



Duale Hochschule Baden-Württemberg Mannheim

Bericht

Data Whispers – Stock Price Prediction

Studiengang Wirtschaftsinformatik

Studienrichtung Data Science

Verfasser:	Michael Greif, German Paul, Finn Münstermann, Hendrik Träber
Matrikelnummer:	5658606, 9973152, 3071508, 6367227
Kurs:	WWI22DSB
Studiengangsleiter :	Prof. Dr. Bernhard Drabant
Dozent:	Marc Benjamin Jung
Bearbeitungszeitraum:	12.12.2023 - 09.02.2024

Inhaltsverzeichnis

1. Unser Produkt.....	3
2. Modelle zum Projekt	5
2.1. Aktivitätsdiagramm	5
2.2. Ereignisliste	6
2.3. Ereignis-Reaktions-Modelle	7
2.4. Kombinierte SWOT-Analyse	7
2.5. Kontextdiagramm.....	8
3. Rahmen des Projektes	9
4. Technische Anforderungen und Entwicklungsüberblick.....	10
4.1. Webanwendungs-Framework: Streamlit	10
4.2. Datenverarbeitung und -visualisierung.....	10
4.3. Text- und Datenextraktion	11
4.4. E-Mail-Integration und Bildverarbeitung	11
4.5. Sicherheit und Benutzerauthentifizierung.....	11
4.6. Allgemeine Programmierkenntnisse.....	11
4.7. Fazit	12
5. Machine Learning Modelle.....	12
5.1. BERTopic.....	12
5.2. Doc2Vec (self-trained).....	13
5.3. Sentence-BERT (pretrained).....	13
5.4. GloVe (pretrained)	14
5.5. WordNet.....	15
6. Kritische Reflektion	15
7. Zusammenfassung und Nutzungserklärung	17
8. Punkteverteilung.....	20

1. Unser Produkt

Im Rahmen der Präsentation des Projekts "Stock Price Prediction" der DataWhispers AG wurde ein innovatives Produkt vorgestellt, das auf der Analyse und Vorhersage von Aktienkursen basiert. Das Konzept des Produkts ist inspiriert von der reichen Mythologie des antiken Griechenlands, wobei der Fokus auf Wissen, Geduld und Vertrauen als Kernprinzipien für erfolgreiche Investitionen liegt.

Erstens, das Prinzip "Wissen ist Macht" spiegelt die Bedeutung einer gründlichen Analyse und Forschung wider. DataWhispers AG betont die Rolle innovativer Datenanalysemethoden und modernster Technologien zur Generierung präziser Marktvorhersagen. Dieser Ansatz unterstreicht die Notwendigkeit, umfassend informiert zu sein, um fundierte Entscheidungen im Finanzmarkt treffen zu können.

Zweitens, das Prinzip "Geduld ist eine Tugend" hebt hervor, dass erfolgreiche Investitionen eine langfristige Perspektive erfordern. Das Unternehmen richtet seinen Fokus auf die Identifikation und Unterstützung von langfristigen Trends, anstatt kurzfristige Gewinne zu priorisieren. Dieses Prinzip betont die Bedeutung von Nachhaltigkeit und Stabilität in der Investitionsstrategie.

Drittens, das Prinzip "Vertrauen ist unverzichtbar" bezieht sich auf die Wichtigkeit von Transparenz und offener Kommunikation zur Stärkung des Kundenvertrauens. DataWhispers AG verpflichtet sich zu einer Arbeitsweise, die maximale Zuverlässigkeit und Verantwortung gegenüber ihren Kunden gewährleistet, um eine vertrauensvolle Beziehung aufzubauen und zu erhalten.

Die Einladung an Investoren, sich auf die innovativen Lösungen der DataWhispers AG einzulassen, symbolisiert nicht nur den Zugang zu fortschrittlichen Analysewerkzeugen, sondern auch eine Partnerschaft, die auf tiefem Verständnis und Vertrauen basiert. Michael Greif, ein Vertreter des Unternehmens, präsentierte zusammen mit einer weiteren Schlüsselperson das Produktportfolio und unterstrich die Bedeutung der angebotenen Dienstleistungen für die Navigation und das erfolgreiche Agieren auf den volatilen Finanzmärkten.

Zusammenfassend präsentiert die DataWhispers AG mit ihrem Produkt "Stock Price Prediction" eine Synthese aus antiker Weisheit und moderner Finanztechnologie. Durch die Anwendung der drei grundlegenden Prinzipien strebt das Unternehmen danach, seinen Kunden nicht nur Werkzeuge zur Vorhersage von Marktbewegungen zu bieten, sondern auch eine Philosophie der Investition, die Wissen, Geduld und Vertrauen in den Vordergrund stellt. Diese Elemente bilden zusammen eine solide Grundlage für die Entscheidungsfindung in der komplexen Welt der Finanzmärkte.

2. Modelle zum Projekt

2.1. Aktivitätsdiagramm

Das Aktivitätsdiagramm illustriert den systematischen Ablauf, den ein Nutzer bei der Interaktion mit der Website der Data Whispers AG durchläuft. Es beginnt mit dem Aufrufen der Website und führt den Betrachter durch eine logische Abfolge von Entscheidungen und Aktionen. Das Diagramm verdeutlicht, dass bei erfolgreicher Verbindung verschiedene Prozesse wie Website-Anzeige, Web-Scraping, Aktienvorhersage und die Darstellung zugehöriger Nachrichtenartikel stattfinden. Bei Fehlern werden entsprechende Benachrichtigungen ausgegeben. Dieses Diagramm ist ein Schlüsselwerkzeug, um den Entwicklern und Projektbeteiligten zu helfen, den Workflow zu verstehen und sicherzustellen, dass alle Eventualitäten im Benutzererlebnis berücksichtigt werden.

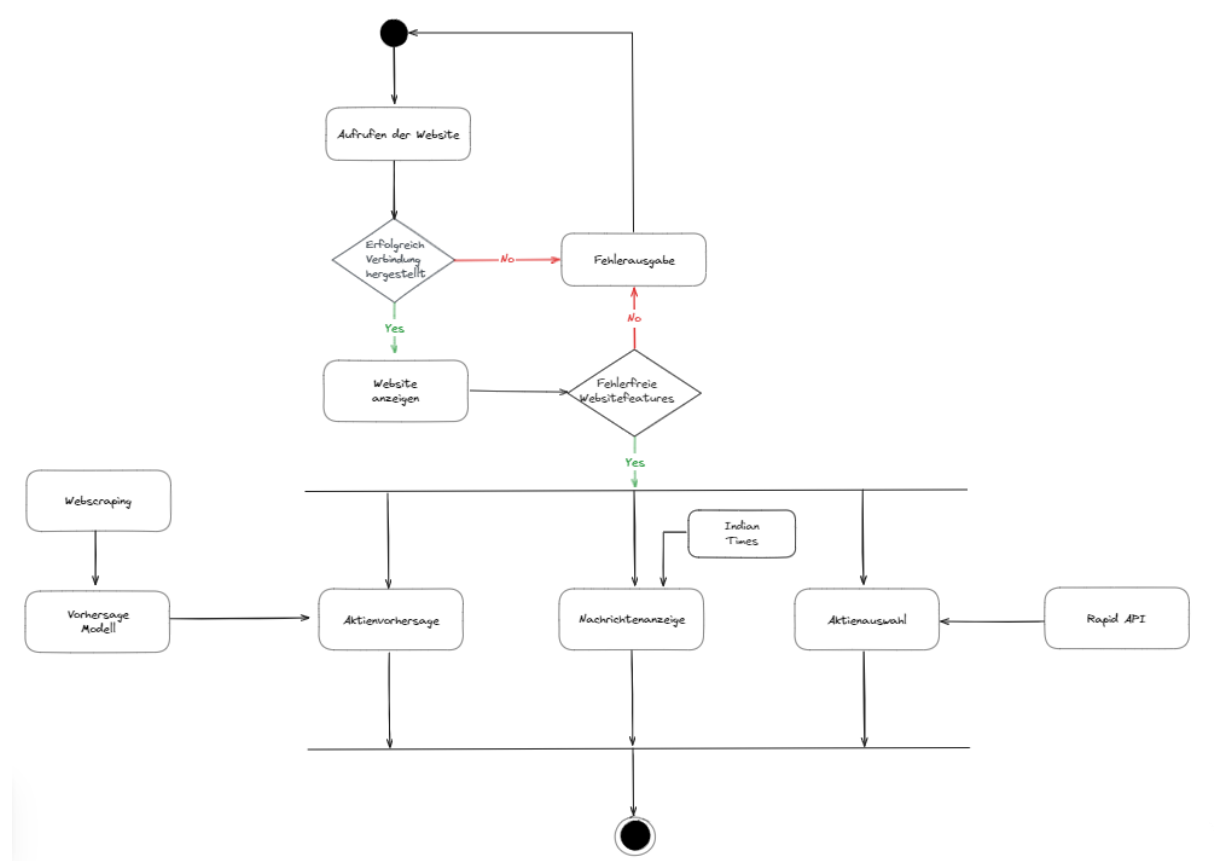


Abb. 1: Aktivitätsdiagramm

2.2.Ereignisliste

Die Ereignisliste dokumentiert detailliert die Ereignisse, die während der Benutzerinteraktion auftreten können. Sie verzeichnet sowohl die Eingaben des Benutzers als auch die Ausgaben des Systems. Zum Beispiel, wenn ein Kunde eine Vorhersage für den Dow Jones Index anfordert, reagiert das System, indem es die Website öffnet und die Vorhersage anzeigt. Sollte dabei ein Fehler auftreten, wird ein Fehlertext ausgegeben. Diese Liste ist nicht nur ein Instrument zur Fehlerbehebung, sondern dient auch dazu, die Benutzererfahrung zu optimieren und sicherzustellen, dass das System intuitiv auf Benutzeranfragen reagiert.

Ereignisname	Zugehörige Datenflüsse
Kunde will Dow Jones Vorhersage	1) Öffnen der Website (I) 2) Vorhersage wird angezeigt (O) 3) Ausgabe eines Fehlertextes (O)
Kunde will angezeigter Zeitraum ändern	4) Einstellen des Zeitraums in Fenster (I) 5) Anpassung des Zeitraumes (O) 6) Ausgabe einer Fehlermeldung (O)
Kunde will zugehörige Newsartikel ansehen	7) Anklicken der verlinkten News (I) 8) Weiterleitung an Newsartikel (O) 9) Fehlermeldung (O)
Kunde will Index und Wert wissen	10) Kunde hovort über angezeigten Graph (I) 11) Anzeige des Datums und Wertes (O) 12) Keine Ausgabe von Werten (O)
Kunde will verwandte Aktienkurse sehen	13) Öffnen der Website (I) 14) Automatisch aktualisierte Werte (O) 15) Fehlermeldung

Abb. 2: Ereignisliste

2.3. Ereignis-Reaktions-Modelle

Die Ereignis-Reaktions-Modelle zeigen die direkten Auswirkungen der Benutzerinteraktionen auf das System. Diese Modelle bilden eine Brücke zwischen Benutzeraktionen wie dem Anklicken von verlinkten Nachrichtenartikeln und den Systemreaktionen wie dem Anzeigen von Artikeln oder dem Ausgeben von Fehlermeldungen. Die Visualisierung dieser Interaktionen ist entscheidend für die Entwicklung einer benutzerfreundlichen Oberfläche und stellt sicher, dass das System so reagiert, wie es die Endbenutzer erwarten.

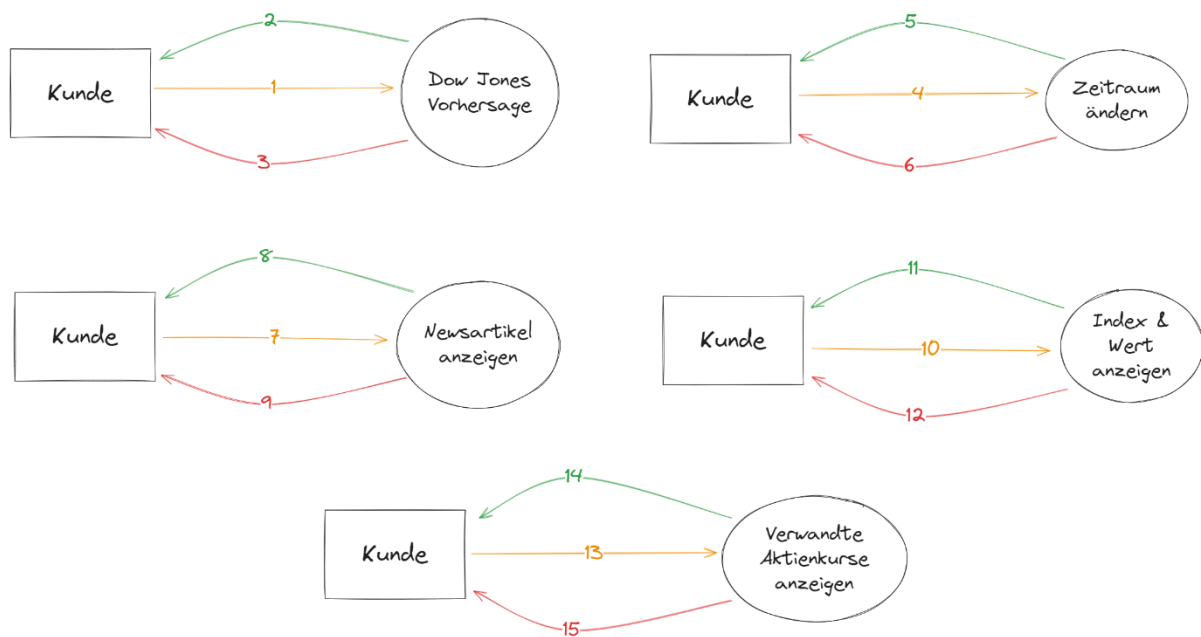


Abb. 3: Ereignis-Reaktions-Modelle

2.4. Kombinierte SWOT-Analyse

Die kombinierte SWOT-Analyse ist ein strategisches Werkzeug, das genutzt wurde, um die Stärken, Schwächen, Chancen und Risiken des Vorhersagemodells für den Aktienmarkt zu bewerten. Diese Analyse hebt hervor, dass die Nutzung von maschinellem Lernen und die Analyse von Zeitungsartikeln und Aktienkursen die Genauigkeit der Vorhersagen erhöhen und somit einen Vorteil für Investoren und Finanzanalysten darstellen. Die SWOT-Analyse unterstreicht die Notwendigkeit, sich auf die Stärken des Modells zu konzentrieren und gleichzeitig eine Strategie zu entwickeln, um Schwächen anzugehen und Risiken zu minimieren.

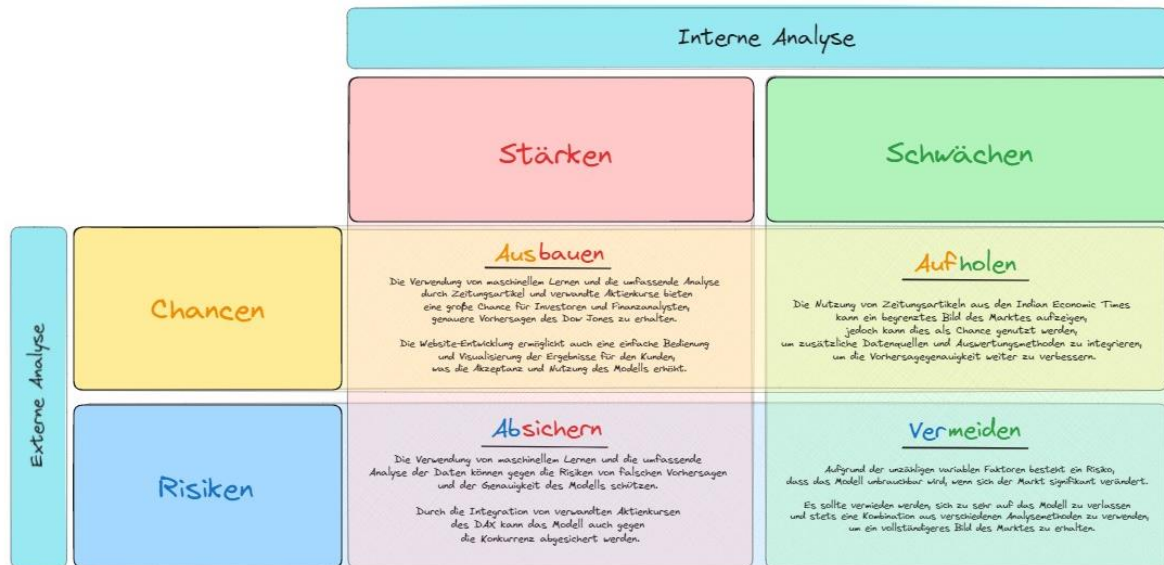


Abb. 4: Kombinierte SWOT-Analyse

2.5. Kontextdiagramm

Das Kontextdiagramm bietet eine ganzheitliche Ansicht des Systems und seiner Interaktionen mit externen Akteuren. In diesem Fall illustriert das Diagramm, wie der Kunde mit der Website der Data Whispers AG interagiert. Es veranschaulicht die vielfältigen Datenflüsse zwischen dem Kunden und dem System und zeigt auf, wie Eingaben des Benutzers zu entsprechenden Systemreaktionen führen. Dieses Diagramm ist entscheidend, um zu verstehen, wie externe Faktoren das System beeinflussen, und stellt sicher, dass das Systemdesign alle Benutzeranforderungen erfüllt.

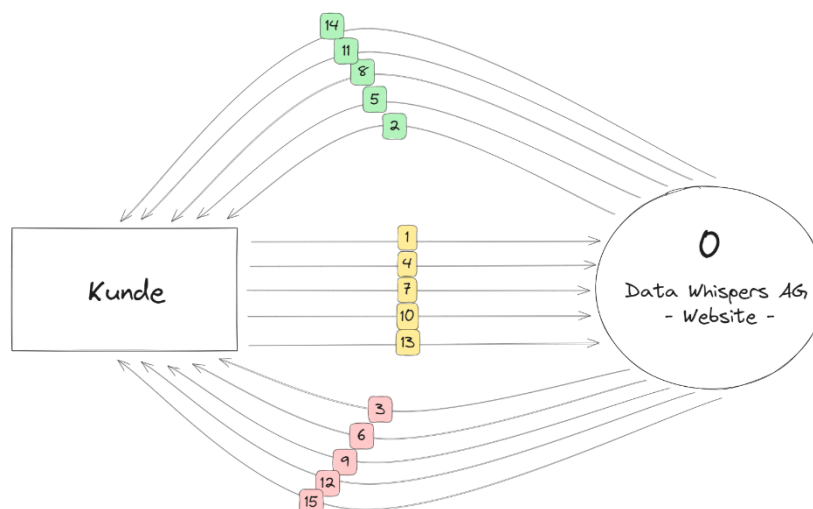


Abb. 5: Kontextdiagramm

3. Rahmen des Projektes

Im Rahmen der fortwährenden Bestrebungen, innovative Lösungen im Bereich der Finanztechnologie zu entwickeln, hat sich ein bemerkenswertes Konsortium zusammengefunden, um die Möglichkeiten der künstlichen Intelligenz für die Vorhersage von Aktienmarktrends zu erforschen. Diese Zusammenarbeit, bestehend aus der Data Whispers AG, Marc Benjamin Jung und der Five Head AG, zielt darauf ab, ein hochmodernes Machine Learning (ML)-Modell zu entwickeln. Dieses Modell soll in der Lage sein, die zukünftige Performance des Dow Jones Index vorauszusagen, basierend auf einer Analyse historischer Aktienkurse sowie aktueller Nachrichtenartikel aus den Indian Economic Times.

Das Projekt, offiziell betitelt als "Stock-Price-Prediction", verspricht, eine neue Ära der Aktienmarktanalyse einzuläuten. Die Data Whispers AG übernimmt dabei die technische und fachliche Leitung des Projekts, verpflichtet sich zur eigenständigen Durchführung und gewährleistet die Bereitstellung einer funktionsfähigen Prediction-Engine bis zum festgelegten Fertigstellungsdatum am 23. Januar 2024. Im Gegenzug verpflichtet sich Marc Benjamin Jung, die Ergebnisse des Projekts anhand einer vordefinierten Bewertungsskala zu bewerten, als Form des Arbeitsaufwandsersatzes. Die Five Head AG unterstützt das Projektteam durch seelischen Beistand bei eventuell auftretenden personellen und persönlichen Unannehmlichkeiten.

Ein wesentlicher Bestandteil des SLA sind die festgelegten Service-Level-Kennzahlen, die den Erfolg des Projekts quantifizieren. Diese Kennzahlen orientieren sich an der prozentualen Abweichung der Modellvorhersagen im Vergleich zu den tatsächlichen Werten des Dow Jones, gemessen durch den RSME-Score. Zudem sind spezifische Meilensteine definiert, die bis zu vorgegebenen Zeitpunkten erreicht werden sollen, beginnend mit dem ausgearbeiteten Systementwurf bis hin zur Präsentationserstellung und Marketingstrategie.

Um die Einhaltung der Projektziele zu überwachen, ist ein regelmäßiges Monitoring und Reporting vorgesehen. Bei terminlich gesteckten Zusammenkünften präsentiert der Dienstleister den aktuellen Ist-Zustand des Projekts im Vergleich zum Soll-Zustand. Für den Fall, dass die Projektergebnisse hinter den Erwartungen zurückbleiben, sieht das Eskalationsmanagement die Hinzunahme einer höheren Instanz vor, um eine Lösung zu finden.

Die Laufzeit dieses Vertrags erstreckt sich vom 27. November 2023 bis zum 23. Januar 2024, eine Zeitspanne, die sowohl die Entwicklung als auch die abschließende Bewertung des Projekts umfasst. Dieses SLA markiert einen bedeutenden Schritt vorwärts in der Anwendung von ML-Technologien im Finanzsektor und setzt neue Maßstäbe für die Zusammenarbeit zwischen technologischen Innovatoren und Finanzexperten.

4. Technische Anforderungen und Entwicklungsüberblick

In der heutigen digitalen Ära ist die Entwicklung von Webanwendungen, die sowohl leistungsfähig als auch benutzerfreundlich sind, von entscheidender Bedeutung. Diese Dokumentation zielt darauf ab, einen umfassenden Überblick über die technischen Anforderungen und die damit verbundenen Technologien zu geben, die für die Entwicklung und Wartung der Anwendung benötigt werden.

4.1. Webanwendungs-Framework: Streamlit

Streamlit ist ein innovatives Framework, das speziell für die Entwicklung von Datenwissenschafts- und ML-Anwendungen konzipiert wurde. Es ermöglicht Entwicklern, mit wenigen Codezeilen interaktive Webanwendungen zu erstellen. Die Verwendung von Streamlit in unseren Dateien zeigt, wie effizient und schnell Prototypen und vollwertige Anwendungen erstellt werden können. Entwickler müssen mit den Grundlagen von Streamlit vertraut sein, einschließlich der Einrichtung von Seitenlayouts, der Integration von Widgets und der Verwaltung von Benutzerinteraktionen.

4.2. Datenverarbeitung und-visualisierung

Die Fähigkeit, Daten effektiv zu verarbeiten und zu visualisieren, ist ein Kernaspekt moderner Webanwendungen. Mit Bibliotheken wie Pandas und NumPy können komplexe Datenmanipulationen und -analysen mit minimaler Anstrengung durchgeführt werden. Plotly erweitert diese Fähigkeiten, indem es leistungsstarke Visualisierungswerkzeuge bereitstellt, die interaktive Graphen und Charts für Webanwendungen ermöglichen. Ein tiefes Verständnis dieser Bibliotheken und ihrer Anwendung in Echtzeit-Datenvisualisierungen ist für die Entwicklung nützlicher und ansprechender Anwendungen unerlässlich.

4.3. Text- und Datenextraktion

Die Extraktion von Text und Daten aus verschiedenen Quellen ist eine häufige Anforderung in vielen Projekten. Die Nutzung von NLTK für die Textanalyse ermöglicht es Entwicklern, Einblicke in Textdaten zu gewinnen, indem sie Aufgaben wie Tokenisierung, Tagging und Klassifizierung durchführen. Selenium, ein Tool für automatisierte Webtests, ist ebenfalls entscheidend, wenn dynamische Inhalte von Webseiten extrahiert werden müssen. Kenntnisse in diesen Bereichen sind für die Implementierung von Features wie Artikel-Extraktion und automatisiertem Web-Scraping von großer Bedeutung.

4.4. E-Mail-Integration und Bildverarbeitung

Die Integration von E-Mail-Funktionalitäten und die Verarbeitung von Bildern sind wichtige Komponenten, die zur Verbesserung der Benutzererfahrung beitragen. Durch die Verwendung von Yagmail wird der Prozess des E-Mail-Versands vereinfacht, was besonders nützlich ist, um mit Benutzern zu kommunizieren oder Benachrichtigungen zu senden. Die Python Imaging Library (PIL) bietet umfangreiche Möglichkeiten zur Bildbearbeitung, die für die Gestaltung von Benutzeroberflächen und die Präsentation von Inhalten verwendet werden können.

4.5. Sicherheit und Benutzerauthentifizierung

Die Sicherheit der Anwendung und der Schutz der Benutzerdaten sind von höchster Wichtigkeit. Die Implementierung von Authentifizierungsmechanismen, wie sie in den Dateien für Premium-Kunden vorgesehen sind, stellt sicher, dass nur berechtigte Benutzer Zugang zu bestimmten Funktionen haben. Entwickler müssen mit den Best Practices für Sicherheit und Authentifizierung vertraut sein, einschließlich der sicheren Speicherung von Passwörtern und der Verwaltung von Benutzersitzungen.

4.6. Allgemeine Programmierkenntnisse

Neben den spezifisch benutzten Technologien und Frameworks sind allgemeine Programmierkenntnisse in Python unerlässlich für die effektive Arbeit an diesem Projekt. Ein solides Verständnis der Programmierprinzipien, Code-Organisation und Softwareentwicklungsmethoden ist erforderlich, um wartbaren und erweiterbaren Code zu schreiben.

4.7.Fazit

Die Entwicklung und Wartung der beschriebenen Webanwendung erfordert eine breite Palette von Fähigkeiten und Kenntnissen in verschiedenen technologischen Bereichen. Von der Webentwicklung mit Streamlit über die Datenverarbeitung und -visualisierung bis hin zur Textextraktion und Sicherheit – jedes Element spielt eine entscheidende Rolle in der Gesamtarchitektur des Projekts. Durch das Beherrschen dieser Technologien können Entwickler robuste, interaktive und benutzerfreundliche Anwendungen erstellen, die den Anforderungen moderner Benutzer entsprechen.

5. Machine Learning Modelle

Im Rahmen des "DataWhispers Stock-Price-Prediction" Projekts wurden verschiedene maschinelle Lernmodelle verwendet, um aus Nachrichtenartikeln Vorhersagen über Aktienpreise zu treffen. Hier ist eine Zusammenfassung der genutzten Modelle und deren Funktionsweise, sowie eine Erweiterung des hier aufgeführten Wissens mit Verweisen zu verwandten Wissenschaftlichen Artikeln:

5.1.BERTopic

BERTopic ist ein unsupervised Machine Learning Modell, das darauf spezialisiert ist, aus großen Mengen an Textdaten Themen zu generieren. Es nutzt dabei Techniken aus dem Bereich des Natural Language Processing (NLP), um Dokumente basierend auf ihrer inhaltlichen Ähnlichkeit zu clustern. Zunächst werden Document Embeddings erstellt, die die Texte in einem hochdimensionalen Vektorraum darstellen. Anschließend wird mittels Dimensionalitätsreduktion und dem Clustering-Algorithmus HDBSCAN eine Gruppierung ähnlicher Dokumente vorgenommen. Die Ähnlichkeit zwischen den einzelnen Embeddings wird berechnet, um thematische Cluster zu identifizieren.

Wissenschaftlicher Artikel: <https://arxiv.org/pdf/1908.10084.pdf>

Sentence-BERT (SBERT) verbessert BERT für effizientes Satz-Embedding, um schnelle Ähnlichkeitssuchen und Clusterbildung zu ermöglichen. Durch den Einsatz von Siamese- und

Triplet-Netzwerkstrukturen generiert SBERT Einbettungen, die semantische Bedeutungen beibehalten und mit der Kosinusähnlichkeit verglichen werden können. Diese Anpassung reduziert die Berechnungszeit von Stunden auf Sekunden für Aufgaben wie semantische Suche, im Vergleich zu herkömmlichen BERT-Ansätzen. SBERT zeigt eine überlegene Leistung bei semantischer Textähnlichkeit und anderen NLP-Aufgaben und bietet eine praktische Lösung für Anwendungen, die ein semantisches Verständnis von Sätzen erfordern.

5.2.Doc2Vec (self-trained)

Doc2Vec ist ein Algorithmus, der es ermöglicht, ganze Dokumente, einschließlich Sätzen und Absätzen, in Vektoren umzuwandeln. Im Gegensatz zu früheren Ansätzen wie Word2Vec, der lediglich Wörter in Vektoren umwandelt, berücksichtigt Doc2Vec den Kontext des gesamten Dokuments. Dies ermöglicht es, die Bedeutung eines Textes in einem mehrdimensionalen Raum darzustellen. Für dieses Projekt wurde Doc2Vec verwendet, um Nachrichtenartikel zu vektorisieren und anschließend die Kosinusähnlichkeit zwischen diesen Vektoren und spezifischen Features zu berechnen.

Wissenschaftlicher Artikel:

<https://proceedings.mlr.press/v32/le14.html?ref=https://githubhelp.com>

Der Artikel von Quoc Le und Tomas Mikolov beschreibt einen Algorithmus zum Erlernen von Vektorrepräsentationen für Sätze und Dokumente, der Bag-of-Words-Modelle in Bezug auf Wortreihenfolge und Semantik übertrifft. Dieser Ansatz zeigt bei Textklassifizierung und Sentimentanalyse bessere Ergebnisse als herkömmliche Methoden.

5.3.Sentence-BERT (pretrained)

Ist eine Modifikation des bekannten BERT-Modells (Bidirectional Encoder Representations from Transformers), die darauf abzielt, die Berechnung semantisch sinnvoller Satz-Embeddings zu beschleunigen. SBERT passt BERT an, um effizienter mit Satzpaaren umzugehen, und ermöglicht es, die semantische Ähnlichkeit zwischen Sätzen direkt zu messen. Dies geschieht durch das Training auf Basis von NLP-Aufgaben, wie z.B. semantische

Textähnlichkeit oder Textklassifizierung, um Embeddings zu generieren, die sich gut für direkte Vergleiche eignen.

Wissenschaftlicher Artikel: <https://arxiv.org/pdf/2203.05794.pdf>

BERTopic ist ein neuartiges, auf maschinellem Lernen basierendes Modell zur Themenmodellierung, das eine klassenbasierte TF-IDF-Prozedur nutzt, um die Schwächen herkömmlicher Bag-of-Words-Modelle zu überwinden. Es verwendet vortrainierte transformer-basierte Sprachmodelle zur Erzeugung von Dokumenten-Embeddings, gruppiert diese und extrahiert dann Themenrepräsentationen. BERTopic ist in der Lage, kohärente Themen zu generieren und liefert neue Spitzenwerte in verschiedenen Benchmarks zur Textklassifizierung und Sentimentanalyse.

5.4. GloVe (pretrained)

GloVe (Global Vectors for Word Representation) ist ein Ansatz für die Vektorrepräsentation von Wörtern, der auf der Co-Occurrence (dem gemeinsamen Auftreten) von Wörtern in einem Korpus basiert. GloVe kombiniert die Vorteile von Wortkontextmatrizen mit den von Word2Vec eingeführten Techniken für effiziente Lernverfahren. Das Ziel ist es, Wörter so in einen Vektorraum einzubetten, dass die Relationen zwischen den Wörtern erhalten bleiben und semantische sowie syntaktische Ähnlichkeiten abgebildet werden können.

Wissenschaftlicher Artikel: <https://aclanthology.org/D14-1162.pdf>

GloVe (Global Vectors for Word Representation) von Pennington, Socher und Manning ist ein Modell zur Erstellung von Wortvektoren, das feingranulare semantische und syntaktische Regularitäten in Vektorräumen durch Vektorarithmetik erfasst. Das Modell kombiniert globale Matrixfaktorisierung mit lokalen Kontextmethoden und trainiert effizient auf Nicht-Null-Elementen einer Wort-Wort-Kookkurrenzmatrix. GloVe übertrifft bestehende Modelle in Wortanalogie-Aufgaben, Wortähnlichkeitstests und Named Entity Recognition, indem es eine sinnvolle Struktur im Vektorraum schafft

5.5. WordNet

WordNet ist keine direkte Machine-Learning-Technik, sondern eine große lexikalische Datenbank der englischen Sprache, die Nomen, Verben, Adjektive und Adverbien als Knoten in einem Netzwerk organisiert. Synonyme werden in Synsets gruppiert, wobei die Beziehungen zwischen diesen Synsets genutzt werden, um semantische Ähnlichkeiten zwischen Wörtern zu finden. Im Kontext des Projekts wurde WordNet möglicherweise verwendet, um die Bedeutung von Wörtern zu verstehen und die semantische Nähe zwischen verschiedenen Begriffen zu ermitteln.

Diese Modelle wurden in Kombination verwendet, um aus den Inhalten von Nachrichtenartikeln Vorhersagen über die zukünftige Entwicklung von Aktienkursen zu treffen, wobei jedes Modell spezifische Aspekte der Textdaten verarbeitet und zur Gesamtleistung des Vorhersagesystems beiträgt.

Wissenschaftlicher Artikel: <https://www.d.umn.edu/~tpederse/Pubs/AAAI04PedersenT.pdf>

Der Artikel "WordNet::Similarity - Measuring the Relatedness of Concepts" von Ted Pedersen und anderen beschreibt ein Softwarepaket, das die semantische Ähnlichkeit oder Verwandtschaft zwischen Konzeptpaaren mittels WordNet misst. Es bietet sechs Ähnlichkeits- und drei Verwandtschaftsmaße, die auf lexikalischen Daten basieren. Diese Maße nutzen Informationen wie Pfadlängen zwischen Konzepten und Informationsgehalt. WordNet::Similarity wird durch Perl-Module implementiert, unterstützt hypothetische Wurzelknoten für umfassende Messungen und kann über eine Befehlszeilenschnittstelle oder Webinterface genutzt werden.

6. Kritische Reflektion

Eine kritische Reflexion über die Vorhersage von Aktienkursen, insbesondere unter Berücksichtigung der Abhängigkeit von Nachrichtenartikeln, dem Sprach- und Schreibstil der Autoren sowie dem Fehlen von Insiderinformationen, offenbart verschiedene

Herausforderungen und Einschränkungen, die bei der Entwicklung und Anwendung von Modellen zur Aktienpreisvorhersage berücksichtigt werden müssen.

Abhängigkeit von Nachrichtenartikeln

Die Nutzung von Nachrichtenartikeln als Datenquelle für die Vorhersage von Aktienkursen birgt ein fundamentales Risiko: die Qualität und Verlässlichkeit der Informationen. Nachrichten sind häufig subjektiv und können durch den Ton, die Auswahl der Themen oder die Perspektive des Autors verzerrt sein. Die Abhängigkeit von solchen zweifelhaften Quellen kann zu fehlerhaften Eingabedaten führen, die die Genauigkeit der Vorhersagemodelle erheblich beeinträchtigen. Zudem spiegeln Nachrichtenartikel oft bereits öffentlich zugängliche Informationen wider, wodurch der Zeitvorteil, den präzise Vorhersagen bieten sollen, minimiert wird. Die Herausforderung besteht darin, Algorithmen zu entwickeln, die die Verlässlichkeit von Nachrichtenquellen bewerten und die Auswirkungen von Nachrichten auf die Aktienpreise unter Berücksichtigung ihrer Glaubwürdigkeit und ihres Einflusses einbeziehen.

Sprach- und Schreibstil der Autoren

Der Sprach- und Schreibstil von Nachrichtenartikeln variiert erheblich zwischen verschiedenen Autoren und Publikationen. Diese Variabilität führt zu einer weiteren Ebene der Komplexität bei der Analyse von Textdaten. Ironie, Subtilität und kulturelle Kontexte können von maschinellen Lernmodellen schwer zu erfassen sein, insbesondere wenn diese Modelle nicht speziell darauf trainiert wurden, solche Nuancen zu erkennen. Die unterschiedlichen Stile können die Extraktion von Stimmungen oder die Bestimmung der Relevanz von Informationen für die Aktienpreisvorhersage erschweren. Eine kritische Betrachtung muss daher auch innovative Ansätze in der Verarbeitung natürlicher Sprache und im maschinellen Lernen berücksichtigen, die in der Lage sind, diese Feinheiten effektiv zu interpretieren.

Fehlen von Insiderinformationen

Ein grundlegendes Problem bei der Vorhersage von Aktienkursen auf der Grundlage öffentlich zugänglicher Informationen, einschließlich Nachrichtenartikeln, ist das Fehlen von

Insiderinformationen. Solche Informationen, die nicht öffentlich bekannt sind, können erhebliche Auswirkungen auf die Aktienkurse haben, sobald sie bekannt werden. Modelle, die ausschließlich auf öffentlichen Daten basieren, können daher nie die vollständige Bandbreite an Faktoren erfassen, die den Markt beeinflussen. Dies begrenzt ihre Vorhersagekraft und unterstreicht die Bedeutung einer diversifizierten Strategie, die verschiedene Datenquellen und Analysemethoden kombiniert, um die Genauigkeit der Vorhersagen zu verbessern.

Fazit

Die Reflexion über die genannten Punkte zeigt, dass die Vorhersage von Aktienkursen eine komplexe Aufgabe ist, die von der Qualität der verwendeten Daten, der Fähigkeit, sprachliche Nuancen zu interpretieren, und dem Zugang zu einer breiten Palette von Informationen, einschließlich nicht öffentlicher Daten, abhängt. Es bedarf kontinuierlicher Forschung und Entwicklung, um Modelle zu verbessern, die diese Herausforderungen überwinden können. Darüber hinaus ist es wichtig, dass Anwender solcher Vorhersagemodelle sich der inhärenten Unsicherheiten bewusst sind und Vorhersagen als einen von vielen Faktoren in einem umfassenderen Entscheidungsfindungsprozess betrachten.

7. Zusammenfassung und Nutzungserklärung

Der Dow Jones Predictor ist ein innovatives Werkzeug, konzipiert zur Analyse historischer Daten des Dow Jones Industrial Average (DJIA) und zur Prognose seiner zukünftigen Entwicklungen mittels fortschrittlicher Algorithmen des maschinellen Lernens. Dieses Projekt stellt einen integralen Bestandteil unseres Studiums an der Dualen Hochschule Baden-Württemberg Mannheim dar und veranschaulicht die praktische Anwendung unserer erworbenen Fachkenntnisse im Bereich der Datenwissenschaft sowie der Webentwicklung.

Kernfunktionen des Dow Jones Predictors:

Umfangreiche Datenanalyse:

Das Fundament des Predictors bildet eine tiefgreifende Analyse und Verarbeitung der Daten mittels Python. Diese ermöglicht es, komplexe Muster und Trends innerhalb des historischen Datenbestands des DJIA zu identifizieren und zu analysieren.

Einsatz von maschinellem Lernen:

Herzstück des Projekts ist die Implementierung hochentwickelter Modelle des maschinellen Lernens, die darauf abzielen, die zukünftige Bewegung des Dow Jones-Indexes mit hoher Präzision vorherzusagen. Diese Modelle werden stetig trainiert und verbessert, um die Genauigkeit der Vorhersagen zu optimieren.

Interaktive Webanwendung:

Durch die Entwicklung einer benutzerfreundlichen, auf Streamlit basierenden Webanwendung wird den Nutzern eine Plattform geboten, auf der sie interaktiv mit dem Predictor agieren können. Diese Schnittstelle präsentiert grafische Darstellungen der Daten und Prognosen, die es den Nutzern erleichtern, die Informationen zu verstehen und zu interpretieren.

Benutzerfreundliches Design:


Das Design der Webanwendung zeichnet sich durch seine intuitive Navigation und Benutzerführung aus, wodurch der Zugang sowohl für technisch versierte Anwender als auch für Einsteiger ohne technischen Hintergrund gewährleistet wird.

Installation und Inbetriebnahme:

Um den Dow Jones Predictor auf einem lokalen System zu installieren, sind folgende Schritte erforderlich:

1. **Klonen des GitHub-Repositorys:** Zunächst sollte das Repository mittels Git geklont werden: `git clone https://github.com/GermanPaul12/DataWhispers-Stock-Price-Prediction-Projekt-DHBW.git`
2. **Navigieren in das Projektverzeichnis:** Anschließend wechselt man in das Verzeichnis des geklonten Projekts: `cd DataWhispers-Stock-Price-Prediction-Projekt-DHBW`
3. **Installation der notwendigen Pakete:** Die für den Betrieb erforderlichen Python-Pakete werden mittels Pip installiert: `pip install -r Code/requirements.txt`

Starten der Webanwendung:

Die Webanwendung kann auf zwei Arten gestartet werden: Entweder durch Ausführung des Befehls: `streamlit run Code/_Home.py` im Terminal, woraufhin Streamlit den Webserver initialisiert und die App im Browser zugänglich macht, oder durch direktes Anklicken eines bereitgestellten Links, der den Nutzer direkt zur Anwendung führt.

Nutzung des Predictors:

Innerhalb der Streamlit-Webanwendung können Nutzer durch verschiedene Sektionen navigieren, um spezifische Funktionen zu nutzen. Interaktive Steuerelemente erlauben es den Anwendern, die Analyse und die daraus resultierenden Vorhersagen individuell anzupassen. Die Ergebnisse werden übersichtlich in Form von Diagrammen und Grafiken dargestellt, sodass sie leicht verständlich und interpretierbar sind.

8. Punkteverteilung

Matrikel Nr. 3071508 Finn: 90%

Matrikel Nr. 9973152 German 110%

Matrikel Nr. 5658606 Michael 110%

Matrikel Nr. 6367227 Hendrik 90%