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 :	19. Dezember 2024

Inhaltsverzeichnis

Abbildungsverzeichnis

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 Abstrakt

Diese Arbeit widmet sich der Forschungsfrage, inwieweit sich ein Transformer-Encoder-Modell mit verhältnismäßig wenig Parametern für eine Sentiment-Analyse eignet. Zusätzlich soll geprüft werden, wie groß der Einfluss des Umfangs an Trainingsdaten auf die Modellleistung ist. Aktuelle große Sprachmodelle nutzen Transformer-Architekturen, bei denen festgestellt wurde, dass deren Leistung besonders hervorsticht, wenn diese mit sehr vielen Daten trainiert wurden und Unmengen an lernbaren Parametern besitzen. Diese Anforderungen lassen sich nur von sehr finanzstarken Unternehmen erfüllen. Die Forschungsarbeit kann dazu beitragen, einen optimalen Punkt bei der Erreichung eines MVP-Modells (minimal funktionsfähige Iteration eines Modells) zu identifizieren. Als Leistungsrichtwert, der übertroffen werden muss, dient ein lexikonbasiertes Modell aus der "NLTK" Python-Bibliothek[3].

2.2 Datenaufbereitung

Als Datengrundlage dienen 3.6Mio Amazonreviews, welche auf Kaggle.com zur Verfügung gestellt werden[4].

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] *Hutto*, *C.*, *Gilbert*, *E.*, 2014.
[4] *Bittlingmayer*, *A.*, 2019.

Bei der Datenaufbereitung resultieren 2 Abwandlungen der Reviews und eine Zahlendarstellung (Abb. 1). Bei ”review_tokens” handelt es sich um eine Schlagwortextraktion, bei ”review_features” um eine reduzierte Form von ersterem (5k mögliche Schlagwörter) und bei ”feature_vec” um eine Zahlendarstellung von zweiterem. Zur Komplexitätsreduktion werden lediglich Zeilen mit mindestens 10 und maximal 40 Review-Features beibehalten. Falls ein Review weniger als 40 Features aufweist, wird der Zahlenvektor durch linksseitiges Padding mit 0 passend aufgefüllt. Die Daten werden in 2 Teilen gespeichert, mit 80% für das Training des Transformer-Encoders und 20% zum Testen.

2.3 Basis Modell

Das Basis Modell erreicht bei einer Anwendung auf Testdaten eine Genauigkeit von ca. 70% (Abb. 2), welche vom Transformer-Encoder-Modell übertroffen werden muss.

Abbildung 2: Basis Modell Performance

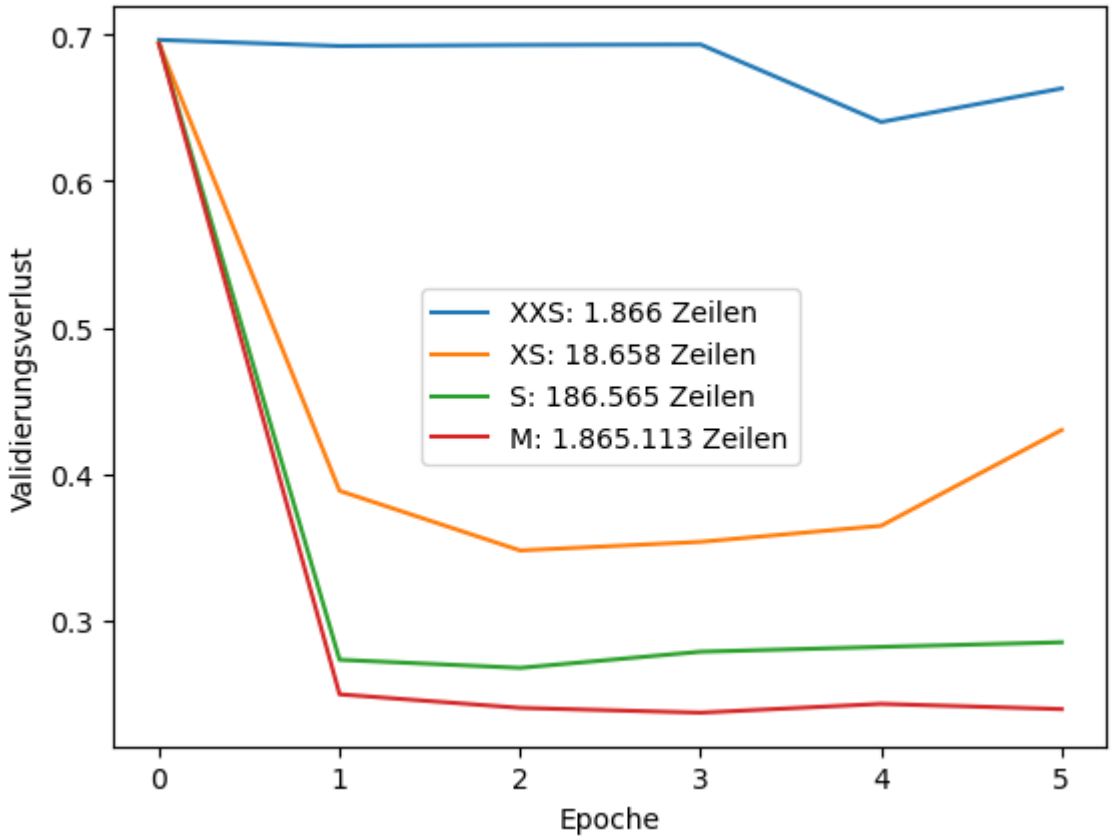
Eingabe	Genauigkeit (%)
review_tokens	69,8
review_features	69,2

Transformer-Encoder

Genaue Architekturdetails des umgesetzten Transformer-Encoders befinden sich im Anhang als Bild ”model_architecture.png” und orientieren sich an einer Vorlage von Andrej Karpathy[5], wobei mit dem Python-Paket ”PyTorch” gearbeitet wird[6]. Als Eingabe für den Transformer-Encoder dienen die Feature-Vektoren, welche eine einheitliche Kontextlänge von 40 aufweisen. Zur Untersuchung des Einflusses der Datenmenge auf Modellleistungen werden 4 Datengrößen angewandt (Abb. 3).

[5] *Karpathy*, *A.*, 2024.
[6] *PyTorch Foundation*, 2024.

Abbildung 3: Verlauf des Validierungsverlusts



Es ergeben sich die in Abbildung 4 dargestellten Genauigkeiten.

Abbildung 4: Transformer-Encoder-Modell Performance

Trainingsdatengröße	Genauigkeit (%)
XXS	63,5
XS	85,4
S	88,8
M	90,3

Daraus geht hervor, dass bereits wenige Trainingsdaten (XS) ausreichen, um ein lexikonbasiertes Verfahren bei einer Sentimentanalyse zu übertreffen. Außerdem wird deutlich, dass die notwendige Steigerung der Trainingsdatenmenge für eine spürbare Verbesserung der Modellleistung exponentiell mit der bereits vorhandenen Performance wächst.

Dabei ist mit einem Sättigungspunkt zu rechnen, wobei man spätestens dann dazu übergehen sollte, die Modellkomplexität durch z.B. zusätzliche Schichten zu erhöhen.

3 Erstellung eines Dokumentenkorpus

Die Umsetzung erfolgt in logischen Schritten und orientiert sich dabei am Prozessmodell **CRISP-DM!** (**CRISP-DM!**). Im ersten Schritt (Kap. 3.1) werden Patentdaten gemäß der Zielvorgabe beschaffen. Wie bereits im Kapitel 1 angedeutet, beschränkt man sich auf die Regionen Europa, Nordamerika und Ostasien, wobei lediglich Daten bezüglich bedeutender Länder einbezogen werden. Als sinnvoller zeitlicher Rahmen in Bezug auf das Veröffentlichungsdatum der zu untersuchenden Patente wurde 2022-Q1 bis einschließlich 2024-Q2 festgelegt. Der Grund dafür ist, dass lediglich aktuelle Innovationen für die Forschungsfrage relevant sind. Außerdem verkürzt sich die Entwicklungszeit bezüglich technischer Neuheiten zunehmend, so dass ein Zeitfenster von ca. $2,5$ Jahren in der Vergangenheit genügend Aussagekraft besitzt.

3.1 Datenbeschaffung

Als Datenquelle wird das in Kapitel **??** behandelte Online-Recherchetool DEPATISnet verwendet. Dabei stehen verschiedene Recherchemodi zur Verfügung. Als besonders zielführend hat sich der Expertenmodus erwiesen, bei dem komplexe syntaktische Suchanfragen möglich sind. Dabei erfolgt die Formulierung der Suchkriterien in Form einer Zeichenfolge, in der eine Vielzahl an nützlichen Operatoren und Platzhaltern zur Effizienzsteigerung einsetzbar sind[7].

Trefferlistenkonfiguration

Über die Trefferlistenkonfiguration lassen sich relevante Patentattribute wählen, welche bei der Suche in den Resultaten erscheinen sollen. Folgende Attribute werden angefordert:

- Veröffentlichungsnummer
- Veröffentlichungsdatum
- IPC-Hauptklasse
- IPC-Neben-/Indexklassen
- Titel
- Zusammenfassung

[7] **dpma_hilfe**.

Suchkriterien

Um gemäß der Forschungsfrage lediglich Patente mit Bezug zum Bereich Robotik als Resultate zu erhalten, wird dementsprechend eine Bedingung formuliert, welche innerhalb des Titel (TI) Attributs und der Zusammenfassung (AB) dahingehend prüft (Code 1, Zeile 1-3).

Code 1: Eingabe für Expertenrecherche auf DEPATISnet

```
1  (TI = (robot? OR telerobot? OR exoskeleton? OR ((bionic
     OR intelligent?)(2A)prosthet?))
2  OR
3  AB = (robot? OR telerobot? OR exoskeleton? OR ((bionic
     OR intelligent?)(2A)prosthet?)))
4  AND (Pub >= <Startdatum> AND Pub <= <Enddatum>)
5  AND AC = (<(OR verknüpfte) Ländercode(s)>)
```

Quelle: Eigene Darstellung

Dies erfolgt unter Zuhilfenahme von Operatoren und dem Platzhalter `?`, welcher 0 bis mehrere beliebige Zeichen ersetzt. Bei `2A` handelt es sich um einen Nachbarschaftsoperator, welcher einen wahren Rückgabewert liefert, falls beide Operanden mit einem maximalen Abstand von 2 Wörtern zueinander im Text vorkommen. Aus dem Suchstring geht hervor, dass nicht nur explizit die Begriffe `robot?` bzw. `telerobot?` zur Filterung verwendet werden, sondern auch die Begriffe `exoskeleton?` und `prosthet?`, wobei bei Letzterem als Bedingung ein Vorkommen des Wortes `bionic` oder `intelligent?` in naher Nachbarschaft festgelegt wird. Diese beiden Themenfelder sind eng mit Robotik verwandt, da diese in vielen technischen Bereichen, wie zum Beispiel mechanische Bewegung, sehr ähnliche Herausforderungen bewältigen.

Die Anzahl der möglichen Resultate ist seitens DEPATISnet auf 10.000 begrenzt[8]. Daher werden diese mit Hilfe des Attributs Veröffentlichungsdatum (Pub) in passenden Intervallen heruntergeladen (Code 1, Zeile 4), so dass das Zielintervall 2022-Q1 bis 2024-Q2 erfüllt ist.

Eine Filterung nach Regionen erfolgt über das Attribut Anmeldeland (AC). Damit werden jeweils die bedeutendsten Länder bzw. Patentämter mittels Codes selektiert (Code 1, Zeile 5).

[8] **dpma_hilfe**.

Tabelle 1: Trefferanzahl bei regionaler Filterung mittels Ländercodes

Ländercodes	Trefferanzahl
Europa	11.421
European Patent Office (EP)	8.073
Deutschland (DE)	1.937
United Kingdom (GB)	727
Frankreich (FR)	558
Spanien (ES)	166
Nordamerika	20.541
United States (US)	
Ostasien	110.120
China (CN)	

Quelle: Eigene Darstellung

Für die Region Europa werden die 5 bedeutendsten Länder bzw. Patentämter berücksichtigt. Aufgrund der Beschränkung auf einen Vertreter aus den Regionen Nordamerika und Ostasien werden diese im Verlauf der Arbeit auch mit USA respektive China referenziert.

Download

Zum Herunterladen der Treffer stehen seitens DEPATISnet die Formate CSV, XLSX und PDF zur Verfügung. Als praktikabel und robust hat sich XLSX erwiesen. Das CSV-Format hingegen ist aufgrund von ”;” als Trennzeichen und textueller Werte ohne Ummantelung (z.B. mit ”) problematisch. Daher werden die Daten aus DEPATISnet im XLSX-Format heruntergeladen.

3.2 Aggregierung

Aufgrund der Begrenzungsproblematik hinsichtlich maximaler Trefferanzahl in DEPATISnet von 10.000, erfolgt nach dem Download noch eine programmatische Aggregierung der Dateninkremente. Dabei wird aus den Excel-Teildaten eine CSV-Datei erzeugt, wobei die Region als differenzierendes Attribut hinzugefügt wird. Der resultierende Dokumentenkorpus hat eine Größe von 160MB.

4 Erstellung von Datenkorpora

Kapitel 4.1 befasst sich mit Datenaufbereitung, bei der es darum geht, den Dokumentenkorpus von Dubletten und anderen Anomalien zu bereinigen. Anschließend werden Datenkorpora passend zum Anwendungszweck abgeleitet. Für die Analyse bezüglich der Forschungsfrage resultiert zum einen ein Datenkorpus mit IPC-Werten (Kap. 4.2) und zum anderen einer mit manuell modellierten Themen (Kap. 4.4).

4.1 Datenaufbereitung

Bei der Datenaufbereitung wird eine Pipeline durchlaufen. Im ersten Schritt werden ungültige Beobachtungen entfernt und anschließend erfolgt eine Filterung bezüglich Veröffentlichungsdatum, so dass lediglich Zeilen im einschließenden Bereich 2022-Q1 bis 2024-Q2 übrig bleiben (Abb. **??**). Letzteres ist notwendig, auch wenn dies bereits bei der Suchanfrage auf **DEPATIS!** (**DEPATIS!**)net umgesetzt wurde (Kap. 3.1 Code 1).

Als Nächstes werden die Textattribute Titel und Abstrakt auf die Sprache Englisch und alternativ Deutsch reduziert. Dies entspricht den Schritten 3-4 in Abbildung **??**.

Anschließend werden Abstrakt-Duplikate entfernt und es erfolgt eine Kumulation der Textattribute zu einem ”text” Attribut, wobei der Titel mit einem Punkt vom Abstrakt getrennt vorangestellt wird (Abb. **??** Schritte 5-6).

4.2 Datenkorpus für IPC!-Analysen

Trotz dessen, dass bei einer **IPC!**-Analyse gemäß der Forschungsfrage keine Titel und Abstrakt Daten von Patenten notwendig sind, wird hierbei an den Schritt 6 aus Abbildung **??** angeknüpft. Denn der vorher stattfindende Vorgang 5 ist wichtig, um logische Patent-Wiederholungen auszuschließen, welche in der Praxis vorgekommen sind.

Als 7. Schritt erfolgt eine Komplexitätsreduktion, indem das Attribut ”ipc” eingeführt wird, welches die **IPC!**-Hauptklasse und alternativ die erste **IPC!**-Nebenklasse repräsentiert. Anschließend erfolgt mit Hilfe eines Python-Pakets namens ”wipo_ipc”[9] eine Titelauflösung von IPC-Symbole, bei der Zeilen ohne resultierenden gültigen Wert entfernt werden (Abb. **??**).

[9] **wipo_ipc**.

Abschließend lässt sich ein Datenkorpus für **IPC!**-Analysen ableiten, welcher aus den Attributen Veröffentlichungsnummer (”id”), Veröffentlichungsdatum (”pub_date”), Region (”region”) und **IPC!**-Titel (”ipc_title”) besteht. Als Speicherformat wird CSV gewählt und die Größe beträgt 12MB.

4.3 Datenkorpus für Themenmodellierung mit BERTopic

In Kapitel **??** wurde bereits über den modularen Aufbau der BERTopic-Vorgehensweise informiert. Demnach werden im ersten Schritt aus Textdaten Embeddings generiert. Dieser Vorgang wird im Falle von Parameter-Finetuning wiederholt[10]. Die Erzeugung von Embeddings ist im Falle der Projektarbeit unabhängig vom Finetuning und kann demnach bereits im Datenkorpus umgesetzt sein, wodurch viel Rechenaufwand eingespart wird. Bevor dieser Umwandlungsschritt vollzogen wird, erfolgt zunächst eine Bereinigung der Textdaten, so dass beispielsweise Sonderzeichen inklusive Klammerausdrücke entfernt sind (Abb. **??**). Das Ziel ist die Herstellung einer sauberen Satzstruktur mit Berücksichtigung von Sprachsonderheiten, damit beim Embedding mittels Sprachmodell bestmögliche Ergebnisse erreicht werden.

Aufgrund dessen, dass sowohl deutsche als auch englische Texte vorkommen, wird das multilinguale Sprachmodell ”distiluse-base-multilingual-cased-v1” verwendet[11]. Dieses erzeugt eine Vektordarstellung mit 512 Dimensionen zu jedem Text. Zudem wird für das Clustering mit BERTopic eine Darstellung mit weniger Dimensionen erzeugt. Dies erfolgt durch Anwendung von **UMAP!** (**UMAP!**)[12], wobei Darstellungen mit 5 Komponenten resultieren. Der letztendliche Datenkorpus für eine Themenmodellierung beinhaltet die Attribute Veröffentlichungsdatum (”id”), Text, 512D-Embeddings (”emb512d”) und 5D-Embeddings (”emb5d”). Als Speicherformat wird CSV gewählt und die Größe beträgt 1.4GB.

4.4 Datenkorpus für Analysen modellierter Themen

Der erste Schritt beinhaltet das Clustering der 5D-Embeddings in dem Datenkorpus aus Kapitel 4.3. Dafür wird **HDBSCAN!** (**HDBSCAN!**) verwendet[13]. Als Mindestgröße für Themenblöcke werden 30 Patente angegeben. Die Anzahl resultierender Gruppierungen ergibt sich algorithmisch. Im vorliegenden Fall werden 171 Themen identifiziert, wobei 29550 Patente keinem Thema zugeordnet sind (Abb. **??**).

[10] **website:bertopic_bestpractices**.
[11] **website:st_bert_models**.
[12] **umap**.
[13] **hdbscan**.

Für die Erzeugung von repräsentativen Themennamen werden zunächst mit Hilfe von **c-TF-IDF!** (**c-TF-IDF!**) themenrelevante Stichwörter ermittelt, wobei nicht auf Dokumentenebene (**TF-IDF!** (**TF-IDF!**)) differenziert wird, sondern auf Cluster-Ebene (**c-TF-IDF!**)[14,15]. Anschließend werden die in Abbildung **??** unter ”Name” aufgelisteten Themenrepräsentationen mit Hilfe des OpenAI-Sprachmodells ”gpt-3.5-turbo” erzeugt[16]. Dies wird von BERTopic unterstützt, wobei intern Anfragen mit Beigabe des Kontexts, in Form von Keywords und repräsentativen Texten zu jedem Cluster, an die OpenAI-API gesandt werden. Als Antwort wird ein passender Themenname zurückgegeben.

Im nächsten Schritt wurden manuell ansprechende Hyperonyme identifiziert, so dass sich die Anzahl an möglichen Topics reduziert, wodurch folgende Analysen übersichtlicher und leichter verdaulich sind (Tab. 2 u. 3). Das anwendungsspezifische Thema ”Militär” wurde nicht direkt erkannt, da Neuerungen diesbezüglich in der Regel einer Geheimhaltung unterliegen. Es soll dennoch untersucht werden, in welchem Ausmaß Patente einen Bezug zu militärischen Zwecken aufweisen.

Tabelle 2: Hyperonym ”Anwendung”

Thema	**Keywords**
Reinigung und Haushalt	Reinigung, Haushalt, waschen, ...
Landwirtschaft und Tierhaltung	Tierfütterung, Tierarztpflege, ...
Militär	Armee, Kampfdrohne, Militär, Waffe, ...
Gesundheit und Wohlbefinden	medizinisch, Wohlbefinden, ...
Sicherheits- und Rettungsdienste	Sicherheit, Rettung, ...
Küchentechnologie und Gastgewerbe	Küche, Gastgewerbe, Kochen, ...
Lagerung und Logistik	Etikettierung, Lagerung, Logistik, ...
Handwerk	Bauprozess, Handwerk, Messen, ...

Quelle: Eigene Darstellung

[14] **website:bertopic_ctfidf**.
[15] **tfidf**.
[16] **website:bertopic_llm**.

Tabelle 3: Hyperonym ”Technologie”

Thema	**Keywords**
Energieversorgung und Ladeinfrastruktur	Laden, Stromversorgung, ...
Teleoperation	Fernkommunikation, Opt-in-Anfrage, ...
Autonomie	Autonomie, selbstheilend, ...
Exoskelett	Exoskelett, Orthese, Prothese, ...
Form und Bewegung	Körperform, Bewegung, humanoid, ...
Physisches Handwerk	flexibler Greifer, Teleskoparm, ...
Sensorisches Handwerk	Sensor, Erkennung, Sonar, ...
Material	strahlungsbeständig, Wärmeableitung, ...

Quelle: Eigene Darstellung

Abschließend werden zu jedem Dokument sowohl anwendungs- als auch technologiespezifische Themen aus den Tabellen 2 und 3 zugeordnet. Dies erfolgt unter Anwendung der Kosinus-Ähnlichkeit zwischen Dokument-512D-Embedding und Thema-Keyword-Embedding (**??**). Aufgrund einer Mindestgröße der Ähnlichkeit als Bedingung können Dokumente zu 0 bis mehreren Themen zugeordnet werden. Dies ist in Anbetracht der gewählten Hyperonyme sinnvoll, da beispielsweise Patente existieren können, die sowohl zum Thema ”Exoskelett” als auch ”Form u. Bewegung” zuordenbar sind. Der resultierende Datenkorpus wird im CSV-Format abgespeichert und hat eine Größe von 5.6MB.

Abbildung **??** fasst den Prozess zur Erstellung eines Datenkorpus für Themenanalysen zusammen.

5 Vergleichende Analysen

5.1 Äquivalenzfaktor

Aus Tabelle 1 im Kapitel 3.1 geht bereits hervor, dass Ostasien hinsichtlich Patentmenge weit vorne liegt. Die zu untersuchenden Regionen mit ihren bedeutendsten Vertretern unterscheiden sich jedoch auch deutlich hinsichtlich der Anzahl an Beschäftigten. Um die Intensität der Robotik-Entwicklung vergleichbarer zu machen, werden Äquivalenzfaktoren auf Basis der Beschäftigtenzahl je Region gebildet (Formel 1). Als Datenquelle für die

Formel 1: Durchschnittliche Beschäftigtenzahlen pro Jahr und Äquivalenzfaktoren

$$\text{China\_avg} = 749121\text{k} \tag{1}$$
$$\text{USA\_avg} = 165665\text{k} \tag{2}$$
$$\text{EU\_avg} = 241456\text{k} \tag{3}$$
$$\text{Total\_avg} = 385414\text{k} \tag{4}$$
$$\text{Equivalence factor} = \frac{\text{Total\_avg}}{\text{Region\_avg}}. \tag{5}$$
$$\text{China} = 0.51 \tag{6}$$
$$\text{USA} = 2.33 \tag{7}$$
$$\text{EU} = 1.6 \tag{8}$$

Quelle: Beschäftigtenzahlen (Alter >14 Jahre) aus ILOSTAT am 30.06.2024

Zahlenangaben dient **ILOSTAT!** (**ILOSTAT!**)[17].

5.2 IPC! Analysen

Als Basis für **IPC!**-Analysen dient der in Kapitel 4.2 beschriebene Datenkorpus.

Ein Blick auf die Verteilung der 8 häufigsten **IPC!**-Titel in Abbildung **??** zeigt deutlich, dass der Bereich Handwerk dominiert. Dies ist wenig überraschend, da mit Robotik in diesem Zusammenhang viel körperliche Anstrengung vermieden werden kann.

In Abbildung **??** erfolgt ein direkter interkontinentaler Vergleich von Anteilen an der Gesamtmenge je Thema. Durch die Gewichtung der Anteile mit den Äquivalenzfaktoren aus Kapitel 5.1 sollen Intensitätsunterschiede deutlich gemacht werden. Es fällt auf, dass China in fast allen Bereichen anteilsmäßig führt. Bemerkenswert ist das Thema Medizin mit USA als offensichtlichem Anteilsführer. Europa ist trotz Gewichtung in allen Bereichen das Schlusslicht.

[17] **website:ilostat**.

5.3 Analysen modellierter Themen

Auf Basis des Datenkorpus aus Kapitel 4.4 werden Analysen bezüglich der Hyperonyme Technologie und Anwendung durchgeführt.

5.3.1 Technologie

Analyse von Bereichsintensitäten

Der anteilsmäßige Vorsprung von China ist in den meisten Bereichen groß (Abb. **??**). Lediglich der Bereich Teleoperation wird von den USA dominiert.

Longitudinale Schwerpunktanalyse

Der Fokus aller drei Regionen liegt auf Technologie mit Bezug zu Handwerk, Autonomie, Form und Bewegung (Abb. **??**-**??**), wobei dieser bei China besonders ausgeprägt ist (Abb. **??**).

Bei Europa fällt auf, dass der Bereich Exoskelett eine schwankende Anzahl an patentierten Neuerungen aufweist und im Schnitt weniger Beachtung findet als in den anderen Regionen (Abb. **??**).

Die Themenentwicklungen von USA weisen überwiegend einen relativ deutlichen Aufwärtstrend auf und in China ist bereichsübergreifend ein sehr starker Abwärtstrend zu erkennen (Abb. **??** u. **??**).

5.3.2 Anwendung

Analyse von Bereichsintensitäten

Im Falle des Hyperonyms ”Anwendung” liegt China in allen Themenbereichen anteilsmäßig vorne, ausgenommen Gesundheit und Wohlbefinden, welches von USA dominiert wird (Abb. **??**). Diese logische Deckungsgleichheit mit der medizinischen Anteilsverteilung in Abbildung **??** aus dem Kapitel 5.2 ist ein guter Indikator für die Korrektheit der Themenmodellierung in Kapitel 4.4.

Longitudinale Schwerpunktanalyse

Der Fokus in allen drei Regionen liegt auf Handwerk und Militär, wobei in China ein besonderer Schwerpunkt auf diesen beiden Anwendungsbereichen liegt (Abb. **??**-**??**).

Die Anzahl an USA-Patenten mit Bezug zu den modellierten Themen steigt relativ stark an, wobei der Bereich Sicherheit und Rettung mit einer Steigerung von 375% im Zeitraum von 2022-Q1 bis 2024-Q2 auffällt (Abb. **??**). Stattdessen sind die Patentmengen der Region China bereichsübergreifend stark fallend.

6 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 1 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[18].

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 **DEPATIS!** Dokumentenarchiv einbezogen. Ein weiterer wichtiger Schritt ist eine Ausweitung der Datenquellen. Dabei sollen mindestens

[18] **vincent2023openai**.

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.

7 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

Bittlingmayer, *Adam* (2019): Amazon Reviews for Sentiment Analysis, <https://www.kaggle.com/datasets/bittlingmayer/amazonreviews> (2019) [Zugriff: 2024-07-02]

Hutto, *C.J.*, *Gilbert*, *E.E.* (2014): VADER Sentiment Analysis, <https://github.com/cjhutto/vaderSentiment> (2014) [Zugriff: 2024-07-03]

Karpathy, *Andrej* (2024): nanoGPT, <https://github.com/karpathy/nanoGPT> (2024-06-03) [Zugriff: 2024-07-03]

PyTorch Foundation (2024): PyTorch, <https://pytorch.org/> (2024) [Zugriff: 2024-07-04]

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)