2022

Sentiment Analysis Report

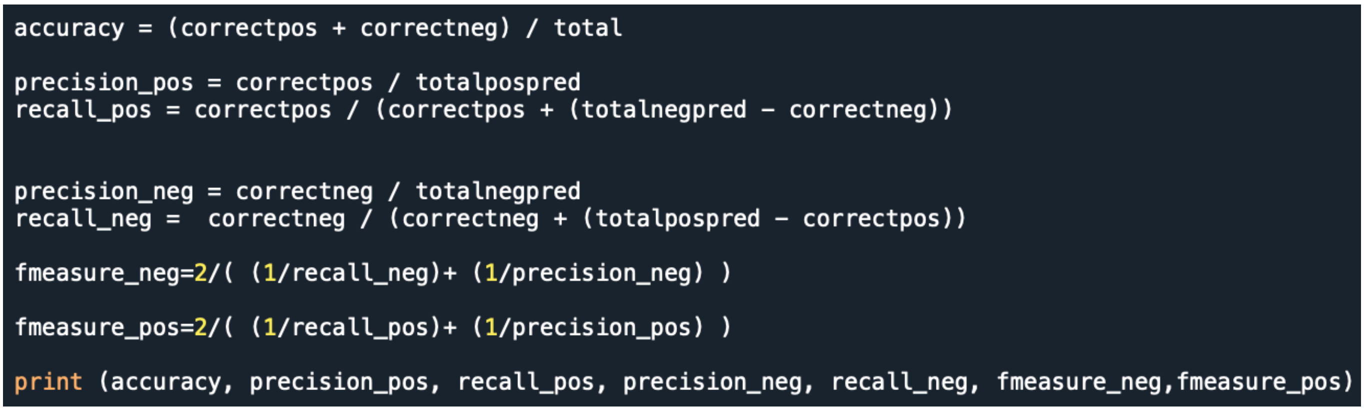
EVANGELIA AMOIRIDOU

Naïve Bayes Classifier (NBC)

Rotten Tomatoes and Nokia Data

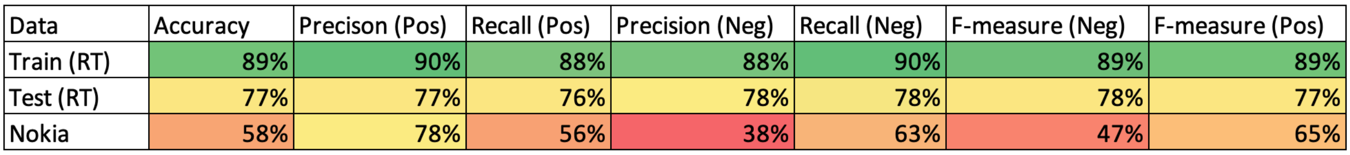
The NBC analyses data on film and phone reviews by learning from a sample of train data. This classifier calculates the p(word|sentiment) and the p(sentiment). The test data comprises of 90% of the Rotten Tomatoes data and 100% of the Nokia data. The outcome is an estimate of the sentences' sentiment. Seven metrics were used (accuracy, recall, precision & f-measure) to evaluate the performance of the model.

Figure 1: Code to create the metrics



```
accuracy = (correctpos + correctneg) / total

precision_pos = correctpos / totalpospred
recall_pos = correctpos / (correctpos + (totalnegpred - correctneg))


precision_neg = correctneg / totalnegpred
recall_neg =  correctneg / (correctneg + (totalpospred - correctpos))

fmeasure_neg=2/( (1/recall_neg)+ (1/precision_neg) )

fmeasure_pos=2/( (1/recall_pos)+ (1/precision_pos) )

print (accuracy, precision_pos, recall_pos, precision_neg, recall_neg, fmeasure_neg,fmeasure_pos)
```

Table 1: Classification results



Data	Accuracy	Precison (Pos)	Recall (Pos)	Precision (Neg)	Recall (Neg)	F-measure (Neg)	F-measure (Pos)
Train (RT)	89%	90%	88%	88%	90%	89%	89%
Test (RT)	77%	77%	76%	78%	78%	78%	77%
Nokia	58%	78%	56%	38%	63%	47%	65%

Key Findings

- The model performs best on the Train data.
- Secondly, it works relatively well on the Test RT data as compared to the Nokia data.
- What stands out are the high positive precision rates across both data sets (Test RT with 77% and Nokia with 78%).
- On the other hand, the precision rates for the negative words are very low for the Nokia data (38%) comparing to the other data sets (Test RT 78%).
- Nokia data scores low on the recall positive calculation as well (56%).

- As a final observation, it’s evident that the Train data for the RT achieve the highest scores, while the Nokia data achieve comparatively lower scores. The Test data for RT seem to maintain a fair performance which hovers at 80%.

Reasoning

Rotten Tomatoes Test Data

False negative

- *“not so much a movie as a picture book for the big screen . this isn't my favorite in the series , still i enjoyed it enough to recommend”*

The above sentence has several negative connotations that cover a large proportion of the sentence. The wrong classification occurs because the model is unable to understand the impact of the positive connotation that comes at the end of the sentence which should overpower the previous negative connotations.

False positive

- *“pays tribute to heroes the way julia roberts hands out awards--with phony humility barely camouflaging grotesque narcissism”*

The extracted sentence has several words with a positive meaning as well as the word “narcissism” which is absent from the negative word dictionary. These two factors cause the false positive result to arise.

Nokia Data

False Negative

- *“the speaker phone is very functional and i use it in the car , very audible even with freeway noise”*

The model is extracting the meaning of the word “noise” and taking it out of context, which results to the wrong prediction.

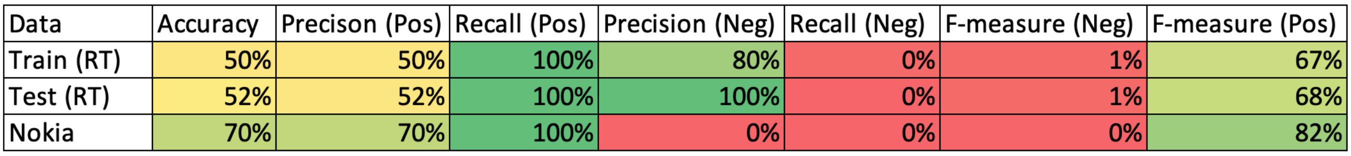
False Positive

- “some of the higher pitched rings are very easy to hear , but not easy to listen to”

In this case the words “but” and “not” are not included in the dictionary. The expectation would be that parts of speech like conjunctions that are used to contrast ideas and to show negation, would give the sentence a negative semantic value.

Rule based approach (RBA)

Table 2: RBA results



Data	Accuracy	Precison (Pos)	Recall (Pos)	Precision (Neg)	Recall (Neg)	F-measure (Neg)	F-measure (Pos)
Train (RT)	50%	50%	100%	80%	0%	1%	67%
Test (RT)	52%	52%	100%	100%	0%	1%	68%
Nokia	70%	70%	100%	0%	0%	0%	82%

Key findings

- The model shows the best performance on the Recall Positive metric and performs poorly on the Recall Negative as there are almost no false negatives.
- As Nokia’s sentence structure is simpler, the model performs relatively well (accuracy).
- The Precision Negative results vary significantly unlike the Precision Positive ones.

Reasoning

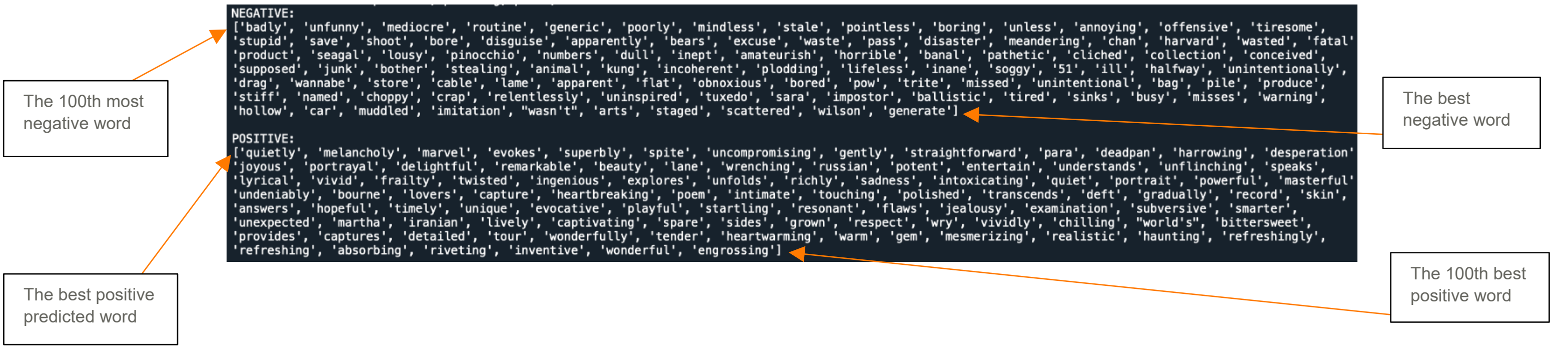
False Positive

- “one more thing , the default ringtones that come with the phone are horrible “

The defined threshold is very low which causes the underestimation of some negative sentences.

Most useful words

Figure 2: Top 100 most useful words for predicting sentiment



```
NEGATIVE:
['badly', 'unfunny', 'mediocre', 'routine', 'generic', 'poorly', 'mindless', 'stale', 'pointless', 'boring', 'unless', 'annoying', 'offensive', 'tiresome',
'stupid', 'save', 'shoot', 'bore', 'disguise', 'apparently', 'bears', 'excuse', 'waste', 'pass', 'disaster', 'meandering', 'chan', 'harvard', 'wasted', 'fatal'
'product', 'seagal', 'lousy', 'pinocchio', 'numbers', 'dull', 'inept', 'amateurish', 'horrible', 'banal', 'pathetic', 'cliched', 'collection', 'conceived',
'supposed', 'junk', 'bother', 'stealing', 'animal', 'kung', 'incoherent', 'plodding', 'lifeless', 'inane', 'soggy', '51', 'ill', 'halfway', 'unintentionally',
'drag', 'wannabe', 'store', 'cable', 'lame', 'apparent', 'flat', 'obnoxious', 'bored', 'pow', 'trite', 'missed', 'unintentional', 'bag', 'pile', 'produce',
'stiff', 'named', 'choppy', 'crap', 'relentlessly', 'uninspired', 'tuxedo', 'sara', 'impostor', 'ballistic', 'tired', 'sinks', 'busy', 'misses', 'warning',
'hollow', 'car', 'muddled', 'imitation', "wasn't", 'arts', 'staged', 'scattered', 'wilson', 'generate']

POSITIVE:
['quietly', 'melancholy', 'marvel', 'evokes', 'superbly', 'spite', 'uncompromising', 'gently', 'straightforward', 'para', 'deadpan', 'harrowing', 'desperation'
'joyous', 'portrayal', 'delightful', 'remarkable', 'beauty', 'lane', 'wrenching', 'russian', 'potent', 'entertain', 'understands', 'unflinching', 'speaks',
'lyrical', 'vivid', 'frailty', 'twisted', 'ingenious', 'explores', 'unfolds', 'richly', 'sadness', 'intoxicating', 'quiet', 'portrait', 'powerful', 'masterful'
'undeniably', 'bourne', 'lovers', 'capture', 'heartbreaking', 'poem', 'intimate', 'touching', 'polished', 'transcends', 'deft', 'gradually', 'record', 'skin',
'answers', 'hopeful', 'timely', 'unique', 'evocative', 'playful', 'startling', 'resonant', 'flaws', 'jealousy', 'examination', 'subversive', 'smarter',
'unexpected', 'martha', 'iranian', 'lively', 'captivating', 'spare', 'sides', 'grown', 'respect', 'wry', 'vividly', 'chilling', "world's", 'bittersweet',
'provides', 'captures', 'detailed', 'tour', 'wonderfully', 'tender', 'heartwarming', 'warm', 'gem', 'mesmerizing', 'realistic', 'haunting', 'refreshingly',
'refreshing', 'absorbing', 'riveting', 'inventive', 'wonderful', 'engrossing']
```

The 100th most negative word

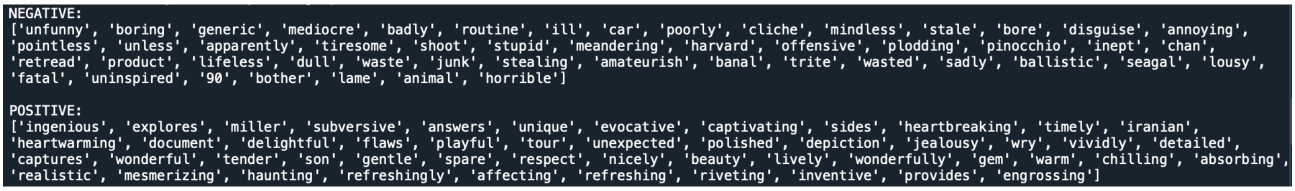
The best negative word

The best positive predicted word

The 100th best positive word

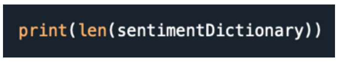
It's been noticed that in the list of words above there have been classified as negative words that don't have a negative connotation and vice versa. In most cases, these words are nouns ("car", "tour"). Many ambiguous words have been included in such as "routine" and "unexpected". Most words are good at predicting positive or negative sentiment, but the model's limitation is that it cannot evaluate the neighboring words' sentiment to make a judgement.

Figure 3: Top 50 most useful words for predicting sentiment



```
NEGATIVE:
['unfunny', 'boring', 'generic', 'mediocre', 'badly', 'routine', 'ill', 'car', 'poorly', 'cliche', 'mindless', 'stale', 'bore', 'disguise', 'annoying',
'pointless', 'unless', 'apparently', 'tiresome', 'shoot', 'stupid', 'meandering', 'harvard', 'offensive', 'plodding', 'pinocchio', 'inept', 'chan',
'retread', 'product', 'lifeless', 'dull', 'waste', 'junk', 'stealing', 'amateurish', 'banal', 'trite', 'wasted', 'sadly', 'ballistic', 'seagal', 'lousy',
'fatal', 'uninspired', '90', 'bother', 'lame', 'animal', 'horrible']

POSITIVE:
['ingenious', 'explores', 'miller', 'subversive', 'answers', 'unique', 'evocative', 'captivating', 'sides', 'heartbreaking', 'timely', 'iranian',
'heartwarming', 'document', 'delightful', 'flaws', 'playful', 'tour', 'unexpected', 'polished', 'depiction', 'jealousy', 'wry', 'vividly', 'detailed',
'captures', 'wonderful', 'tender', 'son', 'gentle', 'spare', 'respect', 'nicely', 'beauty', 'lively', 'wonderfully', 'gem', 'warm', 'chilling', 'absorbing',
'realistic', 'mesmerizing', 'haunting', 'refreshingly', 'affecting', 'refreshing', 'riveting', 'inventive', 'provides', 'engrossing']
```

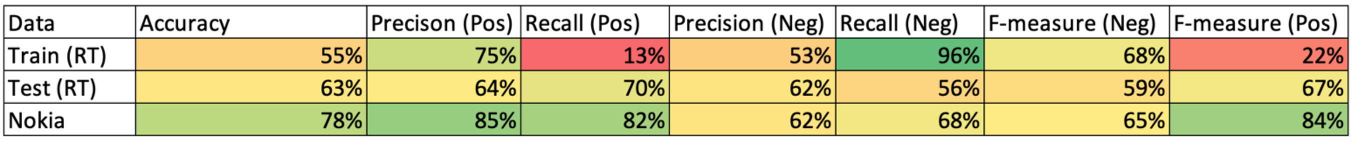
The code below calculated the size of the dictionary. The number of words is 6790.



```
print(len(sentimentDictionary))
```

New rule - based approach (NRBA)

Table 3: NRBA performance



Data	Accuracy	Precison (Pos)	Recall (Pos)	Precision (Neg)	Recall (Neg)	F-measure (Neg)	F-measure (Pos)
Train (RT)	55%	75%	13%	53%	96%	68%	22%
Test (RT)	63%	64%	70%	62%	56%	59%	67%
Nokia	78%	85%	82%	62%	68%	65%	84%

Key Findings

- The NRBA applies 4 additional rules – negation, capitalization, intensifiers and diminishers.
- This new model performs relatively well on simple data – Nokia.
- However, it suffers when deployed on the larger set of RT data (train RT) as compared to test RT which highlights that the model doesn't perform well on a larger set of data.

Reasoning

False negative

- "small size"

The sentence doesn't have a sufficient number of words to help the model and ends up taking the words out of context.

False positive

- "the master of disguise may have made a great saturday night live sketch , but a great movie it is not"

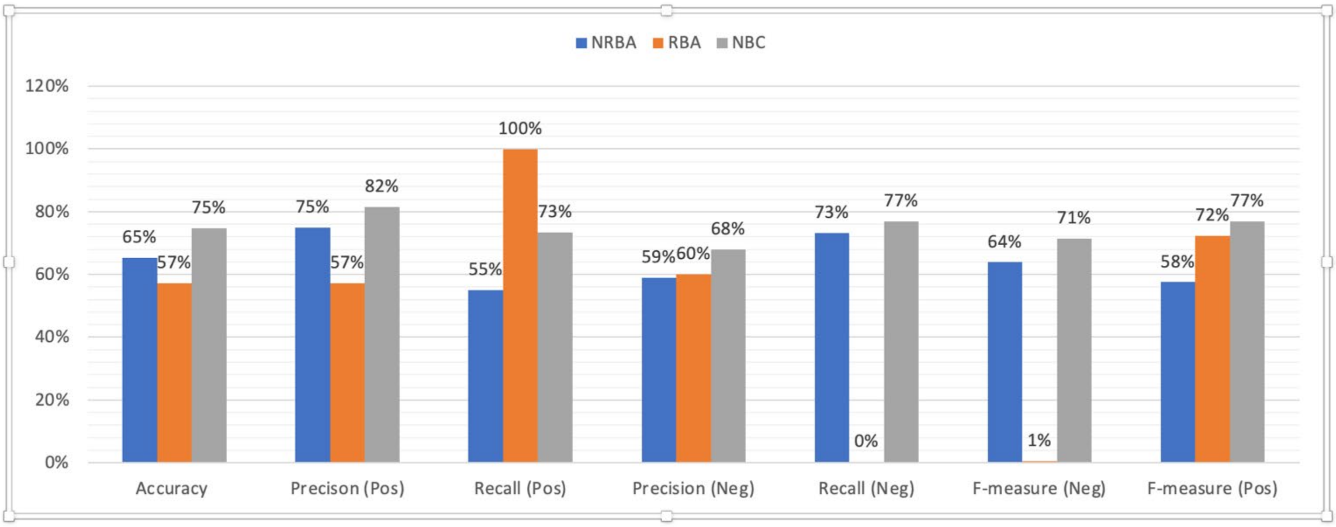
The NRBA still suffers from understanding the importance of the sentiment of the last section of the sentence which affects the overall sentiment.

Conclusion

Table 4: Average scores for all the models.

Model	Accuracy	Precison (Pos)	Recall (Pos)	Precision (Neg)	Recall (Neg)	F-measure (Neg)	F-measure (Pos)
NRBA	65%	75%	55%	59%	73%	64%	58%
RBA	57%	57%	100%	60%	0%	1%	72%
NBC	75%	82%	73%	68%	77%	71%	77%

Figure 4: Comparison of all the models (averages)



Comparison

Although, recall positive for RBA is very high, the model's accuracy is lower because it fails to predict negative sentences appropriately.

Across the metrics, the NRBA model performs on a more consistent basis compared to the RBA model.

Despite the RBA having a higher f-measure as compared to the NRBA, a significant proportion of metrics (4) have risen from the new approach.

Generalisation & Patterns

The statistical approach works well on the RT data. The RBA ,on the contrary, performs well on simpler data (Nokia).

The NBC model generalises relatively well on movie-based review data as its training data is based off a niche domain. Despite demonstrating a low performance across several metrics for a completely unseen data set (Nokia), its positive precision is almost unshakeable.

The RBA approach is untrustworthy as it tends to over predict on positive sentiments resulting in many false positives. It consistently demonstrates this behaviour across all three datasets.

The NRBA approach fails to scale across larger datasets as seen by the scores, which highlights the need to be cautious when deploying this model. This model somewhat scores relatively high on the simpler data (Nokia).