fuse | machines

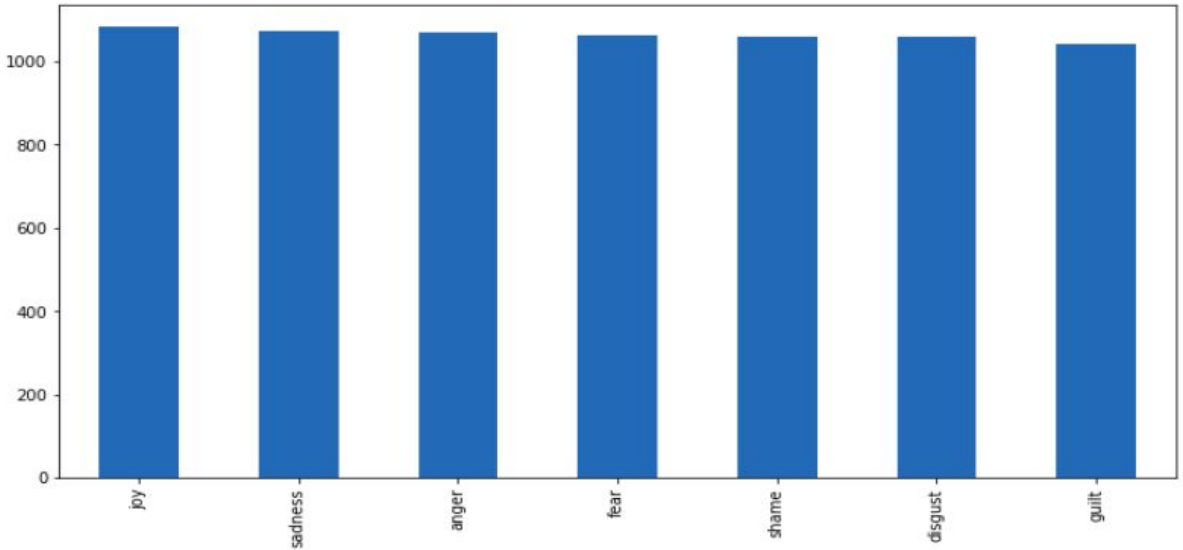
Emotion Classification

Progress Report I

2020.05.14

Aditi Gajurel

Bar plot of the output labels (emotions)



Out of total 7446 rows , there are 7 emotions which are balanced in nature as shown in the bar graph. This is the multiclass classification problem.

Baseline MOdel: Data Preprocessing

Text Preprocessing (CountVectorizer)

- Removal of Punctuations.
- Removal of Stopwords in english

Bigram Representation

- A more sophisticated data representation model is the bigram model where occurrences depend on a sequence of two words rather than an individual one.

Document Term Matrix

- mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

Baseline Model: Stochastic Gradient Descent Classifier

- While gradient descend is powerful, it can be prohibitively expensive when the dataset is extremely large because every single data point needs to be processed.

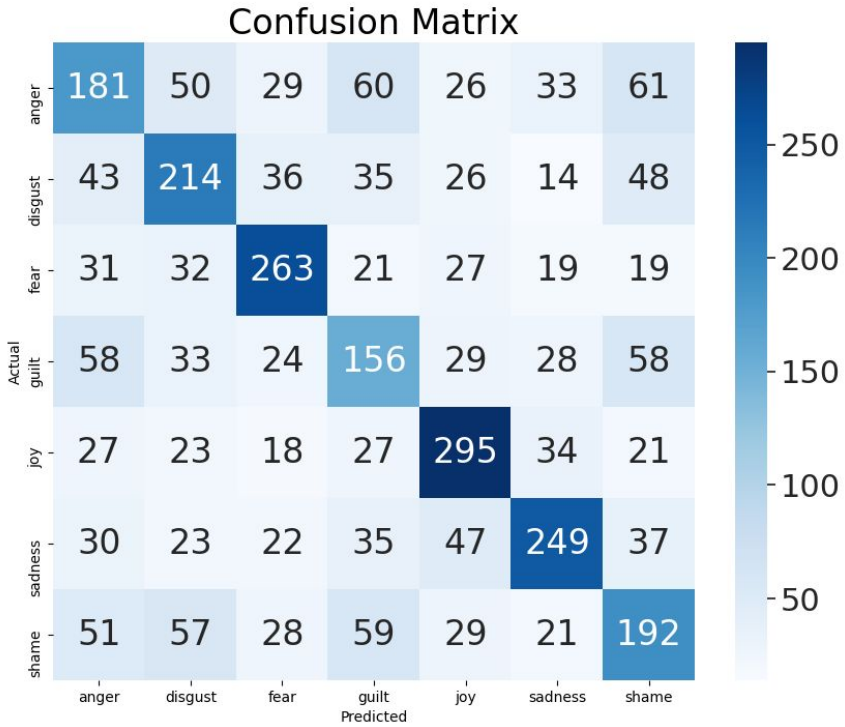
- However, it turns out when the data is large, rather than the entire dataset, SGD algorithm performs just as good with a small random subset of the original data.

This is the central idea of Stochastic SGD and particularly handy for the text data since corpus are often humongous.

Reference for SGD Classifier.
https://en.wikipedia.org/wiki/Stochastic_gradient_descent

.

Baseline Model: Evaluation



Accuracy : 53 %

Confusion Matrix with actual and prediction labels by SGDClassifier.

Further Improvements

- Preprocess the data.
- Try out different classifiers.
- Try different n gram models for document term matrix.
- Increase the accuracy and generalisation of the classifier.