NLP(No DEEP) - Report Lab 2

Eliot LECLAIR, Alex POIRON, Tom THIL, Aurelien VISENTIN

1 The dataset

1.1 How many splits does the dataset has ?

We have 3 splits : train, test, unsupervised

1.2 How big are these splits ?

- train: 25.000 entries
- test: 25.000 entries
- unsupervised: 50.000 entries

1.3 What is the proportion of each class on the supervised splits?

For both train and test we have 12.500 values with positive label and 12.500 values with negative values. For both, we have 50/50 class representation.

2 Naive Bayes Classifier

2.1 Report the accuracy on both training and test set

- Train Accuracy: 0.91284
- Test Accuracy: 0.8172

2.2 Why is accuracy a sufficient measure of evaluation here?

It's a sufficient measure because we just have to know if in a binary classification, the text is a positive label or a negative one.

2.3 What are the top 10 most important words (features) for each class? (bonus points)

Top most important word for negative review (no filter):

- 'the', 'and', 'of', 'to', 'is', 'in', 'this', 'it', 'that', 'br'

Top most important word for positive review (no filter):

- 'the', 'and', 'of', 'to', 'is', 'in', 'it', 'this', 'that', 'br'

Top most important word for negative review (filtered):

- 'movie', 'film', 'one', 'like', 'even', 'good', 'bad', 'would', 'really', 'time'

Top most important word for positive review (filtered):

- 'film', 'movie', 'one', 'like', 'good', 'story', 'great', 'time', 'see', 'well'

2.4 Take at least 2 wrongly classified example from the test set and try explaining why the model failed

First wrongly classified:

- **Text:** 'yep edward g gives us a retro view of the criminal defense world first hes an overzealous prosecutor who sends the wrong man to the chair played passionately albeit briefly by deforrest kelly then hes so filled with remorse his only solace is the bottle throw in a jaded romance a genuinely rapid descent into penury and no qualms about who he defends and next thing you know shazam black leg lawyer god i love that phrase he sees the light just in time to save his jaded beloved from the chair yawnbr br but really the courtroom action is pure melodrama see him punch out a witness see him drink poison see him argue passionately as he clutches a bullet hole in his breast be prepared for melodramabr br the hoot of the film though is jayne russell with curvesdefying the laws of gravity and an iq approached absolute zero she is something to see even sings a bit'
- **Label:** 0

Second wrongly classified:

- **Text:** 'king vladislav angus scrimm of romania is a vampire but a vampire of light who wants nothing more than to live in peace and harmony with mankind but his son radu anders hove is a cruel creature to his very heart which is pretty obvious as soon as you see him three female students have come to study local folklore but find themselves drawn into the vampires legends at just the wrong time vladislav has been killedbr br who can say anything bad about a film featuring a cameo from angus scrimm i canti mean i had some low expectations after seeing other full moon pictures puppet master in particular and demonic toys but despite the really bad animated effects of the demons this film was actually really well done and very fun to watch plenty of blood a good plot and backstory the bloodstone story was surprisingly refreshing and even some new angles on the vampire mythos which youd think would be dead by now maybe im wrong but this is probably the first film to feature rosary beads being fired from a gun aside from vampires and blood you get a share of nudity gratuitous but welcome and i had to notice the excellent score from the composers not sure who deserves credit but those involved include stuart brotman richard kosinski william levine michael portis and john zeretzka this is horror 101 all the way heck you even get two sequels which is the sign of a true horror film of course some bad films get sequels too did i mention puppet master the romanian theme was welldone and the film even seems to have been made by romanians if i am guessing their name origins correctly and the score the music really stood out for me as a nice change of pace very moodsetting i like richard band but im glad another composer was given a shot because he nailed the atmosphere on the head if you like vampire films and want a slight variation one of the eastern european variety this is worth seeing'
- **Label:** 1

In the first example, the user uses sarcasm to review the film, which is difficult to interpret for the model. Moreover, he also gives some positive reviews which can distort the result a bit. The second one is more complicated to interpret, its construction is more complex and does not only consist in giving its opinion but also its feeling and its expectations before seeing the film, which the model has difficulty in perceiving.

3 Stemming and Lemmization

3.1 Are the results better or worse? Try explaining why the accuracy change

New accuracy with stemming:

- Train Accuracy: 0.8828
- Test Accuracy: 0.80516

Old accuracy:

- Train Accuracy: 0.91284
- Test Accuracy: 0.8172

We can see that results are a little bit worse. The stemming step may be not adapted to the model we have and the classification we want to do.