauguration	147	1	0	359	1	0	344	14	0

Table 4.3: Frame count for number of modalities per emotion

Trump's speech at Davos on 05-01-2020 consists of 53 frames. On careful observation, it is seen that the sum of the number of frames accounting for the presence of happiness, anger and

sadness in a single modality is 106, which is twice the actual number of frames in the integrated data for the speech. This suggests that there is an overlap of emotions across one or more modes of communication and can be used as a way to prove leakage of emotions.

The next part of the analysis will deal with the correlation of each modality with the presence of each of the three emotions. This will help us in identifying which modality contributes the most to an emotion being present.

4.2.1 Correlation of Modalities and Emotions

This part of the analysis helps in finding out which mode of communication contributes the most to an emotion in a speech or public appearance for a politician. Table 4.4 shows the correlation between the happiness present in 1, 2 and 3 modes of communication.

Date and Occasion	**Facial**			**Speech**			**Text**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
05-01-2020 in Davos	42.76	35.63	0	0	0	0	-16.51	69.98	0
11-01-2017 First press	25.89	42.47	0	-10.85	20.56	0	-6.11	58.49	0
17-02-2017 Press conf.	85.91	12.10	0	12.40	-0.69	0	28.15	31.90	0
12-03-2020 Covid address	28.47	0	0	16.38	0	0	92.99	0	0
30-03-2018 on Easter	-21.25	72.64	0	-16.64	38.04	0	60.02	33.70	0
07-04-2017 On Syria	65.90	23.14	13.10	-15.24	-3.39	100	0.76	54.16	30.66
18-04-2017 Speech	6.74	57.73	0	0	0	0	33.63	42.81	0
15-08-2017 Charlottesville	40.41	46.29	0	-7.35	34.29	0	48.12	36.76	0
09-11-2016 Victory Speech	31.40	53.73	0	0	0	0	52.25	43.04	0
20-01-2017 Inauguration	40.37	15.42	0	6.50	-0.25	0	79.33	8.07	0
Average	**34.66**	**35.92**	**1.31**	**-1.47**	**8.85**	**10**	**37.26**	**37.89**	**3.06**
Standard Deviation	**29.37**	**22.93**	**4.14**	**11.17**	**15.89**	**31.62**	**36.42**	**21.47**	**9.69**

Table 4.4: Correlation of all modalities with Happiness

The Facial, Speech and Text columns indicate the modality in which happiness is being correlated. The columns 1, 2 and 3 indicate percentage of correlation with respect to each modality when happiness is present in 1, 2 or 3 modes of communication. For example, the column Facial-1 indicates the percentage of correlation between Facial expressions and happiness when happiness is present only in one modality. Similarly, Facial-2 indicates the percentage of correlation between happiness and facial expressions when happiness is present in two modalities. The table tries to signify the contribution of each modality to happiness in multi-modal communication.

The values were calculated using Pearson's correlation. This statistical technique is used to find if there is a linear relationship between two quantitative variables. Python's Scipy.stats library consists of a function called 'pearsonr' that finds the correlation between any two quantitative variables. The value of Pearson correlation lies between -1 and +1. Values between 0.9 and 1.0 indicate very high correlation. Correlation coefficients where the value is between 0.7 and 0.9 indicate high correlation. Correlation coefficients where the value is between 0.4 and 0.7 indicate moderate correlation. Correlation coefficients where the value is between 0.3 and 0.4 indicate low correlation, and correlation coefficients where the value is less than 0.3 indicate very low correlation. Negative values indicate anti-correlation between two variables. The values in the table are indicative of the percentage of correlation by multiplying the original correlation coefficient values by 100.

From Table 4.4, it can be observed that on average, facial expressions and text have similar values of correlation with happiness when happiness is present in one or two modalities, with text having a slightly higher value of correlation. However, the standard deviation of the correlation coefficients for text is larger than that of facial expressions which indicates that the contribution of text towards happiness is more varied in both cases, i.e., where happiness is present in one mode and where happiness is present in two modes.

It can also be seen that speech on average is anti-correlated with happiness when happiness is present in only one modality and there is very low correlation when it is present in two modalities.

The last noteworthy detail is that none of the modalities show any significant level of correlation when happiness in all three modalities exists. This is because in most of the tapes from the dataset where Donald Trump makes a speech or appearance, there are no instances of happiness being present in all modalities. However, presence of speech is directly correlated with happiness when happiness is present in all three modalities.

Similarly, Table 4.5 and Table 4.6 represent the correlation of anger and sadness with the different number of modalities respectively.

As discussed in previous subsection, presence of anger in more than one modality is negligible. Using that information along with Table 4.5 suggests that Donald Trump expresses anger through

facial expressions as there is a very high percentage of correlation between facial expressions and anger when anger exists in only one modality. The low standard deviation in Facial-1 column suggests that the the values are not highly varied.

	Facial			Speech			Text		
	1	2	3	1	2	3	1	2	3
Average	**96.22**	**0.31**	**0**	**-8.20**	**17.06**	**0**	**1.78**	**0**	**0**
Standard Deviation	**9.32**	**0.66**	**0**	**22.07**	**36.63**	**0**	**5.63**	**0**	**0**

Table 4.5: Correlation of different modalities with Anger - Donald Trump

It is also apparent from Table 4.6 that Trump expresses sadness mostly through facial expressions when there is evidence of sadness in a single modality. For the few frames where sadness exists across two modalities, the evidence signifies that there is higher correlation between speech and sadness compared to facial expressions and text. However, the standard deviation suggests that this value is moderately varied.

	Facial			Speech			Text		
	1	2	3	1	2	3	1	2	3
Average	**81.63**	**6.68**	**0.52**	**-29.93**	**61.94**	**4.84**	**-0.41**	**1.20**	**-0.62**
Standard Deviation	**17.98**	**4.92**	**1.66**	**28.88**	**37.20**	**15.31**	**10.59**	**5.08**	**1.97**

Table 4.6: Correlation of different modalities with Sadness - Donald Trump

The multi-modal analysis performed in this section concludes that there is negligible evidence of an emotion being present in all modalities at the same time. The evidence also suggests that facial expressions contribute more to the presence of an emotion compared to the other modalities.

4.3 Detecting Leakage of Emotion

Emotional leakage can be detected when evidence of multiple emotions exists in the same or multiple modalities at the same time interval. Evidence for *Happiness* and *Anger* is observed to identify emotional leakage through facial expressions. By observing the behavior for Happiness Vs Anger in Donald Trump's facial expressions in his appearance at Davos on 05-01-2020, a time series was constructed which is displayed in Figure 4.2.

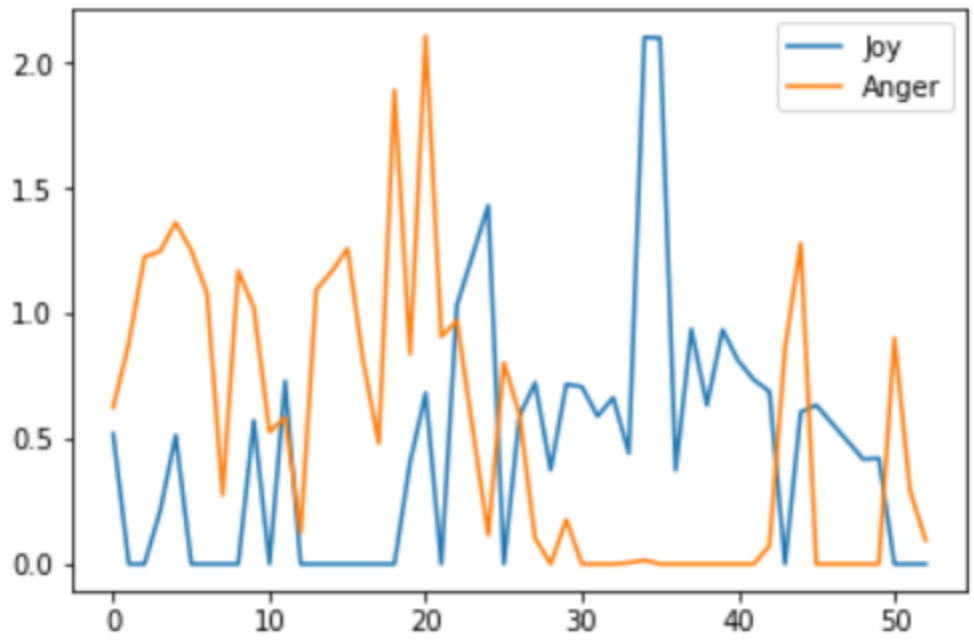

Figure 4.2: Leakage of Emotions between Joy and Anger

From Figure 4.2 it can be observed that there is a lot of positive evidence where anger and joy appear at the same time. From the first frame until the 20th frame, the value of Anger on average is higher than that of Joy. Then the dominant emotion in facial expressions changes to joy while the evidence of Anger disappears almost completely till the the 42nd frame. The transition in emotions in facial expressions may denote the shifting moods during the speech. The positive evidence of anger and joy simultaneously indicates leakage of emotion through facial expressions.

4.3.1 Detecting Emotional Leakage between Facial Expression and Speech

To detect leakage of emotion across multiple modalities, each frame of an individual appearance of a politician can be examined. In this case, evidence of Happiness across facial expressions is looked up against evidence of Sadness in speech during the same time intervals. Table 4.7 shows evidence of these two emotions in the two modes during the last 5 frames of Trump's speech at Davos.

On examining Frame 49, we see that evidence for facial expressions that indicate Joy is not accompanied by significant evidence for sadness as the threshold for an emotion to exist in speech is 0.3. However, on further observation, it can be seen that in frame 50 that positive evidence for Joy in facial expressions is accompanied by a high evidence of Sadness in speech. The presence of two separate emotions across two modalities in this frame suggests that there is a possible leakage of emotion.

Frame	Joy Facial Evidence	Sadness Speech Evidence	Leakage
49	53	3	5.6
50	0.42	0.92	Yes
51	0	0.08	No
52	0	0.10	No
53	0	0.11	No

Table 4.7: Emotional Leakage - Joy/Sadness

Table 4.8 contains information of possible leakage of emotion for joy in facial expressions and sadness in speech, for all of Donald Trump's appearances in the dataset. The table displays a count of the total number of frames where positive evidence of Joy in facial expressions is accompanied by Sadness in speech.

Date	No. of Frames	Frames with Joy/Sadness leakage
05-01-2020 at Davos	53	3
11-01-2017 First press conf.	213	1
17-02-2017 Press Conference	501	4
12-03-2020 Covid Address	281	0
30-03-2018 Easter	39	0
07-04-2017 on Syria	53	4
18-04-2017 Speech	42	0
15-08-2017 Charlottesville	120	0
09-11-2016 Victory Speech	189	4
20-01-2017 Inauguration	387	2

Table 4.8: Donald Trump overall emotional leakage in Joy/Sadness

The results from Table 4.8 identify that there is possible leakage of emotion in 6 out of the 10 public appearances of Donald Trump. Similar results were obtained for Boris Johnson. However, in the case of Gina Raimondo, possible leakage of emotion for Joy/Sadness was observed in only 1 out of the 5 appearances. And in the case of Nancy Pelosi, possible leakage of emotion was observed in all 5 of her appearances from the dataset.

4.3.2 Detecting Emotional Leakage across multiple modalities

Another way of detecting possible leakage of emotion is by observing evidence of anti-correlated emotions across all three modalities. A few frames of such information can be displayed in Table 4.9 depicting possible leakage of emotion.

Frame	Joy Facial	Sadness Speech	Positive Words	Negative Words	Leakage
91	1.23	0.89	1	0	Yes
92	2.43	0.05	2	0	No
93	2.90	0.01	2	0	No
94	2.33	0.03	2	0	No
95	1.80	0.01	1	0	No
96	3.18	0.03	1	1	Yes

Table 4.9: Emotional Leakage across all modalities- Joy/Sadness

It can be clearly observed that there is high evidence of Joy being present in facial expressions from frame number 91 to 96. In frame 91, the text consists of a positive word which reinforces the emotion of happiness because of its presence in two modalities. However, there is high evidence of sadness in speech and therefore, there is possible leakage of emotion in Frame 91. From frame 92 to 96, the emotion of happiness is consistent across facial expressions and text and there is no evidence of sadness in speech and text, except in the case of frame 96 where a negative word is present. Presence of opposing emotions in facial expressions and text also indicate possible leakage of emotion.

Chapter 5

Conclusions

This study was aimed at analyzing facial expressions and speech acoustics as the physical correlates of emotion as well as text from speech using a lexicon-based approach. The distribution of emotions across the three modalities for all politicians in the dataset was found. It was observed that Donald Trump and Boris Johnson displayed similar emotions except in case of sadness through facial expressions, while Nancy Pelosi expressed more happiness than any other politician in the dataset and Gina Raimondo expressed the most sadness through facial expressions.

Correlation between emotions and their presence in different modalities was calculated and it was found that an emotion was only present in one modality in majority of the cases and there was very low evidence of an emotion being present in all modalities at the same time. It was also found that the contribution of speech towards the presence of an emotion was the lowest. Facial expressions were found to be the most correlated with the emotions.

Possible leakage of emotion was detected in uni-modal communication through facial expressions. Positive evidence of Joy and Anger was observed and it was found that there was an overlap of these emotions during a number of frames. Possible leakage of emotion was also detected by observing presence of different emotions across multiple modalities in the fused data. The happiness and sadness evidence for Donald Trump was observed and the presence of both emotions in the different modalities indicated leakage of emotion.

The evidence from the study suggests that emotions are not always consistent across different modalities.

Chapter 6

Future work

There is substantial scope for future works in the field of Emotion Recognition. Multi-modal analysis using other modalities such as gestures, head movement, and body language could be used together with the physical correlates of emotions such as facial expressions and sound acoustics.

Other emotion recognition systems such as Affectiva for facial expressions, and DAVID or Praat for speech recognition can be used ensuring the consistency of the outputs. The accuracy of sentiment analysis systems can be improved by applying other dictionaries such as Whissell's Affect dictionary which denotes the pleasantness and imagery of a word. Similar solutions with respect to other modalities can be applied and Data Fusion between all such modes of communication can be performed for better understanding of the correlation between them.

Furthermore, more politicians can be included as part of the study to understand their behaviors and make comparisons between politicians from different parts of the world.

Bibliography

[1] P. Ekman, "A Methodological Discussion of Nonverbal Behavior," *The Journal of Psychology*, vol. Vol. 43, no. No. 1, pp. 141–149, 1957.

[2] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental psychology and nonverbal behavior*, vol. Vol. 1, no. No. 1, pp. 56–75, 1976.

[3] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Annual Review of Psychology*, vol. Vol. 30, no. No. 1, pp. 527–554, 1979.

[4] C. Izard, *Human Emotions*. New York: Plenum Press, 1977.

[5] D. Sauter, F. Eisner, P. Ekman, and S. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 2408–12, 02 2010.

[6] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," 01 1999.

[7] P. Ekman and W. V. Friesen, "Facial Action Coding System," *CA: Consulting Psychology Press*, 1978.

[8] C. V. Ward, "Hands. By John Napier (Revised by Russell H. Tuttle). Princeton, NJ: Princeton University Press.," *American Journal of Physical Anthropology*, vol. Vol. 93, no. No. 1, pp. 137–139, 1994.

[9] J. Bono and R. Ilies, "Charisma, positive emotions and mood contagion," *The Leadership Quarterly*, vol. Vol. 17, pp. 317–334, 2006.

[10] I. L. Janis and S. Feshbach, "Effect of fear-arousing communications.," *Journal of Abnormal Psychology*, vol. Vol. 48, no. No. 1, pp. 78–92, 1953.

[11] C. Frith, "Introduction," *Philosophical Transactions of Royal Society London - Series B Biological Science*, vol. 358(1431), pp. 431–434, 2003.

[12] K. D. Craig, S. A. Hyde, and C. J. Patrick, "Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain.," *Pain*, vol. Vol. 46, no. No. 2, pp. 161–171, 1991.

[13] P. Ekman, W. V. Friesen, and R. C. Simons, *Is the startle reaction an emotion?*, vol. 49(5), pp. 1416–1426. 1985.

[14] M. Archinard, V. Haynal-Reymond, and M. Heller, "Doctor's and patients' facial expression and suicide reattempt risk assessment," *Journal of psychiatric research*, vol. 34, pp. 261–2, 05 2000.

[15] J. J. Christy, I. T. J. Swamidason, S. Selvam, and A. Xavier, "Emotion detection in text and acoustic based applications," *Journal of Green Engineering*, vol. Vol. 10, pp. 3006–3020, 2020.

[16] D. Messinger, M. Mahoor, S.-M. Chow, and J. Cohn, "Automated measurement of facial expression in infant-mother interaction: A pilot study," *Infancy : the official journal of the International Society on Infant Studies*, vol. 14, pp. 285–305, 05 2009.

[17] T. Kanade, J. F. Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pp. 46–53, 2000.

[18] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," pp. 94 – 101, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[19] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 298–305, IEEE, 2011.

[20] Bartlett, M. Stewart, G. Littlewort, Ford, J. Movellan, I. Fasel, and M. Frank, "Automated Facial Action Coding System. United States US8798374B2, filed August 26, 2009, and issued August 5, 2014."

[21] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California*, 2005.

[22] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," in *Electronics and Communications*, vol. Vol. 53A, (Japan), pp. 36–43, 1970.

[23] D. Klatt, “Review of the ARPA speech understanding project,” *Journal of the Acoustical Society of America*, vol. VOl. 62, pp. 1345–1366, 1977.

[24] F. Jelinek, L. Bahl, and R. Mercer, “Design of a linguistic statistical decoder for the recognition of continuous speech,” *IEEE Transactions on Information Theory*, vol. VOl. 21, no. No. 3, pp. 250–256, 1975.

[25] P. N. Juslin and K. R. Scherer, “Speech emotion analysis,” *Scholarpedia*, vol. Vol. 3, no. No. 10, p. 4240, 2008.

[26] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters,” *Procedia Technology*, vol. Vol. 9, pp. 1112–1122, 2013.

[27] J. P. Teixeira and P. O. Fernandes, “Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices,” *Procedia Technology*, vol. Vol. 16, pp. 1228–1237, 2014.

[28] M. K. Pichora-Fuller, K. Dupuis, and P. Van Lieshout, “Importance of F0 for predicting vocal emotion categorization,” *The Journal of the Acoustical Society of America*, vol. Vol. 140, no. No. 4, p. 3401, 2016.

[29] X. Li, J. Tao, M. Johnson, J. Soltis, A. Savage, K. Leong, and J. Newman, “Stress and emotion classification using jitter and shimmer features,” vol. 4, pp. IV–1081, 05 2007.

[30] L. Rachman, M. Luini, P. Arias, A. Lind, P. Johansson, L. Hall, D. Richardson, K. Watanabe, S. Dubal, and J.-J. Aucouturier, “David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech,” *Behavior research methods*, pp. 323–343, 2018.

[31] K. Scherer, “Vocal communication of emotion: A review of research paradigms. speech communication 40, 227-256,” *Speech Communication*, vol. VOl. 40, pp. 227–256, 2003.

[32] P. N. Juslin and P. Laukka, “Communication of emotions in vocal expression and music performance: different channels, same code?,” *Psychological Bulletin*, vol. Vol. 129, no. No. 5, pp. 770–814, 2003.

[33] T. Bänziger, S. Patel, and K. R. Scherer, “The Role of Perceived Voice and Speech Characteristics in Vocal Emotion Communication,” *Journal of Nonverbal Behavior*, vol. Vol. 38, no. No. 1, pp. 31–52, 2014.

[34] P. Laukka, P. Juslin, and R. Bresin, “A dimensional approach to vocal expression of emotion,” *Cognition and Emotion*, vol. Vol. 19(5), pp. 633–653, 2005.

[35] W. C. S. KN, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. Vol. 52, pp. 1238–1250, 1972.

[36] M. M. H. E. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. Vol. 44, pp. 572–587, 2011.

[37] F. Eyben and B. Schuller, "Opensmile:): The munich open-source large-scale multimedia feature extractor," *SIGMultimedia Rec.*, vol. Vol. 6, no. No. 4, p. 4–13, 2015.

[38] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," 08 2012.

[39] K. Ahmad, J. Han, E. Hutson, C. Kearney, and S. Liu, "Media-expressed negative tone and firm-level stock returns," *Journal of Corporate Finance*, vol. 37, 12 2015.

[40] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, vol. Vol. 62, no. No. 3, pp. 1139–1168, 2007.

[41] A. Devitt and K. Ahmad, "Sentiment polarity identification in financial news: A cohesion-based approach," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 984–991, 2007.

[42] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. Vol. 37, pp. 267–307, 2011.

[43] M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment analysis of tweets using machine learning approach," pp. 1–3, 08 2018.

[44] D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. Vol. 44, no. No. 8, pp. 1077–1087, 2014.

[45] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. Vol. 3, no. No. 1, pp. 33–48, 2010.

[46] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. Vol. 37, pp. 98–125, 2017.

[47] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. Vol. 57, no. No. 2, pp. 137–154, 2004.

[48] P. Stone, R. Bales, J. Namenwirth, and D. Ogilvie, "The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information," *Behavioral Science*, vol. Vol. 7, pp. 484 – 498, 2007.

[49] Scholastic, "https://www.scholastic.com/teachers/articles/teaching-content/vocabulary-political-words," *Vocabulary: Political Words*.