Technische Hochschule Ingolstadt

Spam Detection

Spam vs. non-Spam

Dear Professor Georges,

I have a few more questions about task sheet 1.

Should we create an FST in Python ourselves to normalize the text, or may we also use other normalization methods?

Spam or ham?

Email containing non-sense (Spam)

Spam or ham?

https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering

Spam filter with Bayes classifier

$$SF(\underline{w}) = \text{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w})$$

Rings a bell?

https://www.gunfist.com/language-detection-wth-n-grams/

Spam filter with Bayes classifier

Let <u>w</u> be the word sequence.

Question: How to model a spam filter using Bayes classifier?

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

Let $\underline{w}$ be the word sequence.

Question: How to model a spam filter using Bayes classifier?

Answer: A **S**pam **F**ilter, denoted by **SF**($\underline{w}$), taking $\underline{w}$ as input, can be modeled as:

$$SF(\underline{w}) = \ldots$$

Spam filter with Bayes classifier

$$SF(\underline{w}) = \text{argmax}_{S=\{spam,no\text{-}spam\}} \ldots ?$$

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

$$SF(\underline{w}) = \mathrm{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w})$$

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

$$
\begin{aligned}
SF(\underline{w}) &= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\, P(S|\underline{w}) \\
&= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)/P(\underline{w})
\end{aligned}
$$

https://www.gunfist.com/language-detection-wth-n-grams/

Spam filter with Bayes classifier

$$\begin{aligned} SF(\underline{w}) &= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w}) \\ &= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)/P(\underline{w}) \\ &= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S) \end{aligned}$$

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

$$\mathrm{SF}(\underline{w}) = \mathrm{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w})$$
$$= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)/P(\underline{w})$$
$$= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)$$

Question: What needs to be estimated?

https://www.gunfist.com/language-detection-wth-n-grams/

Spam filter with Bayes classifier

$$\begin{aligned}
SF(\underline{w}) &= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(S|\underline{w}) \\
&= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)/P(\underline{w}) \\
&= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)
\end{aligned}$$

$$\left.\begin{aligned}
P(\underline{w}|spam) &= P_{spam}(\underline{w}) \\
P(\underline{w}|no\text{-}spam) &= P_{no\text{-}spam}(\underline{w})
\end{aligned}\right\}$$

Question: How to compute that?

https://www.gunfist.com/language-detection-wth-n-grams/

Spam filter with Bayes classifier

$$SF(\underline{w}) = \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w})$$
$$= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)/P(\underline{w})$$
$$= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)$$

$$\left.\begin{aligned} P(\underline{w}|spam) &= P_{spam}(\underline{w}) \\ P(\underline{w}|no\text{-}spam) &= P_{no\text{-}spam}(\underline{w}) \end{aligned}\right\}$$

How about an n-gram language model?

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

$$
\begin{aligned}
SF(\underline{w}) &= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\, P(S|\underline{w}) \\
&= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)/P(\underline{w}) \\
&= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)
\end{aligned}
$$

$$
\left.
\begin{aligned}
P(\underline{w}|spam) &= P_{spam}(\underline{w}) \\
P(\underline{w}|no\text{-}spam) &= P_{no\text{-}spam}(\underline{w})
\end{aligned}
\right\}
$$

How about an n-gram language model?

Question:
Which quantity is missing?

https://www.gunfist.com/language-detection-wth-n-grams/

Spam filter with Bayes classifier

$$\begin{aligned} SF(\underline{w}) &= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(S|\underline{w}) \\ &= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S)/P(\underline{w}) \\ &= \text{argmax}_{S=\{spam,no\text{-}spam\}}\, P(\underline{w}|S)P(S) \end{aligned}$$

$$\left.\begin{aligned} P(\underline{w}|spam) &= P_{spam}(\underline{w}) \\ P(\underline{w}|no\text{-}spam) &= P_{no\text{-}spam}(\underline{w}) \end{aligned}\right\}$$

How about an n-gram language model?

$P(S)$ => Question: How to compute that?

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

$$
\begin{aligned}
\mathrm{SF}(\underline{w}) &= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(S|\underline{w}) \\
&= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)/P(\underline{w}) \\
&= \operatorname{argmax}_{S=\{spam,no\text{-}spam\}} P(\underline{w}|S)P(S)
\end{aligned}
$$

$$
\left.
\begin{aligned}
P(\underline{w}|spam) &= P_{spam}(\underline{w}) \\
P(\underline{w}|no\text{-}spam) &= P_{no\text{-}spam}(\underline{w})
\end{aligned}
\right\}
$$

How about an n-gram language model?

$$P(spam) + P(no\text{-}spam) = 1.0$$

Question: How to compute that?

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Spam filter with Bayes classifier

Predictor Prior Probability

$$\mathrm{SF}(\underline{w}) = \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\ P(S|\underline{w})$$
$$= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\ P(\underline{w}|S)P(S)/P(\underline{w})$$
$$= \mathrm{argmax}_{S=\{spam,no\text{-}spam\}}\ P(\underline{w}|S)P(S)$$

Likelihood

$$P(\underline{w}|spam) = P_{spam}(\underline{w})$$
$$P(\underline{w}|no\text{-}spam) = P_{no\text{-}spam}(\underline{w})$$

How about an n-gram language model?

Class Prior

$$P(spam) + P(no\text{-}spam) = 1.0$$

can be estimated

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

Language Identification

Language Identification (LID)

„Wenn wir das nehmen und darauf aufbauen, dann können wir einen Schritt weiter gehen: Wenn das Meer nicht glücklich ist, ist keiner glücklich. “

English or German?

„And if we just take that and we build from there, then we can go to the next step, which is that if the ocean ain't happy, ain't nobody happy. “

English or German?

https://object.pouta.csc.fi/OPUS-TED2013/v1.1/moses/de-en.txt.zip

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \operatorname{argmax}_{L=\{de,en\}} P(L|\underline{w})$$

https://www.gunfist.com/language-detection-wth-n-grams/
https://textcatdemo.toolforge.org/

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \mathrm{argmax}_{L=\{de,en\}}\, P(L|\underline{w})$$

$$= \mathrm{argmax}_{L=\{de,en\}}\, P(\underline{w}|L)P(L)/P(\underline{w})$$

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

https://en.wikipedia.org/wiki/Bayes%27_theorem
https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \mathrm{argmax}_{L=\{de,en\}}\, P(L|\underline{w})$$
$$= \mathrm{argmax}_{L=\{de,en\}}\, P(\underline{w}|L)P(L)/P\cancel{(\underline{w})}$$
$$= \mathrm{argmax}_{L=\{de,en\}}\, P(\underline{w}|L)P(L)$$

Independent of $P(\underline{w})$

https://en.wikipedia.org/wiki/Bayes%27_theorem
https://www.gunfist.com/language-detection-wth-n-grams/

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \mathrm{argmax}_{L=\{de,en\}}\, P(L|\underline{w})$$

$$= \mathrm{argmax}_{L=\{de,en\}}\, P(\underline{w}|L)P(L)/P(\underline{w})$$

$$= \mathrm{argmax}_{L=\{de,en\}}\, P(\underline{w}|L)P(L)$$

$$P(\underline{w}|L) = ? \qquad P(L) = ?$$

https://en.wikipedia.org/wiki/Bayes%27_theorem
https://www.gunfist.com/language-detection-wth-n-grams/

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \mathrm{argmax}_{L=\{de,en\}} P(L|\underline{w})$$
$$= \mathrm{argmax}_{L=\{de,en\}} P(\underline{w}|L)P(L)/P(\underline{w})$$
$$= \mathrm{argmax}_{L=\{de,en\}} P(\underline{w}|L)P(L)$$

$$\left.\begin{aligned} P(\underline{w}|de) &= P_{German}(\underline{w}) \\ P(\underline{w}|en) &= P_{English}(\underline{w}) \end{aligned}\right\}$$

How about an n-gram language model?

https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

LID with Bayes Classifier

$$\mathrm{LID}(\underline{w}) = \operatorname{argmax}_{L=\{de,en\}} P(L|\underline{w})$$
$$= \operatorname{argmax}_{L=\{de,en\}} P(\underline{w}|L)P(L)/P(\underline{w})$$
$$= \operatorname{argmax}_{L=\{de,en\}} P(\underline{w}|L)P(L)$$

$$\left.\begin{aligned} P(\underline{w}|de) &= P_{German}(\underline{w}) \\ P(\underline{w}|en) &= P_{English}(\underline{w}) \end{aligned}\right\}$$

How about an n-gram language model?

$$P(de) + P(en) = 1.0$$

Usually uniformly distributed, depends on application

https://www.gunfist.com/language-detection-wth-n-grams/

LID with Bayes Classifier

$$P(\underline{w}|de) = P_{German}(\underline{w}) = P_{German}(F(\underline{w}))$$

$$P(\underline{w}|en) = P_{English}(\underline{w}) = P_{English}(F(\underline{w}))$$

$F(\underline{w})$:= Text preprocessing

- Tokenization
- Character Set
- Punctuation
- ....

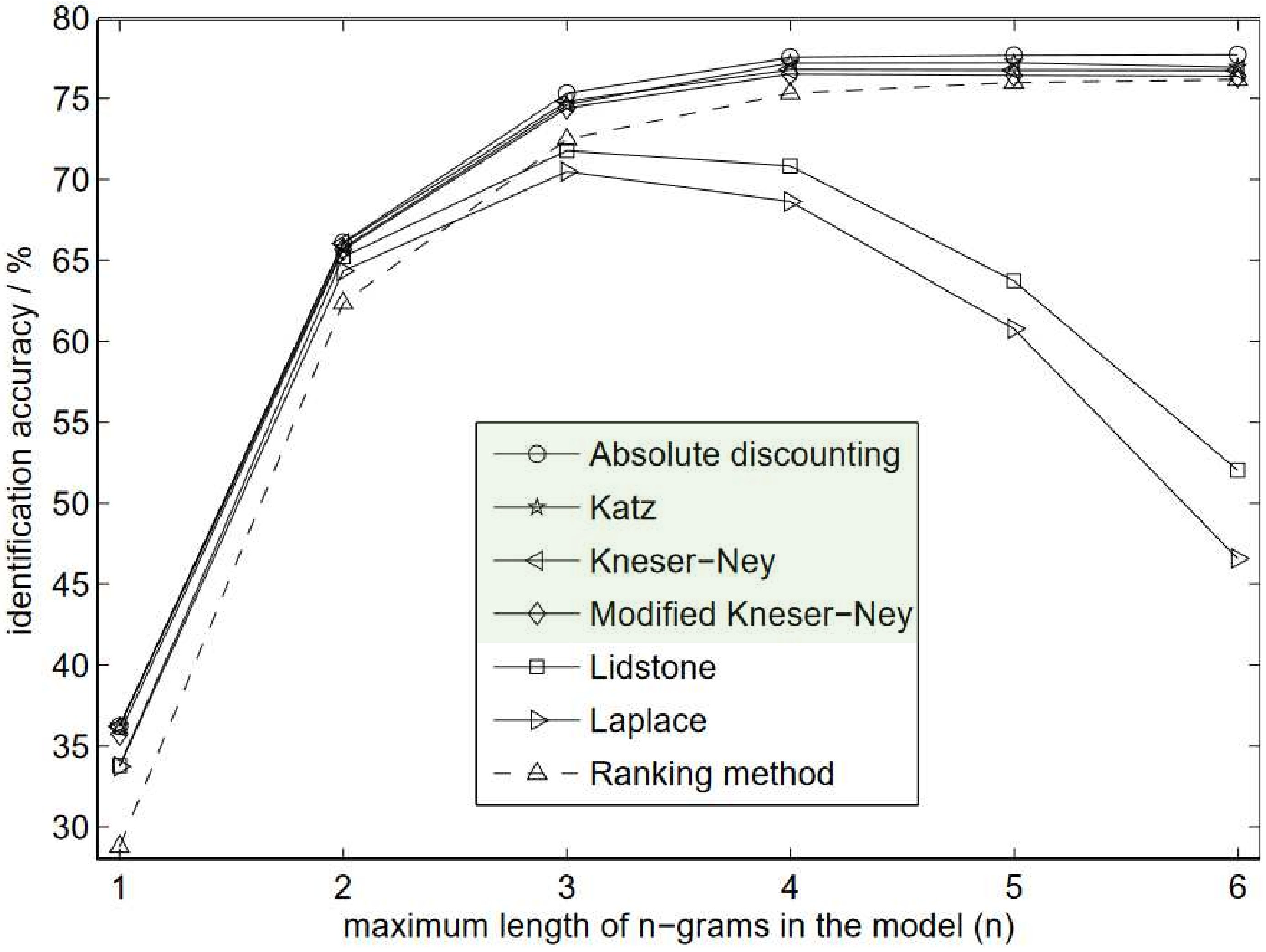
https://www.gunfist.com/language-detection-wth-n-grams/

Technische Hochschule Ingolstadt

LID with n-gram Language Models

Different character-based n-gram models

Language	Character set size (characters)
English	57
French	60
German	70
Spanish	61
Portuguese	64
Finnish	57
Greek	68
Russian	67
Tagalog	57
Ashéninca	60
Cashinahua	59
Japanese	506
Chinese	**539**
Kikongo	**44**
⋮	⋮
Average	66.0



http://www.lrec-conf.org/proceedings/lrec2010/pdf/279_Paper.pdf

LID too booring?

Multi-lingual …
- **Automatic Speech Recognition**
- **Spoken Language Understanding**

Confusion Matrix (accuracy)

	Play Music	Increase	Decrease	Stop	Reference
Play Music	0.8	0	0	0.2	
Increase	0	0.5	0.4	0.1	
Decrease	0	0.4	0.5	0.1	
Stop	0.2	0.1	0.1	0.6	
Prediction					

Technische Hochschule Ingolstadt

Confusion Matrix (accuracy)

	Play Music	Increase	Decrease	Stop	Reference
Play Music	**0.8**	0	0	0.2	
Increase	0	0.5	0.4	0.1	
Decrease	0	0.4	0.5	0.1	
Stop	0.2	0.1	0.1	0.6	
Prediction					

„Play Music“ was correctly recognized in 80% of all cases

Technische Hochschule Ingolstadt

Confusion Matrix (accuracy)

	Play Music	**Increase**	Decrease	Stop	Reference
Play Music	0.8	0	0	0.2	
Increase	0	**0.5**	0.4	0.1	
Decrease	0	0.4	0.5	0.1	
Stop	0.2	0.1	0.1	0.6	
Prediction					

„Increase“was correctly recognized in 50% of all cases

Technische Hochschule Ingolstadt

Confusion Matrix (accuracy)

	Play Music	**Increase**	**Decrease**	Stop	Reference
Play Music	0.8	0	0	0.2	
Increase	0	**0.5**	**0.4**	0.1	
Decrease	0	0.4	0.5	0.1	
Stop	0.2	0.1	0.1	0.6	
Prediction					

„Decrease“ was confused with „Increase“ in 40% of all cases.

Confusion Matrix (accuracy)

	Play Music	Increase	Decrease	Stop	Reference
Play Music	**0.8**	0	0	0.2	
Increase	0	**0.5**	0.4	0.1	
Decrease	0	0.4	**0.5**	0.1	
Stop	0.2	0.1	0.1	**0.6**	
Prediction					

Ideally:
each diagonal
element in matrix
equal to 1.0

Technische Hochschule Ingolstadt

Confusion Matrix (accuracy)

	Play Music	Increase	Decrease	Stop	Reference
Play Music	**0.8**	0	0	0.2	
Increase	0	**0.5**	0.4	0.1	
Decrease	0	0.4	**0.5**	0.1	
Stop	0.2	0.1	0.1	**0.6**	
Prediction					

Wanted:
Metric telling
„how diagonal is
the matrix“?

Technische Hochschule Ingolstadt

Evaluation

“A systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards.”

Technische Hochschule Ingolstadt

True/False Positive/Negative

True Positive: A (spam) E-Mail is correctly recognized as spam

True Negative: A (non-spam) E-Mail is correctly recognized as non-spam

False Positive: A (non-spam) e-mail is incorrectly recognized as spam

False Negative: A (spam) e-mail is incorrectly recognized as non-spam

Technische Hochschule Ingolstadt

Accuracy

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

condition positive (P) := TP + FN

condition negative (N) := TN + FP

True Positive: A (spam) E-Mail is correctly recognized as spam

True Negative: A (non-spam) E-Mail is correctly recognized as non-spam

False Positive: A (non-spam) e-mail is incorrectly recognized as spam

False Negative: A (spam) e-mail is incorrectly recognized as non-spam

Technische Hochschule Ingolstadt

Precision

“Rate of relevant instances among the retrieved instances.”

Positive Predictive Value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

True Positive: A (spam) E-Mail is correctly recognized as spam

True Negative: A (non-spam) E-Mail is correctly recognized as non-spam

False Positive: A (non-spam) e-mail is incorrectly recognized as spam

False Negative: A (spam) e-mail is incorrectly recognized as non-spam

Technische Hochschule Ingolstadt

Recall

"Rate of the total amount of relevant instances that were actually retrieved."

Sensitivity, Hit Rate, True Positive Rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

condition positive (P) := TP + FN

True Positive:	A (spam) E-Mail is correctly recognized as spam
True Negative:	A (non-spam) E-Mail is correctly recognized as non-spam
False Positive:	A (non-spam) e-mail is incorrectly recognized as spam
False Negative:	A (spam) e-mail is incorrectly recognized as non-spam

Technische Hochschule Ingolstadt

Fall-Out

False Positive Rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

condition negative (N) := TN + FP

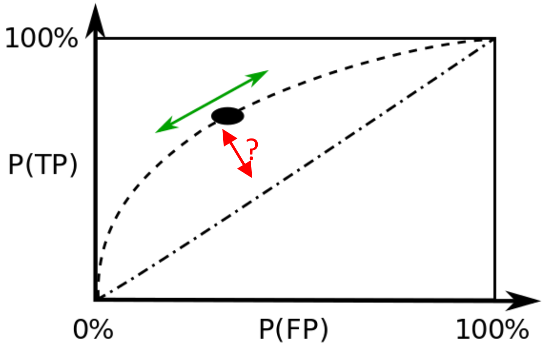
True Positive: A (spam) E-Mail is correctly recognized as spam
True Negative: A (non-spam) E-Mail is correctly recognized as non-spam
False Positive: A (non-spam) e-mail is incorrectly recognized as spam
False Negative: A (spam) e-mail is incorrectly recognized as non-spam

Technische Hochschule Ingolstadt

F_β - Score

F1 Score for β = 1 (or Sørensen–Dice Coefficient)

$$F_\beta = (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$



$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

P : Precision

R : Recall

F_β -Score

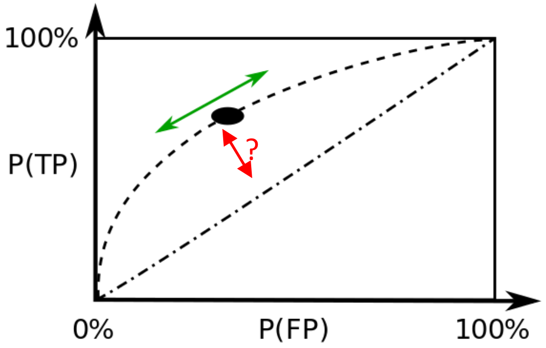
Use it only if your test set contains an equal amount of samples per class!

If your test set is imbalanced w.r.t. class sample size, the score is misleading!

Matthews Correlation Coefficient

Robustness despite imbalanced data set (w.r.t. #samplesPerClass)

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$



True Positive:	A (spam) E-Mail is correctly recognized as spam
True Negative:	A (non-spam) E-Mail is correctly recognized as non-spam
False Positive:	A (non-spam) e-mail is incorrectly recognized as spam
False Negative:	A (spam) e-mail is incorrectly recognized as non-spam

https://en.wikipedia.org/wiki/Matthews_correlation_coefficient
https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0041882&type=printable