**Name: Abrar Syed**
**Student_ID - 3035529743**

**#For original clean reviews with unigram setting and max features 1000**

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=1000)
```

```
The training accuracy is:  0.9999
 The validation accuracy is:  0.8216
```

**# For original clean reviews with bigram setting and max features 1000**

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=2, max_features=1000)
```

```
The training accuracy is:  0.99995
 The validation accuracy is:  0.8208
```

## Question 1

**# For lemmatized reviews with unigram setting and max features 1000**

```
 original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=1000)
```

```
The training accuracy is:  1.0
 The validation accuracy is:  0.82
```

**#For stemmed reviews with unigram setting and max features 1000**

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=True)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=1000)
The training accuracy is:  1.0
 The validation accuracy is:  0.82
```

**Observation :**

There is a slight change in the training accuracy in lemmatized and stemmed reviews in comparison to the original reviews  but the validation accuracy almost remains the same for both when compared to the original reviews.

## Question -2

```
#For lemmatized reviews with bigram setting and max features 1000
original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=2, max_features=1000)
```

```
 The training accuracy is:  1.0
 The validation accuracy is:  0.8166
```

```
#For stemmed reviews with bigram setting and max features 1000
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=True)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=2, max_features=1000)
```

```
The training accuracy is:  0.99995
 The validation accuracy is:  0.8242
```

**Observation-**

The training accuracy remains the same when comparing the stemmed and original reviews. But the validation accuracy changes a bit by 2 or 3 decimal places in lemmatized and stemmed reviews in comparison with original reviews.

## Question 3

```
# For lemmatized reviews with unigram setting and max features 10
 original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=10)
```

```
The training accuracy is:  0.87145
 The validation accuracy is:  0.5594
```

```
# For lemmatized reviews with unigram setting and max features 100
 original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=100)
```

```
The training accuracy is:  0.99995
 The validation accuracy is:  0.7228
```

```
# For lemmatized reviews with unigram setting and max features 1000
 original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=1000)
The training accuracy is:  1.0
 The validation accuracy is:  0.8256
```

```
# For lemmatized reviews with unigram setting and max features 5000
 original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews, y=train["sentiment"],
ngram=1, max_features=5000)
```

```
The training accuracy is:  1.0
 The validation accuracy is:  0.8394
```

## Observation-

The training accuracy keeps increasing as the max_features increases but later changes by a bit for max _ features = 1000 and 5000. The validation accuracy increases for max _features =10 and 100 and then increases slightly for max_features = 1000 and 5000.