NLP IN MENTAL HEALTH DETECTION: ANALYSIS & CLASSIFICATION

Contributors: *Minh Hieu Bui, Chun-Yen Lin, Hsin Yu Liao, Nehal Bhachu, Oprah Winfrey Ewurafua Ampah, Roopa Sree Gopalappa Seenappa, Suraj Shaik, Thao Nguyen.*

Class: *MSBA 265 – Special Analytics Topics*

Lecturer: *Professor Shyla Solis*

TABLE OF CONTENTS

PROJECT OVERVIEW

WHY?

Gen Z faces a mental health crisis, with high rates of anxiety, depression, and self-harm driven by social media pressures, global crises, and academic and financial stress. Traditional support systems are often overwhelmed, leaving a critical need for scalable, real-time solutions.

WHAT?

This project leverages NLP tools to detect and analyze mental health concerns through digital communications, enabling early, non-intrusive intervention tailored to Gen Z's unique language and communication styles.

HOW?

We aim to create an app that uses NLP to analyze mental health through user-shared stories in voice, text, or images. It provides real-time emotional analysis, empathetic responses, cyberbullying detection, and flags high-risk language for early intervention and tailored support.

NLP MODELS ANALYSIS

Data Preparation

- The dataset consists of text messages labeled as cyberbullying (1) or non-cyberbullying (0).
- Data is structured and loaded into a pandas DataFrame for inspection and preparation.

Modeling and Feature Engineering

- Logistic Regression, Random Forest, Support Vector Machine (SVM), and BERT-based Transformer.
- Feature engineering includes TF-IDF, n-grams, and sentiment analysis using NLTK.
- Models are trained to classify and predict cyberbullying with improved pattern detection.

Preprocessing and Augmentation

- Text preprocessing steps include cleaning (removing URLs, special characters, etc.), tokenization, stop word removal, and lemmatization.
- TF-IDF vectorization is used to convert text into numerical features.
- SMOTE is applied to address class imbalance, generating synthetic samples for the minority class (cyberbullying).

Training and Evaluation

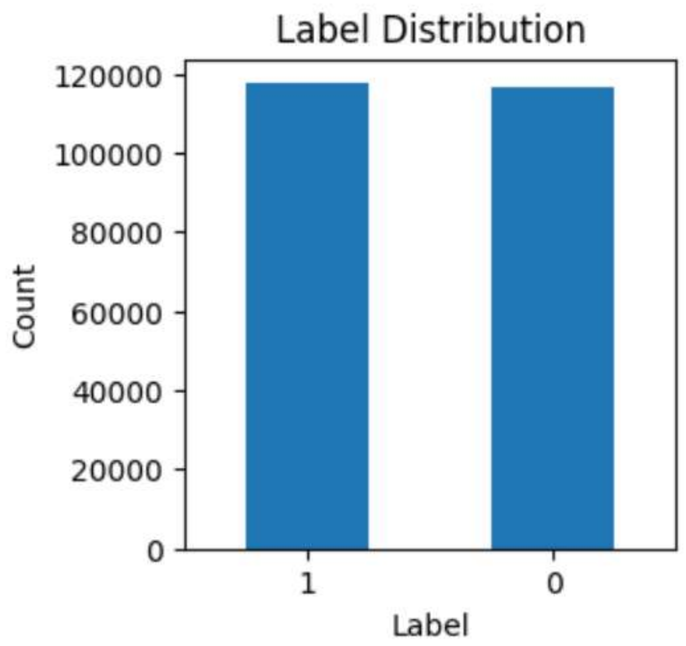
- The dataset is split into 80% training and 20% testing, and hyperparameter tuning is done using GridSearchCV.
- k-fold cross-validation ensures model stability.
- Evaluation metrics include accuracy, precision, recall, F1 score, and confusion matrix to assess model performance and reliability.

NLP MODELS ANALYSIS

```
Word Count Stats:
count    234060.000000
mean         24.558071
std          35.583312
min           1.000000
25%           7.000000
50%          13.000000
75%          27.000000
max         607.000000
Name: Word_Count, dtype: float64

Character Count Stats:
count    234060.000000
mean        173.061894
std         264.407795
min           1.000000
25%          46.000000
50%          89.000000
75%         188.000000
max        9965.000000
Name: Char_Count, dtype: float64

Most Common Words in Cleaned Text:
[('article', 38774), ('page', 35165), ('like', 32543), ('dont', 29782), ('one', 28371), ('wikipedia', 26639), ('would', 2608
7), ('please', 25105), ('im', 22625), ('think', 22472)]
```



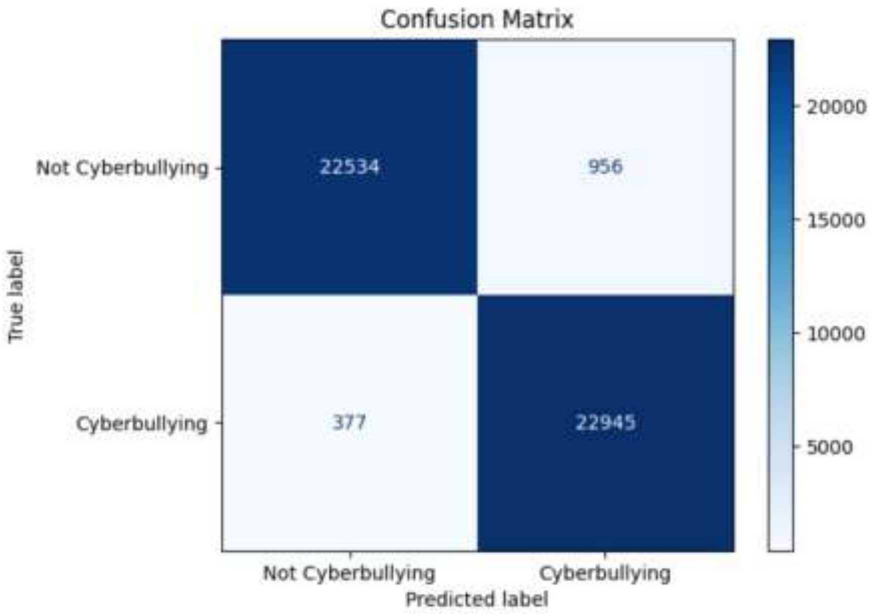
	Text	Label	Cleaned_Text
0	u0 lmao wow fuck you too	0	lmao wow fuck
1	a white dress and red lipstick make everything...	0	white dress red lipstick make everything better
2	this has been a trend since <number> of course...	0	trend since course wall street assumed eternal...
3	<user> <user> babies in cages destroying envir...	0	baby cage destroying environment rolling back ...
4	<user> more good neighbours yes that working o...	0	good neighbour yes working well crime skyrocke...

NLP MODELS ANALYSIS

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     23490
           1       0.96      0.98      0.97     23322

    accuracy                           0.97     46812
   macro avg       0.97      0.97      0.97     46812
weighted avg       0.97      0.97      0.97     46812
```



BERT-Based Model Performance

- Achieved 97% accuracy and a balanced F1-score of 0.97.
- Captured contextual nuances in text with minimal false positives and false negatives.
- Correctly classified 22,945 out of 23,322 cyberbullying messages and 22,534 out of 23,490 non-cyberbullying messages.
- Precision: 0.96, Recall: 0.98 for cyberbullying; vice versa for non-cyberbullying.

Comparison with Other Models

- **Random Forest**: Solid performance, capturing non-linear relationships but slightly lower recall than BERT.
- **Logistic Regression**: Provided reliable baseline results but struggled with textual complexity.
- **SVM and Naive Bayes**: Delivered reasonable performance but were less effective than BERT or Random Forest.

NLP MODELS ANALYSIS

Data Preparation

- Dataset: 11,000 Twitter messages, each with an emotion label (Sadness, Joy, Love, Fear, Anger, Surprise).
- Fields: Message, Source, Label (numeric, 0-5), Emotion (emotion classification).
- The dataset is used for text analysis, emotion classification, and sentiment analysis.

Modeling and Feature Engineering

- Logistic Regression (baseline model), Random Forest (ensemble model), and Support Vector Machine (SVM).
- TF-IDF vectorization, n-grams (unigrams and bigrams), sentiment analysis using NLTK's SentimentIntensityAnalyzer.
- Hyperparameter tuning using GridSearchCV.
- K-fold cross-validation for model stability and overfitting prevention.

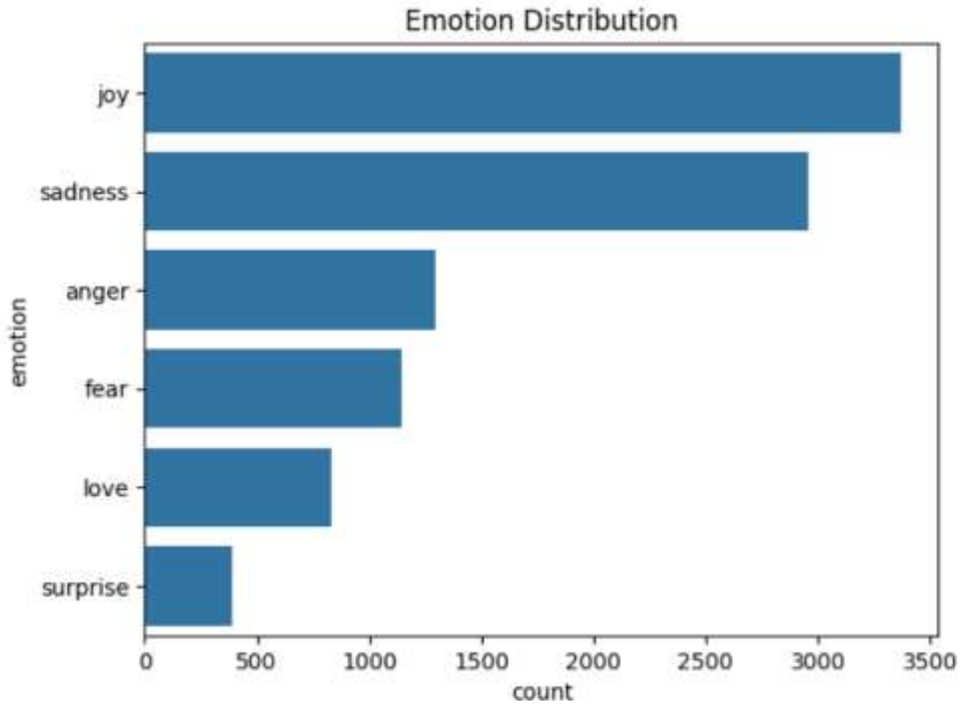
Preprocessing and Augmentation

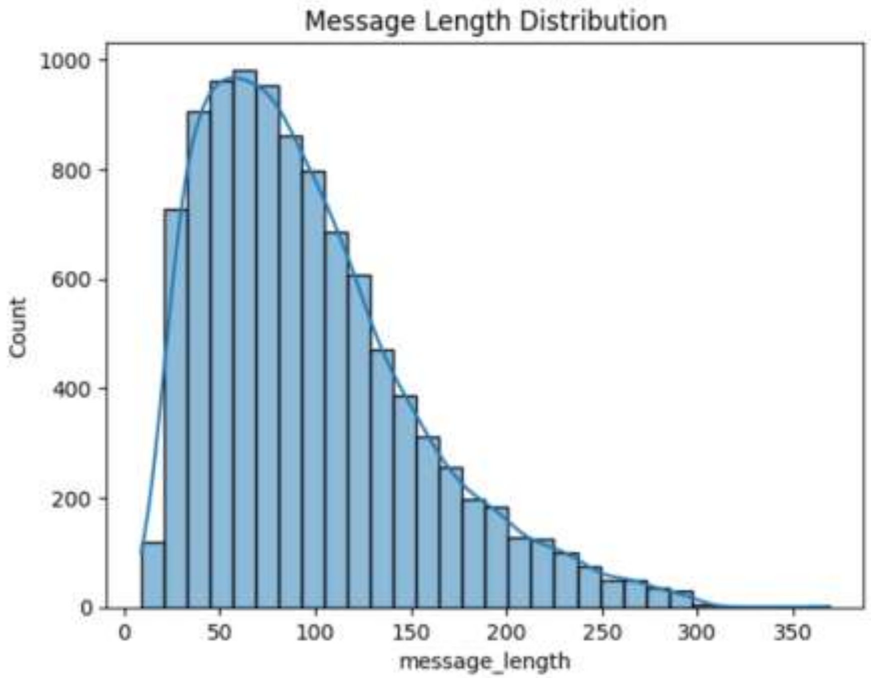
- Text normalization (lowercase, removal of special characters and punctuation).
- Tokenization, stop word removal, stemming, and lemmatization.
- TF-IDF vectorization with unigrams and bigrams for feature extraction.
- Applied SMOTE to address class imbalance, particularly for underrepresented emotions (Fear and Surprise).
- Synthetic samples generated for balanced dataset classes.

Training and Evaluation

- Accuracy, precision, recall, and F1-score to evaluate model performance.
- Focus on class imbalance, with metrics assessing prediction accuracy and ability to identify true positives.

NLP MODELS ANALYSIS





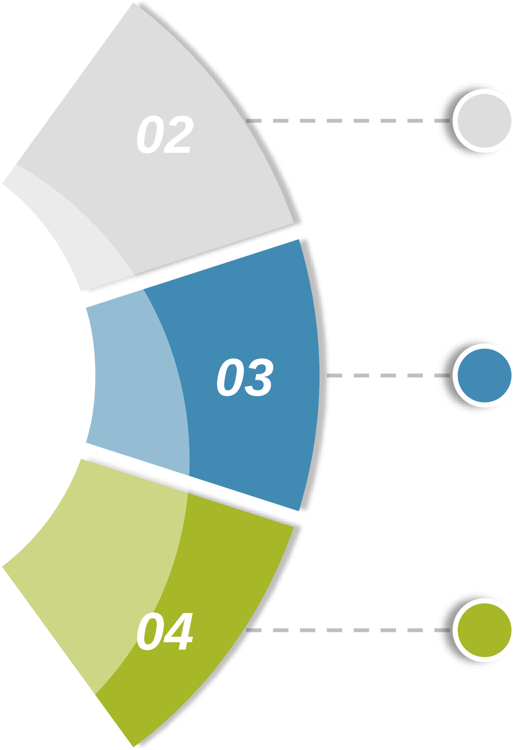
NLP MODELS ANALYSIS

Best Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.91	0.84	0.88	675
1	0.82	0.86	0.84	674
2	0.88	0.94	0.91	600
3	0.94	0.91	0.93	600
4	0.93	0.96	0.95	600
5	0.98	0.96	0.97	600
accuracy			0.91	3749
macro avg	0.91	0.91	0.91	3749
weighted avg	0.91	0.91	0.91	3749

Best Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.93	0.89	0.91	675
1	0.94	0.85	0.89	674
2	0.90	0.98	0.94	600
3	0.93	0.96	0.95	600
4	0.92	0.96	0.94	600
5	0.95	0.95	0.95	600
accuracy			0.93	3749
macro avg	0.93	0.93	0.93	3749
weighted avg	0.93	0.93	0.93	3749



Logistic Regression vs. Random Forest

Logistic Regression achieved higher accuracy (93%) compared to Random Forest (91%). While both models performed well, Logistic Regression demonstrated slightly better overall classification accuracy across all emotion categories.

TF-IDF Vectorization and Sentiment Analysis

TF-IDF vectorization combined with sentiment analysis using NLTK, improved the model’s ability to understand the emotional tone in Twitter messages, leading to more accurate predictions of emotional states.

Future Work

Future work could explore more advanced machine learning models, such as deep learning approaches (e.g., LSTM, BERT, or transformer-based models), to further improve emotion detection accuracy. These models could capture more complex patterns in text, handle larger datasets, and offer deeper contextual understanding of user emotions.

NLP MODELS ANALYSIS

Data Preparation

- The CREMA-D dataset contains 7,442 audio clips from 91 actors, featuring six emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) at four intensity levels.
- Audio file paths and emotion labels are extracted and standardized for consistency. Exploratory Data Analysis (EDA) is conducted to examine dataset structure and distribution.

Modeling and Feature Engineering

- CNN is used to efficiently process time-series audio data, with convolutional and pooling layers to extract spatial features and reduce dimensionality.
- The architecture combines 1D CNNs for time-domain with LSTMs for sequence modeling and 2D CNNs for frequency-domain with optimizers like SGD and Adam.

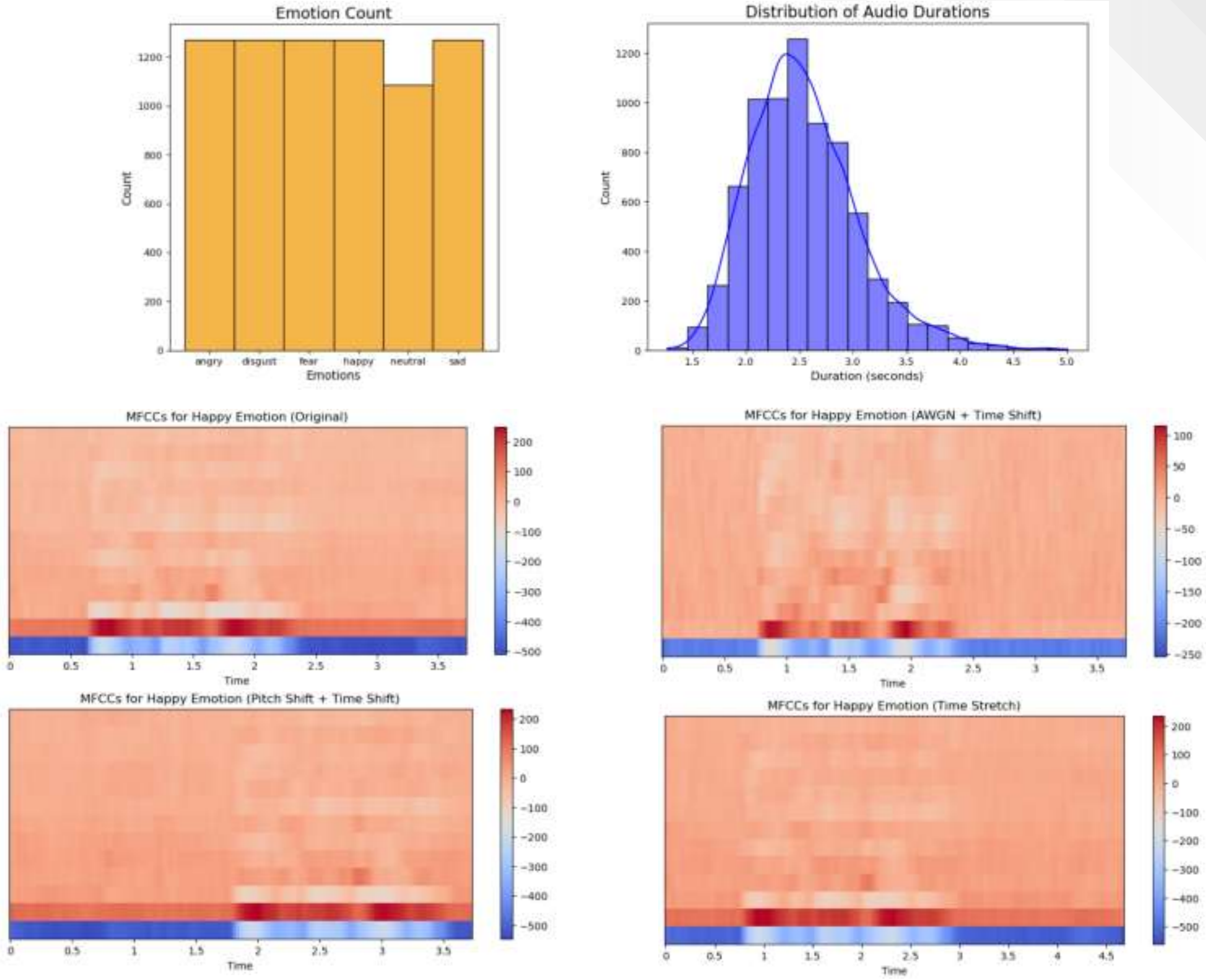
Preprocessing and Augmentation

- Audio features (sampling rate, length, and raw data) are processed, and labels are converted to numeric values using label encoding and one-hot encoding.
- Data augmentation are applied to diversify training data and improve model generalization.
- Features are extracted to represent temporal and frequency-domain characteristics of the audio.

Training and Evaluation

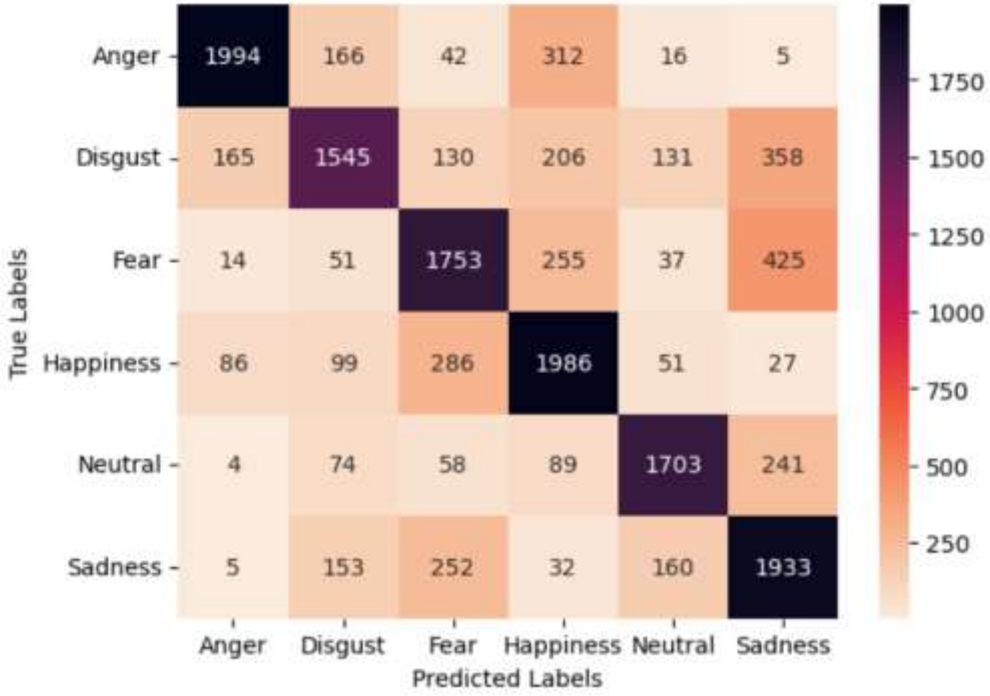
- Accuracy, precision, recall, and F1-score to evaluate model performance.
- Confusion Matrix: Visualize misclassifications across emotion classes, providing insights into precision and recall for each emotion.

NLP MODELS ANALYSIS



NLP MODELS ANALYSIS

	precision	recall	f1-score	support
0	0.88	0.79	0.83	2535
1	0.74	0.61	0.67	2535
2	0.70	0.69	0.69	2535
3	0.69	0.78	0.73	2535
4	0.81	0.79	0.80	2169
5	0.65	0.76	0.70	2535
accuracy			0.74	14844
macro avg	0.74	0.74	0.74	14844
weighted avg	0.74	0.74	0.74	14844



Class-Level Performance and Accuracy

Time Domain Model:
- Struggled with underrepresented classes like "Fear" and "Surprise."
- Validation accuracy for these classes remained below 90%, even with augmentation.

Frequency Domain Model:
- Achieved higher class-wise accuracy across all emotions, with validation accuracy exceeding 92% for even the most challenging classes.

Time Domain Model:
- Often misclassified "Neutral" as "Sadness" due to overlapping temporal features.

Frequency Domain Model:
- Fewer misclassifications, for emotions with distinct spectral patterns like "Happiness" and "Anger."
- Minor confusion persisted between "Sadness" and "Neutral".

Confusion Matrix Insights

NLP MODELS ANALYSIS

Data Preparation

- **FER2013**: Contains ~35,000 grayscale facial images (48x48 pixels) categorized into seven emotional labels.
- **Emotion Recognition Dataset**: Provides a diverse set of facial expression images labeled with five emotions.
- The datasets were merged and re-encoded for uniformity, forming a unified dataset for model training, validation, and testing.

Modeling and Feature Engineering

- The model uses **ResNet-50**, a deep convolutional neural network pre-trained on ImageNet.
- Transfer learning was employed to leverage pre-trained weights, reducing the reliance on large labeled datasets and speeding up convergence.
- Adam optimizer was used with a learning rate of 0.001 to update model weights.

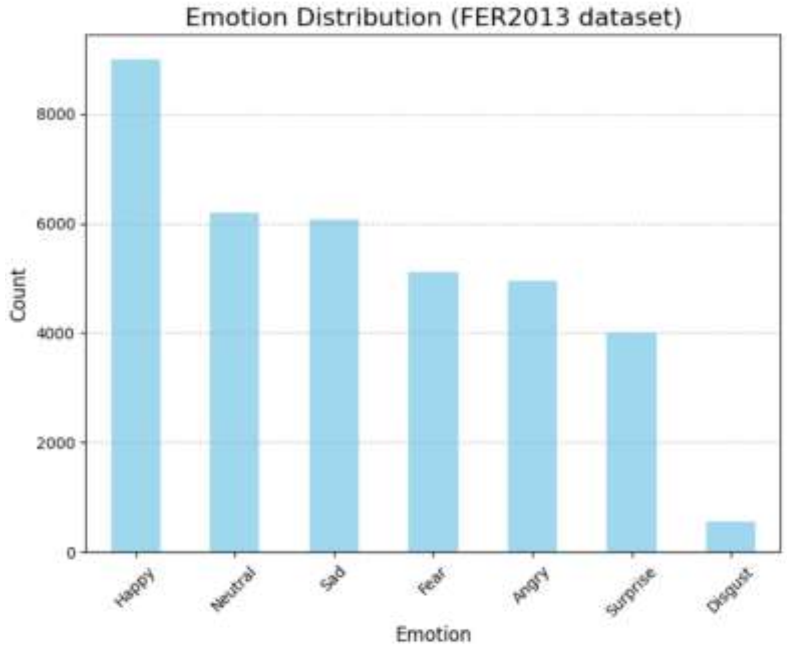
Preprocessing and Augmentation

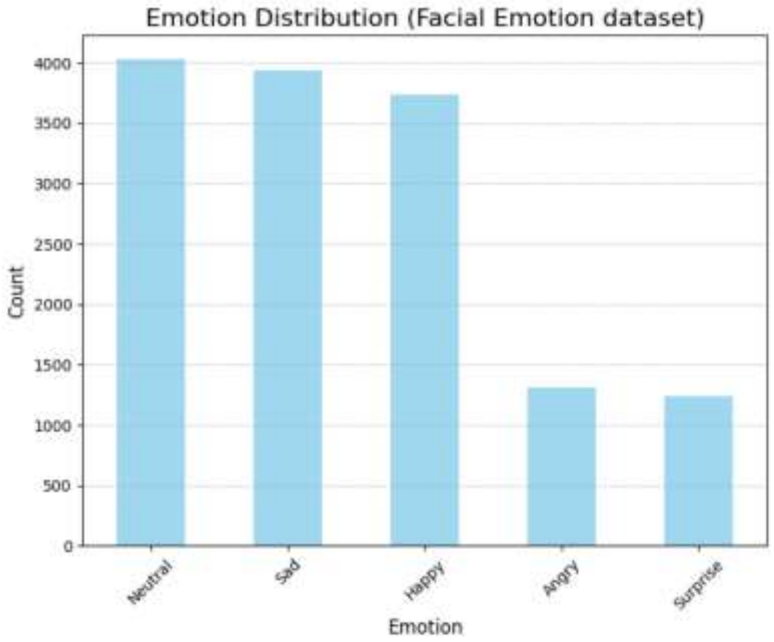
- Pixel values were reshaped into 48x48 matrices and converted into grayscale images.
- Data normalization was applied to scale pixel values to the range [-1, 1].
- Random horizontal flipping was applied to the training set to increase model robustness and improve generalization.

Training and Evaluation

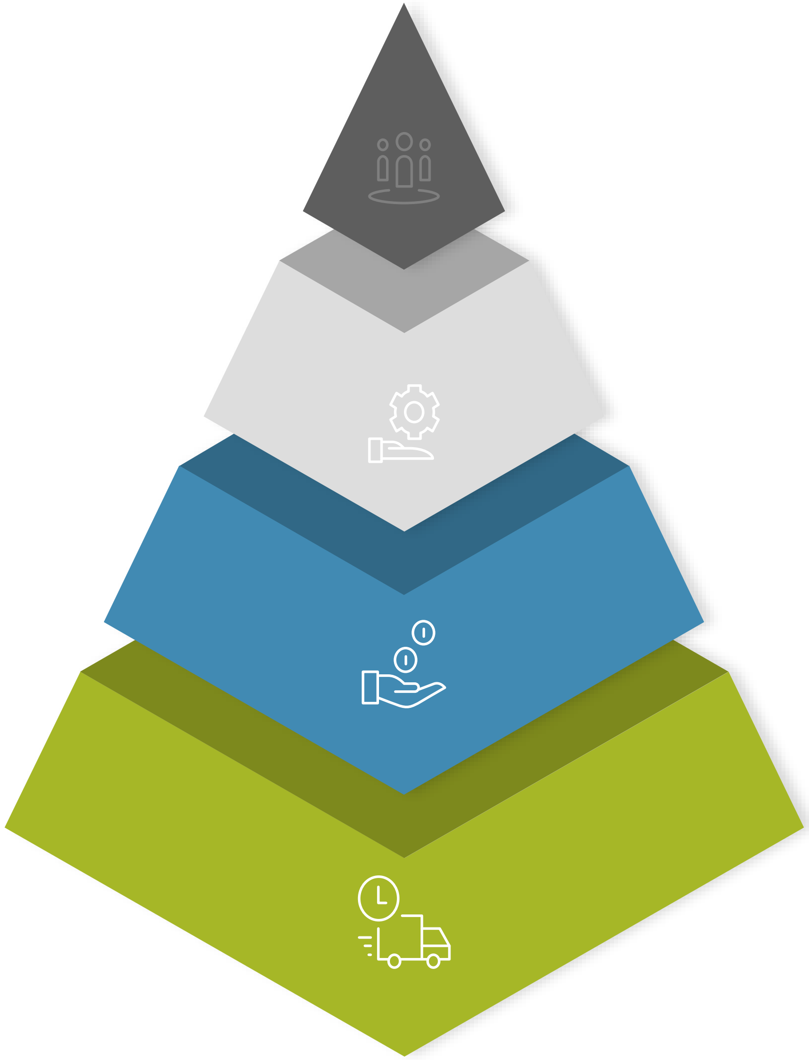
- The model's performance was assessed based on accuracy, measuring the proportion of correctly classified emotional labels.
- This evaluation metric ensured an effective assessment of the model's ability to accurately identify facial expressions.

NLP MODELS ANALYSIS





NLP MODELS ANALYSIS



Model Performance on Test Data

The **ResNet-50** model achieved a test accuracy of 69.69%, correctly classifying emotional states in approximately 7 out of 10 cases, which provides a foundational assessment of its ability to analyze facial expressions in grayscale images.

Evaluation Using Video Data

Processed face images are input into the ResNet-50 model, and emotions are predicted based on the highest probability from five categories (happiness, anger, sadness, surprise, and neutrality).

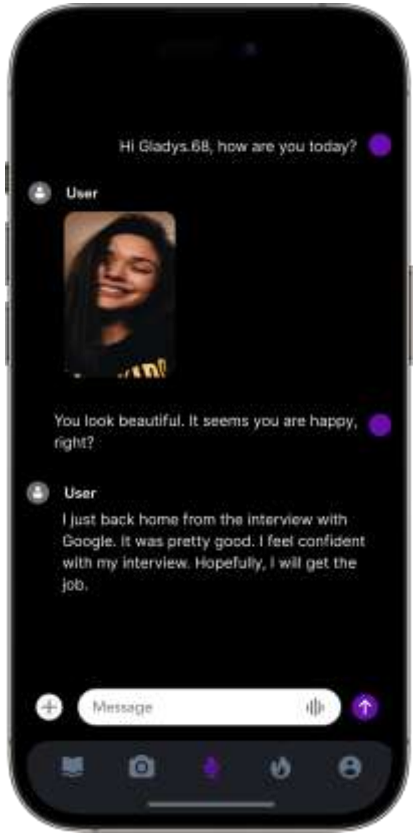
Real-World Testing

The model was tested with both real-time video and pre-recorded videos, including using a computer webcam for live testing.

Results and Observations

Despite challenges with lighting and facial orientation, the model consistently performed well, identifying emotions accurately in dynamic, real-world settings.

NLP MODELS & APPLICATION DEMO



CONCLUSION

This research underscores NLP's transformative potential in mental health monitoring, using advanced models to analyze text, audio, and images, while offering actionable insights for future improvements in data diversity, ethical practices, and real-world deployment.

Demonstrated the effectiveness of advanced NLP models (e.g., BERT, ResNet-50, CNNs) in mental health tasks like cyberbullying detection, emotion classification, and sentiment analysis.

Emphasized the critical role of data preprocessing, augmentation, and feature engineering in enhancing model accuracy.

Highlighted NLP as a scalable, non-intrusive solution for real-world mental health applications despite resource constraints.

FUTURE RECOMMENDATIONS

Innovation

NLP in mental health detection has the potential to revolutionize how we understand and support emotional well-being. Future advancements could focus on improving accuracy, personalization, and ethical considerations while expanding its applications in diverse, real-world settings.

Expand Dataset Diversity

- Incorporate multilingual and culturally diverse datasets for better generalizability.
- Use real-world data from broader platforms to increase ecological validity.

Advanced Model Development

- Explore cutting-edge transformer-based models for contextual understanding.
- Optimize models for real-time use, particularly on edge devices like smartphones.

Real-World Deployment

- Pilot systems with mental health professionals to assess practical effectiveness.
- Conduct longitudinal studies to measure long-term impact on mental health outcomes.

Improved User Interaction

- Design empathetic and context-sensitive AI interfaces to support user needs.
- Incorporate personalization to adapt responses to individual communication styles.

NLP IN MENTAL HEALTH DETECTION: ANALYSIS & CLASSIFICATION

Thank You For Listening