



Abschlussbericht

***BMBF-Verbundprojekt:
Dynamische Textdaten-basierte Output-Indikatoren als Basis
einer neuen Innovationsmetrik
- DynTOBI –***

Justus-Liebig-Universität Gießen

Albina Latifi, David Lenz, Monika Schuhmacher, Peter Winker

**ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH
Mannheim**

Bettina Peters, Georg Licht, Jakob Ehlich, Jan Kinne, Leon Steines

Gießen, Januar 2025

Zuwendungsempfänger:

JLU – Justus-Liebig-Universität Gießen

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH
Mannheim

Förderkennzeichen:

16IFI114 und 16IFI115

Vorhabenbezeichnung:

Dynamische Textdaten-basierte Output-Indikatoren
als Basis einer neuen Innovationsmetrik
- DynTOBI -

Laufzeit des Vorhabens:

15.04.2021 bis 30.06.2024

Berichtszeitraum:

15.04.2021 bis 30.06.2024

Projektteam JLU:

Prof. Dr. Peter Winker (JLU), Projektleiter

Albina Latifi

Dr. David Lenz

Prof. Dr. Monika Schuhmacher

Projektteam ZEW:

Dr. Georg Licht (ZEW), Teilprojektleiter (bis Juni 2023)

Prof. Dr. Bettina Peters, Teilprojektleiterin (ab Juni 2023)

Dr. Jan Kinne

Jakob Ehlich (ab Juni 2023)

Leon Steines (ab Juni 2023)

Dr. Julian Dörr (bis 2022)

Kurzzusammenfassung:

Ziel des DynTOBI-Projekts war es, die Diffusionsprozesse von Innovationen im Zeitverlauf zu beschreiben und Ansätze zur Bestimmung der voraussichtlichen Entwicklung und Verbreitung von Innovationen zu entwickeln. Hierfür wurden innerhalb des Verbundprojekts zwei Datensätze verwendet.

Während das JLU-Team vorrangig mit einem Längsschnittdatensatz auf Basis der Newsticker-Daten des heise-Verlags gearbeitet hat, verwendete das ZEW-Team in Kooperation mit istari.ai webbasierte Daten, die mit Informationen aus Unternehmensdatenbanken für den entsprechenden Zeitraum angereichert wurden. Durch den Rückgriff auf Webarchive wurde dieser Datenbestand für mehrere Zeitperioden erweitert und mit entsprechenden Unternehmensdaten auf Basis des Mannheimer Unternehmenspanel verknüpft. Die Innovationsdynamik auf der Basis dieser „historischen“ Web- und Unternehmensdaten wurde mit der Innovationsdynamik, die im Mannheimer Innovationspanel beobachtet wird, verglichen.

Für die Datenbasis der JLU wurden alle bis zum 18.10.2021 verfügbaren Artikel im Nachrichtenarchiv der Technologiezeitschrift „heise.de“ herangezogen, was insgesamt einem Umfang von 190.722 Nachrichtenartikel entspricht. Diese umfangreiche Textdatenbasis wurde mittels Methoden der natürlichen Sprachverarbeitung analysiert. Zunächst wurden aus diesem Korpus Themen modelliert, welche anschließend mithilfe von Experten in die Kategorien „innovationsbezogen“ und „nicht innovationsbezogen“ eingeordnet wurden.

Die so ermittelten Innovationsfelder wurden in einem nächsten Schritt hinsichtlich der dazu in der Berichterstattung relevanten Entitäten untersucht. Unter einer Entität verstehen wir in diesem Kontext ein Unternehmen, das im Text durch seinen Unternehmensnamen repräsentiert wird. Hierfür wurde ein Eigennamenerkennungsmodell entwickelt. Ziel dieser Analyse war es, potentielle Innovationstreiber den verschiedenen Innovationsfeldern zuzuordnen. Des Weiteren haben wir eine Metrik entwickelt, die ausgehend von den identifizierten Entitäten ein Affinitätsmaß zu bestimmten Themenfeldern liefert. Wir zeigen in unserer Publikation „Identification of innovation drivers based on technology-related news articles“ einige Entitäten, die als innovationsaffin gelten. Des Weiteren zeigen wir für exemplarische Innovationsfelder wie „E-Mobilität“ und „autonome Fahrzeuge“ die dazugehörigen Innovationstreiber. Neben den in diesem Kontext zu erwartenden großen Firmen lassen sich mit unserem Ansatz auch kleinere Innovationstreiber identifizieren.

Zu jedem Themenfeld lag zudem der dazugehörige Verlauf der Relevanz im Zeitablauf vor. Das bedeutet, dass die Themengewichte im Zeitablauf als Zeitreihe dargestellt werden können. Wir haben die Diffusionsverläufe aller Themen mittels Methoden der funktionalen Datenanalyse analysiert, um so prototypische Innovationsdiffusionskurven zu identifizieren. Die Verläufe wurden mit einem globalen Optimierungsalgorithmus, der auf der Threshold Accepting Heuristik basiert, zu Gruppen mit ähnlichen Mustern zugesammengefasst. Unter diesen prototypischen Mustern gab es einige, die besonders häufig Themen mit Innovationsbezug umfassen. Diese unterschiedlichen Muster deuten darauf hin, dass es empirisch unterschiedliche Verläufe der Diffusion der Relevanz in der öffentlichen Wahrnehmung gibt, die nicht alle im gleichen Maße der klassischen Diffusionskurve aus der Innovationsforschung ähneln. Darüber hinaus konnten mit diesem Ansatz für spezifische Themen unterschiedliche Stadien im Verlauf der Innovationsdiffusion aufgezeigt werden.

Der zentrale Gegenstand des ZEW-Teilprojekts bestand in der Untersuchung von Innovationsdynamiken, also der Entwicklung und Veränderung von Innovationsaktivitäten im zeitlichen Verlauf auf Basis von textbasierten Unternehmensdaten. Für die Untersuchung und Generierung von textbasierten Indikatoren zur Messung von Innovationsdynamiken wurde ein Paneldatensatz an Unternehmenswebseiteninhalten generiert und auf Basis dessen webbasierte Innovationsindikatoren über die Zeit ermittelt. Die Datengrundlage hierfür bildeten Webarchive. Das ZEW hat hierzu in Kooperation mit ISTARI.AI im Rahmen des Projekts einen ArchiveScraper entwickelt, welcher effizient die archivierten Webseitenbestände deutscher Unternehmen abfragen kann. Dazu wurde eine Amazon Web Services (AWS) cloudbasierte Softwarelösung entwickelt, mit der parallelisiert und distribuiert auf die beim Internet Archive hinterlegten WARC Files zugegriffen werden kann. Auf Basis einer Input-Liste mit URL-Datum-Kombinationen wurden aus diesen WARC Files die zum gesuchten Datum passenden historischen, archivierten Website-HTML extrahiert und als neue, jahresspezifische WARC-Files auf AWS S3 Cloudspeicher abgelegt. Von dort wurden diese WARC Files heruntergeladen und die historischen Unternehmenswebseiten analysiert und ein jährlicher Indikator generiert, der die Innovationswahrscheinlichkeit auf Unternehmensebene auf Basis von historischen Informationen auf den Webseiten der Unternehmen widerspiegelt. Dieser Paneldatensatz wurde mit Unternehmensdaten des Mannheimer Unternehmenspanel (MUP) und mit Innovationsdaten aus dem Mannheimer Innovationspanel (MIP) verknüpft.

Auf Basis des neu generierten Paneldatensatzes mit webbasierten Innovationsindikatoren auf Unternehmensebene wurde der zeitliche Verlauf des Innovationsgeschehens auf Basis des Webpanels verglichen mit dem zeitlichen Verlauf des Innovationsgeschehens auf Basis des MIP, welches verschiedene Innovationsindikatoren seit nunmehr 32 Jahren auf Basis einer repräsentativen Erhebung bei Unternehmen mit Sitz in Deutschland und mit mindestens 5 Beschäftigten erhebt. Der Vergleich basierte sowohl auf aggregierten Zeitreihen als auch auf mikroökonomischen Analysen auf Ebene der Unternehmen. Darauf aufbauend wurde eine abschließende Beurteilung der Nutzung webbasierter Innovationsindikatoren für eine Längsschnittanalyse vorgenommen und ein Konzept zur Implementation und Kostenabschätzung, wie ein permanentes Nah-Echtzeit-Innovations-Monitoring-System für Deutschland aufgebaut werden kann, entwickelt.

Inhalt

1	Projektergebnisse	7
1.1	Wissenschaftliche Papiere	7
1.1.1	Papier 1: Identification of innovation-drivers based on technology-related news articles (JLU)	7
1.1.2	Papier 2: Finding Shapes of Diffusion of Innovation Curves by Clustering Time Series of Topic Weights (JLU).....	12
1.1.3	Papier 3: Innovation Panel Surveys in Germany: the Mannheim Innovation Panel (ZEW)	15
1.1.4	Papier 4: Innovation Dynamics using Web-based Innovation Indicators (ZEW).....	16
1.2	Technische Dokumentation	21
1.3	Präsentationen auf wissenschaftlichen Konferenzen und Workshops	21
1.4	Öffentlichkeitsarbeit	22
1.5	Abgeschlossene Dissertationen im Rahmen des DynTOBI Projekts	22
1.6	Externe Entwicklung des Forschungsfelds	22
2	Folgeverwertung	24
2.1	Startup istari.ai.....	24
2.2	Folgeprojekte	24
3	Literaturliste.....	26

1 Projektergebnisse

1.1 Wissenschaftliche Papiere

Insgesamt entstanden aus dem DynTOBI Projekt im Teilvorhaben der JLU zwei wissenschaftliche Papiere.

1.1.1 Papier 1: Identification of innovation-drivers based on technology-related news articles (JLU)

Es ist allgemein bekannt, dass innovative Unternehmen zum Wirtschaftswachstum beitragen durch z.B. der Einführung neuer Produkte oder Dienstleistungen (Carree und Thurik, 2010). Wirtschaftspolitische Entscheidungsträger sind insbesondere darin interessiert, die treibenden Kräfte von Innovationsaktivitäten – sprich Innovationstreiber – zu erkennen, um diese nötigenfalls mit gezielten Fördermaßnahmen unterstützen zu können. Existierende Methoden, um diese Art von Informationen zu erhalten sind z.B. die Durchführung von Umfragen, die Analyse von Patenten oder die Bewertung aggregierter Maße zu den Ausgaben für Forschungs- und Entwicklung. Die etablierten Methoden haben den essenziellen Nachteil, dass diese in der Regel einen nachlaufenden Charakter haben; z.B. kommen lange Veröffentlichungsfristen für Erhebungs- und Patentdaten zum Tragen. Zudem ist die Analyse von Umfragedaten mit hohen Kosten verbunden (Kinne und Axenbeck, 2020). Mit unserem auf aktuell verfügbare Textdaten gestützten Ansatz zielen wir darauf ab, kurzfristig und mit hoher Granularität, potentielle Innovationstreiber für spezifische Innovationsfelder aufzudecken.

Unsere Analyse basiert auf einem umfangreichen Textkorpus aus dem Archiv der technologiebezogenen Zeitschrift "heise.de" von 1996 bis 2021. Der Textkorpus hat einen Umfang von 190.722 Nachrichtenartikeln. Um Innovationsfelder zu identifizieren, wurden aus diesem Korpus Themen modelliert. Im nächsten Schritt haben wir die Themen in Form von Wortwolken, wie diese in Abbildung 1 veranschaulicht sind, jeweils mehreren Experten vorgelegt, die diese in die Kategorien "innovationsbezogen" und "nicht innovationsbezogen" eingeordnet haben. Für jedes Themenfeld erhalten wir den Innovationsgrad als Anteil der Kategorisierung der 8 Expertenmeinungen. Wortwolke 67 repräsentiert z.B. das Thema "E-Mobilität" und wurde von allen Experten als innovationsbezogen kategorisiert (100% innovationsbezogen).

Topic #67



Topic #90



Abbildung 1: Wortwolken, die als Themenfelder zu interpretieren sind. Wortwolke 67 umfasst das Themenfeld „E-Mobilität“ und Wortwolke 90 umfasst das Themenfeld „autonome Fahrzeuge“.

Die Präsentation der folgenden Ergebnisse fokussieren sich auf dieses Themenfeld, die dabei eingesetzte Methode ist aber analog für andere Themenfelder geeignet. Um aus dem Nachrichtendatenkorpus Innovationstreiber zu ermitteln, wurde zunächst ein Eigennamenerkennungsmodell entwickelt, um Unternehmensnamen aus dem Textkorpus zu identifizieren. In Abbildung 2 sind die aus dem Modell identifizierten Eigennamen exemplarisch an einem kurzen Nachrichtenartikel dargestellt. In gelber Farbe hervorgehoben sind die identifizierten Unternehmensnamen, die im gesamten Korpus 1.246.524 Mal auftauchen. Nach weiteren Aufbereitungsschritten und einem Abgleich mit einer Unternehmensdatenbank wurden schließlich 7.946 einzigartige DACH-bezogene Unternehmen über ihren Namen identifiziert, auf die sich die weitere Auswertung fokussiert.

Durch die Kombination der Informationen über die vorherrschenden Themen und Entitäten (Unternehmensnamen) in einem Artikel, lassen sich zwei Folgerungen ziehen:

Zum einen erhalten wir ein Maß für die Topicaffinitäten (Themenfelder) jeder Entität (Unternehmen). Durch die vorherige Klassifizierung der Themenfelder in innovationsbezogen/nicht-innovationsbezogen, erhalten wir über die Aggregation der Themenaffinität der innovationsbezogenen Themenfelder ein Maß für

die Innovationsaffinität eines Unternehmens. In Abbildung 3 ist die Affinitätsverteilung über die Top-10-Themenfelder der Entität Bython dargestellt. Orangenfarbige Balken symbolisieren innovationsbezogene Themenfelder. Die höchste Affinität der Entität „BYTON“ ist mit 18,24 % für das Thema 101 (Start-Ups) gegeben. Gefolgt von zwei innovationsbezogenen Themenfeldern: Thema 67 (E-Mobilität mit 9,7%) und Thema 90 (autonome Fahrzeuge mit 8.6%). Somit können wir „BYTON“ als eine Entität mit einer relativ hohen Affinität zu innovationsbezogenen Themen einstufen.

Die Hildesheimer Firma Pios, Lizenznehmer von BeOS und deutscher Vertrieb der BeBox, will jedenfalls auch weiterhin für den Support ihrer rund 200 deutschen Kunden geradestehen, “solange es die BeBox und die Firma Pios gibt”, so der Geschäftsführer Geerd-Ulrich Ebeling gegenüber c’t.

Abbildung 2: Durch das Eigennamenerkennungsmodell identifizierte Entitäten. Gelb hervorgehoben sind die erkannten Unternehmensentitäten.

In Abbildung 4 sind die vier Themenfelder dargestellt zu denen, die Entität Bython die höchste Affinität hat. Es handelt sich neben den bereits genannten Themen 101 (Start-ups), 67 (E-Mobilität) und 90 (autonome Fahrzeuge) noch um das Thema 63 (Unternehmensübernahme). In Anbetracht der Tatsache, dass es sich bei dem betrachteten Unternehmen um ein inzwischen insolventes Start-up im Bereich Elektroautos handelt, erscheinen diese empirischen Ergebnisse überzeugend.

Zum anderen können wir aus der Kombination der Informationen über die vorherrschenden Themen und Unternehmensnamen in einem Artikel, die potentiellen Innovationstreiber für jedes Innovationsfeld ableiten. In Tabelle 1 sind die Top-15 der potentiellen Innovationstreiber für das Themenfeld der E-Mobilität gelistet. "NVIDIA" ist der wichtigste Innovationstreiber mit einem Anteil von 17,8%, gefolgt von "GM" mit einem Anteil von 13,7% und "AMD" mit einem Anteil von 4,2%. Neben den bekannten Unternehmen ist auch das ehemalige Start-Up "BYTON" mit einem Anteil von 3,5 % als Innovationstreiber aufgeführt. Aus qualitativer Sicht, basierend auf menschlichem Urteilsvermögen, erscheinen die identifizierten Unternehmen sinnvoll und umfassen nicht nur die üblichen Verdächtigen für die jeweiligen Innovationsfeldern.

Wir stellen auch fest, dass viele der als Innovationstreiber identifizierten Unternehmen in der DACH-Region ansässige Tochterunternehmen großer internationaler Konzerne sind. Für diese Unternehmen ist es nicht einfach zu entscheiden, ob der Innovationsbeitrag aus dem in der DACH-Region lokalisierten Tochterunternehmen oder vom internationalen Konzern stammt, da der Algorithmus eine solche Unterscheidung nicht zulässt. Zukünftige Forschungsbestrebungen könnten darauf abzielen, durch die Anreicherung relevanter Metadaten eine bessere Differenzierung in dieser Hinsicht zu ermöglichen, was auch für wirtschaftspolitische Entscheidungen relevant sein könnte.

Unternehmensname	Potentielle Innovationstreiber (%)
NVIDIA	17,8
GM	13,7
AMD	4,2
LG	3,5
BYTON	1,9
BYD	1,9
HP	1,7
NPE	1,6
BOSCH	1,5
GROHMANN	1,3
GRUENE LIGA BRANDENBURG	1,2
MUSIC UNLIMITED	1,2
RIM	1,2
DHL FREIGHT	1,1
BBC	1,1

Tabelle 1: Potentielle Innovationstreiber für das Themenfeld E-Mobilität

1.1.2 Papier 2: Finding Shapes of Diffusion of Innovation Curves by Clustering Time Series of Topic Weights (JLU)

Die Innovationsdiffusionstheorie (Diffusion of Innovation Theory – DOI) wurde 1962 von Rogers entwickelt (Rogers, 2003). Im Allgemeinen erklärt die Theorie die Geschwindigkeit, mit der sich neue Ideen und Technologien in einem sozialen System verbreiten. In Abbildung 5 ist die aus dieser Theorie abgeleitete

schematische Innovationsdiffusionskurve zu sehen. Die Diffusionskurve weist idealtypisch einen glockenförmigen Verlauf auf und lässt sich in insgesamt fünf verschiedene Adoptionsstufen unterteilen.

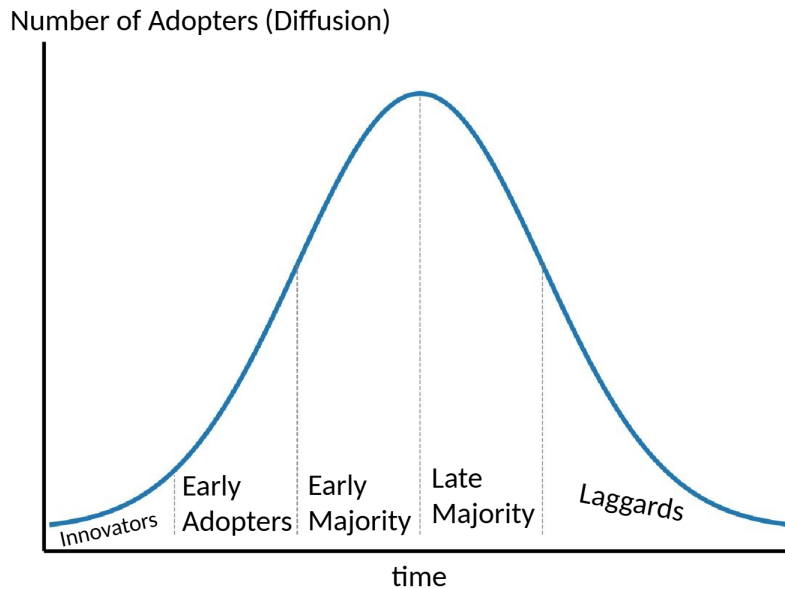


Abbildung 5: Schematische Innovationsdiffusionskurve basierend auf Rogers (1962)

Wir beleuchten diese Theorie mit Hilfe eines datengestützten Ansatzes, der auf den technologiebezogenen Korpus der heise-Zeitschrift, welcher bereits in 1.1.1 präsentiert wurde, basiert.

Durch die in Abschnitt 1.1.1 dargestellte Themenmodellierung erhalten wir verschiedene Themenfelder, teilweise mit und teilweise ohne expliziten Innovationsbezug. Die Diffusion der Berichterstattung über diese Themenfelder lässt sich ebenfalls aus den Themengewichten ableiten, die durch die Aggregation von Themenwahrscheinlichkeiten für die einzelnen Texte im Zeitablauf konstruiert werden.

Unser Ziel ist es, auf der Grundlage einer datengestützten Analyse prototypische Diffusionskurven zu ermitteln, die für Themen mit Innovationsbezug charakteristisch sind. Dafür werden die Themengewichtszeitreihen in einem ersten Schritt aufbereitet und auf monatlicher Basis aggregiert. Danach wenden wir Methoden der funktionellen Datenanalyse an, um sich ähnelnde Diffusionsverläufe in verschiedenen Clustern zu gruppieren. Im Detail basiert der entwickelte

Algorithmus auf der Heuristik der Schwellenwertakzeptanz (Threshold Accepting) in Kombination mit dem so genannten Time-Warping, das berücksichtigt, dass Beginn und Ende der jeweiligen Kurven auf historisch unterschiedliche Zeitpunkte fallen können. Für jedes Cluster lässt sich ein prototypischer Diffusionsverlauf ableiten. In Abbildung 6 sind die Ergebnisse des Zeitreihen-Clustering unter der Annahme von 16 Clustern dargestellt. Sich ähnelnde Diffusionsverläufe werden durch den entwickelten Algorithmus jeweils in einem Cluster zusammengefasst.

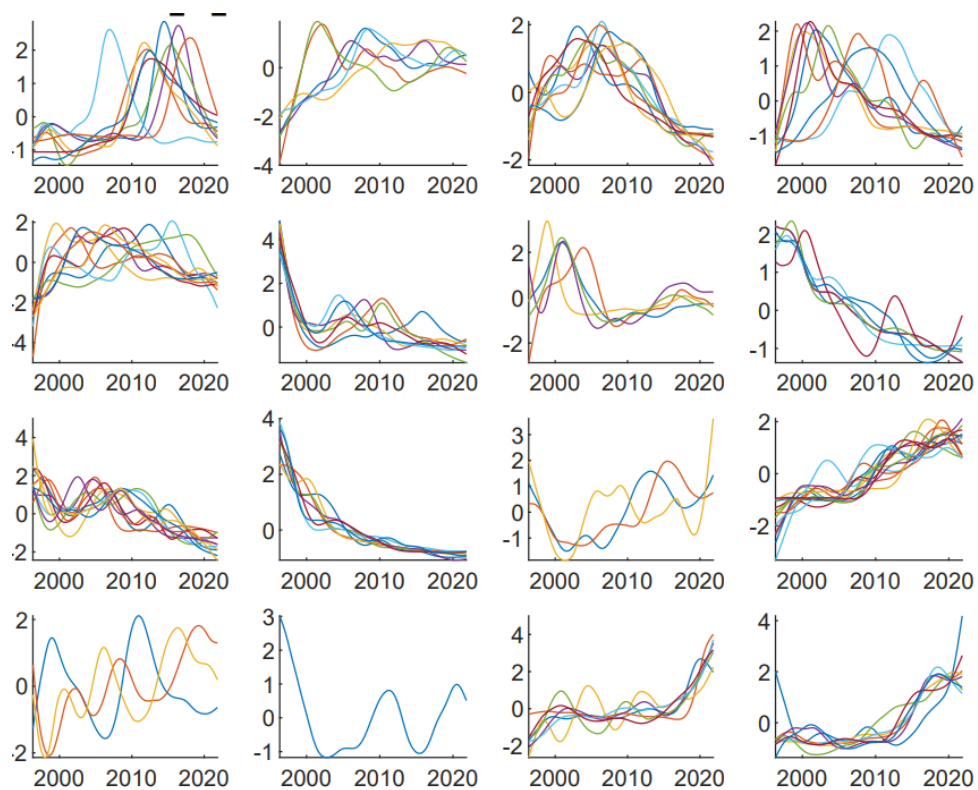


Abbildung 6: Gruppierung der Diffusionsverläufe in 16 Clustern

Als ein Beispiel kann das Cluster in der unteren Reihe rechts betrachtet werden. Darin ist die prototypische Diffusionskurve diejenige für das Themenfeld „E-Mobilität“. Das Cluster scheint generell Verläufe zu Themenfeldern zu zeigen, die sich im Adoptionsstadium „frühe Mehrheit“ (early majority) befinden, was angesichts der Entwicklung der E-Mobilitätsbranche (in der zeitlichen Betrachtung bis 2021) eine plausible Schlussfolgerung darstellt. Das Cluster insgesamt besitzt 8 Mitglieder. Der durchschnittliche Innovationsgrad aller Themenfelder, die diesem Cluster angehören beträgt jedoch nur 25%.

Cluster 6 und 10 (das jeweils zweite Cluster von links in der 2. und 3. Reihe) besitzen ebenfalls jeweils 8 Mitglieder und weisen mit einem durchschnittlichen Innovationsgrad von 63% die höchste Innovationsausprägung auf. Der prototypische Verlauf der Kurve repräsentiert eher die „späte Mehrheit“ (late majority) bzw die „Nachzügler“ (Laggards). In diesen Clustern sind Themenfelder wie Windows-Betriebssysteme, Apple Smartphones und Markenstrategie, Webbrowser (Cluster 6) bzw. Domain-Systeme, I-Mac Computer, Smartphones, Internet und Medienmarkt (Cluster 10) vertreten. Die innovationsbezogenen Themenfelder in diesem Cluster lassen auf eine fast vollkommene Innovationsdiffusion schließen.

Cluster 3 (10 Mitglieder), 4 (9 Mitglieder) und 7 (5 Mitglieder) zeigen vergleichbare Verläufe wie die idealtypische Diffusionskurve nach Rogers (2003). Cluster 3 und 7 haben dabei einen eher geringen Innovationsgrad von 20% und Cluster 4 einen durchschnittlichen Innovationsgrad von 33%. Eine idealtypische Diffusionskurve von Themenfeldern lässt demnach nicht automatisch auf Innovationen schließen. Eine weitere Differenzierung von Diffusionsverläufen z.B. durch die Anreicherung mit relevanten Metadaten ist Gegenstand weiterer Forschungsbestrebungen.

1.1.3 Papier 3: Innovation Panel Surveys in Germany: the Mannheim Innovation Panel (ZEW)

Peters und Rammer (2023) analysieren in diesem Papier, das im Handbook of Innovation Indicators and Measurement erschienen ist, den zeitlichen Verlauf des Innovationsgeschehens über 30 Jahre auf Basis des MIP. Neben einfachen aggregierten Analysen des Innovatorenanteils und des Anteils der Unternehmen Produkt- und Prozessinnovationen (gesamt und nach Branchen und Größenklassen) wurden insbesondere Übergangsraten auf Unternehmensebene analysiert, um die Persistenz der Unternehmen zu untersuchen. Die Analysen zeigen eine hohe Persistenz im Innovationsverhalten auf Basis der surveybasierten MIP-Daten. So sind 84 % aller Produktinnovatoren eines Jahres auch im Folgejahr Produktinnovator. Dies impliziert, dass nur 16 % aller Produktinnovatoren im Folgejahr aus Produktinnovationsaktivitäten aussteigen (Exitrate). Für Prozessinnovatoren liegt die Persistenz mit 79 % ähnlich hoch. Die hohe Persistenz wird in der Literatur damit begründet, dass erfolgreiche Innovatoren die Innovationserlöse zur Finanzierung zukünftiger, riskanter Innovationsprojekte investieren können. Darüber hinaus sind Unternehmen, die erstmalig FuE-Aktivitäten durchführen, mit hohen versunkenen Einstiegskosten konfrontiert, die sie eher dazu bewegen, auch zukünftig in Innovationen zu investieren. Schließlich fördert das angesammelte Wissen aus vergangenen Innovationsaktivitäten die

Entwicklung weiterer Innovationen. Die Innovationstätigkeit einzelner Unternehmen begünstigt wiederum Technologietransfers zu anderen Wirtschaftsakteuren.

Darüber hinaus analysiert das Paper mittels panelökonometrischer Methoden die Auswirkungen kontinuierlichen bzw. wechselnden Innovationsverhaltens auf die Unternehmensperformance (Produktivität). Unternehmen, die kontinuierlich innovativ sind, weisen eine signifikant höhere Produktivität auf als Unternehmen, die kontinuierlich nicht innovativ sind. Die Produktivitätseffekte eines kontinuierlichen Innovationsverhaltens sind bei Produktinnovationen stärker als bei Prozessinnovationen und am höchsten bei Marktneuheiten. Bei Prozessinnovationen weisen Unternehmen mit einem gelegentlichen Innovationsmuster (Wechsel von Innovator zu Nicht-Innovator und zurück) höhere Produktivitätsgewinne auf als anhaltende Prozessinnovatoren. Dies gilt insbesondere für kostensenkende Prozessinnovationen. Diese Ergebnisse implizieren, dass es keine Produktivitätsvorteile durch die kontinuierliche Aktualisierung von Geschäftsprozessen gibt.

1.1.4 Papier 4: Innovation Dynamics using Web-based Innovation Indicators (ZEW)

In diesem Papier, das auf Grund von Verzögerungen in der Datengenerierung derzeit noch in der Fertigstellung ist, wird die zeitliche Entwicklung der webbasierten Innovationsindikatoren für den Zeitraum 2010-2023 für alle Unternehmen des Mannheimer Innovationspanels (MIP), die eine Webseite haben, untersucht. Erstens wird auf aggregierter Ebene die zeitliche Entwicklung des webbasierten Innovationsindikators mit den Entwicklungen verschiedener surveybasierter Innovationsindikatoren aus dem MIP verglichen. Abbildung 7 verdeutlicht, dass der webbasierte Innovationsindikator insgesamt die zeitliche Entwicklung des Innovationsgeschehens, welches sich auf Basis des MIP ergibt, gut reflektiert. Dies gilt insbesondere für den Produktinnovatorenanteil und für den Zeitraum 2010 bis 2020. In diesem Zeitraum liegt der MIP-Produktinnovatorenanteil nur leicht über dem webbasierten Innovatorenanteil und beide zeigen eine ähnliche zeitliche Entwicklung. Dies führt zu einer Korrelation beider Zeitreihen in dem Zeitraum 2010-2020 von hohen 0,82. Niedriger ist dagegen die Korrelation mit dem MIP-Prozessinnovatorenanteil, die bei 0,44 liegt. Dies impliziert, dass der derzeitige webbasierte Innovationsindikator besser geeignet ist, Produktinnovationen zu identifizieren und weniger gut Prozessinnovatoren erkennen kann. Hier sollte der nächste Schritt in der Entwicklung eines speziellen webbasierten Prozessinnovationsindikators liegen. Betrachtet man die Korrelation des webbasierten Innovatorenanteils mit dem MIP-Innovatorenanteil (Produkt- oder Prozessinnovationen) über die Zeit liegt diese bei 0,70. Auffällig

ist, dass nach dem Ausbruch der Corona-Pandemie und mit der Energiekrise 2022 der webbasierte Indikator erstmals größer ist als der Produktinnovatorenanteil im MIP. Dies führt dazu, dass die Korrelation für den Gesamtzeitraum auf 0,54 sinkt. Dieses Ergebnis spricht dafür, das zugrundeliegende Modell zur Berechnung der webbasierten Innovationsindikatoren wenn möglich permanent oder in gewissen Abständen neu zu trainieren, um verändertes Innovationsverhalten besser abbilden zu können.

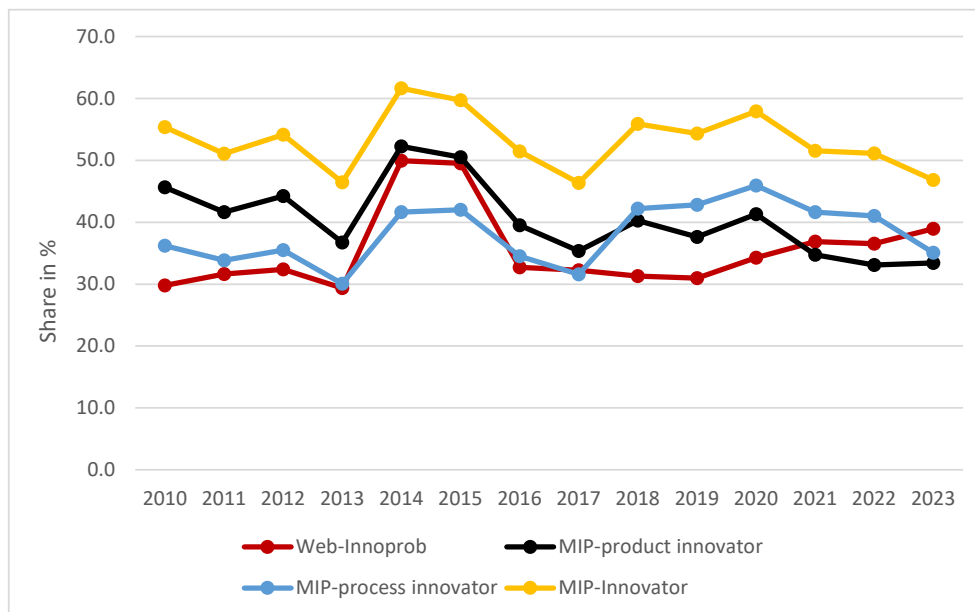


Abbildung 7: Zeitliche Entwicklung der Innovationsanteile

Zweitens untersucht das Papier auf Mikroebene die Persistenz im Produktinnovationsverhalten und im Nicht-Produktinnovationsverhalten deskriptiv mittels Übergangsmatrizen. Diese Analysen erlauben gleichzeitig Aussagen über den Einstieg (Entry) in bzw. den Ausstieg (Exit) aus Produktinnovationsaktivitäten zu treffen. Im Hinblick auf die Güte der webbasierten Innovationsdaten, lassen sich die zentralen Ergebnisse wie folgt zusammenfassen:

Der webbasierte Innovationsindikator weist über den gesamten Zeitverlauf 2010-2023 eine recht hohe Genauigkeit (Accuracy) auf: Zwischen 63 und 71 % der Beobachtungen in den einzelnen Jahren werden korrekt vorhergesagt, im Durchschnitt sind es 68,4 % korrekte Vorhersagen (Abbildung 8). Die Genauig-

keit steigt im Zeitverlauf. Zu den korrekten Vorhersagen tragen die Nicht-Produktinnovatoren stärker bei als die Produktinnovatoren, wobei berücksichtigt werden muss, dass es auch mehr Nicht-Produktinnovatoren als Produktinnovatoren gibt.

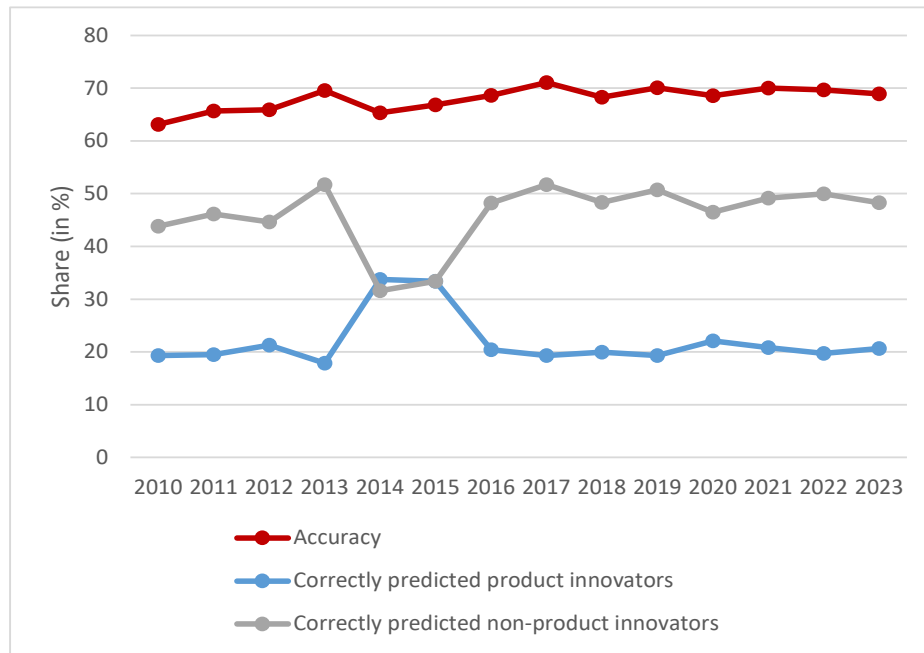


Abbildung 8: Güte des webbasierten Innovationsindikators im Zeitverlauf 2010-2023

Die Precision der Produktinnovatoren, d.h. der Anteil der wahren (=MIP) Produktinnovatoren an allen Unternehmen, die das Modell auf Basis der Webseiten als Produktinnovatoren klassifiziert, liegt bei 61 %. Die Precision der Nicht-Produktinnovatoren, d.h. der Anteil der wahren (=MIP) Nicht-Produktinnovatoren an allen Unternehmen, die das Modell auf Basis der Webseiten als Nicht-Produktinnovatoren klassifiziert, liegt deutlich höher bei 72%. Die Precision ist für beide Produkt- als auch Nicht-Produktinnovatoren relativ stabil im Zeitverlauf, allerdings sehen wir auch hier, dass der webbasierte Indikator seit 2021 weniger gut die wahren Produktinnovatoren vorhersagt, während die Precision für die Nicht-Produktinnovatoren zunimmt. Insgesamt liegen die Werte der Precision aber unterhalb der Werte im Tobi-Projekt. Dort lag die Precision bei rund 80 % für beide Größen.

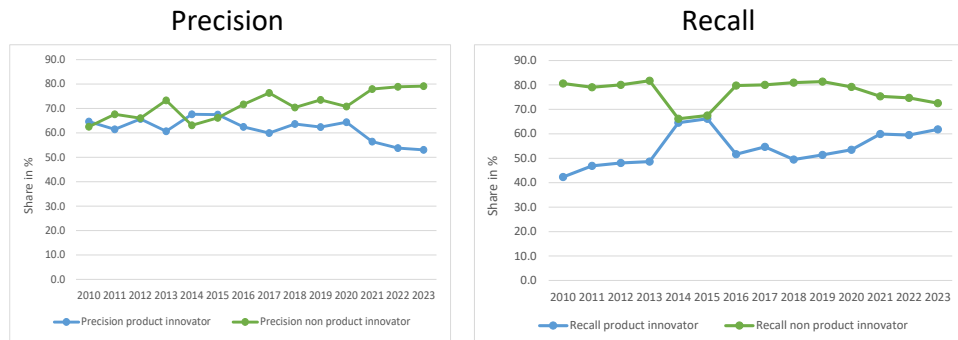


Abbildung 9: Precision und Recall des Produktinnovatorenstatus auf Basis webbasierter Innovationsdaten

Der Recall liegt für Produktinnovatoren bei 54 %. Der Recall wird auch als Sensitivität bezeichnet und misst den Anteil der als Produktinnovator klassifizierten Unternehmen auf Basis der Web-Daten in der Gruppe der wahren Produktinnovatoren, die sich auf Basis der MIP-Daten ergeben. Der Recall für Nicht-Produktinnovatoren liegt wiederum deutlich höher bei 78 %. Der letztere Wert liegt ähnlich hoch wie im Tobi-Projekt. Der geringere Wert des Recalls für Produktinnovatoren war auch bereits im Tobi-Projekt beobachtet worden. Die Ergebnisse können dahingehend interpretiert werden, dass es sehr gut gelingt, Nicht-Produktinnovatoren mit den webbasierten Innovationsindikatoren zu identifizieren. Die Identifikation der Produktinnovatoren ist dagegen generell schwieriger. Es zeigt sich jedoch im zeitlichen Verlauf, dass der Recall der Produktinnovatoren deutlich zugenommen hat von 42 % im Jahr 2010 auf 62 % im Jahr 2023.

Es zeigt sich eine hohe Persistenz des Produktinnovations-Status und des Nicht-Produktinnovations-Status sowohl im MIP als auch mit den webbasierten Daten über den Zeitverlauf. Mit dem MIP-Produktinnovationsindikator finden wir, dass 30 % aller Unternehmen im MIP sowohl Produktinnovator im Jahr t als auch im Jahr $t+1$ sind. Auf Basis der webbasierten Daten liegt der entsprechende Anteil bei 31 %. Für die Nicht-Produktinnovatoren ist die Persistenz etwas höher mit den webbasierten Daten. 59 % aller Unternehmen sind persistente Nicht-Produktinnovatoren, während dieser Anteil mit den surveybasierten MIP-Daten bei 50 % liegt.

Dies bedeutet im Umkehrschluss, dass sowohl Entry als auch Exit in Produktinnovationsaktivitäten mit den MIP-Daten und mit den webbasierten Innovationsdaten beobachtet wird. Allerdings sind Entry und Exit häufiger zu beobachtende Ereignisse im MIP (20 % aller Beobachtungen) als in den webbasierten

Daten (11 % aller Beobachtungen). Sowohl die Entry- als auch die Exit-Rate liegen im MIP etwa doppelt so hoch wie auf Basis des webbasierten Indikators.

Konkret bedeutet dies, dass die 1-Jahres-Entry-Rate im MIP 16 % beträgt, während sie auf Basis des webbasierten Indikators bei 9 % liegt. D.h. im MIP werden 16 % aller Nicht-Produktinnovatoren in Jahr t im Jahr $t+1$ zu Produktinnovatoren, während wir 9 % der webbasierten Nicht-Produktinnovatoren im Jahr t als Produktinnovatoren im Jahr $t+1$ identifizieren.

Ein Ausstieg aus Produktinnovationsaktivitäten ist häufiger zu beobachten als ein Einstieg in diese Aktivität. Im MIP hören 27 % der Produktinnovatoren im Jahr t im Jahr $t+1$ mit Produktinnovationsaktivitäten auf. Mit den webbasierten Daten liegt die Exit-Rate bei 13 %.

Berücksichtigt man die Tatsache, dass der Produktinnovationsindikator im MIP ein 3-Jahres-Indikator ist, und betrachtet statt der 1-Jahres-Übergangswahrscheinlichkeiten stattdessen 2- oder 3-Jahres-Übergangswahrscheinlichkeiten, um eine Überlappung des Indikators zu vermeiden, dann zeigt sich, dass die Entry-Rate im MIP bei 2- oder 3-Jahres-Übergangswahrscheinlichkeiten leicht steigt auf 19 % bzw. 20 %. Auf Basis des webbasierten Indikators steigt die Entry-Rate dagegen deutlicher von 9 % auf 13 % bei 2-Jahresübergängen und auf über 16 % bei 3-Jahresübergängen an. Sie kommt damit der MIP-Entry-Rate recht nahe.

Eine Annäherung der Exit-Rate auf Basis der webbasierten Daten an die der surveybasierten MIP-Daten kann man ebenfalls beobachten, wenngleich in abgeschwächter Form. So steigt im MIP die Exit-Rate im Zeitverlauf an, d.h. von 27 % bei 1-Jahresübergängen auf 30 % bei 2-Jahresübergängen und auf 33 % bei 3-Jahresübergängen. Auf Basis des webbasierten Innovationsindikators identifizieren wir zwar ebenfalls einen Anstieg von 13 % auf 17 % bzw. 19 %. Die Exit-Rate bleibt aber fast halb so groß wie im MIP. Insgesamt zeigen diese Analysen, dass es schwieriger ist, den Wechsel des Status von Produkt- zu Nicht-Produktinnovator mit webbasierten Innovationsindikatoren vorherzusagen.

Drittens untersuchen wir jeweils die Dynamik webbasierter und surveybasierter Innovationsindikatoren mittels ökonometrischer Paneldatenmodelle und vergleichen, inwieweit der Innovationsstatus in beiden Datensätzen durch die gleichen Faktoren erklärt werden kann. Die Ergebnisse weisen in hier auf eine hohe Übereinstimmung der Faktoren hin, die den Innovationsstatus in beiden Datensätzen erklären. Den stärksten Einfluss hat jeweils der vergangene Innovationsstatus, dies weist auf eine hohe Pfadabhängigkeit im Innovationsverhalten hin. Daneben beeinflussen die Unternehmensgröße, das Unternehmensalter, Finanzierungsmöglichkeiten (Bonität), Branche, Rechtsform, Region und das Jahr signifikant den webbasierten und surveybasierten Innovationsstatus.

Insgesamt zeigen die Analysen in diesem Papier, dass das entwickelte Modell zur Identifikation von Produktinnovatoren auf Basis von textbasierten Webdaten auch im Zeitverlauf geeignet ist, um Produktinnovatoren zu identifizieren und somit ein nützliches Instrument darstellen, um aktuelle Trends in Produkt-Innovationsaktivitäten auch im Zeitverlauf abzubilden. Dies gilt trotz gewisser Schwierigkeiten, Produktinnovatoren mit webbasierten Daten zu identifizieren und dann in der Folge auch den Exit von Produktinnovationsaktivitäten korrekt vorherzusagen.

1.2 Technische Dokumentation

Neben den wissenschaftlichen Papieren haben wir eine technische Dokumentation (Meilenstein AP 4) verfasst, die einerseits die Systemarchitektur des Archive Scrapers und seine Komponenten beschreibt, die Konfiguration, Anpassung und Skalierbarkeit des Archive Scrapers erläutert und mögliche zukünftige Weiterentwicklungen aufzeigt. Andererseits enthält die technische Dokumentation eine Beschreibung zur Implementation und Kostenabschätzung, wie ein permanentes Nah-Echtzeit-Innovations-Monitoring-System für Deutschland aufgebaut werden kann.

1.3 Präsentationen auf wissenschaftlichen Konferenzen und Workshops

Die Ergebnisse des DynTOBI-Projekts wurden auf folgenden internationalen Fachkonferenzen und Seminaren zur Präsentation angenommen.

1. **24th International Conference on Computational Statistics (COMPSTAT 2022)**, Bologna, Italien, 2022
2. **CSDA & EcoSta Workshop on Statistical Data Science (SDS 2022)**, Bologna, Italien, 2022
3. **Doktorandenseminar Bergische Universität Wuppertal**, Wuppertal, Deutschland, 2023
4. **HiTEc Spring Course**, Limassol, Zypern, 2023
5. **Sixth International and Interdisciplinary Conference on the Quantitative and Computational Analysis of Textual Data (COMPTEXT 2024)**, Amsterdam, Niederlande, 2024
6. **26th International Conference on Computational Statistics (COMPSTAT 2024)**, Gießen, Deutschland, 2024
7. **Jahrestagung Verein für Socialpolitik (VfS 2024)**, Berlin, Deutschland, 2024

1.4 Öffentlichkeitsarbeit

Wir waren von Beginn des Projekts an darum bemüht, neben dem Fachpublikum auch eine breitere Öffentlichkeit zu erreichen, was insbesondere über die Sichtbarkeit der Ausgründung aus dem Vorgängerprojekt istari.ai gelungen ist.

1.5 Abgeschlossene Dissertationen im Rahmen des DynTOBI Projekts

Im Rahmen des DynTOBI Projekts konnte die am Projekt beteiligte Doktorandin Albina Latifi wesentliche Teile der Forschungsarbeiten in ihrer Dissertation integrieren. Albina Latifis Dissertation mit dem Arbeitstitel „Supporting data-driven policy making: Combining Natural Language Processing and Time Series Analysis Methods“ soll im Jahr 2025 an der Justus-Liebig-Universität Gießen eingereicht werden und basiert in Teilen auf die Forschungsarbeiten im Rahmen des DynTOBI Projekts.

Es ist außerdem vorgesehen, dass Leon Steines, der nach dem Weggang von Julian Dörr im Juni 2023 zum ZEW-Projektteam gestoßen ist, das in Abschnitt 1.1.4 vorgestellte Papier als Teil seiner Dissertation einbringen wird. Die Dissertation wird an der TU München von Prof. Dr. Hanna Hottenrott betreut und wird voraussichtlich in der ersten Hälfte 2027 abgeschlossen.

1.6 Externe Entwicklung des Forschungsfelds

Im Teilvorhaben der JLU, welches sich auf die namentliche Identifikation von potentiellen Innovationstreibern und prototypischen Innovationsdiffusionsverläufen auf Basis von Nachrichtenartikel fokussiert hat, sind kaum externe Entwicklungen ähnlicher Bestrebungen bekannt.

Die einzige verwandte Studie, die die namentliche Identifikation von Innovationstreibern beabsichtigt, ist Chen u. a. (2021). Sie wenden einen ähnlichen Ansatz an, um potentielle Innovatoren in der Pharmaindustrie aufzudecken. Hierbei wenden die Autoren u.a. auch große Sprachmodelle an (Large Language Models).

Während der Projektlaufzeit wurden weitere webbasierte Indikatoren entwickelt, z.B. zur KI-Nutzung (Dahlke et al. 2023) und zur Nachhaltigkeit (Schmidt et al. 2022). Andere Forschung zur Dynamik von webbasierten Innovationsindikatoren gab es weder zu Projektbeginn noch zum Projektende.

Generell findet jedoch eine rasante Weiterentwicklung der Methoden der natürlichen Sprachverarbeitung statt. Seit der Einführung des ersten weit bekannten großen Sprachmodells, des „Bidirectional Encoding Representations from

Transformers“ (BERT) von Devlin u. a. (2018) im Jahr 2018, haben große Sprachmodelle viel Aufmerksamkeit erhalten und werden stetig weiterentwickelt.

2 Folgeverwertung

2.1 Startup istari.ai

Das aus dem Vorgängerprojekt TOBI hervorgegangene Startup istari.ai ist mittlerweile als GmbH etabliert und weiterhin auf Wachstumskurs. Darüber werden die in den Projekten TOBI und DynTOBI entwickelten Ansätze auch einem breiteren Kreis von Nutzern aus den Bereichen Wirtschaft, öffentliche Einrichtungen und Forschungsinstitutionen zugänglich. Zugleich werden in diesem Rahmen Weiterentwicklungen mit Blick auf die Bedürfnisse der jeweiligen Kundengruppen vorgenommen.

2.2 Folgeprojekte

Folgende Forschungsprojekte stehen mit dem DynTOBI-Projekt im Zusammenhang:

(1) Integration von Text Mining Verfahren mit multivariater Zeitreihenanalyse (TEXTMOD)

In diesem für den Zeitraum 2020 – 2022 im Rahmen der Kooperation von DFG und NCN geförderten deutsch-polnisches Kooperationsprojekt wurden mehrsprachige Ansätze genutzt, um textbasierte Indikatoren mit realen ökonomischen Entwicklungen in Beziehung zu setzen.

Neben dieser inhaltlichen Fragestellung wurden in dem Projekt gezielt Fragen nach der Robustheit der verwandten Methodik adressiert. Unter anderem konnte ein neues Kriterium für die Wahl der Anzahl von Themen in der Themenmodellierung etabliert werden, das im Rahmen einer Monte Carlo Studie gegenüber den bislang benutzten Kriterien vorteilhaft erscheint. Außerdem wurden Fragen der Datenaufbereitung (Entfernung seltener Worte) und der Darstellung der Schätzunsicherheit in neuen Varianten der üblicherweise verwendeten Wortwolken adressiert.

Durch den Austausch der Projektmitarbeiterinnen von DynTOBI und TEXTMOD ergaben sich eine Reihe von Synergien.

(2) Makroökonomische Wirkungen der Fiskalpolitik: Neue Evidenz aus der textanalytischen Auswertung der Debatten im Deutschen Bundestag (MaFiText)

In diesem 2021 – 2024 von der DFG geförderten Projekt wurden Methoden der automatisierten Textverarbeitung auf die Reden im Deutschen

Bundestag und Zeitungsbeiträge mit Bezug zu fiskalpolitischen Themen entwickelt. Das Ziel bestand in der Generierung von Indikatoren zum fiskalpolitischen Sentiment der Redebeiträge (also ob diese eher expansive oder restriktive Argumente vortrugen). Diese Indikatoren konnten durch die Zuordnung der Entitäten auf Parteebene aggregiert werden, so dass entsprechende Stimmungsindikatoren für den Bundestag insgesamt und einzelne Parteien ausgewiesen werden konnten. Durch die Einbindung der Indikatoren in makroökonomische Modelle konnte gezeigt werden, dass Schocks im fiskalpolitischen Sentiment der Bundestagsreden Wirkungen in der realen Ökonomie entfalten.

In einer Variante des Modells wurde untersucht, wie stark sich das fiskalpolitische Sentiment zwischen Regierung und Opposition einerseits und zwischen den Regierungsparteien andererseits unterscheidet. Es zeigt sich, dass insbesondere eine Uneinigkeit innerhalb der Regierungskoalition negative Wachstumswirkungen aufweist, die neben dem normalen Budgetkanal wirken.

Über die Ergebnisse dieses Projektes wurde u.a. in der Frankfurter Allgemeinen Zeitung und der Süddeutschen Zeitung berichtet.

Bei der Bearbeitung dieses Projektes ergaben sich ebenfalls Synergien zum DynTOBI Projekt, in dem einzelne technische Lösungen im jeweils anderen Projekt übernommen werden konnten.

- (3)** Der am ZEW generierte Paneldatensatz soll anderen Forschern im Rahmen des Forschungsdatenzentrums am ZEW für weitere Forschungen zur Verfügung gestellt werden. So plant die TU München die Daten im Projekt Hidden Regions, welches durch die Dieter Schwarz Stiftung gefördert wird, einzusetzen um regionale webbasierte Innovationsindikatoren abzubilden.
- (4)** Neben der weiteren Nutzung des neuen Paneldatensatzes für Forschungszwecke haben wir in der technischen Dokumentation ein Konzept zur Implementation und Kostenabschätzung entwickelt, wie ein permanentes Nah-Echtzeit-Innovations-Monitoring-System für Deutschland aufgebaut werden kann. Aus diesem Konzept kann bei entsprechender Förderung ein Folgeprojekt entstehen.

3 Literaturliste

Carree, M. und Thurik, R. (2010). The impact of entrepreneurship on economic growth, in Z. J. Acs and D. B. Audretsch (eds), *Handbook of Entrepreneurship Research*, Vol. 1, Springer New York, NY, pp. 557–594.

Chen, K., Cosgro, B., Domfeh, O., Stern, A., Korkmaz, G., Kattampallil, N. A. (2021). Leveraging google bert to detect and measure innovation discussed in news articles, *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 1–6.

Dahlke, J., Beck, M, Kinne, J., Lenz, D., Dehghan, R., Wörter, M., Ebersberger, B. (2023), Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption, *Research Policy* 53 (2), 104917.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*

Kinne, J. und Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study, *Scientometrics* 125: 2011 – 2041.

Latifi, A., Lenz, D., Winker, P. (2025a). Identification of innovation drivers based on technology-related news articles. *Journal of Open Innovation: Technology, Market, and Complexity* 11 (1), 100475.

Latifi, A., Lenz, D., Winker, P. (2025a). Finding shapes of diffusion of innovation curves by clustering time series of topic weights. *MAGKS Discussion Papier*. Forthcoming.

Peters, B., Ehlich, J., Steines, L., Kinne, J. (2025), Innovation Dynamics using Web-based Innovation Indicators, mimeo.

Peters, B., Kinne, J., Lenz, D. (2024), Dynamische Textdaten-basierte Output-Indikatoren als Basis einer neuen Innovationsmetrik - DynTOBI – Technische Dokumentation, Mannheim.

Peters, B., Rammer, C. (2023), Innovation panel surveys in Germany: the Mannheim Innovation Panel, in: Gault, F., Arundel, A., Kraemer-Mbula, E., Handbook

of Innovation Indicators and Measurement, Second Edition, Edward Elgar Publishing, Cheltenham, Northampton, 54-87

Rogers, E. (2003). The diffusion of innovations. Fifth Edition. *The Free Press, New York*.

Schmidt, S., Kinne, J., Lautenbach, S., Blaschke, T., Lenz, D., Resch, B. (2022), Greenwashing in the US metal industry? A novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining, *Science of The Total Environment*, 835,155512.