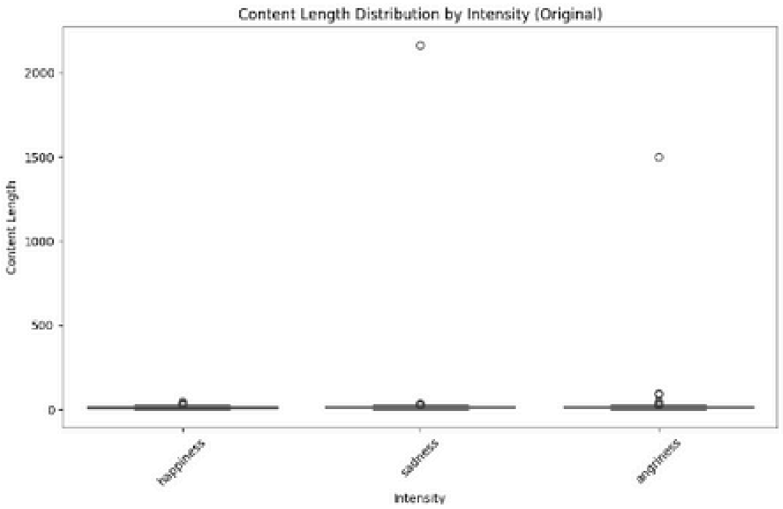
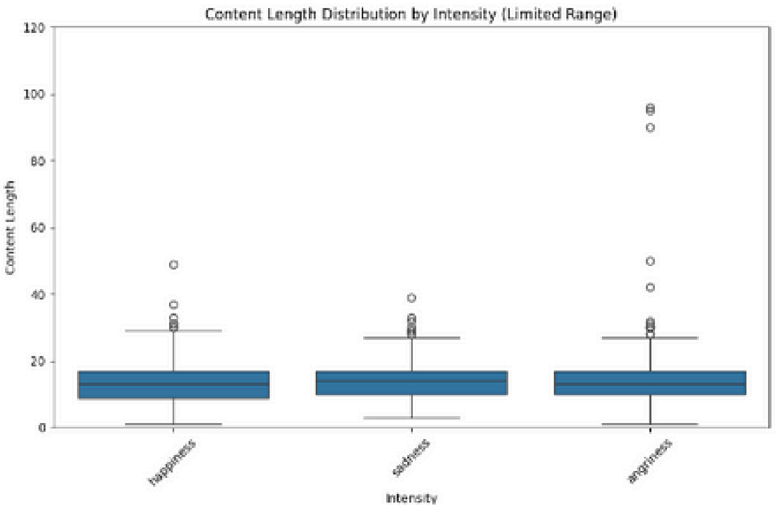
EMOTION CLASSIFICATION

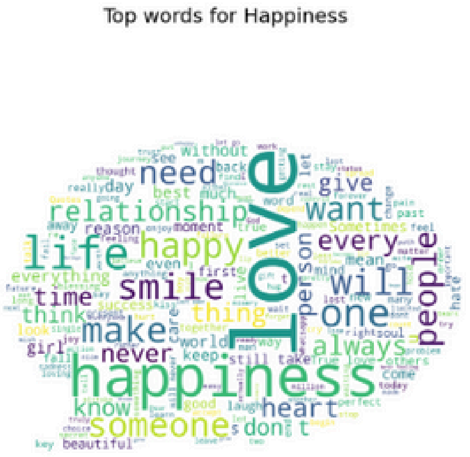
Vishal Naik

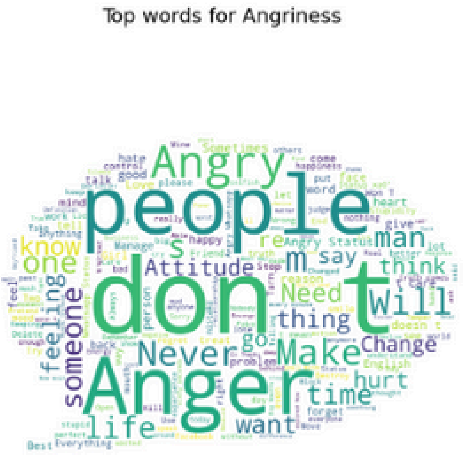
TABLE OF CONTENTS

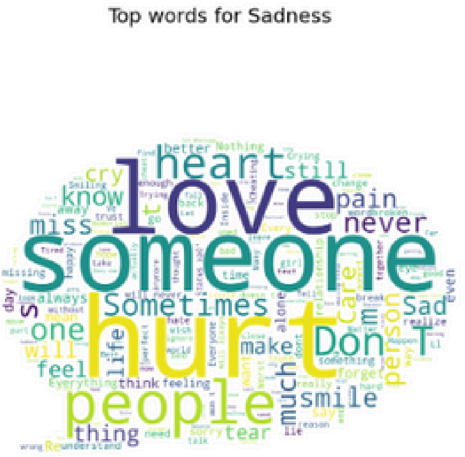
DATASET OVERVIEW

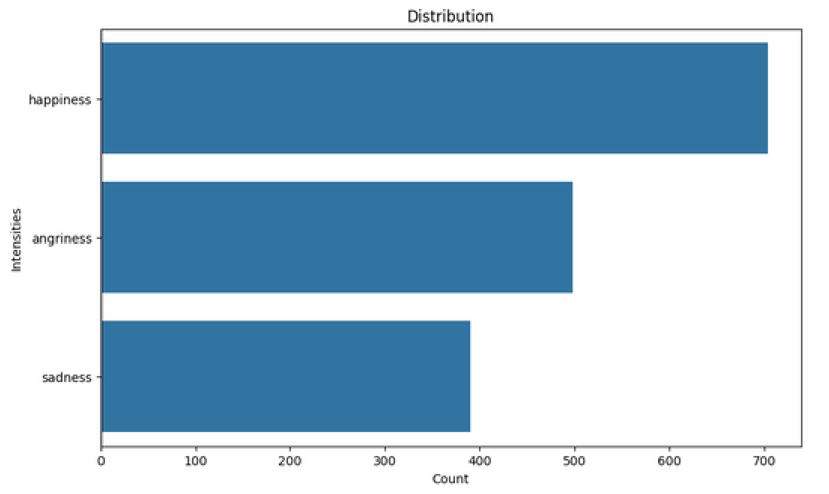






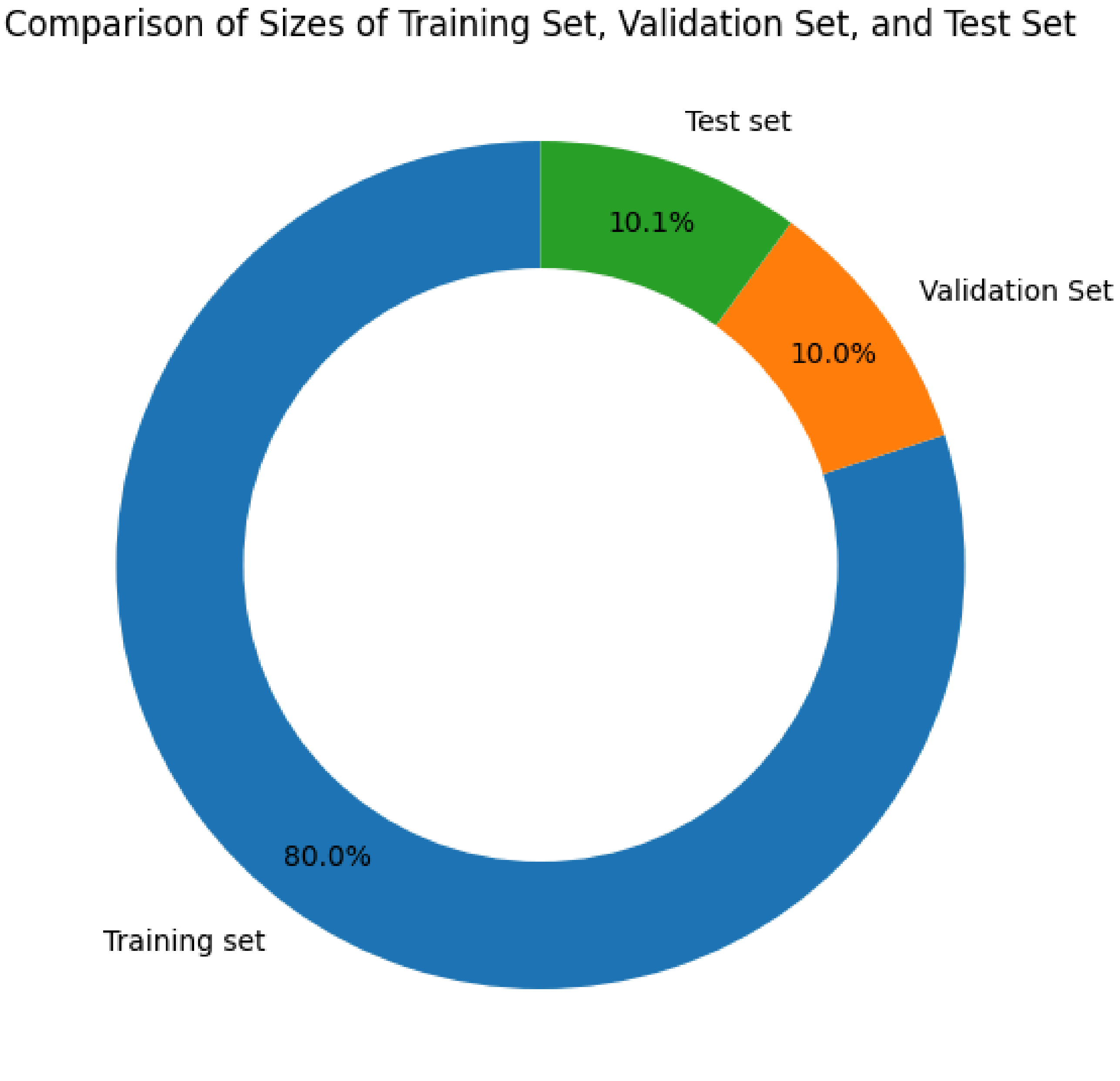




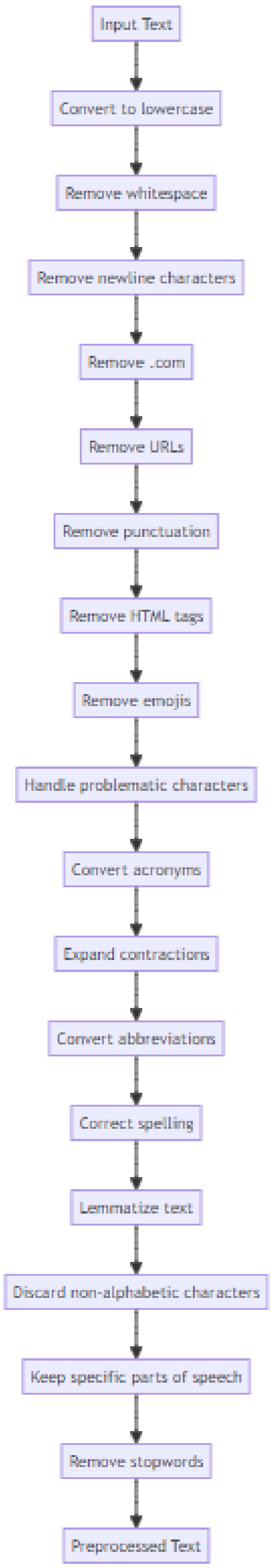


The dataset contains more categories of Happiness followed by Angriness and Sadness. This implies that when humans are happy, they tend to send longer and more messages compared to when they are angry or sad.

SPLITTED DATA

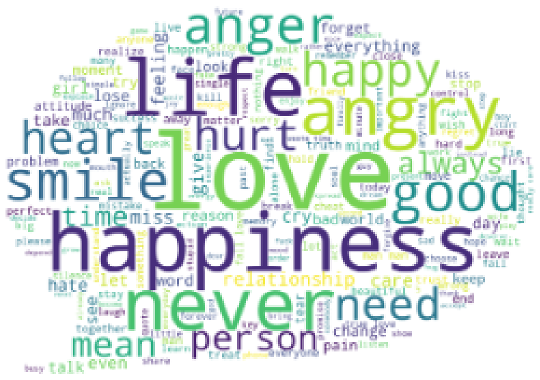


TEXT NORMALIZATION

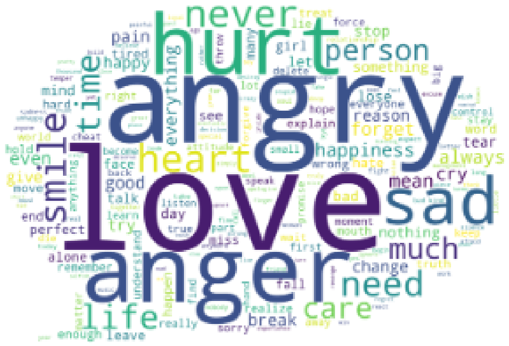


Input Text
Convert to lowercase
Remove whitespace
Remove newline characters
Remove .com
Remove URLs
Remove punctuation
Remove HTML tags
Remove emojis
Handle problematic characters
Convert acronyms
Expand contractions
Convert abbreviations
Correct spelling
Lemmatize text
Discard non-alphabetic characters
Keep specific parts of speech
Remove stopwords
Preprocessed Text

mormalized words for Training

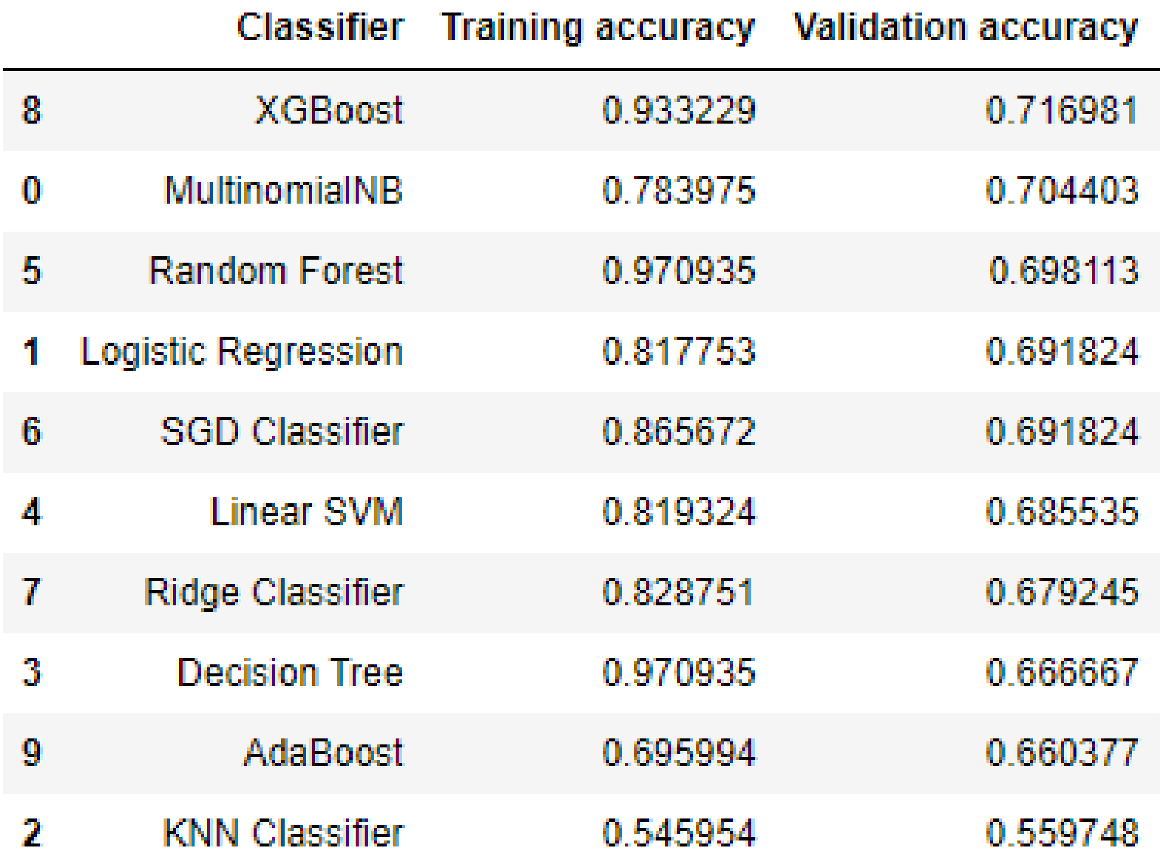


mormalized words for Validation



mormalized words for Validation

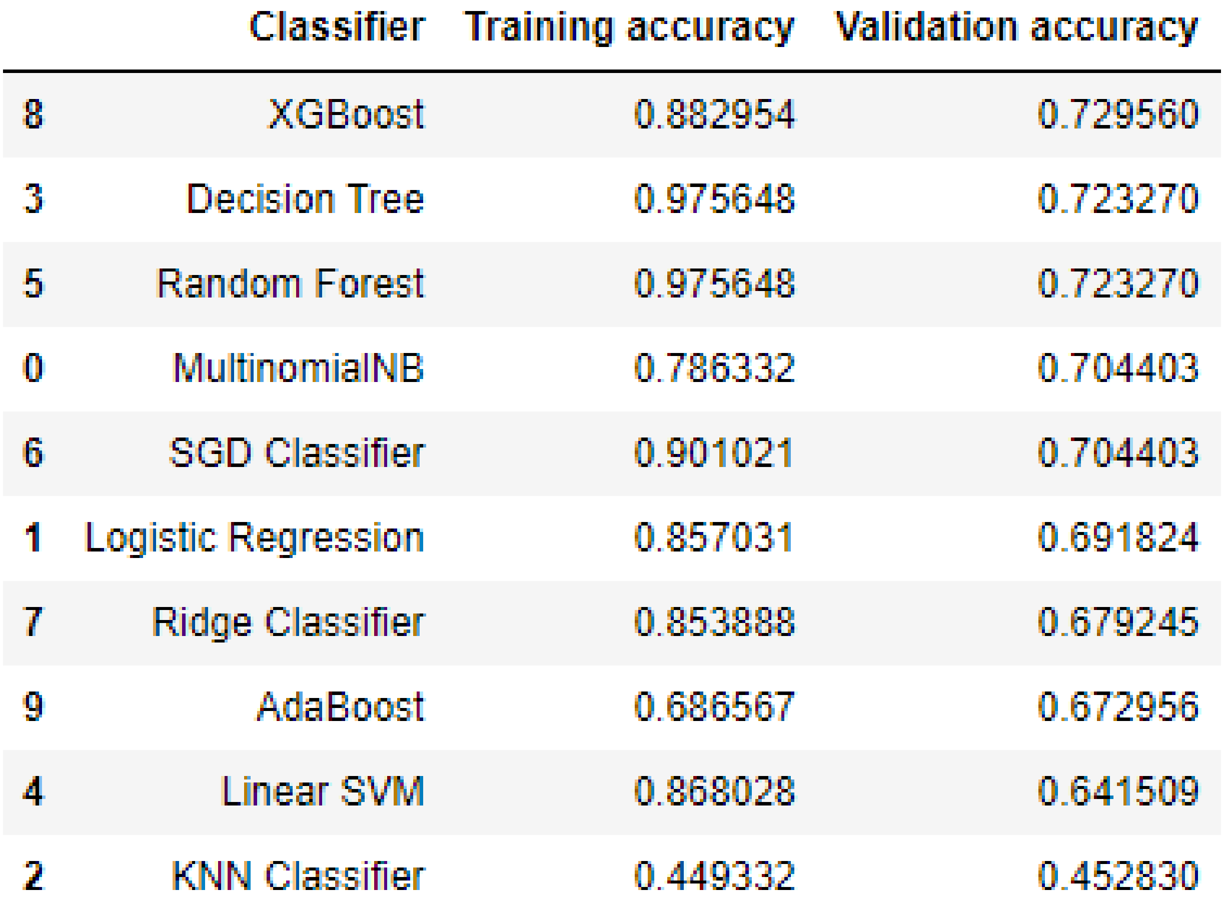
TF-IDF MODELS



	Classifier	Training accuracy	Validation accuracy
8	XGBoost	0.933229	0.716981
0	MultinomialNB	0.783975	0.704403
5	Random Forest	0.970935	0.698113
1	Logistic Regression	0.817753	0.691824
6	SGD Classifier	0.865672	0.691824
4	Linear SVM	0.819324	0.685535
7	Ridge Classifier	0.828751	0.679245
3	Decision Tree	0.970935	0.666667
9	AdaBoost	0.695994	0.660377
2	KNN Classifier	0.545954	0.559748

XGBoost performed well on Training data but not on validation data, it is clearly Overfitting!!

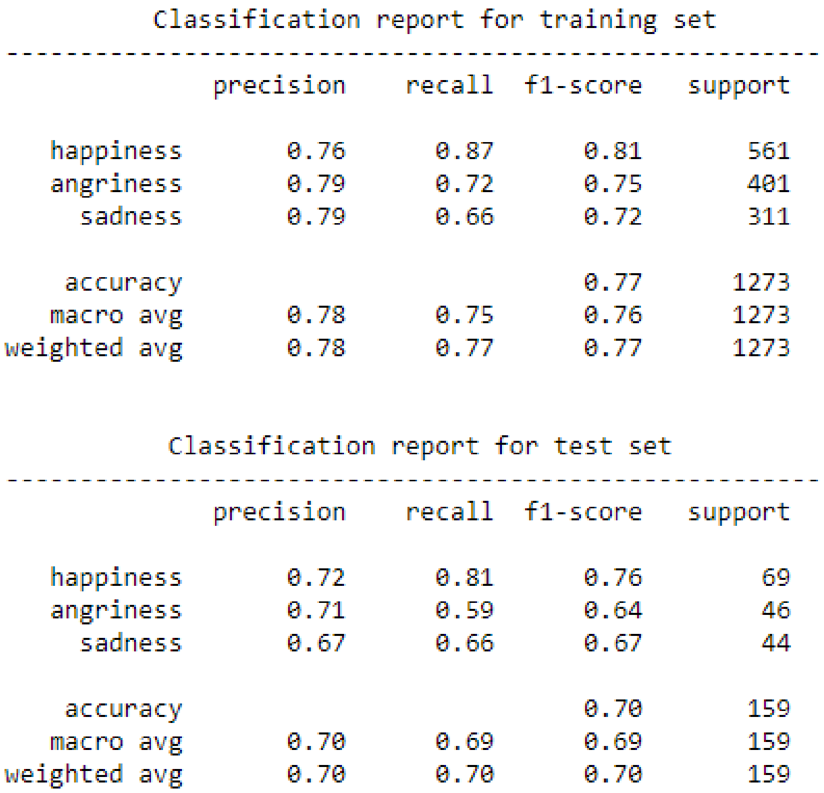
BoW MODELS



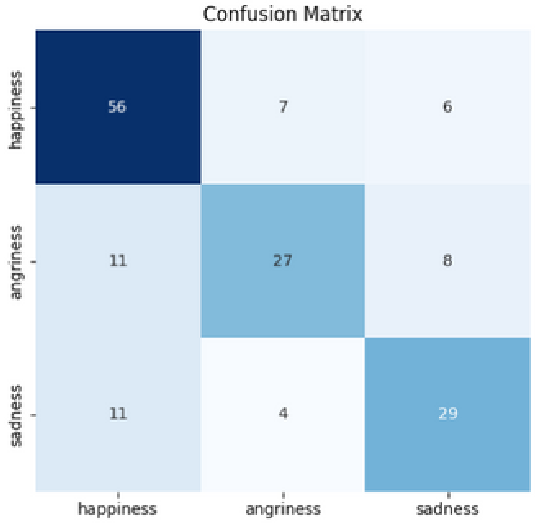
	Classifier	Training accuracy	Validation accuracy
8	XGBoost	0.882954	0.729560
3	Decision Tree	0.975648	0.723270
5	Random Forest	0.975648	0.723270
0	MultinomialNB	0.786332	0.704403
6	SGD Classifier	0.901021	0.704403
1	Logistic Regression	0.857031	0.691824
7	Ridge Classifier	0.853888	0.679245
9	AdaBoost	0.686567	0.672956
4	Linear SVM	0.868028	0.641509
2	KNN Classifier	0.449332	0.452830

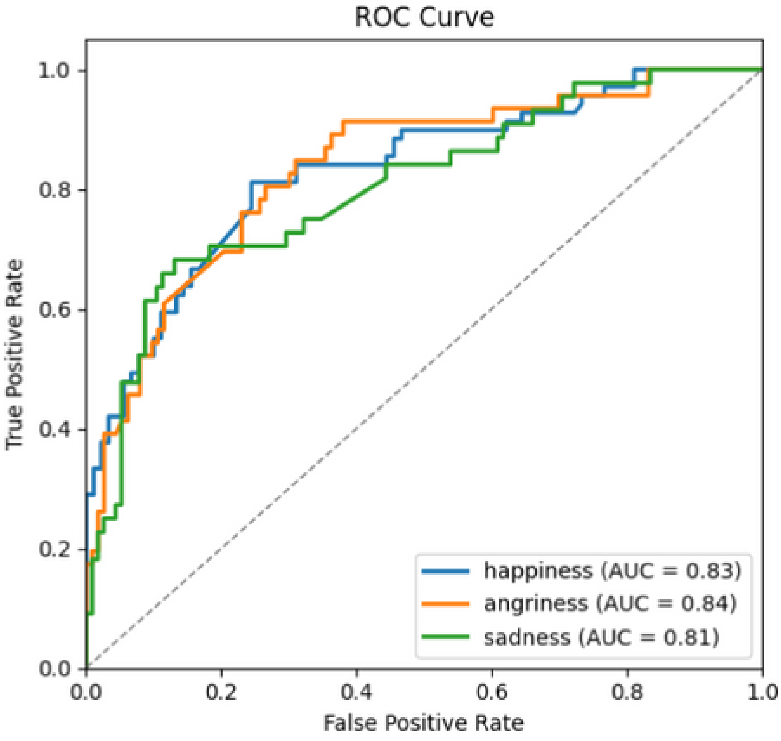
Comparing the results of TF-IDF vs. BoW, models fitted using CountVectorization performed better. Therefore, we will only perform hyperparameter tuning on the best-performing model, which is XGBoost using CountVectorizer.

BoW HYPERPARAMETER TUNING



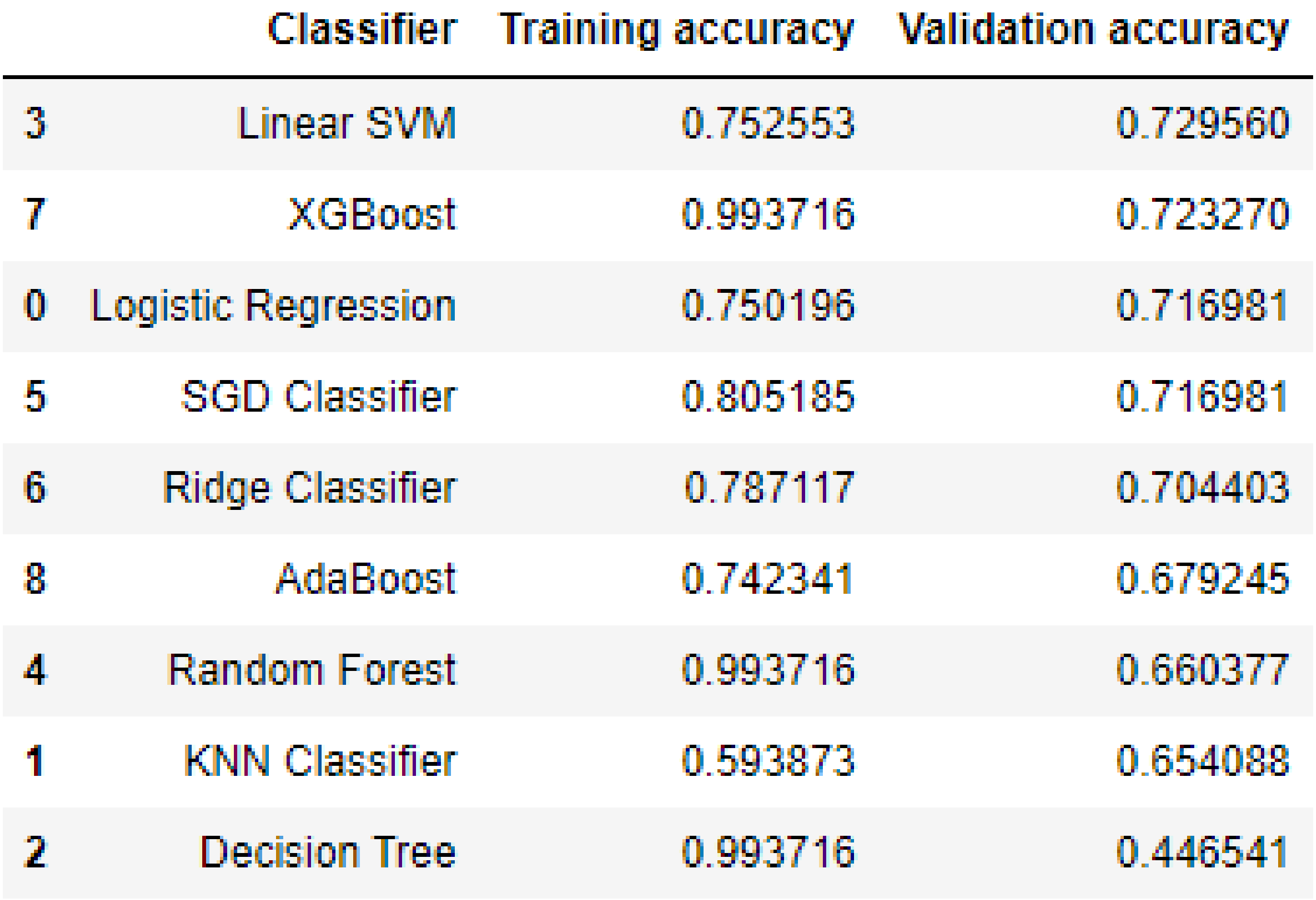
```
              Classification report for training set
-------------------------------------------------------
              precision    recall  f1-score   support

   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
```





The model performs well on the training set and reasonably on the test set, indicating good generalization. "Happiness" is easier to identify, while "angriness" and "sadness" are more challenging. The slight drop in test accuracy suggests some overfitting, but overall, the results are promising.

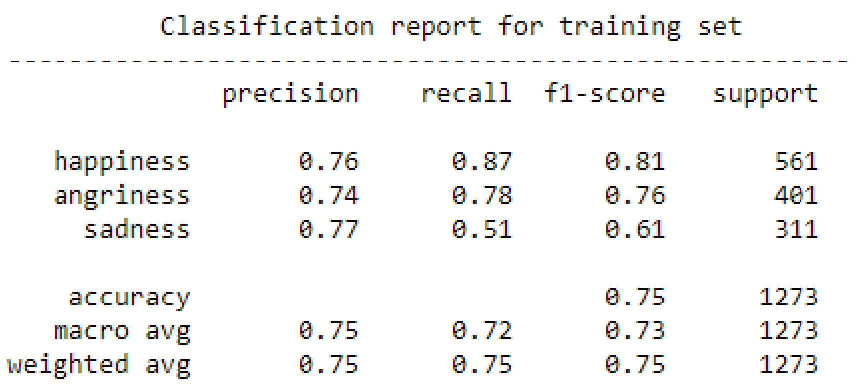
Word2Vec M0DELS



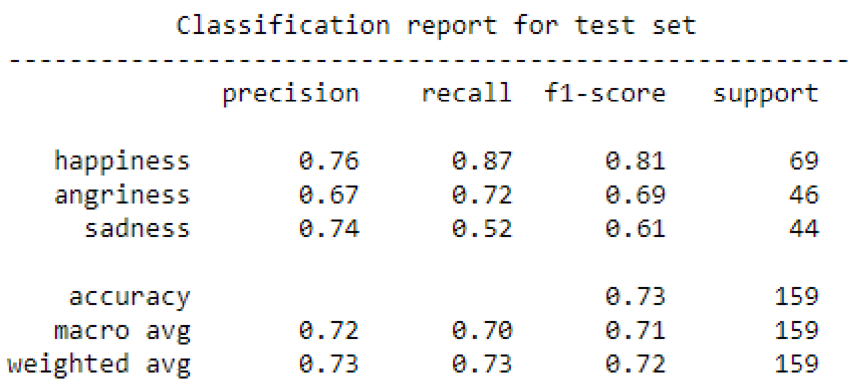
	Classifier	Training accuracy	Validation accuracy
3	Linear SVM	0.752553	0.729560
7	XGBoost	0.993716	0.723270
0	Logistic Regression	0.750196	0.716981
5	SGD Classifier	0.805185	0.716981
6	Ridge Classifier	0.787117	0.704403
8	AdaBoost	0.742341	0.679245
4	Random Forest	0.993716	0.660377
1	KNN Classifier	0.593873	0.654088
2	Decision Tree	0.993716	0.446541

Linear SVM performed better compared to others

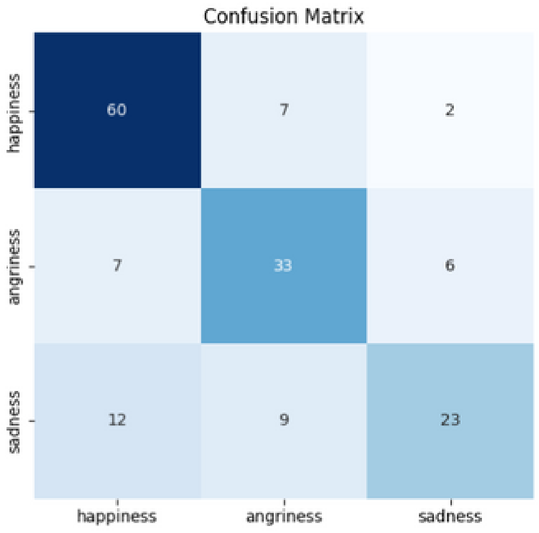
Word2Vec HYPERPARAMETER TUNING



```
            Classification report for training set
-------------------------------------------------------
              precision    recall  f1-score   support

   happiness       0.76      0.87      0.81       561
  angriness       0.74      0.78      0.76       401
     sadness       0.77      0.51      0.61       311

    accuracy                           0.75      1273
   macro avg       0.75      0.72      0.73      1273
weighted avg       0.75      0.75      0.75      1273
```

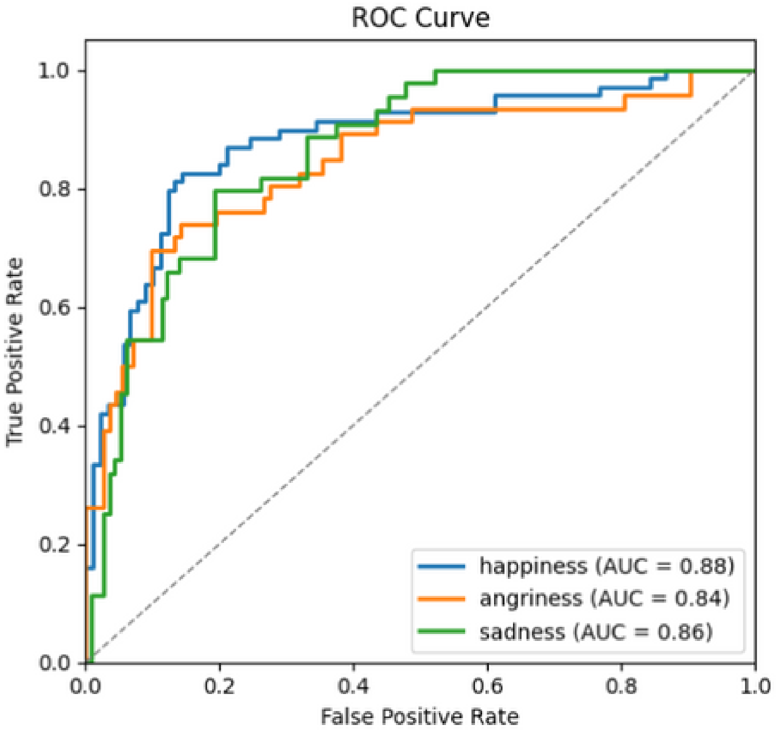


```
             Classification report for test set
-------------------------------------------------------
              precision    recall  f1-score   support

   happiness       0.76      0.87      0.81        69
  angriness       0.67      0.72      0.69        46
     sadness       0.74      0.52      0.61        44

    accuracy                           0.73       159
   macro avg       0.72      0.70      0.71       159
weighted avg       0.73      0.73      0.72       159
```



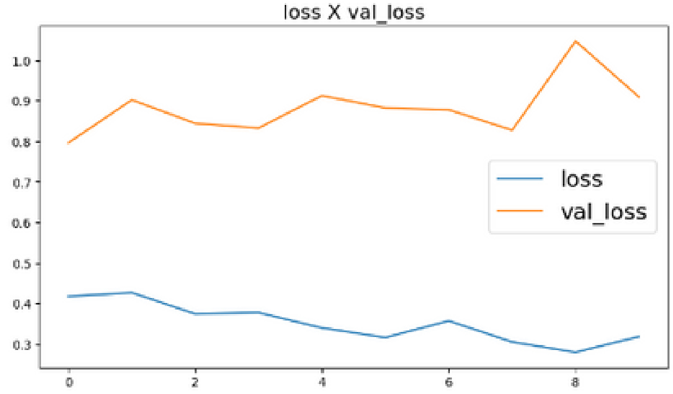
Confusion Matrix

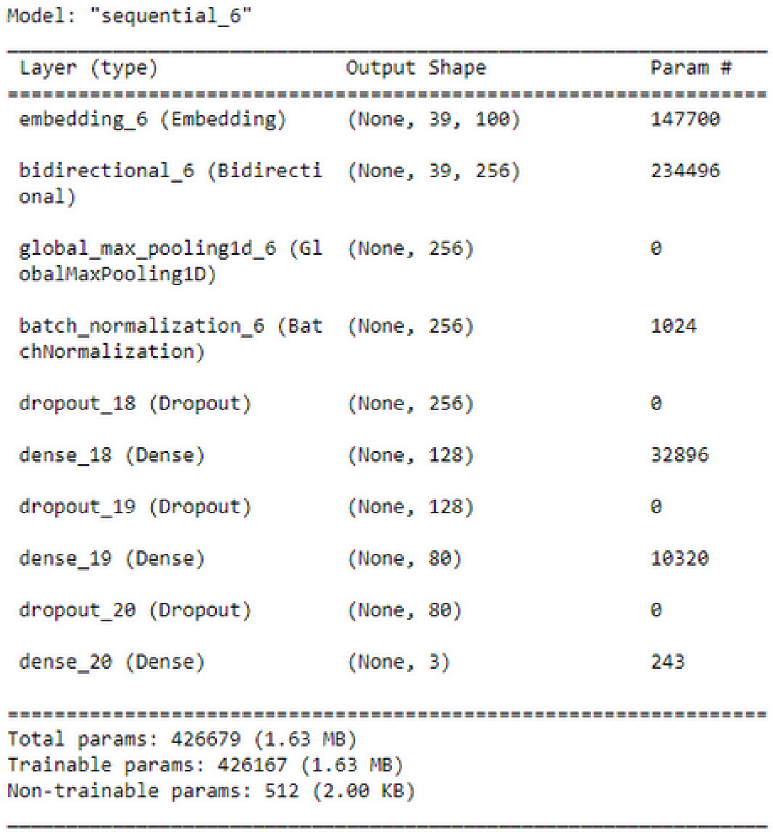
	happiness	angriness	sadness
happiness	60	7	2
angriness	7	33	6
sadness	12	9	23



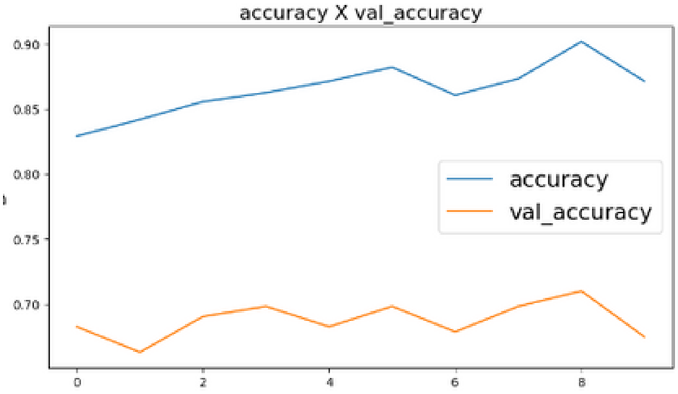
After doing hypertunning, the Linear SVM model is performing well, got accuracy: **73%, minimal overfitting.**

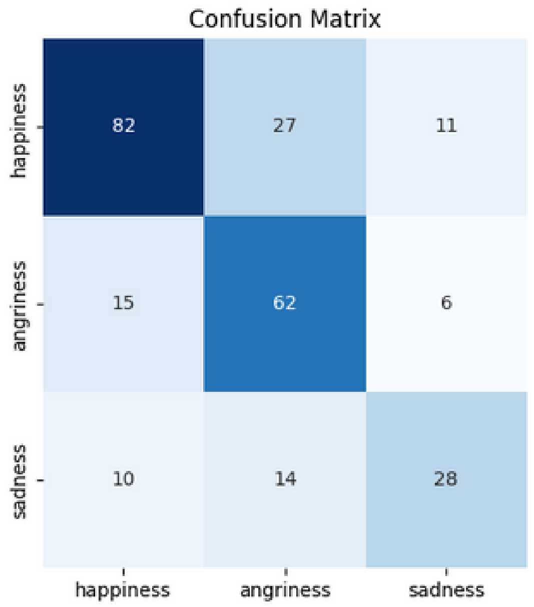
GloVe LSTM



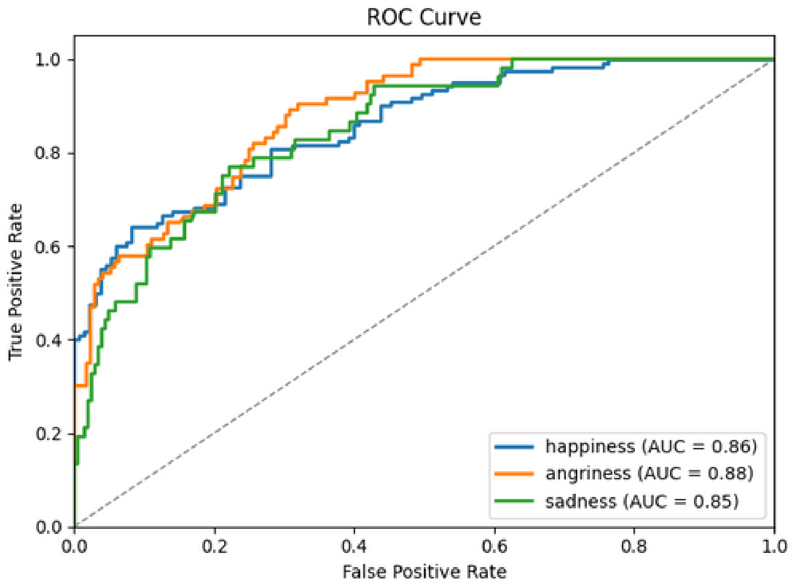


```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_6 (Embedding)     (None, 39, 100)           147700

 bidirectional_6 (Bidirecti  (None, 39, 256)           234496
 onal)

 global_max_pooling1d_6 (Gl  (None, 256)               0
 obalMaxPooling1D)

 batch_normalization_6 (Bat  (None, 256)               1024
 chNormalization)

 dropout_18 (Dropout)        (None, 256)               0

 dense_18 (Dense)            (None, 128)               32896

 dropout_19 (Dropout)        (None, 128)               0

 dense_19 (Dense)            (None, 80)                10320

 dropout_20 (Dropout)        (None, 80)                0

 dense_20 (Dense)            (None, 3)                 243

=================================================================
Total params: 426679 (1.63 MB)
Trainable params: 426167 (1.63 MB)
Non-trainable params: 512 (2.00 KB)
_________________________________________________________________
```



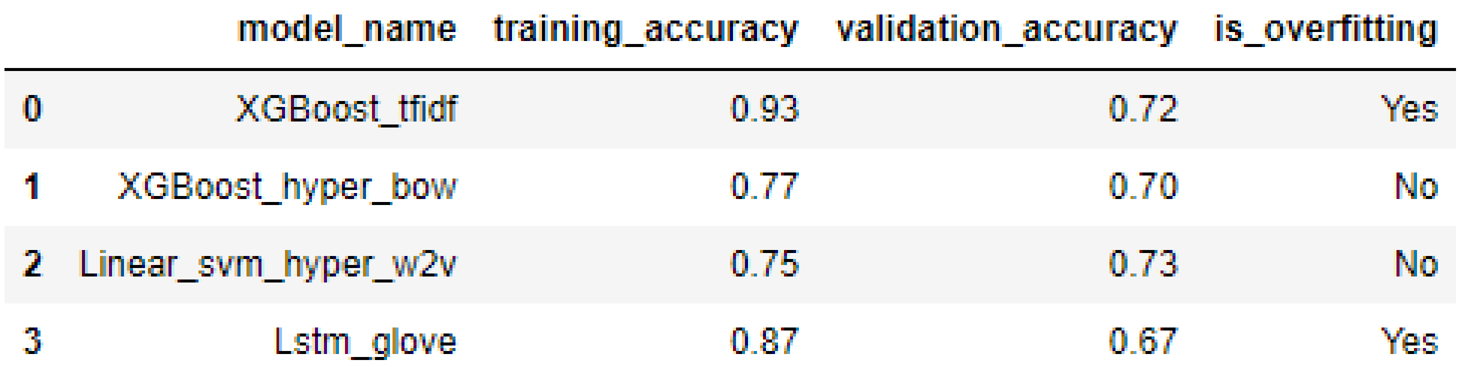


	happiness	angriness	sadness
happiness	82	27	11
angriness	15	62	6
sadness	10	14	28

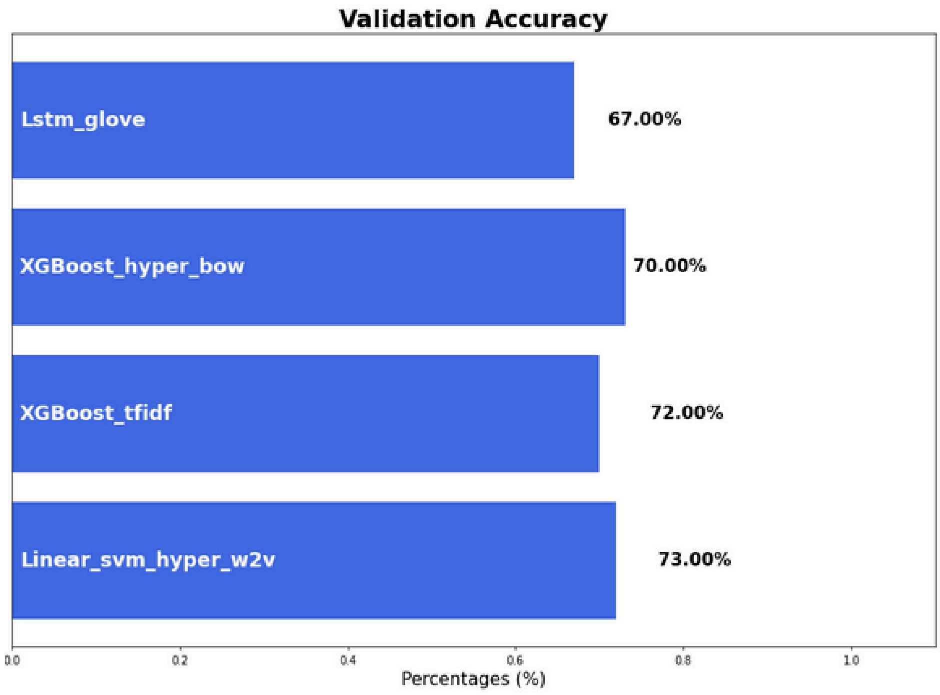


The performance is not that good, given our small dataset and the need for more data for optimal deep learning performance, we plan to collect more data to improve emotion prediction and pattern identification.

CONCLUSION



	model_name	training_accuracy	validation_accuracy	is_overfitting
0	XGBoost_tfidf	0.93	0.72	Yes
1	XGBoost_hyper_bow	0.77	0.70	No
2	Linear_svm_hyper_w2v	0.75	0.73	No
3	Lstm_glove	0.87	0.67	Yes



The Linear SVM with Word2Vec is the best model, showing minimal overfitting and a high validation accuracy of 73%, effectively predicting emotions on new data. Despite the challenge of Indian users typing native language slang in English, such as "Khatarnaak" meaning "angry," we achieved 73% accuracy.

THANK YOU

Thank you for reading! I hope your emotion is "happy" after going through this report! 😊