First interesting observation:

The a simple naïve bayes classifier trained on English/german data and then run on the other language is surprisingly good:

On translated dataset:

Accuracy of English classifier on English test data: 0.9824144486692015

Accuracy of German classifier on German test data: 0.9771863117870723

Accuracy of English classifier on German test data: 0.966254752851711

Accuracy of German classifier on English test data: 0.7656844106463878

As we can see an English trained classifier performs surprisingly well on german test data -> Why is that so?

On the other hand, why does a german trained classifier on an English test only have 76% accuracy?

These are the results of the original dataset:

Accuracy of English classifier on English test data: 0.9883355764917003

Accuracy of German classifier on German test data: 0.9869896814715119

Accuracy of English classifier on German test data: 0.9793629430237775

Accuracy of German classifier on English test data: 0.7164647824136384

These pictures are made from the multinomial binomial classifier trained on the “text” from the file translatedTextGerman. The English text.

These are most important features from the English trained classifier.A graph with blue lines

Description automatically generated

A graph of blue and white lines

Description automatically generated

Takeaway:

NOW COMPARISN FROM THE MOST 20 important words of the German trained classifier:

A graph of words with blue lines

Description automatically generated with medium confidence

Why is there a difference in languages?

Claim translated can be beanspruchen, anspruch, beanspruchscode Rueckforderung, Aufforderung,

A screenshot of a computer

Description automatically generated

Same stats from the kaggle tutorial:

A graph with blue lines

Description automatically generated

A graph with blue and white lines

Description automatically generated

Running tests on the hindi language:

When running tests on hindi language the English and german classifier had about 93% result, which was about the same for a hindi trained classifier. Therefore, there was no further investigation towards hindi language done, since I assume there must be some extra step to be able to use the different alphabet and words. Also, I do not trust the machine translation, since there are some English words still in there. The assumption can be made, that the classifier only works regarding the englsish words in the hindi text, which are a result of poor translation. Furthermore, backwards translation from hindi text was tried to confirm suspicion, and it was confirmed that translation was poor.

Crazy results of French classifier running on

* Accuracy of French classifier on English text data: 0.5683776022972002
* Accuracy of French classifier on German text data: 0.9637473079684135

Examine why French classifier so much better on English than it is on german??

Short answer.

A graph with blue and white lines

Description automatically generated

A graph with blue lines

Description automatically generated

Then try different method to actually detect language.

Content based filtering is usually used to create automatic filtering rules and to classify emails using [machine learning approaches](https://www.sciencedirect.com/topics/computer-science/machine-learning-approach), such as Naïve [Bayesian classification](https://www.sciencedirect.com/topics/computer-science/bayesian-classification), [Support Vector Machine](https://www.sciencedirect.com/topics/mathematics/support-vector-machine), [K Nearest Neighbor](https://www.sciencedirect.com/topics/nursing-and-health-professions/k-nearest-neighbor), [Neural Networks](https://www.sciencedirect.com/topics/chemical-engineering/neural-network).

<https://www.sciencedirect.com/science/article/pii/S2405844018353404>

use a support vector machine? Or a k-nearest neighbor?

ChaGPT as a spam filter:

What to ask it:

        messages=[

            {"role": "system", "content": "You are a multilingual informal text message spam filter. Please classify this message as spam or ham. If you are unsure make it ham"},

            {"role": "user", "content": "Is this message spam or ham? Please classify it as spam or ham. Respond with either 'spam' or 'ham': " + text}

        ]

Without the “If you are unsure make it ham” it performs significantly worse

Correctly classified ham: 791

Correctly classified spam: 147

Incorrectly classified: 62

Without the unsure make it ham:

Correctly classified ham: 436

Correctly classified spam: 151

Incorrectly classified: 413