TU Clausthal

Kolloquium zur Masterarbeit

Philipp Haan, B.Sc.

„Employing natural language processing to discover structured information from unstructured texts for market research"

In einem immer schnelllebigeren Marktumfeld ist eine genaue Beobachtung des Marktes, insbesondere der Veränderungen im Technologie-, Kunden- und Konkurrenzumfeld von besonderer Wichtigkeit. Durch die zunehmende Darstellung aller Marktakteure im Internet auf Social-Media-Kanälen, Presseportalen oder domänenspezifischen Online-Magazinen hat sich die zur Verfügung stehende Informationsmenge in den letzten Jahren exponentiell vergrößert.
Die klassischen Methoden der Marktbeobachtung sind jedoch nur bedingt in der Lage, diese permanent wachsende Menge an unstrukturierten Textdaten zu erfassen, zu analysieren und zur Entscheidungsfindung aufzubereiten.
Natural Language Processing (NLP) ist ein Teilgebiet des maschinellen Lernens, das sich mit der maschinellen Verarbeitung menschlicher Sprache beschäftigt. Ziel dieser Arbeit ist es herauszufinden, ob es mithilfe von Methoden des Natural-Language-Processing (NLP) möglich ist, automatisiert strukturierte Informationen aus diesen unstrukturierten Textdaten mit einer solchen Genauigkeit zu extrahieren, dass hierdurch eine Entscheidungsfindung ermöglicht werden kann.
Hierfür wurde ein Prozess entwickelt, der es ermöglicht, auf Basis konkreter, branchenspezifische Fragestellungen Textstellen zu identifizieren und die zur Beantwortung benötigten Informationen zu extrahieren. Die Arbeit führt von der Identifizierung der relevanten Textstellen unter Verwendung von Methoden der automatischen Sprachanalyse und Named-Entity-Recognition (NER) über die Extraktion der fragenrelevanten Antworten mithilfe von Transformer-Modellen hin zur Ergebnisdarstellung. In diesem Zusammenhang werden die theoretischen Grundlagen, die den verwendeten NLP-Methoden zugrunde liegen behandelt. Der Fokus liegt hierbei auf den Transformer-Modellen, im speziellen auf „Bidirectional Encoder Representations from Transformers" (BERT).
Um diesen Prozess zu testen, wurden Fragen des lokalen Energiebetreibers N-ERGIE aus Nürnberg zum Thema der Entwicklung von Elektromobilität im öffentlichen Nahverkehr verwendet. Die Genauigkeit wird anhand der Fragen nach Anzahl, Hersteller, Betreiber und Einsatzort von Elektrobussen auf einem Testdatensatz berechnet. Hierbei wird die Verbesserung auf die Ergebnisse verschiedener Transformer-Modelle für das Question-Answering durch den in dieser Arbeit entwickelten Lösungsansatz gemessen. Das ursprüngliche BERT-Base-Modell, das für die Beantwortung von Fragen trainiert wurde, wird als Referenzwert verwendet.
Ohne Veränderung der Transformer-Modelle lag der Beste gemessene F1-Wert bei 63,88 (Referenz BERT-Base F1: 44,74). Nach Anwendung des hier vorgestellten Prozesses lag der F1-Wert bei 90,06 (Referenz BERT-Base F1: 71,83). Im Vergleich hierzu liegt die Genauigkeit des Menschen auf dem bekannten SQuAD2.0 Datensatz bei einem F1-Wert von 89,45.
Diese Arbeit zeigt, dass es mittels eines Zusammenspiels von verschiedenen NLP-Methoden möglich ist, mit einer menschenähnlichen Genauigkeit Antworten auf konkrete, branchenspezifische Fragen aus unstrukturierten Texten zu extrahieren.

Freitag, 11.12.2020, 10:00 Uhr

Videokonferenz: BigBlueButton

Technische Universität Clausthal • Institut für Informatik • Julius-Albert-Straße 4 • 38678 Clausthal-Zellerfeld
Tel +49 (0)53 23 72-7140• Fax +49 (0)5323 72-7199 • http://www.in.tu-clausthal.de • info@in.tu-clausthal.de