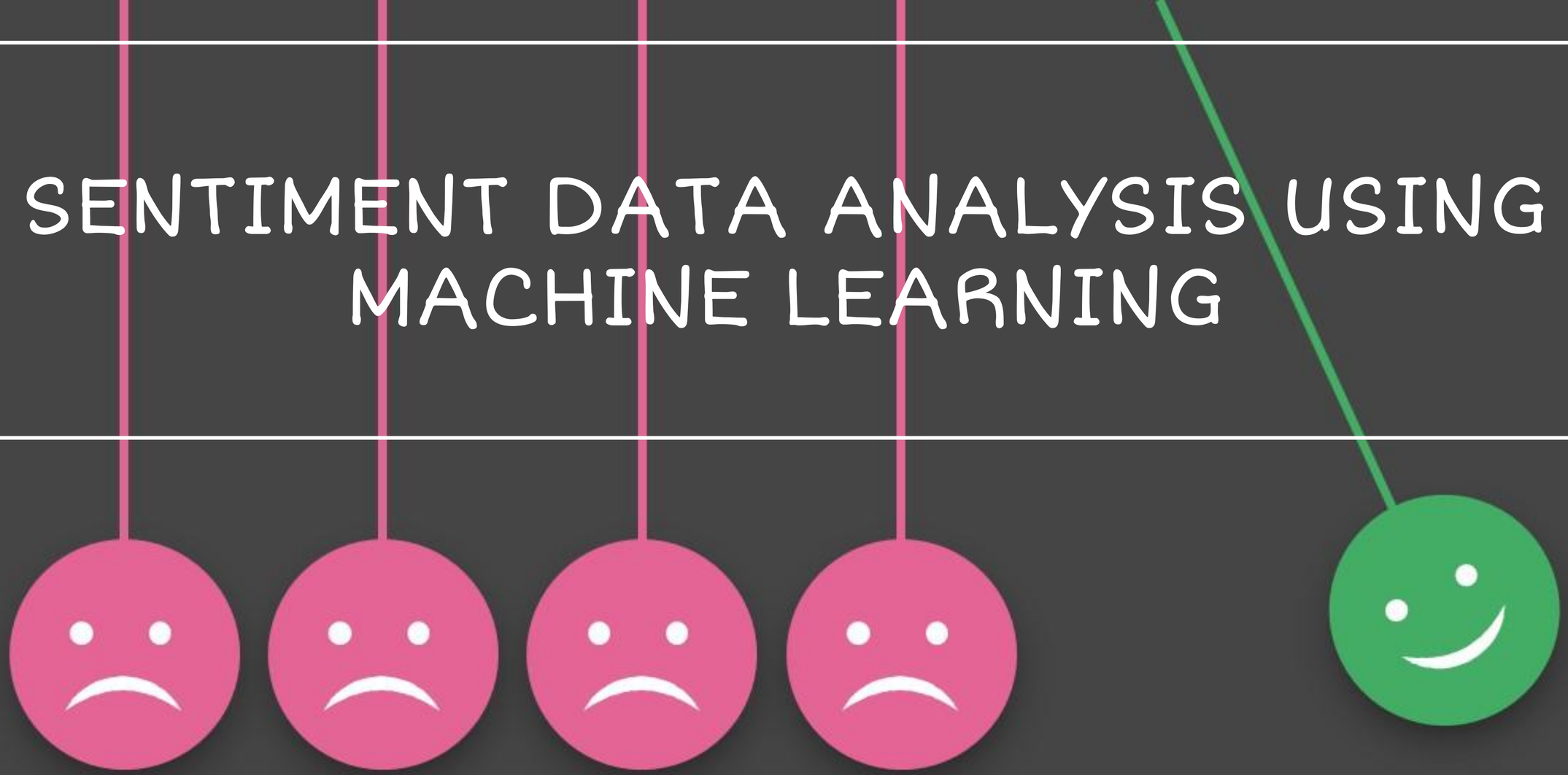
SENTIMENT DATA ANALYSIS USING MACHINE LEARNING



David Nguyen-Huu (26659330)
Shagana Mahendrarajah(40015699)
Mikaeil Sarkis (40037086)
Henry Dang(40131548)

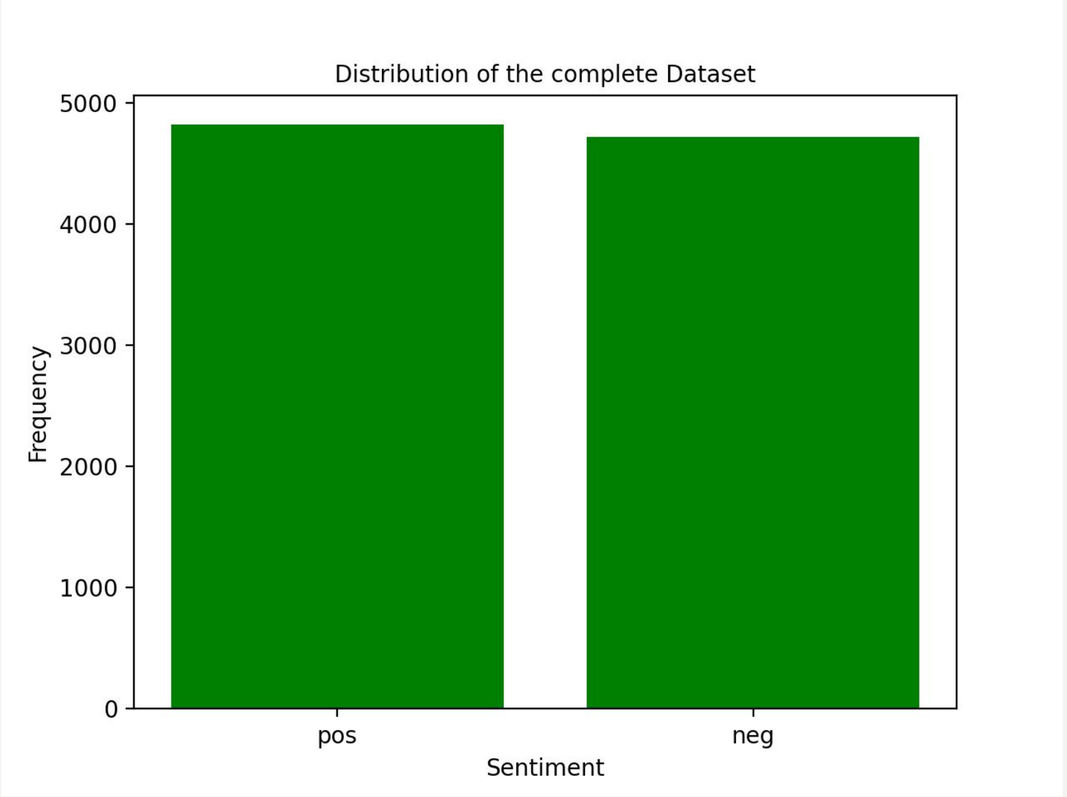
Goal

TRAIN DIFFERENT MACHINE LEARNING MODELS

COMPARE EACH MODEL'S ACCURACY

INITIAL DATA

Data	Frequency
POS	4817
NEG	4714

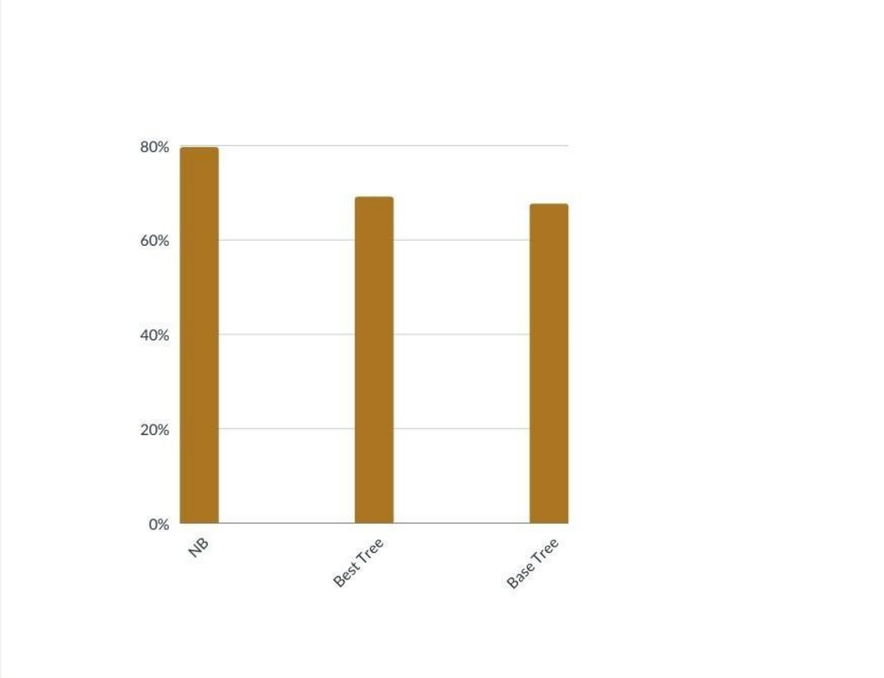


DATA SET ANALYSIS

- Huge number of records > 10000 reviews
- Only 2 labels to predict “pos , neg”
- The only features that can be used are the words in each review which is associated to a label. In similar way to filtering spam emails
- Different length between records which can effect the probability of predicating the correct label since shorter reviews will have a higher probability to get the wrong label
- The data distribution between the 2 labels is very close which can be a good factor while calculating the accuracy of the learning model

MACHINE LEARNING DATA MODELS

- We have used 3 models to predict the labels for each review in the evaluation set so we got a different results for each learning model.
- The models ranking based on their performance:
- The Naive Bayes classifier which was the most accurate model
- The Best decision tree which was less accurate than NB model but more accurate than the Base decision tree
- The Base decision tree which was the least accurate model of the three

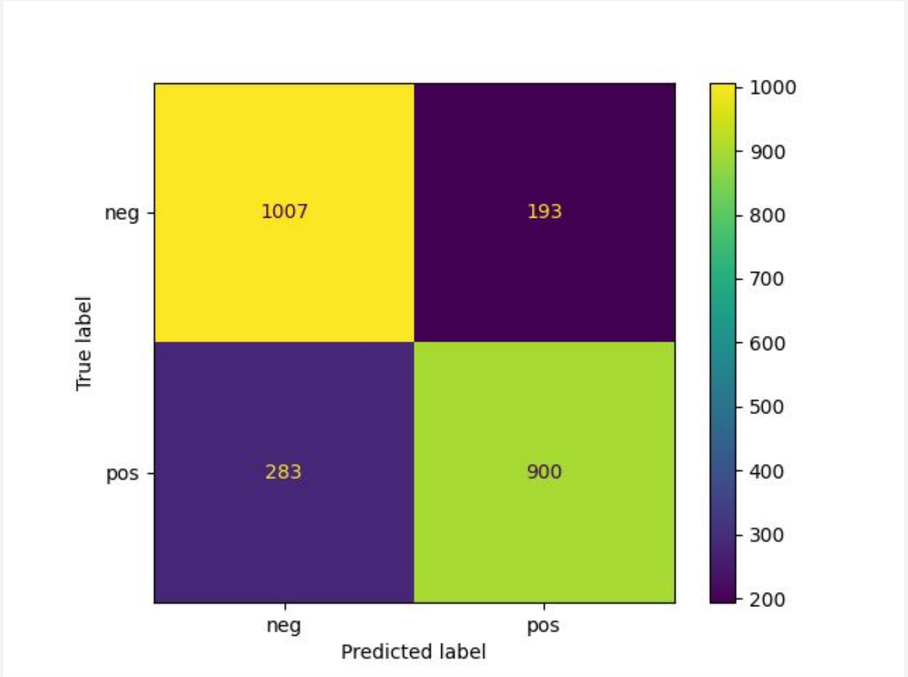


NAIVE BAYES CLASSIFIER

- The results for this model were the best in comparison to the other 2 models

Accuracy	Precision	Recall	Fscore
0.8002517834662191	0.8018682944986723	0.8002517834662191	0.7999123633383931

The confusion matrix



ANALYZING THE RESULTS:

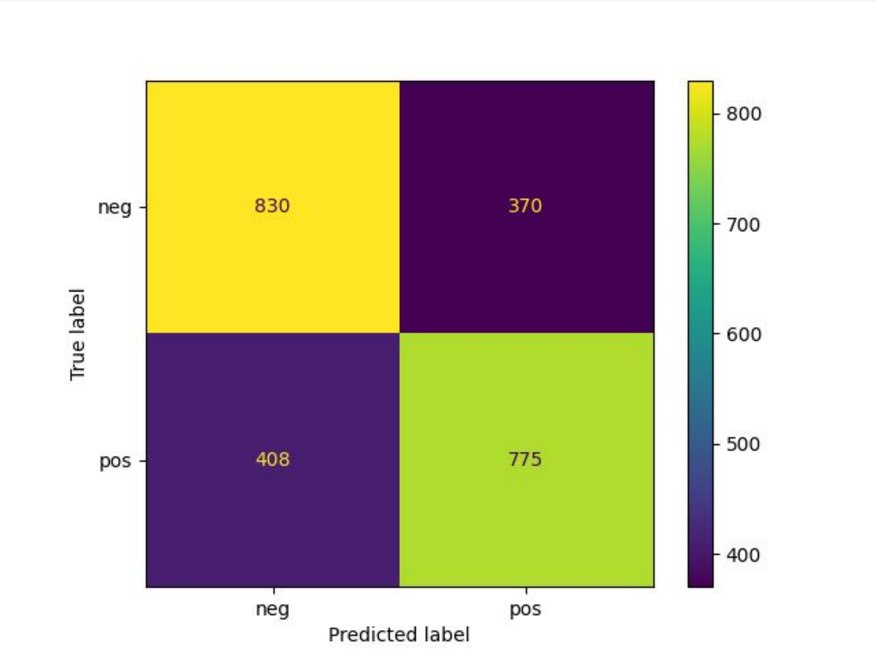
- The model was able to predicate the correct label for 80% of the reviews when we used 0.9 smoothing
- For the pos label the system predicated 900 reviews that were pos and 193 reviews that were neg
- For the neg label the system predicated 1007 reviews that were neg and 283 reviews that were pos
- The reason why the system failed in classifying the right label for some of the reviews can be related the number of words in these reviews since we are using these words as our features so the reviews that have a fewer words are going to be harder to classify
- If we take an example where a review have a small number of words then when we will have less features thus when calculating the probabilities to rank the hypothesis we will get a close results which can lead to a wrong predication

BASE DECISION TREE

- The results for this model were not as good as in the other 2 models

Accuracy	Precision	Recall	Fscore
0.686109945446915 6	0.6862280090045658	0.6861099454469156	0.6859943589333348

The confusion matrix



ANALYZING THE RESULTS:

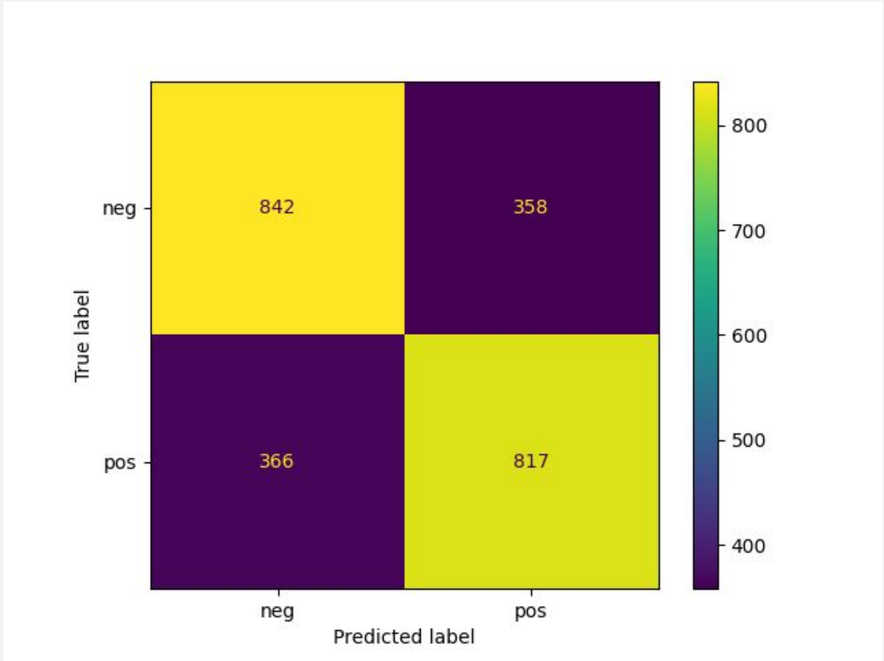
- The model was able to predicate the correct label for over 68% of the reviews
- For the pos label the system predicated 775 reviews that were pos and 370 reviews that were neg
- For the neg label the system predicated 830 reviews that were neg and 408 reviews that were pos
- Unlike the Naive Bayes classifier where all the features were treated equally and used in calculating the probabilities to predicate the correct label we had in the Decision tree to rank the features.
- The reason why the Base tree was less accurate than the Naive Bayes can be related to how discriminate the feature that was chosen as the root in the tree so in case we have a few number of features that are not providing a precise classification then we can end up with the wrong label.

BEST DECISION TREE

- The results for this model were better than the Base tree but not as good as NB

Accuracy	Precision	Recall	Fscore
0.6961812840956777	0.6961755745913101	0.6961812840956777	0.6961705826060184

The confusion matrix



ANALYZING THE RESULTS:

- The model was able to predicate the correct label for over 69% of the reviews
- For the pos label the system predicated 775 reviews that were pos and 370 reviews that were neg
- For the neg label the system predicated 830 reviews that were neg and 408 reviews that were pos
- The results were slightly better than the base tree since we used to calculate the entropy of each feature to determine the information gain which will help in choosing the root for the tree so we can get a smaller tree
- But as in the previous models if we have a small number of features that are not discriminate like the case when we try to classify short reviews then even with using the entropy as a guide to choose the tree root we will still get a wrong label since the entropy will be high and this will increase the ambiguity about the correct label of the given review.

THE PRECISION , THE RECALL , THE FSCORE

- The results of The Precision and the recall and the fscore were similar to the accuracy and the reason of these results is the distribution of the data since the 2 labels are reparented equally as we saw in the distribution plot
- When calculating the Precision, recall, fscore we used the weighted Harmonic mean which takes into account the weight of each class distribution and as we mentioned since we had an equal distribution the results were similar to the accuracy in the three models.

Error Analysis

- Naive Bayes Classifier assumes that features are independent from each other

  878.txt: “i am a huge Thoman Hardy fan , and i was not disappointed”

  The model predicted neg when it should be pos

- Books/Movies

  167.txt: “jmaes joyce takes us on stephen daedalus ' interior journey from pre-teen as son of a country gentleman to young adult who wanders the streets of dublin , struggling with sin”

  Decision Tree models predicted neg when it should be pos

- Short reviews/Reviews starting with a positive statement and ends with a negative statement and vice versa

Contribution

Name	Student ID	Contributions
Mikaeil Sarkis	40037086	• Worked on task 0 and task 2 • Helped with presentation slides
Shagana Mahendrarajah	40015699	• Worked on task 0 and task 1 • Helped with presentation slides
David Nguyen-Huu	26659330	• Worked on task 3 • Helped with presentation slides
Henry Dang	40131548	• Worked on task 4 • Helped with presentation slides