COM6115 Text Processing

Sentiment Analysis

03/12/2023

Mohammad Eizeddin

ACP23ME

Word-Count

Excluding tables, headings, and appendix

1071

Table of Contents

Step 2 Naive Bayes (Film Dataset):

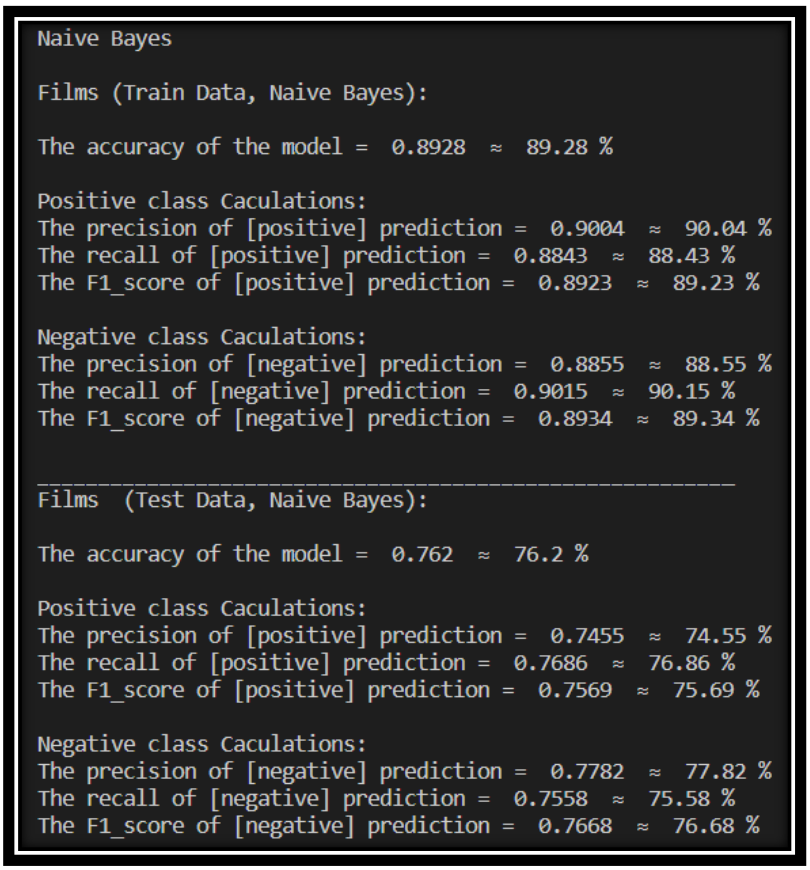


Figure 01: Results of running the Naive Bayes model on the film dataset.

Each time the assignment code is executed, classification metrics change slightly. To avoid this issue, a fixed random seed of (42) is used to produce the results.

Film dataset:

The model achieves high accuracy in predicting the overall sentiment of the Film dataset. It also achieves high precision scores, close to the accuracy score, meaning the model can accurately identify the targeted sentiment class (positive/negative). The recall score is likewise sharp and balanced with precision suggesting the correct classification of most sentiments from the correct total with not much of a precision-recall trade-off. The F1-scores indicate good and balanced performance across both classes.

Film training data vs Film testing data:

Figure 01 shows that the model performs better on the training data than on the testing data indicating a degree of overfitting. Overfitting is when a model fits almost exactly against its training data. The model could benefit from better generalisation and regularization techniques.

Step 3 Naive Bayes (Nokia vs Films Datasets):

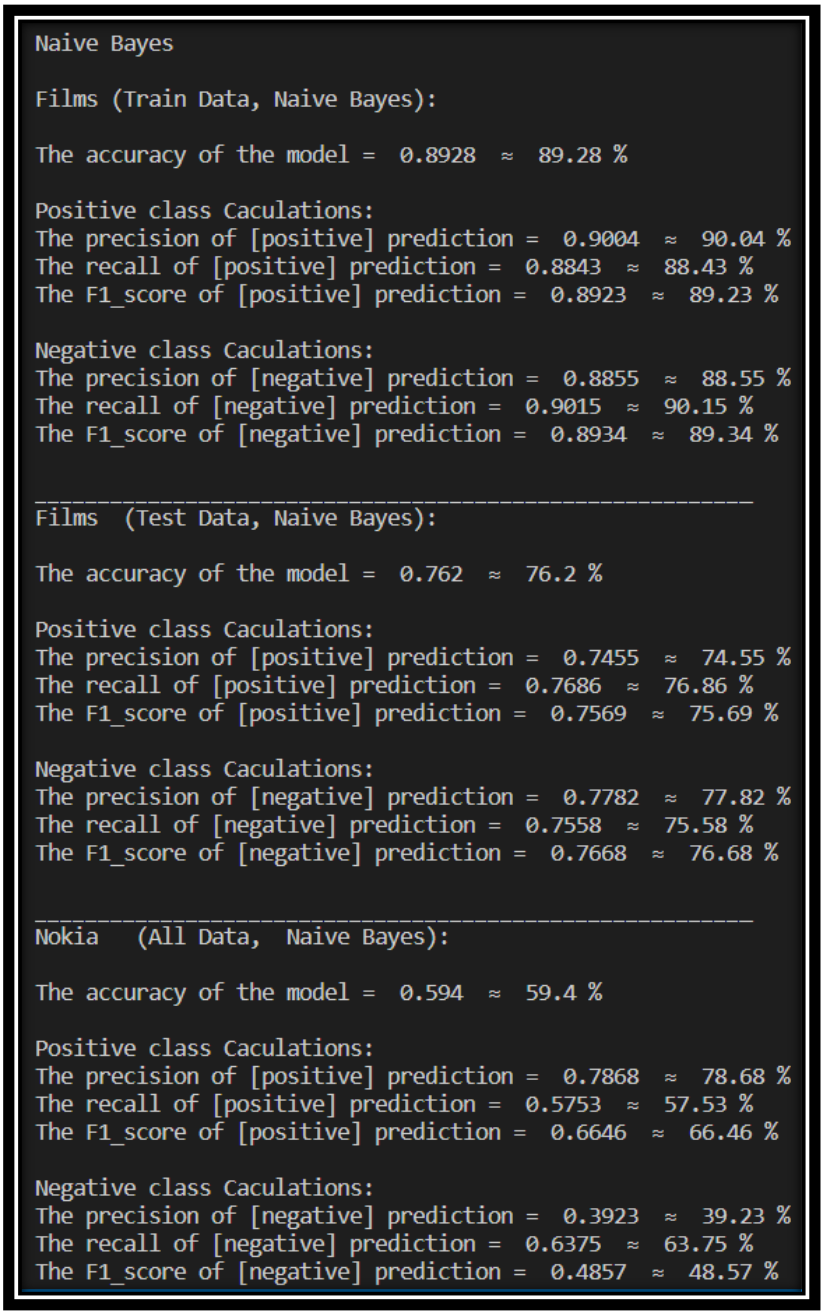


Figure 02: Results of running the Naive Bayes model on both the film and Nokia datasets.

Film data vs Nokia data:

The Bayes model was trained on the Films dataset while ignoring the Nokia dataset. Thus, the model is expected to perform better on the Films dataset. The Nokia scores are average but drastically lower. It is inefficient to generalise the Films training data to work across both domains as the manner of expressing sentiment might vary between domains.

Nokia dataset:

While the model's overall accuracy with Nokia's dataset is above average, the precision, recall, and f1-score for positive sentiment are all relatively higher than for negative sentiment, indicating that the model performs better with positive sentiments. The slightly high negative recall score could simply reflect the presence of many negative flags; however, precision indicates the majority of those flags are incorrect.

Step 4 Naive Model Analysis:

Most useful words:

According to the $mostUseful$ function, the most useful words for predicting positive sentiment are words with an extremely low probability of being negative or words with a high overall probability of being positive ($pWordPos[word]$) / ($pWordPos[word] + pWordNeg[word]$) . Conversely, the most useful words for predicting negative sentiment are those with the lowest probability of being positive, given that the most useful words for negative sentiments are taken from the head of a list sorted in ascending order of positive prediction power.

```
The most useful word for classifing positive and negative sentiments:

NEGATIVE:
['routine', 'generic', 'mediocre', 'badly', 'unfunny', 'waste', 'poorly', 'disguise', 'mindless', 'pointless', 'bore', 'stale', 'devoid', 'offensive',
 'shoot', 'boring', 'annoying', 'tiresome', 'dull', 'stupid', 'unless', 'meandering', 'apparently', 'plodding', 'worse', 'pass', 'horrible', 'wasted',
 'paper', 'harvard', 'pinocchio', 'supposed', 'junk', 'amateurish', 'banal', 'chan', 'kung', 'maudlin', 'pathetic', 'product', 'incoherent', 'seagal',
 'lifeless', 'lame', 'ill', 'stealing', 'fatal', 'pile', 'trite', 'lousy']

POSITIVE:
['timely', 'intimate', 'surreal', 'respect', 'jealousy', 'sly', 'current', 'record', 'subversive', 'answers', 'smarter', 'flaws', 'warm', 'evocative',
 'startling', 'format', 'captivating', 'spare', 'sides', 'grown', 'captures', 'tender', 'playful', 'son', 'iranian', 'detailed', 'tour', 'polished', '
wonderful', 'challenging', 'wonderfully', 'wry', 'powerful', 'glimpse', 'lively', 'vividly', 'heartwarming', 'chilling', 'gem', 'portrait', 'delicate'
, 'realistic', 'mesmerizing', 'refreshingly', 'triumph', 'riveting', 'refreshing', 'inventive', 'provides', 'engrossing']
```

Figure 03: Most useful words as printed from the mostUseful function.

Most useful words (count & analysis):

```
Count of the most useful words present in the sentiment dictionary =  57
Percentage present in sentimentDictionary =  57 %
```

Figure 04: The number of the most useful words present in the sentiment dictionary.

Some of the words in the list, such as (Iranian, Son, Tour, Format, Paper, Harvard, Pinocchio), have no sentimental value and must not be used to determine sentiment. Furthermore, words such as (jealousy, flaws) in the positive list must be in the negative. Most of the words are good sentiment indicators, but the model takes some liberties that are not always correct/accurate.

Step 5 Naïve Bayes vs Rule-Based:

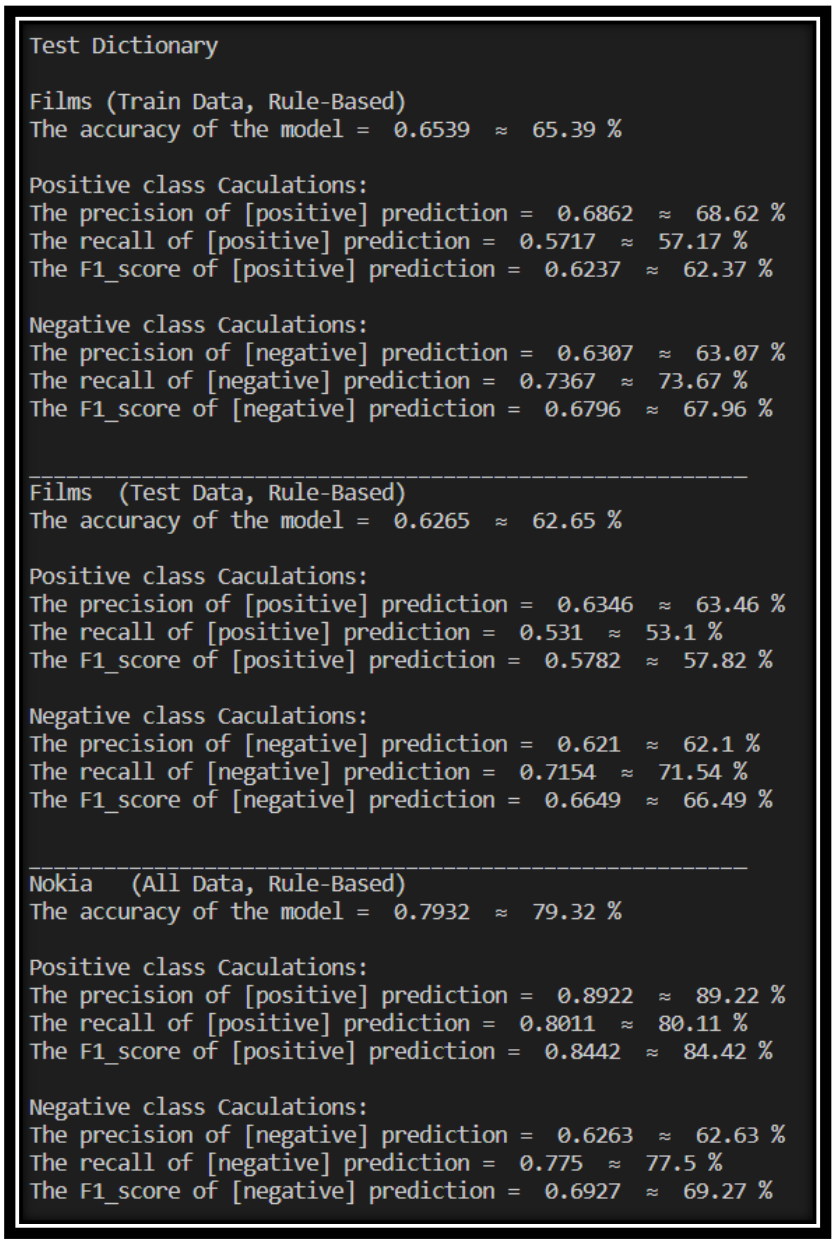


Figure 05: Results of running the Test Dictionary model on both the film and Nokia datasets.

From **Figures (02, 05)**, it is evident that the rule-based approach scored lower on average on the film set, but considerably higher on the Nokia set.

The Bayesian approach is trained on the Films dataset, while the Nokia dataset is ignored. Consequently, it performs better on the Films dataset.

It is worth noting that, as shown in **Figure 05**, the rule-based approach performs nearly identically on training and test data from the Films dataset due to the absence of overfitting.

To estimate the sentiment, Naive Bayes uses supervised training sets to generate statistics about the word's sentiment, as well as learning patterns, linguistic nuance, and word relationships. It performs well with sentiment analysis and can, to some extent, handle complex language structures. However, it requires a substantial amount of data to perform well across all domains as evident from the issue with the Nokia set. Furthermore, overfitting might cause the model to learn noise patterns impacting its quality.

The Dictionary-based approach uses predefined sentiment dictionaries, scores, and rules, making decisions based on a threshold. The approach is simple to implement, performs well where language patterns and rules are clearly defined, does not require a vast amount of data, and allows for manual control over language rules and patterns. However, it lacks the sophistication to process nuances and context effectively.

The trade-off lies between the statistical model's adaptability to diverse data, cost, and complexity, and the rule-based model's precision but limited scope of predefined rules.

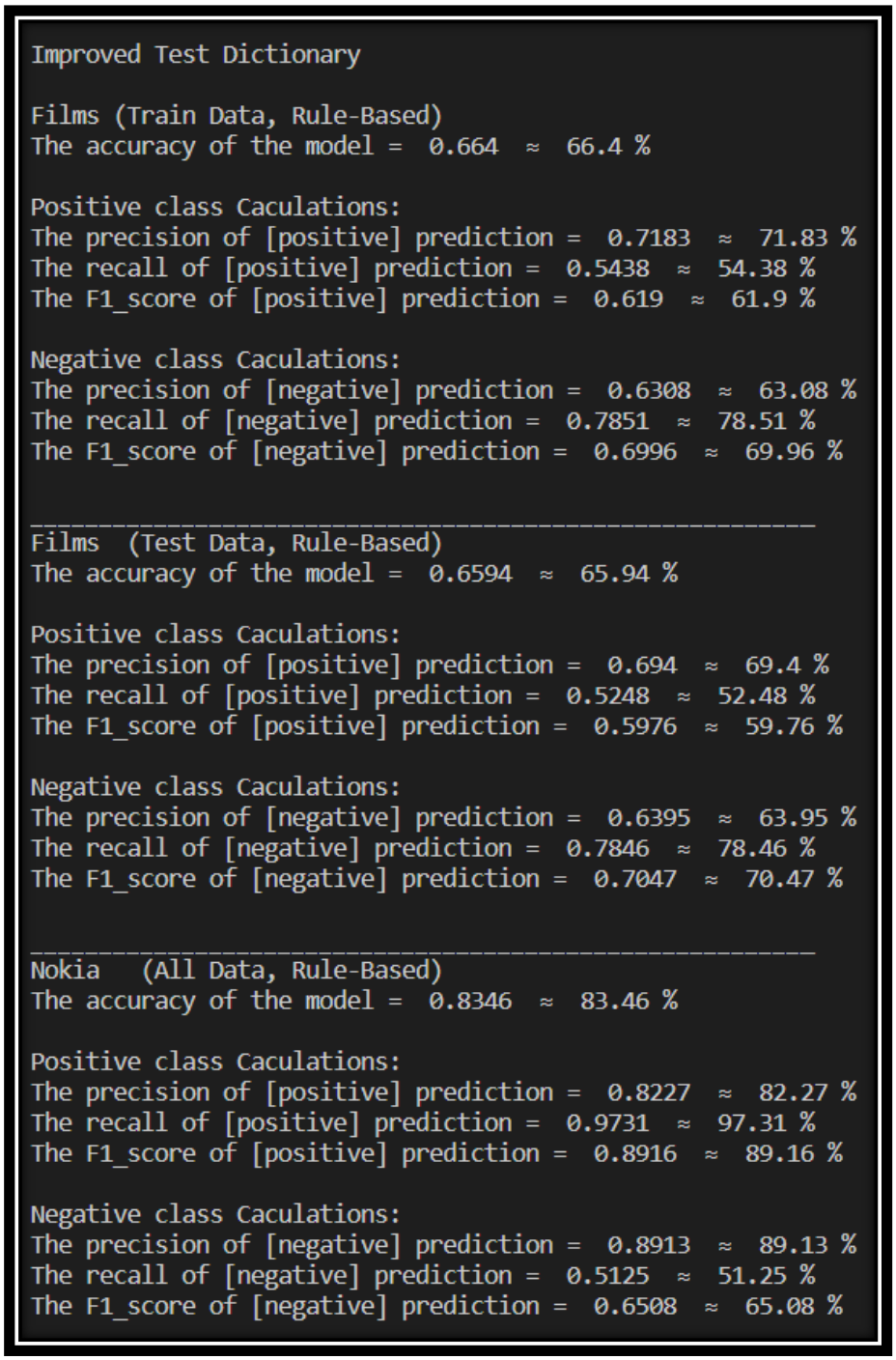


Figure 06: Results of running the improved Test Dictionary model on both the film and Nokia datasets.

The new implementation $improved\_testDictionary()$ considers rules like negation, diminishers, amplifiers, and coordinate-conjunction(but). It attempts to better analyse sentiments by adding or subtracting to sentence sentiment score when faced with diminishers or amplifiers. It negates the score when faced with negation and resets the score after facing coordinate-conjunction(but) neglecting sentiments before the word but.

Figure 06 shows an improvement of 1%-4% over both sets for the new function. The improvement might be considered small; however, language context/nuance is complex and requires an equally complex model for substantial improvement.

Step 6 Analysing Errors:

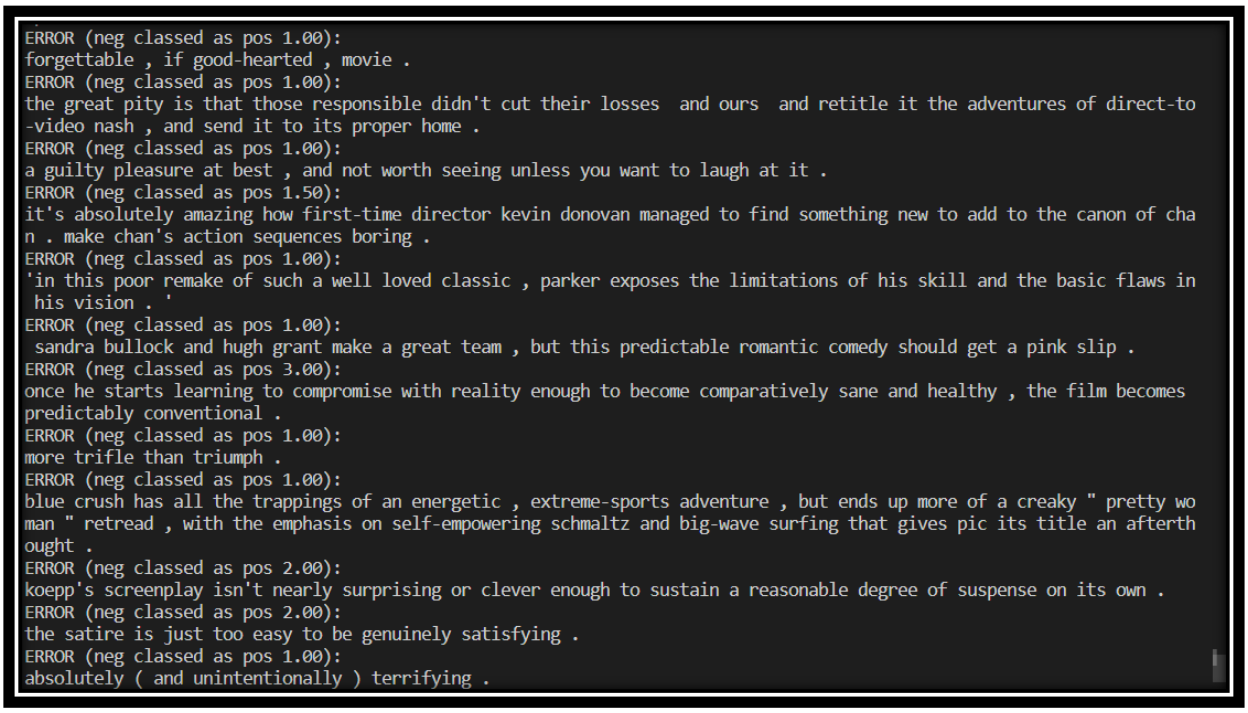


Figure 07: Test Dictionary errors as printed in the terminal.

Both models generate numerous errors, some of which are displayed on the terminal, **Figure 07**. For ease of reading and analysis, the error tables below contain samples of the errors produced by both models.

#	Score	Naive Bayes Error Table (Positive Classified as Negative)
1	0.47	manages to be original, even though it rips off many of its ideas.
2	0.00	despite the long running time, the pace never feels slack -- there's no scene that screams " bathroom break! "
3	0.01	equilibrium is what george orwell might have imagined had today's mood-altering drug therapy been envisioned by chemists in 1949.
4	0.28	cool? this movie is a snow emergency.
5	0.04	a very pretty after-school special. it's an effort to watch this movie, but it eventually pays off and is effective if you stick with it.

Table 01: shows Bayes approach errors of positive sentimental sentences classified as negative.

#	Score	Naïve Bayes Error Table (Negative Classified as Positive)
1	0.99	it's a decent glimpse into a time period, and an outcast, that is no longer accessible, but it doesn't necessarily shed more light on its subject than the popular predecessor.
2	0.77	must-see viewing for anyone involved in the high-tech industry. others may find it migraine-inducing, despite moore's attempts at whimsy and spoon feeding.
3	0.62	none of birthday girl's calculated events take us by surprise . . .
4	0.80	absolutely (and unintentionally) terrifying.
5	0.51	something must have been lost in the translation.

Table 02: shows Bayes approach errors of negative sentimental sentences classified as positive.

Scores in the Naïve Bayes are very close to the threshold (0.5) meaning the margin of error is not too significant and the model might largely benefit from a few tweaks.

Sentences[1, 2] Table(01) & sentence[2] Table(02):

The model struggles when sentences are divided into positive and negative sections and when sentences are positive in context but structured using words with negative sentiment.

Sentence[3] Table(01) & sentence[3, 5] Table(02):

The model struggles with complex contexts involving slang/idioms/sarcasm.

Sentence[4] Table(01):

Ambiguous sentences, and wrong punctuation.

Sentence[5] Table(01) & sentence[1] Table(02):

The model struggles to classify sentences with coordinate-conjunction(but), maybe the training data should include more complex sentences with coordinate conjunctions.

Sentence[4] Table(02):

The use of consequent multiple amplifiers and special characters might be affecting the model's decision.

#	Score	Rule-Based Error Table (Positive Classified as Negative)
1	-1.00	if the very concept makes you nervous... you'll have an idea of the film's creepy, scary effectiveness.
2	0.00	'compleja e intelectualmente retadora , el ladrón de orquídeas es uno de esos filmes que vale la pena ver precisamente por su originalidad . '
3	0.00	the year's happiest surprise, a movie that deals with a real subject in an always surprising way.
4	-2.00	if director michael dowse only superficially understands his characters, he doesn't hold them in contempt.
5	-3.00	if the film has a problem, its shortness disappoints: you want the story to go on and on.

Table 03: shows Test Dictionary errors of positive sentimental sentences classified as negative.

#	Score	Rule-Based Error Table (Negative Classified as Positive)
1	1.00	absolutely (and unintentionally) terrifying.
2	2.00	the satire is just too easy to be genuinely satisfying.
3	1.00	sandra bullock and hugh grant make a great team, but this predictable romantic comedy should get a pink slip.
4	3.50	the problem, amazingly enough, is the screenplay.
5	4.50	it won't harm anyone, but neither can i think of a very good reason to rush right out and see it. after all, it'll probably be in video stores by christmas, and it might just be better suited to a night in the living room than a night at the movies.

Table 04: shows Test Dictionary errors of negative sentimental sentences classified as positive.

The scores for the rule-based approach deviate considerably from the threshold(1), signifying a considerable margin of error. Consequently, substantial tweaks might be necessary to enhance the model. The approach has no consideration for context, it simply relies on the presence of sentimental words in a dictionary and their score; thus, errors must be analysed accordingly.

Sentence[2] Table(01), is written in Spanish, the dictionary only includes English words, hence, the score is 0.

Sentence[3] Table(01), even though the sentence is written in English, the dictionary does not include any of the words in the sentence, hence, the score is 0.

The Test Dictionary is prone to errors for three reasons:

1. The count of positive words is greater than the negative and the context is negative or vice versa.
2. The dictionary does not include all the sentimental words in the sentence, or the words are of different languages.
3. The rules are insufficient to infer the context of the sentence. This is only true if rules like negation, diminisher, etc are included.

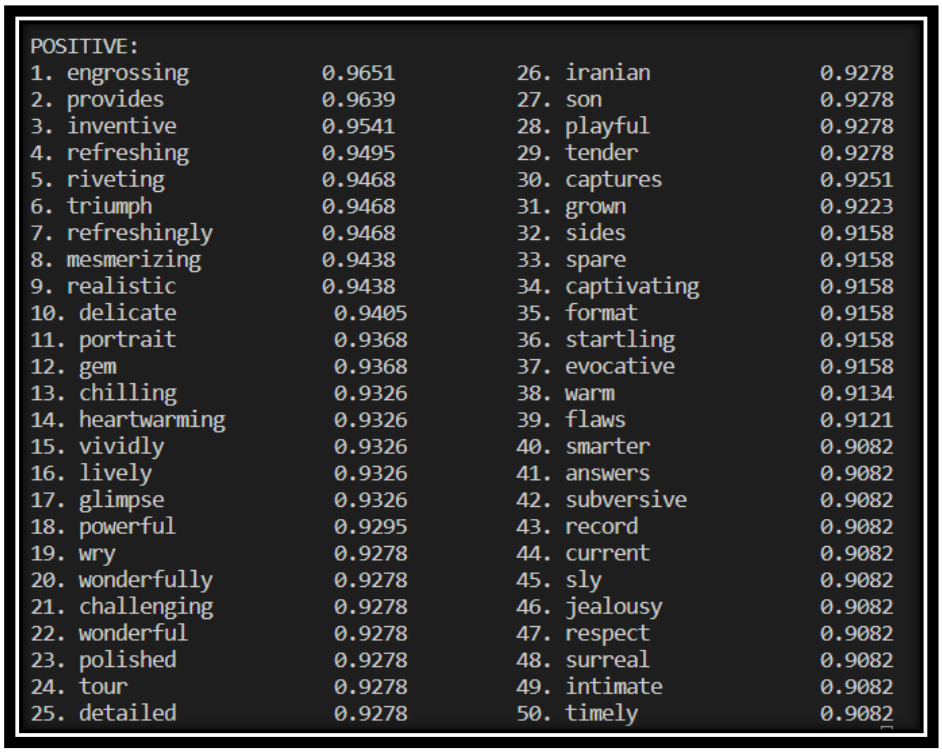
Appendix

Step 4 Extra

A suggested fix is to reverse the list of the most useful words for positive sentiment, as derived from the tail of an ascending order list. A new function, $improved\_mostUseful()$, does this, as well as edit the output style and include the probability score to make the result more understandable.

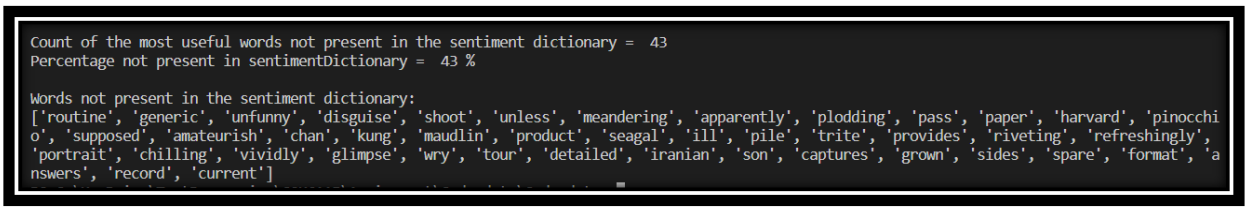
```
------------------------------------------------------------
The most useful word for classifing positive and negative sentiments:

NEGATIVE:
1. routine          0.03957      26. pass           0.07612
2. generic          0.03957      27. horrible       0.07612
3. mediocre         0.04122      28. wasted         0.07612
4. badly            0.04122      29. paper          0.07612
5. unfunny          0.04301      30. harvard        0.07612
6. waste            0.04946      31. pinocchio      0.07612
7. poorly           0.05207      32. supposed       0.07917
8. disguise         0.05496      33. junk           0.08247
9. mindless         0.0582       34. amateurish     0.08247
10. pointless        0.0582      35. banal          0.08247
11. bore             0.0582      36. chan           0.08247
12. stale           0.06184      37. kung           0.08247
13. devoid          0.06184      38. maudlin        0.08247
14. offensive       0.06184      39. pathetic       0.08247
15. shoot           0.06184      40. product        0.08247
16. boring          0.06316      41. incoherent     0.08247
17. annoying        0.06597      42. seagal         0.08247
18. tiresome        0.06597      43. lifeless       0.08247
19. dull            0.06872      44. lame           0.08606
20. stupid          0.07068      45. ill            0.08998
21. unless          0.07068      46. stealing       0.08998
22. meandering      0.07068      47. fatal          0.08998
23. apparently      0.07068      48. pile           0.08998
24. plodding        0.07068      49. trite          0.08998
25. worse           0.07612      50. lousy          0.08998
```



```
POSITIVE:
1. engrossing          0.9651        26. iranian            0.9278
2. provides            0.9639        27. son                0.9278
3. inventive           0.9541        28. playful            0.9278
4. refreshing          0.9495        29. tender             0.9278
5. riveting            0.9468        30. captures           0.9251
6. triumph             0.9468        31. grown              0.9223
7. refreshingly        0.9468        32. sides              0.9158
8. mesmerizing         0.9438        33. spare              0.9158
9. realistic           0.9438        34. captivating        0.9158
10. delicate            0.9405       35. format             0.9158
11. portrait           0.9368        36. startling          0.9158
12. gem                0.9368        37. evocative          0.9158
13. chilling           0.9326        38. warm               0.9134
14. heartwarming       0.9326        39. flaws              0.9121
15. vividly            0.9326        40. smarter            0.9082
16. lively             0.9326        41. answers            0.9082
17. glimpse            0.9326        42. subversive         0.9082
18. powerful           0.9295        43. record             0.9082
19. wry                0.9278        44. current            0.9082
20. wonderfully        0.9278        45. sly                0.9082
21. challenging        0.9278        46. jealousy           0.9082
22. wonderful          0.9278        47. respect            0.9082
23. polished           0.9278        48. surreal            0.9082
24. tour               0.9278        49. intimate           0.9082
25. detailed           0.9278        50. timely             0.9082
```

Figure 08: Most useful words as outputted by the improved_mostUseful function.



```
Count of the most useful words not present in the sentiment dictionary =  43
Percentage not present in sentimentDictionary =  43 %

Words not present in the sentiment dictionary:
['routine', 'generic', 'unfunny', 'disguise', 'shoot', 'unless', 'meandering', 'apparently', 'plodding', 'pass', 'paper', 'harvard', 'pinocchi
o', 'supposed', 'amateurish', 'chan', 'kung', 'maudlin', 'product', 'seagal', 'ill', 'pile', 'trite', 'provides', 'riveting', 'refreshingly',
'portrait', 'chilling', 'vividly', 'glimpse', 'wry', 'tour', 'detailed', 'iranian', 'son', 'captures', 'grown', 'sides', 'spare', 'format', 'a
nswers', 'record', 'current']
```

Figure 09: the number and words of the most useful words absent from the sentiment dictionary.

Figure 09 is an addition showing the words missing from the sentiment dictionary. As expected, the list includes all the words mentioned above as having no sentimental value, indicating that the problem is with the model rather than the dictionary.