Chirag Shah **Text Analytics COMP47600** **16/11/2019**
19200072 **Practical Number 9**

1) Human Ratings Task:

- **a) Get 3 classmates (opinion holders) to write three different opinions about their phone**
- **b) Get 3 different people (raters) to rate these comments as positive, negative, neutral or can't-say**
- **c) Take this 3 x 3 matrix and find the inter-rater reliability between your 3 raters using Kappa**
- **d) If you wanted to get the correlation between raters (using Pearson's rho) what would you do?**

Following are the opinions from 3 classmates about their phones -

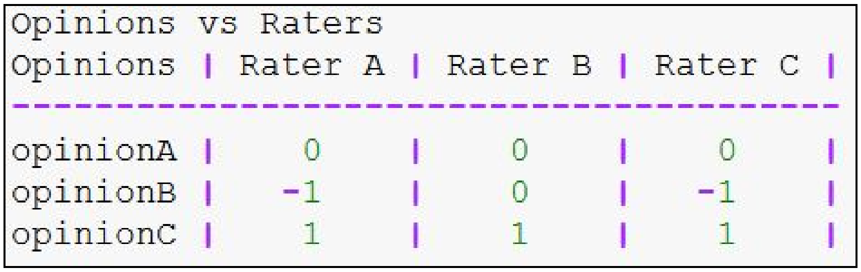
opinionA => 'OnePlus has good battery life and brilliant processing capabilities'
opinionB => 'iPhones is not good in display and camera, but it is better than Samsung'
opinionC => 'OnePlus has brilliant video streaming quality because of amazing graphics processor'

Rating numerical representation -
Positive rating => +1
Negative rating => -1
Neutral rating => 0

Following are the ratings given by 3 people over the above opinions -

raterA => [0,-1,1]
raterB => [0, 0, 1]
raterC => [0,-1,1]

Following is the 3x3 matrix representation for the above -



```
Opinions vs Raters
Opinions | Rater A | Rater B | Rater C |
----------------------------------------
opinionA |    0    |    0    |    0    |
opinionB |   -1    |    0    |   -1    |
opinionC |    1    |    1    |    1    |
```

Cohen's Kappa -

Cohen's kappa coefficient (κ) is used to **measure inter-rater and intra-rater reliability** for categorical items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance. Inter-rater reliability is the **degree of agreement among raters**. It is a score of how much homogeneity or consensus exists in the ratings given by various judges. Cohen's Kappa has the assumption that the raters are deliberately chosen.

The formula to calculate Cohen's kappa for two raters is:

$$\boxed{\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}}$$

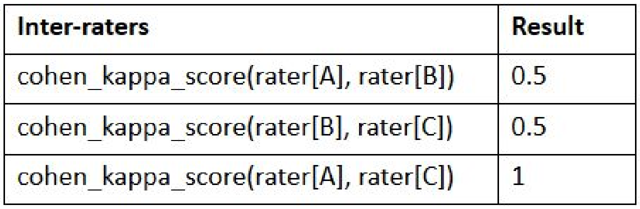
where, P_o = the relative observed agreement among raters. P_e = the probability of chance agreement. The interpretation of a kappa coefficient is the amount of observed agreement divided by the possible amount of non-chance agreement. The numerator represents the discrepancy between the observed probability of success and the probability of success under the assumption of an extremely bad case.

Possible interpretation of Kappa -

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

Kappa is **always less than or equal to 1**. A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement. In rare situations, Kappa can be negative. This is a sign that the two observers agreed less than would be expected just by chance.

To calculate Kappa score, we perform comparison between 2 raters at one time and do this computation for all the raters. The cohen_kappa_score() from sklearn module helps us in calculating the Kappa score. We pass in two raters list at one time to the above function. Following are the computation -



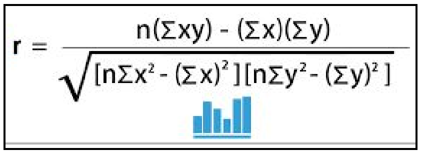
Inter-raters	Result
cohen_kappa_score(rater[A], rater[B])	0.5
cohen_kappa_score(rater[B], rater[C])	0.5
cohen_kappa_score(rater[A], rater[C])	1

From the results it is clearly observed that rater A and rater B are at **moderate agreement** for all the opinions. Similarly rater B and rater C are also at **moderate agreement**s for all the opinions. However, rater A and rater C have **strong agreements** over all the opinions.

Pearson's correlation is number between "**+1**" to "**-1**" which represents how strongly the two variables(in our case raters) are associated. Or to put this in simple words, it states the measure of the strength of **linear association** between two variables. It attempts to draw a line to best fit through the data of the given two variables, and the Pearson correlation coefficient "**r**" indicates how far away all these data points are from the line of best fit. The value of "**r**" ranges from +1 to -1 where:

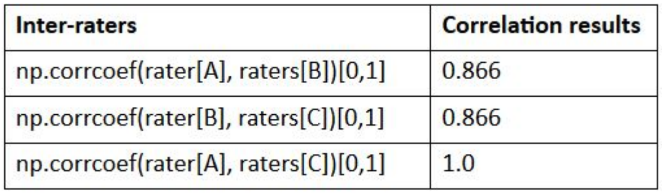
- **r**= +1/-1 represents that all data points lie on the line of best fit only i.e there is no data point which shows any variation from the line of best fit.
- **r** = 0 means that there is no correlation between the two variables.
- The values of **r** between +1 and -1 indicate that there is a variation of data around the line.
- The closer the values of **r** to 0, the greater the variation of data points around the line of best fit.

The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable.



$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

To compute pearson correlation we use the numpy function corrcoef() as follows -



Inter-raters	Correlation results
np.corrcoef(rater[A], raters[B])[0,1]	0.866
np.corrcoef(rater[B], raters[C])[0,1]	0.866
np.corrcoef(rater[A], raters[C])[0,1]	1.0

From the above results, it is evident that rater A and rater B have **moderate to strong correlation** between them. Similarly rater B and rater C also have **moderate to strong correlation** between them. However, rater A and rater C have the **strongest correlation** between them.

2) Do, some searches and find 3 sentiment lists that are commonly used in previous research. For 2 of these lists, select 10 positive and 10 negative words (randomly). Evaluate each word, discussing whether it is really positive/negative; for each one try to find a sentential context in which it might be interpreted with the opposite valence.

Following are the sentiment lists that are commonly used in previous research -

1. Liu and Hu opinion lexicon - It contains around 6800 positive and negative opinion words. (Liu, B. (2019). *Opinion Mining, Sentiment Analysis, Opinion Extraction*)
2. SentiWordNet is a lexical resource for opinion mining that assigns three sentiment scores: positivity, negativity, and objectivity (Sentiwordnet.isti.cnr.it. (2019). *Text Learning Group*)
3. The MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon is a list of subjectivity clues that is part of OpinionFinder. (Mpqa.cs.pitt.edu. (2019). *Subjectivity Lexicon*)

10 +ve and 10 -ve words are chosen from the list 1. (Liu and Hu opinion lexicon) and list 3. (MPQA subjectivity lexicon) respectively. Then accordingly sentences are constructed that imply the same valence and opposite valence.
Following are tabular representations of the results obtained -

*10 **Positive words** from the **Liu and Hu opinion lexicon sentiment list** -*

words	same valence(+ve context)	opposite valence(-ve context)
amenable	The immigration restrictionists around Trump seemed amenable to the choice	The Doctor said he had scar tissue and would be amenable to pneumonia in the injured lung.
flashy	He wears flashy clothes	The flashy marketing events conducted by the company was the primary reason why they couldn't achieve success
toil	Lindi has achieved her comfortable life only after years of hard toil.	I and my friends toiled up the very long path on steep hill because of which we fainted midway.
smitten	He was so smitten by her that he promised to move to Argentina to be near her.	The contractor may smite our land because of excessive fights over the car parking issue
abide	If you abide by the guidelines, you are sure to succeed with the project.	If there is one thing I cannot abide it is a lack of discipline
prefer	He prefers watching football to playing it.	She succumbed to depression because she preferred to stay aloof from everyone
piety	In this capacity his sincere piety and amiable character gained him great influence.	Being too piety about religious beliefs, the man couldn't maintain relations with family and friends
relent	Her parents eventually relented and let her go to the party.	She was going to refuse his request for going to the party, but later relented due to other commitments
beckon	He beckoned to me, as if he wanted to speak to me.	From the time he was a child, the haunting stories of the jungle beckoned him.
exceptionally	He is an exceptionally talented teacher.	The weather became exceptionally windy because of sudden storms
obsession	He has an obsession with cleanliness.	Too much of money became an obsession for him which made him a thief

10 ***Negative words*** *from the* ***Liu and Hu opinion lexicon sentiment list -***

words	same valence (-ve context)	opposite valence(+ve context)
anxious	Life was empty without him and no one seemed to be anxious to replace him	The company was anxious to avoid any trouble
hamper	Fierce storms have been hampering rescue efforts	Last night she had noticed a few clothes in a hamper in the laundry room
ambiguity	There is no ambiguity in calling conditionals with true antecedents "true" or "false."	The ambiguity in the messages helped the police crack the unsolved mystery of murder
abysmal	Red-clay is an abysmal formation, occurring in the sea bottom.	The abysmal failures of Jack Ma helped him become the CEO of Alibaba group
wobble	The tree should not wobble at the top of the root ball.	They have been wobbling in their support of the president's policies
wanton	He displayed a wanton disregard for the facts.	It was his custom on all these trips to make little wanton sketches of landscape and buildings.
abrasive	If the particles are not tiny enough, they will have an abrasive effect on the skin.	The wood got cleaned as it should be rubbed down with fine abrasive paper
variable	British weather is perhaps at its most variable in the spring.	The variable interest rates helped the bank gain good profits in the early summer quarters
sick	Lucy felt sick the morning after the party.	The plan was to surprise the heck out of the grizzly by sick the dog on him
defensive	The team’s defensive strategy was ineffective as they suffered heavy defeats due to lack of attacking play	He was very defensive about that good side of his life
throttle	Sometimes he annoys me so much that I could throttle him.	The airplane was flying at full throttle midway in the sky.

10 ***Positive words*** *from the MPQA Subjectivity Lexicon* ***sentiment list -***

words	same valence (+ve context)	opposite valence(-ve context)
ferocious	The stories of his ferocious savagery exceed belief.	The president came in for a ferocious criticism and he was forced to resign
vulnerable	The scheme will help charities working with vulnerable adults and young people	Start-ups are very vulnerable in the business world.
fortuitous	The collapse of its rivals was a fortuitous opportunity for the company.	The collapse of the ruler was the main reason for the empire to lose its astounding fame
friction	There’s less friction in relationships when you use teamwork.	It was impossible to reach an agreement because of the friction between the two sides.
understand	I wanted to make sure that we want the same things, that we really understand each other.	He tried a lot to understand the views of public; but couldn't win the elections.
feisty	She’s a feisty kid who is not afraid to challenge authority.	The people from communal party launched a feisty attack on the common people and injured many of them
torrid	Summers in the tropics are torrid.	He'd been given a pretty torrid time by the nation's voters
simple	The instructions were written in simple English.	The children gave too much of time on simple problems and hence failed in the exams
outshine	Ben Palmer easily outshone his rivals in the 200 metre freestyle.	he can outshine the deal and leave the other person stranded
appropriate	The EU has issued guidelines on appropriate levels of pay for part-time manual workers.	The accused had appropriated the property from the landlord

*10 **Negative words** from the MPQA Subjectivity Lexicon **sentiment list -***

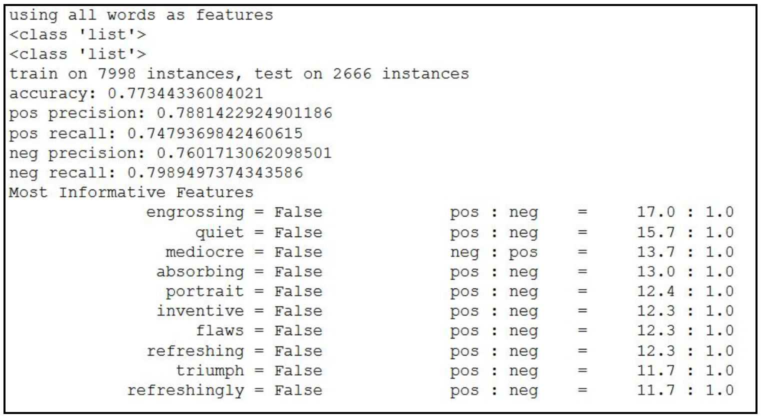
words	same valence (-ve context)	opposite valence(+ve context)
wicked	Telling lies is a wicked thing to do.	Sophie makes wicked cakes
overshadow	My happiness was overshadowed by the bad news.	His competitive nature often overshadows the other qualities
exhort	He wrongly exhorted me to follow the path of Engineering which I regret doing even today	I exhorted her to be a good child
swagger	He walked with a swagger and would laugh and joke confidently on the street, even with strangers.	They strolled around the camp with an exaggerated swagger
overawe	The savagery of his thoughts overawed him	The small kid was overawed by the atmosphere
audacity	She had the audacity to suggest I'd been carrying on with him	He whistled at the sheer audacity of the plan
sympathetic	The doctor was not at all sympathetic towards her patients	The current government are very sympathetic towards environmental issues.
susceptible	Patients with liver disease may be susceptible to infection	Joss Whedon's movies are susceptible to various interpretations.
supremacy	The two powers went beyond the extent of humanity to get the supremacy of state	This victory clearly proves the supremacy of the Brazilians in football.
rhetoric	The opposition parties have nothing but just an empty rhetoric	I was swayed by her rhetoric into donating all my savings to the charity
rhapsodize	The over rhapsodizing behavior of husband forced the wife to lodge a complaint against him	He began to rhapsodize about Gaby's beauty and charm

3) Bromberg's Sentiment Program: Have a look at the simple program that does sentiment analysis. So, take a look at the program and see what is happening in the different variables, but adding print statements on its variables.
a) Now consider ways to improve the training. Eg if you removed stop-words from the inputs what do you think might happen? b) Implement this or another solution in the program and report what happens to the precision and recall of the classifier.

a)

By using the Bromberg's Sentiment program we observe that the accuracy obtained is - **77%**

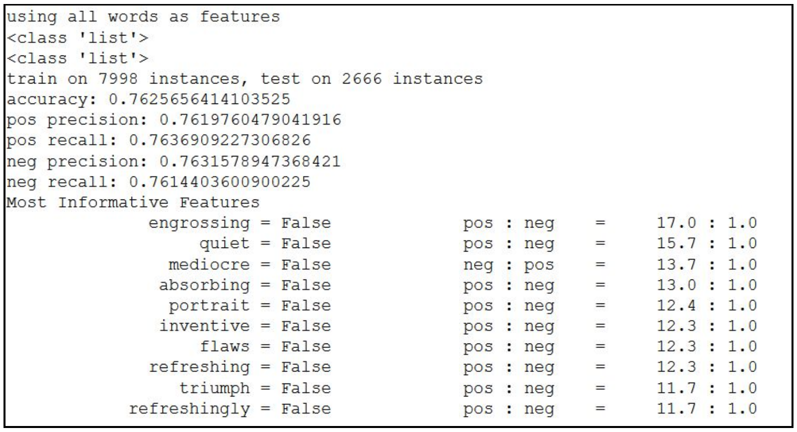
Following is the screenshot of results obtained after running the program -



```
using all words as features
<class 'list'>
<class 'list'>
train on 7998 instances, test on 2666 instances
accuracy: 0.77344336084021
pos precision: 0.7881422924901186
pos recall: 0.7479369842460615
neg precision: 0.7601713062098501
neg recall: 0.7989497374343586
Most Informative Features
              engrossing = False              pos : neg    =     17.0 : 1.0
                   quiet = False              pos : neg    =     15.7 : 1.0
                mediocre = False              neg : pos    =     13.7 : 1.0
               absorbing = False              pos : neg    =     13.0 : 1.0
                portrait = False              pos : neg    =     12.4 : 1.0
               inventive = False              pos : neg    =     12.3 : 1.0
                   flaws = False              pos : neg    =     12.3 : 1.0
              refreshing = False              pos : neg    =     12.3 : 1.0
                 triumph = False              pos : neg    =     11.7 : 1.0
            refreshingly = False              pos : neg    =     11.7 : 1.0
```

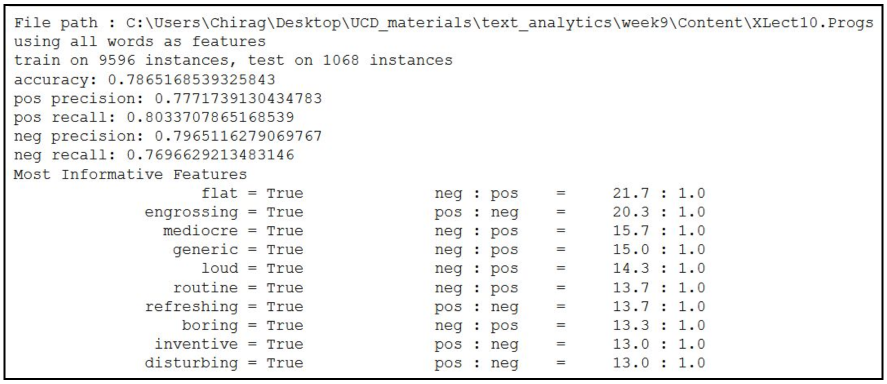
To improve the training following approaches were undertaken -

1. Stop words were removed - **Here, the accuracy decreased to 76%**
   Result snippet -



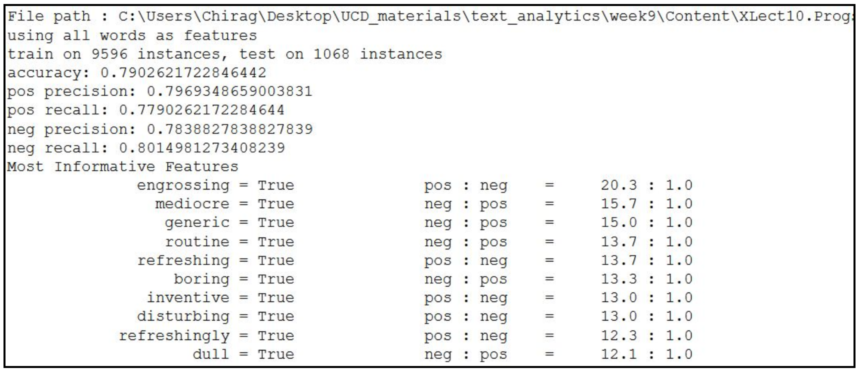
```
using all words as features
<class 'list'>
<class 'list'>
train on 7998 instances, test on 2666 instances
accuracy: 0.7625656414103525
pos precision: 0.7619760479041916
pos recall: 0.7636909227306826
neg precision: 0.7631578947368421
neg recall: 0.7614403600900225
Most Informative Features
              engrossing = False            pos : neg    =     17.0 : 1.0
                   quiet = False            pos : neg    =     15.7 : 1.0
                mediocre = False            neg : pos    =     13.7 : 1.0
               absorbing = False            pos : neg    =     13.0 : 1.0
                portrait = False            pos : neg    =     12.4 : 1.0
               inventive = False            pos : neg    =     12.3 : 1.0
                   flaws = False            pos : neg    =     12.3 : 1.0
              refreshing = False            pos : neg    =     12.3 : 1.0
                 triumph = False            pos : neg    =     11.7 : 1.0
            refreshingly = False            pos : neg    =     11.7 : 1.0
```

2. Increasing the size of training set to 80% and removing the stop words. - **Here, the accuracy increased to 78%**
   Result snippet -



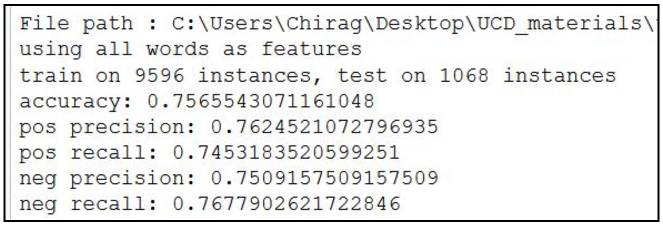
```
File path : C:\Users\Chirag\Desktop\UCD_materials\text_analytics\week9\Content\XLect10.Progs
using all words as features
train on 9596 instances, test on 1068 instances
accuracy: 0.7865168539325843
pos precision: 0.7771739130434783
pos recall: 0.8033707865168539
neg precision: 0.7965116279069767
neg recall: 0.7696629213483146
Most Informative Features
                    flat = True             neg : pos    =     21.7 : 1.0
              engrossing = True             pos : neg    =     20.3 : 1.0
                mediocre = True             neg : pos    =     15.7 : 1.0
                 generic = True             neg : pos    =     15.0 : 1.0
                    loud = True             neg : pos    =     14.3 : 1.0
                 routine = True             neg : pos    =     13.7 : 1.0
              refreshing = True             pos : neg    =     13.7 : 1.0
                  boring = True             neg : pos    =     13.3 : 1.0
               inventive = True             pos : neg    =     13.0 : 1.0
              disturbing = True             pos : neg    =     13.0 : 1.0
```

3. Only increasing the training set. **Here, the accuracy increased to 79%**

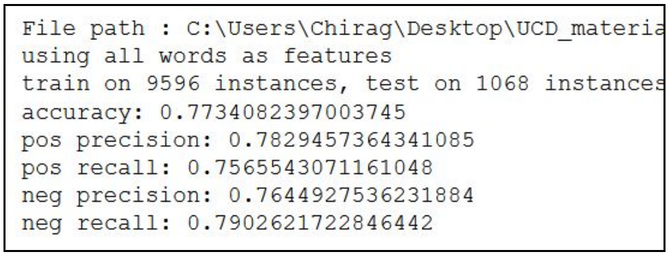


```
File path : C:\Users\Chirag\Desktop\UCD_materials\text_analytics\week9\Content\XLect10.Prog
using all words as features
train on 9596 instances, test on 1068 instances
accuracy: 0.7902621722846442
pos precision: 0.7969348659003831
pos recall: 0.7790262172284644
neg precision: 0.7838827838827839
neg recall: 0.8014981273408239
Most Informative Features
              engrossing = True             pos : neg    =     20.3 : 1.0
                mediocre = True             neg : pos    =     15.7 : 1.0
                 generic = True             neg : pos    =     15.0 : 1.0
                 routine = True             neg : pos    =     13.7 : 1.0
              refreshing = True             pos : neg    =     13.7 : 1.0
                  boring = True             neg : pos    =     13.3 : 1.0
               inventive = True             pos : neg    =     13.0 : 1.0
              disturbing = True             pos : neg    =     13.0 : 1.0
            refreshingly = True             pos : neg    =     12.3 : 1.0
                    dull = True             neg : pos    =     12.1 : 1.0
```

4. Using different classifiers
   For SVM, **the accuracy decreased to 75%** and for logistic regression, **the accuracy was 77%**



```
File path : C:\Users\Chirag\Desktop\UCD_materials\
using all words as features
train on 9596 instances, test on 1068 instances
accuracy: 0.7565543071161048
pos precision: 0.7624521072796935
pos recall: 0.7453183520599251
neg precision: 0.7509157509157509
neg recall: 0.7677902621722846
```



```
File path : C:\Users\Chirag\Desktop\UCD_materia
using all words as features
train on 9596 instances, test on 1068 instances
accuracy: 0.7734082397003745
pos precision: 0.7829457364341085
pos recall: 0.7565543071161048
neg precision: 0.7644927536231884
neg recall: 0.7902621722846442
```

From the above results we can infer that by increasing the size of training set, increasing the size of training and removing stop words; the accuracy obtained is better than the original accuracy. In all other scenarios; the accuracy achieved was less than the original accuracy.

b)

	Positive precision	Positive recall	Negative precision	Negative recall
Original method	0.78	0.74	0.76	0.79
Increasing training set and removing stop words	0.77	0.80	0.79	0.76
Increasing training set	0.79	0.77	0.78	0.80

Here we can consider recall because recall refers to the percentage of total relevant results correctly classified by your algorithm. From the above results we can infer that since the data is more clean after removing stop-words, the positive recall is higher.

References:-

Liu, B. (2019). *Opinion Mining, Sentiment Analysis, Opinion Extraction*. [online] Cs.uic.edu. Available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon [Accessed 17 Nov. 2019].

Sentiwordnet.isti.cnr.it. (2019). *Text Learning Group*. [online] Available at: http://sentiwordnet.isti.cnr.it/ [Accessed 17 Nov. 2019].

Mpqa.cs.pitt.edu. (2019). *Subjectivity Lexicon | MPQA*. [online] Available at: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ [Accessed 17 Nov. 2019].