Drug Review dataset exploration report

1. **Libraries**
    - Keras
    - Pandas
    - Numpy
    - Nltk
    - Seaborn
    - Matplotlib
    - Sklearn

2. **Processing text comment review**
    - Split overall rating into 3 categories (Positive/Neutral/Negative) to improve performances and the accuracy:
        - Positive – rating from 7-10
        - Negative – rating from 1-4
        - Neutral – rating from 5-6

```
a.head(20)
```

	rating	class_label
0	4	Negative
1	1	Negative
2	10	Positive
3	3	Negative
4	2	Negative
5	1	Negative
6	9	Positive
7	10	Positive
8	10	Positive
9	1	Negative
10	7	Positive
11	8	Positive
12	8	Positive
13	9	Positive
14	4	Negative
15	8	Positive
16	6	Neutral
17	1	Negative
18	8	Positive
19	6	Neutral

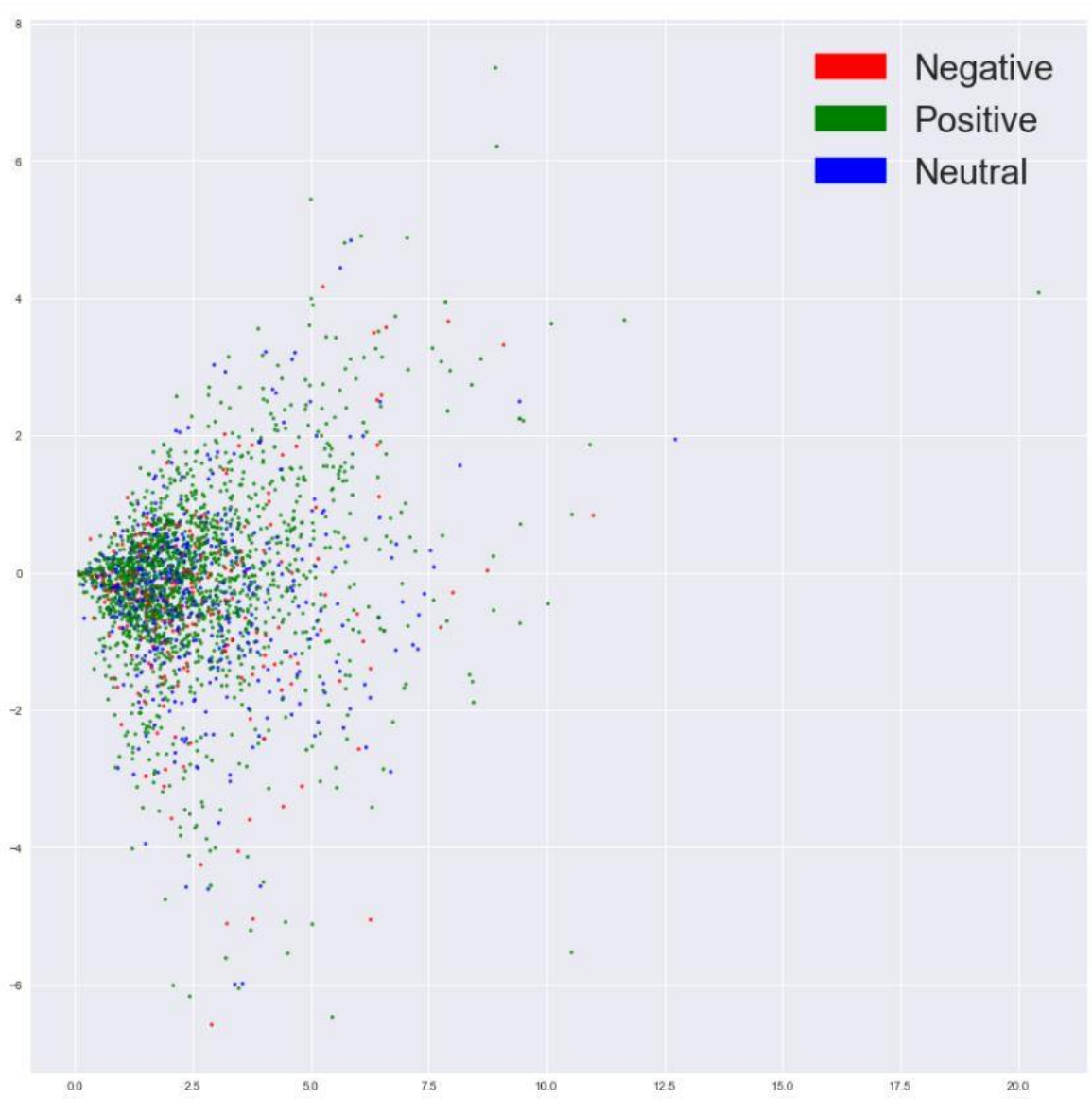
    - Concatenate 3 columns Reviews: commentsReview, benefitsReview and sideEffectsReview

```
In [9]: df_train.commentsReview[0]

Out[9]: 'monitor blood pressure , weight and asses for resolution of fluid slowed the progression of left ventricular dysfunction into
overt heart failure \r\r\nalone or with other agents in the managment of hypertension \r\r\nmangagement of congestive heart fai
lur cough, hypotension , proteinuria, impotence , renal failure , angina pectoris , tachycardia , eosinophilic pneumonitis, tas
tes disturbances , anusease anorecia , weakness fatigue insominca weakness'

In [10]: df_train.commentsReview[1]

Out[10]: "I Hate This Birth Control, I Would Not Suggest This To Anyone. Although this type of birth control has more cons than pros, it
did help with my cramps. It's also effective with the prevention of pregnancy. (Along with use of condoms as well) Heavy Cycle,
Cramps, Hot Flashes, Fatigue, Long Lasting Cycles. It's only been 5 1/2 months, but i'm concidering changing to a different bc.
This is my first time using any kind of bc, unfortunately due to the constant hassel, i'm not happy with the results."
```

- Preprocessing steps:
    - Lowercase all the characters to avoid the different between uppercase and lowercase
    - Remove all punctuations, special characters and numbers
    - Remove all stopwords include common English stopwords and drugname.
    - Remove some top frequent words
- Tokenize text of commentsReview to apply in Bag of Words implementation and TF-IDF implementation

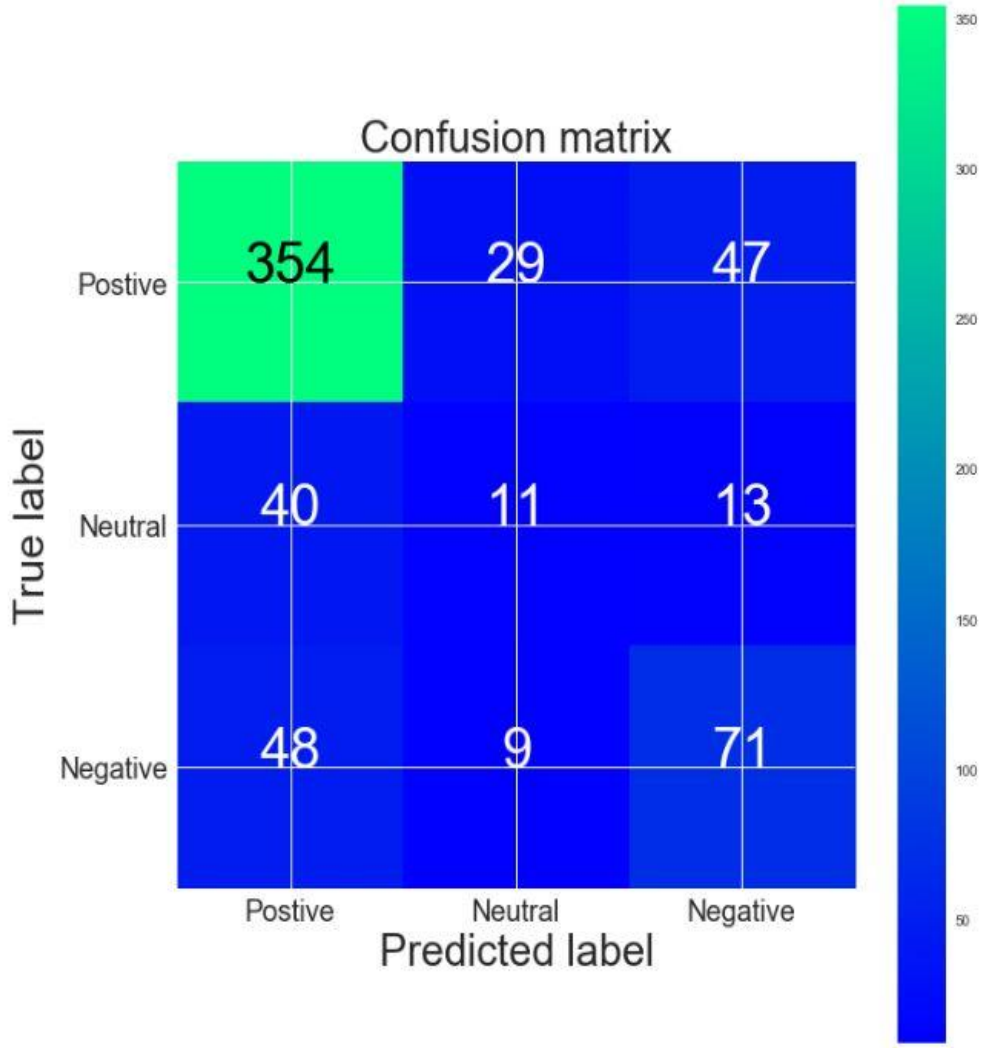
3. Apply Bag of Word, embedding visualization, importance words

- Embedding visualization plot after applying BoW



The embedding plot don’t look too separated but there are many outliers to be removed. This is a small dataset so I do not remove them.

- Plot confusion matrix

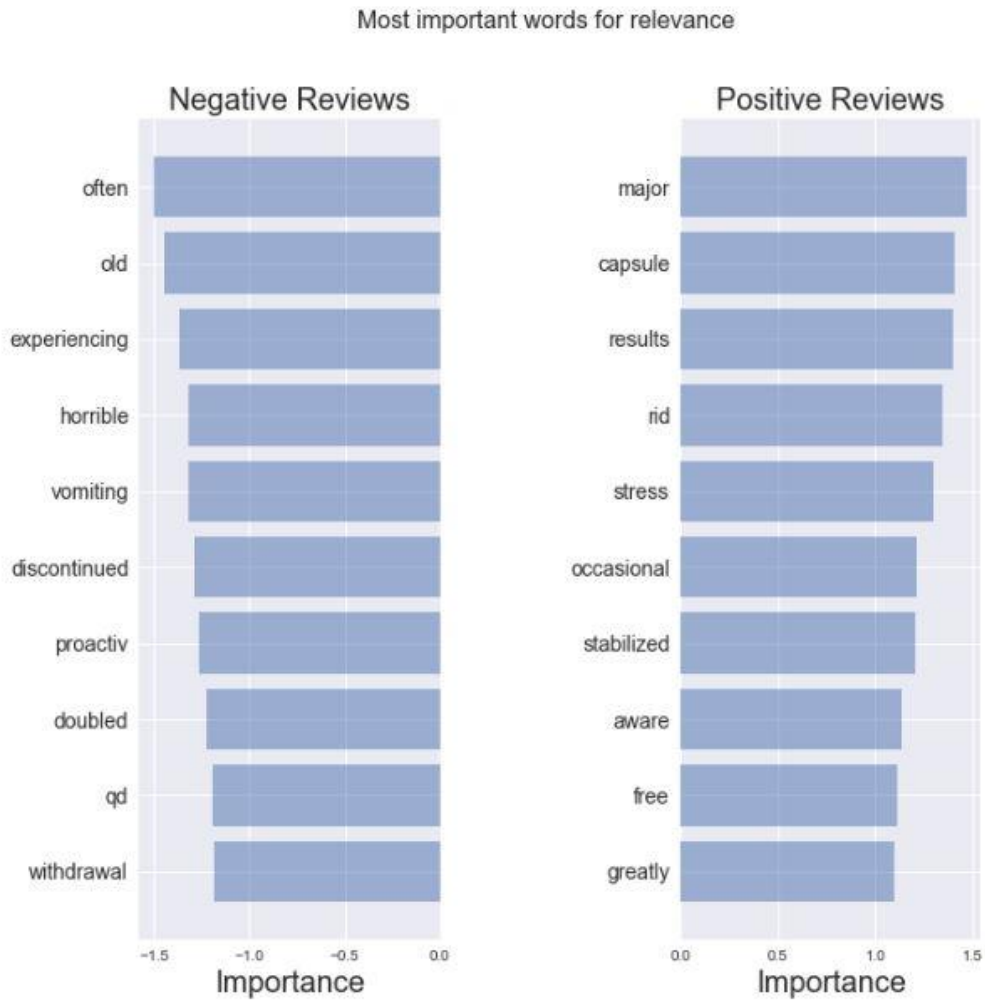



```
Test Data Value Counts:
 1    430
0    128
2     64
Name: 0, dtype: int64
```

The classifier predicted class 3 - Neutral is not so good but other 2 are ok ...

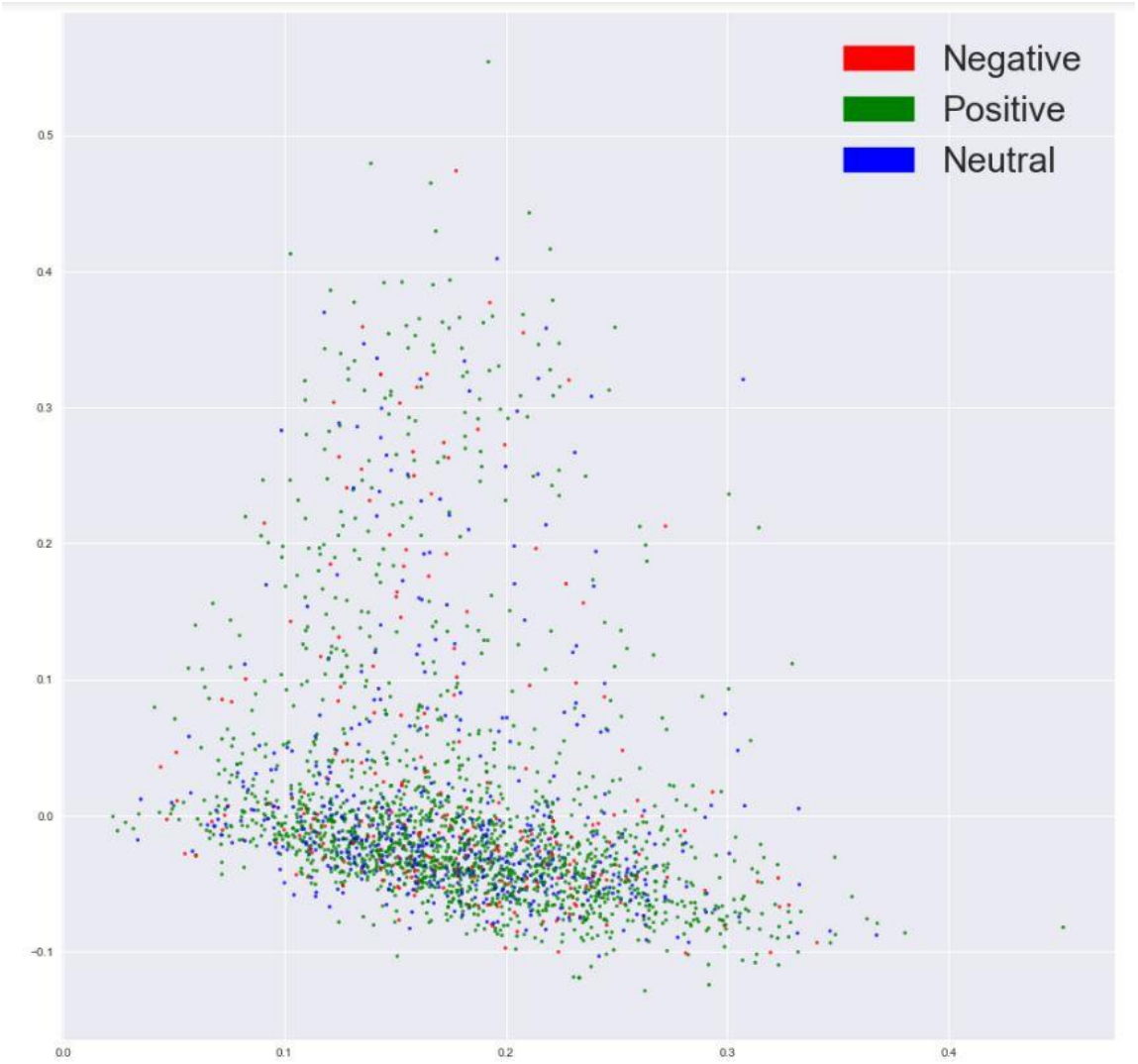
The classifier performed very bad at Neutral class in validation set but others two are ok.

- Plot of negative and positive importance words of train set



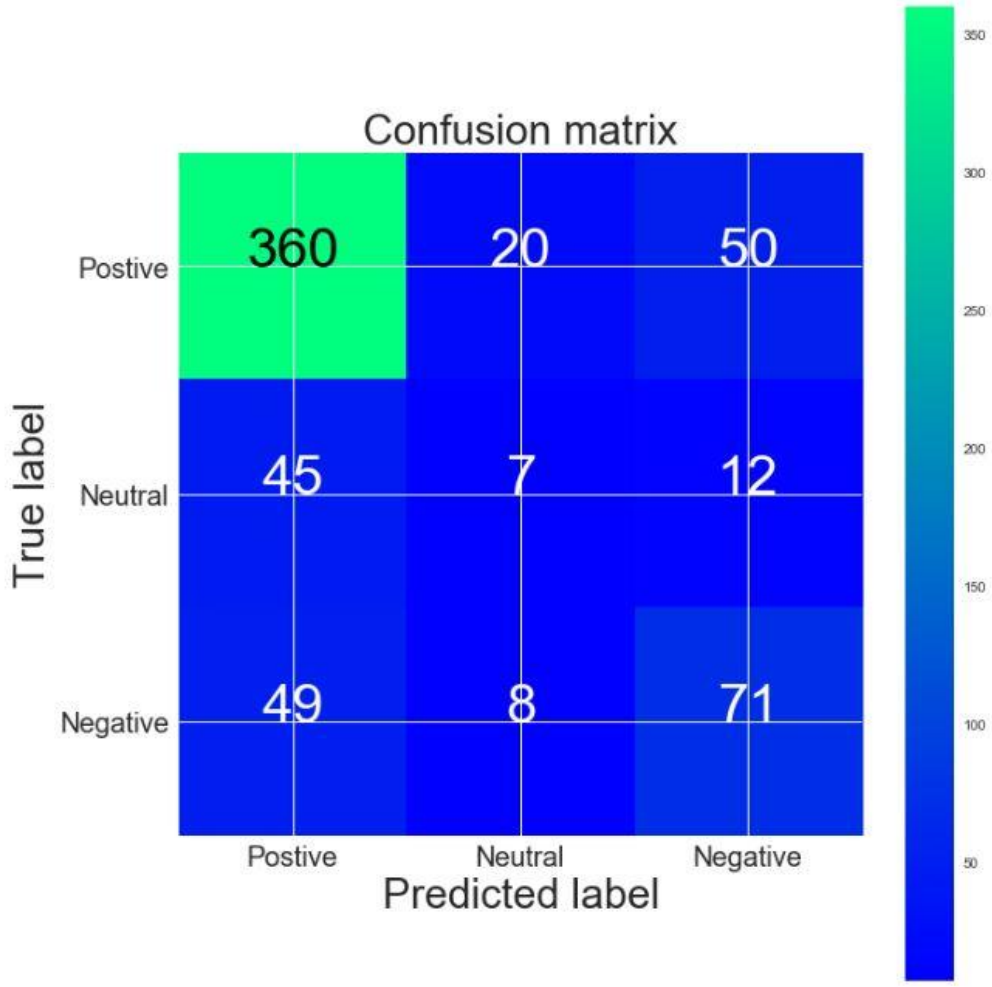
Classifier picked up some bad words in negative review (horrible, discontinued, vomiting) and it also picked some good words in positive review (free, greatly, stabilized)

4. **Apply TF-IDF, embedding visualization, importance words**
    - Embedding visualization plot after applying TF-IDF



This embedding look much more separated.

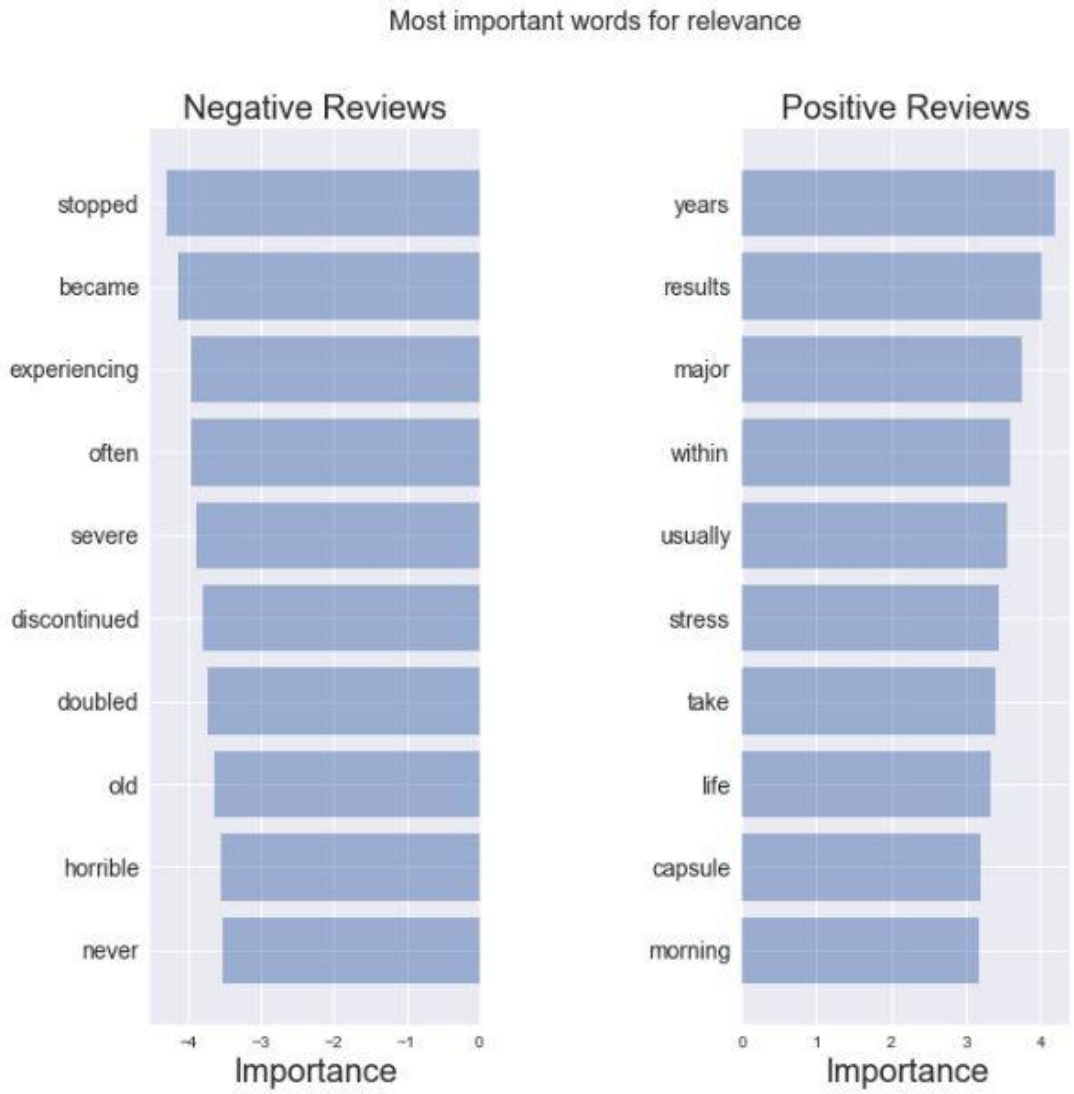
- Plot confusion matrix




```
Test Data Value Counts:
 1    430
0    128
2     64
```

This model performed not as good as BoW model.

- Plot of negative and positive importance words of train set



Classifier picked up some bad words in negative review (horrible, discontinued, never, old) and no good words in positive review.

5. Apply LSTM model and predict result on test set

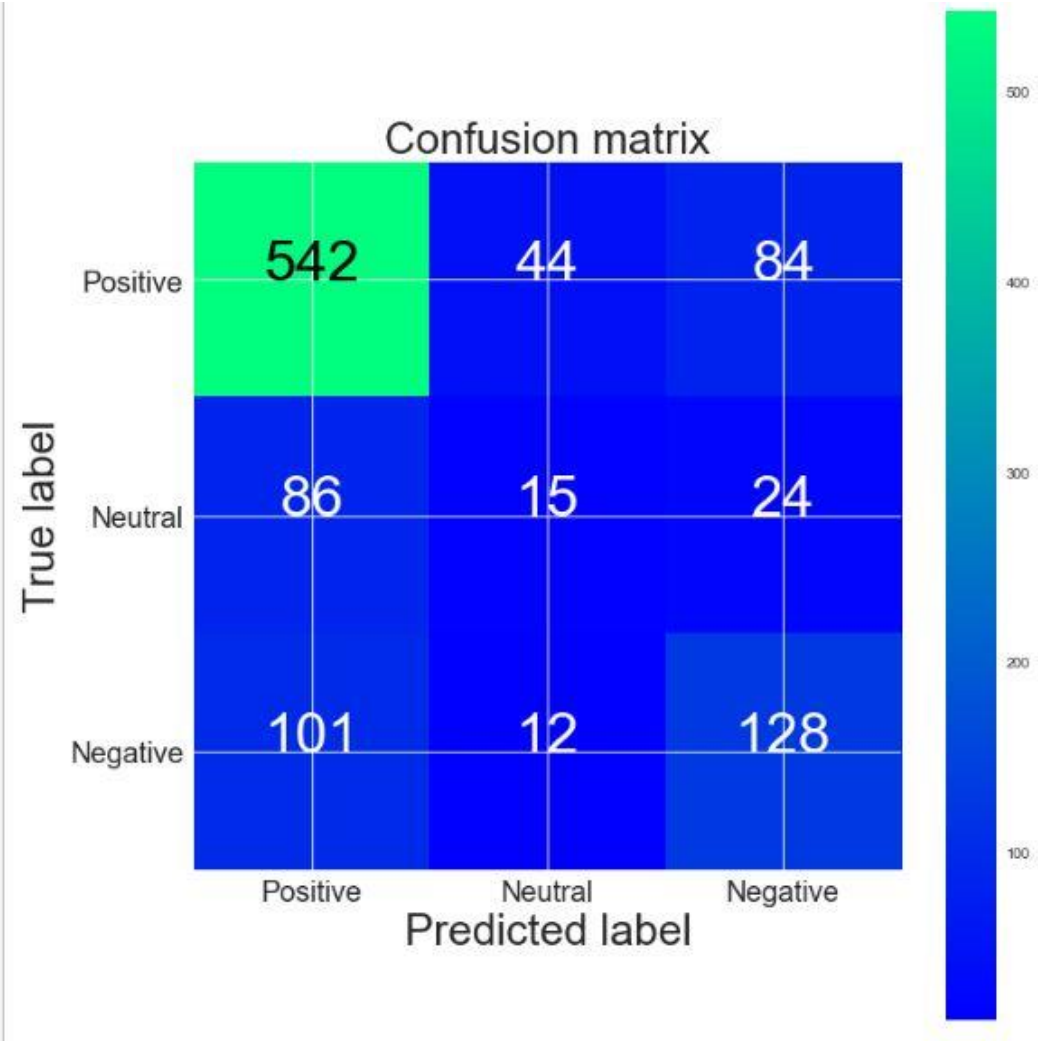
- Setup a LSTM model include dropout layers

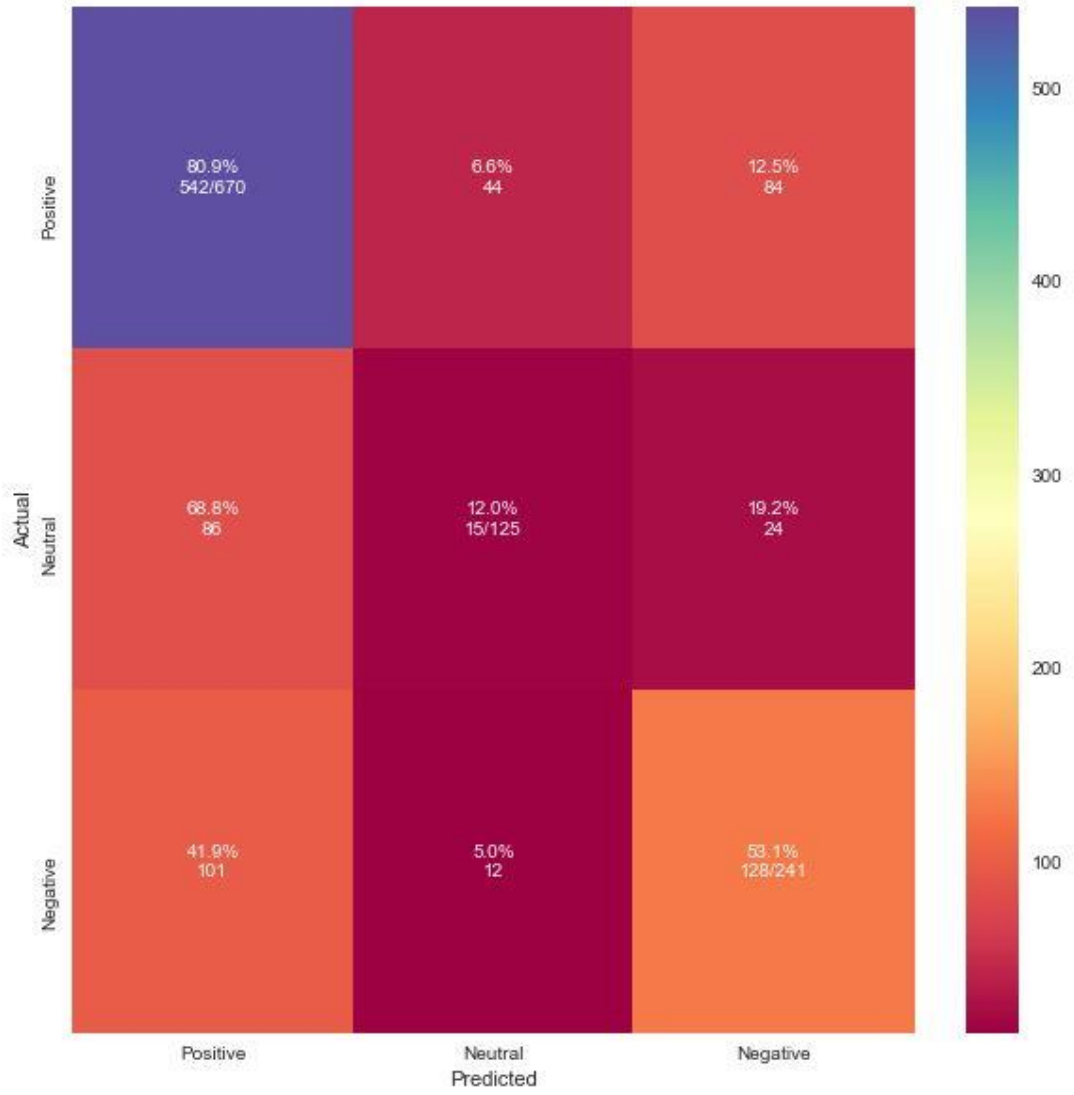
```
model = Sequential()
model.add(Embedding(nb_words, EMBEDDING_DIM, input_length=lstm_out))
model.add(Dropout(0.2))
model.add(LSTM(100))
model.add(Dropout(0.2))
model.add(Dense(3, activation = 'softmax'))
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
```

In [73]:
```
model.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 100, 160)          480000
_________________________________________________________________
dropout_1 (Dropout)          (None, 100, 160)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               104400
_________________________________________________________________
dropout_2 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 303
=================================================================
Total params: 584,703
Trainable params: 584,703
Non-trainable params: 0
_________________________________________________________________
```

- Plot confusion matrix in test set





The model did good in predicting Positive and Negative labels, still very bad on Neutral labels

6. Conclusion

The LSTM model performed good in test set but I'm still wonder why it predicted very bad on Neutral class. Maybe the cause is the ratio of binned classes I have set for both dataset. I am happy with this LSTM model and the final result, I have more clearly understanding about processing text data in after this practice.