



Mittelstand 4.0
Kompetenzzentrum
Rostock



PRAXISBEISPIEL

Machine Learning zur Qualitätssicherung

Langfristige Qualitätssicherung der Plattform-Inhalte mittels Datenanalyse und
Automatisierung von Kategoriezuweisungen durch Text Mining

Mittelstand-
Digital



Gefördert durch:



Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages



Abb. 1: Screenshot der Endkunden-Webseite www.auf-nach-mv.de des TMV e.V.

Problemstellung

Durch die zunehmende Digitalisierung werden immer mehr Daten erfasst und gespeichert. Dabei gilt: Je höher die Qualität ist, desto wertvoller sind diese Daten und die daraus resultierenden Analyseergebnisse.

In einem Pilotprojekt wurde gemeinsam mit dem Tourismusverband Mecklenburg-Vorpommern e.V. die Datenbank der *Point of Interests* (Abk. POIs; hier: touristisch interessante Orte) und Veranstaltungen für das gesamte Bundesland überprüft und analysiert. Ziel dieser Untersuchung war die Bewertung möglicher Automatisierungen, um auf diesem Weg eine Arbeitserleichterung im Bereich Qualitätssicherung und -entwicklung der POI-Daten zu erzielen.

Das Fundament des Projektes war die Einschätzung zur Qualität der Dateninhalte. Darauf aufbauend soll den Regional- und Branchenverbänden sowie den touristischen Unternehmen Mecklenburg-Vorpommerns die Möglichkeit aufgezeigt werden, die Datenqualität zu den touristischen Angeboten im Land durch Machine Learning auf langfristige Sicht zu sichern und zu erhöhen.

Allgemeine Vorgehensweise

Vor der Anwendung von Machine Learning Verfahren ist es zunächst in einem **ersten Schritt** erforderlich, die zu untersuchenden Daten mit Hilfe von deskriptiven Methoden zu überprüfen, sodass die erforderliche gute Datenqualität gewährleistet werden kann.

Machine Learning

Bei Machine Learning Anwendungen sucht ein Computeralgorithmus in Trainingsdaten nach Strukturen und Mustern. Anschließend werden diese Erkenntnisse automatisiert auf die zu analysierenden Daten übertragen und ermöglichen auf diese Weise die Problemlösung.

Anschließend wird in einem **zweiten Schritt** das interessierende Zielmerkmal der einzelnen Stammdaten (in diesem Projekt die POI-Kategorie, also welche Art von POI vorliegt) betrachtet. Es wird bewertet, ob und mit welchem Machine Learning Verfahren die Zuweisung automatisiert geschehen kann und gegebenenfalls Restriktionen zur Realisierung bestimmt.

In einem **dritten Schritt** wird das angepasste Verfahren bewertet und die weitere Vorgehensweise gemeinsam mit dem Unternehmen geplant, sodass zukünftig die Kategorien automatisiert zugeordnet werden können. Aufgrund der gewonnenen Erkenntnisse lässt sich dauerhaft die Qualität der Datenbankeinträge verbessern.

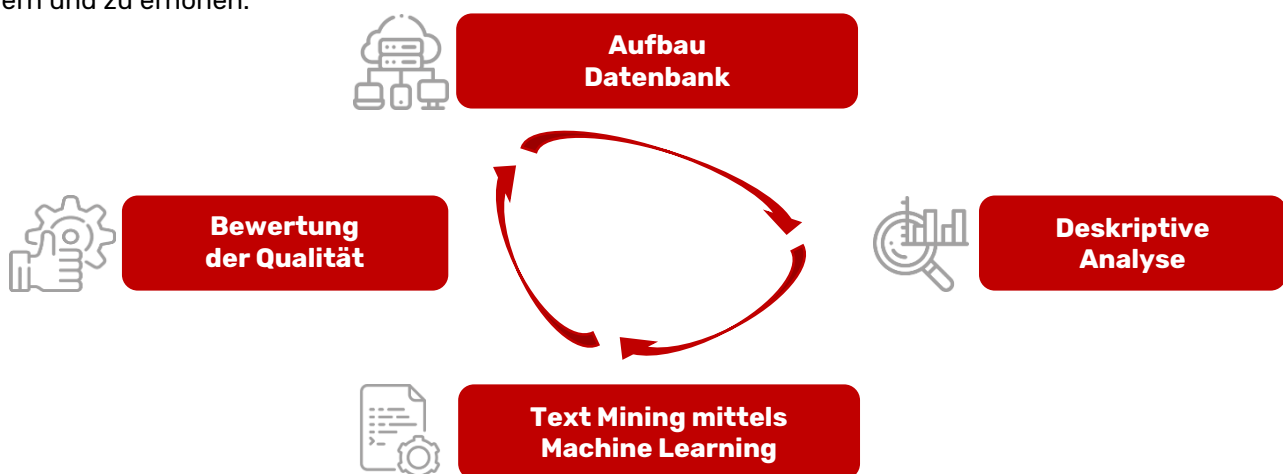


Abb. 2: Vorgehensweise zur Qualitätssicherung von Datenbankinhalten

Die Bewertung der Datenqualität und die Entscheidung über den möglichen Einsatz von Machine Learning Methoden zur automatisierten Kategorienzuweisung der POI-Stammdaten erfolgte mit Hilfe der *Software R*. Diese Open Source Statistiksoftware ist eine kostenlose Programmiersprache und -umgebung (unter der „GNU S“-Lizenz) mit einer großen Community und eine der derzeit verbreitetsten Statistik-Programmierungsumgebungen mit diversen fachspezifischen Erweiterungen.

Anschließend erfolgte die Untersuchung der zu bestimmenden Kategorien (z.B. Aktivität, Herkunft, Kultur, ...) im Hinblick auf Unterschiede in den POI-Beschreibungen. Dabei wurden für die einzelnen Kategorien jeweils Signalwörter ermittelt und auf Eindeutigkeit geprüft. In der in Abb. 3 exemplarisch dargestellten Wordcloud sind die für die Kategorie *Aktivität* charakteristischsten Wörter zu erkennen. Weiterhin wurden die für die Zuordnung wichtigen Einflussvariablen bestimmt.

Auf Grundlage der Analyse erfolgte die Diskussion verschiedener Strategien für die Zukunft.

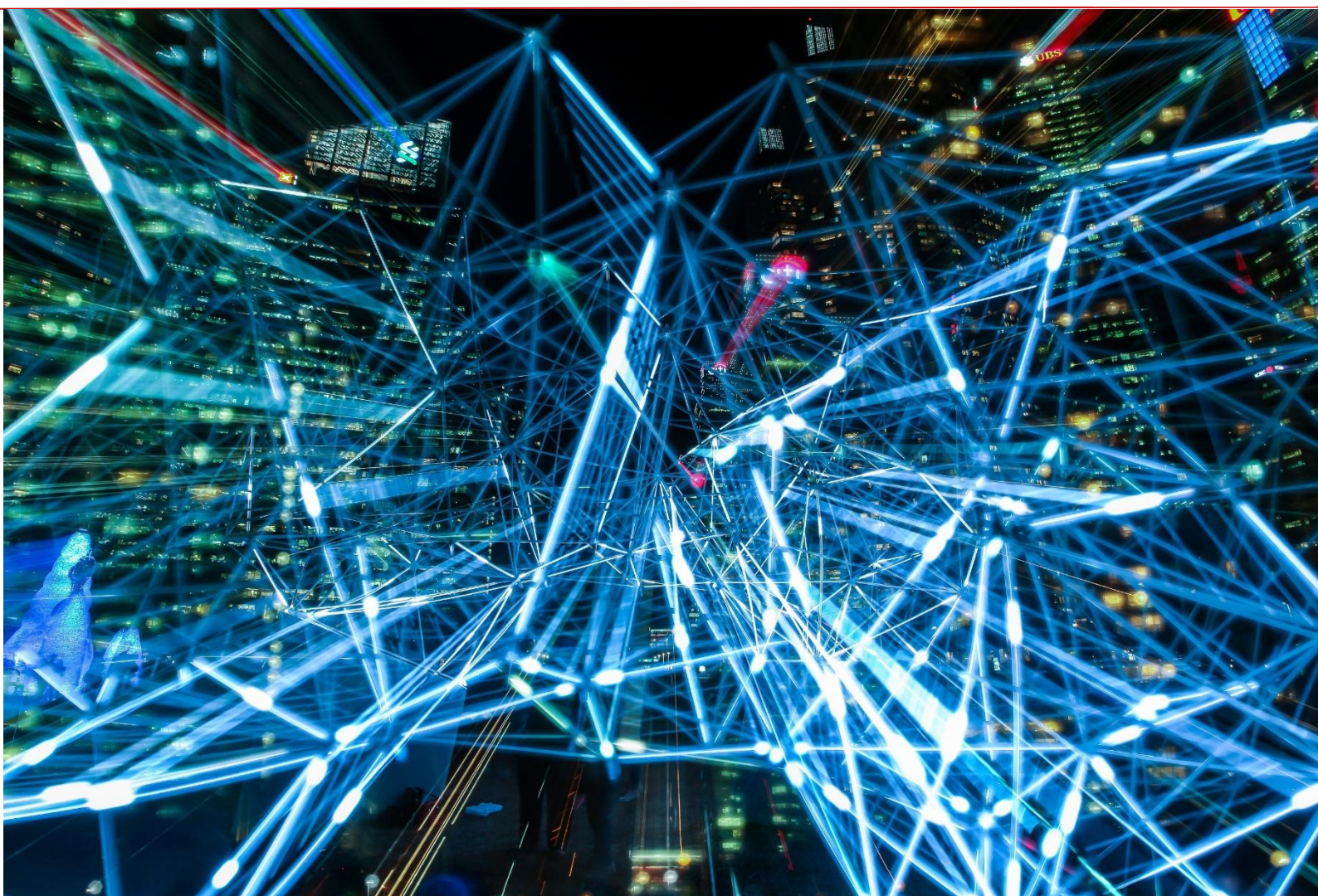
Beim Machine Learning werden Verfahren mit Hilfe von Trainingsdaten angepasst. Um die Qualität des Modells zu bewerten, wird anschließend das angepasste Modell auf neue (also nicht in das Modell eingegangene) Validierungsdaten angewendet und der Output des Modells mit dem gewünschten Resultat verglichen. So kann die sogenannte Güte des Modells bewertet werden.



Die Analyse der POI-Datenbankinhalte zeigte das Potential einer automatisierten Verarbeitung auf. Weiterhin konnten durch Analyseergebnisse konkrete Ansatzpunkte zur Steigerung der Datenqualität ermittelt werden.

Abb. 4: Screenshot der Suchmaske nach POIs (Webseite: www.auf-nach-mv.de)





Was ist Mittelstand-Digital?

Mittelstand-Digital informiert kleine und mittlere Unternehmen über die Chancen und Herausforderungen der Digitalisierung. Die geförderten Kompetenzzentren helfen mit Expertenwissen, Demonstrationszentren, Best-Practice-Beispielen sowie Netzwerken, die dem Erfahrungsaustausch dienen. Das Bundesministerium für Wirtschaft und Energie (BMWi) ermöglicht die kostenfreie Nutzung aller Angebote von Mittelstand-Digital.

Der DLR Projektträger begleitet im Auftrag des BMWi die Projekte fachlich und sorgt für eine bedarfs- und mittelstandsgerechte Umsetzung der Angebote. Das wissenschaftliche Institut für Infrastruktur und Kommunikationsdienste (WIK) unterstützt mit wissenschaftlicher Begleitung, Vernetzung und Öffentlichkeitsarbeit.

Weitere Informationen finden Sie unter:
www.mittelstand-digital.de

Impressum

Herausgeber:
Mittelstand 4.0-Kompetenzzentrum Rostock
c/o Hochschule Stralsund
Zur Schwedenschanze 15
18435 Stralsund

Verantwortlich: Prof. Dr. Gero Szepannek
Gestaltung: Rabea Aschenbruck
Bildnachweise: Aschenbruck, www.auf-nach-mv.de,
Icons made by Freepik from www.flaticon.com
Stand: August 2019
www.kompetenzzentrum-rostock.digital