Anvith Vobbilisetty

Findings and Comparisons of Binary Classification Models

Results

Model Results on UDHR

Model	Vectorizer (EN vs NL)	Train Acc	Dev Acc	Test Acc	Test Confusion Matrix *(rows=true, cols=pred; labels [EN=0, DU=1])*
Linear SVM	TF-IDF char n-grams (3–5)	1.00	1.00	1.00	`[[20, 0], [0, 20]]`
Logistic Regression	TF-IDF char n-grams (3–5)	1.00	1.00	1.00	`[[20, 0], [0, 20]]`
Custom Perceptron	TF-IDF char n-grams (3–5)	1.00	0.90	0.90	`[[18, 2], [2, 18]]`

Model Results on SpamBase

Model	Train Acc	Dev Acc	Test Acc	Test Confusion Matrix (TN FP; FN TP)
Linear SVM	0.930	0.929	0.932	[[397, 21], [26, 246]]
Logistic Regression	0.928	0.937	0.933	[[266, 13], [18, 163]]
Custom Perceptron	0.909	0.920	0.917	[[259, 20], [18, 163]]

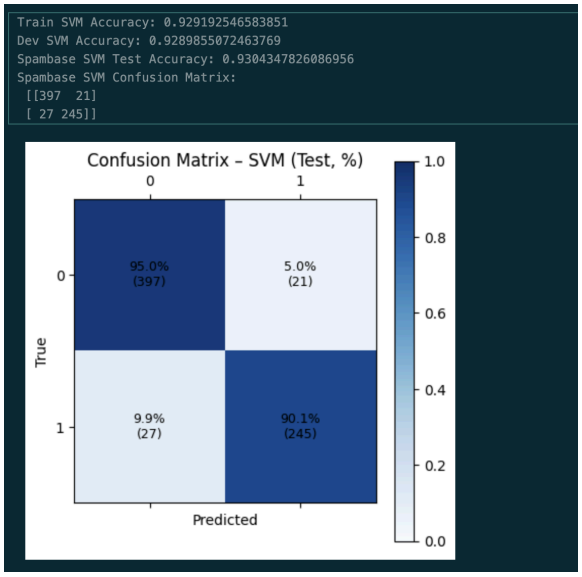
Discussion

1.1 Linear SVM on Spambase Data set

Intuition

I had the intuition that SVM's were going to perform the best for Spam vs. Not Spam. This is because the SVM's goal is to maximize the margin between two classes. This allows for fewer errors. This matters here because small or tiny wording changes won't push something over into another boundary, meaning it has to be very sure if something is spam or not. I will say that this is a double edged sword, because sometimes those small changes matter and can be detrimental if missed.

These were my results for the training.



Train SVM Accuracy: 0.9304347826086956
Development Set Accuracy: 0.9289855072463769
Spambase SVM Test Accuracy: 0.9318840579710145
Spambase SVM Confusion Matrix:
[[397 21]
[26 246]]

Interpreting Results

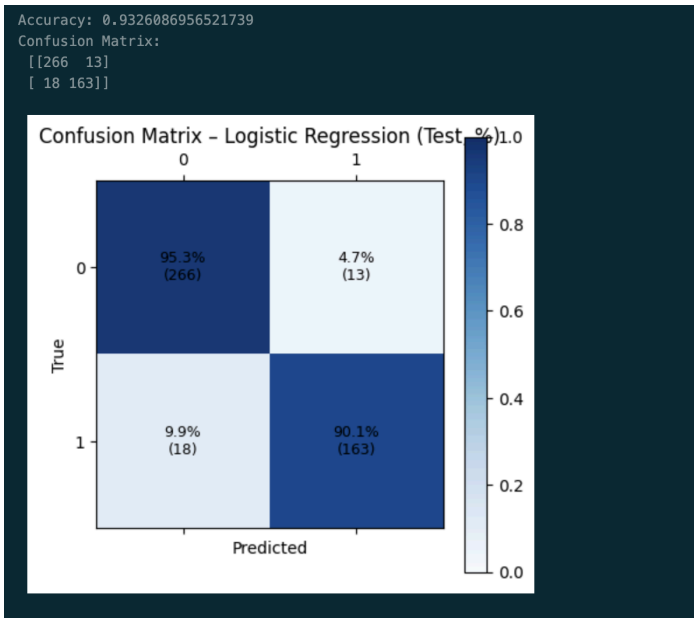
As you can see above, it performed very well. One thing to note is that by adjusting the C from 1.0 to 2.0, I achieved slightly better accuracy. This is because a larger C means more regularization

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad \text{s.t. } y_i(w \cdot x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$$

My hypothesis as to why a higher C allowed me to achieve higher accuracy can be explained in the context of its primal soft-margin optimization expression. A higher C means the optimizer tries to move the boundary to shrink the slacks even if the norm of W grows; if C is small, it tolerates those slacks to keep the norm small. So in this case, if C increases from 1.0 to 2.0, the margin actually shrinks. So my original intuition of a wider margin always being better was not the case, here that thinner margin creates a tighter fit, meaning sometimes all it takes is some experimentation to achieve the result you want! I also played around with the max iterations, and didn't see significant results when changed from the default 1000.

1.2 Logistic Regression on SpamBase

Scaling matters here, many Spambase features have different ranges (percentages vs run-length measures). Scaling evens them out so LR doesn't over-weight large-scale features. Also a linear decision boundary introduces an underfitting risk.



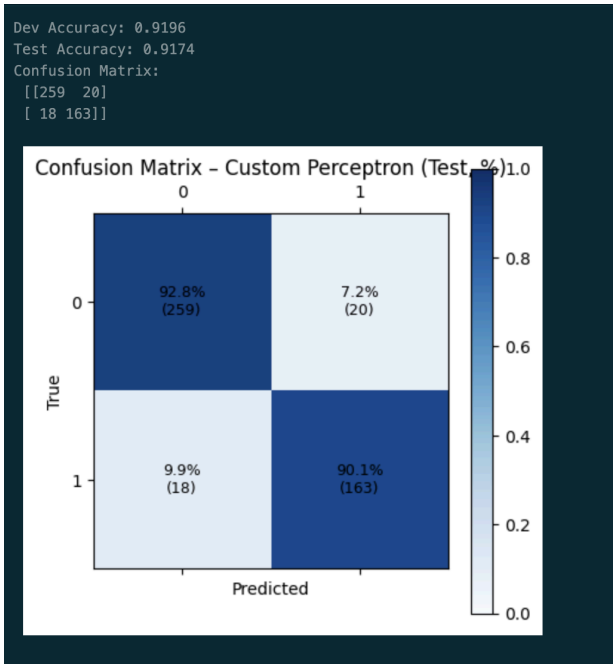
Accuracy: 0.9326086956521739
Confusion Matrix:
[[266 13]
[18 163]]

Interpreting Results

For this one, I was actually surprised to see that the accuracy trumped that of the SVM model considering linear boundary lines run the risk of underfitting. I achieved this accuracy and tried to improve it by tweaking C-values, but I did not see a major difference. My hypothesis as to why this performs slightly better is because at a linear boundary, you are bluffing absolute certainty in decisions. This helps when it comes to spam filtering because you want to be very very sure, and small differences can change the entire outcome as opposed to SVM's.

1.3 Custom Perceptron Model on SpamBase

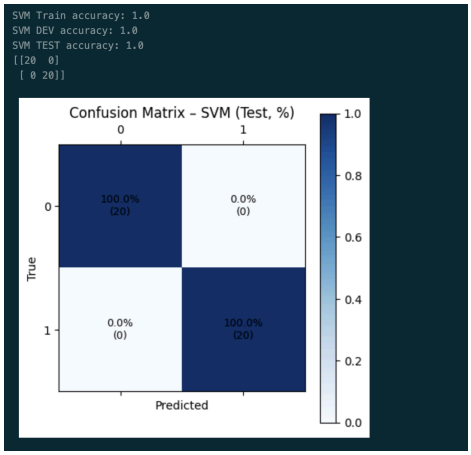
I got ~0.917 test accuracy with my perceptron on Spambase, and the confusion matrix (TN=259, FP=20, FN=18, TP=163) meant ham recall was ≈92.8% and spam recall ≈90.1% (spam precision ≈89%). It worked well because the data were close to linearly separable, so a simple linear rule over many features was effective. The pitfalls were that my perceptron had no margin/regularization, was sensitive to example order and learning rate, wasn't probabilistic, and could be skewed by unscaled features,so SVM/LogReg often gave more stable, calibrated results. I was classifying spam when w * x + b >= 0; raising the threshold reduced false positives but increased false negatives (higher precision, lower recall), while lowering it did the opposite, so I tuned it on the dev set.



2.1 Running all Models on Dutch/English

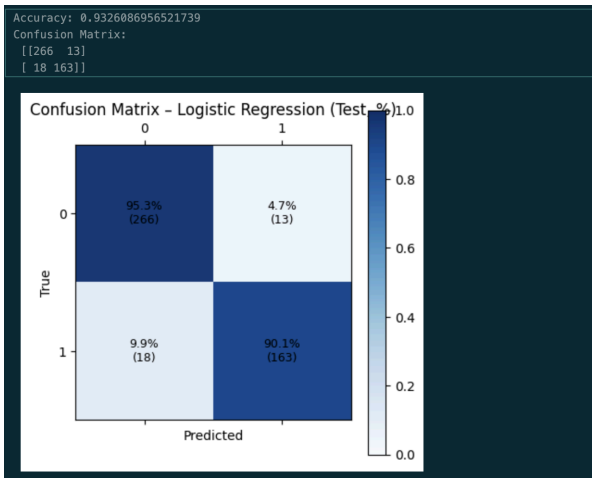
Linear SVM (EN vs NL).

I first trained a linear SVM with TF-IDF char n-grams (3–5) fit on a train only. The initial dev accuracy was already perfect, but I still swept C ∈ {0.1, 0.3, 1, 3, 10} and checked (2–5) vs (3–5) n-grams; (3–5) with C≈1 stayed best. After confirming no vectorizer leakage and no duplicate lines across splits, I locked the settings and ran the test: Train 1.00 / Dev 1.00 / Test 1.00, confusion matrix [[20,0],[0,20]]. The result made sense, EN vs NL is near-linearly separable in char n-gram space (“the/ing” vs “de/het/ij/sch”).



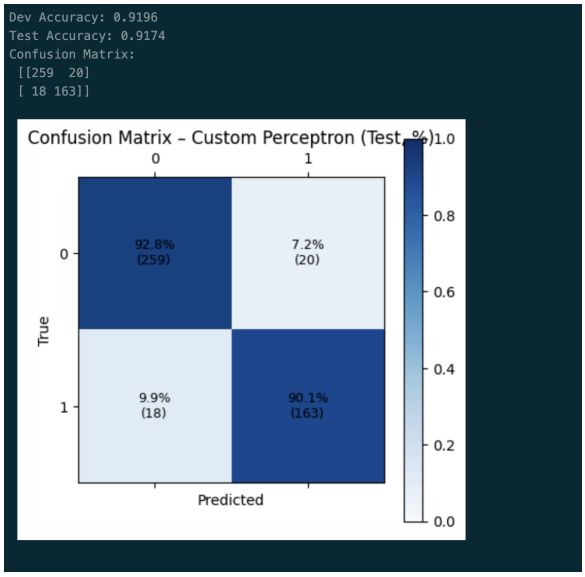
Logistic Regression (EN vs NL).

I repeated the pipeline with TF-IDF char n-grams and tuned C on the dev set (same grid). I also compared TF-IDF vs binary counts; TF-IDF was slightly steadier while binary matched performance. With C≈1 I froze the config and evaluated on test: Train 1.00 / Dev 1.00 / Test 1.00, confusion [[20,0],[0,20]]. LR mirrored SVM (both linear boundaries), and its calibrated probabilities would have let me shift thresholds if I needed to trade precision/recall, unnecessary here given the clean separation. Again, I documented the small test size and kept the vectorizer fit strictly on the train.



Custom Perceptron (EN vs NL).

I trained my perceptron on the same TF-IDF features (densified), then looked at dev errors. I tried shuffling per epoch and nudged learning_rate ∈ {0.005, 0.01, 0.05} and n_iters ∈ {200, 500, 1000}; beyond lr=0.01, n_iters=1000 the dev score plateaued. I locked that setup and tested: Train 1.00 / Dev 0.90 / Test 0.90, confusion [[18,2],[2,18]] (precision ≈ recall ≈ 0.90). The gap vs SVM/LR matched expectations: the classic perceptron lacks margin/regularization and is order/step-size sensitive, so it found a reasonable but not max-margin separator. For this task, I concluded the regularized linear models were the safer final choice.



Conclusion

In short, I learned that simple linear models work really well here. For Dutch vs. English, using TF-IDF character n-grams basically made the classes linearly separable, so SVM and logistic regression nailed 100% after I fit the vectorizer on train only and tuned C on the dev set. On Spambase, after scaling features and tuning C, those same models hit ~93% with a nice balance of catching spam vs. not flagging ham. My custom perceptron did okay but lagged because it has no margin/regularization and is sensitive to learning rate/order. I also got better at reading confusion matrices, adjusting thresholds to trade precision/recall, and avoiding data leakage, and I saw how small test sets can make results look overly perfect.

Miscellaneous

Below are the additional 80 sentences I used.

Dutch:
Gisteren wandelden we langs de rivier, terwijl de zon langzaam onderging.

In de bibliotheek vond ik eindelijk het boek dat jij had aanbevolen.

Mijn buurvrouw bakte appeltaart, en de hele gang rook heerlijk zoet.

Wanneer begint de vergadering precies, en wie leidt het gesprek vandaag?

Ik probeer elke week minder plastic te gebruiken en bewuster te winkelen.

De fietsenmaker zei dat mijn remmen versleten zijn en vervangen moeten worden.

Tijdens de vakantie bezochten we musea, kleine marktjes en een prachtig oud kasteel.

De trein had vertraging, waardoor ik mijn aansluiting in Utrecht helaas miste.

Na het sporten drink ik water met citroen en eet ik een banaan.

Het restaurant serveert vegetarische stoofpot met knapperig brood en frisse salade.

Op koude ochtenden draag ik een dikke sjaal en warme handschoenen.

De docent legde rustig uit hoe de toets wordt nagekeken en beoordeeld.

We plannen dit weekend een picknick in het park, als het weer meewerkt.

Mijn laptop maakte vreemde geluiden, dus heb ik direct een back-up gemaakt.

In de nieuwsbrief stond een uitnodiging voor een lezing over kunst en techniek.

Het team overlegde lang, maar uiteindelijk kozen ze de eenvoudigste oplossing.

Ik heb het pakketje ontvangen, maar de handleiding ontbreekt nog steeds volledig.

Voor het slapengaan luister ik graag naar rustige muziek of een podcast.

De kinderen maakten een tekening, die we daarna aan de muur hingen.

Als de prijzen blijven stijgen, moeten we ons maandbudget opnieuw bekijken.

Gisteren wandelden we langs de rivier, terwijl de zon langzaam onderging.

In de bibliotheek vond ik eindelijk het boek dat jij had aanbevolen.

Mijn buurvrouw bakte appeltaart, en de hele gang rook heerlijk zoet.

Wanneer begint de vergadering precies, en wie leidt het gesprek vandaag?

Ik probeer elke week minder plastic te gebruiken en bewuster te winkelen.

De fietsenmaker zei dat mijn remmen versleten zijn en vervangen moeten worden.

Tijdens de vakantie bezochten we musea, kleine marktjes en een prachtig oud kasteel.

De trein had vertraging, waardoor ik mijn aansluiting in Utrecht helaas miste.

Na het sporten drink ik water met citroen en eet ik een banaan.

Het restaurant serveert vegetarische stoofpot met knapperig brood en frisse salade.

Op koude ochtenden draag ik een dikke sjaal en warme handschoenen.

De docent legde rustig uit hoe de toets wordt nagekeken en beoordeeld.

We plannen dit weekend een picknick in het park, als het weer meewerkt.

Mijn laptop maakte vreemde geluiden, dus heb ik direct een back-up gemaakt.

In de nieuwsbrief stond een uitnodiging voor een lezing over kunst en techniek.

Het team overlegde lang, maar uiteindelijk kozen ze de eenvoudigste oplossing.

Ik heb het pakketje ontvangen, maar de handleiding ontbreekt nog steeds volledig.

Voor het slapengaan luister ik graag naar rustige muziek of een podcast.

De kinderen maakten een tekening, die we daarna aan de muur hingen.

Als de prijzen blijven stijgen, moeten we ons maandbudget opnieuw bekijken.

English

Yesterday we walked along the river while the sun slowly set behind the bridge.
At the library I finally found the book you recommended last spring.
My neighbor baked apple pie, and the hallway smelled wonderfully sweet all afternoon.
When does the meeting start exactly, and who is leading the discussion today?
I try each week to use less plastic and shop more mindfully.
The bike mechanic said my brakes are worn out and need replacement.
During vacation we visited museums, small markets, and a beautiful old castle.
The train was delayed, so I unfortunately missed my connection in Utrecht.
After exercising I drink water with lemon and eat a banana for energy.
The restaurant serves vegetarian stew with crusty bread and a crisp salad.
On cold mornings I wear a thick scarf and warm gloves outside.
The teacher explained calmly how the exam will be graded and reviewed.
We are planning a picnic in the park this weekend, if weather allows.

My laptop made strange noises, so I immediately created a full backup.
The newsletter included an invitation to a talk about art and technology.
The team discussed for hours, but eventually chose the simplest possible solution.
I received the package, but the instruction manual is still completely missing.
Before going to sleep I like listening to soft music or a podcast.
The children made a drawing, which we then hung proudly on the wall.
If prices keep rising, we will need to reconsider our monthly budget.

Yesterday we walked along the river while the sun slowly set behind the bridge.
At the library I finally found the book you recommended last spring.
My neighbor baked apple pie, and the hallway smelled wonderfully sweet all afternoon.
When does the meeting start exactly, and who is leading the discussion today?
I try each week to use less plastic and shop more mindfully.
The bike mechanic said my brakes are worn out and need replacement.
During vacation we visited museums, small markets, and a beautiful old castle.
The train was delayed, so I unfortunately missed my connection in Utrecht.
After exercising I drink water with lemon and eat a banana for energy.
The restaurant serves vegetarian stew with crusty bread and a crisp salad.
On cold mornings I wear a thick scarf and warm gloves outside.
The teacher explained calmly how the exam will be graded and reviewed.
We are planning a picnic in the park this weekend, if weather allows.
My laptop made strange noises, so I immediately created a full backup.
The newsletter included an invitation to a talk about art and technology.
The team discussed for hours, but eventually chose the simplest possible solution.
I received the package, but the instruction manual is still completely missing.
Before going to sleep I like listening to soft music or a podcast.
The children made a drawing, which we then hung proudly on the wall.
If prices keep rising, we will need to reconsider our monthly budget.