Textdaten-basierte Output-Indikatoren als Basis einer neuen Innovationsmetrik (TOBI)

Ziel des Projektes ist es, Innovationsindikatoren durch die automatisierte und regelmäßige Auswertung digitaler Text-Massendaten zu generieren. Im Teilvorhaben des ZEW basieren die Innovationsindikatoren auf der Analyse von Unternehmenswebseiten, während im Teilvorhaben der JLU News-Artikel aus IT-Fachzeitschriften analysiert werden. Die neuartigen Innovationsindikatoren werden mit Hilfe von bereits etablierten Referenzindikatoren kalibriert und validiert. Dadurch sollen Innovationsindikatoren von sehr hoher Aktualität und in feiner regionaler und sektoraler Gliederung generiert werden.

Der Ausgangspunkt des Teilvorhabens am ZEW ist es, dass die Aktualisierung einer Unternehmenswebseite in Verbindung mit Innovationen stehen kann: Es werden hier z.B. Informationen zu neuen Produkten oder Dienstleistungen veröffentlicht und beworben. Mithilfe von computerlinguistischen Methoden (Text Data Mining) werden diese Veränderungen vom ZEW identifiziert und nach ihrem Innovationsbezug sowie der Art der Innovation klassifiziert. Die Textanalyse kann dabei durch zusätzliche Informationen zu Unternehmensmerkmalen wie Geschäftstätigkeit, Größe, Organisationsform, Unternehmensalter und Standort optimiert werden.

Dem ZEW liegen hierfür Webseiten-Adressen sowie Informationen zu Unternehmensmerkmalen von rund 1,3 Mio. Unternehmen in Deutschland aus dem Mannheimer Unternehmenspanel (MUP) vor. Dadurch ist es uns möglich, eine neue Form von Innovationsstatistik auf Unternehmensebene zu etablieren, die Unternehmen der gewerblichen Wirtschaft großflächig abbildet. Um die Validität der gewonnenen Innovationsindikatoren zu überprüfen und den Textanalysealgorithmus zu kalibrieren, werden die Ergebnisse der Webseitenanalyse mit den Innovationsindikatoren des Mannheimer Innovationspanels (MIP) und anderen Indikatoren (z. B. aus mit dem MUP zusammengeführte Patentdatenbanken) verglichen. Das MIP dient jedoch nicht nur zum Vergleich der Ergebnisse. Das Teilvorhaben plant über die Metrik des MIP hinauszugehen und diese in Hinblick auf Aktualität, Genauigkeit der sektoralen und regionalen Gliederung und insbesondere der „fachlichen“ und „sachlichen“ Tiefe zu ergänzen.

Vorgehensweise: Im ersten Schritt wird ein Webscraper entwickelt, welcher die Textinhalte der Unternehmenswebseiten extrahiert. Durch regelmäßiges Webscraping soll eine Panel-Textdatenbank aufgebaut werden. Anschließend werden mittels computerlinguistischen Methoden aus den gespeicherten Texten Innovationsindikatoren auf Unternehmensebene abgeleitet. Unter anderem soll hierfür ein LDA-Modell (*Latent Dirichlet Allocation)* verwendet werden. Dieses identifiziert sogenannte *latente Topics,* welche auf Unternehmenswebseiten vorkommen. Gemeinsam auftretende Worte werden gruppiert bzw. geclustert und hierdurch in Topics mit Bezug zu Unternehmensaktivitäten sortiert. Das kontinuierliche Monitoring von Topics mit spezifischem Innovationsbezug erlaubt das Tracking von Innovationen, also deren erstmaliges Auftreten in Unternehmen und anschließende Verbreitung über Unternehmen, Branchen und Regionen hinweg.

An der JLU werden die News-Artikel mittels Topic-Modeling in Hinblick auf relevante, d.h. innovationsnahe, Themen analysiert. Die so identifizierten Innovationen werden, bspw. mittels *Named-Entity Recognition* (NER) Methoden, auf die handelnden Entitäten (Unternehmen) untersucht, welche wiederum mit Angaben aus dem MUP und anderen, mit dem MUP verbundene, Datenbanken verknüpft werden. Aus den gesammelten Daten können schließlich

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Innovationsmetriken extrahiert werden, die unter Benutzung *Vektor-Autoregressiver* (VAR) Modelle auf ihre Vor- bzw. Nachlaufeigenschaften untersucht werden können. Abgesehen von der Identifikation von Vorlaufeigenschaften, die für ein Indikatorensystem als hilfreich eingeschätzt werden, können auch Indikatoren identifiziert werden, die zusätzliche Informationen liefern, die bislang nicht im Indikatorensystem enthalten sind und somit eine relevante Erweiterung der Innovationsmetrik darstellen. Das permanente Monitoring der aktuellsten News-Artikel erlaubt dabei eine hohe Aktualität der generierten Indikatoren.

Neben methodischen Überschneidungen (Innovationsindikatoren aus Text-Massendaten mittels Text Data Mining) können die beiden Teilvorhaben auch dahingehend voneinander profitieren, dass beispielsweise die trainierten Topic-Modelle des jeweils anderen Teilvorhabens zur Klassifizierung der eigenen Textdaten herangezogen werden. So lassen sich beispielsweise Innovationsfelder, die in IT-Fachzeitschriften identifiziert wurden, auch auf den Webseiten von IT-Unternehmen erkennen.

Publikationen

Kinne, Jan und Janna Axenbeck (2020), Web Mining for Innovation Ecosystem Mapping: A Framework and a Large-scale Pilot Study, Scientometrics. Link

Mirtsch, Mona, Jan Kinne und Knut Blind (2020), Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis, IEEE Transactions on Engineering Management. Link

Kinne, Jan, Miriam Krüger, David Lenz, Georg Licht und Peter Winker (2020), Corona-Pandemie betrifft Unternehmen unterschiedlich, Tagesaktuelle Webseiten-Analyse zur Reaktion von Unternehmen auf die Corona-Pandemie in Deutschland, ZEW-Kurzexpertise Nr. 20-05, Mannheim. Download

Krüger, Miriam, Jan Kinne, David Lenz und Bernd Resch (2020), The Digital Layer: How Innovative Firms Relate on the Web, ZEW Discussion Paper No. 20-003, Mannheim. Download

Kinne, Jan und David Lenz (2019), Predicting Innovative Firms Using Web Mining and Deep Learning, ZEW Discussion Paper No. 19-001, Mannheim. Download

D. Lenz, P. Winker (2020), "Measuring the Diffusion of Innovations with Paragraph Vector Topic Models" PLOS ONE. 2020;15(1):1-18. Link

D. Eugenidis, D. Lenz, C. Leser, F. Schleer-van Gellecom und P. Winker (2020), "Text-mining basierte Analyse der Kapitalmarktreaktionen auf Ad-hoc-Mitteilungen" CORPORATE FINANCE, 2020, 09-10. Link

Axenbeck, Janna und Patrick Breithaupt (2019), Web-Based Innovation Indicators – Which Firm Website Characteristics Relate to Firm-Level Innovation Activity?, ZEW Discussion Paper No. 19-063, Mannheim. Download

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
ZEW Zentrum für Europäische Wirtschaftsforschung GmbH

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung