Quentin Le Helloco, Etienne Sharpin et Ségolène Denjoy

NLP Report

Cinema Reviews Sentiment Analysis

See subject here : https://github.com/mvonwyl/epita/tree/master/NLP/03
and there : https://github.com/mvonwyl/epita/tree/master/NLP/04

Table of Content

1. Naive Bayes model

We chose to use the naive Bayes model from scikit-learn because of it's high simplicity.
https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

1.1 Results analysis

We can see that overall, preprocessing only worsen the results.

The accuracy is 1% less when removing stop-words than without. Same with stemming, accuracy plummets by about 4% and 3% for lemming.

Nonetheless, the model gains 2% when we only check for the presence of words instead of counting them.

The best model we achieved does not use preprocessing and uses binary Naive Bayes. It has 82% precision, recall, f1-score and accuracy (same for all).

```
In [11]: get_gnb_score(basic, binary=True)
```

```
              precision    recall  f1-score   support

           0       0.82      0.82      0.82     12500
           1       0.82      0.82      0.82     12500

    accuracy                           0.82     25000
   macro avg       0.82      0.82      0.82     25000
weighted avg       0.82      0.82      0.82     25000
```

1.2. Wrongly classified samples

I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.

I first saw this movie on IFC. Which is a great network by the way to see underground films. I watched this movie and was thinking it was going to be pure drama and a story line that doesn't hold water. But it really was a worth while watch. The main character is in such rough shape, and you hate to see him deny help, but no matter what you just can't hate him. His devotion to The Beatles and John Lennon is a great metaphor for his life and the helplessness he feels.

The atmosphere of the film is also great. At times, you feel like you can see what he sees, feel what he feels in some situations. This movie does not leave you wanting to know more, or disliking a loophole in the plot. There are NO loopholes (in my opinion). I have always been a fan of foreign films, especially now with movies being made so poorly in America. I really enjoy the foreign settings because I feel it can take you on a trip, and sometimes understand a different culture. This movie did all those things to me and more. Please watch this movie and if you're new to foreign films, this is a great start.

Both are classified as positive when they should be negative.
If we take those two samples that were mispredicted, we can see that they use negative words as well so their feelings are more implicit.w

2. Logistic Regression

2.1. Preprocessing

Initial conditions :

Regularization = l2

Solver = saga

Preprocessing	Mean Accuracy	AccPos	AccNeg	recallPos	recallNeg	f1scorePos	f1scoreNeg	Notes
None	0.7034	0.69454434	0.71310033	0.72616	0.68064	0.71000039	0.69649216	
stopswords	0.71272	0.70143939	0.72533898	0.74072	0.68472	0.72054475	0.70444444	best strategy
stemming	0.68688	0.68333072	0.69056942	0.69656	0.6772	0.68988194	0.68381937	
lemming	0.70844	0.70136796	0.71602686	0.726	0.69088	0.71347144	0.7032287	
lemming + stopwords	0.70412	0.69360346	0.71584468	0.73128	0.67696	0.71194361	0.69585955	we thought it could be the best strategy but no

The best strategy is the stop-words removal. It's quite logical as some words with no meaning could be taken into account even with the Vader lexicon and it's sentiment rating.

Small improvement with lemming but not the 'giant step' of the stopwords removal. We thought using the lemming and stopword-removal together could yield better results but no.

2.2. Already existing features

We fine-tuned the threshold value of the features sentiment-related.

Initial conditions :

Without preprocessing

Regularization = l2

Threshold	Accuracy	Notes
1	0.69668	
1.8	0.69488	
1.9	0.70248	Best strategy Solver : saga
2	0.6978	
2.2	0.69508	
3	0.64512	
1.2 / -2	0.679	

2.3. More features

We added 3 more features :

Number of "."
Number of words of length < 3 characters
Number of ":" (simplist version of a smiley, could do a regexp)

2.3.a. Accuracy when adding the new features

Initial conditions :

Without preprocessing
Regularization = l2
Threshold = 2

* Accuracy without any feature = 0.6978
* Accuracy with feature "." = 0.69776
* Accuracy with feature word len < 3 = 0.69724
* Accuracy with feature simplist smiley = 0.69808

So we can think that the feature word len is not really good but :
* Accuracy with feature "." and smiley = 0.69748
* Accuracy with all new features = 0.69912

2.4. Regularization and solver

2.4.a. Regularization

We can do it directly in the logistic regression of sklearn as the parameter "penalty"

	None	l2
Accuracy	0.70244	0.70248

Solver : by-default

2.4.b. Solver

Solver	None	l1	l2	Notes
lbfgs	0.70244	/	0.70248	is the by-default solver
Saga	0.70336	0.7034	0.70444	max_iter = 100000
newton-cg	0.70244	/	/	l2 couldn't converge
sag	0.70296	/	0.70296	

liblinear	/	0.70276	0.70288	

2.5. Wrongly classified samples

2.5.a. Examples

Four things intrigued me as to this film - firstly, it stars Carly Pope (of "Popular" fame), who is always a pleasure to watch. Secdonly, it features brilliant New Zealand actress Rena Owen. Thirdly, it is filmed in association with the New Zealand Film Commission. Fourthly, a friend recommended it to me. However, I was utterly disappointed. The whole storyline is absurd and complicated, with very little resolution. Pope\'s acting is fine, but Owen is unfortunately under-used. The other actors and actresses are all okay, but I am unfamiliar with them all. Aside from the nice riddles which are littered throughout the movie (and Pope and Owen), this film isn\'t very good. So the moral of the story is...don\'t watch it unless you really want to.

and

David Bryce\'s comments nearby are exceptionally well written and informative as almost say everything I feel about DARLING LILI. This massive musical is so peculiar and over blown, over produced and must have caused ruptures at Paramount in It cost million dollars! That is simply irresponsible. DARLING LILI must have been greenlit from a board meeting that said "hey we got that Pink Panther guy and that Sound Of Music gal... lets get this too" and handed over a blank cheque. The result is a hybrid of GIGI, ZEPPELIN, HALF A SIXPENCE, some MGM song and dance numbers of a style (daisies and boaters!) so hopelessly old fashioned as to be like musical porridge, and MATA HARI dramatics. The production is colossal, lush, breathtaking to view, but the rest: the ridiculous romance, Julie looking befuddled, Hudson already dead, the mistimed comedy, and the astoundingly boring songs deaden this spectacular film into being irritating. LILI is like a twee mega musical with some vulgar bits to spice it up. STAR! released the year before sadly crashed and now is being finally appreciated for the excellent film is genuinely is... and Andrews looks sublime, mature, especially in the last half hour......but LILI is POPPINS and DOLLY frilly and I believe really killed off the mega musical binge of the 60s..... and made Andrews look like Poppins again... which I believe was not Edwards intention. Paramount must have collectively fainted when they saw this: and with another $million festering in CATCH and $million in ON A CLEAR DAY and $million in PAINT YOUR WAGON....they had a financial abyss of CLEOPATRA proportions with $million tied into 4 films with very uncertain futures. Maybe they should have asked seer Daisy Gamble from ON A CLEAR DAY ......LILI was very popular on immediate first release in Australia and ran in 70mm cinemas for months but it failed once out in the subs and the sticks and only ever surfaced after that on one night stands with ON A CLEAR DAY as a Sunday night double. Thank god Paramount had their simple $1million (yes, ONE MILLION DOLLAR) film LOVE STORY and that $4 million dollar gangster pic THE GODFATHER also ready to recover all the $million in just the next two years....for just $5m.... incredible!

Both are classified as positive when they should be negative.

The first one starts really well but concludes on something negative which is not taken into account since we do not take word order and context into account.

The second one is purely ironic.

This is why those examples are misclassified.

2.6. Conclusion

As the last strategy we implemented here was the preprocessing, we computed all the results with the solver saga. Maybe it is not the best. We should redo all the tests like in the Solver section but with the stopwords.

To conclude, the best strategy we saw was :

- preprocessing : stopwords removal

- solver : saga
- regularization : l2
- new features : all

With an accuracy of 71%.

To better improve our model, we could add even more features.

3. FastText

FastText is a library providing a text classifier. We will use it to get a better score for our sentiment prediction project.

Once our data was correctly formatted as a FastText input file, we trained a first model using default settings of FastText in supervised learning mode.

3.1. First model

The first train was made with an untreated dataset, and the accuracy was 0.86.

	precision	recall	f1-score	support
0	0.86	0.86	0.86	12500
1	0.86	0.86	0.86	12500
accuracy			0.86	25000
macro avg	0.86	0.86	0.86	25000
weighted avg	0.86	0.86	0.86	25000

To increase these results, we first tried to get better hyperparameters for the model.

3.2. Hyperparameters tuning

Fortunately FastText possesses an autotuning method which only requires a validation set out of the full train dataset to set the best score possible.

We used of the train dataset as validation set and after 5 minutes of autotuning, the scores increased by 0.01%

	precision	recall	f1-score	support
0	0.87	0.88	0.87	12500
1	0.87	0.87	0.87	12500
accuracy			0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

In order to increase this score, we could let autotune execute longer but the best option is to apply treatment to the input data to remove stopwords and use lemming/stemming.

3.3. Dataset pretreatment

Pretreating the data may be the most complicated part because of the large volume of samples in the dataset. Applying stopword removal, stemming and/or flemming can be very long to execute.

3.3.a. Stop words removal

First we tried to only remove stop words which leads to slightly better results as with our first model.

	precision	recall	f1-score	support
0	0.88	0.87	0.88	12500
1	0.87	0.88	0.88	12500
accuracy			0.88	25000
macro avg	0.88	0.88	0.88	25000
weighted avg	0.88	0.88	0.88	25000

3.3.b. Stemming

Then we tried to use stemming on top of removing stop words which didn't increase the overall results.

	precision	recall	f1-score	support
0	0.88	0.87	0.87	12500
1	0.87	0.88	0.87	12500
accuracy			0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

3.3.c. Lemming

Finally we tried to use lemming (always with stop words removed) which also didn't increase the results (slightly lower scores than stemming).

	precision	recall	f1-score	support
0	0.87	0.87	0.87	12500
1	0.87	0.87	0.87	12500
accuracy			0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

We tried to tune hyperparameters for those 3 models for 5 minutes each, but none get better results from this method.

3.4. Wrongly classified sample analysis

If we look at some of the mispredicted samples, we can reveal some pattern that may indicate why the classification failed.

For example this sample was predicted as positive but labeled as negative:
'Four things intrigued me as to this film - firstly, it stars Carly Pope (of "Popular" fame), who is always a pleasure to watch. Secdonly, it features brilliant New Zealand actress Rena Owen. Thirdly, it is filmed in association with the New Zealand Film Commission. Fourthly, a friend recommended it to me. However, I was utterly disappointed. The whole storyline is absurd and complicated, with very little resolution. Pope\'s acting is fine, but Owen is unfortunately under-used. The other actors and actresses are all okay, but I am unfamiliar with them all. Aside from the nice riddles which are littered throughout the movie (and Pope and Owen), this film isn\'t very good. So the moral of the story is...don\'t watch it unless you really want to.'
This review contains a lot of positive words when describing the overall acting performance such as "*always a pleasure to watch*", "*brilliant*", "*recommended*", "*are all okay*", " *nice riddles*" which may lead to a positive final sentiment. But as the label states and we can also clearly see at the end "*this film isn't very good*", this review is negative.

Another example is this sample which was predicted negative but labeled as positive:
'The plot of this movie is as dumb as a bag of hair. Jimmy Smit plays a character that could have been upset by the ridiculousness of the story. He is evil and a wife beater. It's a character as far from his NYPD and LA Law roles as you could possibly get. If you've thought he had the looks and the acting chops to play the really bad boy role, her's your present.But!!!!!!!! Mary Louis Parker wears black miniskirts and little black minidresses throughout the movie. She has always had some of the greatest legs in the history of the movies. This makes the movie well worth it for this leg admirer. I'd buy the DVD for this reason only if it was available.'
The problem with this review is that the reviewer only uses irony to describe his sentiment. It leads to a lot of negative word like "is as dumb", "ridiculousness" and not much positive words.

To sum up, if the review contains irony or too many positive words, it can be mispredicted.

3.5. Improving model

To improve the previous score, we tried to take pretrained embedding from FastText english model and apply the average of each word vector as value for our classification model (SVM)

	precision	recall	f1-score	support
0	0.84	0.84	0.84	12500
1	0.84	0.84	0.84	12500
accuracy			0.84	25000
macro avg	0.84	0.84	0.84	25000
weighted avg	0.84	0.84	0.84	25000

We get a much worse score than using the FastText model directly.
To improve this score, we can either use our self-trained FastText instead of the pretrained one, change the average vector method used to create sample vectors or change the model used to classify (RandomForest, Linear regression, etc.).

3.6 Conclusion

FastText is a powerful library that gives better results than logistic regression and naive Bayesian with out of the box models. Pretreatment over the data doesn't increase the score too much nor does finding better hyperparameters. The best way to improve the model seems to play with the embedding but even there, if you look at the inside of FastText supervised_train() method, it's already taking embedding on our dataset and using a classifier to find results (and it does it much faster than our handmade algorithms). Using other models such as Word2Vec or GloVe could be a way to improve our prediction score.