financial news headline **sentiment analysis**

Tim Körppen,
Hannes Weichelt

Agenda

- **Motivation und Einführung**
- Preprocessing
- Model Selection
- Evaluation
- Bewertung und Potenziale
- Nächste Schritte

Motivation

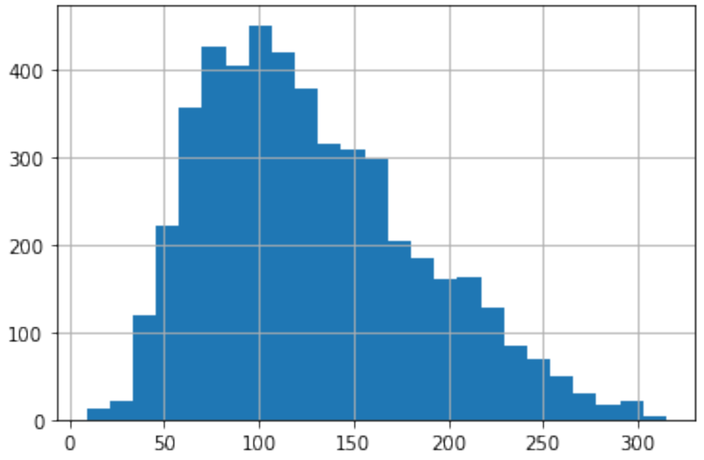
- NLP und Word-Tokenizing steigende Relevanz (insb. Wegen Stimmerkennung und KI-Chatbots)
- Automatisch Sentiment/Gefühle von großen Menschmengen/Datensätzen erkennen kann entscheidende Vorteile bieten
- Finanzsektor bietet viele Daten und Möglichkeiten der Weiterentwicklung

Exploratory data analysis: Einführung in den Datensatz

Absolute und relative Häufigkeiten der Klassen

	Negativ	Neutral	Positiv	Gesamt
n	604	2879	1363	4846
h	12,5 %	59,4 %	28,1 %	

Häufigkeitsverteilung über Stringlänge



Beispiele der Klassen

Text	Sentiment
Sales in Finland decreased by 10.5 % in January , while sales outside Finland dropped by 17 % .	Negativ
According to Gran , the company has no plans to move all production to Russia , although that is where the company is growing.	Neutral
In the third quarter of 2010 , net sales increased by 5.2 % to EUR 205.5 mn , and operating profit by 34.9 % to EUR 23.5 mn .	Positiv

Quelle: Datensatz; Vaughan, D. 2020

Agenda

- Motivation und Einführung
- **Preprocessing**
- Model Selection
- Evaluation
- Bewertung und Potenziale
- Nächste Schritte

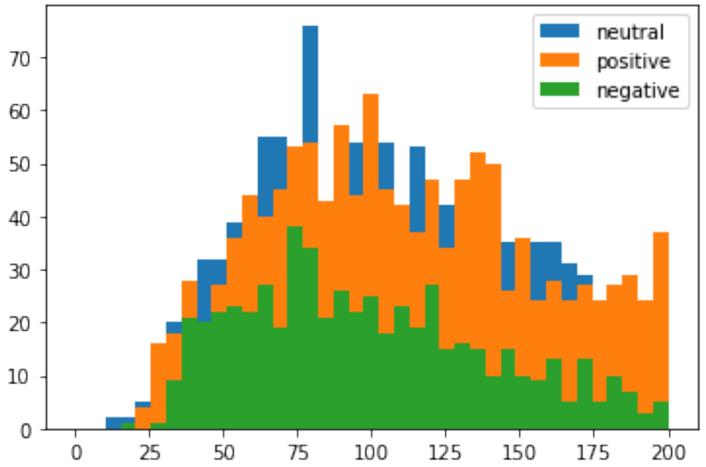
Preprocessing Prozess: Data Cleaning

Originaler Text	Störende Elemente	Zeichensetzung	Tokenizing	Stopwords	Lemmatizing	Final	Anz. Chars
„Technopolis“ plans to develop in 2 mio. stages an area…	„technopolis“ plans to develop in stages an area...	technopolis plans to develop in stages an area...	[technopolis, plans, to, develop, in, stages, ...	[technopolis, plans, develop, stages, area, le...	[technopolis, plan, develop, stage, area, …	technopolis plan develop stage area ...	178

Preprocessing Prozess: Verworfene Preprocessing-Ideen

- Klassenbias durch Löschen aufgeben; Verworfen, weil zu viele Samples verloren gingen
- Ersetzen von Störenden Elementen (statt löschen); Verworfen, weil künstlicher Bias erzeugt wurde
- (Feature Creation) Anz. Chars verwenden; Verworfen, weil nicht aussagekräftig

Häufigkeitsverteilung über Anzahl Chars in Klassen



Preprocessing Prozess: Vectorizing

1

N-Gram

- Reihenfolge der Wörter berücksichtigen
- K-Tupel für aufeinanderfolgender Wörter s.d. k <= N (N-Gram länge)

2

TF-IDF Vectorizer

- Anzahl des Vorkommens mit inverser Dokumentfrequenz Gewichten (häufige Wörter abgewichten)
- Vektor für späteres Modell

$$IDF(wort_i) = \log \frac{\#\text{Dokumente}}{\#\text{Dokumente , in denen Wort}_i \text{ vorkommt}}$$

$$TFIDF(\mathrm{x}) = \frac{1}{|\mathrm{x}|} \begin{pmatrix} TF(Wort_1) \cdot IDF(Wort_1) \\ \vdots \\ TF(Wort_n) \cdot IDF(Wort_n) \end{pmatrix}$$

Agenda

- Motivation und Einführung
- Preprocessing
- **Model Selection**
- Evaluation
- Bewertung und Potenziale
- Nächste Schritte

Preelimiary Model Consideration: Intuition

Modell	Descision Trees	Linear Classification	Neural Networks	Bayesian Models
Intuition	• Mächtige (Ensemble) Methode • Gut nachvollziehbar (bspw. Feature Importances) • Hochdimensionierte TF-IDF Datensätze gut lösbar • Nicht-Lineare Zusammenhänge erkennbar	• Einfache, lineare Methode (komplementär zu Decision Trees) • Schnell in Primaler und Dualer (kernelized) Sicht anwendbar • Potenziell Vorteile durch Kernel, weil deutlich mehr Features als Samples	• Hochdimensionierter Datensatz benötigt großes Modell • Viel Rechenleistung • Wäre für V2 des Modells (ggf. auf verteilten Systemen) geeignet	• Könnte statt dem Linearen Modell verwendet werden (SVC) • Kein Vor-/Nachteil erwartet
Entscheidung	Angenommen; RandomForest Classifier	Angenommen; SVM Classifier	Abgewiesen	Abgewiesen

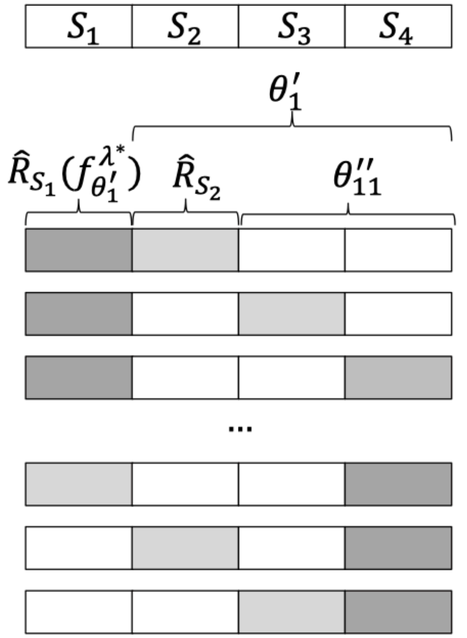
Model Selection; Hyperparametertuning

- Nested Cross Validation und Grid Search
- Evaluierung der Vectorizing-Strategie und Modelparametern
- Verwendung der AUC (Area under the ROC Curve) als Messinstrument bei der GridSearch
  - Klassenbias könnte Precision künstlich in die Höhe treiben
  - AUC ist weniger von Klassenbias beeinflusst (Verhältnis von True-Positive Rate und False-Positive Rate bei verändertem Klassifikations-Threshold)

$$r_{TP} = \frac{n_{TP}}{n_{TP}+n_{FN}}$$

$$r_{FP} = \frac{n_{FP}}{n_{FP}+n_{TN}}$$

Nested Cross Validation; Prinzip



Model Selection; Support Vector Classifier

$$\boldsymbol{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\max\left(0, 1 - y_i \mathbf{x}_i{}^{\mathrm{T}} \boldsymbol{\theta}\right) + \frac{\lambda}{n} \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\theta} \right]$$

Rank	Fit-Time	AUC	C	Dual	N-Gram
1	7.325071	0.855051	1.5	True	(1,2)
2	5.867231	0.854483	1	True	(1,2)
3	8.328862	0.853669	2	True	(1,2)
Grid			0.5 / 1 / 1.5 / 2	True / False	(1,2) (1,3) (2,3)

Model Selection; Random Forest Classifier

C4.5/C5.0 Algorithmus für kontinuierliche Attribute mit **Gini** als Split-Kriterium

Rank	Fit-Time	AUC	n-Estimators	Max-Depth	N-Gram
1	3.612475	0.829797	100	150	(1,2)
2	3.899151	0.829081	150	150	(1,2)
3	2.447473	0.827331	100	100	(1,2)
Grid			50 / 100 / 150	100 / 150 / 200	(1,2) (1,3) (2,3)

Agenda

- Motivation und Einführung
- Preprocessing
- Model Selection
- **Evaluation**
- Bewertung und Potenziale
- Nächste Schritte

Model Evaluation: Strategie

- GridSearch mit AUC (statt Accuracy) liefert das beste Modell (trainiert auf Gesamtdatensatz)
- Holdout-Set erstellen
- Confusion Matrix zur Prüfung der Vorhersagen und Identifikation des Bias (Ausschluss der Accuracy-Score) auf Holdout-Set
- ROC Kurven, um SVC und RFC miteinander zu vergleichen (auf Holdout-Set)
    - Multi-Class ROC Kurven (Receiver Operating Characteristic): Averaging, um mit Multi-Class Setting umzugehen

$$Pr_{micro} = \frac{TP_1 + TP_2 + \ldots + TP_k}{(TP_1 + TP_2 + \ldots + TP_k) + (FP_1 + FP_2 + \ldots + FP_k)}$$

Wenn Klassenbias groß

$$Pr_{macro} = \frac{Pr_1 + Pr_2 + \ldots + Pr_k}{k} = Pr_1 \frac{1}{k} + Pr_2 \frac{1}{k} + \ldots + Pr_k \frac{1}{k}$$

Unterrepräsentierte Klasse stärker gewichten

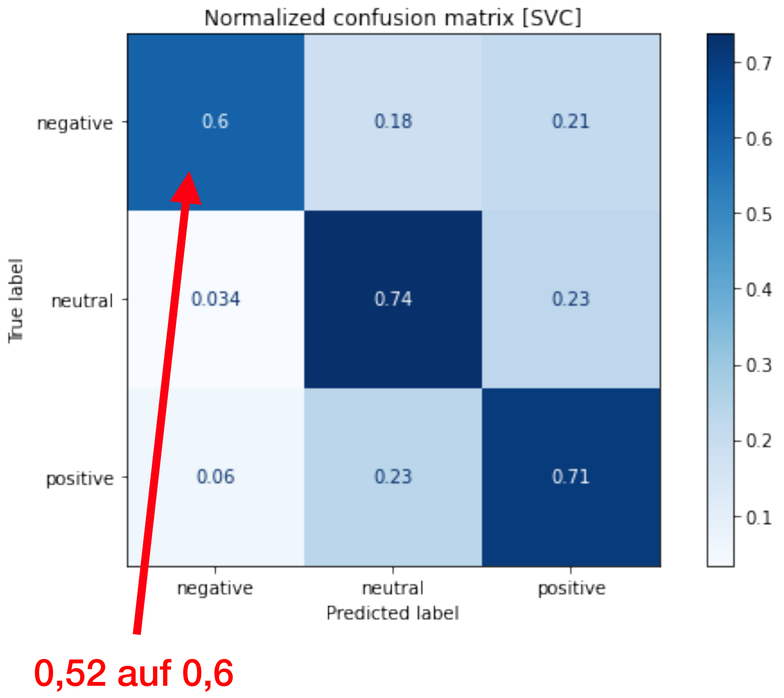
Quelle: Vaughan, D. 2020

Model Evaluation: Ergebnisse

Top 10 Features nach Wichtigkeit für Klassifizierung (RFC)

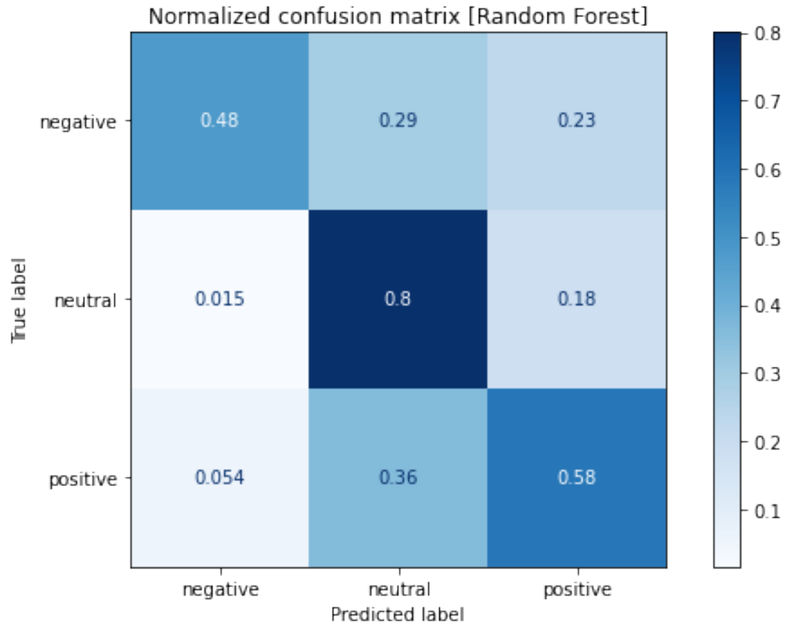
Rank	Feature
1	profit
2	decreased
3	rose
4	year
5	operating
6	fell
7	operating profit
8	increase
9	sale
10	said

Confusion Matrix SVC



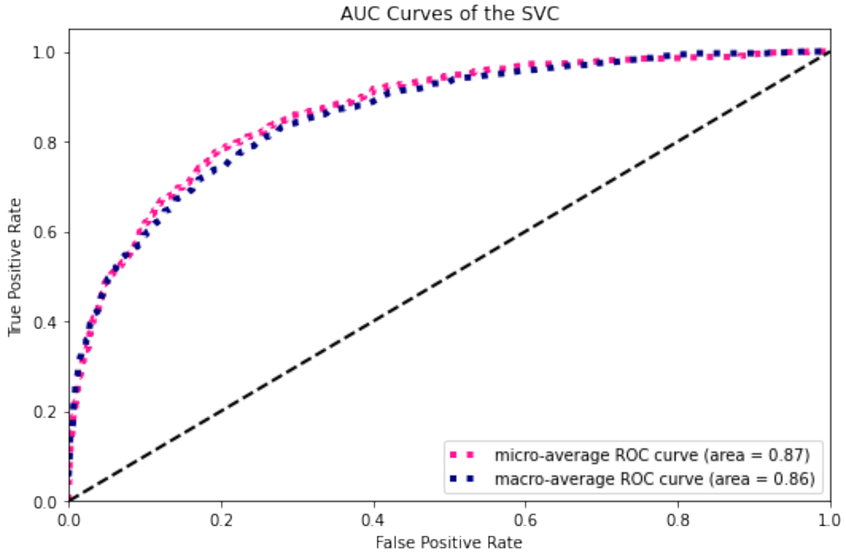
0,52 auf 0,6

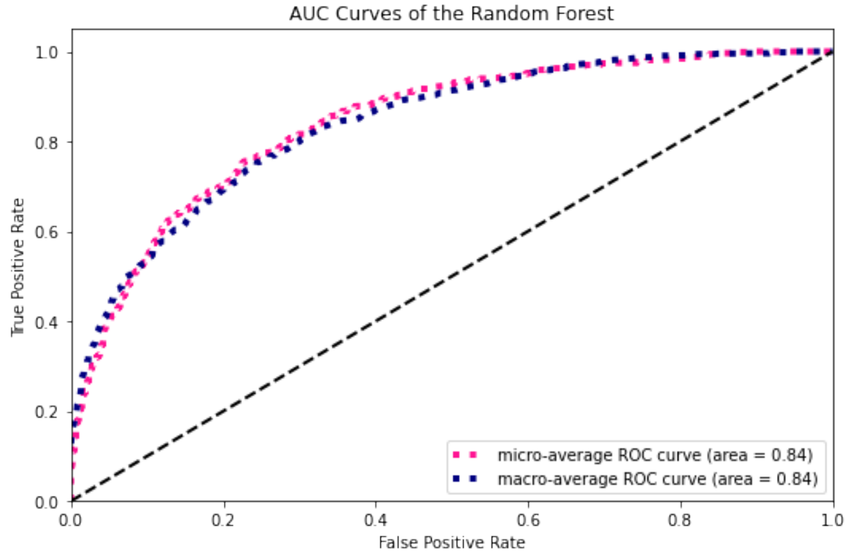
Confusion Matrix RFC

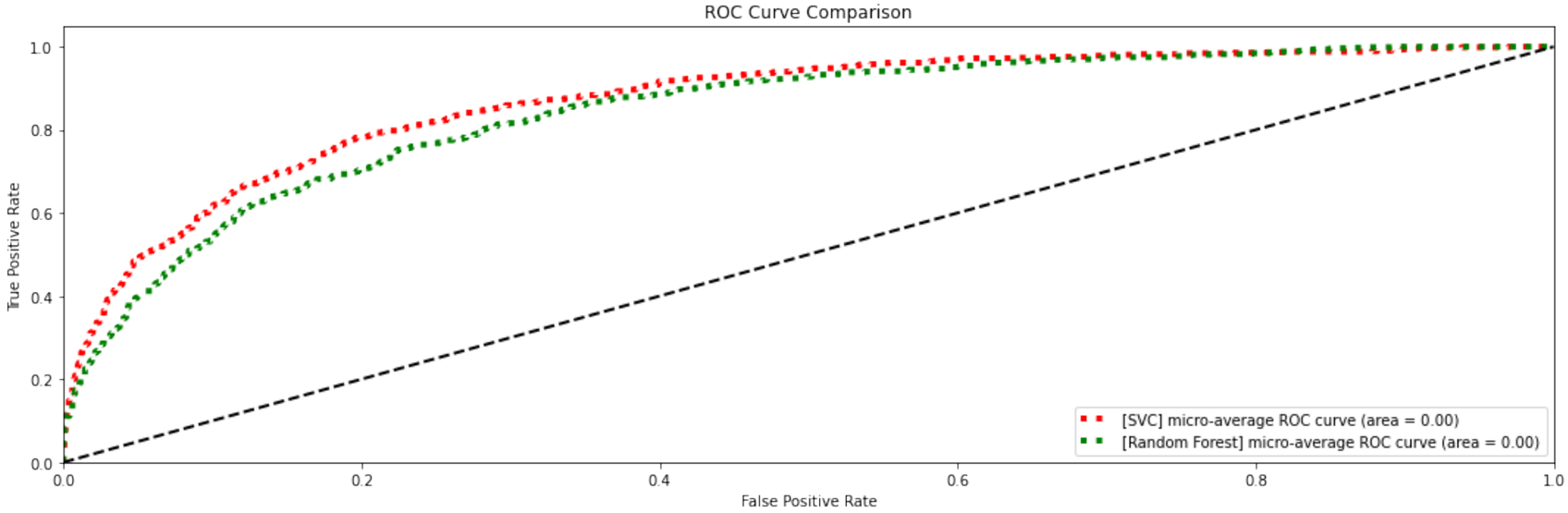


Model Evaluation: Ergebnisse

- Jeweils wie erwartet Micro-Average größer als Macro-Average
- SVC deutlich bauchiger als ROC-Kurve des RFC
- Folglich AUC bei SVC größer als bei RFC







Model Evaluation: Real World Testing

- Test des Modells auf echten, aktuellen Finanzheadlines
- Manuelle Erstellung einer Decision-Pipeline
  - `step_0 = remove_numbers(string.lower())`
  - `step_1 = remove_punctuation(step_0)`
  - `step_2 = nltk.word_tokenize(step_1)`
  - `step_3 = remove_stopwords(step_2)`
  - `step_4 = lemmatizing(step_2)`

Agenda

- Motivation und Einführung
- Preprocessing
- Model Selection
- Evaluation
- **Bewertung und Potenziale**
- Nächste Schritte

Bewertung und Potenziale

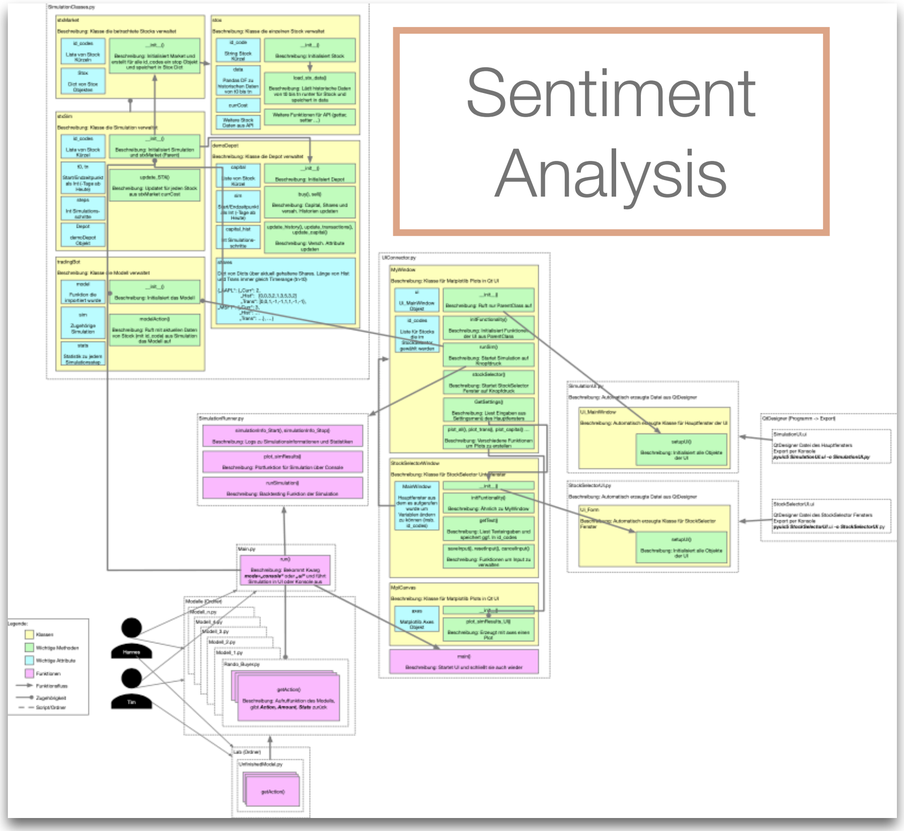
- Klassenbias entfernen:
  - SMOTE (Synthetic Minority Oversampling Technique)
  - Händisch auffüllen/Webscrapen
- Strategie um numerische Werte zu interpretieren
- Verbindung mehrerer Modelle und Ensemble Methoden (v.a. für SVC) anwenden:
  - Boosting, um negative Klassen besser zu klassifizieren
- Transformermodelle (NN) anwenden

Bsp.: Relevanz numerischer Werte

Text	Sentiment
Compared with the FTSE 100 index , which rose 36.7 points (or 0.6 %) on the day , this was a relative price change of **-0.2 % .**	Negativ

Nächste Schritte

- Modulares Projekt zur technischen Analyse von Wirtschaftskonjunktur und Aktienverläufen
- Sentiment Analysis (historisch) mit Kursentwicklung verbinden
- Korrelation von Sentiment aus anderen Bereichen (ggf. „Normalen“ Nachrichten) mit Finanznachrichten und Auswirkungen untersuchen
- Zeitreihenanalysen/-vorhersagen und mit Sentimentanalysis verbinden



Key-Takeaways

1. Multi-Layer Preprocessing und N-Gram als TF-IDF Vector

2. RFC und SVC evaluiert und AUC auf über 0,85 getunt 

3. Multi-Klassen ROC Kurven und Confusion-Matrices um Verfälschung durch Bias auszuschließen

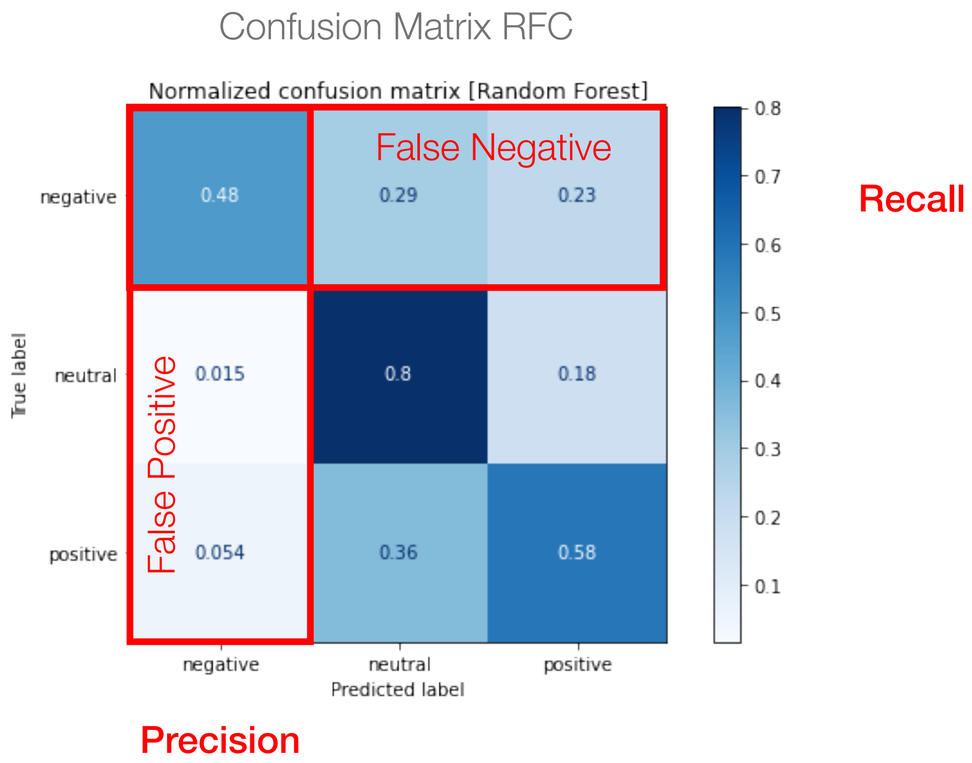
Quellen

- VanderPlas, Jake: Python Data Science Handbook (2016)
- Sinha, A.: (Datensatz für Projekt) Sentiment Analysis for Financial News (2020), https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news (Aufgerufen: 10.10.2020)
- Gunjit, B.: A guide to Text Classification(NLP) using SVM and Naive Bayes with Python (2018), Web: https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34 (Aufgerufen: 9.9.2020)
- Precision-Recall, Web: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html (Aufgerufen: 9.9.2020)
- Goonewardana, H.: Evaluating Multi-Class Classifiers (2019), https://medium.com/apprentice-journal/evaluating-multi-class-classifiers-12b2946e755b (Aufgerufen: 9.9.2020)
- Hilfestellungen zur Multi-Class ROC Kurve mit Scikit-Learn, Web: https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html (Aufgerufen: 01.10.2020)
- Vaughan, D.: Multiclass Averaging (2020), Web: https://cran.r-project.org/web/packages/yardstick/vignettes/multiclass.html (Aufgerufen: 01.10.2020)
- How to increase the speed for SVM classifier using Sk-learn, https://stackoverflow.com/questions/34939683/how-to-increase-the-speed-for-svm-classifier-using-sk-learn (Aufgerufen: 9.9.2020)
- SVC classifier taking too much time for training, https://stackoverflow.com/questions/53940258/svc-classifier-taking-too-much-time-for-training (Aufgerufen: 9.9.2020)

Model Evaluation: Bsp. Precision Recall für Klasse Negativ

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

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



Recall

Precision