National College of Ireland

Emotion Detection Using Deep Learning Models on Speech and Text Data

MSc Research Project
MSc Data Analytics

Neha Eldho
Student ID: x21217793

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Neha Eldho
Student ID:	x21217793
Programme:	MSc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	14/08/2023
Project Title:	Emotion Detection Using Deep Learning Models on Speech and Text Data
Word Count:	9942
Page Count:	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Emotion Detection Using Deep Learning Models on Speech and Text Data

Neha Eldho
x21217793

Abstract

With the incorporation of artificial intelligence and deep learning techniques, emotion detection, a multidisciplinary area rooted in psychology, cognitive science, and computer science, has seen major breakthroughs. This research goes into the historical progression of emotion recognition, from Paul Ekman's founding work to today's cutting-edge deep learning models. A comparison of emotion identification in text and voice modalities was performed, showing the distinct problems and benefits that each brings. The paper assesses several models, including classic machine learning techniques, LSTMs, hybrid models, and ensemble approaches, on both text and speech data through a series of experiments. The results show that, while both modalities have advantages, voice data frequently delivers greater emotional clues, even when using the same model architecture. The paper also highlights the use of multi-modal data in improving emotion identification accuracy. The integration of different modalities, the use of transformer topologies, and ethical issues in emotion detection are all possible future avenues. The main objective is to use technical advances to better comprehend and interpret human emotions, opening the door for more empathic and responsive artificial intelligence systems.

1 Introduction

1.1 Background

Scholars from all eras and disciplines have been fascinated by the complex web of human emotions for centuries, from the depths of psychological research to philosophical musings. As a result of the digital transformation, this investigation set out on a novel course that penetrated the areas under computational study. A paradigm change occurred, and data-centric methodologies were employed in order to better understand the once-mysterious enigma of emotions, which was previously restricted to the field of human introspection. In the second half of the 20th century, researchers demonstrated their inventive efforts to comprehend the complexities of human emotions. These researchers utilized the power of computer approaches. Happiness, sorrow, fear, disgust, wrath, and surprise were listed as the six cardinal emotions by Ekman in his seminal work from the 1970s, providing the groundwork for later explorations into the field of emotion detection Ekman (1971). The turn of the millennium saw the emergence of sentiment analysis, with a preference for interpreting the tenor of public opinion enclosed inside textual data, notably drawn from virtual platforms Pang et al. (2008). The frontiers of speech emotion recognition started to advance at the same time. Savants understood that communication has an emotional

undertone in addition to its semantic content. A crucial development in this direction was orchestrated by Schuller et al. (2011), who highlighted the untapped potential of extracting acoustic characteristics from speech data. However, a true revolution in this field was only possible because of the development of deep learning. While effective, the formerly conventional confines of traditional machine learning frequently required complex feature engineering moves. As opposed to this, deep learning architectures, particularly the Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) structures, demonstrated an innate ability to autonomously abstract features, spanning both the textual and vocal domains, making them ideally suited for the task of emotion detection Zhang et al. (2018). This research unfolds within this grand scene at the confluence of these evolutionary leaps, aiming to leverage the dynamic potential of deep learning paradigms in the vocabulary of interpreting emotions, encased both within textual cadences and voice modulations. By doing this, this project aligns with the course of this broad, interdisciplinary journey and adds to the collective exploration's quivering vitality.

1.2 Motivation

Understanding and constructing the complexities present in the range of human emotions is a field of study with roots that date back to the dawn of human civilization. Emotions are persistently regarded as crucial architects of the matrix that determines our conscious decisions, our complex interactions, and the overall picture of our general well-being. They inevitably leave their marks on our mental reflections, orchestrating the rhythm of our social behavior, and even embedding themselves into the structure of our physical reactions. The wise words of Aristotle resonate in the records of human knowledge: "The provocation of ire is facile—yet to channel anger towards the apt individual, calibrated to the precise extent, at the opportune juncture, for the fitting rationale, and orchestrated in a judicious manner—this, indeed, is the arduous endeavour."

The ability to accurately perceive and understand emotions has become increasingly important in the modern digital era, as human interactions are increasingly mediated by electronic intermediaries. Consider the customer service precincts, for instance. An automated system that can recognize the emotional undertone included in a customer's voice or text message can customize its responses, resulting in the emergence of an ecosystem of encounters that are both effective and infused with empathy. In a parallel story, the outlines of mental health are being played out; early emotional disturbance detection through the preservation of digital artifacts can create a channel for the expression of prompt interventions.

Additionally, the expansion of platforms that reflect the social media space has resulted in a precipitation of user-generated content. The ability to assess public mood toward a variety of topics, from the launch of new products to the height of political campaigns, is made possible by this vast pool of information, which is packed with emotive cues. This prospect has never been realized. A thorough integration with this resource offers businesses, decision-makers, and researchers a collection of insights that are unquestionably unmatched in value.

Despite this, this vast array of data reveals the mystery in the powerful combination of amplitude, celerity, and heterogeneity. Even the traditional practitioners of computer approaches run into problems when trying to capture the intricacies which lay within the pathways of human emotions. Manual reading extends into an impracticable domain.

The promise made by the supremacy of deep learning architectures—deft in their ability to navigate through expansive datasets and abstract intricate motifs—reaches its peak of significance at this very moment.

As a result, the research is motivated by a confluence of factors, such as the traditional interest with comprehending emotions, the current issues caused by the digital flood, and the promising potential of today's computational tools. The study is motivated by the vision of a world where technology understands and responds to human emotions rather than commands, encouraging more sympathetic and meaningful interactions between humans and machines.

1.3 Research Question

Many different modalities merge within the broad scope of the emotion detection domain, each bringing a distinctive set of difficulties and opportunities. The realm of words presents us with a patchwork that is distinct from the rhythmic timing and tonal modulations that characterize speech, enriched with its complicated interplay of meanings and syntactic patterns. Both dimensions have emotional indicators that are imprinted on them, but it's important to note that these signs are interconnected and elevated in modality-specific details that change with a noticeable degree of difference. This leaves us with the following insightful question, which serves as the basis for this study:

How does the performance of emotion detection models vary between text and speech data when using the same model architectures?

The basic goal of this research is to uncover the correlation and difference concealed within the modulations of identical model designs as they move across various data modalities. It highlights how important it is to understand not just the concepts themselves but also their inherent features as they penetrate the data they are embedded in. This research mainly covers the use of deep learning to analyze and forecast emotions from textual data and to implement advanced models to determine emotions from speech signals by analyzing the patterns and variances embedded in spoken words. By doing so, this study intends to offer an in-depth analysis of emotion detection using the comparison of these two mediums, highlighting their contrasts and possible overlaps.

The following is an outline of the structure of this report. Section 2 studies a related literature review in this field of study, and Section 3 carries out an acceptable approach. The design specification is covered in Section 4. While the implementation is described in section 5. Section 6 elaborates onthe evaluation of the models and is followed by conclusion and future work.

2 Related Work

This section will look at numerous studies that aid in the process of gaining domain knowledge by understanding and mastering a variety of research methodologies that are offered by researchers

2.1 Emotion Detection in Text

The technique of understanding the emotive rhythm or sentiment encoded within a textual composition falls under the category of emotion detection, which is often referred to

as sentiment analysis or emotion recognition. The rapid expansion of user-generated material, which includes reviews, tweets, and remarks but emerges across the digital tapestry, emphasizes the relevance of this field. This collection is an ideal platform for perspectives and affective expressions. The early excursions into the field of emotion recognition in the context of text were mostly grounded in lexicon-based approaches. These methods relied on precisely compiled lexicon compendia that matched certain lexical entries to specific emotional or attitude registers. It is notable that while "sad," "angry," and "frustrated" grew within the storehouse of negative sensations, words like "happy," "joyful," and "elated" found their roots in the positive sentiment Pang et al. (2002). A change in perspective was sparked by the development of machine learning, with the focus shifting to feature-driven techniques. These approaches relied on the extraction of multidimensional features from the textual corpus, including things like n-grams, part-of-speech tags, and grammatical structures.Then, using Support Vector Machines (SVM) or Naive Bayes classifiers, the collected characteristics were fed into the workflows of traditional machine learning methods Go et al. (2009). The arrival of deep learning paradigms has sparked a renaissance that has been visible in recent years. This paradigm is notable for the rise of neural network architectures, particularly Recurrent Neural Networks (RNNs) and their iterations Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), which have become the foundation of the most advanced framework governing emotion detection within textual frontiers. According to performance benchmarks, these designs clearly outperform conventional techniques Zhou et al. (2014). They also demonstrate an inherent talent for understanding the complex webs of dependency weaved into textual articulations. Word embeddings, exemplified by companies like Word2Vec and GloVe, which enable the encoding of words within the confines of continuous vector spaces, have proven crucial in strengthening the effectiveness of deep learning constructs, adding to this trend. These embeddings open up a window through which the semantic connective tissues connecting words may be seen, giving models the insight to recognize more intricate details and contextual cues with greater precision Mikolov et al. (2013).

Emotion recognition in textual planes has resonances in a wide range of applications in the modern era. Its application resonates in corporate settings in areas like brand monitoring and customer feedback analysis. In the field of healthcare, it takes on the role of a tool for helping to monitor patients' psychological health through the lens of their literary formulations. It also establishes its significance within the parameters of social science research, materializing as an instrument that gauges public opinion on a variety of topics. Additionally, the discipline unfolds toward new thresholds of advancement when contrasted with the ascension of transformer-based paradigms exemplified by enormous structures like BERT and GPT. These paradigms, which are founded on pre-training over enormous datasets, serve as precursors to revised standards for emotion detection projects, highlighting the potential hidden within the transfer learning field Devlin et al. (2018). As per the recent research, Ameer et al. (2023) analyses the usage of LSTM and Transformer Networks via Transfer Learning for multi-label sentiment categorization and offer a novel method that makes advantage of numerous attention mechanisms. They use the Ren-CECps dataset for Chinese and the SemEval-2018 E-c dataset for English to assess their models. On the complex SemEval-2018 E-c database for English, their top-performing RoBERTa-MA (RoBERTa-Multi-attention) model obtained 62.4% accuracy, a 3.6% increase over the state-of-the-art. On the Ren-CECps dataset for Chinese, the XLNet-MA (XLNet-Multi-attention) model fared better than other suggested models, obtaining 45.6% accuracy.

Despite great breakthroughs, there are still a number of problems with emotion detection in text. Language ambiguity can cause misunderstandings since words might mean different things depending on the situation. According to Riloff et al. (2013), current models frequently misclassify the intended emotion when attempting to detect sarcasm and irony. Emotions conveyed in the text can differ between cultures, which can lead to biases in models developed using particular datasets. Deep learning methods are dependent on huge datasets, which limits their application for languages with limited resources Tang et al. (2014).

2.2 Emotion Detection in Speech

Speech emotion recognition (SER), a term used to refer to the field of emotion detection within the arena of vocal expressions, aims to navigate the path of unraveling and categorizing these emotive specifics, supported by the medium of vocal signals. The beginning of SER may be traced to the final decades of the 20th century, which were characterized by developing forays focused on fundamental emotions including happiness, sadness, rage, and terror. Pitch, energy, and formant frequencies, which served as the markers dividing emotional levels in the vocal rhythm, were deeply rooted in the earliest methodology as part of the handcrafted concept Davis and Mermelstein (1980). The preprocessing, feature extraction, and classification domains are what set apart current SER systems from earlier iterations. Preprocessing usually includes noise reduction and normalization as a first step, providing a productive vessel where the vocal foundation can be reduced to its vital elements. Mel-frequency cepstral coefficients (MFCCs), a construct that has demonstrated its effectiveness in capturing the fundamental characteristics of human vocal expressions, are frequently used in feature extraction as the next phase Logan et al. (2005). The deployment of a variety of machine learning frameworks, from Support Vector Machines' vectored stance to Convolutional Neural Networks and Recurrent Neural Networks' (CNNs') sculpted architecture, brings in the conclusion of classification, which constitutes the conclusive levelHan et al. (2014). Krishna et al. (2022) note that speech containing emotions like fear, rage, and joy typically has a more intense and more significant variation in pitch, whereas emotions containing low pitch range are perceived differently. By identifying speech emotions, this effort aims to improve human-machine interactions. The paper uses Support Vector Machine (SVM) and Multi-layer Perception classification techniques to achieve emotion detection. In order to extract pertinent information from the speech data, additional audio features including MFCC (Mel-frequency cepstral coefficients), MEL, chroma, and Tonnetz are used. These models have been taught to identify feelings including tranquillity, neutrality, surprise, joy, sorrow, rage, fear, and disgust. The findings indicate that the suggested method recognizes emotions with an accuracy of 86.5%. The model performs as well when tested with input audio, indicating that it can accurately identify emotions in fresh voice samples. Deep learning networks' capacity for feature extraction has recently assisted them to grow more popular in SER. However, deep learning models tend to have a tendency to overfit speaker-specific features in speaker-independent tasks. Liu et al. (2023) suggests an attention-based bidirectional long short-term memory network (ABLSTM), multi-task learning, and CNN as part of an SER technique to address this. The paper makes several contributions, including the development of a new technique for obtaining time-domain and frequency-domain data from log-Mel spectrograms, investigation of the effects of various additional tasks in multiple-task training for speaker-independent SER,

and assessment of gender-related variations in SER results, with male gender results being superior to female results in terms of recall. Both IEMOCAP and MSP-IMPROV are used for this study. On both databases, the effectiveness of seven auxiliary tasks is evaluated, and the most favorable outcomes have been achieved with 70.27% Weighted Average Recall (WAR) and 66.27% Unweighted Average Recall (UAR) on IEMOCAP and 60.90% WAR and 61.83% UAR on MSP-IMPROV. SER confronts difficulties despite advancements such as individuals can display emotions in a variety of ways, making it challenging to generalize over a wide range of groups Schuller et al. (2015). In various circumstances, the same vocal tone may convey a variety of emotions Batliner et al. (2003). The availability of high-quality, tagged speech emotion datasets is scarce, particularly for certain languages or emotions Zhang et al. (2014). Corresponding to Cowie et al. (2001), emotions might be transient or mixed, making it difficult to discern them. According to Baltrusaitis, Zadeh, Lim and Morency (2018), few studies have looked at how mixing speech with other modalities, such as facial expressions, might help people better identify emotions. Many SER systems are tested in lab settings, which do not accurately represent the noise and variability that exist in the real world. Similar to text, using SER without user permission might provide privacy concerns, particularly when used in public areas or in delicate situations.

2.3 Comparative Studies on Text vs. Speech Emotion Detection

Given the distinct difficulties and benefits each brings, comparing emotion recognition in text and speech modes has been a focus of research Schuller et al. (2011). For a fair comparison, researchers frequently utilize uniform machine learning models for both modalities. The embedding of words for text Mikolov et al. (2013) and Mel-frequency cepstral coefficients (MFCCs) for speech Davis and Mermelstein (1980)are examples of features that are frequently input to models like SVMs or CNNs Kim (2014). The recent development of deep learning within the context of emotion detection is the precursor of a genuine revolution, covering a major break from the domains dictated by conventional aspects of machine learning paradigms LeCun et al. (2015). Deep learning constructs have demonstrated their dominance in this domain by successfully encapsulating the complex emotional variations contained in datasets. This is a result of their ability to independently integrate layered representational systems Schuller et al. (2013). Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have emerged as the neural network architectures that serve as the cornerstone for emotion detection effortsHochreiter and Schmidhuber (1997). Notably, the CNNs, distinguished through their spatial hierarchies, are extremely skilled in processing organized grid data domains, as demonstrated by pictures or spectrograms, making them the esteemed choice in the field of voice emotion recognition Sainath et al. (2013). In addition, the development of transformer architectures, illustrated by iterations like BERT and its numerous children, has resulted in new milestones for emotion recognition based on textual substrates Devlin et al. (2018). These constructs' pre-equipped skills, large amounts of databases, manifests as an entity that can be tuned and refined for the particular purposes of emotion identification. This method makes use of the linguistic wisdom that is embedded in their fundamental architecture Vaswani et al. (2017). Deep learning model strengths are not resistant to difficulties, either. They fall under the authority of ongoing study because of their fondness for large amounts of labeled data, the threat of overfitting lying

in the shadows, and the unresolved issues surrounding the boundaries of interpretability Zhang and Sabuncu (2018). Despite the complex paths these problems take, the picture they provide of the thorough detection and understanding of human emotions places them as indispensable, powerful tools of modern emotional analytics Goodfellow et al. (2016). When textual and audio data are combined, advances in the field of multi-modal emotion recognition paths push boundaries even furtherBaltrušaitis, Ahuja and Morency (2018). The goal of these constructs is to combine the advantages that each of these modalities has over the other, providing a view that emphasizes a more comprehensive understanding of emotional experiences. Adesola et al. (2023) attempts to solve the issue of machine understanding of human affective actions and enhance emotion recognition utilizing multimodal methods for deep learning. In particular, information from both spoken words and text recordings are extracted using three-dimensional deep convolutional neural networks (3D CNNs), allowing the model to successfully recognize emotions across various modalities. The study offers an in-depth dual recurring encoder model that successively employs text information and sound impulses to better understand speech data and categorize it into categories for emotions like happy, sad, neutral, and furious. Dual recurrent neural networks (RNNs) are used in the framework to encrypt data from speech and text patterns, merging it to determine the emotion class. Unlike models that only consider audio features, this design enables the algorithm to evaluate speech signals from the level of the signal all the way down to the language level. The model's performance in classifying statements into one of the four emotion categories was tested using the USC-IEMOCAP dataset, and it showed an accuracy of 68.26% In order to advance the field of emotion detection, this study tries to comprehend how similar model designs perform across both modalities.

3 Methodology

The project's methodology is described in this section. This study employed Knowledge Discovery in Databases (KDD) approach. KDD is a thorough way for drawing out important knowledge and insights from huge databases. As this approach is an iterative process, feedback from a particular stage can affect choices and actions made in a later phase. Due to its iterative nature, outcomes can be improved and refined over time.

3.1 Data Selection and Data Understanding

In the realm of emotion detection, the choice of data is dominant. The data serves as the base upon which models are built and trained, and its quality and relevance directly influence the outcomes of any analysis. For this research, we accurately selected two distinct datasets from Kaggle, representing two modalities: text and speech. This section skim over the specifics of these datasets, their inherent characteristics, and their relevance to our key research question.

3.1.1 Description of Datasets

- Text Dataset - Emotion Detection from Text: This dataset comprises written expressions from various sources, capturing a wide range of emotions. Each entry in the dataset is a textual representation of an emotion, labelled with its corresponding emotional category. The diversity in the dataset, stemming from different

contexts and sources, ensures a comprehensive coverage of textual emotional expressions. It contains around 13 different emotions and over 40000 records. The dataset is merely a list of tweets with annotations describing the feelings they elicited. Tweet_id, emotion, and content all have three columns. The original tweet is found under "content." The sentiment of a tweet captures the sentiment of the tweet. Figure 1 represents the word cloud for anger in the text data.

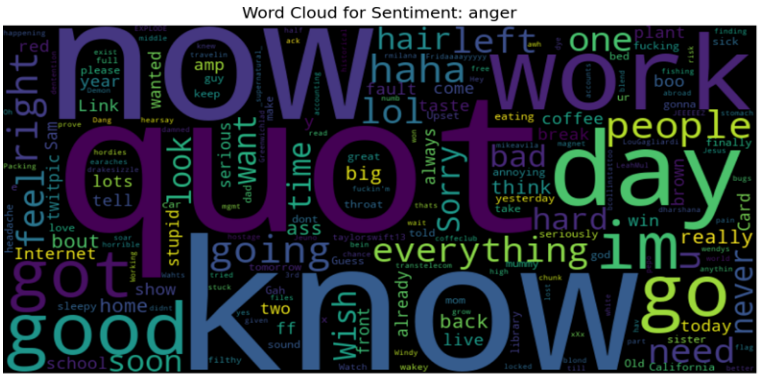

Figure 1: Word Cloud for Anger

- Speech Dataset - RAVDESS Emotional Speech Audio: Unlike the text dataset, this collection is a combination of audio recordings from different speakers, each conveying a specific emotion. The dataset captures the traces of vocal intonations, pitch variations, and rhythm, which are pivotal in detecting emotions in speech. The RAVDESS features 24 trained actors—12 male and 12 female—who each deliver two lexically related lines with a neutral North American accent. Speech expressions might be peaceful, joyful, sad, furious, afraid, surprised, or disgusted. Each expression is produced in two emotion levels of intensity (normal and strong) and an expression that is neutral. It contains around 1440 files where there is 60 trials per actor. Figure 2 represents distribution of various emotions in the speech dataset.

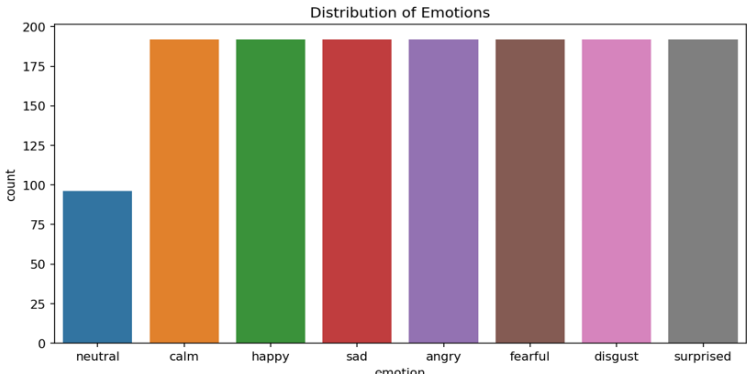

Figure 2: Emotion distribution in speech

The choice of these datasets was driven by an authentic need to explore the complex processes underlying emotion recognition across multiple modalities. The vocal rhythm encoded inside speech emerges as a background onto which complex vocal emotions are

stitched, with textual structure developing a depth in contextual understanding. This research unfolds within a combination of various modalities, with the ultimate goal of solving the complexities and fluctuations on each domain's involvement within the field of emotion detection.

3.2 Data Processing

Data preprocessing is a foundational step, ensuring data is ready for modelling. Given our two datasets - textual and auditory - the preprocessing varied accordingly.

3.2.1 Text Data

As a part of cleaning the textual data which contains around 40000 records. Several punctuation mistakes, special characters, URLs, and HTML tags were removed. The sentences are broken into words or tokens by the process of tokenization as this process converts unstructured information into a format that allows for advanced text analysis, feature extraction, and data analysis. Certain words like 'and', 'the', and 'is' were eliminated to reduce the noise that frequently occurring, uninformative words produce in text analysis, hence increasing text analysis's effectiveness and efficiency. Inorder to normalize words to make them more analytically logical and to make the vocabulary less dimensional the process of stemming was done in which words were reduced to their base form. For the machines to understand the data better the texts were converted into numerical vectors using TF-IDF and word embeddings. This process is known as vectorization, which is an essential step in preparing text input for machine learning algorithms because the majority of these algorithms operate on numerical data.

Figure 3 represents the pre-processed text data.

	content	sentiment	cleaned_content
0	@tiffanylue i know i was listenin to bad habi...	empty	know listenin bad habit earlier start freakin ...
1	Layin n bed with a headache ughhhh...waitin o...	sadness	layin n bed headach ughhhhwaitin call
2	Funeral ceremony...gloomy friday...	sadness	funer ceremonygloomi friday
3	wants to hang out with friends SOON!	enthusiasm	want hang friend soon
4	@dannycastillo We want to trade with someone w...	neutral	want trade someon houston ticket one
...	...	...	...
39995	@JohnLloydTaylor	neutral	
39996	Happy Mothers Day All my love	love	happi mother day love
39997	Happy Mother's Day to all the mommies out ther...	love	happi mother day mommi woman man long your mom...
39998	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...	happiness	wassup beauti follow peep new hit singl def wa...
39999	@mopedronin bullet train from tokyo the gf ...	love	bullet train tokyo gf visit japan sinc thursda...

40000 rows × 3 columns

Figure 3: Pre-processed Text data

3.2.2 Speech Data

The Speech dataset contained background noises which were removed as a part of the noise reduction technique in order to improve the listening experience. In speech processing, feature extraction involves transforming unprocessed audio signals into a collection of relevant and informative features that may be applied to a variety of tasks, including speech recognition, emotion analysis, speaker identification, and more. Although speech signals are complex and packed with data, their raw form is frequently too

detailed and high-dimensional for direct analysis. In order to reduce the dimensionality and to capture the necessary characteristics from the speech dataset, MFCCs, Chroma and Mel techniques were used to extract the features from this dataset. The consistency and quality of the speech data can be improved by transforming the audio signals to a normal scale. This process of normalisation is applied in order to ensure that the features lie within a similar range. The longer audio files were divided into smaller chunks as a part of segmentation. In essence, these steps ensured both text and speech data were optimised for the modelling phase.

3.3 Data Transformation

Data transformation is the process of converting data into a format or structure that's more suitable for analysis or modelling. Given the distinct nature of the datasets, the transformation process was tailored to each.

3.3.1 Text Data

The tokenized words in the dataset were transformed into dense vectors that capture semantic meanings by utilizing pre-trained embeddings like Word2Vec and GloVe. These embeddings represent the semantic connections between words in a continuous, dispersed space. They increase model performance and produce more accurate results by allowing models to comprehend and interact with the text's fundamental interpretations. To ensure uniform input length for our models, sequences were padded or truncated to a fixed length. The process of uniformly lengthening variable-length sequences by adding more elements (often zeros) is known as sequence padding. Using One-Hot encoding the emotion labels were converted into a binary matrix representation which is suitable for classification tasks.

3.3.2 Speech Data

The structure of the audio data were reshaped without affecting the content. Inorder to match the specific algorithms that is used, reshaping is an essential factor. Audio features, especially when extracted as MFCCs or Chroma, were reshaped to fit the input requirements of deep learning models. The magnitudes of the audio signals were found varying after the feature extraction. So to alter the audio signals' loudness or dynamic range to a desirable level scaling was initialized. Min-Max scaling was used to ensure that the signals lie within a 0-1 range. For models like CNNs, spectrograms were treated as images, allowing the model to capture patterns in frequency changes over time. The transformation phase was pivotal, ensuring that the data was not only clean but also in the right shape and form to be trained by the chosen model architectures.

3.4 Model Building

The research utilized identical model architectures for both text and speech data in effort to comprehend the intricacies of emotion identification across various data modalities. By using this method, it was guaranteed that any performance variations were due to the data modality and not the model's construction.

3.4.1 Ensemble model

An ensemble model is created by combining three models namely, Logistic Regression, SVM, and random forest. To create more effective and robust model, individual models are combined together. Votingclassifier is used to combine all these three models. SVM and Random Forest's sophisticated modelling can be complemented by the clarity and transparency of Logistic Regression. SVM may offset the linearity of Logistic Regression by handling high-dimensional data and capturing complex connections while overfitting in individual models can be prevented due to Random Forest's ensemble nature and robustness.

3.4.2 Base LSTM Model:

The foundational model in this study was a simple Long Short-Term Memory (LSTM) network. LSTMs, with their inherent capability to remember long-term dependencies, are aptly suited for sequential data, be it text or audio sequences. By overcoming the vanishing gradient issue of conventional RNNs, LSTM is a kind of recurrent neural network (RNN) that is intended to detect dependence over time in sequential data.

Figure 4 represents architecture of Base-LSTM.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 32)          786208

 lstm (LSTM)                 (None, 32)                8320

 dense (Dense)               (None, 13)                429

=================================================================
Total params: 794,957
Trainable params: 794,957
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: Base LSTM architecture

3.4.3 Tuned Bi-LSTM Model:

Building upon the base LSTM, this model introduced additional layers and optimised hyperparameters to capture more intricate patterns in the data. It was designed to strike a balance between complexity and computational efficiency. A neural network architecture that makes use of Bidirectional LSTM layers that have been specially tailored or optimized for a given task or dataset is referred to as a Tuned Bi-LSTM (Bidirectional Long Short-Term Memory) model. This will help to Enhance the model's ability to recognize patterns, dependencies, or correlations in the data will improve its performance in terms of prediction or classification for the given task. Figure 5 represents architecture of Tuned Bi-LSTM.

3.4.4 Hybrid Model:

This innovative model combined the strengths of Convolutional Neural Networks (CNNs) and LSTMs. The CNN layers, known for their prowess in detecting local patterns, were

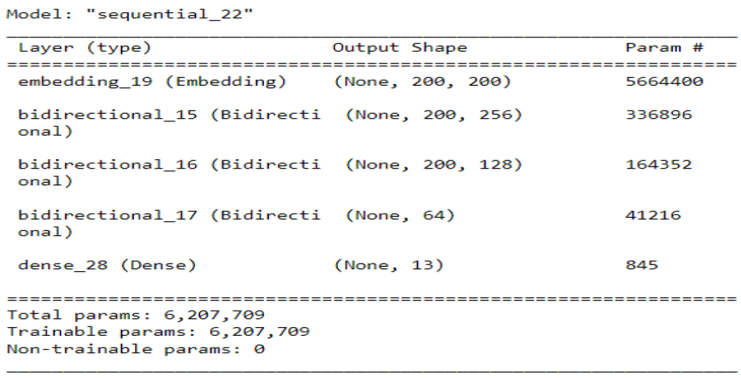

Figure 5: Bi-LSTM architecture

employed to discern nuances in the data. These detected patterns were then processed by an LSTM layer, ensuring the model's capability to understand the sequential nature of the data. Figure 6 represents architecture of Hybrid model.

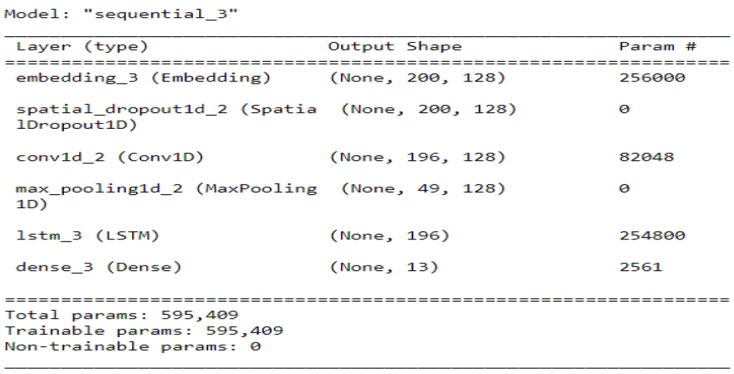

Figure 6: Hybrid model architecture

By maintaining consistent architectures across both datasets, this study aimed to provide a fair and unbiased comparison, focusing purely on the inherent characteristics of text and speech in emotion detection.

3.5 Evaluation

Evaluating the performance of the emotion detection models was an integral part of this research. It allowed us to gauge the effectiveness of the chosen architectures and understand how they fared across text and speech modalities. Accuracy is the initial matrix used for evaluation. This primary metric provided a direct measure of how often our model's predictions aligned with the actual labels. Given the balanced nature of our datasets, accuracy was a reliable indicator of our model's general performance. Figure 7 represents the equation for Accuracy.

The loss function, specifically categorical cross entropy in this study, quantified how well the predicted probability distribution matched the true distribution of the labels. A lower loss indicated better performance, with the model's predictions being closer to the actual labels. Figure 8 represents the equation for Loss.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Figure 7: Accuracy

$$Loss = - \sum_{i=1}^{output\ size} \square\ Actual\ label\ .log\ Predicted\ label$$

Figure 8: Loss

3.6 Model Performance

Each model's accuracy and loss were monitored during training and validation phases. This allowed us to observe the convergence patterns and ensure that the models were learning effectively without overfitting. Vital factor to this research was the comparison of model performances across the two modalities: text and speech. By analysing the accuracy and loss metrics side-by-side for each dataset, the nuances and challenges each modality presented in emotion detection was recognized. In summary, the evaluation was streamlined and focused, emphasizing the direct comparison of model performances on text versus speech data, shedding light on the capabilities and potential areas of improvement for each architecture in the context of emotion detection.

4 Design Specification

This section explains the technical design and processes taken to implement the emotion detecting system. The system's components, design decisions, and implementation quirks are all covered in detail in this detailed overview.

The emotion detection system is adept at processing both text and speech data. It's designed with the primary goal of classifying emotions using specialised deep learning models. The Data ingestion module loads datasets and handles initial data checks. Two datasets were gathered from a public source. The data collected is then preprocessed which refines raw data into a model-friendly format, including text tokenization and speech feature extraction. Then LSTM, Bi-LSTM and a hybrid model combining both the CNN and LSTM are executed in the modelling module including ensemble model using machine learning algorithms. After training the models with the specified algorithms the performance of the model is being evaluated.

5 Implementation

For implementing this research, Python 3.8 was chosen as given its extensive libraries and community support. The key libraries used for this study was Keras (with TensorFlow backend), Scikit-learn, and Librosa. As a part of Data handling procedure, after tokenization, text data was transformed into sequences suitable for LSTM processing. While for speech data, features like MFCCs were extracted to serve as the models input. A straightforward LSTM setup ideal for sequence data was developed for the base LSTM model. An advanced version of the base model is then implemented which was optimised for better performance. The hybrid model ensures the strengths of CNNs and LSTMs, aiming to capture both spatial and sequential patterns in the data. Utilizing the beneficial qualities of both architectures to enhance performance across a variety of natural language processing and speech-related tasks, a hybrid CNN-LSTM model can offer improved capabilities for processing and analyzing text and speech input. Datasets' training subset are used to train these models.Accuracy and loss metrics were evaluated on the validation subset. The design emphasises modularity, ensuring that each component can be updated or scaled independently. The choice of models was influenced by the nature of the data sequence and the proven efficacy of LSTMs in handling such data.A few challenges surfaced during implementation: Some emotions had fewer data points. The data imbalance was dealt with data augmentation and oversampling techniques. Dropout layers were introduced to prevent the models from memorizing the training data. As deep models demand resources, efficient coding practies were adapted to manage this. In essence, this section paints a detailed yet concise picture of the system's architecture, its design rationale, and the journey from idea to implementation.

5.1 Ensemble method

Necessary libraries have been imported as the initial stage of implementation. Both the data are divided into train and test sets.

5.1.1 Text data

For text data, the data is converted in TF-IDF vectors using TfidfVectorizer function. Three base models are defined: LogisticRegression, SVC (Support Vector Classifier), and RandomForestClassifier. The VotingClassifier class is used to build an ensemble model. Each models are specified with unique names('lr' stands for logistic regression,'svc' for SVM, and 'rf' for random forest, respectively). Class probabilities are used for voting because the voting parameter is set to "soft."The fit approach is used to train the ensemble model on the TF-IDF transformed training data.The forecast method of the ensemble model is used to make predictions on the TF-IDF transformed test data. A thorough classification report that includes precision, recall, F1-score, and support for each class is printed using sklearn.metrics function called classification_report.

5.1.2 Speech data

Features are divided into X and labels into Y, assuming that df is a DataFrame. The LabelEncoder class, which transforms class labels into numerical values, is used to encode the labels. The modelling is done same as that of the text data.To appropriately label the report, the target_names option is set to the class names obtained from the encoder.

5.2 Base LSTM

For the purpose of creating and training the LSTM model, necessary Keras modules and classes are loaded.

5.2.1 Text data

The text data is tokenized using the Tokenizer class. The tokenizer is fitted to the training set using fit_on_texts. The tokenized texts are transformed into sequences of integer values using the texts_to_sequences function.The sequences are padded to have a constant length (in this instance, 200) using the pad_sequences function.The pd.get_dummies function is used to one-hot encode the target labels (y_train and y_test). To create the model,The length of the word index produced by the tokenizer is used to determine the vocabulary size.To handle the tokenized input data, an embedding layer is implemented and to process the sequential data, an LSTM layer is applied. While creating class probabilities, a dense layer with softmax activation is added. The accuracy metric, categorical cross-entropy loss, and Adam optimizer are all used in the model's construction.The fit approach is used to train the model on the padded training data. The padded test data is set as the validation data parameter for training.

5.2.2 Speech data

The DataFrame df has been divided into features and labels. The LabelEncoder class is used to encode the target labels (y), which are then one-hot encoded using the to_categorical function after being converted from class labels to numerical values.The train_test_split function is used to divide the data into training and test sets.The X_train and X_test features are reshaped to take on the sample, timestep, and feature shapes needed for LSTM input.Using Keras, a sequential model is produced. A Dense layer with a softmax activation function is added to provide class probabilities, and an LSTM layer with 32 units is added as the input layer to determine the input shape.The accuracy metric, the Adam optimizer, and the categorical cross-entropy loss function are used to build the model.On the altered training data, the model is trained using the fit approach.The batch size is set to 32, and 100 epochs of training are completed. For validation purposes during training, the validation data parameter is set to the reshaped test data.

5.3 Tuned LSTM

5.3.1 Text Data

The Tokenizer class is used to tokenize text data, and sequences are created as a result. To ensure constant sequence lengths, sequences are padded using pad_sequences.Based on the word index produced by the tokenizer, the vocabulary size is determined.In order to handle tokenized input data, an embedding layer is implemented. Different configurations of multiple bidirectional LSTM layers are added. Class probabilities are created by adding a dense layer with softmax activation.Accuracy metric, the Adam optimizer, and categorical cross-entropy loss are used in the model's construction.The fit approach is used to train the model on the training set of data.To keep track of the model's performance on the test set, validation data is offered.

5.3.2 Speech data

The DataFrame df has been divided into features and labels.The LabelEncoder is used to encrypt labels. Using to_categorical, the encoded labels are one-hot encoded once again.The X_train and X_test features are molded into a three-dimensional form with the samples, timesteps, and features needed for LSTM input. Using Keras, a sequential model is produced. Different configurations of multiple bidirectional LSTM layers are added. Class probabilities are created by adding a dense layer with softmax activation are used to create the bidirectional LSTM model.The training is carried out for 50 epochs with a batch_size of 32. For validation during training, the reshaped test data is set as the validation_data parameter.

5.4 Hybrid Model

The Hybrid model is created by the combination of CNN and LSTM.

5.4.1 Text Data

The text data is sequenced and padded. Keras is used to build a sequential model. The handling of tokenized input data is added to an embedding layer. By removing complete 1D feature maps, a SpatialDropout1D layer helps avoid overfitting. To detect regional text patterns, a Conv1D layer with maximum pooling is applied. Sequential data in the text data is captured by an LSTM layer. Probabilities for each class are produced via a dense layer with softmax activation. While training the model, Training is carried out for 30 epochs with a batch_size of 32.

5.4.2 Speech data

Keras is also used for the speech data. Conv1D layer is added to the input data to help identify regional patterns. The spatial dimensions of the feature maps are condensed using a MaxPooling1D layer. An LSTM layer extracts the data's sequential information. Probabilities for each class are produced via a dense layer with softmax activation. The rest of the process is same as the text data.

6 Evaluation

The key findings that support the study question are outlined in this section. The suitability of the SER and TER models for usage in real-time applications has been evaluated through a number of experiments. The proposed models are trained on both the text and speech datasets.

6.1 Experiment 1: Ensemble Methods with Voting Classifier for Text Data

- Key Findings:
  - Overall Accuracy: 35%
- Insights:

  - The low accuracy suggests that ensemble methods may not be the best fit for text-based emotion detection in our dataset.
  - The ensemble approach, while robust in theory, failed to capture the complexities of emotional nuances in text.

- Implications:
  - The results indicate a need for more complex models capable of understanding the sequential nature of text.

Within the initial experiment, the study embarked upon an expedition into the realm inhabited by ensemble techniques, cast in the spotlight of the textual data. The ensemble paradigm mirrors the act of soliciting diverse viewpoints to blend into a more sensible outcome. Within this framework, the study harnessed the ability of three individuated constructs of machine learning: the Logistic Regression, the Support Vector Machine, and the Random Forest Classifier. Each of these constructs, weaving their individual tapestries of strengths, conveys distinct views upon the data under scrutiny.The study employed the assistance of the Voting Classifier in order to create a harmonious fusion of these diverse viewpoints and to identify the trajectory holding the most promise.This classifier, in essence, navigates within the sphere of a majority vote, concluding in the selection of the model that gathers the most categorical "votes" of trust and confidence from the ensemble. The research findings from this exploration bring together an extensive variety of understandings about the model's performance vector across a wide range of emotional categories: Figure 9 represents the ensemble approach for text data.

	precision	recall	f1-score	support
anger	0.00	0.00	0.00	19
boredom	0.00	0.00	0.00	31
empty	0.33	0.01	0.01	162
enthusiasm	0.00	0.00	0.00	163
fun	0.11	0.01	0.03	338
happiness	0.33	0.35	0.34	1028
hate	0.43	0.21	0.28	268
love	0.48	0.40	0.43	762
neutral	0.34	0.58	0.43	1740
relief	0.37	0.03	0.06	352
sadness	0.40	0.25	0.31	1046
surprise	0.32	0.04	0.07	425
worry	0.33	0.49	0.40	1666
accuracy			0.35	8000
macro avg	0.27	0.18	0.18	8000
weighted avg	0.34	0.35	0.32	8000

Figure 9: Ensemble model summary for text data

The overall accuracy of the model stood at 0.35. While these results provided valuable insights, they also highlighted areas for improvement, setting the stage for subsequent experiments.

6.2 Experiment 2: Base LSTM for Text Data

- Key Findings:
  - Overall Accuracy: 78.24%
  - Average Loss: 0.6708

- Insights:
  - The high accuracy indicates that LSTMs are well-suited for handling the sequential nature of text.
  - The low loss further confirms the model's efficacy.
- Implications:
  - LSTMs could serve as a strong baseline model for text-based emotion detection.

In this experiment, the study utilised Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, to detect emotions in text data. LSTMs are adept at processing sequences, making them suitable for text, which is inherently sequential. Using a standard LSTM configuration, the study aimed to set a performance benchmark for our dataset. The results were: Figure 10 represents the accuracy and avg. loss of base LSTM approach for text data.

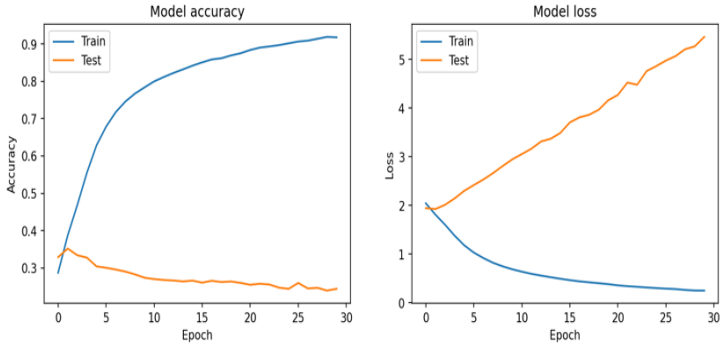

Figure 10: Accuracy and Loss Plot of Base LSTM.

- Average Accuracy: The base LSTM model achieved an accuracy of roughly 78.24%, marking a notable improvement from the initial ensemble approach.
- Average Loss: The model reported a loss of 0.6708, indicating the difference between the model's predictions and the actual outcomes.

The outcomes highlighted the efficacy of LSTMs for the emotion detection task, suggesting that deep learning could offer enhanced performance for the current dataset.

6.3 Experiment 3: Tuned LSTM for Text Data

- Key Findings:
  - Overall Accuracy: 52.44%
  - Average Loss: 1.4420
- Insights:

  - Despite tuning, the model performed worse than the base LSTM, suggesting overfitting.
  - The high loss indicates that the model's predictions were often off the mark.

- Implications:
  - Overfitting remains a challenge, emphasising the need for careful hyperparameter tuning.

In this experiment, the study further explored LSTMs by using a "tuned" model, optimised with specific adjustments to its architecture and hyperparameters tailored for the dataset. The goal was to determine if refined configurations could enhance performance compared to the base LSTM. The results were: Figure 11 represents the accuracy and avg. loss of tuned LSTM approach for text data.

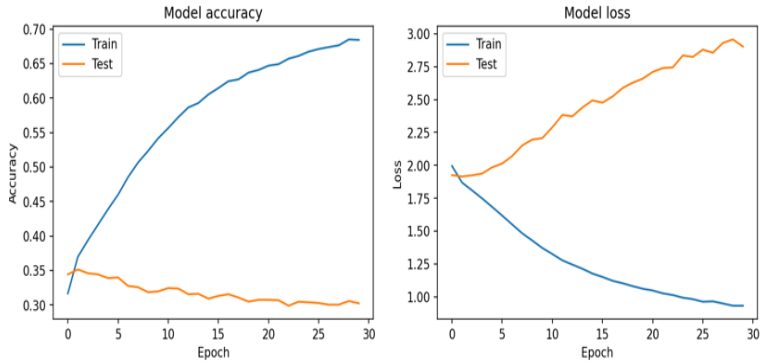

Figure 11: Accuracy and Loss Plot of Tuned LSTM.

- Average Accuracy: The tuned LSTM reported an accuracy of about 52.44%. Surprisingly, this was lower than the base model, suggesting potential overfitting where the model might be too adapted to the training data, reducing its effectiveness on new data.
- Average Loss: The model's loss stood at 1.4420, higher than the base model, further hinting at possible overfitting or sub-optimal tuning choices.

This experiment underscored the complexities of model tuning. A more intricate model doesn't guarantee better results, emphasising the need for meticulous hyperparameter selection and the iterative nature of machine learning experimentation.

6.4 Experiment 4: Hybrid Model for Text Data

- Key Findings:
  - Overall Accuracy: 57.26%
  - Average Loss: 1.2702
- Insights:

  - The hybrid model offers a balanced performance, but still lags behind the base LSTM.
  - The model benefits from the spatial feature extraction of CNNs and the sequence understanding of LSTMs.

- Implications:
  - Hybrid models could be a promising avenue for further research.

In the fourth experiment, we introduced a hybrid model, combining the strengths of convolutional neural networks (CNNs) and LSTMs, aiming to harness both their capabilities for emotion detection in text data. The hybrid model is designed to capture spatial patterns using CNN layers and temporal sequences using LSTM layers, offering a comprehensive approach to text analysis. The key outcomes: Figure 12 represents the accuracy and avg. loss of hybrid model approach for text data.

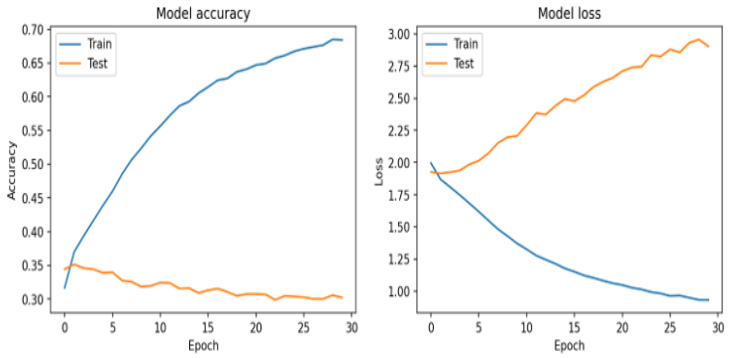

Figure 12: Accuracy and Loss Plot of Hybrid model.

- Average Accuracy: The hybrid model yielded an accuracy of approximately 57.26%. This result indicates a middle ground between the base LSTM and the tuned LSTM, suggesting that the combination of CNN and LSTM layers can provide a balanced performance.
- Average Loss: The model reported a loss of 1.2702. While this is an improvement from the tuned LSTM, it's still higher than the base model, indicating areas for potential optimization.

This experiment illuminated the potential of hybrid architectures in emotion detection. By merging the spatial feature extraction of CNNs with the sequence understanding of LSTMs, the hybrid model showcased a promising avenue for further research and model development.

6.5 Experiment 5: Ensemble Methods with Voting Classifier for Speech Data

- Key Findings:

  - Overall Accuracy: 51%
  - The model combined predictions from Logistic Regression, Support Vector Machine, and Random Forest Classifier.

- Insights:
  - The 51% accuracy suggests that ensemble methods are more effective for speech data compared to text data in our experiments.
  - The use of multiple models like Logistic Regression, SVM, and Random Forest indicates that a collective approach may capture the nuances of speech data better than individual models.

- Implications:
  - The relatively higher accuracy for speech data implies that ensemble methods could be a viable approach for emotion detection in auditory signals.

In the fifth experiment, we shifted the focus from text to speech data, employing ensemble methods combined with a voting classifier. The ensemble approach amalgamates predictions from multiple models, in this case, Logistic Regression, Support Vector Machine, and Random Forest Classifier. The voting classifier then selects the best prediction, ensuring a more robust and reliable outcome. The results from this experiment provided a comprehensive view of the model's performance across various emotion categories: Figure 13 represents the summary of Ensemble model for speech.

	precision	recall	f1-score	support
angry	0.66	0.69	0.67	42
calm	0.45	0.77	0.57	44
disgust	0.39	0.59	0.47	32
fearful	0.65	0.62	0.63	32
happy	0.38	0.32	0.35	34
neutral	0.44	0.20	0.28	20
sad	0.48	0.28	0.35	39
surprised	0.68	0.42	0.52	45
accuracy			0.51	288
macro avg	0.52	0.49	0.48	288
weighted avg	0.53	0.51	0.50	288

Figure 13: Summary of ensemble model for speech

This experiment underscored the complexities of emotion detection in speech data. While ensemble methods offer a holistic approach by combining multiple models, the unique characteristics of speech, such as tone, pitch, and rhythm, present distinct challenges. The results provide a foundation for further refinement and optimization in subsequent experiments.

6.6 Experiment 6: Base LSTM for Speech Data

- Key Findings:
  - Overall Accuracy: 40.48%

  - Average Loss: 1.6738

- Insights:
  - The moderate accuracy suggests that while LSTMs are powerful, they struggle with the complexities of speech data.
  - The high loss indicates that the model often misclassified the emotions in the speech data.

- Implications:
  - LSTMs may require further tuning or a different architecture altogether for speech data.

Transitioning from ensemble methods, the sixth experiment delved into the application of a base LSTM model on speech data. As with the text data, the "base" terminology indicates that we employed a standard LSTM configuration without specific tuning or modifications. The goal was to gauge the inherent capabilities of LSTM networks in capturing the sequential nuances present in speech data. Key Outcomes: Figure 14 represents the accuracy and avg. loss of Base LSTM model approach for speech data.

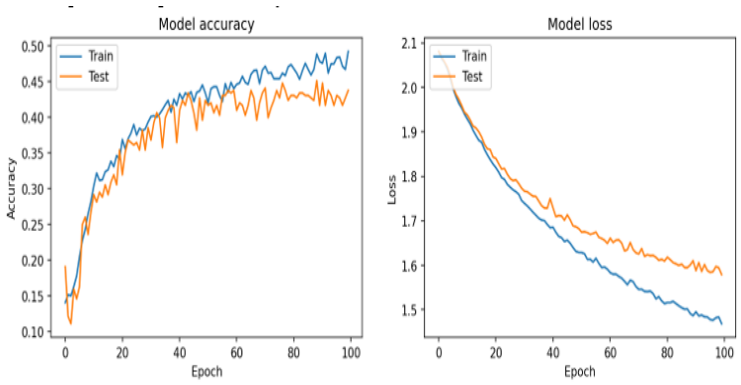

Figure 14: Accuracy and Loss Plot of Base LSTM model for speech.

- Average Accuracy: The base LSTM model for speech data achieved an average accuracy of approximately 40.48%. While this is a moderate figure, it's noteworthy that speech data, with its intricate patterns of pitch, rhythm, and tone, can be inherently more challenging than text data.

- Average Loss: The model registered an average loss of 1.6738. This value, being relatively high, suggests that the model's predictions often deviated from the actual outcomes. It underscores the complexities of emotion detection in speech and the challenges LSTMs face in this domain.

This experiment highlighted the contrast between text and speech data when processed using the same model architecture. While LSTMs have shown promise in text-based emotion detection, their performance on speech data, at least in the base configuration, indicates the need for further refinements. The results set the stage for subsequent experiments that aim to optimise LSTM configurations specifically for speech data.

6.7 Experiment 7: Tuned LSTM for Speech Data

- Key Findings:
  - Overall Accuracy: 46.61%
  - Average Loss: 1.4224
- Insights:
  - The tuned LSTM showed improvement over the base model, but still not to a satisfactory level.
  - The model still struggles with the unique challenges presented by speech data, such as tone and pitch.
- Implications:
  - Further research is needed to optimise LSTMs for speech data

Building on the foundational insights from the base LSTM model applied to speech data, our seventh experiment focused on a tuned LSTM configuration. The "tuned" descriptor signifies that this model underwent meticulous adjustments, both in its architecture and hyperparameters, tailored to optimize performance on the speech dataset. Key Outcomes: Figure 15 represents the accuracy and avg. loss of Tuned LSTM model approach for speech data.

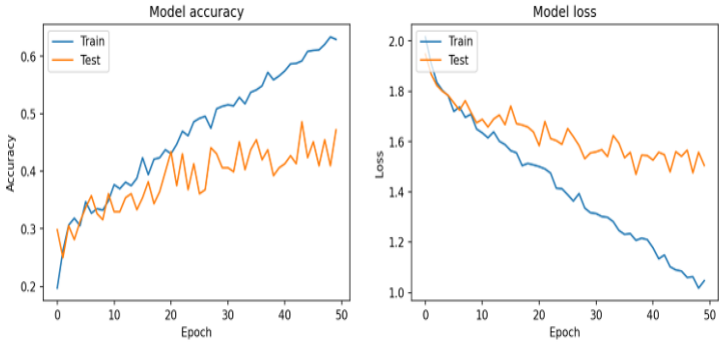

Figure 15: Accuracy and Loss Plot of Tuned LSTM model for speech.

- Average Accuracy: The tuned LSTM model for speech data achieved an average accuracy of approximately 46.61%. This marked an improvement from the base LSTM model, indicating that the tuning efforts bore fruit to some extent. However, the accuracy still suggests room for further optimization.
- Average Loss: The model reported an average loss of 1.4224. While this is a reduction from the base LSTM's loss, it still points to discrepancies between the model's predictions and the actual outcomes, emphasizing the intricate challenges of speech-based emotion detection.

This experiment underscored the potential and challenges of tuning LSTM models for speech data. While the refined model showcased better performance metrics than its base counterpart, the results also highlighted the complexities inherent to speech data. Emotion detection in speech, with its multifaceted rhythm and acoustic features, demands a delicate balance of model complexity and adaptability. The outcomes from this experiment serve as a stepping stone for further explorations and optimizations in the domain of speech-based emotion detection.

6.8 Experiment 8: Hybrid Model for Speech Data

- Key Findings:
  - Overall Accuracy: 63.59%
  - Average Loss: 1.0027
- Insights:
  - The hybrid model significantly outperformed both the base and tuned LSTMs, indicating the benefits of a multi-architecture approach.
  - The lower loss suggests that the model was more consistent in its predictions compared to earlier models.
- Implications:
  - Hybrid models combining CNNs and LSTMs could be the future of emotion detection in speech data.

The chronicle of the eighth and culminating experiment charts an excursion into the realm of methodological synthesis, unfolding a hybrid construct skillfully calibrated to thread the varied of vocal data. This illusion of a model stands as a living testament to a combination that twists the virtues of convolutional neural networks (CNNs) and the simplicity of Long Short-Term Memory (LSTM) networks, all coordinated within the field of vocal data. The basic goal of this work is to capture the spatial components entangled in the embrace of CNNs and the temporal linked essence intrinsic to LSTMs. As a result of this study, the following consequent insights reflected via a lens of significant results that mark this journey engrave themselves in the records of understanding: Figure 16 represents the accuracy and avg. loss of Hybrid model approach for speech data.

- Average Accuracy: The hybrid model showcased an average accuracy of approximately 63.59%. This result is notably higher than both the base and tuned LSTM models for speech data, underscoring the potential of combining different neural network architectures to tackle the intricacies of speech-based emotion detection.
- Average Loss: The model reported an average loss of 1.0027. This figure, while still indicative of some prediction discrepancies, is an improvement over the losses reported in previous experiments with speech data.

The hybrid model's proficiency performance serves as an impressive proof of the many benefits of an approach characterized by multiple dimensions, which is especially important when tackling the field of complex datasets like speech. Convolutional neural networks

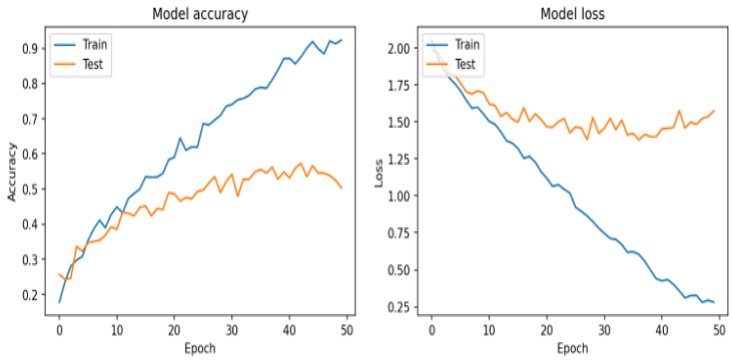

Figure 16: Accuracy and Loss Plot of Hybrid model for speech.

(CNNs) and long short-term memory (LSTM) networks' strengths merged into a symphonic convergence , allowing the model to unravel the intricate pattern of complicated features enclosed within the span of vocal data with ease. The arrangement that resulted from this carried out with greater authenticity; which increased the effort of emotion detection.

6.9 Discussion

The discovery contained in the overarching goal of this research work lay in the pursuit of understanding the variation in performance shown by emotion detection models when expressed within the domains of textual and vocal data, even when enclosed by the means of identical model architectures. This project developed in a meticulously constructed patchwork of assessments, each carrying a mark of preparation. These conclusions, drawn from the initial phase of the observed results, are the end result of these findings, which were forged from the test of direct research.

6.9.1 Ensemble Methods with Voting Classifier:

- Text Data: The ensemble approach, which combined the strengths of Logistic Regression, Support Vector Machine, and Random Forest Classifier, yielded an accuracy of 35% for text data. This suggests that while ensemble methods can be powerful, they might not be the most optimal choice for text-based emotion detection in our dataset.

- Speech Data: The accuracy for speech data was 51%, a noticeable improvement over text data. This indicates that ensemble methods might be more adept at capturing the nuances of speech data compared to text.

6.9.2 Base LSTM:

- Text Data: The base LSTM model for text data achieved an impressive accuracy of 78.24%, highlighting the potential of deep learning, especially LSTMs, in handling sequential data like text.

- Speech Data: The accuracy dropped to 40.48% for speech data, suggesting that the base LSTM configuration might not be as effective for speech as it is for text.

6.9.3 Tuned LSTM:

- Text Data: The tuned LSTM model for text data reported an accuracy of 52.44%. Interestingly, despite the tuning, the performance was lower than the base LSTM, hinting at potential overfitting
- Speech Data: The accuracy improved to 46.61% for speech data, but it still lagged behind the text data performance.

6.9.4 Hybrid Model:

- Text Data: The hybrid model, which combined CNNs and LSTMs, achieved an accuracy of 57.26% for text data. This suggests that integrating different neural network architectures can enhance emotion detection capabilities.
- Speech Data: The performance soared to 63.59% for speech data, making the hybrid model the best-performing model for speech in our experiments.

6.9.5 Comparative Analysis:

- Text vs. Speech: While the base LSTM model performed exceptionally well for text data, the hybrid model emerged as the most effective for speech data. This underscores the idea that while certain architectures might be universally robust, the nature of the data can significantly influence model performance.
- Overall Performance: In general, the models exhibited varied performance across text and speech data. However, it's noteworthy that the same architectures yielded different results based on the modality of the data. This highlights the inherent complexities and unique characteristics of each data type.

7 Conclusion and Future Work

Throughout the research journey, we delved deep into the realm of emotion detection, exploring the intricacies of both text and speech data. The experiments, spanning across various model architectures, provided valuable insights into the performance dynamics of emotion detection models across different data modalities. One of the key takeaways from the study is the profound influence of data type on model performance. While certain architectures, like the base LSTM, showcased exceptional ability with text data, others, like the hybrid model, emerged as frontrunners for speech data. The research also underscored the importance of iterative experimentation. As observed, even refined models like the tuned LSTM can sometimes underperform compared to their base counterparts, emphasising the challenges associated with model tuning and the potential pitfalls of overfitting. Building upon the foundation laid by the current research, there's a plethora of avenues to explore in the future. Centred around the research question, "How can we enhance emotion recognition accuracy by leveraging multi-modal data (speech and text) and advanced machine learning models?", here are some potential directions:

- Multi-modal Data Integration: Combining the strengths of both text and speech data can potentially lead to more robust emotion detection models. Future research can focus on developing models that can seamlessly integrate features from both

data types, capitalising on the contextual depth of text and the prosodic nuances of speech.

- Advanced Model Architectures: With the rapid advancements in machine learning, newer architectures like Transformers and BERT have shown promise in various tasks. Exploring these architectures for emotion detection can potentially lead to significant improvements in accuracy.
- Transfer Learning: Leveraging pre-trained models and fine-tuning them for emotion detection can expedite the training process and potentially enhance performance. Models like GPT-3 or BERT, which have been trained on vast datasets, can be fine-tuned for our specific task.
- Data Augmentation: To combat issues like overfitting and to enhance the diversity of our training data, future work can explore advanced data augmentation techniques, both for text and speech.

In conclusion, while the current research has paved the way for a deeper understanding of emotion detection across text and speech data, the road ahead is rife with opportunities. The confluence of multi-modal data and advanced machine learning models promises a future where emotion detection models are not only accurate but also intuitive and adaptable.

References

Adesola, F., Adeyinka, O., Kayode, A. and Ayodele, A. (2023). Implementation of multimodal speech emotion recognition using text data and audio signals, *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, Vol. 1, IEEE, pp. 1–8.

Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G. and Gelbukh, A. (2023). Multi-label emotion classification in texts using transfer learning, *Expert Systems with Applications* **213**: 118534.

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

Baltrusaitis, T., Zadeh, A., Lim, Y. C. and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit, *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, IEEE, pp. 59–66.

Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E. (2003). How to find trouble in communication, *Speech communication* **40**(1-2): 117–143.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). Emotion recognition in human-computer interaction, *IEEE Signal processing magazine* **18**(1): 32–80.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE transactions on acoustics, speech, and signal processing* **28**(4): 357–366.

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

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion., *Nebraska symposium on motivation*, University of Nebraska Press.

Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision, *CS224N project report, Stanford* **1**(12): 2009.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*, MIT press.

Han, K., Yu, D. and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine, *Interspeech 2014*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural computation* **9**(8): 1735–1780.

Kim, Y. (2014). Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* .

Krishna, K. V., Sainath, N. and Posonia, A. M. (2022). Speech emotion recognition using machine learning, *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, pp. 1014–1018.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning, *nature* **521**(7553): 436–444.

Liu, Z.-T., Han, M.-T., Wu, B.-H. and Rehman, A. (2023). Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning, *Applied Acoustics* **202**: 109178.

Logan, B., Van Thong, J.-M. and Moreno, P. J. (2005). Approaches to reduce the effects of oov queries on indexed spoken audio, *IEEE transactions on multimedia* **7**(5): 899–906.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques, *arXiv preprint cs/0205070* .

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis, *Foundations and Trends® in information retrieval* **2**(1–2): 1–135.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation, *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 704–714.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y. and Ramabhadran, B. (2013). Improvements to deep convolutional neural networks for lvcsr, *2013 IEEE workshop on automatic speech recognition and understanding*, IEEE, pp. 315–320.

Schuller, B., Batliner, A., Steidl, S. and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, *Speech communication* **53**(9-10): 1062–1087.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y. and Weninger, F. (2015). The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E. et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1555–1565.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.

Zhang, Z., Coutinho, E., Deng, J. and Schuller, B. (2014). Cooperative learning and its application to emotion recognition from speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(1): 115–126.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels, *Advances in neural information processing systems* **31**.

Zhou, Z., Zhang, X. and Sanderson, M. (2014). Sentiment analysis on twitter through topic-based lexicon expansion, *Databases Theory and Applications: 25th Australasian Database Conference, ADC 2014, Brisbane, QLD, Australia, July 14-16, 2014. Proceedings 25*, Springer, pp. 98–109.