BEST PREPARATION AND METHOD :

Best preparation was using the Word2Vec after the tweets were tokenized (using NLTK trained model), removed of mentions, hashtags and other symbols common in tweets. Then the corresponding values for each word was weighted using TFID vectorization and averaged over. This preparation was employed to give unbalanced, more natural weighing for each word in the sentences based on context without the noise of symbols or stopwords of the tweets . Then TruncatedSVD is applied to reduce the number of components to a more manageable number.

The best method was a Neural network, with variations in metrics, epochs, optimizers and more. Metrics specifically for F1 score were used to increase that parameter. Dropouts were used to reduce overfitting, which was visible in the epoch logs previously.

OTHER PREPS ANALYZED :

1)
First stopwords from NLTK corpus and other symbols were removed from the tweets to reduce noise. Later general TfidVectorizer with a weighing system based on frequency than context was used to vectorize the tweets. Due to the less pre trained data of Word2Vec, this vectorization provided a slightly less attractive accuracy and F1 score .

OTHER METHODS ANALYZED :

1)
SVC (SVM model for Classification) was tested for different kernels and C values.

2)
Logistic Regression was tested for different C values

3)
Neural networks with different number of layers and Regularizations were tested.