technische universität
dortmund

Masterthesis

# Multilabel-Klassifizierung von Nachrichten Schlagzeilen

Vergleich zwischen neuronalen Netzen und baumbasierten Algorithmen auf verschiedenen Repräsentationen von Wörtern

Fakultät Statistik
Lehrstuhl Statistical Methods for Big Data

Betreuer: Prof. Dr. Andreas Groll
Verfasser: Marc Schmieder

# Inhaltsverzeichnis

# 1 Einleitung

- nutzen der klassifikation für die Redaktion
- oft nur binäre klassifikation, multi selteneres topic
- multiclass auch bei next best offer
- bag of words kein guten Ruf, bei welchen Datensätzen lohnen sich Neural nets überhaupt.

# 2 Datensatz und Problemstellung

In diesem Kapitel wird der für diese Thesis relevante Datensatz vorgestellt. Nach dessen Bereinigung erfolgt eine Exploration und anschließend die Darlegung der Zielstellung dieser Thesis.

## 2.1 Initialer Datensatz

Der Datensatz trägt den Titel *News Category Dataset* (Misra, 2018) und stammt von der Machine Learning Plattform *Kaggle*. Er umfasst 200853 Beobachtungen, die Informationen in englischer Sprache über Artikel der US-Amerikanischen Onlinezeitung *Huffpost* enthalten. Der Zeitraum, in dem die Veröffentlichungen stattgefunden haben, erstreckt sich vom 28.01.2012 bis zum 25.05.2018, also über eine Zeitspanne von über 6 Jahren. Die Inhalte der Artikel sind lediglich verlinkt und nicht direkt im Datensatz enthalten. Für jeden Artikel ist die Nachrichtenschlagzeile des Artikels angegeben. Zusätzlich zu dem Link des Artikels ist für jeden Datenpunkt das Veröffentlichungsdatum, der Name des Autors, eine Kurzbeschreibung und die Nachrichtenkategorie gegeben. Letzteres ist die Zielvariable (genauere Erläuterung in Kapitel 2.4), die 41 verschiedene Ausprägungen annimmt. Die Kurzbeschreibung enthält ähnliche Informationen wie die Nachrichtenüberschrift und ist nur teilweise vorhanden. Für die Beantwortung der Fragestellung (Kapitel 2.4) soll nur die Schlagzeile als abhängige Variable in die Modellierung eingehen. Bevor eine Exploration des Datensatz erfolgt, werden im nächsten Abschnitt vorgenommene Änderungen an den relevanten Spalten Nachrichtenkategorie und Nachrichtenschlagzeile aufgelistet und begründet.

## 2.2 Änderungen am Datensatz

In der US-amerikanischen Sprache spielt die Groß- und Kleinschreibung außer bei der Nutzung von Personalpronomen keine Rolle. Deshalb wurden in den Texten alle Großbuchstaben zu Kleinbuchstaben konvertiert. Auf diese Weise werden in der Modellierung beispielsweise die Wörter *Teacher* und *teacher* nicht unterschiedlich behandelt.

Die Artikel wurden vermutlich von einigen Autoren in unterschiedlichen Ländern geschrieben, denn die Texte enthalten unterschiedliche Zeichensätze. Bei der verwendeten `utf8` Enkodierung entstanden bei unbekannten Zeichen Konvertierungsfehler (z.b. der Form "a@S"). Diese wurden durch Leerzeichen ersetzt. In dem Wissen, dass die Wörter des Textkorpus mit den *Global Word Vectors* (Kapitel 3.2.5) abgeglichen werden, wurden einige Begriffe so ersetzt, dass bestimmte Wörter in den *Global Word Vectors* gefunden werden. Zuerst erfolgte eine Entfernung von Sonderzeichen wie beispielsweise "©" oder "™". Dann folgte die Ersetzung von Verneinungen wie zum Beispiel "n't" durch " not". Analog wurden "'ll" durch " will" und "'ve" durch " have" ersetzt. Kurzformen der Form "here's" wurden zu "here is" geändert, da sonst die Wörter mit Endung "'s" so als eigenständige Wörter repräsentiert werden und nicht sinngemäß als Tupel. Häufig vorkommende Eigennamen mit analoger Endung "trump's" wurden durch "trump his" ausgetauscht. Nachdem Vorkommnisse der Art "here's" entfernt worden, können nun Vorkommnisse der Art "john's son" durch "john its son" ersetzt werden. So ist bei Wörtertupeln dieser Art zwar nicht das Geschlecht von John bekannt, aber zumindest offensichtlich, dass der Sohn John zugehörig ist. Nach der Bereinigung des Textkorpus wurden letztendlich noch 6 Schlagzeilen entfernt, die keine Wörter mehr enthalten. Es verbleiben nun also insgesamt 200847 Beobachtungen.

Nach der umfangreichen Bereinigung des Schlagzeilen-Textkorpuses liegt nun die Zielvariable Nachrichtenkategorie im Fokus. Bei genauerer Betrachtung der 41 Kategorien fällt auf, dass diese teilweise bereits namentlich sehr ähnlich ausfallen. In Tabelle 2.2 sind beispielhaft 4 Schlagzeilen der Kategorien *parents* und *parenting* aufgeführt.

| Beispiel | Kategorie *parents* | Kategorie *parenting* |
|---|---|---|
| 1 | 40 tweets that sum up life with 4-year-olds | a baby book of disasters |
| 2 | these were the trendiest baby names in the late '80s | it is time to find your tribe |
| 3 | these quotes from kids are hilarious, adorable and oddly insightful | help huffpost parents win a webby award! |
| 4 | 30 'star wars'-inspired names parents are giving their babies | why our 'imperfect' moments are perfect to our children |

Tabelle 1: Beispiele für Schlagzeilen der Kategorien *parents* und *parenting*

Anhand der Beispiele wird deutlich, dass es schwierig ist, diese mit menschlicher Intuition eine der beiden Kategorien eindeutig zuzuordnen. Als zusätzlicher Indikator, der für die Verschmelzung zweier Kategorien spricht, erfolgte die Betrachtung der relativen Schnittmenge der gemeinsamen häufigsten Wörter. Die häufigsten Wörter pro Kategorie wurden ermittelt, indem die kompletten Daten auf die entsprechende Kategorie gefiltert werden, anschließend Symbole und *stopwords* (Wörter wie "he", "is" oder "through", die komplette Liste ist im Anhang unter (todo) zu finden) entfernt und die Wörter nach der gesamten Anzahl ihres Auftretens sortiert werden. Tabelle 2.2 zeigt die relative Schnittmenge der 50 häufigsten Wörter für ausgewählte Paare an Kategorien.

| Kategorie 1 | Kategorie 2 | relative Schnittmenge |
|---|---|---|
| *parents* | *parenting* | 0.74 |
| *arts* | *culture & arts* | 0.52 |
| *culture & arts* | *arts & culture* | 0.38 |
| *arts* | *arts & culture* | 0.40 |
| *the worldpost* | *worldpost* | 0.46 |
| *style* | *style & beauty* | 0.52 |
| *green* | *environment* | 0.50 |
| *wellness* | *black voices* | 0.16 |
| *politics* | *home & living* | 0.06 |

Tabelle 2: Relative Schittmenge der 50 häufigsten Wörter für Paare an Kategorien

Die ersten 7 Zeilen beinhalten Kategorien, die eine relative Schnittmenge von gemeinsamen Wörtern von 0.38 oder höher haben. In den letzten beiden Zeilen ist zu sehen, dass inhaltlich verschiedene Kategorien eine vergleichbar geringe Schnittmenge an häufigsten Wörtern haben.

Mit den Argumenten der menschlichen Intuition und der Ergebnisse aus Tabelle 2.2 wurden in folgenden Fällen die Kategorien zusammengelegt:

Die Kategorien *parents* und *parenting* wurde zu *parents*, *arts*, *culture & arts* und *arts & culture* zu *arts & culture*, *the worldpost* und *worldpost* zu *worldpost*, *style* und *style & beauty* zu *style & beauty* sowie *green* und *environment* zu *green & environment*. Beispiele analog zu 2.2 für die anderen zusammengelegten Kategorien finden sich im Anhang (todo verlinken, auflisten). Die 41 Kategorien wurden somit auf 35 Kategorien reduziert, welche inhaltlich mit menschlicher Intuition unterscheidbar sind. Nach den Modifikationen folgt im nächsten Abschnitt eine Exploration des Datensatzes.

## 2.3 Exploration des bereinigten Datensatzes

Von großem Interesse ist die Verteilung der Nachrichtensparten im kompletten bereinigten Datensatz. Abbildung 2.3 zeigt die absoluten Anzahlen der Datenpunkte pro Nachrichtenkategorie.

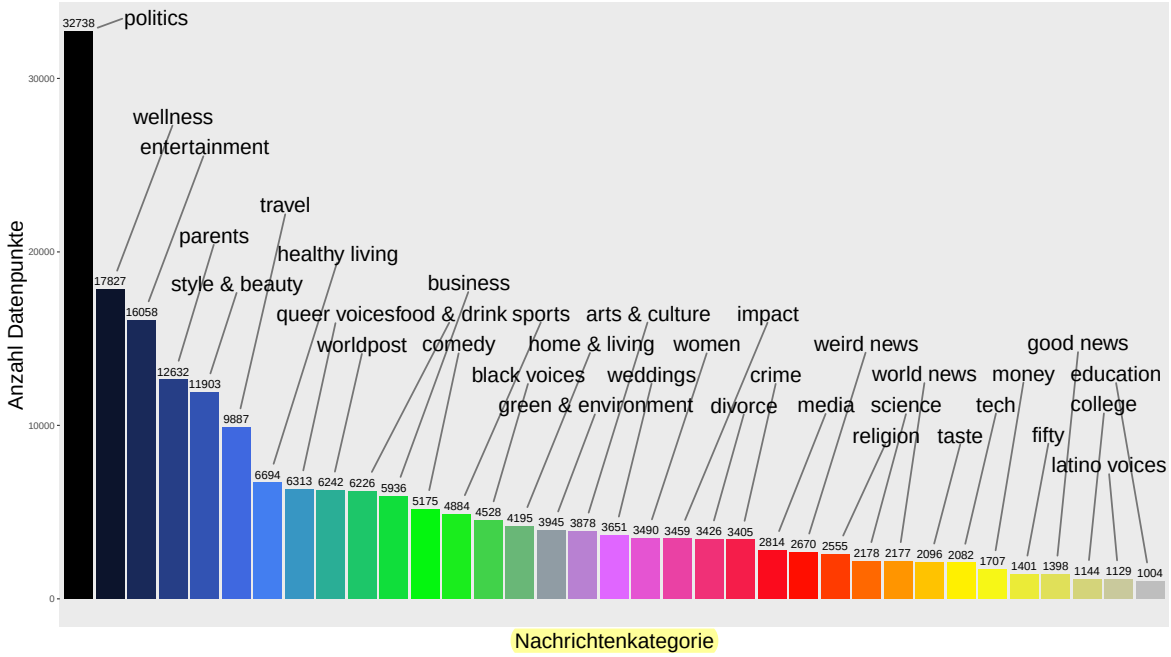


Abbildung 1: Anzahl Datenpunkte pro Nachrichtenkategorie

todo: box um die labels größer machen.

Es ist festzustellen, dass die Kategorien keineswegs ausgeglichen vorliegen. Die häufigste Kategorie stellt *politics* dar mit 32738 Datenpunkten. Zweit- und dritthäufigste Kategorien sind *wellness* und *entertainment* mit 17827 und 16058 Beobachtungen. Die Nachrichtensparten mit den wenigsten Artikeln bilden *college*, *latino voices* und *education* mit 1144, 1129 und 1004 Beobachtungen.
Die ersten 6 Kategorien stellen bereits 50.3 Prozent der gesamten Beobachtungen dar. Durchschnittlich beinhaltet eine Kategorie 5738.5 Nachrichtenschlagzeilen.

Es folgt nun eine gesamtheitliche Exploration des Textkorpus der Schlagzeilen. Im Rahmen der Analyse zählen Symbole sämtlicher Art auch als Wörter. Die kürzeste Überschrift des Datensatzes enthält nur 1 Wort, während die längste 91 Wörter umfasst. Durchschnittlich enthält eine Schlagzeile 11.147 Wörter. Das Vokabular aller Schlagzeilen umfasst 67938 Wörter, wobei “the” das häufigste Wort ist und in 54033 Artikelüberschriften vorkommt. 31274 Wörter kommen nur einmal vor. In der Betrachtung der mittleren Wortanzahlen pro

Kategorie fällt auf, dass diese differieren. Die Kategorie mit der höchsten durchschnittlichen Anzahl von 12.863 Wörtern ist *style & beauty*. Die Kategorie, bei der sich die Autoren durchschnittlich am kürzesten fassen, ist *food & drink* mit 9.328 Wörtern. Eine weitere interessante Fragestellung ist, ob in den Kategorien bestimmte Sonderzeichen oder Symbole besonders häufig oder selten vorkommen. Abbildung 2.3 zeigt die relative Anzahl der Vorkommnisse verschiedener Symbole in den Schlagzeilen pro Kategorie.

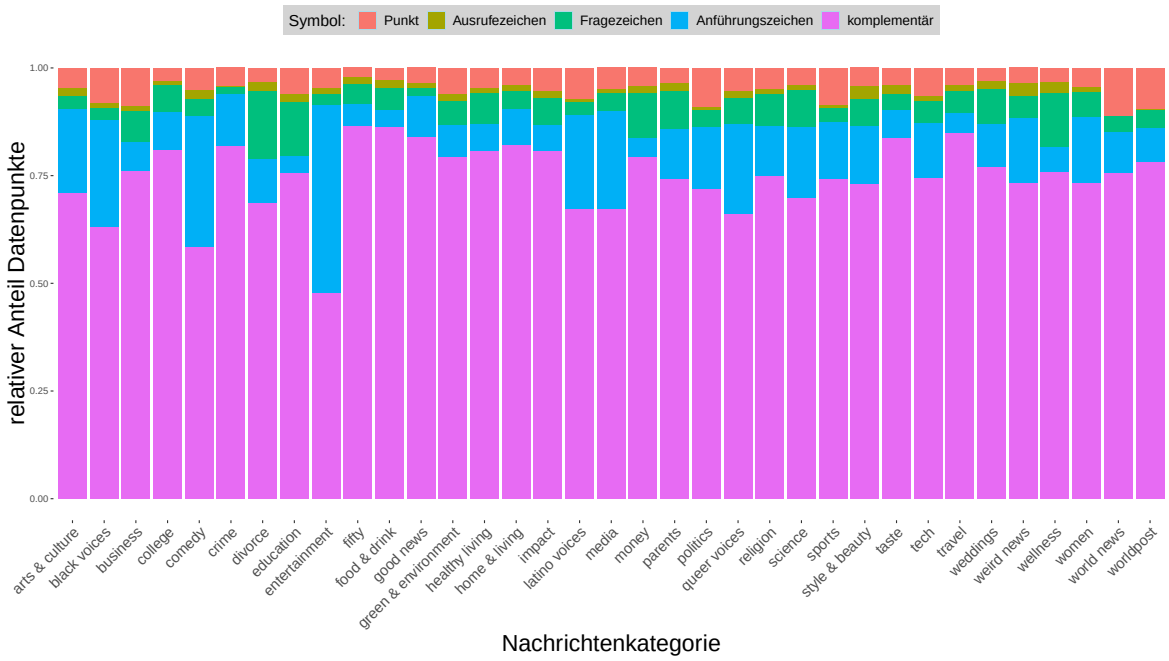


Abbildung 2: Relativer Anteil Datenpunkte mit Sonderzeichen pro Kategorie

Ein Wert von beispielsweise 0.15 für eine bestimmte Nachrichtensparte in der Grafik ist so zu interpretieren, dass in 15 Prozent aller Schlagzeilen dieser Kategorie das entsprechende Symbol mindestens einmal aufgetaucht ist. Die im Folgenden beschriebenen Durchschnitte sind Mittelwerte über die Kategorien und nicht über den gesamten Datensatz. Betrachtet sei nun das Vorkommen eines Punktes in einer Schlagzeile. Dieses kann so interpretiert werden, dass eine Schlagzeile 2 mehrere Sätze enthält. Dies ist insgesamt mit einem durchschnittlichen relativen Anteil von 0.051 selten der Fall. In der Sparte *world news* kommen mehrere Sätze am häufigsten vor, in *fifty* am wenigsten. Der Mittelwert für Ausrufezeichen beträgt 0.014 und die Kategorie *style & beauty* nimmt das Maximum, die Kategorie *world news* das Minimum an. Fragezeichen kommen in durchschnittlich 0.059 der Schlagzeilen vor, dabei am häufigsten in *divorce* und am seltensten in *crime*. Anführungszeichen sind mit durchschnittlich 0.127 von den hier betrachteten Satzzeichen am meisten vertreten. Sie wurden besonders oft in der *entertainment* Sparte genutzt und kamen am seltensten in *education* zum Einsatz. Die Rubrik "komplementär" gibt an, zu welchem relativen Anteil keine der betrachteten Symbole

vorkommt. Hier ist zu sehen, dass Kategorien wie *entertainment*, *comedy*, *black voices* oder *divorce* häufig Symbole beinhalten, die einen dramatischen Charakter ausdrücken. Sparten wie *fifty*, *food & drink* und *travel* bleiben mit weniger Symbolen sachlicher. Abbildung 2.3 zeigt insgesamt, dass Symbole für die Nachrichtensparten eine Rolle spielen und für die in Kapitel 3.2.1 und 3.2.2 beschriebenen *bag-of-words* und *tfidf* (todo, hier entfernen, falls die Algos doch nicht genutzt werden) Ansätze nicht entfernt werden sollten.

Es folgt nun eine Betrachtung der vorkommenden Wörter im Datensatz. Abbildung 2.3 zeigt eine *wordcloud*, in der die häufigsten 100 Wörter aller Kategorien (nach Entfernung von Symbolen und *stopwords*) abgebildet sind.

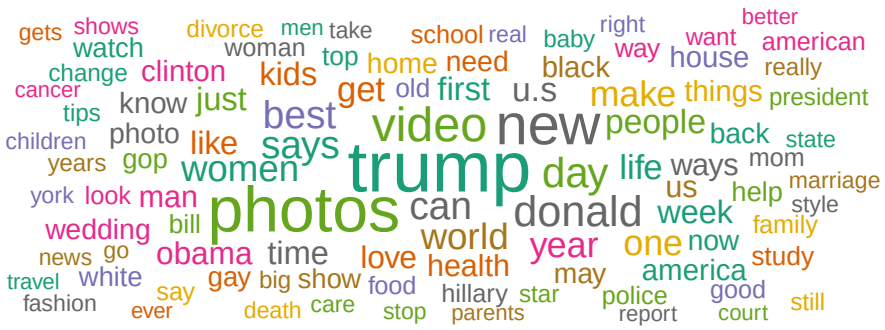


Abbildung 3: *Wordcloud* für die häufigsten 100 Wörter aller Kategorien

todo: grafik gescheit ausfüllen

Je größer ein Wort, desto häufiger kommt es insgesamt im Textkorpus vor. Es ist erstaunlich, dass "trump" sich über den gesamten Textkorpus als häufigstes Wort etabliert hat, in Anbetracht dessen, dass sich der Zeitraum der Daten auf über 6 Jahre erstreckt. Die 3 häufigsten Wörter danach sind "photos", "new" und "video". Es ist zu sehen, dass viele Namen und Begriffe aus der Politik zu sehen sind, was einleuchtend ist, da *politics* zu größte Sparte darstellt. Auffallend ist außerdem, dass viele der auftauchenden Begriffe identisch oder fast identisch zu den Namen einiger Kategorien sind. Beispiele dafür sind "travel", "wedding", "style" oder "parents". In Abbildung 2.3 ist eine weitere *Wordcloud* der zweitgrößten Sparte *wellness* zu sehen.


Abbildung 4: *Wordcloud* für die häufigsten 100 Wörter der Kategorie *wellness*

In dieser Kategorie ist zu beobachten, dass oft über Schlaf, Yoga, Meditation und Sport geschrieben wird. Es werden aber auch Krankheiten wie Krebs ("cancer"), Übergewichtigkeit ("obesity") und Diabetis angesprochen, die genauso für die Kategorie *healthy living* stehen könnten. "yoga", "breast", "diet" oder "pounds" könnten auch legitime Schlagwörter für die "women" Sparte sein. Genauso ist zu vermuten, dass die Wörter "eat", "foods" und "healthy" ebenso in "food & drink" oft vorzufinden sind. Mit dem Blick auf die *Wordclouds* und auch in Anbetracht der Kategorienzusammenlegung in Kapitel 2.2 fällt auf, dass es nicht immer einfach ist, die Kategorien nur über ihre häufigsten Schlagwörter auseinanderzuhalten. Auf die verschiedenen Möglichkeiten, wie die Schlagzeilen in ein numerisches Datenformat überführt werden können, wird in Kapitel 3.2 detailiert eingegangen. Nachdem nun der Datensatz ausführlich exploriert wurde, folgt im nächsten Unterabschnitt die Zielstellung dieser Thesis.

## 2.4 Zielstellung

Ziel dieser Arbeit soll es sein, verschiedene Algorithmen auf verschiedenen Arten der Repräsentation der Wörter miteinander zu vergleichen. Dabei liegt ein *multilabel*-Klassifikationsproblem mit 35 Klassen vor, bei der jede Beobachtung genau einer Klasse zugehörig ist. Die Zielvariable liegt im kategoriellen Format vor. Dabei soll für die Vorhersagen $f(x_i)$ der trainierten Modelle gelten, dass

$$f(x_i) = p_i = (p(c_1), ...., p(c_{35})) \qquad \text{, mit } \sum_{i=1}^{35} p(c_i) = 1,$$

dabei sind $p(c_j)$, $j = 1, ..., 35$ die modellierten Wahrscheinlichkeiten der Zugehörigkeit der Beobachtung $x_i$ zur Klasse $c_j$, die in Summe 1 ergeben müssen. Die eindeutige prognostizierte Klasse $\hat{y}$ wird dann zugeordnet durch

$$\hat{y}_i = \underset{c_1,...,c_{35}}{argmax}\ p(c_j).$$

Bezüglich der Güte der Modelle soll dann mit den Gütemaßen aus Kapitel 3.1 ein Vergleich erfolgen. Dabei ist nicht nur von Interesse, wie gut die Modelle insgesamt abschneiden, sondern auch ob manche Modelle bestimmte Kategorien trennschärfer identifizieren können und welche Gründe es dafür gibt. Des Weiteren soll analysiert werden, in welche Nachbarklassen die Beobachtungen bei einer Fehlklassifikation eingeordnet werden und ob diese inhaltlich nahe an der richtigen Kategorie ist. Eine weitere Untersuchungsfrage für die Modellgüte ist, inwiefern die Repräsentation der Wörter ausschlaggebend sind und welches Gewicht der verwendete Algorithmus dabei einnimmt. Nun sollen im nächsten Abschnitt die statistischen Methoden beschrieben werden. (todo: diesen Abschnitt im Laufe der Arbeit ergänzen)

# 3 Statistische Methoden

Dieses Kapitel ist in 3 Abschnitte unterteilt. Zuerst werden die zur Bewertung der Modelle herangezogenen Gütemaße beschrieben. Es folgt die Darlegung der verschiedenen Methoden zur numerischen Repräsentation von Wörtern. Zuletzt folgt die ausführliche Beschreibung der in dieser Arbeit benutzten *machine-learning* Algorithmen.

## 3.1 Gütemaße zur Evaluation der Modelle

Eine Bewertung der Modelle erfolgt dann über verschiedene Kennzahlen. Einige davon werden unter Verwendung der *confusion matrix* (oder auch Klassifikationsmatrix) berechnet. Diese in Tabelle 3.1 zu sehende Matrix stellt nach dem Anwenden des Modells auf den Testdatensatz die resultierenden Richtig- und Falschklassifikationen übersichtlich dar.

| | Prognostizierte Klasse | | | |
|---|---|---|---|---|
| Wahre Klasse | Kategorie $c_1$ | ... | Kategorie $c_n$ | Zeilensumme |
| Kategorie $c_1$ | $h_{11}$ | $\dots$ | $h_{1n}$ | $\sum_{j=1}^{n} h_{1j}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| Kategorie $c_n$ | $h_{n1}$ | $\dots$ | $h_{nn}$ | $\sum_{j=1}^{n} h_{nj}$ |
| Spaltensumme | $\sum_{i=1}^{n} h_{i1}$ | $\vdots$ | $\sum_{i=1}^{n} h_{in}$ | $N = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}$ |

Tabelle 3: Übersicht über eine *confusion matrix* für $n$ Klassen (vgl. Backhaus, Erichson, Plinke und Weiber, 2016, S. 238

Auf der bauen dann *accuracy*, *Sensitivität*, *Spezifität* und *f1-score*

- unter anderem accuray maximal
- f1 score wegen imbalanced
- mlogloss/cross entropy, damit nachbarklassen nicht stark bestraft werden

### 3.2 Repräsentationen der Wörter und Sätze (mit preprocessing der tokens

#### 3.2.1 Bag-Of-Words

#### 3.2.2 Term Frequency Inverse Document frequency

#### 3.2.3 Sequentielle Darstellung

#### 3.2.4 Word-To-Vec Überwacht, Summe von Word-To-Vec

#### 3.2.5 Glove Embeddings

### 3.3 Algorithmen und Verfahren

#### 3.3.1 Extreme Gradient Boosting

#### 3.3.2 Random Forest

#### 3.3.3 Neuronales Netz: Multi-Layer-Perception

#### 3.3.4 Neuronales Netz: Convolutional Neural Net

#### 3.3.5 Neuronales Netz: Long-Short-Term-Memory Neural Net

- grafik über kombination von word embeddings und Algorithmen (was kann mit was verwendet werden)

## 4 Statistische Auswertung

### 4.1 Vorauswahl der Verfahren

- grafik mit framework zu train/test/validation (10 prozent vorauswahl benchmarken, dann auf 90 prozent traintest/tuning. Die selben Modelle dann auf 10 prozent valdata validieren? 10 prozent war, weil dann in der kleinsten klasse noch mindestens 100 Beobachtungen sind.
- außerdem welche tokens entfernt wurden.
- tabelle mit allem was ich ausprobiert habe

### 4.2 Anwendung der Modelle

#### 4.2.1 Performanzmaße

- accuracy, mse etc
- accuracy by class comparison
- confidence vs accuracy plots

- maß wie sicher ist sich das Verfahren, wenn es die richtige klasse ist?

#### 4.2.2 Explaining der besten Modelle

- beobachtungen verändern, wörter wegnehmen, hinzufügen, reihenfolge ändern und schauen ob das verfahren stabil /sensitiv zur reihenfolge
- nachbarklassen identifizieren
- convolutional filters holen und ähnlichkeiten zu combinationen aus word vectors erhalten

#### 4.2.3 Anpassung des besten Modell auf den gesamten Datensatz

# 5 Zusammenfassung

## 5.1 Ergebnisse

## 5.2 Fazit und Ausblick

## Literaturverzeichnis

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (14. Aufl.). Gabler Verlag.

Misra, R. (2018). News Category Dataset. doi:10.13140/RG.2.2.20331.18729