Intelligenter Text Klassifikator

Schutz der Verbraucher mit mobilen Endgeräten via maschinellem Lernen (ML)

Xiao Wang

09.06.2022

Agenda

- Use Case Betrachtung - SMS Spam Classifier
- Beschreibung des SMS-Datensatzes
- Datenaufbereitung
- Bewertungsmetrik
- Evaluation der Modellperformance
- Benchmarking von ML Algorithmen
- Schlussfolgerung

Use Case Betrachtung - Spam Classifier

Motivation:

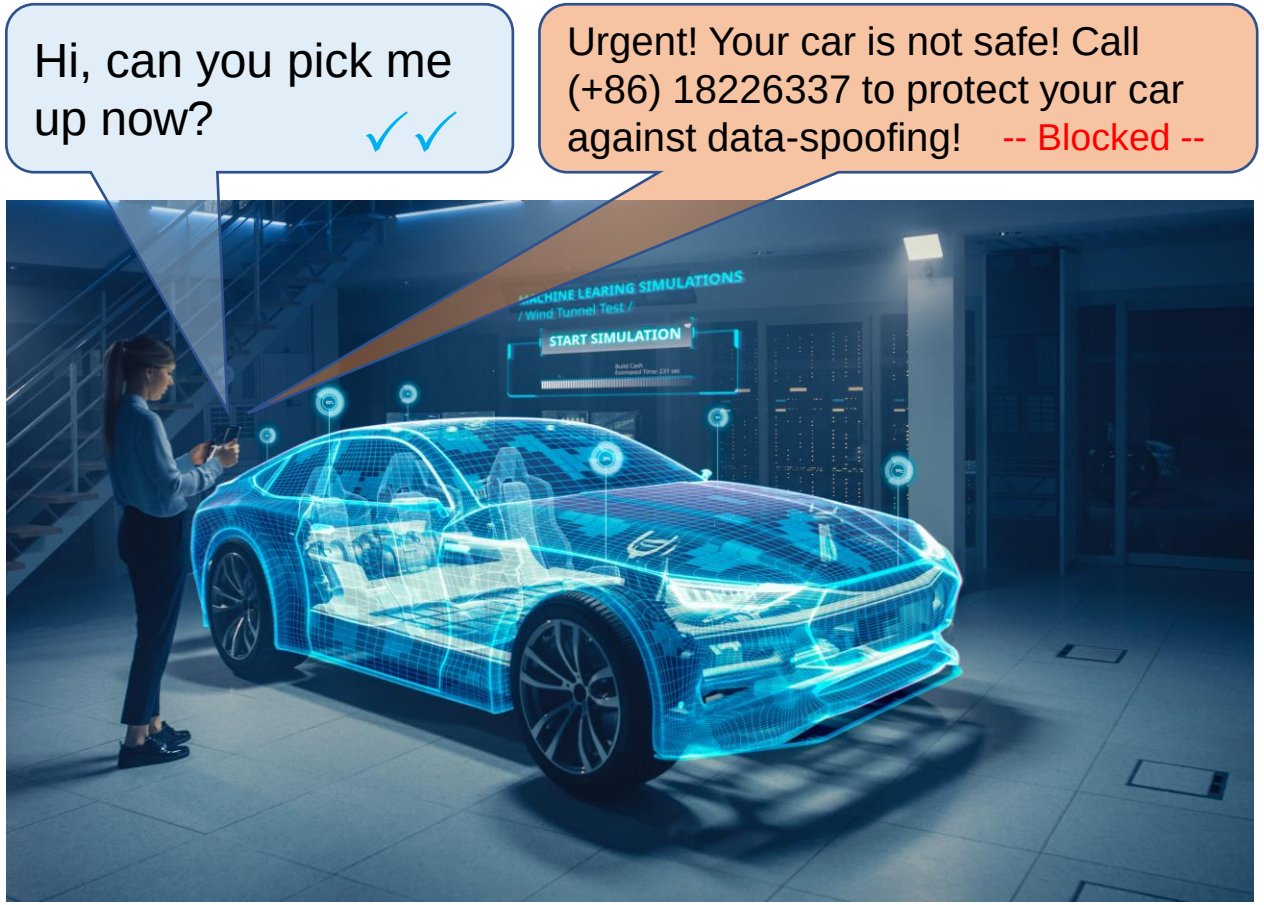
Klassifikation der SMS Spams von legitimen Nachrichten (Ham).

Ziel:

Schutz der Verbraucher mit mobilen Endgeräten vor SMS Spams.

Methodik:

Anwendung von ML Modellen zur binären Klassifikation von Spam und Ham Daten.

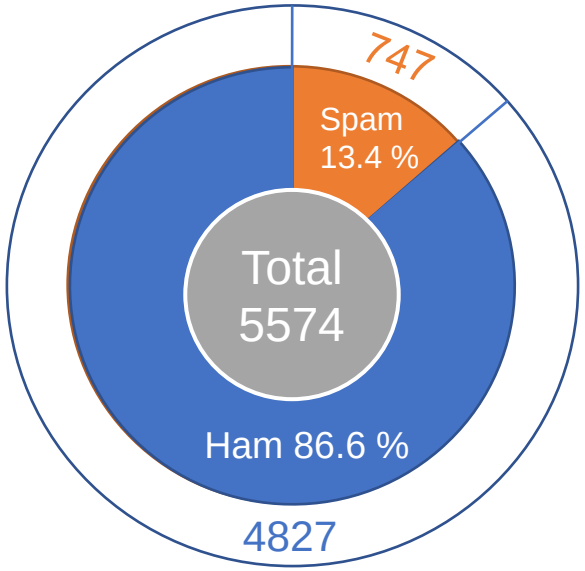


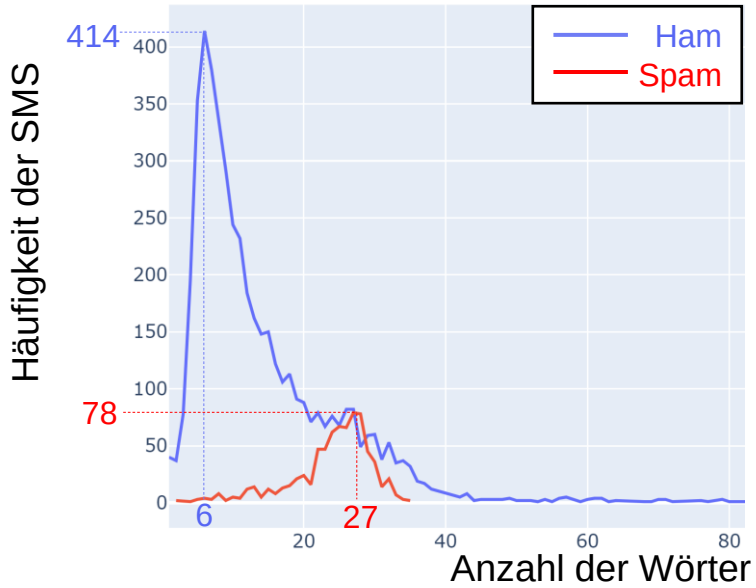
Short Message Service = SMS

Beschreibung des SMS-Datensatzes

Informationen über Datenbasis [1]:

- Zusammengesetzt aus diversen Subsätzen
- Insgesamt 5574 Datensätze: 4827 Hams und 747 Spams

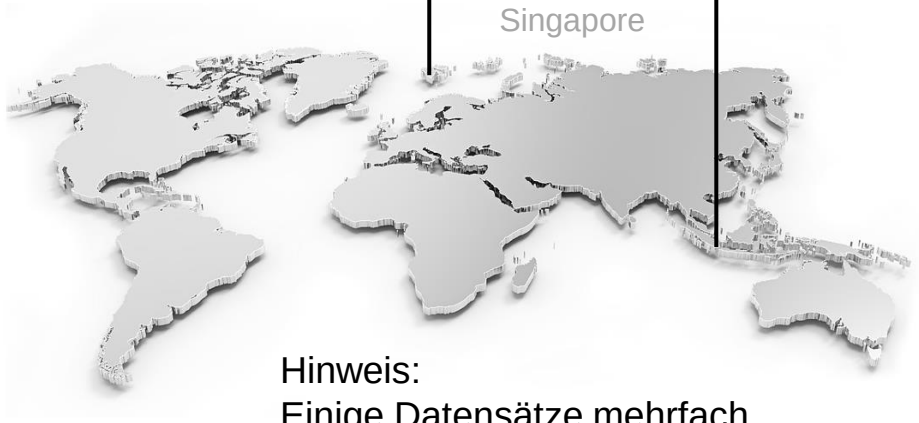




450 Hams
- Dissertation C. Taag, Univ. Birmingham [2]

1002 Hams,
747 Spams
- Handynutzer(innen)
- Beschwerden über SMS Spam
- Grumbletext

3375 Hams
- Student(innen)
- Nat. Univ. Singapore



Hinweis:
Einige Datensätze mehrfach,
einige länger als 160 Zeichen.

[1] Almeida, T.A., Gómez Hidalgo, J.M., Yamakami, A. Contributions to the study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (ACM DOCENG'11), Mountain View, CA, USA, 2011.
[2] Tagg, Caroline. "A corpus linguistics study of SMS text messaging." (2009).

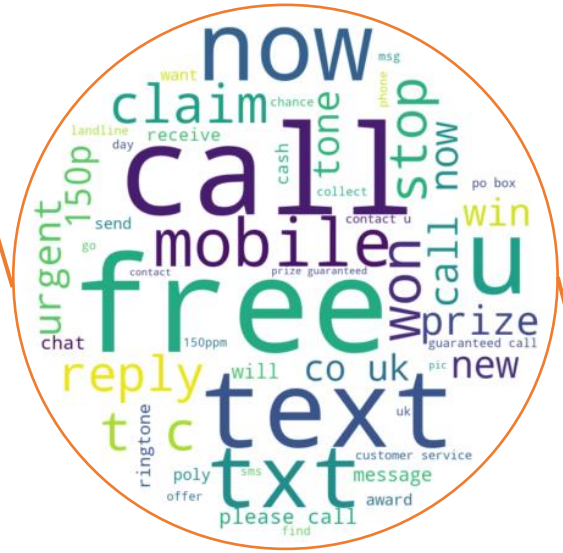
Datenaufbereitung

- Umwandlung der Textbausteine in Tokens
- Aufbereitung der Tokens als Input für verschiedene ML Algorithmen
  - Tokens als Einträge in Wörterbüchern (Bag of Words)
  - Tokens als Word-Vectors (Word2Vec) mit Kontext

50 häufigsten Wörter in Ham

Wörter aus dem Alltag

50 häufigsten Wörter in Spam



Wörter aus Werbespot

Evaluation der Modellperformance

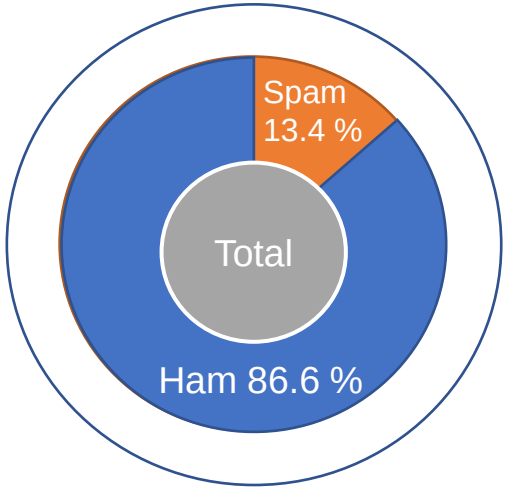
Bewertungsmetriken:

- **Konfusionsmatrix** erlaubt Bewertung der Qualität der Modelle

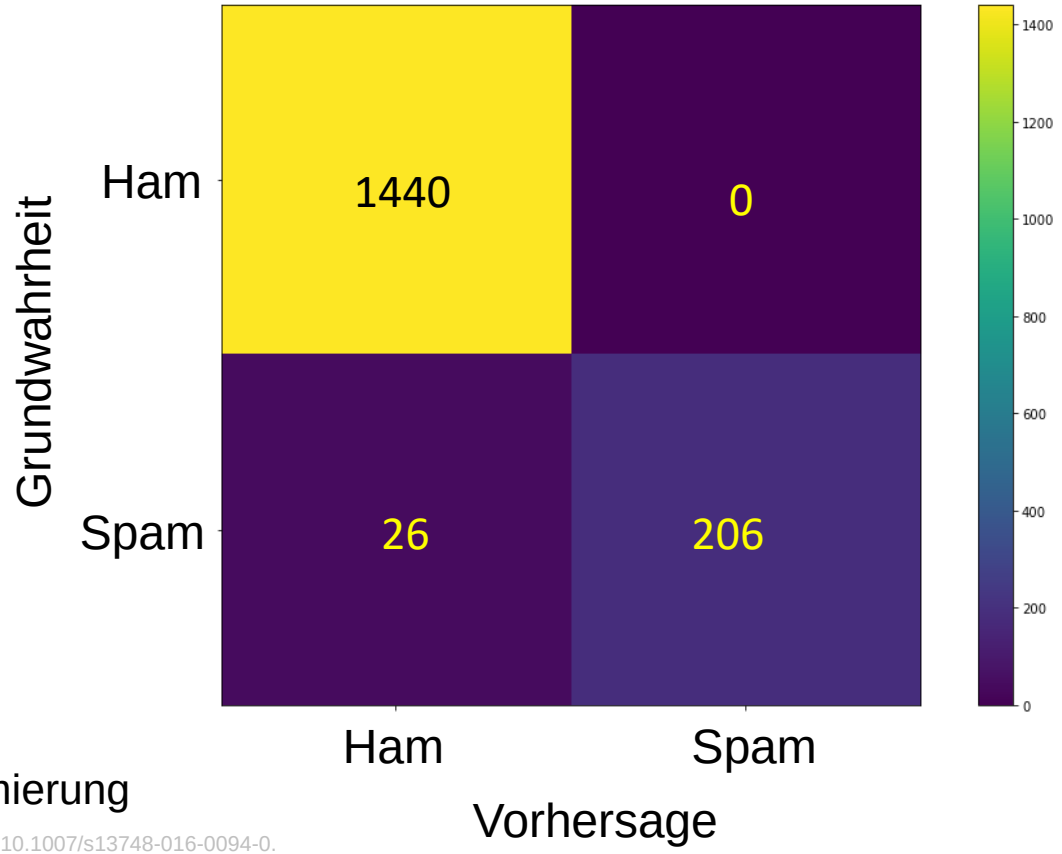
0 oder möglichst klein

Ham richtig als Ham klassifiziert	Ham falsch als Spam klassifiziert
Spam falsch als Ham klassifiziert	Spam richtig als Spam klassifiziert

möglichst groß



Imbalanced dataset [1]



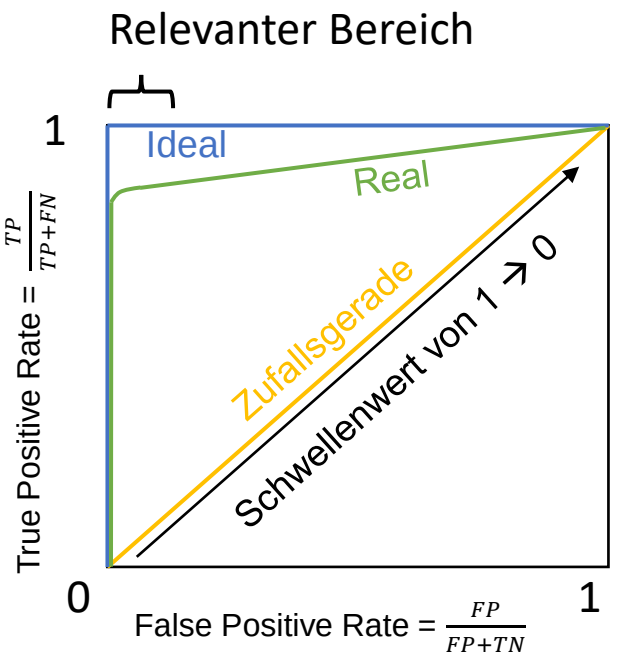
- Receiver Operating Characteristic (ROC) Plot zeigt Parameteroptimierung

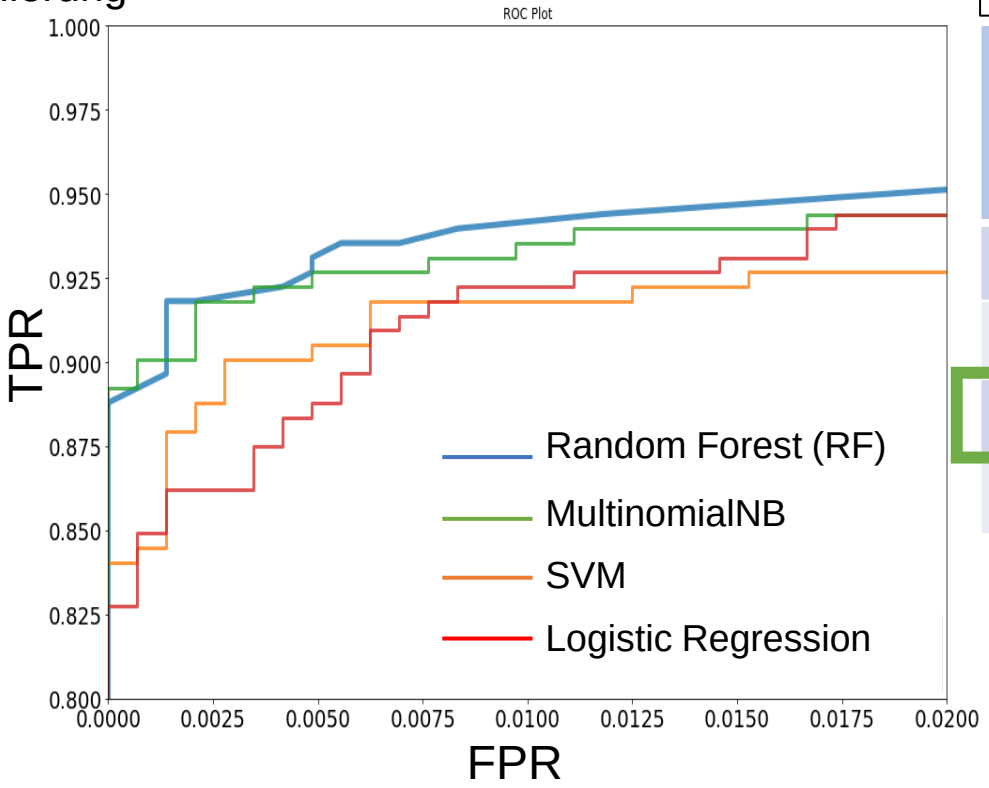
[1] Krawczyk, Bartosz. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence. 5. 10.1007/s13748-016-0094-0.

Benchmarking von ML Algorithmen

Bewertungsmetriken:

- **ROC Plot** für Parameteroptimierung





Modell liefert Zahl zwischen 0 … 1

Schwelle	Ham falsch als Spam klassifiziert	Spam richtig als Spam klassifiziert
0.25	6	214
0.30	2	211
0.35	0	206
0.40	0	199

Optimum bei Schwelle 0.35

Schlussfolgerung

- Von den untersuchten ML Algorithmen zeigte **Random Forest** das beste Ergebnis

- **Optimierung** der Modellperformance kann über breitere **Hyperparametervariation** erreicht werden

- Weitere Verbesserungen für größere Datensätze lassen sich mit aufwändigeren Methoden erzielen (z.B. BERT von Google) [1][2][3]

[1] Xu, Yang et al. “Many vs. Many Query Matching with Hierarchical BERT and Transformer.” NLPCC (2019).
[2] Lin, Jimmy J. et al. “Pretrained Transformers for Text Ranking: BERT and Beyond.” Proceedings of the 14th ACM International Conference on Web Search and Data Mining (2021).
[3] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (cite arxiv:1810.04805Comment: 13 pages).

Vielen Dank für Ihre Aufmerksamkeit !