No 11

To load the dataset, I used pandas. The code for loading the data is as follows:

```
train_df = pd.read_csv('train.tsv', sep='\t')
x_train, y_train = train_df['Phrase'], train_df['Sentiment']
test_df = pd.read_csv('test.tsv', sep='\t')
label = pd.read_csv('sampleSubmission.csv')
x_test, y_test = test_df['Phrase'], label['Sentiment']
```

SampleSubmission.csv contains the label of the test data.

After loading the data, I converted train and test input into vector using CountVectorizer by sklearn. The commands are as follows:

```
vec = CountVectorizer(stop_words='english')
x_train_vec = vec.fit_transform(x_train).toarray()
x_test_vec = vec.transform(x_test).toarray()
```

The model that I used to solve this problem is Multinomial Naïve Bayes. This is because it is commonly used in text classification where the data are represented as word vector counts. I implement this model by using sklearn library as follows:

```
from sklearn.naive_bayes import MultinomialNB
gnb = MultinomialNB()
y_pred = gnb.fit(x_train_vec, y_train).predict(x_test_vec)
```

The result was as follows:

```
Number of mislabeled points out of a total 66292 points : 22577
Accuracy:  0.6594310022325469
```