1 Exercise 1

1. For this question, I asked three friends to provide their opinion on iPhones.

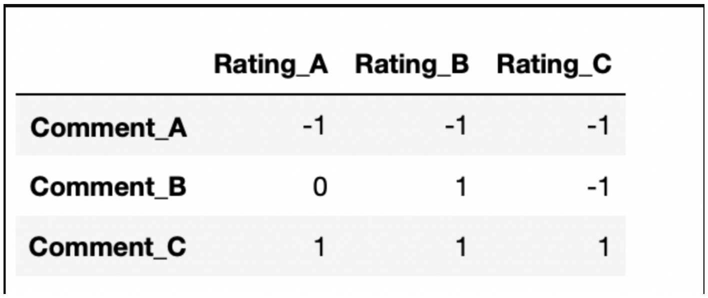
   A. The following comments are collected

   - Comment_A: "I hate iPhones. For me, they're too locked down and don't allow enough customisation for me. That, and they're far too expensive."
   - Comment_B: "I don't really care what phone I have as long as it's reliable. I like iPhones but I've broken the screen a few times so tend to go for cheaper phones now."
   - Comment_C: "I love my iPhone. Its really reliable and I've had my 6S for 5 years at this stage and outside of a slow battery it's still going strong. I'll probably get an iPhone 13 at some point this year after graduation."

   B. I asked three friends to rate the comments above, on a scale of positive, neutral, or negative which I have classified in the below as +1 for positive, -1 for negative, and 0 for neutral. The following ratings were given where the order corresponds with the order of the comment in the above. Overall, there was a consensus that the first comment was negative and the third comment was positive, however there was disagreement on whether the middle comment should be positive, negative, or neutral with each commentator commenting on a different aspect of the comment as the reason for their rating ('not care what phone' as neutral, 'I like iPhones' as positive, and 'I've broken the screen a few times so tend to go for cheaper phones now' as negative')

   - Rating_A=[-1,0,1]
   - Rating_B=[-1,1,1]
   - Rating_C=[-1,-1,1]

   - The ratings can be summarised in the below matrix



	Rating_A	Rating_B	Rating_C
Comment_A	-1	-1	-1
Comment_B	0	1	-1
Comment_C	1	1	1

Figure 1: Question 1 Dataframe - Ratings

   C. In the next section, I calculate the Cohen Kappa Score and Pearson Score for the ratings of the comments above.

   The Cohen Kappa score is a correlation metric in the range $[-1, 1] \subset \mathbb{R}$ designed measure the degree to which two rates agree while factoring in the probability that the two rates agree by chance. It is calculated as: $\kappa = \frac{p_0 - p_e}{1 - p_e}$: p_0 := Relative Observed agreement among raters, p_e := Probability of chance agrement

   Cohen suggested that the metric can be interpretted as a score less than 0 indicating no agremeent, 0.01 to 0.2 as no to slight agreement, 0.21 to 0.4 as fair agreement, 0.41 to 0.6 as moderate agreement, 0.61 to 0.8 as substantial agreement, and 0.81 to 1 as almost perfect agreement. The use of this metric is one which is controversial in literature, with the widely documented Kappa paradox documenting how the Kappa metric assumes low values despite high agreement [**paradox**] a key point in its reliability for imbalanced data, and in the case of using it as a classification metric instances can be generated where a poorer classifier produces a higher Kappa metric [**cohenbad**] which are undesirable outcomes for a performance measure, the metric is designed for only comparing 2 raters rather than n-raters, and the metric is a function of the number of subjects and categories chosen. Finally, the interpretation of the Kappa score which was mentioned was subjective and based on personal opinion rather than one which was produced based on data and therefore is itself controversial (particularly in the medical and legal field). Despite these factors, the usage of the measure is widespread and common.

   The Pearson score is similarly a correlation metric and is the 'standard' correlation coefficient taken which is a normalised measure of the covariance (and thus, the linear correlation) of two variables. It captures the distance between the line of best fit and the datapoints. Broadly, an absolute value of the correlation close to zero represents no linear correlation between the two variables, while an absolute value of the correlation close to 1 represents a strong linear correlation between the two variables, while the sign of the correlation indicates whether the correlation is positive or negative respectively.

Over a population, the correlation coefficient for random variables X, Y is defined as:
$p_{X,Y} := \frac{\text{cov}(X,Y)}{\sigma_X, \sigma_Y}$: $\sigma_X :=$ std. deviation of X, $\sigma_Y :=$ std. deviation of Y

While for a sample the sample correlation coefficient for a random variable X, Y is defined as

$r_{X,Y} := \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{(n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2)^{1/2}(n\sum_{i=1}^{n}(y_i)^2 - (\sum_{i=1}^{n} y_i)^2)^{1/2}}$: $x_i, y_i :=$ indexed samples of X and Y, $n :=$ no. of samples

Using the matrix above, I calculate the total Cohen Kappa score for each pair of ratings as follows using the cohen_kappa_score from SKLearn's Metrics library:

- score(Rating_A, Rating_B) = 0.5
- score(Rating_A, Rating_C) = 0.5
- score(Rating_B, Rating_C) = 0.4

We see that while each of the ratings is classified as moderate agreement, B and C are rated lower in agreement than A,B and A,C. This is driven by the rating of the second comment.

I also calculate the Pearson score for each pair of ratings as follows using numpy's 'corrcoef' function

- corrcoef(Rating_A, Rating_B) = 0.8660254037844385
- corrcoef(Rating_A, Rating_C) = 0.8660254037844385
- corrcoef(Rating_B, Rating_C) = 0.5

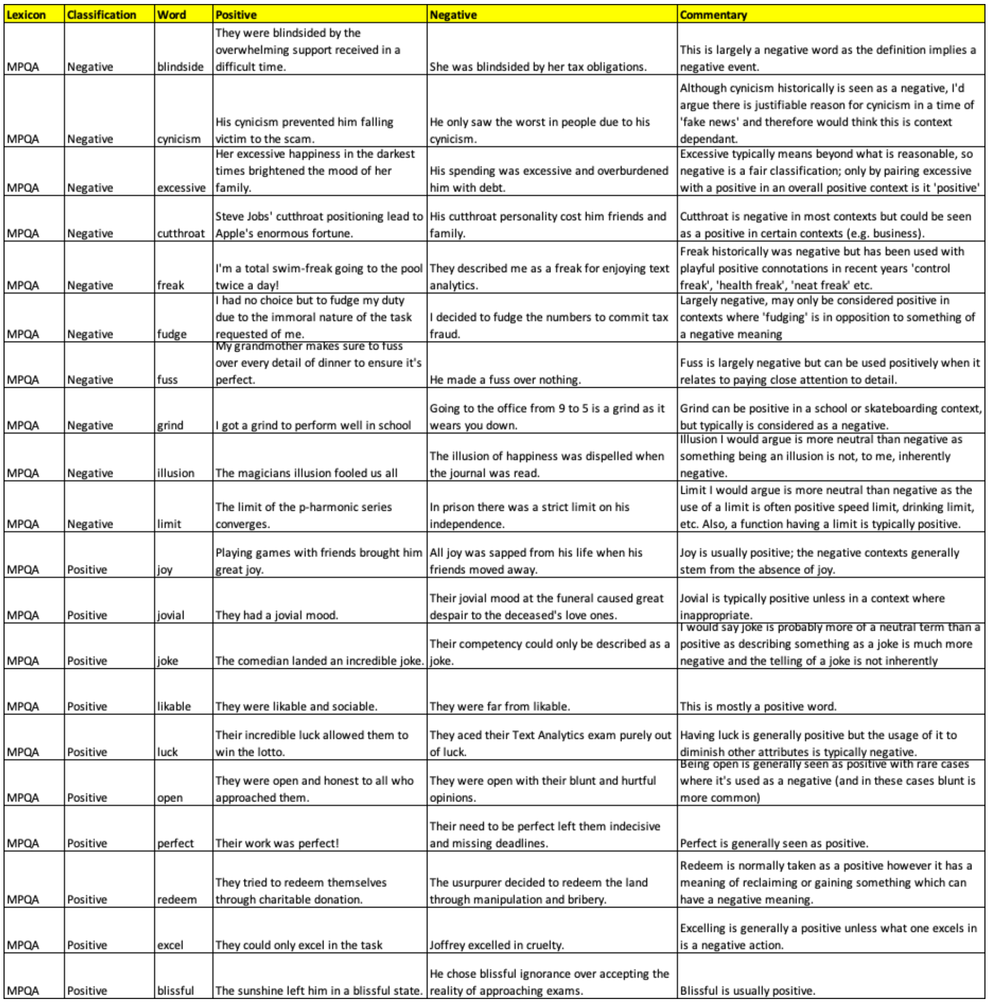
We see that while there is a strong correlation between Rating A and Rating B, there is a weakly positive correlation between Rating B and Rating C. As in the previous example when we calculated the Cohen Kappa Score, this is primarily driven by the difference in opinion between Rating B and C in regards to the second comment, and the absolute value of the distance between their ratings in comparison to the equivalent for A,B and A,C.

2 Exercise 2

1. I observe the following three lexicons which are used in literature:

   - SentiWordNet
   - Harvard General Inquirer
   - MPQA Subjectivity Lexicon

   For the MPQA, I chose the following negative and positive words and generated potential positive and negative contexts for them. The Classification is how they are classified in the MPQA, while the word is the word, and the positive and negative columns are different contexts where the word controbutes to semantic meaning of the sentence which is positive or negative.



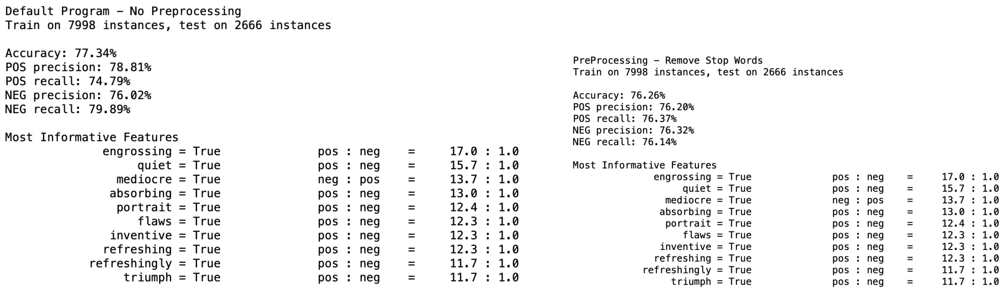
Lexicon	Classification	Word	Positive	Negative	Commentary
MPQA	Negative	blindside	They were blindsided by the overwhelming support received in a difficult time.	She was blindsided by her tax obligations.	This is largely a negative word as the definition implies a negative event.
MPQA	Negative	cynicism	His cynicism prevented him falling victim to the scam.	He only saw the worst in people due to his cynicism.	Although cynicism historically is seen as a negative, I'd argue there is justifiable reason for cynicism in a time of 'fake news' and therefore would think this is context dependant.
MPQA	Negative	excessive	Her excessive happiness in the darkest times brightened the mood of her family.	His spending was excessive and overburdened him with debt.	Excessive typically means beyond what is reasonable, so negative is a fair classification; only by pairing excessive with a positive in an overall positive context is it 'positive'
MPQA	Negative	cutthroat	Steve Jobs' cutthroat positioning lead to Apple's enormous fortune.	His cutthroat personality cost him friends and family.	Cutthroat is negative in most contexts but could be seen as a positive in certain contexts (e.g. business).
MPQA	Negative	freak	I'm a total swim-freak going to the pool twice a day!	They described me as a freak for enjoying text analytics.	Freak historically was negative but has been used with playful positive connotations in recent years 'control freak', 'health freak', 'neat freak' etc.
MPQA	Negative	fudge	I had no choice but to fudge my duty due to the immoral nature of the task requested of me.	I decided to fudge the numbers to commit tax fraud.	Largely negative, may only be considered positive in contexts where 'fudging' is in opposition to something of a negative meaning
MPQA	Negative	fuss	My grandmother makes sure to fuss over every detail of dinner to ensure it's perfect.	He made a fuss over nothing.	Fuss is largely negative but can be used positively when it relates to paying close attention to detail.
MPQA	Negative	grind	I got a grind to perform well in school	Going to the office from 9 to 5 is a grind as it wears you down.	Grind can be positive in a school or skateboarding context, but typically is considered as a negative.
MPQA	Negative	illusion	The magicians illusion fooled us all	The illusion of happiness was dispelled when the journal was read.	Illusion I would argue is more neutral than negative as something being an illusion is not, to me, inherently negative.
MPQA	Negative	limit	The limit of the p-harmonic series converges.	In prison there was a strict limit on his independence.	Limit I would argue is more neutral than negative as the use of a limit is often positive speed limit, drinking limit, etc. Also, a function having a limit is typically positive.
MPQA	Positive	joy	Playing games with friends brought him great joy.	All joy was sapped from his life when his friends moved away.	Joy is usually positive; the negative contexts generally stem from the absence of joy.
MPQA	Positive	jovial	They had a jovial mood.	Their jovial mood at the funeral caused great despair to the deceased's love ones.	Jovial is typically positive unless in a context where inappropriate.
MPQA	Positive	joke	The comedian landed an incredible joke.	Their competency could only be described as a joke.	I would say joke is probably more of a neutral term than a positive as describing something as a joke is much more negative and the telling of a joke is not inherently
MPQA	Positive	likable	They were likable and sociable.	They were far from likable.	This is mostly a positive word.
MPQA	Positive	luck	Their incredible luck allowed them to win the lotto.	They aced their Text Analytics exam purely out of luck.	Having luck is generally positive but the usage of it to diminish other attributes is typically negative.
MPQA	Positive	open	They were open and honest to all who approached them.	They were open with their blunt and hurtful opinions.	Being open is generally seen as positive with rare cases where it's used as a negative (and in these cases blunt is more common)
MPQA	Positive	perfect	Their work was perfect!	Their need to be perfect left them indecisive and missing deadlines.	Perfect is generally seen as positive.
MPQA	Positive	redeem	They tried to redeem themselves through charitable donation.	The usurpurer decided to redeem the land through manipulation and bribery.	Redeem is normally taken as a positive however it has a meaning of reclaiming or gaining something which can have a negative meaning.
MPQA	Positive	excel	They could only excel in the task	Joffrey excelled in cruelty.	Excelling is generally a positive unless what one excels in is a negative action.
MPQA	Positive	blissful	The sunshine left him in a blissful state.	He chose blissful ignorance over accepting the reality of approaching exams.	Blissful is usually positive.

Figure 2: Negative and Positive Words - MPQA - Please Zoom In to View

   Out of space constraints it is not possible to also classify another twenty words for another lexicon and also include it in the report (referring to Molly's comments in the breakout room that twenty words would be sufficient if size-constrained as the discussion was most important). For the twenty words randomly selected, ten positive and ten negative, I've tried to capture contexts in which the word may have both valences. In general, most of the words were pretty accurately classified in the lexicon (capturing most of the usages of the word and the most standard and common definition) however there were some instances where I felt that the word was not appropriately classified. In particular, the word 'cynicism' was classified as negative but I believe particularly in a modern context rather than a historical one it is probably more appropriate to capture this as neutral in the current climate of information exchange. Similarly, illusion is classified as a negative however I don't necessarily agree that something being described as an illusion or illusory is inherently a negative description; to take an example, while 'the illusion of peaces' implies an undercurrent threat and hence illusion could be taken as a negative, I would argue there is also an aspect whereby the illusion itself could be comforting and a positive (as an escape from the reality of a persistent threat). In general, in my opinion the classification of most of the results makes sense, with the opposite valences typically stemming from edge-cases or atypical usage of the word.

3 Exercise 3

1. I format the print statements in the provided text and do two runs with an without stop work removal to produce the following results:



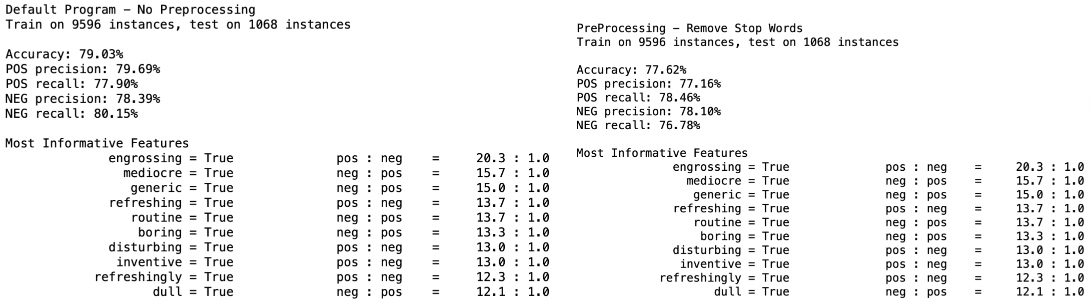
```
Default Program - No Preprocessing
Train on 7998 instances, test on 2666 instances

Accuracy: 77.34%
POS precision: 78.81%
POS recall: 74.79%
NEG precision: 76.02%
NEG recall: 79.89%

Most Informative Features
              engrossing = True              pos : neg    =     17.0 : 1.0
                   quiet = True              pos : neg    =     15.7 : 1.0
                mediocre = True              neg : pos    =     13.7 : 1.0
               absorbing = True              pos : neg    =     13.0 : 1.0
                portrait = True              pos : neg    =     12.4 : 1.0
                   flaws = True              pos : neg    =     12.3 : 1.0
               inventive = True              pos : neg    =     12.3 : 1.0
              refreshing = True              pos : neg    =     12.3 : 1.0
            refreshingly = True              pos : neg    =     11.7 : 1.0
                 triumph = True              pos : neg    =     11.7 : 1.0
```

```
PreProcessing - Remove Stop Words
Train on 7998 instances, test on 2666 instances

Accuracy: 76.26%
POS precision: 76.20%
POS recall: 76.37%
NEG precision: 76.32%
NEG recall: 76.14%

Most Informative Features
              engrossing = True              pos : neg    =     17.0 : 1.0
                   quiet = True              pos : neg    =     15.7 : 1.0
                mediocre = True              neg : pos    =     13.7 : 1.0
               absorbing = True              pos : neg    =     13.0 : 1.0
                portrait = True              pos : neg    =     12.4 : 1.0
                   flaws = True              pos : neg    =     12.3 : 1.0
               inventive = True              pos : neg    =     12.3 : 1.0
              refreshing = True              pos : neg    =     12.3 : 1.0
            refreshingly = True              pos : neg    =     11.7 : 1.0
                 triumph = True              pos : neg    =     11.7 : 1.0
```

Figure 3: Initial run vs With Stop Word Removal - 75% Test

2. We see that the top ten most important words remain the same in both instances (which is unsurprising as none of these words are stop words in the initial run) however what is initially unexpected to me is that the overall classification accuracy decreases while the positive precision decreases, positive recall increases, negative precision increases, and negative recall decreases in between the two runs. After thinking on this further, it makes sense why we are receiving these results the accuracy potentially decreases because we're removing junk data (the stop words) which may be present in volume and 'easy' to classify which is resulting in an inflated accuracy in the first case. The positive recall increases as with the removal of the stop words the number of predicted results is decreased while the number of true positives remains 'static' (or at least, what we want to consider as a true positive) causing a higher precision.

   In an attempt to improve the accuracy, I tried increase the test size to be 90% of the data, with and without preprocessing, and again we observe that the overall accuracy for the non-stop word version is higher with a similar trend in other metrics to the default setup, however both versions produce strong results than the baseline run (at 75% test size with no stop word removal).

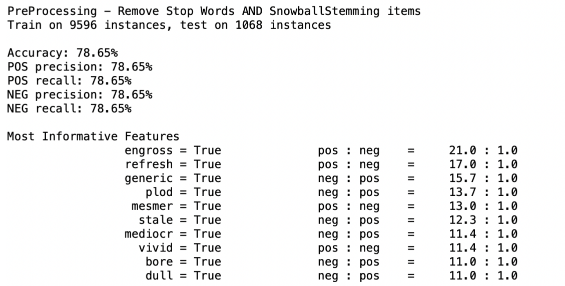


```
Default Program - No Preprocessing
Train on 9596 instances, test on 1068 instances

Accuracy: 79.03%
POS precision: 79.69%
POS recall: 77.90%
NEG precision: 78.39%
NEG recall: 80.15%

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
              refreshing = True              pos : neg    =     13.7 : 1.0
                 routine = True              neg : pos    =     13.7 : 1.0
                  boring = True              neg : pos    =     13.3 : 1.0
              disturbing = True              pos : neg    =     13.0 : 1.0
               inventive = True              pos : neg    =     13.0 : 1.0
            refreshingly = True              pos : neg    =     12.3 : 1.0
                    dull = True              neg : pos    =     12.1 : 1.0
```

```
PreProcessing - Remove Stop Words
Train on 9596 instances, test on 1068 instances

Accuracy: 77.62%
POS precision: 77.16%
POS recall: 78.46%
NEG precision: 78.10%
NEG recall: 76.78%

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
              refreshing = True              pos : neg    =     13.7 : 1.0
                 routine = True              neg : pos    =     13.7 : 1.0
                  boring = True              neg : pos    =     13.3 : 1.0
              disturbing = True              pos : neg    =     13.0 : 1.0
               inventive = True              pos : neg    =     13.0 : 1.0
            refreshingly = True              pos : neg    =     12.3 : 1.0
                    dull = True              neg : pos    =     12.1 : 1.0
```

Figure 4: Initial run vs With Stop Word Removal - 90% Test

   Interestingly, we see that the bottom half of the important features are changed from our initial run. One aspect which we see is that the 'refresh*' stem appears multiple times in the most important features. Because of this, I investigate the impact which stemming would have choosing to use a Snowball Stemmer as although stemming is crude and not often the best approach to take in pre-processing, I was interested in seeing the impact this would have: We observe that it causes an 'unusual' situation where the positive precision, negative precision, positive recall and negative recall are all equal to two decimal places, and the most informative features become significantly altered (as it now captures positive and negative stems rather than the full words).



```
PreProcessing - Remove Stop Words AND SnowballStemming items
Train on 9596 instances, test on 1068 instances

Accuracy: 78.65%
POS precision: 78.65%
POS recall: 78.65%
NEG precision: 78.65%
NEG recall: 78.65%

Most Informative Features
                 engross = True              pos : neg    =     21.0 : 1.0
                 refresh = True              pos : neg    =     17.0 : 1.0
                 generic = True              neg : pos    =     15.7 : 1.0
                    plod = True              neg : pos    =     13.7 : 1.0
                  mesmer = True              pos : neg    =     13.0 : 1.0
                   stale = True              neg : pos    =     12.3 : 1.0
                 mediocr = True              neg : pos    =     11.4 : 1.0
                   vivid = True              pos : neg    =     11.4 : 1.0
                    bore = True              neg : pos    =     11.0 : 1.0
                    dull = True              neg : pos    =     11.0 : 1.0
```

Figure 5: Stop Word Removal - 90% Test with Stemming

3. To improve the accuracy over the initial results, we see that ultimately remoing stop words, although it 'appears' to lower the accuracy and other metrics, in fact produces more relevant and 'accurate' results by being based on a cleaner set. To improve solely upon the accuracy measure, we see that both increasing the training size is sufficient to improve on the default accuracy, while increasing both the training size and removing stop words beats the default implementation. Finally, we see stemming the words, increasing the training set, and removing stop words produces higher accuracy and improves upon most metrics rather than removing stop words and increasing the training set size alone, however I would be cautious about such an implementation in reality because of the aggressive nature of the Snowball stemmer, and I believe ultimately that cleansing the stop words and increasing the training dataset size produces the best results in the attempts made to improve the default implementation's accuracy.

References

[1] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9. doi: 10.1016/0895-4356(90)90158-l. PMID: 2348207.

[2] Delgado R, Tibau XA (2019) Why Cohen's Kappa should be avoided as performance measure in classification. PLOS ONE 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916

[3] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2200-2204. European Language Resources Association.

[4] Stone, Philip J; Dexter C. Dunphry; Marshall S. Smith; and Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press.

[5] Wiebe, Janyce; Theresa Wilson; and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39: 165-210.