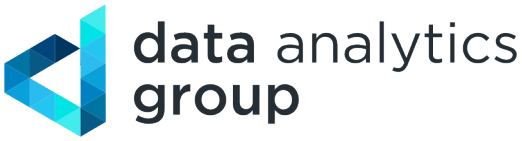

Universität Paderborn
Fakultät für Wirtschaftswissenschaften
Department Wirtschaftsinformatik

Studienarbeit

Predictive Analytics
Daily Financial News for 6000+ Stocks

von

Benjamin Kasten
Matrikelnummer: 7154784
Warburger Str. 100, 33098 Paderborn
bkasten@mail.uni-paderborn.de

und

Dominik Höhr
Matrikelnummer: 7157684
Warburger Str. 100, 33098 Paderborn
dhoehr@mail.uni-paderborn.de

15. August 2020

Eidesstattliche Erklärung

Hiermit erkläre ich, Benjamin Kasten, an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Paderborn, den 15. August 2020

_____

Benjamin Kasten

Eidesstattliche Erklärung

Hiermit erkläre ich, Dominik Höhr, an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Paderborn, den 15. August 2020

_____

Dominik Höhr

Abstrakt

@bkasten füg mal bitte das Exposé hier ein!

Stichworte: data science, machine learning

Inhaltsverzeichnis

1 Business Understanding **1**

2 Data Understanding **6**

3 Data Preparation **10**

4 Modeling **16**

4.1 Sentiment Analyse, Lineare Regression 16

4.1.1 Model bauen 16

4.1.2 Model bewerten 17

4.2 Bag-Of-Words und TF-IDF mit Random Forest Regression 19

4.2.1 Model bauen 19

4.2.2 Modell Bewertung 21

4.3 Auswertung 21

5 Evaluation **22**

5.1 Ergebnissevaluierung im Hinblick auf die Geschäftsziele 22

5.2 Evaluierung des Prozesses 23

5.3 Ausblick 23

6 Deployment **24**

6.1 Strategie 24

6.2 Implementation 25

6.3 Überwachen und Pflegen 25

Bibliography **26**

Abbildungsverzeichnis

2.1 Überblick über den Datensatz, Sieben zufällige Einträge 6
2.2 Boxplot, unaufbereitet 8
2.3 Top 50 Wörter im gesamten Datensatz, unaufbereitet 9

3.1 Zwischenergebnis nach Tokenization, POST und Kleinschreibung 13
3.2 Zwischenergebnis Snyset Format 13
3.3 Überblick über den konstruierten Datensatz, Sieben zufällige Einträge (url, publisher ausgeblendet) 15
3.4 Top 50 Wörter im gesamten Datensatz, bereinigt 15

4.1 Vergleich von vier Linearen Regressionen 18

Tabellenverzeichnis

4.1 Hyper Paramter Tuning 21

1 Business Understanding

Motivation

Seit der Digitalisierung und dem Zugang zum Internet für den Großteil der Weltbevölkerung werden Nachrichten fast schon in Echtzeit auf Onlineportalen veröffentlicht und frei zugänglich gemacht. So ist es jeder Person mit Internetzugang möglich, aktuelle Nachrichten fast ohne Zeitverzug parallel zu einer riesigen Menge anderen Menschen zu konsumieren, und sich selber mittels Kurznachrichten in sozialen medien selbständig überall schnell zu äußern, um seine Meinung oder eine Information öffentlich zu machen (Agarwal et al., 2016). Darunter fallen auch Nachrichten bezüglich Aktien (im Folgenden auch Aktiennews). Aktiennews sind Nachrichten, die sich mit der Entwicklung von Aktienpreisen beschäftigen. Zu einer Aktiennews gibt es immer eine Headline. In dieser Headline versuchen die Autor:innen des Artikels, das Thema der Nachricht kurz und knapp zu beleuchten und zusammenzufassen. So beinhalten viele Headlines unter anderem die Änderung des Aktienpreises und das zugehörige Unternehmen. Durch die Headline wissen die Besucher eines jeweiligen Nachrichtenportals, wie sich ein Aktienkurs entwickelt, ohne den gesamten Artikel lesen zu müssen. Personen, die in Aktien, ETFs etc. investieren und Aktiennews lesen, reagieren möglicherweise direkt auf die aktuellste Headline, ohne den Artikel dazu zu lesen. Dabei haben Headlines häufig eine bestimmte Intention, die entweder postiv oder negativ zu einem bestimmten Thema steht. Solche Headlines können natürlich direkt die gesellschaft an sich beeinflussen (Agarwal et al., 2016) weshalb die Analyse von Aktiennews interessante Ergebnisse liefern kann. Beispielsweise kann auf eine negative Headline zu einem Unternehmen, dessen Aktien ein Kunde besitzt, zum schnellen Verkauf dieser führen. Positive Nachrichten könnten zu einem Anstieg der Nachfrage nach den Aktien des Unternehmens führen und somit dessen Wert erhöhen. Doch nicht immer muss die Headline die Entwicklung des Aktienkurses offenbaren, jedoch wird der Inhalt der Nachricht oft anhand der Headline bewertet. Eben dieser Umstand kann zu fehl Interpretationen führen (Dor, 2003)
Der Fokus auf die Headline an sich, ohne den restlichen Artikel, ist gerade deshalb so interessant, weil Menschen häufig einige Headlines überfliegen, statt den kopmletten Inhalt von Nachrichten zu lesen (Dor, 2003). Headlines spielen also eine entscheidene

Rolle bei der Meinungsbildung, sowohl auf die Sichtweise als auch die Emptionen, die in Zusammhang mit einem Produkt oder Szenario stehen. Selbst kürzeste Headlines spielen eine nicht zu vernachlässigende Rolle bei der Beurteilung des (möglichen) Inhalts einer Nachricht (Agarwal et al., 2016)

Warum sollte jetzt ein Computer sich damit beschäftigen, Artikelheadlines zu analysieren, und zu bewerten, ob sich der Aktienkurs der in der Headline genannten Unternehmen verändern könnte? Natürlich könnten auch Menschen sich die Headlines durchlesen, und entsprechend darauf reagieren, bspw. mit einem Kauf oder Verkauf der Aktie. Jedoch können Computer um einiges schneller Texte verarbeiten, und somit auch schneller reagieren. Ein Computer kann also diese Arbeit für den Menschen übernehmen, und somit seinen Erfolg als Anleger erhöhen. Zudem ist es Möglich das ein entsprechender Algorithmus weitere Zusammenhänge erkennt, so wie zum Beispiel das Verhalten der Aktienkurse zu den veröffentlichten Überschriften. Es ist durch aus vorstellbar, das eben aufgrund der subjektive affektiven Handlung der Leser die Einschätzung zum Kaufen bzw. Verkaufen der Aktie aus historischer und kalkulierter Sicht falsch sind.

Forschungsfrage

Wir wollen untersuchen, ob es einen Zusammenhang zwischen den vorkommenden Wörtern in den Headlines und dem Aktienkurs nach der Veröffentlichung der Headline gibt. Lässt sich also anhand der Wörter und dem Sentiment einer Headline die Veränderung des Aktienkurs, auf die sich die Headline bezieht, vorhersagen?

Die Sentiment Analyse bezieht sich darauf, wie die Wörter eines Satzes (hier die Überschrift) aus Sicht des Menschen empfunden wird (vgl. Agarwal et al., 2016). Die Sentiment Analyse wird auf Basis von Sentiment-Dictionarys durchgeführt, diese beinhalten eine Zuordnung von Wörtern zu einem Score, welche die Gefühlslage des Wortes ausdrücken soll. Die Skala des Scores kann dabei ganz unterschiedlich aufgebaut sein, unter anderem: Positiv bis Negativ (5 bis -5) (Årup Nielsen, 2015) oder verschiedenen Gefühle entsprechend (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) (Mohammad, 2020). Innerhalb unserer Möglichkeiten prüfen wir, mit welchem Preis eine Aktie an dem Tag schließt, an dem die Headline veröffentlicht wurde. Grunflage dafür wäre ein Korrelation zwischen Sentiment Score und Stockpreisentwicklung.

Ziele und Profit

Ziel ist es zukünftige Überschriften einordnen zu können. Damit die Auswirkungen einer veröffentlichten Headline einzuschätzen sind und gegebenenfalls Maßnahmen ergriffen werden können.
Dies hat sowohl für Unternehmen als auch für Inhaber der Aktien Vorteile. Die Unternehmen können so schnell den Überschriften entgegen wirken und den Aktienkurs somit stabil halten. Die Anleger haben aufgrund unseres Algorithmus eine gewisse Grundlage um nicht vorschnell, durch ihre Subjektive Auffassung der Überschrift, zu handeln.

Verwandte Literatur

Das Thema Stock Price Prediction hat in der Literatur in den letzten Jahren einen großen Zuwachs bekommen. Hier listen wir einige Quellen auf, die in Zusammenhang mit unserem Forschungsthema stehen.
Lárló Nemes, Attila Kiss (2021) haben vier verschiedene Ansätze zur Sentiment Analyse von Economic News genutzt. Um die verschiedene Ansätze vergleichen zu können, wurde zu erst BERT genutzt, und dann die Tools Vader, Textblob und ein RNN. Es hat sich herausgestellt, dass BERT und RNN im Vergleich zum Vader Tool und Textblob deutlich einen deutlich besseren Senitment Score ermitteln konnten, ohne neutrale Ergebnisse. Darauf basierend konnte durch den Abgleich des Sentiments und der Stockpriceentwicklung der Moment festgestellt werden, der den Effekt der Headline auf den Aktienkurs abbildet.
Arul Agarwar (2020) hat mit Hilfe des Python Tools VADER eine Analyse der Nachrichten von einzelnen Unternehmen vorgenommen. Es wurden ganze HTML Seiten heruntergeladen,aus diesen die wichtigen Informationen herausgefiltert. Dann wurde ein wörterbuchbasierter Ansatz mithilfe des VADER Tools verwendet, um den einzelnen Nachrichten eine Sentiment Score zuzuweisen. Dem VADER LExicon wurden zudem weitere Wörter hinzugefügt bzw. das Sentiment geändert, um eine Missinterpretation (wegen dem Finanzmarkt) zu verhinden. Zwischen dem Sentiment der Nachrichten und dem Stock Price konnte eine starke Korrelation erkannt werden, die spätestens am nächsten Werktag auftrat.
Branko Kavsek (2017) stellte sich die Frage, ob es ein Wort oder eine Wortkombination gibt, deren Vorhandensein und deren Abwesenheit etwas über die Stockprice entwicklung aussagt. Dafür wurden unterschiedliche Modelle genutzt, die Teil des WE-

KA Tools sind. Bei allen Ansätzen wurde ein Overfitting festgestellt, da es einen starken Abfall der Accuracy beim Testset gab. Jedoch haben der PART, C4.5 und Random Forest Algorithmus sowohl eine 90%tige Accuracy auf dem Trainingsset und jeweils mehr als 70%tige Accuracy auf dem Testset. Das Ziel der Untersuchung war, bestimmte Wortkombinationen zu finden, im Anschluss wurden Entscheidungsregeln untersucht, die für die Bewertung der Stockpreise ausschlaggebend waren.
Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018 (2018) haben für den Kontext der Vorhersage von Stockpreisen sogenannte Hybrid Attention Networks modelliert und implementiert. Grundlage hierfür sind die Schritte, die Menschen kognitiv durchführen, um mit chaotischen News umzugehen.
Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao (2016) haben unter der Annahme, das Stock News den Stockprice bestimmen, die Stockpriceentwicklung eines Unternehmen (Apple Inc.) anhand von Klassifikationsmodellen betrachtet. News (gesamte Artikel) wurden als positiv oder negativ eingestuft anhand ihrer Polarität. Genutzt wurde Random Forest, Naive Bayes und SVM. Mit einer Accuracy von 88% bis 92% abhängig von verschiedenen Test-Szenarien, wie unterschiedlicher Cross Validation, angepasstem Data Split oder ganz neuer Testdaten, konnte mittels Random Forest durch den Sentiment des Artikels der Stockpricetrend am besten vorhergesagt werden.
Saees Seifollahi und Mehdi Sharjari (2018) haben sich damit beschäftigt, Wörter zuerst in ihrem Kontext zu betrachten, bevor eine konventionelle Sentiment Analyse durchgeführt wird. Unter der Annahme, dass die Einordnung eines Wortes in den Kontext, der "Wortsinn", das Sentiment des jeweiligen Wortes überhaupt richtig identifiziert werden kann.

Nächste Schritte

Im folgenden wollen wir zunächst einen Überblick über die vorhandenen Daten geben. Also den Inhalt sowie die Aussagekraft der Daten darlegen und erläutern. Außerdem auch abgrenzen was die Daten eben nicht aussagen. Dafür werden sowohl Metadaten des Datensatzes wie auch die eigentlichen Attribute des Datensatzes genau betrachtet. Hier werden noch keine Daten gefiltert, verändert oder ergänzt.
Vorab sei gesagt, dass der gegebene Datensatz keine Zielvariable enthält und diese somit noch nachträglich ergänzt werden muss. Dies führt zu einigen Einschränkungen. Diese Einschränkungen müssen entsprechend bewertet und eingeschätzt werden um die weitere Evaluation des zu trainierenden Modell zu ermöglichen.
Im Anschluss des Datenverständnis werden die Daten, im Zuge der Datenvorbereitung,

so aufbereitet, dass diese vergleichbar und einheitlich sind. So dass das Modell die Daten entsprechend versteht. Außerdem wird der Datensatz so gefiltert, dass keine Lücken, Fehler oder nicht verwendbare Daten verbleiben. Dies geschieht vor allem aufgrund von Einschränkungen durch die Schnittstelle, welche benötigt wird um die fehlende Zielvariable zu ergänzen. Hier ist speziell darauf zu achten, das die Filterung des Datensatzes keinen, bzw. nur einen geringen, semantischen Einfluss aufweist.
Neben der Schnittstelle zu "Polygon.io", zum beziehen des Aktienkurses wird außerdem der Sentiment-Score der Headline im Datensatz ergänzt.
Nun gilt es, im Rahmen dieser Studienarbeit, ein Modell auf zubauen, welches unserer Forschungsfrage entspricht. Es wird also ein Modell entwickelt bzw. trainiert welches auf Basis des Sentiment-Scores einer Headline die Veränderung des Aktienkurses ins positive oder negative vorhersagen soll.
Zur Evaluation des Modells wird unter anderem die Accuracy betrachtet. Dies geschieht unter heranziehen unterschiedlicher Sentiment-Dictionarys.
Die gewonnen Erkenntnisse so wie das Modell werden im letzten Schritt hinsichtlich ihres Gewinn und Einsatzes für Wirtschaft und Wissenschaft analysiert.

2 Data Understanding

Der Datensatz

Gegeben ist der Datensatz "Daily Financial News for 6000+ Stocks" mit mehr als einer Millionen Einträgen von englischen Überschriften im Bereich der Aktiennews aus den USA. Die Überschriften sind dabei den jeweiligen Aktien zu geordnet, über welche der Artikel beziehungsweise die Überschrift berichtet.

Wir konzentrieren unsere Anstrengungen hinsichtlich der Daten getriebenen Analyse auf den Datensatz "raw_analyst_ratings.csv".
Der Datensatz wird von bot_developer auf Kaggle.com zur Verfügung gestellt.
(siehe Developer, 2020) Der oben genannte Datensatz, rawanalystratings, besteht aus sechs Spalten:

- **id**: Eindeutige ID
- **headline**: Überschrift
- **url**: Webseite des Artikels
- **publisher**: Immer Benzinga, daher hier Autor
- **date**: Datum der Veröffentlichung
- **stock**: Die Aktie auf die sich die Überschrift bezieht

Eine Überschrift ist immer genau einer Aktie zugeordnet. Die Spalte des Publisher stellt

	id	headline	url	publisher	date	stock
869155	873555	Stocks That Hit 52-Week Highs On Monday	https://www.benzinga.com/news/19/08/14292264/s...	Lisa Levin	2019-08-19 00:00:00	MUSA
502450	505137	Shares of several basic materials companies ar...	https://www.benzinga.com/wiim/20/04/15756537/s...	Benzinga Newsdesk	2020-04-07 00:00:00	FMC
24497	25003	10 Stocks To Watch For May 22, 2019	https://www.benzinga.com/news/earnings/19/05/1...	Lisa Levin	2019-05-22 00:00:00	ADI
1213310	1219128	Telefónica and Ericsson Sign Four to Six Year ...	https://www.benzinga.com/news/19/02/13253595/t...	Charles Gross	2019-02-27 00:00:00	TEF
1390349	1396831	Yelp Option Alert: Nov 15 $37 Calls Sweep (39)...	https://www.benzinga.com/markets/options/19/10...	Charles Gross	2019-10-28 00:00:00	YELP
1023140	1028203	UBS Maintains PPG Industries at Neutral, Raise...	https://www.benzinga.com/analyst-ratings/price...	Juan Lopez	2012-08-08 00:00:00	PPG
318165	319933	Top 4 Small-Cap Stocks In The Farm Products In...	https://www.benzinga.com/trading-ideas/small-c...	Monica Gerson	2011-01-28 00:00:00	CVGW

Abbildung 2.1: Überblick über den Datensatz, Sieben zufällige Einträge

nach Aussage von bot_developer den Autor da, Publisher ist immer Benzinga (übersetzt Developer, 2020). Die *id* dient der eindeutigen Identifizierung, da Überschriften und Aktien mehrfach vorkommen und nur die Kombination der beiden features einen eindeutigen Key bilden. Die *headline* beschreibt den zugrunde liegenden Artikel. Die Artikel selbst sind im Datensatz nicht vorhanden, wie im vorangegangenem Kapitel jedoch erwähnt wollen wir insbesondere die Aussagekraft der Überschrift bewerten. Die Überschrift bezieht sich auf die Veränderung des, ebenfalls im Datensatz genannten *stock*, also der Aktie. Hier ist abzugrenzen, dass die Headlines keine, bzw. nur selten, wirtschaftlichen Geschehnisse des Unternehmens darstellen. Die *url* und den *publisher* der jeweiligen Headline wird im folgenden nicht weiter betrachtet, verbleibt jedoch im Datensatz zur besseren Nachvollziehbarkeit. Insbesondere zur Abgrenzung der Qualität des Datensatzes, da die fehlende Varietät des Publishers die Allgemeingültigkeit des Modells einschränkt. Die Spalten *date* gibt das Datum der Veröffentlichung der Aktiennews an und ist im folgenden, zusammen mit der Aktienkennung, der Aktie auf die sich die Headline bezieht, essentiell um die Auswirkung der Headline auf eben diese zu evaluieren.
Auffällig ist, dass der Datensatz verfügt über keine wertenden Spalten, so ist eine Vorhersage jeglicher Art unmöglich. Es bleibt also im folgenden eine Zielvariable zu bestimmen und für jeden Eintrag zu evaluieren, dies geschieht im Kapitel Data Preparation.

Überblick

Insgesamt sind 1407328 Zeilen zu verzeichnen. Diese teilen sich auf 845770 eindeutige Überschriften und 6204 eindeutige Aktien zu 39957 unterschiedlichen Daten auf. Überschriften werden also mehrfach genannt, wenn sich der Artikel auf mehrere Aktien bezieht.
Die News stammen aus den Jahren von 2009 bis 2020. Pro Datum fallen etwa 600 bis 1000 Headlines an. Im Schnitt etwa 117277.3 Headlines pro Jahr. Hier zu sehen ist also die fülle von Aktiennews, und dies alleine bei einem Publisher. Ein Wandel der Anzahl über die Jahre ist nur schwer zu erkennen und somit zu vernachlässigen.
Die Verteilung der Anzahl der Aktien ist ungleichmäßig so kommen einzelne Aktien wie die "AADR", Dorsey Wright ADR ETF, ein ETF der AdvisorShares Investments, LLC, nur zwei Mal vor und andere wie die "MRK", die Aktie des Pharmaunternehmens Merck & Co., bis zu 3238 Mal.
Wie in der obenstehenden Grafik zu sehen sind viele Ausreißer der Verteilung zu erken-

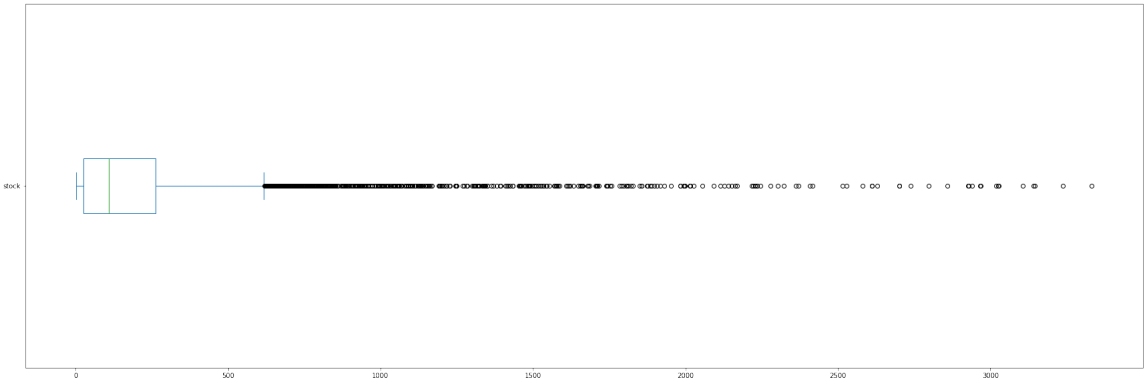

Abbildung 2.2: Boxplot, unaufbereitet

nen, die Durchschnittliche Erwähnung einer Aktie im Zeitraum von 11 Jahren liegt bei unter 200.
Die Ungleichverteilung zeigt ein weiteres Spannungsfeld im Bereich des Forschungsthema, wird jedoch im folgenden keinen semantischen Unterschied darstellen, da jede Headline separat in Ihrem Bezug zum jeweiligen Aktienkurs betrachtet wird. Denkbar sind allerdings Zusammenhänge zwischen den Auswirkungen und der Häufigkeit der Veröffentlichung einer Headline zur Aktie. Auch die Betrachtung eines Einflussfaktors der Häufigkeit der Erwähnung einer Aktie in einem kurzen Zeitintervall ist hier zu nennen.

Betrachten wir nun die Überschriften ohne diese vorher aufbereitet zu haben, so lassen sich einige Begriffe feststellen, die spezifisch am Aktienmarkt sind, dies muss auf jeden Fall im weiteren Umgang mit den Daten beachtet werden. So zum Beispiel ”ESP”, was für ”Earnings per share” steht, also dem Gewinn pro Aktie. Diese Kennzahl wird häufig zur externen Finanzanalyse eines Unternehmens verwendet, hier kann jedoch keine Gewichtung basierend auf dem Wort, bzw. der Abkürzung, getroffen werden. Vor allem auch Begriffe wie ”Price”, ”Stock”, ”Market” und ”Bezinga” geben zwar den Kontext der Aktiennews an, da dieser aber ohne hin schon feststeht, wird bei diesen Begriffen keine Gewichtung zum Aktienkurs erwartet. Zumal gängige Sentiment- und Stopword-Dictionarys keine dieser Wörter beinhalten.

Abbildung 2.3: Top 50 Wörter im gesamten Datensatz, unaufbereitet

Qualität

Es wurde kein Data Mining betrieben, die Daten stammen aus einer externen Quelle, wie dem obigen Abschnitten zu entnehmen ist. Die Daten liegen schon in einer bereinigter Form vor und beinhalten keine null-Werte, Duplikate oder Werte die im Kontext keinen Sinn ergeben. Die Zeitstempel allerdings bestehen zum Großteil nur aus tagesgenauen Daten, so sind die Uhrzeiten in 1351341 von 1407328 Fällen (96,02%) mit ”OO:OO:OO” angegeben.
Mit Hinblick auf die Forschungsfrage, schränkt dieser Tatbestand die Auswahl eines möglichen Vergleichsintervall der potentiellen Zielvariable zur Veröffentlichung der Headline massiv ein. Um einen möglichen Bezug zu der Headline sicher zustellen muss entsprechend der kürzeste Zeitraum um die Veröffentlichung des Artikels gewählt werden.
Die Überschriften sind mit reinen ASCII-Zeichen lesbar.
Laut dem Ersteller des Datensatzes stellt ”Bezinga” den einzigen Publisher, jedoch sind 1034 verschiedene Publisher zu verzeichnen. Dabei handelt es sich um Autoren bei Bezinga Insights.
Während des nächsten Schrittes, der Data Preparation, müssen die Überschriften verarbeitet werden, so dass diese durch maschinenlesbare Attribute abgebildet werden. Hier muss darauf geachtet werden, dass die gängigen Verfahren der Sentiment- oder Stopword-Dictionarys die aktienmarkt-spezifischen Wörter nicht vollständig abdecken werden. Insbesondere die Wechselwirkungen auf dem Aktienmarkt werden durch Wörter abgebildet, die durch das Verfahren des Stopword-Removal entfernt werden (Zimmermann, 2019). Darüber hinaus ist eine Evaluation des Zeitraumes zwischen einer möglichen Zielvariable und der Veröffentlichung der Aktiennews nur tageweise möglich.

3 Data Preparation

Verbleibend sind die unbereinigten Überschriften. Außerdem ist bereits auffällig geworden, dass eine möglich Zielvariable im vorhanden Datensatz fehlt. Alle vorliegenden Daten sind bislang unbewertet. Im Sinne unserer Forschungsfrage gilt es im Schritt der Konstruktion der Daten die entsprechende Zielvariable zu evaluieren und zu ergänzen, um so eine geeignete Bewertung bzw. Gewichtung der Überschriften herleiten zu können. Als Zielvariable erscheint der Aktienkurs geeignet. Dieser bietet ein objektives Abbild des Wertes der Aktie zum gegebenen Zeitpunkt. Somit verbleibt es dem Modell eine Verbindung zwischen der Veröffentlichten Headline und der Veränderung des Aktienkurses herzustellen.

Bereinigen

Aufgrund der Beschränkungen durch die kostenfreie Version, der zugrunde Liegenden API, zur Ergänzung der Aktienpreise zu den entsprechenden Headlines, sind wir gezwungen den Datensatz auf die letzten Zwei Jahre zu verkürzen. Dies hat kein semantischen Auswirkungen. Nicht abzustreiten, das entsprechende Modell würde ungemein genauer werden, wenn mehr historische Daten zur Verfügung stehen würden. Doch semantisch stellt die Anzahl der verwendeten Einträge und Insbesondere die Anzahl der verwendeten Stocks keinen Einflussfaktor dar, es werden keine möglichen Predictoren entfernt. Auf der anderen Seite allerdings ist es durchaus möglich dass das Modell, aufgrund des starken und schnellen Wandels des Aktienmarktes, eine zeitlich begrenzte Gültigkeit besitzt.
Denkbar ist eine Gewichtung des Einflusses von Headlines auf Aktien, welche nur sehr wenig oder sehr oft, über einen längeren Zeitraum, erwähnt werden. Dies ist sicherlich ein Interessantes Forschungsthema, jedoch nicht Teil dieser Arbeit.
Da nur 4% der Daten Uhrzeiten beinhalten, werden die Zeitstempel auf Tage reduziert. Somit ist nur eine tagesgenaue vorhersage möglich. Nach der Bereinigung der Daten besteht der Datensatz nun mehr aus rund 164698 Einträgen mit einer Zeitspanne vom 21.August 2019 bis zum 11.Juni 2021.

Konstruieren

Wie einleitend schon erwähnt gilt es nun die Zielvariable zu ergänzen. Hierfür nutzen wir eine API die zu den jeweiligen Aktienkennungen (Ticker) die entsprechenden Aktienpreise liefert. Bei Angabe des Tickers, eines Zeitraumes und eines Intervalls, erhält man jeweils die Börsenpreise zu Börsenbeginn (open), Börsenschluss (close), den Maximalpreis (high) und den Tiefspreis (low) zum entsprechenden Zeitpunkt.
Verwendet wird "Polygon.io". Aufgrund der nur tagesgenauen Daten verwenden wir für jede eindeutige Aktie, nach oben beschriebenen Kriterien, ein Intervall von einem Tag innerhalb der letzten Zwei Jahre zwischen dem 21.August 2019 und 31. Dezember 2020. Die jeweiligen Aktienpreise werden dann zum Datensatz ergänzt. Hier gilt es ein entsprechendes Intervall vor und nach Veröffentlichung heranzuziehen, die Größe des Intervalls wird später erörtert. Überschriften, zu denen die API keine Aktienpreise nennen kann werden aus dem Datensatz entfernt. Meist sind solche Überschriften zu den Schließzeiten der Börse erschienen, um die Auswirkungen der Headline jedoch entsprechend zuordnen zu können, werden diese Einträge nicht weiter betrachtet.
Nach dem grundsätzlichem Bereinigen des Datensatzes muss nun jede einzelne Überschrift bestimmten Schritten unterzogen werden, die später für das Modelling benötigt werden, im folgenden Headline Cleaning bezeichnet. Für das Preprocessing einer Headline greifen wir auf übliche Methoden zurück, und orientieren uns im folgenden an den Algorithmen angelehnt an Agarwal et al. (2016)

Algorithmus 1: Preprocessing jedes Wortes

1. Einzelne Headline eingeben
2. Headline mit POS-Tagger taggen
3. Lemmatisierung jedes Wortes
4. Ausgeben jedes Wortes als Input für eine weitergehnde Unteruschung mit SentiWordNet

Algorithmus 2: Analyse einer Headline

1. Alle Headlines H der Nachrichten identifizieren
2. Jede einzelne Headline mit dem Algorithmus 1 3 für SentiWordNet Analyse vorbereiten
3. Für jedes Wort jeder Headline mit SentiWordNet einen Sentiment Score berechnen (score = postiv - negativ)
4. Falls score < 0, setzte Sentiment auf -1
5. falls score > 0, setze Sentiment auf +1
6. sonst: setze score = 0

Die im Algorithmus 1 präsentierten Schritte ergänzen wir im Folgenden um kleinere, aber sehr häufig durchgeführte Normalisierungsschritte. Grundsätzlich werden aber folgende Schritte durchgeführt: Tokenization, Part-Of-Speech Tagging (im Folgenden POST), Stopword Removal und Lemmatization sowie Stemming. Die Tokenization spaltet einen Satz in einzelne Worteinheiten (die Tokenz) auf. Wir verwenden standardmäßig immer an Leerzeichen, sodass ein Token meisten ein einzlens Wort ist. Die Tokenization ist notwendig, um in den nächsten Schritten jedes einzelne Wort betrachten zu können. POST ist ein Verfahren zur Vorbereitung eines Opinion Mining (auch häufig Sentiment Analyse genannt). Mit POST kann analysiert werden, welcher Wortart ein Wort zugehörig ist. Dabei unterscheiden wir zwischen Nomen, Adjektiven, Adverben und Verben. Durch das explizite POST kann die sich anschließende Lemmatisierung der Wörter effizienter und genauer durchgeführt werden. Im weiteren werden Stopwords entfernt, die anährend keine Bedeutung für den Bedeutung einer Aussage (in unserem Fall einer Headline) haben. Dazu gehören unter anderem Wörter wie that", ïs", ä". Daran anschlißend führen wir eine Lemmatization durch. Jedes Wort wird damit in seine Wörterbuchform überführt.

Im ersten Schritt, vor der Tokenization, konvertieren wir nun also alle Wörter zur Vereinheitlichung in Kleinbuchstaben. Anschließend führen wir eine Tokenization durch. Dafür wird die im NLTK Modul enthaltene und von den Entwickler:innen des Moduls empfohlene Funktion word_tokenize verwendet. In Agarwal et al. (2016) wird ein POS-Tagger mit einer Accuracy von 97.24% verwendet. Aus Gründen der Einheitlichkeit und der Zielsetzung verzichten wir an dieses Stelle darauf, diesen POS Tagger einzubinden, sondern nutzen den im NLTK Package enthaltenen und von dessen Entwickler:innen empfohlenen POS Tagger. In Abbildung 3.1 ist ein Ausschnitt des Zwischenergebnisses des Data Cleanings zu sehen.

Im nächsten Schritt werden die Stopwords entfernt. Auf diesen Schritt folgt nun die Lemmatizierung mittels dem Wordnet Lemmatizer. Wordnet ist ein lexikalisch-semantisches

	id	headline	url	publisher	date	stock	headlines_cleaned
0	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	2020-06-05	A	[(stocks, n), (that,), (hit, v), (52-week, a)...
1	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	2020-06-03	A	[(stocks, n), (that,), (hit, v), (52-week, a)...
2	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	2020-05-26	A	[(71,), (biggest, a), (movers, n), (from,), ...
3	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	2020-05-22	A	[(46,), (stocks, n), (moving, v), (in,), (fr...
4	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	2020-05-22	A	[(b, n), (of,), (a, n), (securities, n), (mai...

Abbildung 3.1: Zwischenergebnis nach Tokenization, POST und Kleinschreibung

	id	headline	url	publisher	date	stock	headlines_cleaned
0	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	2020-06-05	A	[stock.n.01, hit.v.01, high.n.01, friday.n.01]
1	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	2020-06-03	A	[stock.n.01, hit.v.01, high.n.01, wednesday.n.01]
2	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	2020-05-26	A	[large.a.01, mover.n.01, friday.n.01]
3	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	2020-05-22	A	[stock.n.01, travel.v.01, friday.n.01, session...
4	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	2020-05-22	A	[security.n.01, neutral.n.01, technology.n.01,...

Abbildung 3.2: Zwischenergebnis Snyset Format

Netz, das seit 1985 an der Princteon University entwicklet wird. Es verfügt über zusammenhänge zwsichen Wörtern (Miller, 1995), und ist die Grundlage für viele andere Lexika, wie bspw. SentiWordNet (Baccianella et al., 2010). Durch das POST weiss die Lemmatizer Funktion, um welche Wortart es sich handelt, und kann demetsprechend das jeweilige Wort auf die jeweilige Wörterbuchform zurückführen. Zu Vereinfachung behandeln wir alle Wörter, die beim POST keiner Wortart zugeordnet werden konnten, wie Nomen. Dahinter steht die Überlegung, keine Wörter zu verlieren, die im späteren Verlauf durch das SentiWordNet mit einem Senitment belegt werden könnten. Wir entfernen anschließend erneut weitere Wörter, nämlich jene, die weniger als drei Buchstaben haben sowie solche, die aus Zahlen bestehen, also numerisch sind.

Mithilfe von SentiWordNet werden nun die Sentiments für jede Headline bestimmt. Dabei ergibt sich ein Senitment Score durch die Differenz zwischen poitivem Sentiment und negativem Sentiment. Diese Werte werden duch SentiWordNet geliefert und an die entsprechende Zeile ergänzt.

Integrieren

Die bislang zwei Datensätze der Headlines mit dem Stockticker sowie dem SentimentScore und der Stockprices müssen, zur Verwendung im Modell, vereint werden. Dazu ergänzen

wir jedem Eintrag des ersteren Datensatz mit dem *open*-Kurswert und dem *close*-Kurswert der jeweiligen Aktie am Tag der Veröffentlichung des Artikels. Um die Auswirkungen möglichst eindeutig zuweisen zu können, sollte ein entsprechend kleines Intervall verwendet werden. Um jedoch alle Auswirkungen zu erfassen, darf dieses auch nicht zu klein sein. Wir gehen davon aus, dass im Mittel, die Artikel nach Eröffnung des Aktienhandelsplatz und vor Schließung dessen publiziert werden. Dies ist auch auf der Webseite von Benzinga.com nach zu vollziehen. So wählen wir zu jeder Headline den Aktienkurs beim öffnen des Aktienhandelsplatz als Referenzwert und den Preis beim schließen als Zielvariable. In dem nun vereinten Datensatz ergänzen wir zusätzlich noch, aus dem Datensatz berechenbare, Werte wie den *Stockprice_Change* und *Senti_Binary*. Die jeweiligen Werte des Stockprice_Change bilden die Veränderung des Stockprices am Tag der Veröffentlichung der Headline vom Eröffnungskurs zum Kurs beim Schließen der Börse ab. Dabei nimmt die Spalte des Stockprice_Change eine 1 an wenn sich der Wert der Aktie um mehr als 1% steigert, -1 wenn die Aktie um mehr als 1% fällt. Sonst eine 0, wir gehe davon aus, das die Aktie sich nicht Nennenswert geändert hat, bzw. die Headline als Ursache für die Veränderung des Aktienpreises ausgeschlossen werden kann. Die jeweiligen Werte des Senti_Binary ergänzen den Datensatz um eine dreituple Betrachtung des Sentiment. Ist der verrechnete Sentiment, aus der Anzahl der Wörter mit Negativen und Positiven Sentiment, insgesamt positiv, also größer 0, so ist Senti_Binary eine 1. Ist der Sentiment negativ, also kleiner 0, so ist der Senti_Binary -1. Sonst 0.

Formatieren

Wie bereits ausführlich beschrieben, wurde das Datumsformat auf yyyy-mm-dd reduziert. Weitere Formatierungen wurden nicht unternommen, die Daten verbleiben in der Form wie sie konstruiert wurden.

Ergebnis

Der anfangs schon betrachtete Datensatz wurde nun um die bereinigten Überschriften, den Sentiment-Scores und den Aktienpreisen erweitert. Außerdem wurden die Einträge stark reduziert, unter anderem durch die Festlegung einer Datumsspanne zwischen dem 21.August 2019 und 31. Dezember 2020.
64456 headlines haben einen durchschnittlichen Sentiment von 0, davon haben 59718

	id	headline	date	stock	headline_cleaned	senti_pos_score	senti_neg_score	senti_score	open	close	stockPrice_Change	senti_Binary
16972	136506	90 Biggest Movers From Yesterday	2020-03-27	BAK	[large.a.01, mover.n.01, yesterday.n.01]	0.250	0.125	0.125	7.81	7.0600	-1	1
6652	55556	10 Basic Materials Stocks Moving In Thursday's...	2020-03-05	AKS	[basic.a.01, material.n.01, stocks.n.01, trave...	0.000	0.000	0.000	2.24	2.1700	-1	0
88436	786400	10 Restaurants Stocks Moving In Monday's Session	2020-01-27	LUB	[restaurant.n.01, stocks.n.01, travel.v.01, mo...	0.000	0.000	0.000	2.36	2.4000	1	0
56059	486944	Uber Shares Hit The Curb After Earnings. But T...	2019-11-05	FDX	[share.n.01, hit.v.01, curb.n.01, net_income.n...	0.625	0.375	0.250	164.60	164.8500	0	1
111200	973993	UPDATE: Holly, Plains All American Report Shar...	2019-10-03	PAA	[update.n.01, complain.v.01, american.a.01, re...	0.250	0.750	-0.500	20.00	20.2600	1	-1
36905	307819	Castlight Health Highlights Launch Of Director...	2020-03-26	CSLT	[health.n.01, highlight.n.01, launch.n.01, dir...	0.750	0.000	0.750	0.73	0.8101	1	1
63979	561584	Mid-Day Market Update: Dow Falls Over 500 Poin...	2020-04-13	GME	[market.n.01, update.n.01, falls.n.01, point.n...	0.125	0.000	0.125	4.25	4.7400	1	1

Abbildung 3.3: Überblick über den konstruierten Datensatz, Sieben zufällige Einträge (url, publisher ausgeblendet)

Abbildung 3.4: Top 50 Wörter im gesamten Datensatz, bereinigt

weder einen positiven noch einen negativen Sentiment-Score. Hingegegen haben 97022 Headlines einen Sentiment-Score gesetzt der ungleich 0 ist. 45005 positive, 52017 negative. Dabei haben 50671 jeweils positive und negative zu bewertende Wörter.

Es verbleiben 3722 Stocks und 76162 eindeutige Überschriften. Insgesamt also 14 Spalten auf 161478 Einträge. Die folgende Abbildung bietet einen Überblick über den aktuellen Datensatz, welcher für das Modeling verwendet wird. Zu Übersichtszwecken wurden die Spalten *url* und *publisher* ausgeblendet.
Auch in den Überschriften selbst sieht man nach der Bereinigung einige Unterschiede. Ersichtlich ist, dass die Headlines vorrangig über die Lage des Handels berichten, so erscheinen Wörter wie ”Trade” in Kombination mit ”high” und ”low” häufig. Jedoch erkennt man auch, dass einige Wörter wie ”company”, ohne weitere Aussagekraft, immer noch häufig verwendet werden. Vermutlich aufgrund des verwendeten, nicht spezifischen Stopword-Dictionary im Schritt des Stopword-Removal.

4 Modeling

Im folgenden werden verschiedene Modelle aufgebaut, um auf dem gegeben Datensatz aus dem vorangegangen Kapitel vorherzusagen, wie sich der Aktienpreis ändert. Dabei wird die Aktienpreisänderung fokussiert, und nicht der eigentlichen Aktienpreis. Also (1) steigend, (-1) fallend oder (0) unverändert. Die Zielvariable ist also stehts der im Datensatz vorhandene stockPrice_Change.

4.1 Sentiment Analyse, Lineare Regression

Zunächst betrachten wir die Änderungen des Aktienpreises anhand des im Kapitel der Data Preparation berechneten Sentiment auf wörterbuchbasis.

4.1.1 Model bauen

Durchschnittlich liegt der Sentiment-Score der Headlines bei -0.056242, der kleinste Sentiment-Score beträgt -4.2500, der größte 3.375. Nach kurzer Betrachtung der Binären Zuordnungen stockPrice_Change und senti_Binary ist zu erkennen, dass 54086 Sentiments der Headlines von 161478 positiv mit einem Aktienanstieg korrelieren. Das entspricht etwa 33%. Das bedeutet, bei positiven Sentiment steigt die Aktie, bei negativen Sentiment fällt diese. Hingegen korellieren rund 20% der Headlines, bzw. deren Sentiment, negativ mit einem Aktienanstieg.
Um dieses Vorgehen weiter zu verfeinern, wird ein lineares Regressionsmodell aufgebaut (Stojiljkovicr, 2019). Zunächst nehmen wir den Sentiment-Score der Headline als Predictor, als Input-Varible, die Zielvariable bleibt unverändert. Das lineare Regressionsmodell berechnet nun die mögliche Gewichtung der Predictoren um die Zielvariable herzuleiten (Yahya Eru Cakra, 2015). Durch die Differenzierung des Sentiment-Score in positiv und negativ, erhalten wir zwei Predictoren zur selben Zielvariable. In einem weiteren Modell wird zusätzlich der Aktienpreis zum Zeitpunkt der Börseneröffnung als Predictor herangezogen, so werden also weitere äußere Umstände in das Modell mit einbezogen. Letztlich ergänzen wir dem Datensatz die Anzahl der Veröffentlichungen,

also der Headlines, zur selben Aktie an diesem Tag und nutzen diese zusammen mit dem positiven und negativen Sentiment-Score als Predictoren. So können Zusammenhänge zwischen dem Aktienpreis und der wiederholten Erwähnung einer Aktie in das Modell mit einfließen.

4.1.2 Model bewerten

Die reine Korrelation ist wenig Aussagekräftig, da nur 33% einen direkten Zusammenhang aufweisen. Dies liegt primär an den 61614 Headlines, welche einen Sentiment Score von 0 besitzen und somit nicht zu deuten sind. Dies widerum liegt ist vermutlich auf das nicht auf den Finanzmarkt spezialisierte Sentiment-Dictionary zurück zuführen (Feldman, 2013) Um die entstandenen Modelle nun mit einander zu vergleichen und zu bewerten, wenden wir die jeweiligen Modelle, trainiert auf einem zufälligem train_set von rund 70% des gesamt Datensatzes, auf einem test_set (die restlichen 30% des Datensatzes) an. Somit enthält der test_set Datensatz vier Predictions und den tatsächlichen StockPrice_Change.
Wie in 4.1 zu erkennen steigt die Accuracy von Modell zu Modell, liegt aber immer unter 40%. Die Accuracy sagt aus, wie viele der Vorhersagen den tatsächlichen Werten entsprechen. Konkret:

$$accuracy = \frac{Anzahl\ korrekter\ Vorhersagen}{Gesamtanzahl\ Vorhersagen} \tag{4.1}$$

Eine solch niedrige Accuracy zeigt also, das die verwendeten Predictoren, nicht genau genug sind oder nicht ausreichen, um zuverlässig den tatsächlichen Wert zu berechnen. Dies wird auch durch den RMSE, dem Root Mean Squared Error, und dem R^2 gezeigt. Der hohe RMSE un der niedrige R^2 sagen aus, dass das Modell eben nicht gut an die tatsächlichen werte angepasst ist. An den Koeffizienten ist auch zu sehen, das jeder der Eingabevariablen nur einen geringen Einfluss auf die Zeilvariable hat (Gülden Kaya Uyanik, 2013).
Schlussendlich ist zu erkennen, das der Zusammenhang des Sentiments und der Aktienpreis Entwicklung unter den hier herrschenden Bedingungen nicht eindeutig nachzuweisen ist.

	intercept	coefficients	R^2	accuracy	rmse
senti_score	-0.01787722455488698	[0.11376727]	0.003929287192528164	0.3307323920402939	0.9323585033822364
senti_pos_score, senti_neg_score	-0.05387680752728273	[0.2320418 -0.08361624]	0.008009315107042392	0.37554702336718687	1.1330999288525196
with open	-0.05395554811535282	[2.32018857e-01 -8.35894390e-02 8.71147636e-07]	0.008013142949997332	0.37560895054083066	1.1330726020833533
with PublishCount	-0.04415531911622637	[0.23284092 -0.08631886 -0.0026853]	0.00819052319750202	0.3649987614565271	1.060557992234393

Abbildung 4.1: Vergleich von vier Linearen Regressionen

4.2 Bag-Of-Words und TF-IDF mit Random Forest Regression

Neben der wörterbuchbasierten Methode, betrachten wir in unserer zweiten Analyse den Datensatz auf Grundlage des Bag-Of-Words Modells. Hierbei wird ähnlich wie bei der Dictionarybasierten Methoden jede Headline als eine Reihe (eine Observation, ein Dokument), und jedes Wort als eine Spalte (eine Variable) betrachtet (Das and Chakraborty, 2018). Das Vorkommen eines Wortes in einer Headline wird durch den Wert in der entsprechenden Spalte ausgedrückt.
Ebenso wir für den dictionary-based Ansatz müssen dieselben Preprocessing Steps durchgeführt werden. Zur Vereinfachung verwenden wir hier den schon bereinigten Datensatz aus dem ersten Modell. Hier wurde wie oben beschrieben bereits eine Tokenization, ein POS, Lemmatization und Stopword Removal durchgeführt. Wir nehmen an, dass dieser Datensatz ausreichend bereinigt ist, und gewährleisten durch die VErwendung deselben Daten in beiden MOdellen eine höhere Vergleichbarkeit der Ergebnisse.

4.2.1 Model bauen

Für die Analyse des BOW-Modells nutzen wir wie bereits eingeleiette als Grundlage zur Berechnung die TF-IDF, die Term Frequency - Inverse Document Frequency. Im ersten Schritt berechnen wir die Term Frequency. Die Term Frequency gibt an, wie häufig ein Wort (term) in einer Headline (eine Observation) vorkommt. Das Ergebnis ist eine Tabelle, in der Die Häufigkeit eines Worts im jeweiligen Dokument eingtragen ist. Siehe 4.2 aus (Das and Chakraborty, 2018)

$$TF(Wort, Dokument) = \frac{Häufigkeit\, Wort \in Dokument}{Anzahl\, Wörter \in Dokument} \tag{4.2}$$

Der IDF Wert eines Wortes gibt an, wie häufig er im gesamten Korpus zu finden ist. Umso näher der IDF Wert bei 0 ist, desto häufiger kommt das Wort vor. Wenn das Wort also in vielen Dokumenten vorkommt, ist der Wertnahe an 0, wen das Wort näher an 1. Berechnen lässst sich das IDF Wert mithr an 1. Berechnen lässst sich der IDF Wrt folgendermaßen:

$$IDF(Wort) = \log_e(1 + \frac{Anzahl\, Dokumente}{Anzahl\, Dokumente\, mit\, Wort}) \tag{4.3}$$

Den TF-IDF Wert einees TF eertes und des IDF Wertes eines Wortes:

$$TF - IDF(Wort, Dokument) \; = \; TF(Wort, Dokument) * IDF(Wort) \qquad (4.4)$$

Bei der berechnung der TF und IDF Werte sei erwähnt, dass das sklearn Package, mit dem im weiteren Verlauf gearbeitet wird, zum Teil die zugrundeliegenden Formeln leicht anpasst und für die Anwendung optimiert hat. Siehe dazu Dokumentation sklearn TF-IDF Vectorizer und Dokumentation sklearn TFIDF Transformer Durch die Berechnung der TF-IDF Matrix haben wir nun einen numerischen Datensatz, aufgrund dessen sich verschiedene Analysen durchfrühren lassen. Wir wollen mit dem TF-IDF Rf nun untersuchen, ob wir die Analyseergebnisse gegenüber der Sentiment Analyse verbessern können. Dafür wenden wir auf den numerischen Werten eine Random Forest Regression an. Die Analyse von Stock News Headlines auf Basis von TF-IDF Werten und mit einer Random Forest Regression wurde in der Literatur bisher nicht behandelt. Wir wollen trotzdem aufgrund der Eigenschaften von Random Forest hier eben so eine Analyse durchführen.
Aufgrund der Größe unserer Datensatzes und der damit einhergehenden hohen Laufzeit der meisten Algroithmen müssen wir vor der Anwendung der Random Forest Regression einige Vorbereitungen durchführen.
Zuerst entfernen wir insbesondere aus Laufzeitgründen jene Spalter von Wörtern, die weniger als 100 mal oder mehr als 2000 mal im gesamten Korpus vorkommen. Damit verringert sich die Anzahl der Features um mehr als 6000. Anschließden führen wir auf einem zufällig gewähltem Sample ein Hyper Paramtertuning durch. Das Sample besteht aus 5000 zufällig ausgweählten Headlines des zugrunde liegenden Datensatzes. Für das Hyper Parameter Tuning und die anschließende Random Forest Regression nutzen wir nutzen wir das sklearn Package (Pedregosa et al., 2011). scikit-learn ist ein open source Tool für Machine Learning in Python, die verschiedne Regressions-, Klassifikations- und Clustering Algroithmen enthält.

In Tabelle 4.1 sind die Parameter gelistet, für die wir ein Tuning durchgeführt haben. In der Spalte M̈öglichkeitenb̈efinden sich die Werte, die getestet wurden. In der Spalte Ërgebnisïst der jeweils beste aus dem Tuning hervorgegangene Wert zu finden. Nun führen wir zur anschließenden Evaluierung des Modells einen Trainings-Test-Split durch, wobei das Trainingsset aus 80 Prozent der Observationen besteht. Das Modell wir mithilfe der fit Funktion des RandomForestRegressor mit den aus dem Hyper Parameter Tuning vorgeschlagenen Werten trainiert. Anschließend wird eine Predcition auf

Parameter	Möglichkeiten	Ergebnis
Anzahl an Decision Trees (n_estimators)	30, 37, 45, 53, 61, 68, 76, 84, 92, 100	37
Anzahl der Features, die bei jedem Split benötigt werden	'auto', 'sqrt'	'sqrt'
Maximale Tiefe jedes Desion Trees	10, 20, 30, 40, 50, Keine	10
Mindestanzahl an Observationen, die für einen weiteren Split gebraucht werden	2, 3	2
Mindestanzahl an Observationen, die in einem Leaf Knoten enthalten sein müssen	1, 2	2
Bootstrap als Auswahlmethode für jeden Decision Tree	ja, nein	ja

Tabelle 4.1: Hyper Paramter Tuning

dem Test set durchgeführt.

4.2.2 Modell Bewertung

Zur Bewertung des TF-IDF Random Forest Regression Models ziehen hier den Mean Absolute Error heran. Der Mean Absoult Error ist eine häufig eingesetze Metrik zur Bewertung von Regressions Modellen. Er bezieht aich auf das Testset und gibt an, wie weit im durschnitt die vorhergesagten Werte von den tatsächlichen realen Werten abweichen.

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n} \tag{4.5}$$

$$\tag{4.6}$$

In unserem TF-IDF Random Forest Regression Modell erhalten wir einen MEA von 3.92. Das bedeutet, dass der vorhergesagte Änderung des Stock Preises im durschnitt um 3.92 von dem realen Wert abweicht.

4.3 Auswertung

5 Evaluation

5.1 Ergebnissevaluierung im Hinblick auf die Geschäftsziele

Wie im Kapitel Business Understanding beschrieben, sind die Auswirkungen von Headlines sowohl für die betroffenen Unternehmen als auch (private) Aktienhändler interessant. Blicken wir zusätzlich auf die Forschungsfrage zurück, sind mehrere Ergebnisse zu evaluieren.
Grundsätzlich können wir mit unseren Ergebniss mit einer etwa 80-prozentigen Wahrscheinlichkeit anhand einer Headline bestimmen, ob sich der Aktienkurs positiv oder negativ entwickelt. Wir können jedoch nur mit deutlich geringerer Wahrscheinlichkeit die genaue Änderung des Aktienpreises vorhersagen.
Die Frage ist nun, ob es für die Geschäftsziele ausreicht, den Trend vorherzusagen. Für Anleger, die möglicherweise das Anlegen für andere übernehmen, wird die Analyse sehr wahrscheinlich nicht ausreichend sein. Eine begründete Entscheidung gegenüber dem Kunden für den Kauf einer Aktie kann möglicherweise nicht ausreichend aussagekräftig begründet werden. Für private Anleger, kann die schnelle Analysemöglichkeit jedoch zur Entscheidungsunterstütung herangezogen werden. Insbesondere unerfahrende Anleger können profitieren. Für Unternehmen, die die Auswirkung etwaiger Headlines einschätzen möchten, zeigt die Trendanalyse schnell und unkompliziert kritische Headlines, auf die mit unter anderem Marketing reagiert werden kann. Eine genaue Kalkulation des Aktienpreises ist jedoch nicht ohne weiteres möglich, konkrete Marktanalyse anhand der Headlines sind also nicht möglich. Außerdem zeigt der geringe Zusammenhang, des Sentiments, also der zugewiesenen Gefühle der Headline, zum Aktienpreis, dass letztendlich die Empfindung solcher Aktiennews subjektiv bleiben.
In Abgrenzung zur vorangegangen Literatur haben wir nur die Überschriften von Aktiennews betrachtet, mit einem offensichtlich schlechteren Ergebis als andere, die meist die vollständigen Artikel begutachteten. Daraus lässt sich ableiten, dass die Headline nicht ausreicht um eine fundierte Datenbasierte Entscheidung zutreffen, obwohl, wie zuvor literarisch gezeigt, Leser primär anhand der Headline agieren. Schlussendlich können mit unseren Analysen grobe Empfehlungen gemacht werden, für die risikobehaftete Wirtschaft allerdings sind diese noch nicht ausreichend.

5.2 Evaluierung des Prozesses

Wir wenden in diesem Projekt das CRISP-DM Modell an. Durch die klare Differenzierung der einzelnen Schritte entsteht ein einheitliches und strukturiertes Vorgehen. Wir haben uns bei den einzelnen Schritten an der ausführlichen Beschreibung des CRISP-DM Modells aus (Chapman et al., 2000) orientiert. Die einzelnen Schritte werden teils sehr ausführlich definiert, weshalb wir teilweise einige Abschnitte zusammengefasst haben oder ausließen. Im Gegensatz zu dem im CRISP-DM Modell vorgesehen Iterationen, haben wir nur sprunghaft Änderungen durchgeführt, die sich im Zusammenhang mit der Literaturrecherche und Empfehlungen einiger Kurse an der Universität ergeben haben.

5.3 Ausblick

Die beiden vorgestellten Ansätze zur Analyse von Headlines können noch deutlich intensiver einzeln betrachtet und weiter optimiert werden. Bei der Sentimentanalyse mittels eines Dictionary-Ansatzes, kann ein extra für den Finanzmarkt angepasstes Dictionary verwendet werden. Dieses müsste sich aber explizit auf die Aktiennews beziehen, und nicht auf Nachrichten aus der Allgemeinpresse. Für die darauf aufbauenden Predictions könnten unterschiedliche Algorithmen verwendet werden oder weitere Features mit einbezogen werden. Jedoch haben wir wie im Kapitel geschildert, feststellen müssen, dass ein wörterbuchbasierter Ansatz zur Vorhersage der Aktienpreisentwicklung eher ungeeignet ist. Der TF-IDF Ansatz mit Random Forest Regression schneidet deutlich besser ab. Hier lässt sich ein weitere Feintuning durchführen. Grundsätzlich könnten auch beide Ansätze kombiniert in einem Algorthimus betrachtet werden. Abschließend bleibt zu erwähnen, dass die Datenbasis der evaluierten Modelle stark eingeschränkt sind. Sowohl die Historie als auch die Genauigkeit der Uhrzeiten waren dem Modell, aufgrund äußerer Umstände, nicht gegeben. Rein spekulativ ist auch der verwendete Datensatz der Headlines nicht vollständig aussagekräftig gegenüber dem Aktienmarkt, da hier nur ein Publisher betrachtet wurde und die Formulierungen der künstlerischen Freiheit der Journalisten hierdurch nicht die Gesamtheit abbildet.

6 Deployment

Schlussendlich sollen die Ergebnisse des Modells den Unternehmen und den privat Anlegern zur Verfügung stehen. Es soll Möglich sein, die Vorhersage des Modells zur Entwicklung des Aktienpreises schnell und unkompliziert bei Eingabe einer Headline zu erhalten. Dies bietet sowohl den Unternehmen als auch den privat Anlegern den Vorteil eine objektive, datenbasierte Einschätzung der veröffentlichten Headline in ihre Entscheidungen am Markt einfließen zu lassen.

6.1 Strategie

Um eine möglichst einfache Handhabung zu gewährleisten, wird die Anwendung einfach gehalten: Eine Eingabe, eine Ausgabe. Um außerdem den Unternehmen die Möglichkeit zu bieten, die Vorhersage in deren Unternehmensprozesse einzubinden, wird eine serverbasierte Anwendung erstellt, die Ihren Dienst über eine Schnittstelle zur Verfügung stellt. Denkbar ist es, dass Unternehmen neue Aktiennews automatisch sammeln und stetig durch das Modell bewerten lassen. So erhalten Unternehmen schnelles Feedback zu negativen Nachrichten und können entsprechend reagieren.
Durch die einfache API ist es ebenfalls möglich ein Grafisches-User-Interface zu entwickeln und auf die API zu legen, so können auch die private Anleger durch die Lösung erreicht werden. Allerdings ist die Auslieferung und die Gestaltung der GUI nicht Teil dieser Studienarbeit.
Erreichbar die API durch einen HTTP-POST auf **/predict**. Übergeben Sie die Headlines als Liste von Strings per *POST, application/json* in folgender Form:

```
{
    'headlines': [
        'Dies ist eine Headline.',
        'Dies ist noch eine Headline!'
    ]
}
```

Die Antwort erfolgt ebenfalls per JSON in folgender Form:

```
{
    'prediction': [-0.627, 0.9965],
    'stockChange': [-1, 1]
}
```

6.2 Implementation

Nach den Schritten des Modelling, wird das trainierte Modell exportiert und statisch zur Verfügung gestellt. Ein einfacher Flask-Server basierend auf Python lädt dieses und stellt eine API zur Verfügung.
Wird nun eine oder mehrere Headlines an die Anwendung gesandt, so durch läuft jeder dieser Überschriften die selben Pre-Processing Schritte wie im vorangegangenen Kapitel beschrieben. Nach folgend werden die entsprechenden numerische Werte aus der Headline berechnet und in das Modell eingesetzt. Das Modell liefert eine Vorhersage. Diese wird, zusammen mit der standartisierten Vorhersage (-1, 0 oder 1) als Response auf die Anfrage verpackt und versendet.
Zum produktiven Einsatz, insbesondere der Skalierung, wird die Verwendung eines Anwendungsserver empfohlen. Hier wird uWSGI verwendet, in kombination mit einem Nginx Webserver. Die Tatsache, dass der Server in einem Docker-Container läuft, erleichtert die Auslieferung und die Skalierung. Dennoch ist es möglich, die Anwendung ebenfalls lokal auszuführen.

6.3 Überwachen und Pflegen

Um die stetige Weiterentwicklung des Modells zu ermöglichen, benötigt die Anwendung lediglich die exportierte Version des Modells, alle weiteren Schritte der API bleiben auch bei veränderten Modellen gleich. Dies ermöglicht das Austauschen und Verbessern des Modells. Außerdem ermöglicht die Tatsache der reinen containerbasierten Serveranwendung eine gute Skalierbarkeit. Das eigentliche Überwachen der Anwendung obliegt hier dem potenziellen Host.

Literaturverzeichnis

Agarwal, A., Sharma, V., Sikka, G., and Dhir, R. (2016). Opinion mining of news headlines using SentiWordNet. *Symposium on Colossal Data 2016*.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. `https://www.the-modeling-agency.com/crisp-dm.pdf`.

Das, B. and Chakraborty, S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation.

Developer, B. (2020). Daily Financial News for 6000+ Stocks. `https://www.kaggle.com/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests`.

Dor, D. (2003). On Newspaper Headlines as Relevance Optimizers. *Journal of Pragmatics 35*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, Vol. 56, No. 4*.

Gülden Kaya Uyanik, N. G. (2013). A study on multiple linear regression analysis. *4th International Conference on New Horizons in Education*.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*.

Mohammad, S. M. (2020). NRC Word-Emotion Association Lexicon. `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau,

D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Stojiljkovicr, M. (2019). Linear Regression in Python. `https://www.beoptimized.be/pdf/LinearRegressioninPython.pdf`.

Yahya Eru Cakra, B. D. T. (2015). Stock price prediction using linear regression based on sentiment analysis. *ICACSIS 2015*.

Zimmermann, V. (2019). A new way to sentiment-tag financial news. `https://towardsdatascience.com/a-new-way-to-sentiment-tag-financial-news-9ac7681836a7`.

Årup Nielsen, F. (2015). AFINN-en-165. `https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt`.