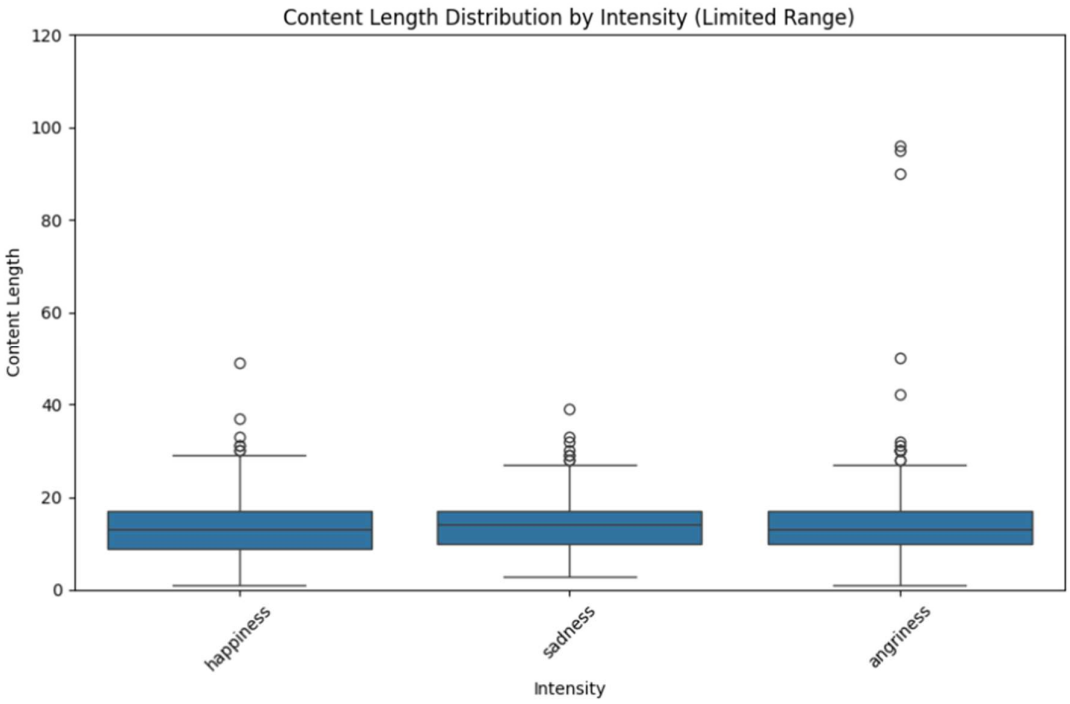
INTENSITY ANALYSIS

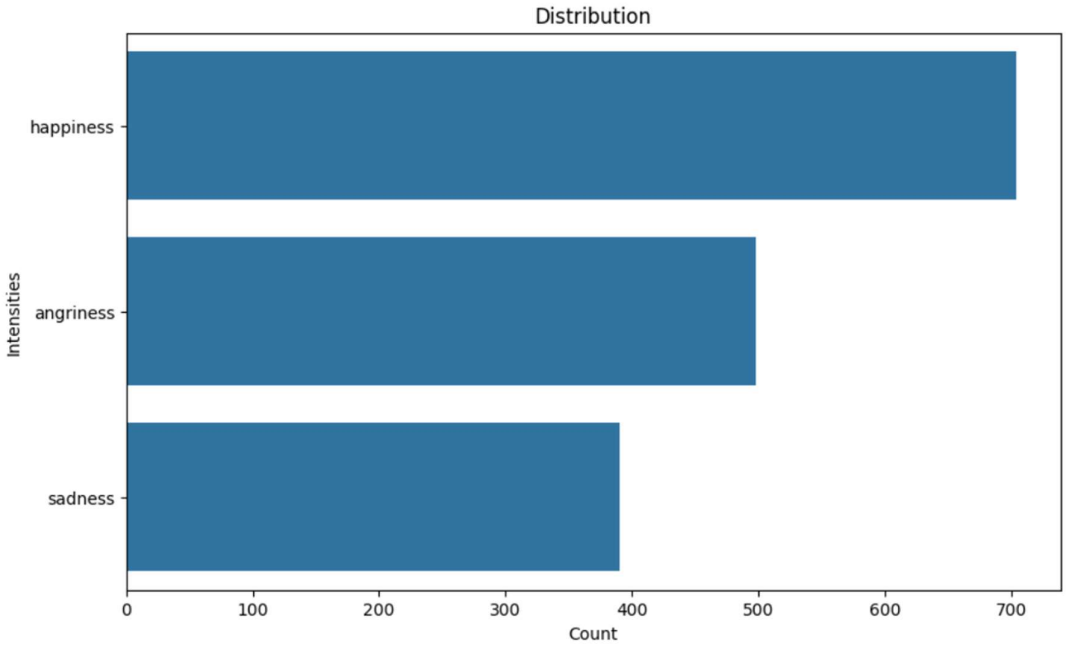
➢ **Rashmi Ranjan**

TABLE OF CONTENTS

1.DATASET OVERVIEW

The dataset contains more categories of Happiness followed by Angriness and Sadness. This implies that when humans are happy, they tend to send longer and more messages compared to when they are angry or sad.



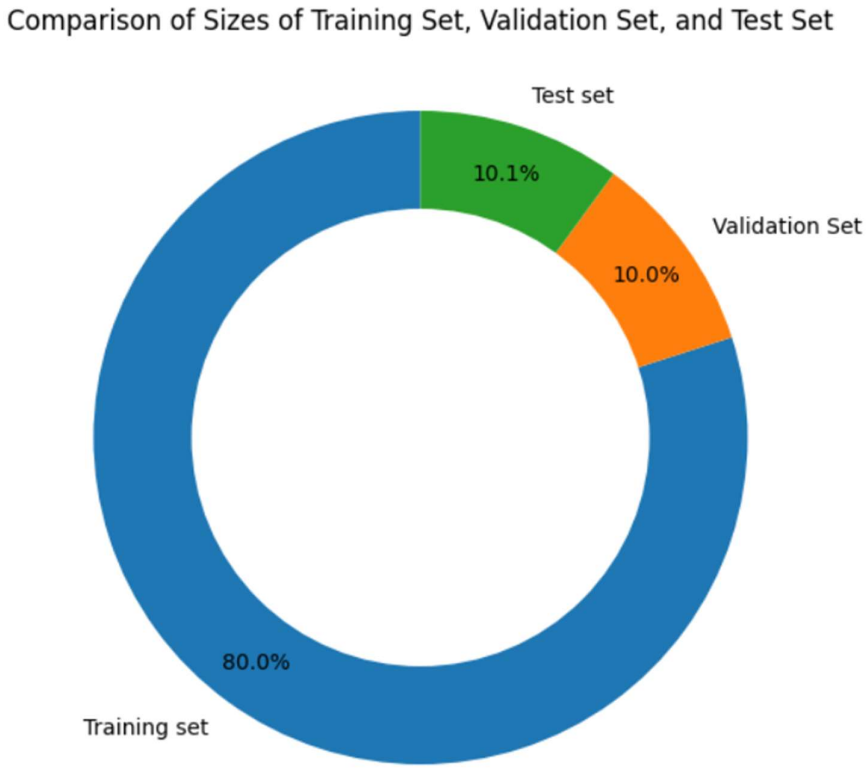


Top words for Happiness

Top words for Angriness

Top words for Sadness

2.SPLITTED DATA



3.TEXT-NORMALIZATION

The text preprocessing pipeline includes the following steps:

1. Convert to lowercase
2. Remove whitespace
3. Remove newline characters
4. Remove ".com" substrings
5. Remove URLs
6. Remove punctuation
7. Remove HTML tags
8. Remove emojis
9. Handle problematic characters in words
10. Convert acronyms
11. Expand contractions
12. handle slangs and abbreviations
13. Correct spelling
14. Lemmatize text
15. Discard non-alphabetic characters
16. Keep specific parts of speech
17. Remove stopwords

Normalized words for Test

Normalized words for Training

Normalized words for Validation

4.TF-IDF MODELS

- Best Model: Random Forest achieved the highest validation accuracy (0.698) and training accuracy (0.971), making it the most suitable classifier.
- Balanced Performance: Linear SVM (Training: 0.840, Validation: 0.654) and XGBoost (Training: 0.933, Validation: 0.654) showed relatively good performance but some signs of overfitting.
- Overfitting Models: Decision Tree (Training: 0.971, Validation: 0.635) displayed high training accuracy but lower generalization.
- Moderate Performance: Logistic Regression, SGD Classifier, and Ridge Classifier had moderate training accuracy (~0.82–0.88) but validation accuracy ≤0.60.
- Underperforming Models: KNN Classifier and AdaBoost performed poorly, with validation accuracy below 0.60.

	Classifier	Training accuracy	Validation accuracy
4	Random Forest	0.970935	0.698113
3	Linear SVM	0.839749	0.654088
7	XGBoost	0.933229	0.654088
2	Decision Tree	0.970935	0.635220
0	Logistic Regression	0.825609	0.597484
5	SGD Classifier	0.881383	0.597484
8	AdaBoost	0.568735	0.597484
6	Ridge Classifier	0.838963	0.578616
1	KNN Classifier	0.567164	0.547170

5.BoW MODELS

- Best Model: Random Forest achieved the highest training accuracy (0.973) and the highest validation accuracy (0.686), making it the top-performing classifier.
- Good Performance: XGBoost (Training: 0.874, Validation: 0.679) and Decision Tree (Training: 0.973, Validation: 0.673) showed strong training results but signs of overfitting.
- Balanced Option: Multinomial Naive Bayes (Training: 0.799, Validation: 0.660) demonstrated consistent performance with less overfitting.
- Moderate Models: Linear SVM, Logistic Regression, and SGD Classifier had moderate training accuracy (~0.86–0.91) and validation accuracy around 0.61.
- Underperforming Models: AdaBoost, Ridge Classifier, and KNN Classifier struggled, with validation accuracy below 0.61, particularly KNN (Validation: 0.491).

	Classifier	Training accuracy	Validation accuracy
5	Random Forest	0.972506	0.685535
8	XGBoost	0.874313	0.679245
3	Decision Tree	0.972506	0.672956
0	MultinomialNB	0.798900	0.660377
4	Linear SVM	0.875884	0.622642
1	Logistic Regression	0.868814	0.610063
6	SGD Classifier	0.912019	0.610063
9	AdaBoost	0.566379	0.610063
7	Ridge Classifier	0.851532	0.572327
2	KNN Classifier	0.448547	0.490566

6.BoW HYPERPARAMETER TUNING

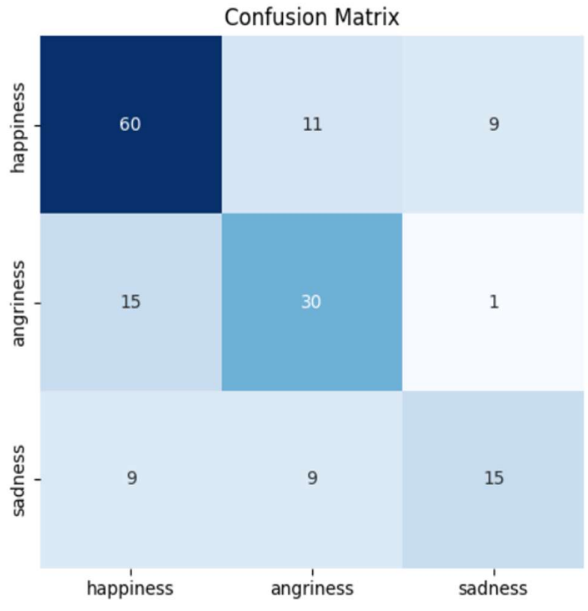
Classification report for training set

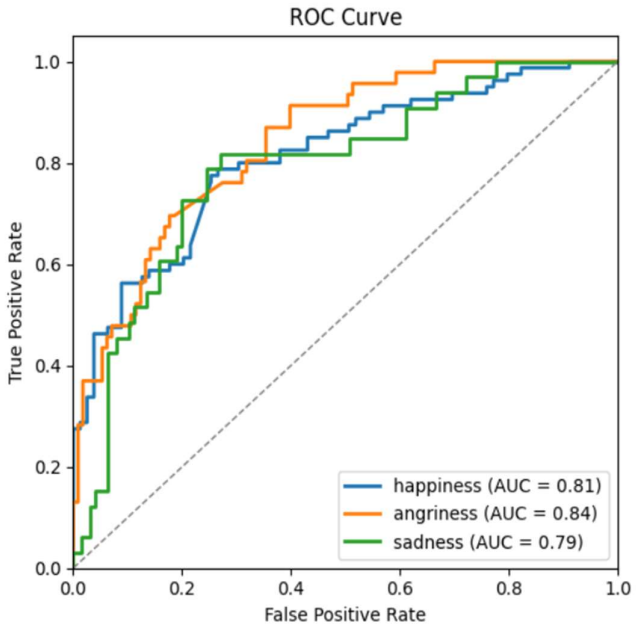
	precision	recall	f1-score	support
happiness	0.74	0.87	0.80	555
angriness	0.78	0.70	0.74	402
sadness	0.79	0.65	0.71	316
accuracy			0.76	1273
macro avg	0.77	0.74	0.75	1273
weighted avg	0.77	0.76	0.76	1273

Classification report for test set

	precision	recall	f1-score	support
happiness	0.71	0.75	0.73	80
angriness	0.60	0.65	0.62	46
sadness	0.60	0.45	0.52	33
accuracy			0.66	159
macro avg	0.64	0.62	0.62	159
weighted avg	0.66	0.66	0.66	159

After doing hypertunning, the Linear SVM model is performing well, got accuracy: 71%, minimal overfitting





7.Word2Vec MODELS

- **SGD Classifier** achieved the highest balance between training and validation accuracy (**82.6%** and **69.8%**, respectively), making it the most reliable for generalization.
- **Random Forest** and **XGBoost** showed very high training accuracies (both **99.3%**) but lower validation accuracies (**67.9%** and **67.3%**), indicating overfitting.
- **Linear SVM** had moderate performance with a validation accuracy of **66.6%** but lower training accuracy (**75.0%**), showing potential with further tuning.
- **Ridge Classifier** and **Logistic Regression** demonstrated average performance, while **KNN Classifier**, **AdaBoost**, and **Decision Tree** performed poorly, with **Decision Tree** having the lowest validation accuracy (**50.3%**), heavily overfitting.

	Classifier	Training accuracy	Validation accuracy
5	SGD Classifier	0.826394	0.698113
4	Random Forest	0.992930	0.679245
7	XGBoost	0.992930	0.672956
3	Linear SVM	0.750982	0.666667
6	Ridge Classifier	0.791830	0.647799
0	Logistic Regression	0.765907	0.616352
1	KNN Classifier	0.625295	0.610063
8	AdaBoost	0.698350	0.591195
2	Decision Tree	0.992930	0.503145

8.Word2Vec HYPERPARAMETER TUNING

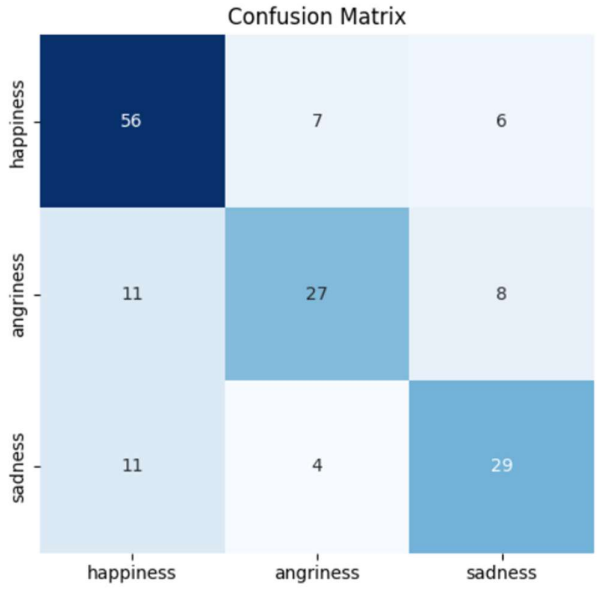
Classification report for training set

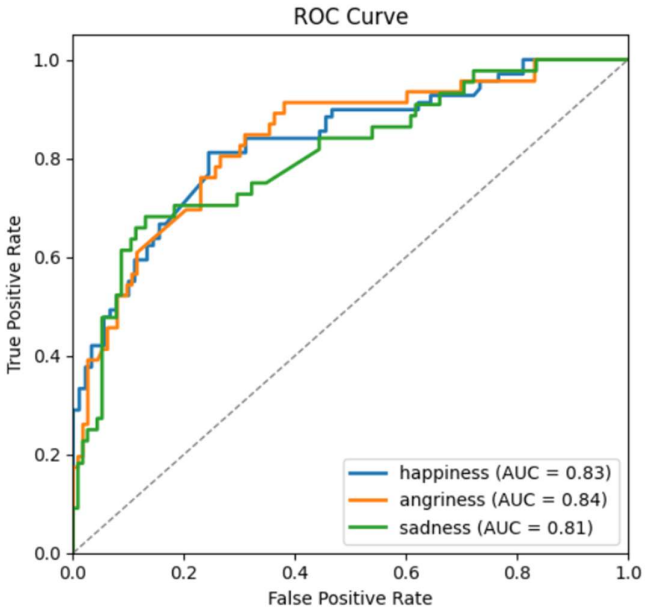
	precision	recall	f1-score	support
happiness	0.76	0.87	0.81	561
angriness	0.79	0.72	0.75	401
sadness	0.79	0.66	0.72	311
accuracy			0.77	1273
macro avg	0.78	0.75	0.76	1273
weighted avg	0.78	0.77	0.77	1273

Classification report for test set

	precision	recall	f1-score	support
happiness	0.72	0.81	0.76	69
angriness	0.71	0.59	0.64	46
sadness	0.67	0.66	0.67	44
accuracy			0.70	159
macro avg	0.70	0.69	0.69	159
weighted avg	0.70	0.70	0.70	159

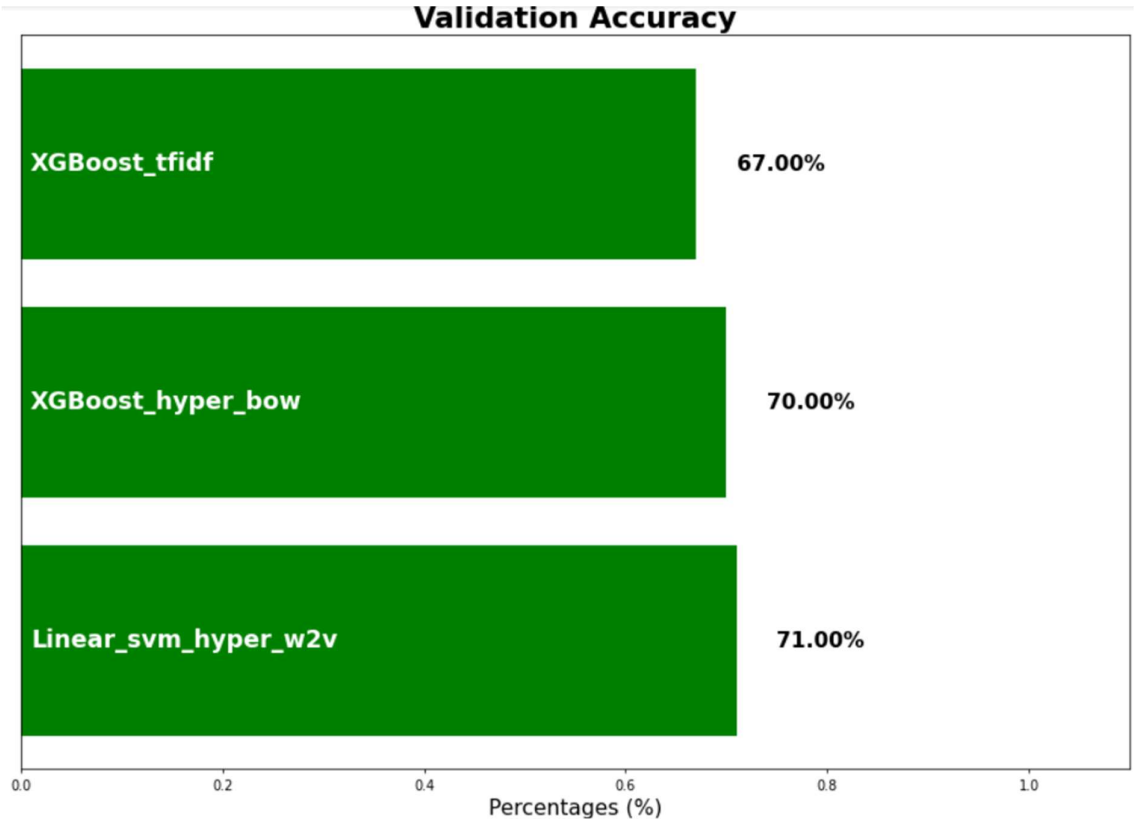
After doing hypertunning, the model is performing well, got accuracy: 71%, minimal overfitting.





9.CONCLUSION

Model Name	Training Accuracy	Validation Accuracy	Overfitting
Linear SVM with Word2Vec	84%	71%	No
XGBoost with Bag-of-Words	82%	70%	No
XGBoost with TF-IDF	77%	67%	No



The **Linear SVM with Word2Vec** model achieved the highest validation accuracy of **71%**, with minimal overfitting, making it the best-performing model for emotion classification. Despite the limited dataset, this model demonstrates strong generalization and reliability for predicting emotions on unseen data.

The data, sourced from **WhatsApp chat analysis of Indian users**, highlights the potential of domain-specific insights. With more extensive data, accuracy is expected to improve further.

THANK YOU 😊