End Semester Presentation

TEAM NAME

Hardly Humans

Hardi Kadia

AU1841059

Tejas Chauhan

AU1841093

Khush Kalavadia

AU1841115

Jimil Desai

AU1841147

End Semester Presentation

Depression detection using Twitter data

Depression is a mood disorder that cannot be easily recognized. Due to the lockdown, there is a surge in the cases of depression.

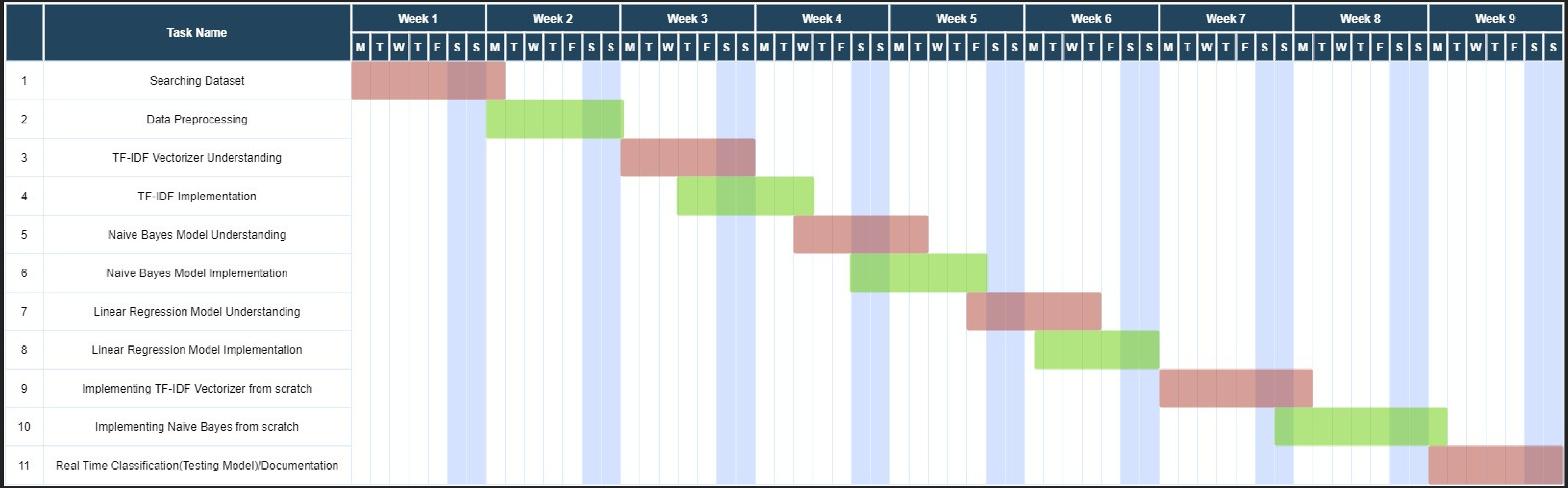
Twitter, unlike other social media platforms, gives emphasis on the interaction through "tweets" which majorly includes text. Analyzing the text is simpler than analyzing audio, video, or photo.

Problem Statement

We are using the Twitter dataset to identify whether the user's tweet is reflecting positive or negative sentiment.

We are using a dataset of 1.6 million tweets which are assigned a label of positive or negative sentiment. We are applying the machine learning model to the words that present in the tweets after preprocessing the same. This way we would be able to classify the sentiment of our tweet.

GANTT Chart



	Task Name	Week 1							Week 2							Week 3							Week 4							Week 5							Week 6							Week 7							Week 8							Week 9						
		M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
1	Searching Dataset	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥																																																								
2	Data Preprocessing								🟩	🟩	🟩	🟩	🟩	🟩	🟩																																																	
3	TF-IDF Vectorizer Understanding															🟥	🟥	🟥	🟥	🟥	🟥	🟥																																										
4	TF-IDF Implementation																		🟩	🟩	🟩	🟩	🟩	🟩	🟩																																							
5	Naive Bayes Model Understanding																								🟥	🟥	🟥	🟥	🟥	🟥	🟥																															
6	Naive Bayes Model Implementation																											🟩	🟩	🟩	🟩	🟩	🟩	🟩																														
7	Linear Regression Model Understanding																																		🟥	🟥	🟥	🟥	🟥	🟥	🟥																								
8	Linear Regression Model Implementation																																					🟩	🟩	🟩	🟩	🟩	🟩	🟩																					
9	Implementing TF-IDF Vectorizer from scratch																																												🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥													
10	Implementing Naive Bayes from scratch																																																		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
11	Real Time Classification(Testing Model)/Documentation																																																										🟥	🟥	🟥	🟥	🟥	🟥	🟥

Existing body of work

- A lot of research and time has gone into Sentiment Analysis and it is handled as NLP task at many levels of granularity. Sentiment analysis traditionally has been considered as a 3 layered approach. Starting from document-level classification, it has been handled at a sentence level and more recently at an abstract/phrase level.
- As far as the sentiment classification techniques are concerned there have been research going around mainly 3 topics:
    - Machine Learning
    - Lexicon Based Approach
    - Combination/Hybrid Approach

Our approach

Dataset identification

Using Twitter API to seek tweets

Using dictionary to find sentiment of the tweets

Using dataset which had 1.6 million tweets with sentiment labels

Preprocessing data

Removal of URL, tags, non-alphabetic characters, etc

Stopwords removal

Word tokenization

Lemmatization

Logistic Regression

Preparing optimisation function logistic regression

Achieving parameters from optimization function

Finding probability for classification using weight of tweets and parameters

Our approach

TF-IDF Vectoriser

Preparing the vocabulary using the training dataset

Finding term frequency based on the frequency of words in each tweet

Finding inverse document frequency by counting the tweets in which a word was present

Preparing tf-idf sparse matrix

Naive Bayes Classification

Calculate prior probability of class label

Likelihood Probability of each attribute for each class

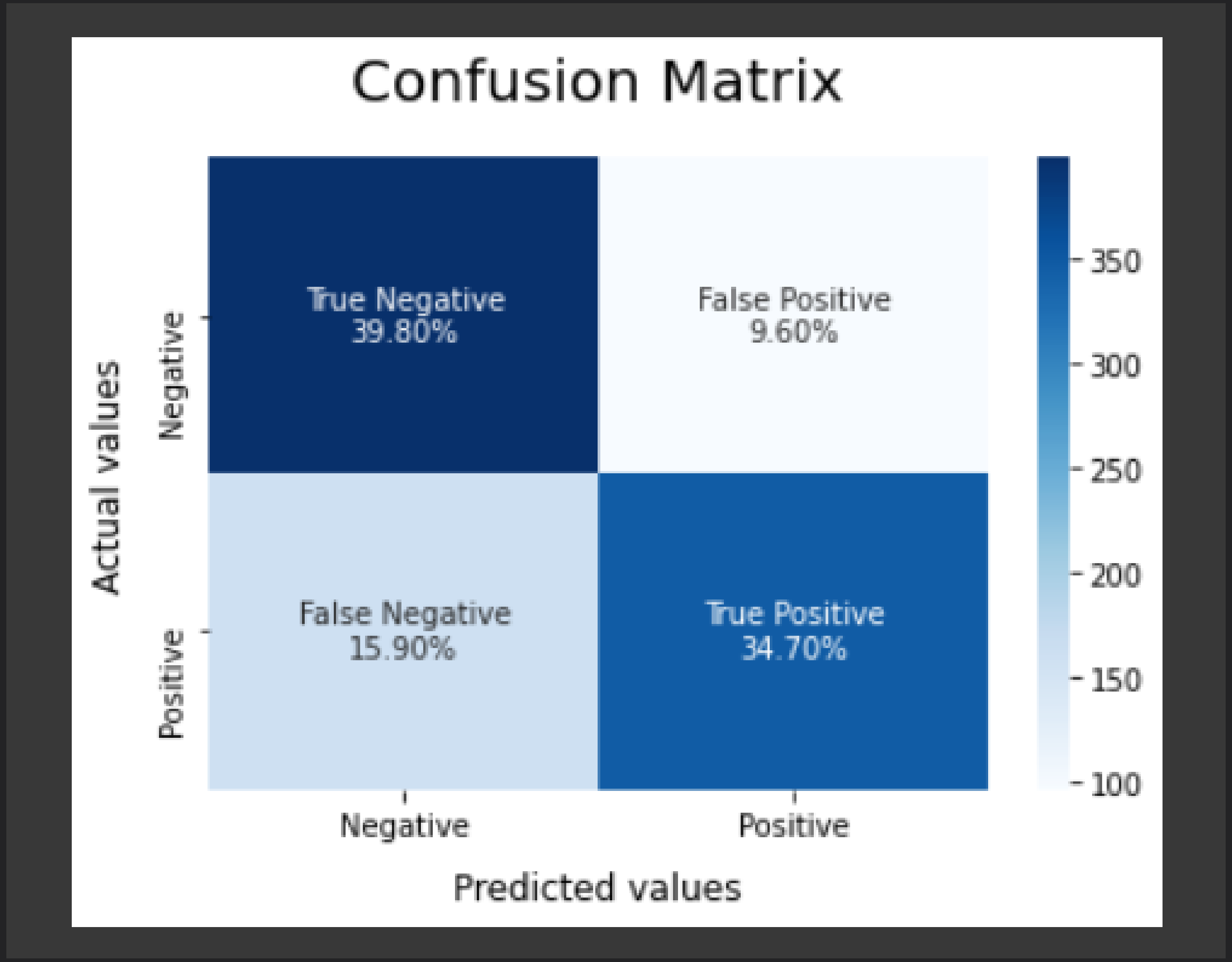
Posterior probability using Bayes Formula

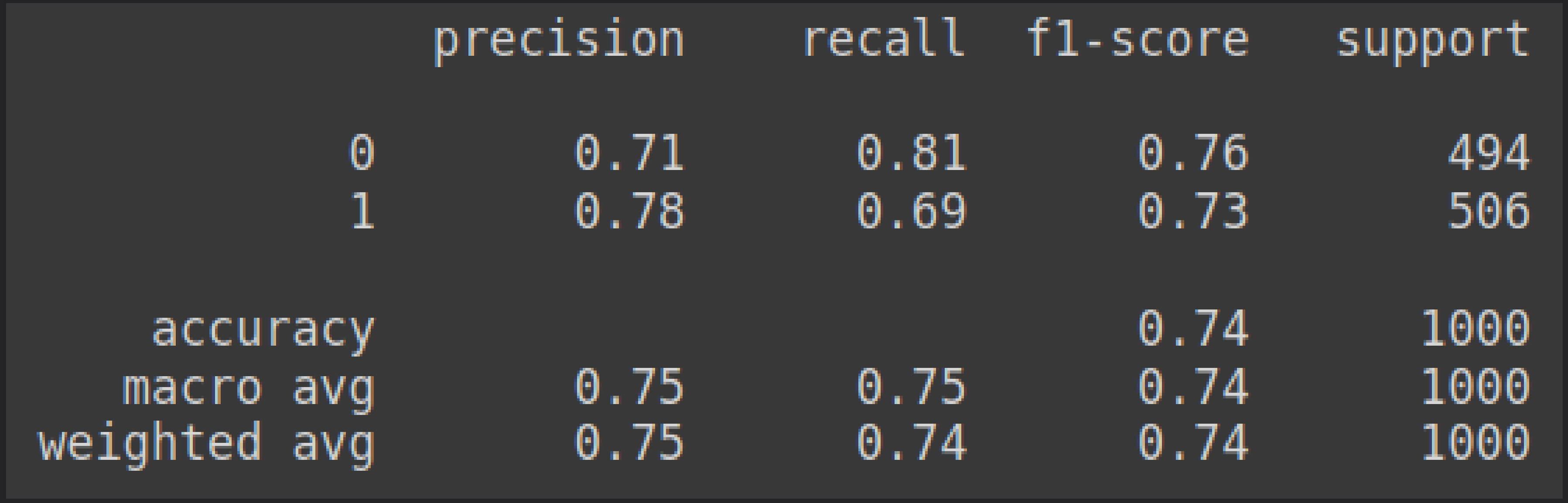
Highest probability class would be our prediction

Result 1

TF-IDF scratch implimentation

TF-IDF was applied from scratch to build up the vocabulary and calculate the weight of each word from the training set. This tf-idf matrix was passed on to the naive bayesian classifier to predict the sentiments. This algorithm was built considering the value of ngram to be 1.



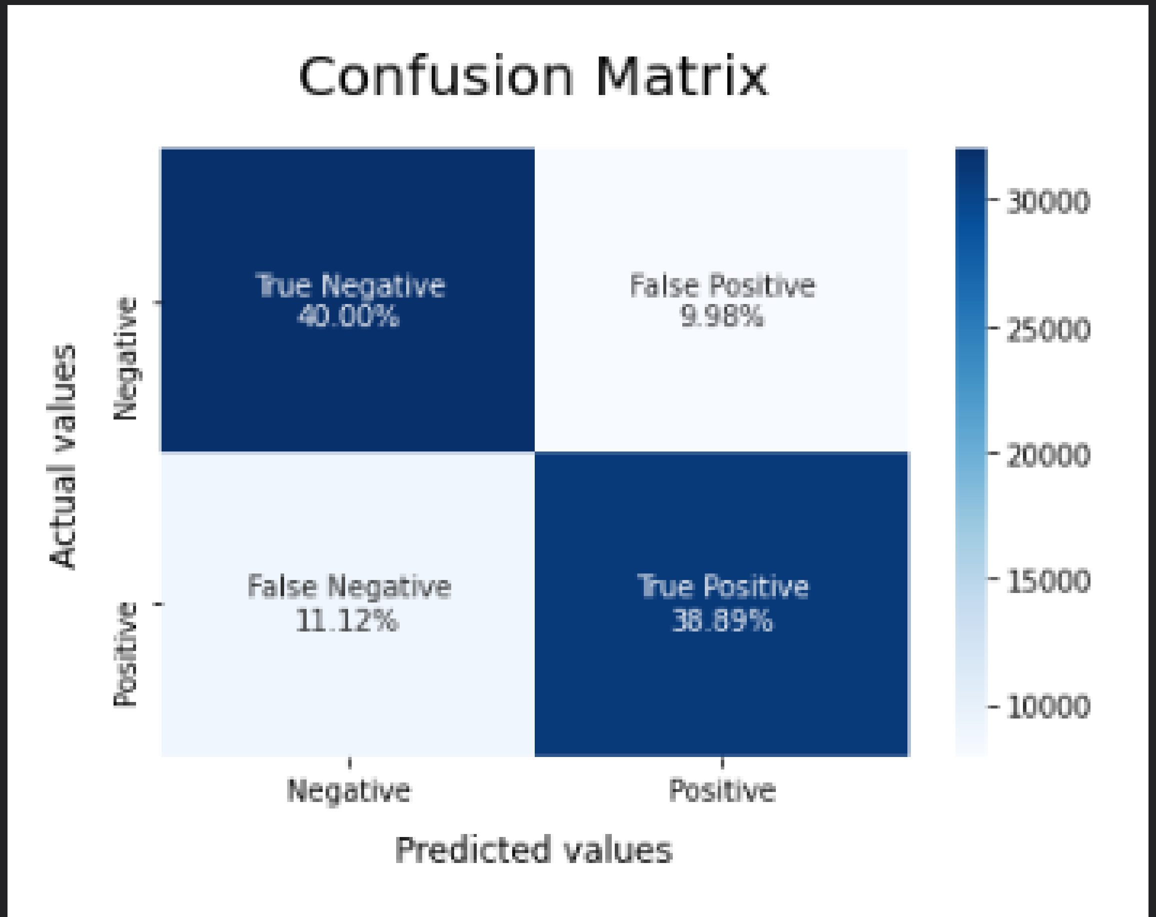


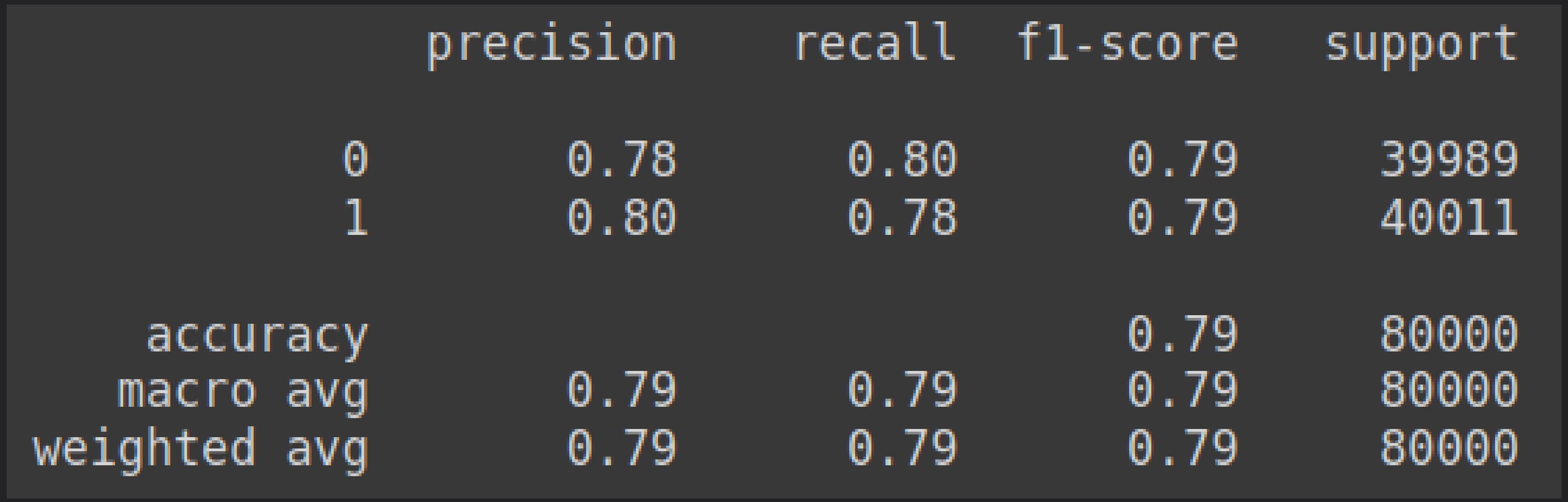
```
              precision    recall  f1-score   support

           0       0.71      0.81      0.76       494
           1       0.78      0.69      0.73       506

    accuracy                           0.74      1000
   macro avg       0.75      0.75      0.74      1000
weighted avg       0.75      0.74      0.74      1000
```

Depression detection using Twitter data





```
              precision    recall  f1-score   support

           0       0.78      0.80      0.79     39989
           1       0.80      0.78      0.79     40011

    accuracy                           0.79     80000
   macro avg       0.79      0.79      0.79     80000
weighted avg       0.79      0.79      0.79     80000
```

Result 2

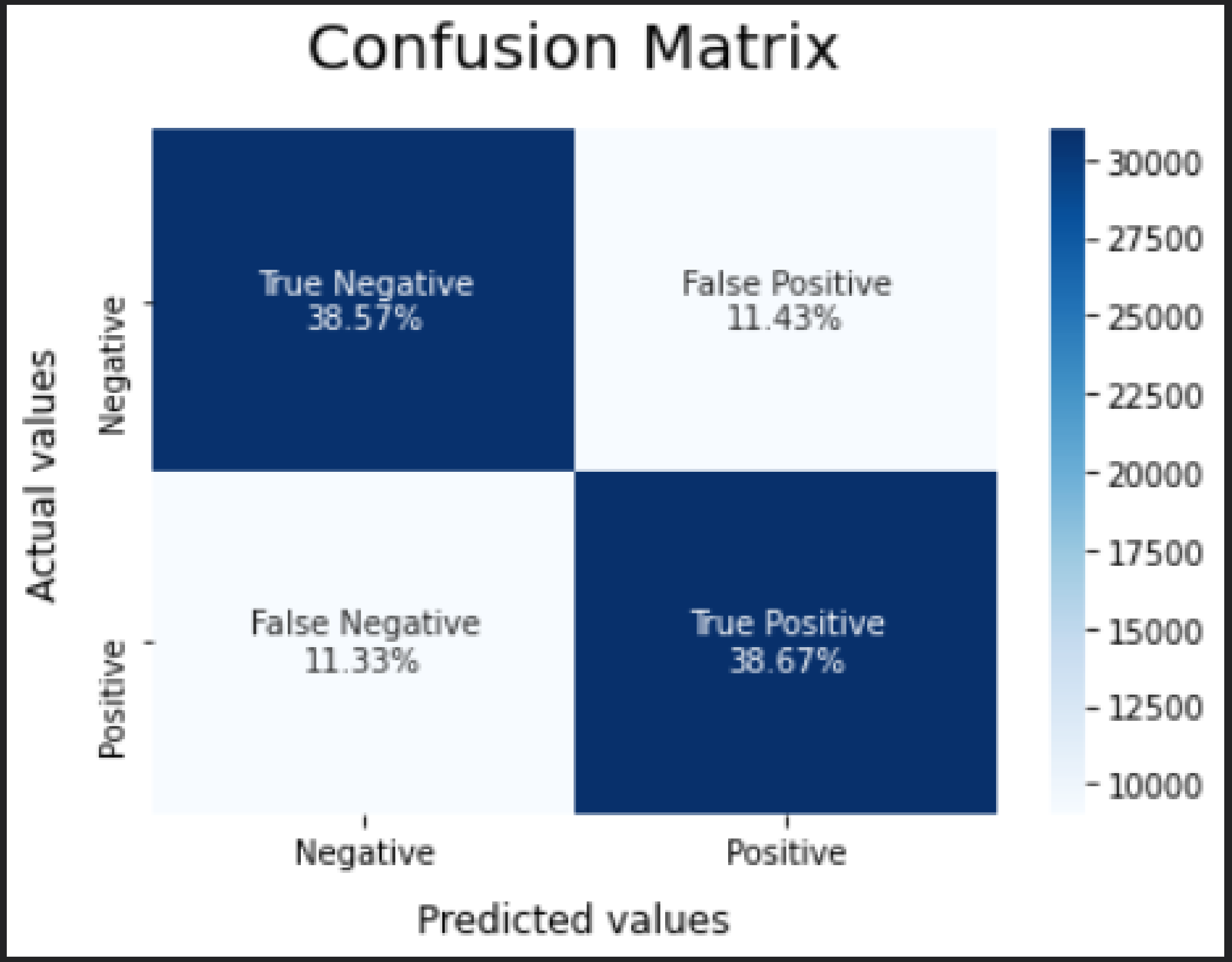
Multinomial Naive Bayes - *Inbuilt*

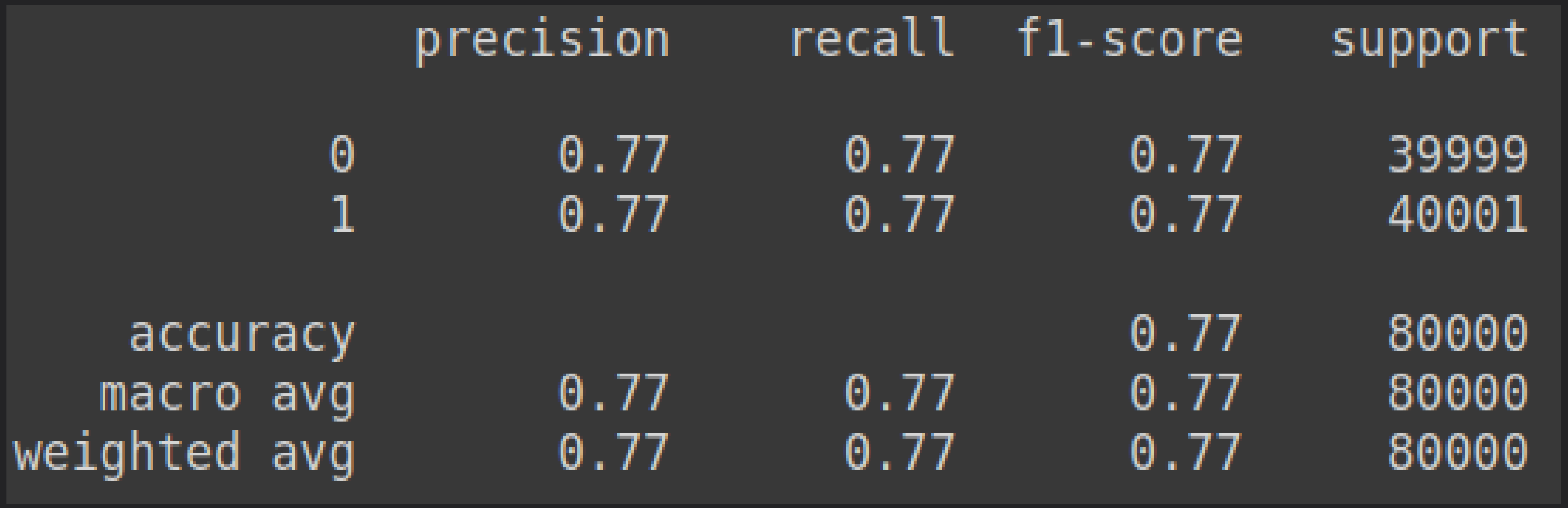
Multinomial NB Classifier was applied to the training data and was tested on the testing data. The multinomial NB model performs decently well giving us an overall accuracy of 79% while classifying the sentiment of a tweet. Training accuracy was 82.6% and testing accuracy was 78.9%. Both the accuracy are close to each other, which is a good sign indicating that the model is not overfitting. The confusion matrix of Multinomial NB is shown in the figure.

Result 3

Multinomial Naive Bayes - *Scratch*

Multinomial Naive Bayesian Classifier was applied from scratch on the training data and tested on the testing data. This model performs decently well giving us the overall accuracy of 77.24%. This is accuracy is close to that of the inbuilt Naive Bayesian classifier indicating that our model behaves similar to the inbuilt sklearn Naive Bayesian Classifier.



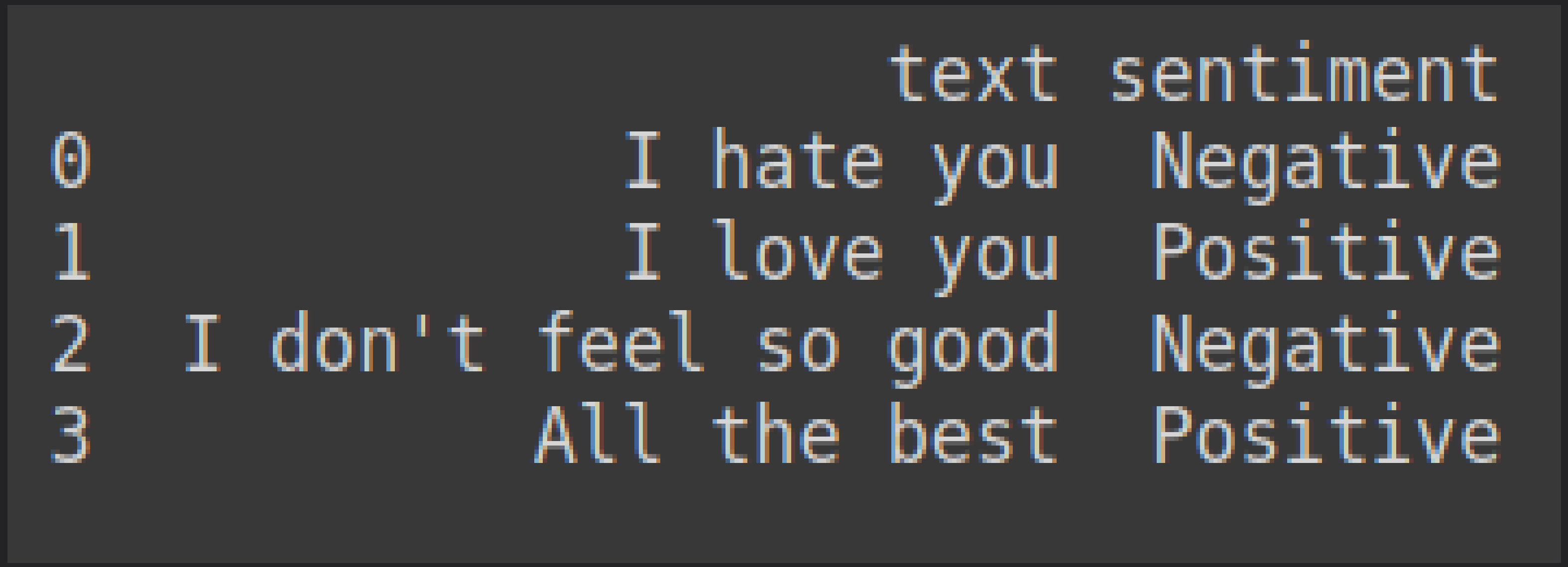


```
              precision    recall  f1-score   support

           0       0.77      0.77      0.77     39999
           1       0.77      0.77      0.77     40001

    accuracy                           0.77     80000
   macro avg       0.77      0.77      0.77     80000
weighted avg       0.77      0.77      0.77     80000
```

Result 4



```
                   text  sentiment
0            I hate you   Negative
1            I love you   Positive
2  I don't feel so good   Negative
3          All the best   Positive
```

Real time Classification

We have used pickle library to save our vectorizer and model to predict sentiments of random tweets in real time. These random tweets are first pre-processed, vectorized and fed into the model at real time to get the sentiment.

Conclusions

During the course of the project, we have successfully implemented the Naive Bayes Classifier and TF-IDF vectorizer from scratch on the twitter dataset to predict sentiments of unseen data. The Naive Bayesian Classifier gave us the accuracy of 77.24%. SKLearn’s Naive Bayes model was also applied on the twitter dataset, which gave us the accuracy of 79%. At last we have done real time classification to test our model.

Role of each group member

Task	Hardi	Tejas	Khush	Jimil
TF-IDF from Scratch	✓		✓	
Naive Bayes from scratch	✓	✓		✓
Real time classification			✓	✓
Report Writing		✓		✓
Presentation	✓		✓	

References

- Depression detection from social network data using machine learning techniques - link
- Identifying Depression on Twitter - link
- A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms - link
- Twitter Sentimental Analysis Using Naive Bayes Classifier(Process Explanation) - link
- Speech and Language Processing. Daniel Jurafsky & James H. Martin - link

Thank you!