FOM Hochschule für Oekonomie & Management

Hochschulzentrum DLS

Projektarbeit

im Studiengang Big Data & Business Analytics

über das Thema

Deep Learning zur Aktienkursprognose mit multimodalen Daten

von

Paul Hornig und Admir Dutovic

Dozent :	M.Sc. Maher Hamid
Matrikelnummer :	701650 und <AdmirNr>
Abgabedatum :	24. Dezember 2024

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

DEPATIS Deutsches Patentinformationssystem

CRISP-DM Cross Industry Standard Process for Data Mining

1 Einleitung

Bereits seid Jahrzenten wird dem Bereich Aktienprognose von Wissenschaftlern als auch Investoren große Aufmerksamkeit gewidmet[1]. Das liegt vor allem daran, dass man mit korrekten Vorhersagen sehr hohe Profite erreichen kann. Die bisher durchgeführte Forschung hat ergeben, dass numerische Aktien-Daten allein lediglich bis zu einem gewissen Grad zur Verbesserung der Leistung von Deep Learning Modellen beitragen[2].

Diese Arbeit widmet sich daher der Untersuchung inwieweit Aktienprognosen durch Betrachtung von multimodalen Daten verbessert werden können. Durch den aktuellen technischen Fortschritt gibt es viele Möglichkeiten nicht-numerische Daten einzubinden. Diese Arbeit fokusiert sich auf die Erprobung von vortrainierten großen Sprachmodellen, mit deren Hilfe Stimmungsdaten erzeugt werden sollen. Ein weiterer Schwerpunkt ist die Ermittlung einer geeigneten DL-Architektur, welche mit Numerik- und Textdaten trainiert wird.

Die Strukturierung dieser Arbeit orientiert sich am Crisp-DM Modell. In Kapitel **??** wird daher zunächst wichtiges Domänenwissen behandelt. Anschließend erfolgt eine Beschreibung der praktischen Umsetzung mitsamt Evaluierung. Darüber hinaus wird auf Möglichkeiten eines Deployment eingegangen. Zum Schluss erfolgt eine zusammenfassende Analyse in Form eines Fazits.

Das Crisp-DM Modell erfüllt unseren Anspruch an Struktur und Vollständigkeit, wobei vor allem das enthaltene iterative Konzept in unserem Anwendungsfall Vorteile mit sich bringt. Denn falls möglich, soll bei unzureichenden Ergebnissen der Prozessanfang bis Ende auf Verbesserungsmöglichkeiten untersucht werden.

[1] *Zhang, Q.* et al., 2022, Kap. Introduction.
[2] Ebd., Kap. Introduction.

2 Theoretische Grundlagen

2.1 Datenverständnis

—

Here's the German translation: Im Finanzbereich werden Aktien in 9 Branchen kategorisiert: Grundstoffe, Konsumgüter, Gesundheitswesen, Dienstleistungen, Versorgungsunternehmen, Mischkonzerne, Finanzwesen, Industriegüter und Technologie. Da Aktien mit hohem Handelsvolumen tendenziell häufiger auf Twitter diskutiert werden, wählen wir die zweijährigen Kursbewegungen von 88 Aktien vom 01.01.2014 bis 01.01.2016 als Ziele aus, bestehend aus allen 8 Aktien der Mischkonzerne und den Top 10 Aktien nach Kapitalvolumen aus jeder der anderen 8 Branchen (siehe ergänzendes Material).

—[3]

2.2 Aktienprognosen mit Deep Learning

2.2.1 LSTM für Kursdaten

2.2.2 LLM für Stimmungsdaten

—

FinancialBERT applies domain-specific language understanding to financial text analysis. Built by ahmedrachid, this model stands alongside other financial sentiment analyzers like finbert-tone and finbert. The model was fine-tuned on the Financial PhraseBank dataset, achieving 98% weighted average precision across sentiment categories. ...

—[4]

[3] *Xu*, *Y.*, *Cohen*, *S. B.*, 2018, Kap. 3.
[4] *Hazourli*, *A. R.*, 2022.

2.2.3 CNN für Stimmungsdaten

Abbildung 1: Aufbereitung von Amazon Kundenbewertungen

label	review
__label__2	Grisham's Best!: Grisham knows just how to keep you reading. His amazing plots are believable and gripping. In 'A Time to Kill', he made you think of what you would d...
__label__1	Horrible buzzing!: The buzzing sound was so loud on this baby monitor that I could not hear the baby at all. I tried different positions for the receiver and monitor ...

label	review_tokens	review_features	feature_vec
1	grisham best grisham know keep reading amazing plot believable gri...	grisham best grisham know keep reading amazing plot believable gri...	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1943 389 1943 2428 2393 3525 136 3254 ...
0	horrible buzzing buzzing sound loud baby monitor could hear baby t...	horrible sound loud baby monitor could hear baby tried different p...	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2114 4122 2622 292 2...

3 Methodik

Die Umsetzung erfolgt in logischen Schritten und orientiert sich dabei am Prozessmodell Cross Industry Standard Process for Data Mining (CRISP-DM).

3.1 Datenbeschaffung

Aufgrund dessen, dass Aktienprognosen mit multimodalen Daten ein sehr belebtes Forschungsfeld ist, gibt es bereits viele sinnvoll zusammengestellte Datensätze. Um diesen Vorteil auszuschöpfen wird ein anerkannter Finanzdatensatz gewählt, welcher in mehreren Wissenschaftsarbeiten zum Einsatz kam[5,6,7]. Die Daten werden auf GitHub unter der MIT-Lizenz zur Verfügung gestellt[8].

Der Datensatz umfasst Kursdaten und Tweets zu 88 Aktien, wobei jeweils eine Rohfassung und eine vorverarbeitete Variante bereitgestellt wird. In dieser Arbeit wird in beiden Fällen die Rohfassung verwendet.

Vom Initial-Beschaffer wird angegeben, dass sich die Daten auf den Zeitraum 01.01.2014 bis 01.01.2016 beziehen[9]. Es gibt jedoch auch Abweichungen, wie im Fall der Aktie "BABA", bei der Kursdaten für den Bereich 19.09.2014 bis 09.01.2017 vorliegen.

Im Rahmen dieser Arbeit werden lediglich Aktien der Branche Technologie einbezogen (Kap. 2.1), welche im entsprechenden Zeitraum unter den Top 10 Aktien nach Handelsvolumen vorkommen.

Kursdaten

Zu jeder Aktie liegt eine `csv`-Datei vor. In Abbildung 2 werden die inkludierten Metriken dargestellt.

[5] *Xu*, *Y.*, *Cohen*, *S. B.*, 2018.
[6] *Xu*, *H.* et al., 2020.
[7] *Zhang*, *Q.* et al., 2022.
[8] *Xu*, *Y.*, *dtaylor-530*, 2017.
[9] *Xu*, *Y.*, *Cohen*, *S. B.*, 2018, Kap. 3.

Abbildung 2: GOOG Kursdaten

```
1   Date,Open,High,Low,Close,Adj Close,Volume
2   2013-12-31,554.043335,558.406982,551.064453,558.262512,558.262512,2725900
3   2014-01-02,555.647278,556.788025,552.060730,554.481689,554.481689,3656400
4   2014-01-03,555.418152,556.379578,550.401978,550.436829,550.436829,3345800
5   2014-01-06,554.426880,557.340942,551.154114,556.573853,556.573853,3551800
6   2014-01-07,560.399475,567.717041,558.486633,567.303589,567.303589,5124300
```

Quelle: Eigene Darstellung

Tweets

Zu jeder Aktie liegt für jeden Tag im jeweiligen Zeitraum eine Datei mit einer Tweet-Liste vor. Abbildung 3 stellt einen Eintrag dieser Liste dar, wobei lediglich relevante Attribute einbezogen werden.

Abbildung 3: GOOG Tweet vom 01.01.2014


```
{"created_at":"Wed Jan 01 03:59:03 +0000 2014", "id":418229860210057216,
"text":"RT @philstockworld: Summary of Yesterday's Webcast Featuring $AAPL $WYNN
$GOOG $LGF #TraderEducation #Options #HedgingStrategies -- http:\/\/\u2026"}
```


Quelle: Eigene Darstellung

3.2 Datenvorverarbeitung

Die Datenvorverarbeitung erfolgt logisch in 4 Teilschritten. Während des 1. Arbeitsschritts werden die Daten aller Aktien in eine `csv`-Datei zusammengefasst, wobei zur Unterscheidung eine zusätzliche Datenspalte mit Marktkürzel als Werte hinzugefügt wird (Abb. 4, Spalte ”stock”).

1. Zeitliche Ausrichtung der Daten

In diesem Arbeitsschritt werden pro Aktie Kursdaten und Tweets zu einer `csv`-Datei zusammengefasst. Dabei werden zunächst alle Kursdaten-Einträge entfernt, dessen Erstellungsdatum außerhalb des Zeitraums liegt, für den Tweets vorliegen. Da diese Abweichung auch entgegengesetzt auftreten kann, werden auch alle Kurznachrichten gelöscht, deren Erstellungsdatum sich außerhalb des Bereichs der Kursdaten befindet. Abbildung 4 zeigt eine resultierende Datenzeile.

Abbildung 4: GOOG Datenzeile nach Schritt 1

```
1   date,open,high,low,close,adj close,volume,tweets,stock
2   2014-01-06,554.42688,557.340942,551.154114,556.573853,556.573853,3551800,"['Google Inc :
    Charleston SEO Company Matthew Rubin Marketing Services ... $GOOG http://t.co/JgvTg01bHJ',
    ""2013's Best Smartphone\nhttp://t.co/ImopbnMjc6 $APPLE $GOOG $FB $TWTR #Samsung"",
    '""@abnormalreturns: Sunday links: reflexive buybacks, economic optimism &amp;amp; short vs
    long-term thinking http://t.co/4pMVcwHOkJ $AMZN $GOOG $FB""', 'Watch out world, Google
    does it again http://t.co/41MlxOacDY #aviateEUROPE $ARM $CSR $GOOG', ...]",GOOG
```

Quelle: Eigene Darstellung

Alle Tweets die außerhalb von Markttagen entstanden sind, werden dem nächstmöglichen Markttag zugeordnet. Somit enthält die Zeile aus Abbildung 4 auch Kurznachrichten vom 04.01 und 05.01.2014.

2. Textbereinigung

Um die Tweets für eine Weiterverarbeitung durch ein großes Sprachmodell vorzubereiten, werden mehrere Bereinigungen durchgeführt. Dabei wird darauf geachtet, dass keinerlei symantische Informationen verloren gehen. Stopp-Wörter wie beispielsweise ”und” werden daher nicht entfernt. In Abbildung 4 ist zu erkennen, dass in Tweets häufig URL’s eingebunden sind. Diese werden mit ”URL” maskiert. Des weiteren treten regelmäßig Referenzierungen mittel ”@” auf. Diese werden mit der Maskierung ”AT_ENTITY” ersetzt. Außerdem erfolgt eine Substituierung von Aktienvorkommen der Form ”$<Kürzel>” mit ”<Kürzel> stock”. In Abbildung 5 ist die bereinigte Form der Tweets-Spalte aus Abbildung 4 zu sehen.

Abbildung 5: GOOG Datenzeile nach Schritt 2

```
1   date,open,high,low,close,adj close,volume,tweets,stock
2   2014-01-06,554.42688,557.340942,551.154114,556.573853,556.573853,3551800,"['Google Inc :
    Charleston SEO Company Matthew Rubin Marketing Services ... GOOG stock URL', ""2013's
    Best Smartphone URL APPLE stock GOOG stock FB stock TWTR stock #Samsung"", '""AT_ENTITY:
    Sunday links: reflexive buybacks, economic optimism &amp;amp; short vs long-term thinking URL
    AMZN stock GOOG stock FB stock""', 'Watch out world, Google does it again URL
    #aviateEUROPE ARM stock CSR stock GOOG stock', ...]",GOOG
```

Quelle: Eigene Darstellung

3. Stimmungsmetriken

Dieser Arbeitsschritt sieht vor, dass pro Datenzeile anhand der enthaltenen Tweet-Liste Stimmungsmetriken extrahiert werden. Zunächst erfolgt eine Ableitung der Relevanz einer Aktie pro Tag quantifiziert als Anzahl veröffentlichter Tweets (Code 1).

Code 1: Relevanz quantifiziert als Anzahl-Tweets

```
df['num_tweets'] = df['tweets'].apply(len)
```

Quelle: Eigene Darstellung

Anschließend wird für jede Tweet-Liste eine durchschnittliche Quantifizierung der Metriken ”positiv” und ”negative” berechnet. Hierfür wird das vortrainierte Sprachmodell ”FinancialBERT” eingesetzt (Kap. 2.2.2). Die genaue Funktionsweise ist im Code 2 beschrieben.

Code 2: Stimmung quantifiziert als Positiv/Negativ-Score

```
# Initialisierung des Sprachmodells
model =
   BertForSequenceClassification.from_pretrained("<Pfad>")
tokenizer = BertTokenizer.from_pretrained("<Pfad>")

# Initialisierung der Klassifikations-Pipeline
nlp = pipeline("sentiment-analysis", model=model,
   tokenizer=tokenizer)

# Funktion zur Extraktion von Sentiment Metriken über eine
   Tweet-Liste
def comp_sent(texts):
  sent_res = []
  for text in texts:
    res = nlp(text)
    # Klassifikation-Resultat für einen Text
    # Format: [{'label': <Wert>, 'score': <Wert>}]
    sentiment = res[0]

    # Umwandlung zu Format: [positive_score,
       negative_score]
    res_formatted = [float(sentiment['label'] ==
       'positive') * sentiment['score'],
    float(sentiment['label'] == 'negative') *
       sentiment['score']
    ]
    # Hinzufügen des Resultats zur Liste
    sent_res.append(res_formatted)

```

```
24  # Berechnung des Durchschnitts der "Tuple"
25  res = list(np.mean(sent_res, axis=0)) if len(sent_res) >
       0 else [0, 0]
26
27  # Rückgabe des Durchschnitts über alle Tweets an diesem
       Tag als Dictionary
28  return {
29    'positive': res[0],
30    'negative': res[1]
31  }
```

Quelle: Eigene Darstellung

4. Stimmung-Embeddings

Zusätzlich zu Stimmungsmetriken soll pro Tweet-Liste ein repräsentative Embedding-Vektor erzeugt werden. Es wird angenommen, dass dadurch für ein neuronales Netz detaillierte Informationen bezüglich Tagesstimmung bereitgestellt sind. Hierfür werden pro Tag alle gesammelten Tweets einer Aktie verbunden und anschließend zur Berechnung des Embeddings an das FinancialBERT-Modell übergeben. Dazu werden zunächst alle Tweets zu einem Text zusammengefasst wobei das Sonderzeichen ”[SEP]” als Bindeglied dient. Anschließend wird dem String das Zeichen ”[CLS]” vorangestellt. Dies ist eine, seitens des zugrundeliegenden ”BERT”-Modells, empfohlene Vorgehensweise für Klassifkationsaufgaben, weshalb das Folgemodell ”FinancialBERT” auch mit dieser Methode optimiert wurde[10]. Anschließend wird der kombinierte Text zu einer Token-ID-Liste umgewandelt (Code 3).

Code 3: Umwandlung von Tweets zu Token-IDs

```
1  # Verbinden der Tweets mit Special-Token zu einem Text
2  text = ' [SEP] '.join(tweet_list)
3  # Klassifikations-Sonderzeichen voranstellen
4  text = '[CLS] ' + text
5
6  # Text zu Token-IDs umwandeln
7  inputs = tokenizer(
8  text,
9  padding=True,
```

[10] *Hazourli, A. R.*, 2022, Kap. 5.3.

```
10  truncation=True,
11  return_tensors="pt"
12  )
```

Quelle: Eigene Darstellung

Die Argumente `padding` und `truncation` in Zeile $8-9$ von Code 3 stellen sicher, dass falls das Resultat die maximale Kontextlänge von 512 Token unter- respektive überschreitet, entweder mit einem Sonderzeichen aufgefüllt wird oder ein entsprechendes Zuschneiden auf 512 IDs erfolgt. Im letzten Schritt werden die Tokens an das Sprachmodell übergeben, welches mit berechneten Embeddings antwortet. Während des iterativen CRISP-DM Prozesses hat sich herausgestellt, dass ein Zuschneiden der ID's bei Längenüberschreitung für eingesetzte neuronale Netzwerke (Kap. 3.3) besser funktioniert, als das Aufteilen in Chunks mit anschließender Durchschnittsberechnung der Teil-Embeddings.

Ergebnis

Der resultierende Datenkorpus ist im `csv`-Format gespeichert und enthält eine Kombination aus Kurs- und Stimmungsdaten von 10 Aktien verteilt auf 5470 Zeilen (6).

Abbildung 6: GOOG Datenzeile im Datenkorpus

```
1  date,open,high,low,close,adj close,volume,stock,num_tweets,positive,negative,tweet_embs
2  2014-01-06,554.42688,557.340942,551.154114,556.573853,556.573853,3551800,GOOG,9,0.2161561581823561,
   0.1096002194616529,"[-1.6030482053756714, 0.28848686814308167, -0.37600037455558777, 0.
   3213549852371216, 1.8825640678405762, 2.0237033367156982, 0.075238898396492, 1.0776742696762085,
   -2.66947603225708, 0.4654025435447693, -1.4391454458236694, -2.0213849544525146, ...]"
```

Quelle: Eigene Darstellung

3.3 Modellierung

3.4 Evaluierung

4 Auswertung

Bei der Trendanalyse ist zu beachten, dass aktuell veröffentlichte Patente (2024-Q2) Innovationen der jüngsten Vergangenheit (bei ungeprüften i.d.R. 18 Monate) darstellen. Die zeitliche Dimension in den Abbildungen **??**-**??** und **??**-**??** ist daher im Hinblick auf Innovationsstärke leicht verzerrt.

Die Tabelle **??** im Kapitel 3.1 macht deutlich, dass China mit Abstand die meisten Innovationen im Robotik-Kontext hervorbringt. Der aktuelle Trend weist jedoch einen plötzlichen, deutlichen Rückgang auf. Diese Entwicklung ist ungewöhnlich und die Gründe dafür können vielfältig sein. Es kann zum Beispiel möglich sein, dass chinesische Unternehmen zunehmend mehr Wert darauf legen, Innovationen im Bereich Robotik und KI vollständig verdeckt zu halten, ähnlich zu der Vorgehensweise von OpenAI mit GPT-4[11].

Europa ist, wie zu erwarten, das Schlusslicht. Auch nach Anwendung eines, auf Beschäftigtenzahlen je Region basierenden, Äquivalenzfaktors blieb die Überlegenheit Chinas in nahezu allen Bereichen bestehen (Abb. **??**, **??** u. **??**). Die Analysen ergaben, dass die USA im Bereich Medizin und Teleoperation im Vergleich zu Europa und China am intensivsten an Neuerungen gearbeitet hatten, so dass dort der größte gewichtete Anteil an Patenten veröffentlicht wurde.

Die Überlegenheit Chinas bei der Patentanzahl muss nicht zwingend bedeuten, dass dieses Land die größte Innovationskraft aufweist. In dieser Arbeit wurden veröffentlichte Patente einbezogen, aber deren Neuartigkeit ist nicht zwangsläufig geprüft und auch eine Bewertung des Einfallsreichtums ist nicht gegeben. So kann es durchaus sein, dass in den USA bei Patenten mehr Wert auf Qualität bzw. Einfallsreichtum der Erfindung gelegt wird, so dass banalere Neuerungen bei der Prüfung abgelehnt werden.

Zukünftige Arbeiten

Die Auswertung führte zur Offenlegung von Limitationen, welche in zukünftigen Arbeiten ausgebessert werden. Ein wichtiger Schritt zur genaueren Vergleichbarkeit der Innovationskraft ist die Eingrenzung der Patente auf bereits Geprüfte. Außerdem soll eine Methode zur Bewertung des Einfallsreichtums der Erfindung umgesetzt werden, denn eine sehr gute Idee kann für mehr Innovation sorgen als viele moderate. In dieser Arbeit wurden lediglich Patentveröffentlichungen aus dem Deutsches Patentinformationssystem (DEPATIS)

[11] **vincent2023openai**.

Dokumentenarchiv einbezogen. Ein weiterer wichtiger Schritt ist eine Ausweitung der Datenquellen. Dabei sollen mindestens all jene Quellen einbezogen werden, welche Bezug zu den involvierten Regionen aufweisen. Zudem sollen auch wissenschaftliche Arbeiten als Indikator für Innovationskraft hinzugefügt werden. Auch eine Erweiterung beziehungsweise Anpassung der einbezogenen Regionen wird für mehr und verlässlicheren Informationsgehalt der Analysen sorgen.

5 Fazit

Im Hinblick auf die Forschungsfrage lässt sich zusammenfassen, dass die untersuchten Regionen in Bezug auf Schwerpunktsetzung viele Gemeinsamkeiten aufweisen und sich nur in Nuancen unterscheiden. Erschreckend, aber nicht überraschend war der interkontinentale Fokus von Robotik mit Bezug zu militärischen Zwecken, wobei dieser in China am ausgeprägtesten ist. Als aktuell innovativste Region, welche zudem auch am intensivsten Patente veröffentlicht, wurde mit deutlichem Vorsprung China ermittelt. Es ist also sehr wahrscheinlich, dass in Zukunft bahnbrechende, massentaugliche Produkte mit Robotik-Bezug auf dem Markt erscheinen, welche zum Großteil auf Innovationen aus China basieren. Die USA haben in den Bereichen Teleoperation und Gesundheit eine dominierende Innovationsintensität bewiesen. Der erhöhte Fokus dieser Region auf Robotik im medizinischen Bereich sollte von anderen Regionen zum Vorbild genommen werden, wobei Europa eine positive Tendenz dahingehend aufweist.

Quellenverzeichnis

Hazourli, *Ahmed Rachid* (2022): FinancialBERT - A Pretrained Language Model for Financial Text Mining, in: Preprint on ResearchGate (2022), CC BY 4.0 License

Xu, *Hongfeng*, *Chai*, *Lei*, *Luo*, *Zhiming*, *Li*, *Shaozi* (2020): Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices, in: Neurocomputing, 418 (2020), S. 326–339

Xu, *Yumo*, *Cohen*, *Shay B.* (2018): Stock Movement Prediction from Tweets and Historical Prices, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018), hrsg. von *Gurevych*, *Iryna*, *Miyao*, *Yusuke*, S. 1970–1979

Xu, *Yumo*, *dtaylor-530* (2017): StockNet Dataset: A Comprehensive Dataset for Stock Movement Prediction from Tweets and Historical Stock Prices, MIT License, <https://github.com/yumoxu/stocknet-dataset> (2017) [Zugriff: 2024-12-20]

Zhang, *Qiuyue*, *Qin*, *Chao*, *Zhang*, *Yunfeng*, *Bao*, *Fangxun*, *Zhang*, *Caiming*, *Liu*, *Peide* (2022): Transformer-based attention network for stock movement prediction, in: Expert Systems with Applications, 202 (2022)