(Intelligentes) Text Mining in der Marktforschung

Stützer, Cathleen (Ed.); Wachenfeld-Schell, Alexandra (Ed.); Oglesby, Stefan (Ed.)

Erstveröffentlichung / Primary Publication
Sammelwerk / collection

KOMPENDIUM DER ONLINE-FORSCHUNG

Band 1

(Intelligentes) Text Mining in der Marktforschung

Cathleen M. Stützer / Alexandra Wachenfeld-Schell / Stefan Oglesby (Hrsg.)

DGOF – Deutsche Gesellschaft für Online-Forschung e. V.

Empfohlene Zitierung / Suggested citation:
Stützer, C. M., Wachenfeld-Schell, A. & Oglesby, S. (2019). (Intelligentes) Text Mining in der Marktforschung. DGOF–Kompendium der Online-Forschung, Band 1. Köln: Deutsche Gesellschaft für Online-Forschung e. V. (DGOF). URL: https://nbn-resolving.org/urn:nbn

ISBN (PDF): 978-3-9815106-8-3

Herausgeber*innen:
Cathleen M. Stützer, ZQA Zentrum für Qualitätsanalyse, TU Dresden, Dresden
Alexandra Wachenfeld-Schell, GIM Gesellschaft für Innovative Marktforschung mbH, Wiesbaden
Stefan Oglesby, data IQ AG, Cham/ Zug

Internet: www.dgof.de
E-Mail: office@dgof.de

Cathleen M. Stützer / Alexandra Wachenfeld-Schell / Stefan Oglesby (Hrsg.)

KOMPENDIUM DER ONLINE-FORSCHUNG

Band 1

(Intelligentes) Text Mining in der Marktforschung

DGOF – Deutsche Gesellschaft für Online-Forschung e. V.

DGOF

Digitale Transformation der Marktforschung

Cathleen M. Stützer, Alexandra Wachenfeld-Schell & Stefan Oglesby

2019

Neue Zugänge mit neuen Herausforderungen

Seit mehr als zwei Dekaden wird sich in der Marktforschung darum bemüht, neue Wege zu erschließen, um der digitalen Transformation und den damit verbundenen gesellschaftlichen Veränderungsprozessen mit geeigneten Forschungsmethoden zu begegnen. Insbesondere in der Online-Marktforschung wird gefragt, inwiefern digitale Ressourcen erschlossen werden können, um zielgruppenorientierte Datenanalysen für (neue) Customer Insights nutzbar zu machen.

Die Informationszugänge sind in der digitalen Welt vielseitig. Von der Online-Befragung bis hin zum Einsatz digitaler Monitoring-Systeme steht ein vergleichsweise großes Portfolio an Instrumenten zur Verfügung. Dennoch werden nach wie vor Fragen zur Verwendung neuer Methoden sowie zur Güte der auf diese Weise (erhobenen) Daten laut. Hierzu wurden indes u. a. ethisch-rechtliche Rahmenbedingungen auf den Prüfstand gestellt (vgl. ICC/ESOMAR-KODEX). Dabei ist offenkundig, dass die Verwendung von Daten aus reaktiven Verfahren (wie Online-Befragungen) auf eine Tradition blickt, in der sich Standards zur Qualitätsbewertung erfolgreich etablieren konnten. Non-reaktive Verfahren hingegen stecken noch weitgehend in der Erprobung und daraus resultierende Ergebnisse werden aufgrund fehlender übergreifender Normierung (noch immer) kritisch begutachtet.

Allerdings ist aktuell ein deutlicher Trend in der Markt- und Sozialforschung erkennbar. Dieser zeigt auf, dass zunehmend kombinierte Verfahren Einzug halten, die Daten aus reaktiven mit Daten aus non-reaktiven Verfahren kombinieren, um an weiterführende Informationen zu gelangen. Hierzu rückt insbesondere der Einsatz von Text Mining-Verfahren in den Blick.

Zwar ist es kein neues Phänomen, sich digitalen Text als Informationsquelle zu erschließen. Dennoch zeigt die Praxis dazu auf, dass die Extraktion von Informationen aus Texten – insbesondere aus unstrukturierten Textdaten wie Foren, Bewertungsportalen bzw. aus offenen Angaben – eine besondere Herausforderung darstellen. Der/die Marktforscher*in von heute braucht hierzu zum einen neues methodisches Know-how, um mit den komplexen Datenbestän-

den sowohl bei der Erhebung wie auch bei der Bewertung dieser umzugehen. Zum anderen müssen im Kontext der digitalen Beforschung von neuen Customer Insights sowohl technische als auch organisationale Infrastrukturen geschaffen werden, um u. a. Geschäftsmodelle in Abläufen und Arbeitsprozessen von Unternehmen, Institutionen und Organisationen etablieren zu können.

Obwohl sich der Ausgestaltung der neuen Rahmenbedingungen in den Marktforschungsunternehmen zunehmend zugewandt und sowohl in die Weiterqualifizierung von Mitarbeiter*innen, in technische Infrastrukturen sowie in die Erweiterungen des jeweiligen Angebotsportfolios investiert wird, scheint nach wie vor eine gewisse Skepsis gegenüber automatisierten Verfahren vorzuherrschen. Diese bezieht sich u. a. auf den Umgang mit den Methoden selbst sowie auch mit den daraus resultierenden Daten(strömen). Eine transparente und nachvollziehbare Darstellung, sowohl beim Datenzugang als auch bei der Ergebnisverwertung, ist für die Einordnung der Aussagekraft der Ergebnisse unabdingbar.

Darüber hinaus ist bei der Interpretation der Ergebnisse sowohl der adressierte Personenkreis als auch dessen Motivationslage, aus der die aktiv geäußerten Beiträge resultieren, zu berücksichtigen. Insbesondere auf Kundenseite wird hierzu nach Gütekriterien sowie nach der Qualität der automatisch extrahierten Daten gefragt, um zum einen kontinuierliche und zum anderen inhaltlich verwertbare Informationsbestände von den jeweiligen Anbietern zu erhalten (vgl. empirische Evidenz).

Ein Plädoyer für den Einsatz intelligenter Methoden in digitalen Kontexten

Sowohl in der Markt- als auch in der Sozialforschung wird aktuell von einer Renaissance der qualitativen Forschung gesprochen. Die Vorteile der qualitativen Beforschung von Sachverhalten liegen dabei insbesondere im Aufdecken neuer Phänomene, weniger in der Überprüfung von Hypothesen. In der automatischen „Qualifizierung" von digitalem Textmaterial liegt nun eine besondere Herausforderung. Zum einen sollen Algorithmen nun nicht mehr nur Trends über Worthäufigkeiten erkennen (z. B. durch die automatische Extraktion von Themen und Schlagwörtern), sondern zunehmend auch Aufschluss über deren Aussagekraft (Meinung, Stimmung etc.) bzw. deren Kontexte liefern. Doch wie kann das gelingen? Welche Ansätze, Methoden und Tools tragen aktuell dazu bei? Und kann intelligentes Text Mining dabei unterstützen, sich den neu aufkommenden Fragen der Markt- und Sozialforschung zu nähern? Der vorliegende Band befasst sich mit diesen Fragen und nimmt insbesondere intelligente Text Mining-Verfahren in den Blick, die zur Kontextualisierung von textbasierten Inhalten beitragen. Die Beiträge des Bandes besprechen hierbei nicht nur vielfältigste Methoden und Verfahren zur automatischen Textextraktion, sondern zeigen hierbei sowohl die Relevanz als auch die Herausforderungen für die Online-Marktforschung auf, die mit dem Einsatz solch innovativer Ansätze und Verfahren verbunden sind.

Über die Herausgeber*innen

Dr. Cathleen M. Stützer
Zentrum für Qualitätsanalyse
TU Dresden
Chemnitzer Straße 48a
01187 Dresden
Tel.: +49 351 463 42224
E-Mail:
Cathleen.Stuetzer@tu-dresden.de

Dr. Cathleen M. Stützer ist Projektleiterin am Zentrum für Qualitätsanalyse (ZQA) der TU Dresden und seit 2015 Vorstandsmitglied der DGOF.

Alexandra Wachenfeld-Schell
GIM Gesellschaft für Innovative Marktforschung mbH
Mainzer Straße 75
65189 Wiesbaden
Tel.: +49 172 6918 745
E-Mail:
A.Wachenfeld-Schell@g-i-m.com

Alexandra Wachenfeld-Schell ist Senior Research Director bei der GIM Gesellschaft für Innovative Marktforschung mbH und seit 2013 Vorstandsmitglied der DGOF.

Dr. Stefan Oglesby
data IQ AG
Alte Steinhauserstrasse 33
6033 Cham/ Zug, Schweiz
Tel.: +41 79 641 0473
E-Mail:
stefan.oglesby@data-iq.ch

Dr. Stefan Oglesby ist Chairman der data IQ AG und Lehrbeauftragter an der Universität Luzern. Seit 2019 ist er Vorstandsmitglied der DGOF.

Weitere Informationen
www.dgof.de

André Lang und Marc Egger, Insius UG

Wie Marktforscher durch kooperatives Natural Language Processing bei der qualitativen Inhaltsanalyse profitieren können

Erkenntnispotenziale und Herausforderungen durch allverfügbare Textdaten

In den vergangenen Jahren haben wir einen immensen Anstieg an verfügbaren Textdaten feststellen dürfen. Nicht nur Verbraucherkommentare in Social Media, Text-Transkripte von Sprachassistenten, Dialoge aus Customer-Feedback und Support Chat (Bots), sondern auch offene Antworten in Umfragen bis hin zu automatisierten Text2Speech-gestützten Audiointerviews sorgen für steigenden Bedarf, Erkenntnisse aus unstrukturierten Textdaten zu gewinnen.

Das Erkenntnisinteresse kann in einer explorativen Fragestellung (wie „Was denken Verbraucher über das Thema X") bzw. bei offenen Antworten in der Abfrage einer konkreten Wahrnehmung eines bestimmten Objektes bzw. Situation bestehen. Der Prozess zur Erkenntnisgenerierung führt in beiden Fällen klassischerweise über eine manuelle Codierung der verfügbaren Textdaten.

Der Wert der kognitiven Leistung, den Forscher hierbei erbringen, um aus textbasierten Daten Informationen bis hin zu nutzbaren Erkenntnissen zu gewinnen, ist unbestreitbar. In der heutigen mit dem Begriff „Big Data" umschriebenen Umgebung, steht dieses Vorgehen jedoch vor großen Herausforderungen. Durch die Menge (Volume) der mittlerweile verfügbaren Textdaten, der Geschwindigkeit, in welcher neue Daten entstehen (Velocity), und den unterschiedlichen Charakteristika (Variety) wie Informationsdichte, Datenqualität und Themenbezogenheit, die bei Tweets anders als in offenen Antworten in einem Fragebogen sind, wird ein manuelles Vorgehen bei der Erschließung inhaltlicher Dimensionen aus Textdaten zunehmend schwieriger.

Gleichzeitig ergeben sich aus dieser Datenflut auch erhebliche Chancen, denn auf der organisationalen Nachfrageseite ist hohes und weiterhin steigendes Interesse an qualitativen Erkenntnissen spezifischer Fragestellungen zu Marken, Produkten oder allgemeinen Themen zu beobachten. Dies erscheint nachvollziehbar, bergen die mittlerweile breit vorliegenden bzw. leicht zu erheben natürlichsprachigen Daten doch ein immenses Potenzial, relevante Insights für das Marketing zu generieren. Dies reicht von Image- und Wahrnehmungsanalysen, der Ermittlung von Bedürfnissen, Kritikpunkten bis hin zur Erkennung von Trends. Andererseits geht mit diesem Interesse auf Kundenseite jedoch nicht unmittelbar ein Verständnis für den notwendigen (manuellen) Analyseaufwand einher oder die Budgets für derartige Analysen sind schlicht nicht vorhanden.

Volume, Velocity und Variety stellen Herausforderungen beim manuellen Erschließen großer Textdatenmengen dar.

Automatisierung und künstliche Intelligenz als Lösung?

Eine mögliche Lösung, um aus Textdaten Erkenntnisse für die Markt- und Verbraucherforschung zu gewinnen und gleichzeitig die Aufwände in durch Zeit- und Kostendruck geprägten Situationen zu reduzieren, wird vielfach in der Verwendung von Methoden aus dem Bereich der künstlichen Intelligenz gesehen. Diese versprechen durch Automation, manuelle Aufwände zu reduzieren.

Gleichzeitig besteht jedoch die Gefahr, Initialaufwände zu übersehen und/oder die lieferbare Erkenntnistiefe der Analyseergebnisse zu über-

schätzen. Hinsichtlich der Initialaufwände ist zu beachten, dass viele Fragestellungen ein vorheriges Anlernen der Algorithmen auf den jeweiligen Untersuchungsgegenstand erfordern (sog. (semi-)supervised learning).

Der Wunsch nach Nutzung von Methoden der künstlichen Intelligenz zur Effizienzsteigerung führt bei der Umsetzung durch externe Dienstleister also häufig zunächst eher zu aufwändigen Beratungsprojekten. Zudem sind die resultierenden Algorithmen und Modelle, beispielsweise solche, die geeignet sind, Textbeiträge anhand verschiedener Themenkomplexe zu klassifizieren, auf spezifische Anwendungsfälle beschränkt.

Die Verfahren erkennen nur die Kategorien (Codes), die ihnen zuvor durch den Forscher beigebracht wurden. Dies erfordert zum einen, Vorannahmen zu treffen oder Vorstudien zur Kategorienermittlung vorzunehmen. Zum anderen schränkt dieses Vorgehen die Möglichkeiten der explorativen Erkenntnisgewinnung ein.

Eine andere, oft angewandte Verfahrensklasse, stellt das automatisierte Clustering dar. Vereinfacht dargestellt erzeugen Clusteringalgorithmen hierbei eine Menge von Begriffswolken (ähnlich Word-Clouds) deren konkrete Anzahl der Forscher allerdings üblicherweise im Vorfeld festlegen muss. Die resultierenden Begriffscluster sind anschließend durch den Forscher zu interpretieren. Zusammen mit der richtigen Aussteuerung der Anzahl der zu erstellenden Cluster kommt diese Aufgabe in der Praxis jedoch häufig eher einem Kaffeesatzlesen gleich und bedeutet ebenfalls nicht zu unterschätzende Aufwände in der Algorithmenaussteuerung. Letztlich sind noch mehr oder minder aufwändige Verfahren des Natural Language Processing (NLP) zu nennen, bei denen in der Regel Häufigkeitszählungen von Worten (oder Wortarten und Bezügen) vorgenommen werden.

In Kombination mit Methoden des Information Retrievals können hier sog. Keyterm-Extraktionen vorgenommen werden, die das Ziel haben, besonders wichtige Begriffe und Phrasen (z. B. sog. Nominalphrasen) zu Tage zu fördern. Hier zeigen sich jedoch oft Probleme in der Erkenntnistiefe der möglichen Ergebnisse. Beispielsweise werden durch Algorithmen Keywords wie „teuer", „Preis", „günstig" etc., die gemeinsam den Themenkomplex der „Preiswahrnehmung" beschreiben, extrahiert.

Die Zusammenführung dieser Keywords zum Themenkomplex kann aber aufgrund von Komplexität und Variantenreichtum (z. B. Schreibweisen, Komposita, Compounds) der Sprache entweder gar nicht oder, z. B. bei Verwendung von Begriffstaxonomien, nur unzureichend automatisch erfolgen. Somit ist auch hier oft ein tiefes Eingreifen in die unterliegenden Algorithmen oder ein aufwändiges Training von Sprachressourcen notwendig, um eine hinreichende Analysetiefe zu gewährleisten.

Jeder macht, was er am besten kann: Kooperative Mensch-Maschine Systeme

Die vorangegangenen ausschnittartigen Ausführungen zeigen, dass der vermeintlich automatisierten Analyse von Textdaten im Vergleich zur manuellen Analyse durchaus Grenzen hinsichtlich der erreichbaren Erkenntnistiefe gesetzt sind.

Hinzu kommt, dass durch die notwendige Algorithmenaussteuerung zusätzliche (Initial-)Aufwände anfallen, die sich nicht proportional zum möglichen Erkenntnisgewinn verhalten. Somit entsteht zwischen den beiden Extremen von manueller Textanalyse (hohe Erkenntnismöglichkeiten / hohe Aufwände) und automatisierter Textanalyse (geringere Aufwände / eingeschränkte Erkenntnismöglichkeiten) eine Lücke. Um diese zu füllen, wird hier für die Entwicklung und Einführung einer neuen Klasse von Textanalysesystemen plädiert.

Die hier als „Kooperatives Natural Language Processing" (Kooperatives NLP) bezeichnete Systemklasse ist per Design darauf angelegt, mit dem Forscher gemeinsam am Erkenntnisgewinn zu arbeiten. Dies bedeutet, dass Aufgaben, die gut maschinell erledigt werden können (z. B. das Erkennen und Extrahieren bestimmter Sprachmuster) mittels automatisierter

Textanalysealgorithmen erfolgen. Aufgaben, die dagegen schwer allgemeingültig automatisierbar sind oder „Weltwissen" erfordern, werden weiterhin durch Menschen erledigt. Hierzu gehören Kontexterkennung, Synonymauflösung oder das Verstehen, welche Themenkomplexe durch welche automatisiert erhobenen Muster abgebildet und dargestellt werden. Zudem besitzen kooperative NLP-Systeme eine Schnittstelle zwischen Forscher und Maschine, die es erlaubt, dass beide voneinander lernen.

Hierzu kann der Forscher einerseits Korrekturen an automatisiert erhobenen Ergebnissen vornehmen, wie z. B. der Korrektur einer falsch automatisiert klassifizierten Tonalität (positiv/negativ/neutral) oder dem System mitteilen, welche automatischen Extraktionen welche Themenfelder bilden. In beiden Fällen hat das kooperative NLP-System die eigenen Systementscheidungen aufgrund des Forscher-Feedbacks zu überprüfen, zu justieren, und dem Forscher ggf. erneut vorzulegen.

Dies wird gemeinhin auch unter dem Begriff „Active Learning" gefasst. Kooperative NLP-Systeme sollten aber über reines Active Learning hinaus gehen und zusätzlich ein sog. Recommender-System beinhalten. Dieses unterbreitet auf Grundlage der durch den Forscher manuell erzeugten Zusammenhänge weitere Vorschläge, welche Zusammenhänge noch in den Daten enthalten sein können. Wir sind der Ansicht, dass ein solches System eine für die Markt- und Verbraucherforschung neue Klasse von Unterstützungssystemen repräsentiert. Ein solches System erlaubt es, manuelle Forschungsarbeit und Automatisierung auf eine für beide Seiten vorteilhafte Weise zu kombinieren.

Sie repräsentieren einen Hebel um Erkenntnistiefe und Automatisierung im Sinne des Forschers auszusteuern. Gemeinsam erreichen also Forscher und Maschine bei explorativen Analysen eine höhere Erkenntnistiefe als komplett automatisierte Systeme, bei gleichzeitig geringerem Aufwand als manuelle Analysen. Im Folgenden wird ein solches System am Beispiel des kooperativen Clusterings vorgestellt.

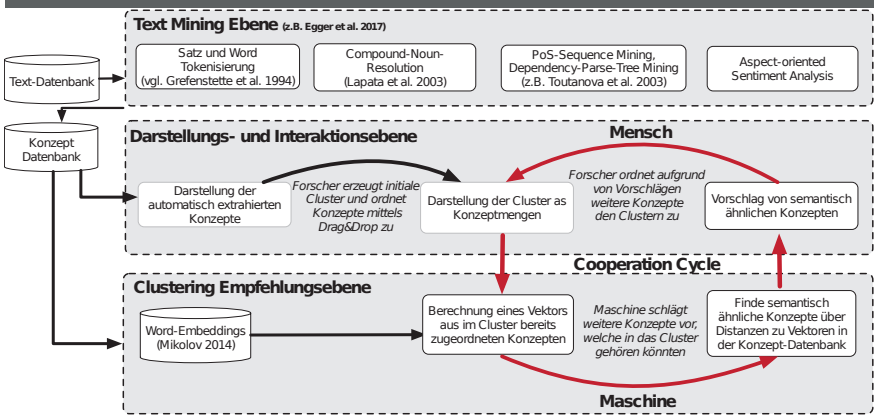
Kooperatives Clustering als Lösung um Themenfelder in Textdaten aufzudecken

Im vorangegangenen Abschnitt wurde ein Aspekt der automatisierten Extraktion von bedeutsamen Worten und Phrasen angesprochen. Wenn automatisierte Textanalysesysteme Worte und Phrasen aus Texten extrahieren, entsteht die Notwendigkeit, diese zu Themenkomplexen zusammenzuführen, um quantitative Aussagen darüber treffen zu können, wie stark diese Themenkomplexe in den Daten repräsentiert sind.

Die einzelnen Extraktionen reichen hierzu nicht aus. Auch die Benennung und Auswahl von Themenkomplexen ergibt sich nicht automatisiert aus den Daten. Diese sind vielmehr durch den Forscher zu bestimmen und in einem in Hinblick auf den Analysezweck geeigneten Abstraktionsgrad zu modellieren – eine konstruktivistische menschliche Leistung, die kein automatisiertes System alleine leisten kann.

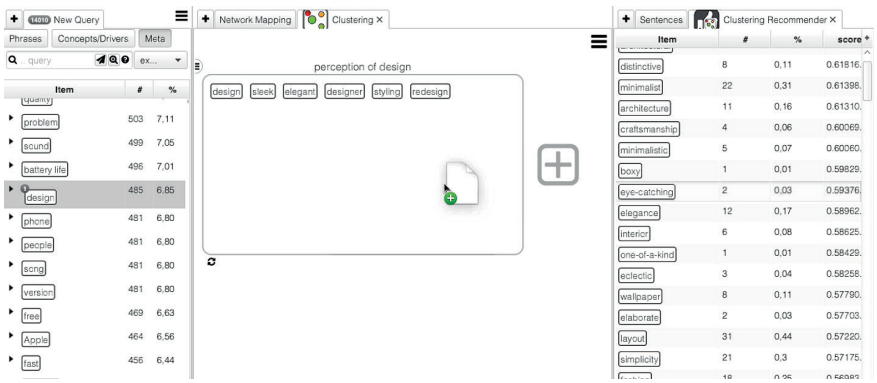
Als Beispiel sei hier erneut die Extraktion der Worte und Phrasen „niedriger Preis", „teuer", „kostengünstig", „Wucherpreis" oder „preiswert" genannt. Diese werden im Folgenden als Konzepte bezeichnet. In diesem Beispiel wäre darin ein Themenkom-

Abbildung 1: Systemarchitektur eines kooperativen NLP Systems



plex „Preiswahrnehmung“ zu erkennen. Allerdings könnte der Forscher auch entscheiden, nicht nur einen Themenkomplex „Preiswahrnehmung“ zu modellieren, sondern nach der Tonalität zu unterscheiden und folglich „positive Preiswahrnehmung“ (kostengünstig, preiswert, niedriger Preis) der „negativen Preiswahrnehmung“ (teuer, Wucherpreis) quantitativ gegenüberzustellen.

Abbildung 2: User-interface Kooperatives Clustering (Quelle: Insius InMap)



Wie der Themenkomplex abzugrenzen und genau zusammenzusetzen ist, obliegt dem individuellen Forschungsinteresse. Durch die Zuordnung der Konzepte zur Preiswahrnehmung kann der Themenkomplex „Preiswahrnehmung“ anschließend quantifiziert werden. Wie bereits erwähnt, könnte die Nutzung von Taxonomien und Synonymlexika einen Ansatz bieten, hier semantische Beziehungen aufzudecken und automatisiert Zuordnungen vorzunehmen.

Allerdings beantworten diese Ansätze (selbst wenn sie perfekt funktionieren würden) nicht die Frage, wie genau Themenkomplexe zu definieren und abzugrenzen sind, um das individuelle Forschungsinteresse zu befriedigen. Daher wird hier ein kooperatives NLP-System illustriert, welches den Forscher durch Mensch-Maschine-Interaktion dabei unterstützt, aus automatisiert extrahierten Konzepten Themenkomplexe zusammenzuführen. Dies wird hier als semi-automatisiertes Clustering bezeichnet.

Das kooperative NLP-System besteht aus drei Ebenen. In der Text Mining-Ebene werden Algorithmen eingesetzt, um aus den in der Text-Datenbank vorliegenden Texten Konzepte zu extrahieren und zusammen mit ihrer Häufigkeit in einer Konzept-Datenbank zu speichern.

Die Ergebnisse der Konzeptextraktion werden in der Darstellung tabellarisch angezeigt und stehen dem Forscher zur Zuordnung zu Clustern (Themenkomplexen) zur Verfügung. Cluster werden wie Ordner angelegt und benannt, die Konzepte werden ihnen durch Drag-and-Drop zugeordnet. Sobald einem Cluster die ersten Konzepte zugewiesen sind, springt das Empfehlungssystem an. Ähnliche Konzepte zu den bereits zugeordneten werden als weitere Kandidaten vorgeschlagen.

Durch die Vorschläge, die mit jedem Hinzufügen präziser werden, lassen sich die Cluster in kurzer Zeit umfassend definieren. Hier trägt die Maschine ihren Teil bei, indem die zeitaufwändige manuelle Suche nach ähnlichen Konzepten entfällt, während der Mensch das Forschungsinteresse über die Clusterauswahl und -zuordnung modelliert. Gemeinsam können Mensch und Maschine so in kurzer Zeit die Themenkomplexe einer großen Datenmenge nachprüfbar und quantifizierbar herausarbeiten.

Anwendungsbeispiel

Auf Grundlage des hier vorgestellten Systems wurde eine illustrative Studie durchgeführt. Hierzu wurden 26.059 deutschsprachige Kommentare und Social Media-Postings mittels eines Web-Crawlers zum Thema „Haushaltsgeräte“ gesammelt, analysiert und mittels des NLP-Systems zu Themen geclustert.

Hierbei war es möglich, innerhalb einer Stunde neun Themenfelder herauszuarbeiten und anhand der jeweiligen Beitragszahlen quantitativ messbar zu machen. Die beiden stärksten Themen sind die verschiedenen Gerätekategorien (Waschmaschine, Kühlschrank, Geschirrspüler

etc.) in 55 % der Texte sowie Nennungen zum Kaufprozess (kaufen, Preis, günstig, verkaufen, bestellen, Lieferung etc.) in 54 % der Texte. Bei den Produktattributen (44 %) werden Preis, Lautstärke (leise), Qualität (u. a. hochwertig) sowie Funktionen und Leistung am stärksten thematisiert. Konkrete Markennennungen, aus denen sich das „Consideration Set" der Verbraucher ableiten lässt, finden sich in 40 % der Beiträge. Weitere diskutierte Themenfelder sind Hilfe und Rat, Verwendungsort (Küche, Wohnung etc.), Farben und Materialien sowie die jeweiligen Retailer (Amazon, Ikea etc.).

Die Studie zeigt, wie mittels kooperativen Clusterings aus einer großen Zahl von Texten in kurzer Zeit relevante Themenfelder herausgearbeitet werden können. Rein automatisierte Verfahren sind hierzu nicht in der Lage, da die kognitive Leistung des Forschers bei der Modellierung der Themenfelder nicht ersetzbar ist. Diese ergeben sich, so auch in dieser Studie, nicht „automatisch" aus den Daten, sondern leiten sich aus den jeweiligen Anforderungen ab, und könnten auch weitaus kleingliedriger gestaltet werden.

Eine Alternative zur detaillierten qualitativen Inhaltsanalyse ist mit dem vorgestellten System dennoch nicht zu erreichen. Vielmehr ermöglicht es in datenreichen, budget- oder zeitkritischen Situationen sowie als Ergänzung zu etablieren Verfahren eine quantifizierbare Sichtweise auf unstrukturierte Textmengen. Ziel ist dabei der bestmögliche Trade-off zwischen Analyseaufwand und Erkenntnistiefe.

Wie kann das kooperative NLP-System genutzt werden?

Als Beitrag formuliert dieser Artikel die Idee von kooperativen NLP-Systemen für die Markt- und Verbraucherforschung. Es wurde ein Design eines solchen Systems zum Clustern von Begrifflichkeiten vorgestellt. Damit ist es möglich, dass Forscher in kurzer Zeit die für das Untersuchungsinteresse wichtigsten Themenkomplexe aus einer großen Textmenge herausarbeiten. Das hierbei vorgestellte kooperative NLP-System kann als Software-as-a-Service getestet bzw. genutzt werden.

INSIUS

INSIUS UG (haftungsbeschränkt)
Eupener Straße 165
50933 Köln
Tel.: +49 221 4558026-0
E-Mail: hello@insius.com

Über die Autoren

André Lang ist Co-Gründer und Geschäftsführer der Firma Insius. Er studierte Wirtschaftsinformatik an der Universität zu Köln.

Dr. Marc Egger ist Co-Gründer und Gesellschafter der Firma Insius. Er studierte und promovierte an der Universität zu Köln.

Weitere Informationen
www.insius.com

Matthias Heurich & Sanja Štajner, Symanto Research

Durch Technologie zu mehr Empathie in der Kundenansprache – Wie Text Analytics helfen kann, die Stimme des digitalen Verbrauchers zu verstehen

Skalierbarkeit semantischer Analysen durch maschinelles Lernen

Sprache stellt unsere Verbindung zur Welt dar – dazu, wie wir die Welt verstehen und mit ihr interagieren. Digitalisierung hat dazu geführt, dass Konsumenten Tag für Tag und in unterschiedlichsten Kanälen digitale, textbasierte Sprachspuren kreieren und hinterlassen.

Ein systematisches Lesen und Verstehen dieser (textlichen) Sprachspuren erlaubt tiefe Einblicke, wie Konsumenten ihre Interaktionen mit Produkten, Marken und Unternehmen wahrnehmen und wie sie ihre Erfahrungen beschreiben. Durch neue technologische Möglichkeiten der Textanalyse erhalten wir authentische, relevante, und, im Gegensatz zu manueller Vercodung, auch skalierbare Einblicke in Bedürfnisse und Wünsche von Konsumenten. Wir können somit ein tiefes Verständnis darüber entwickeln, wie sich Konsumentenerwartungen nicht nur erfüllen, sondern auch übertreffen lassen.

Basierend auf der Text Analytics-Plattform von Symanto soll dieser Artikel einen exemplarischen Einblick in das dynamische Arbeitsfeld der Textanalyse im kommerziellen Anwendungsbereich liefern.

Von Technologie zu Marketing-Entscheidungen

Wir erläutern den Zugang zu Text Analytics anhand der Anwendung der eigenen Plattform im Social Media Kontext. Diese Art von Textdaten hat besondere Relevanz, da Investitionen in Social Media-Management zunehmend steigen, der tatsächliche Einfluss und Nutzen dieser Investitionen häufig jedoch eine unbekannte Größe bleibt.

Die Anwendung fortgeschrittener, tiefer Text Analytics ermöglicht es Unternehmen Social Media-Diskussionen so zu verstehen, dass sie ein tiefes Verständnis von Konsumentengruppen entwickeln können. Darauf aufbauend können Ansätze für eine empathische Konsumentenansprache entwickelt werden, bei denen dieses Wissen um Motivationen, Emotionen und subjektive Bedeutung aus Konsumentenäußerungen berücksichtigt wird, und die damit über Ansätze basierend auf Demographie und reines vergangenheitsbezogenes Verhalten hinausgehen.

Text Analytics in Zeiten von Artificial Intelligence (AI)

Die Basis unserer Text- und Social Media-Analysen ist ein Artificial Intelligence (AI)-System. Bei diesem System handelt es sich um eine skalierbare psycholinguistische Analyseplattform, die mehrere Services umfasst: Spracherkennung, Datenbereinigung und Datenschutzaspekte, Extraktion von Feature-Sentiment, Extraktion von Feature-Sentiment-Beziehung, Psycholinguistik und Profilklassifizierung, Sentiment-Negation, Emotionsklassifizierung, Disambiguations-Klassifizierung, Geschlecht- und Alterseinstufung.

Im Gegensatz zu den üblicherweise verwendeten Textanalyse-Tools legen unsere Systeme besonderen Wert auf den Kontext und das Profil des Autors. Dabei werden die oben genannten Tools zu einem umfassenden Gesamtergebnis kombiniert, bei dem zum Beispiel das Wort „Zeit“ in den folgenden zwei Beispielen mit unterschiedlichem Sentiment besetzt sein wird:

1. „Ich hatte die beste Zeit meines Lebens, als ich dort lebte.“

2. „Ich hatte gestern Zeit, dort hin zu gehen."

In ähnlicher Weise hat das Profil des Autors Einfluss auf die Verarbeitung von Äußerungen. Zum Beispiel sollte die Phrase „mit der Kameraauflösung nicht zufrieden" aus einem Kamera-Review anders interpretiert werden, wenn solch ein Review von einer sehr technischen oder einer sehr emotionalen Person geschrieben wird.

Dazu wurden AI-Systeme entwickelt, um tiefe Einblicke in große Datensätze und Zeitreihen zu erhalten: darunter psycholinguistisches Clustering, Feature-Clustering, Themenerkennung und Empfehlungen, schwach überwachtes sequence labelling, semantic spaces, Sentiment-Erkennung, Depressions-Erkennung, Engagement-Scoring und Natural Language Generation (NLG) mit Stilanpassung.

Die entwickelten Systeme basieren dabei auf Deep Learning, sind fully unsupervised oder weak supervised, erfordern keine manuellen Annotationen und sind leicht an neue Sprachen und Domänen anpassbar. Sie werden zunächst für die englische Sprache entwickelt und dann an andere Sprachen angepasst, wobei der Genauigkeitsverlust bei dieser Übertragung zwischen 2 % und 2,6 % liegt.

„Den neuesten Stand der Technik bilden wir in den Ergebnissen unserer Systeme durch einen Zwei-Phasen-Ansatz ab."

Zwei-Phasen-Ansatz. In der ersten Phase werden generische AI-Modelle aus einer großen Menge von Daten aufgebaut, die im Internet frei verfügbar sind (Konversationen, Rezensionen, Dokumente, Foren, Tweets usw.). Das System lernt Vokabeln, Sprach- und Konversationsstrukturen, Stimmungen, Gefühlsausdrücke, Produktaspekte sowie sprachübergreifende und domänenübergreifende Unterschiede.

In der zweiten Phase werden die generischen AI-Modelle anhand von Kundendaten, spezifischen Geschäftsanforderungen sowie Expertenwissen angepasst, um sowohl kunden- als auch domänenspezifische Anforderungen bei der Analyse zu berücksichtigen.

Derzeit werden eigene Forschungsaktivitäten auf den Aufbau von NLG-Systemen ausgedehnt. Diese können den Stil der Textausgabe an einen bestimmten Benutzertyp oder seinen emotionalen Zustand anpassen. Frühere Arbeiten haben gezeigt, dass die Verwendung des „richtigen" Schreibstils den größten Einfluss darauf hat, Meinungen zu ändern und die beabsichtigte Reaktion beim Leser hervorzurufen. Daher ist es heutzutage eines der Hauptanliegen der Industrie, über NLG-Systeme zu verfügen, die den Stil der Ausgabe an die jeweilige Benutzergruppe anpassen können. Eigene NLG-Systeme zur Stilanpassung basieren auf drei Phasen.

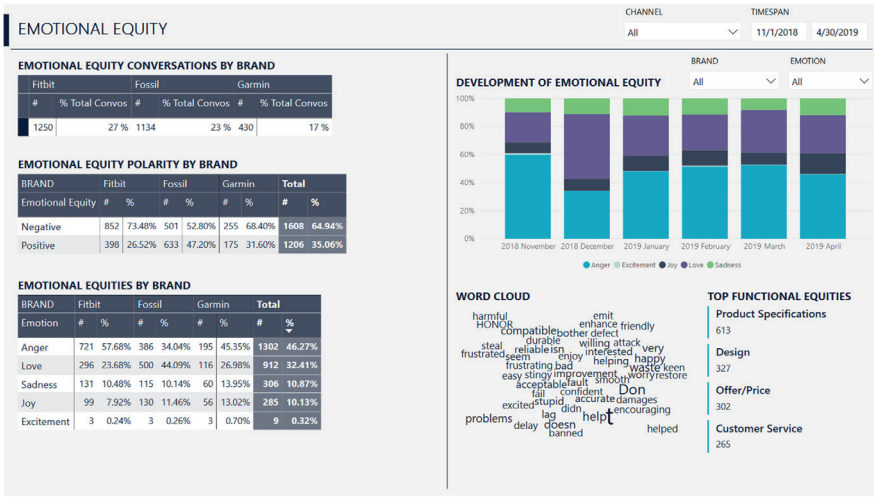
Zuerst werden Benutzer in mehrere Cluster gruppiert, indem eigene grundlegende psycholinguistische Analysewerkzeuge auf die Sammlungen von E-Mails, Metadaten und Benutzerreaktionsinformationen angewendet werden, um ihre psychologischen und emotionalen Merkmale zu extrahieren. Dadurch werden die Unterschiede im Schreibstil erkennbar, die mit verschiedenen Benutzergruppen zusammenspielen, sowie die am stärksten ausgesprochenen Textmerkmale angelernt (z. B. Textlänge, Verwendung von Emojis, Rechtschreibung usw.). Dieses Wissen kann dann verwendet werden, um eine Engine zu trainieren, die Text zwischen psychologischen Profilen anpassen kann.

Social Media, Text Analytics und die Diagnose von Brand Health – Ein Anwendungsfall

Im Kontext von Social Media spielt eine Auswahl der oben dargelegten Analyse-Module zur Messung der Brand Health zusammen. Während weit verbreitete Social Media-Kennzahlen in der Regel Reichweite und Anzahl an Interaktionen mit Konsumenten quantifizieren, zeigen diese nicht den tatsächlichen emotionalen Einfluss von Marken und deren Aktivitäten auf Konsumenten.

Ziel der Entwicklung eines Tools zur Messung der Social Media Brand Health war es, Text Analytics so einzusetzen, um deutlich über oberflächliche Volumen- und globale Sentimentkennzahlen hinauszugehen.

Abbildung: Dashboard



Ein Anspruch ist es, den Grad der emotionalen Verbindung und Empfehlung einer Marke mit Konsumenten aufzuzeigen und zudem, warum diese Dimensionen letztendlich dazu beitragen, dass die systematische und kontinuierliche Beobachtung und Auswertung von Social Media-Aktivitäten zur Verbesserung der sogenannten Brand Health genutzt werden können.

Dabei erstreckt sich der dargestellte Ansatz zur Messung der Brand Health über folgende analytische Dimensionen:

- Topic detection: Worüber reden die Konsumenten im Kontext von Produkten und Marken?
- Advanced sentiment: Was mögen die Konsumenten im Kontext von Produkten und Marken?
- Bedürfnisse: Was möchten Konsumenten?
- Kanäle: Wo tauschen sich Konsumenten aus?
- Persönlichkeit: Welche Rückschlüsse lassen sich auf Basis von angewandter Psychographie über grundlegende Arten von Entscheidungsmustern treffen (emotionales vs. rationales Informationsverarbeitungs- und Entscheidungsverhalten)?
- Motivation: Welche Rückschlüsse lassen sich über faktenbasierte vs. erfahrungs-/selbstzentrierte motivationale Grundhaltungen ziehen?
- Intentionen zur Kommunikation: Inwieweit erfolgt Kommunikation mit dem Ziel eines handlungs- bzw. informationsorientierten Ergebnisses?
- Emotion: Welche Gefühlsmuster lassen sich in den Social Media-Beiträgen im Kontext von Produkten und Marken identifizieren?

Es wurde Social Media Brand Health mithilfe von Maßgrößen kodifiziert, die durch die Kombination aus künstlicher Intelligenz und Psychologie möglich werden, um zu zeigen, wie sich Verbraucher wirklich mit einer Marke und deren Aktivitäten auseinandersetzen, sodass die Marke wiederum schneller und bessere Entscheidungen treffen kann. Durch die Kombination dieser analytischen Dimensionen lassen sich folgende Maßzahlen zur Bewertung der Social Media Brand Equity ableiten:

Functional Equities: Konkrete und grundlegende Elemente der Customer Experience, welche die Qualitätswahrnehmung der Verbraucher durch die Marke beeinflussen.

Emotional Equities: Qualitäten, anhand derer Konsumenten ihre emotionale Erfahrung mit der Marke beschreiben.

Brand Connection: Anteil an Konversationen mit der stärksten Intensität an positiven Emotionen über die Marke.

Brand Recommendation: Wahrgenommener Einfluss, den ein Social Media Post auf andere User haben kann.

Ein Vergleich der Ausprägungen dieser Dimensionen im Kontext von markenspezifischen Diskussionen erlaubt es, in der Analyse konkret he-

rauszuarbeiten, welche Markenprofile sich aus Sicht von Konsumenten in Social Media-Diskussionen herauskristallisieren.

Die Verknüpfung von Emotional Equities und Functional Equities erlaubt es, konkrete Love Marks und Pain Points zu identifizieren und, darauf aufbauend, Marketing-Strategien zu entwickeln, welche die aus Sicht der Konsumenten wirkliche relevanten Themen adressieren und somit Social Media-Strategien intelligenter zu steuern. Als Kompass zur Entwicklung, Steuerung und kontinuierlichen Kursbeobachtung lassen sich die relevanten Maßzahlen in Form von Scorecards und Online-Dashboards zusammenfassen.

Diese ermöglichen (in jeweils unterschiedlichen Ebenen von Detailinformationen) verschiedenen Funktionen im Unternehmen (Marketingstratege, Social Media Team, Geschäftsführung etc.), den jeweils aktuellen Status der Brand Health auch im Vergleich zu den Wettbewerbern zu verstehen, und gleichzeitig die Entwicklung und die Wirkung von beispielsweise neuen Marketing-Kampagnen zu beurteilen.

Die Beispielanalyse zu Smart Watches zeigt hierbei, dass die einzelnen Marken nicht nur auf der funktionalen Ebene mit teils unterschiedlichen Functional Equities besetzt werden (bspw. Service, Price, Sound), sondern es den Marken auch in sehr unterschiedlichem Ausmaß gelingt, eine emotionale Beziehung zu den Konsumenten aufzubauen.

Während sich Garmin als eine überdurchschnittlich funktionale Marke herauskristallisiert, gelingt es Fossil, eine besonders starke emotionale Markenbindung aufzubauen. Für Fossil zeichnet sich Brand Love dabei als dominierende emotionale, stark positive Equity-Dimension in den Social Media-Diskussionen ab.

Integration von Text Analytics in Geschäftsprozesse und -entscheidungen

Die Detailtiefe und Präzision, welche Text Analytics heute bereits liefern, erlaubt es beispielsweise Marketing-strategen in weiteren Analyseschritten herauszuarbeiten, welche Treiber für eine positive, emotionale Markenbeziehung verantwortlich sind und wie diese Treiber konkret genutzt werden können, um das Bild einer Marke zu schärfen und positiv zu beeinflussen.

In Kombination mit der Möglichkeit, Textdaten aus unterschiedlichen Quellen (Social Media, Email, Umfragen etc.) zu konsolidieren und zu analysieren, entstehen somit für Unternehmen über Social Media hinaus weitere innovative Ansätze, um das Wissen über Kunden und Konsumenten über demographische oder auch verhaltensbezogene Daten hinaus auszuweiten. Somit lassen sich neue Dimensionen und Qualitäten für ein emphatisches Kundenbeziehungsmanagement erschließen.

SYMANTO Research
Pretzfelder Strasse 15
90425 Nürnberg
Tel.: +49 911 378466 39
E-Mail: info@symanto.net

Über den/die Autor*in

Matthias Heurich ist Business Developer bei Symanto. Er betreut u. a. Innovations- und Strategieprojekte.

Dr. Sanja Štajner ist Senior Research Scientist bei Symanto. Ihre Themenschwerpunkte sind Natural Language Processing und AI.

Weitere Informationen
www.symanto.net

Gernot Heisenberg, TH Köln & Tina Hees, Questback GmbH

Text Mining-Verfahren zur Analyse offener Antworten in Online-Befragungen im Bereich der Markt- und Medienforschung

Hintergrund, Zielgruppe und Anwendungsfelder

Text Mining ist eine interdisziplinäre Forschungsrichtung, jedoch befassen sich nur sehr wenige Arbeiten bislang mit dem Extrahieren, Analysieren und Erkennen von freien Antworten aus Online-Befragungen.

Dabei haben die Methoden zur inhaltlichen Erfassung und automatischen Analyse von Texten aufgrund der stark gestiegenen Verfügbarkeit und des möglichen Zugriffs zu digitalisierten Textdaten in den letzten Jahren stark an Bedeutung gewonnen. Das Finden relevanter Dokumente für bestimmte Suchanfragen im Bereich Information Retrievals kann bereits sehr gut automatisch ausgeführt werden. Die Herausforderung beim Text Mining liegt darin, dass Informationen in natürlichsprachiger und unstrukturierter Form vorliegen.

Um sie verwenden zu können, müssen diese semantisch gedeutet und die sprachlichen Zusammenhänge erst einmal erkannt werden. Je nach Art der Textdaten und des Informationsbedürfnisses entstehen darüber hinaus – je nach Anwendungsfall – weitere spezifische Herausforderungen.

Text Mining spielt neben offenen Befragungsantworten im Bereich der Markt- und Medienforschung auch für viele andere Anwendungsfelder eine große Rolle.

So wird es zum Beispiel sehr stark im Social Media-Bereich eingesetzt, um dortige Beiträge hinsichtlich ihrer Sentiments auszuwerten, im Customer-Relationship-Management Emails und Posts in Beschwerde- und Nicht-Beschwerde-Emails zu klassifizieren oder Produktbewertungen für Marketingzwecke zu analysieren.

Insbesondere bei großen Textmengen kommt man bei manueller Bearbeitung der Textdaten schnell an die Grenzen eines akzeptablen Aufwandes. Text Mining findet aber auch in weniger zu erwartenden Feldern seine Anwendung.

So zum Beispiel zur Nebenwirkungsforschung in der Medizin (Warrer et al. 2012) oder zur Kommunikationsüberwachung bei der Verbrechensbekämpfung (Keyvanpour, Javideh & Ebrahimi 2011) und dabei insbesondere zur Detektion von „Hate-Speech" in sozialen Netzwerken (Ting et al. 2013).

Text Mining-Ansätze und Methoden

Text Mining ist die Offenlegung und Gewinnung von informativem, nicht-trivialem Wissen aus freiem und unstrukturiertem Text. Dies schließt Methoden von Information Retrieval über die Extraktion von Entitäten und Relationen bis zu Text-Klassifikationen und Clustering mit ein (Kao & Poteet 2007). Es kann als eine Erweiterung des Data Mining auf Texte betrachtet werden (Zeng et al. 2012). Häufig werden dabei auch Methoden aus dem spezifischeren Bereich des Natural Language Processing (NLP) genutzt.

> „Das genaue zu verwendende Methodenset hängt dabei immer von der konkreten Fragestellung ab."

Die Auswertung von freien Antwortfeldern bei Befragungen im Bereich der Markt- und Medienforschung verlangt aber in jedem Fall eine Phrasenextraktion und das Auffinden von wiederkehrenden Themenkategorien (Topics). Sind diese noch in ihrer Stimmung zu bewerten (z. B. bei Mitarbeiterbefragungen bzgl. Arbeitsbedingun-

gen), erfolgt in der Regel noch eine Sentimentanalyse. Im Folgenden werden unterschiedliche Ansätze aus den Bereichen der Phrasenextraktion, Sentimentanalyse und Kategorisierung (Topic Modelling) skizziert.

Phrasenextraktion. Die Extraktion von Phrasen kann auf verschiedene Arten durchgeführt werden, wobei sich als besonders gut die Ansätze Part of Speech-Tagging und Chunking (PoS), Stoppwortgrenzen und Kookkurrenzen erwiesen haben.

PoS-Tagging bedeutet in diesem Zusammenhang, dass der zu analysierende Freitext mit Tags versehen wird, in dem jedem Wort entsprechend seiner morphosyntaktischen Rolle (Kasus, feste Präposition usw.) im Satz eine Wortform wie beispielsweise Nomen, Adjektiv oder Artikel zugeordnet wird. Das Erkennen von Kollokationen, d. h. Ausdrücke aus zwei oder mehr Wörtern, die gemeinsam mehr oder einen anderen Sinn als alleine haben und in einer syntaktischen Beziehung miteinander stehen, wird dann als Chunking bezeichnet.

Eigene Untersuchungen zeigen hierbei, dass das PoS-Tagging und Chunking die relevantesten Phrasen aus größeren Textmengen genauer identifizieren konnte, während die Methode der Stoppwortgrenzen auch aus kleineren Textmustern eine größere Menge relevanter Phrasen extrahieren konnte.

Sentimentanalyse. Für die Sentimentanalyse werden in den meisten Fällen Sentimentlexika genutzt, in denen eine große Sammlung von Wörtern bereits entsprechend ihrer vorrangig assoziierten Wertung mit einem positiven, negativen oder neutralen Label oder mit einem Sentimentwert (Sentimentscore) verknüpft ist.

Die Häufigkeit positiv oder negativ verknüpfter Begriffe in einem Text wird dann als Anhaltspunkt für eine Sentimentbewertung des Inhaltes genutzt (Liu 2012). Darüber hinaus können grammatikalische Textstrukturen für eine zusätzliche Einordnung der Aussagen dienen.

Dabei werden beispielsweise Konditionalsätze, sogenannte Sentiment-Shifter, wie etwa Negationen sowie intensivierende oder abschwächende Formen betrachtet. Die Methoden zur Extraktion der wichtigsten Sentimentmerkmale sind sehr ähnlich zu der bereits aufgeführten Problemstellung der Phrasenextraktion.

Da hier insbesondere die Extraktion der häufigsten Nomen, Nomenkombinationen oder auch Nominalphrasen gefragt ist, wird häufig eine gezielte Extraktion mit Hilfe von PoS-Tags eingesetzt. Die Bestimmung des Sentiments findet dann in der Regel Lexika-basiert statt. Allerdings zeigen eigene Untersuchungen, dass diese Bestimmungen wesentlich akkurater werden, wenn sie um syntaxbasierte Regeln erweitert werden.

Die Anpassungen sollten sowohl allgemeine, als auch spezifische Einflussfaktoren abdecken. Grundlage bildet der in Python implementierte Open Source Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto & Gilbert 2014). Eigene syntaktische Regeln lassen sich darin einfach erweitern.

Topic Modeling. Beim Topic Modeling werden die Muster kookkurrierender Terme betrachtet, um semantische Zusammenhänge zu modellieren, die sich dann als Themen (engl.: Latent Topics) äußern, die den Dokumenten zugrunde liegen.

Ein häufig gewählter Ansatz, um ein Dokument als eine Menge solcher zugrundeliegenden Topics zu modellieren, ist die Latent Dirichlet Allocation (LDA). Bei dieser Form des Topic Modelings bilden probabilistische Modelle die Ausgangsbasis, um die zugrunde liegende Struktur von zusammenhängenden Themen abzubilden.

Jedes Dokument wird als Zusammensetzung aus Latent Topics gesehen und jedes Latent Topic wiederum als Zusammensetzung aus Termen. Die Terme werden in dem Topic jeweils entsprechend ihrer Bedeutung für dieses Topic gewichtet.

Ein Term kann dabei auch mehreren Topics zugeordnet werden. Eine andere sehr gute Methode des Topic Modeling ist die Non-Negative Matrix Factorization (NMF) (Lee &

Seung 1999). Die NMF ist eine non-probabilistische Methode. Sie arbeitet mit dem Prinzip der Matrix-Dekomposition.

Dabei wird die Dokument-Term-Matrix V als Ausgangsbasis benötigt, welche in die beiden Submatrizen W und H zerlegt wird, sodass gilt:

$$V \approx WH$$

W kann dabei als Topic-Term-Matrix interpretiert werden und H als Dokument-Topic-Matrix. Aus dieser zerlegten Repräsentationsform kann, wie beim LDA-Algorithmus, ein Topic durch die entsprechenden Top N der höchstgewichteten Terme aus der Matrix und ein Dokument durch seine Zusammensetzung aus Topics dargestellt werden.

Insbesondere die NMF zeigte bei eigenen Untersuchungen eine stärkere Konvergenz und höhere Überschneidungen mit den Topics, die zuvor manuell erstellt wurden. Auch bezogen auf die Interpretierbarkeit (die Zuordnung von Labels für die gefundenen Topics) wurde eine bessere Leistung im Vergleich zur LDA erzielt.

Quellen

Hutto, C. J. & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International Conference on Weblogs and Social Media (S. 216–225). Palo Alto, CA, USA: AAAI Press.

Kao, A. & Poteet, S. R. (2007). Overview. In A. Kao & S. R. Poteet (Hrsg.), Natural language processing and text mining (S. 1–7). London: Springer Verlag Limited.

Keyvanpour, M. R., Javideh, M. & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. Procedia Computer Science, 3, 872–880.

Lee, D. D. & Seung, S. (1999). Learning the parts of objects by Non-negative Matrix Factorization. Nature, 401 (6755), 788–791.

Liu, B. (2012). Sentiment analysis and opinion mining. San Rafael, CA, USA: Morgan & Claypool Publishers.

Ting, I.-H., Wang, S.-L., Chi, H.-M. & Wu, J.-S. (2013). Content matters: A study of hate groups detection based on social networks analysis and web mining. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (S. 1196–1201). New York, NY, USA: ACM.

Warrer, P., Hansen, E. H., Juhl-Jensen, L. & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. British Journal of Clinical Pharmacology, 73 (5), 674–684.

Zeng, L., Li, L., Duan, L., Lu, K.,Shi, Z., Wang, M., Wu, W. et al. (2012). Distributed data mining: A survey. Information Technology & Management, 13 (4), 403–409.

Über den/die Autor*in

Technology Arts Sciences TH Köln

Prof. Dr. Gernot Heisenberg
TH Köln (TH)
University of Applied Sciences
Claudiusstr. 1
50678 Köln
Tel.: +49 221 8275 3389
E-Mail:
gernot.heisenberg@th-koeln.de

Dr. Gernot Heisenberg ist Professor für Information Research and Data Analytics an der TH Köln.

Weitere Informationen
www.gernotheisenberg.de

questback

Tina Hees
Questback GmbH
Gustav-Heinemann-Ufer 72a
50968 Köln
Tel.: +49 221 27169 729
E-Mail:
Tina.Hees@questback.com

Tina Hees ist Business Intelligence Developer bei Questback GmbH.

Weitere Informationen
www.questback.de

Thomas Reuter, Cogia Intelligence GmbH

Automatische semantische Analysen für die Online-Marktforschung

Ausgangslage

Unternehmen – vor allem im Konsumgüterbereich – sind „datenhungrig". Sie wollen tendenziell alles über ihre Konsumenten und Zielgruppen wissen. Traditionell ist die Beschaffung dieser Daten die Aufgabe der Markt- und Meinungsforschung. Online-Erhebungen, Fragebogen-Aktionen, Telefon-Interviews – solche quantitativen Befragungen gehören zu den Standardverfahren in der Markt- und Meinungsforschung. Mittels kontrollierter, vorgegebener Fragenschemata sollen statistisch valide Aussagen über Marken, Produkte oder Konsumentenverhalten gewonnen werden.

Zunehmend arbeiten quantitative Befragungen jedoch auch mit offenen Fragestellungen. Hier können die Befragten persönliche Eindrücke, Meinungen und Wertungen im Volltext formulieren. Antwortmöglichkeiten werden nicht vorgegeben. Doch wie werden diese offenen Kommentare ausgewertet?

Problem

Die Auswertung kontrollierter Befragungen mit sog. geschlossenen Fragestellungen kann durch den Einsatz entsprechender Statistik-Programme weitgehend automatisiert erfolgen. Anders bei offenen Fragestellungen. Hier gibt es natürlich die Möglichkeit, die Kommentare der Befragten nachträglich manuell zu codieren, um sie ebenfalls automatisch auswerten und quantifizieren zu können.

Doch dies ist zeit- und kostenintensiv und kaum noch zu leisten, wenn die Zahl der Fragebögen die Tausendermarke überschreitet. Daher bietet sich ein anderer Weg an – der Einsatz von Text Mining-Verfahren unter Einbeziehung von künstliche Intelligenz (KI)-Komponenten.

Methoden

> „Hier geht es darum, nicht nur binär zwischen negativ und positiv zu unterschieden, sondern Zwischentöne zu identifizieren."

Das oberste Ziel ist die semantische Analyse der offenen Kommentare hinsichtlich der angesprochenen Themen und des Sentiments, d. h. der emotionalen Orientierung oder Färbung eines Kommentars. Darüber hinaus kommt es darauf an, auch sog. „weak signals" (sog. schwache Signale) zu identifizieren – Hinweise, die auf neue Trends oder einen Trendumschwung deuten. Der aktuelle Hype um künstliche Intelligenz sollte allerdings nicht zu der Annahme verführen, dass entsprechende Systeme gleichsam auf Knopfdruck und selbständig sinnvolle und belastbare Resultate produzieren.

Semantische Analysen laufen automatisiert ab, benötigen aber einen gewissen manuellen bzw. redaktionellen Aufwand für das Setup und die Interpretation der Ergebnisse. In einem ersten Schritt werden die offenen Kommentare nach Themen strukturiert. Hier wird in der Regel mit einer vorgegebenen Taxonomie gearbeitet, anhand derer die Kommentare kategorisiert bzw. codiert werden. Im einfachsten Fall werden dabei die in der Taxonomie abgebildeten Themen durch Keyword-Listen beschrieben und definiert, die dazu genutzt werden, über Matching-Verfahren die Kommentare einzelnen Themen zuzuordnen.

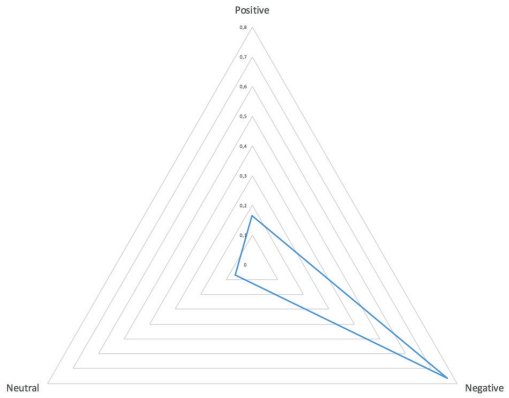
Dieses Vorgehen ist absolut transparent und in jeder Phase kontrollierbar. Alternativ kann maschinelles Lernen eingesetzt werden (man spricht hier auch von „schwacher KI").

Dabei werden für die abzudeckenden Themen Sets aus manuell gelabelten Kommentaren bereitgestellt, die dazu dienen, Klassifikatoren zu trainieren, die sodann auf neue, unbekannte Kommentare angewendete werden können. Für jeden Kommentar wird dabei ein Konfidenzwert ausgewiesen, der angibt, mit welcher Wahrscheinlichkeit er zu einem bestimmten Thema gehört. Ergänzend können die Kommentare einer sog. „Topic detection" unterzogen werden.

Es handelt sich hierbei um den Versuch, ohne taxonomische Vorgaben aus einer gegebenen Menge von Kommentaren Untermengen von Kommentaren zu bilden, die das gleiche Thema behandeln. Berechnet wird hier der Ähnlichkeitsgrad der Kommentare auf Basis ihres Wortbestandes. Dieses Verfahren birgt gewisse Unschärfen, ermöglicht aber die Entdeckung von Themen, die bislang übersehen oder vernachlässigt wurden. Maschinelles Lernen kommt auch zum Einsatz bei der Berechnung des Sentiments der Kommentare.

Auf Basis von gelabelten Kommentaren werden Klassifikatoren darauf trainiert, positive, negative oder neutrale Sätze voneinander zu trennen. Jedem Kommentar werden drei Werte zwischen 0 und 1 zugewiesen, die zeigen, wie stark jede Sentiment-Klasse hier vertreten ist. Vergleichbare Klassifikatoren können auch für einzelne Emotionen trainiert werden.

Abbildung 1: Stark negativer Kommentar



Hier geht es darum, nicht nur binär zwischen negativ und positiv zu unterschieden, sondern Zwischentöne zu identifizieren und die Kommentare hinsichtlich der darin enthaltenen Gefühlsäußerungen wie Ärger, Frustration, Freude oder Lob zu klassifizieren.

Zur Trendanalyse und zur Entdeckung von „schwachen Signalen" eignen sich die sog. Termrelationen oder Begriffsassoziationen. Hier wird auf Satzebene berechnet, wie bestimmte Begriffe miteinander zusammenhängen und mit welcher Wahrscheinlichkeit dieser Kombination auftritt. Da diese Assoziationen auf eine Zeitachse projiziert werden können (vorausgesetzt die Daten haben einen Zeitstempel), lässt sich feststellen, wie sich das semantische Feld rund um ein Produkt, eine Marke oder ein Image verändert und welche Zuschreibungen plötzlich neu auftauchen.

In ihrer Summe führen die vorgestellten Verfahren zur Strukturierung und Aufschlüsselung der offenen Kommentare und leisten damit eine Art Proto-Codierung, die bereits für die Erstellung von Reports genutzt werden kann.

Anwendungsszenarien

Die Einsatzgebiete derartiger automatisierter Verfahren sind so vielfältig wie die Aufgabengebiete der Marktforschung. Sobald mit freien Kommentaren gearbeitet wird, lässt sich hier ein Mehrwert erzielen. Richten wir den Blick nun auf zwei paradigmatische Anwendungsfälle – auf die Bereiche „Kundenzufriedenheitsmanagement" und „Human Resources".

Kundenzufriedenheitsmanagement. Unternehmen erheben regelmäßig Daten zur Zufriedenheit ihrer Kunden mit bestimmten Produkten oder den Serviceleistungen. Zudem eruieren sie bereits im Vorfeld Kundenwünsche, um den Launch neuer oder die Änderung bestehenden Produkte zu steuern.

Hier kommt es entscheidend darauf an, aus den offenen Kommentaren die zentralen Kritikpunkte, Anregungen und

Stimmungen zu destillieren, um das Qualitätsmanagement zu optimieren und das Marketing auf erfolgversprechende Aspekte zu fokussieren.

„Weak Signals" geben Hinweise auf potentielle neue Trends oder Trendumschwünge.

Human Resources. Zum anderen führen vor allem größere Unternehmen kontinuierlich Befragungen zur Mitarbeiterzufriedenheit durch. Hier geht es vor allem darum, die offenen Kommentare auszuwerten hinsichtlich der Kritik an Vorgesetzten, des Arbeitsklimas, der Arbeitsplatz-Ausstattung, der Organisationsstrukturen oder der Umsetzung von Compliance-Regeln. Da diese Befragungen wiederholt werden, lässt sich auch auf der Zeitachse abbilden, ob hier Fortschritte gemacht wurden oder bestimmte Problemfelder nach wie vor bestehen.

Derartige Analysen lassen sich anreichern durch externe Daten aus dem Web. Konsumenten oder Mitarbeiter äußern sich nicht nur intern, sondern posten auch in Social Media, ob in Foren oder Blogs, ob auf Twitter und Facebook, ob in Bewertungsportalen für Produkte oder Arbeitgeber.

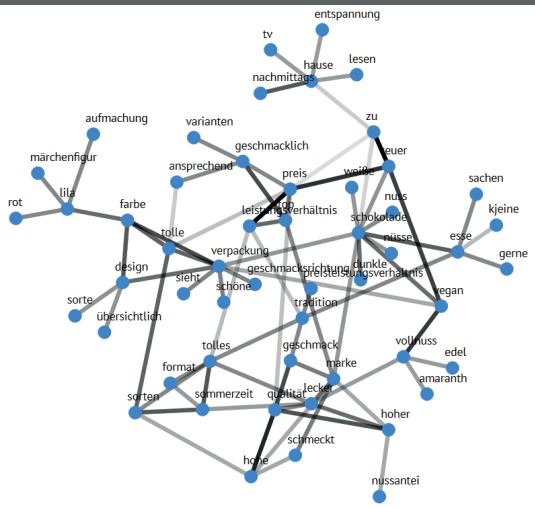
Das Ziel der Einbeziehung externer Daten ist der Vergleich zwischen den freien, internen Kommentaren und User-Kommentaren im Web in Hinsicht auf dieselben Thematiken. Gibt es signifikante Unterschiede in der Einschätzung von bestimmten Produkten, Leistungen oder Problemen? Oder weisen die Kommentare eine ähnliche Tendenz auf?

Die Daten aus Erhebungen und dem Web sowie die Analyse-Ergebnisse können in einem Dashboard zugänglich gemacht werden. Dies eröffnet die Möglichkeit, die jeweiligen Ergebnisse aufeinander zu beziehen und miteinander zu vergleichen. Sollten sie homogen sein, so bestätigen sie sich gegenseitig.

Die Einschätzung von auftretenden Differenzen aber wirft natürlich Fragen auf. Wir sehen uns im Web einer heterogenen Menge von Nutzern gegenüber und wissen nicht, ob deren Kommentare statistisch gesehen repräsentativ sind. Bei den quantitativen Befragungen hingegen können wir davon ausgehen, dass sie zumindest in Hinsicht auf die ausgewählten Fragestellungen und die dazu gehörigen Gruppen (Käufer eines bestimmten Produkts, einer bestimmten Leistung) repräsentativ sind. Doch ungeachtet der mangelnden Transparenz der Gruppe der Web-Nutzer lässt sich aus ihren Kommentaren, sofern deren Anzahl groß genug ist, ein Trend destillieren, der belastbar ist. Trotz der unterschiedlichen Ausgangsdaten können sich so die jeweiligen Analysen und deren Ergebnisse ergänzen.

Die Analysen der offenen Kommentare in Erhebungen sind zweifellos genauer, aber die Analysen der Webdaten können wertvolle Erkenntnisse liefern, insofern sie bestimmte „Schwachstellen" dort kompensieren. Eine Webdaten-Analyse kann in kürzeren Zyklen, bei genügend großen System-Ressourcen sogar in Echtzeit durchgeführt werden.

Abbildung 2: Automatisch erzeugtes semantisches Netz rund um das Thema „Schokaladenpreis".



Sie ist in der Lage, früh die **weak signals**, also erste Anzeichen einer kritischen Beurteilung aufzuspüren.

Sie arbeitet komplett hypothesenfrei, da sie lediglich non-reaktiv beobachtet, was tatsächlich gesprochen bzw. geschrieben wird. Sie kann im laufenden Betrieb an neue Erfordernisse angepasst werden. Beiträge können mit einer Historie bewertet und auf eine Zeitachse projiziert werden. Umgekehrt lassen sich aus diesen Analysen wiederum Hypothesen für künftige Erhebungen bilden. Die dort gestellten Fragen können geschärft und an reale Problemlagen und Meinungsbilder angepasst werden.

Nutzen

Entscheidend für die Entwicklung neuer erfolgreicher Geschäfts- und Produktstrategien ist nicht nur die umfassende Kenntnis des Wettbewerbs und des Marktes, sondern auch die genaue Einschätzung der Kundenerwartungen, Kundenwahrnehmungen oder Mitarbeitermotivationen.

Aus Unternehmenssicht bietet damit ein modernes KI-gestütztes Tool die Chance, große Informationsmengen zu erschließen und intuitiv verständlich aufzubereiten. Es können hier Einsichten gewonnen werden, die im Marketing, im Customer Service, in der Produktentwicklung oder im Personalmanagement für erhöhe Planungs- und Handlungssicherheit sorgen.

cogia intelligence COGIA GmbH

Poststr. 2-4
60329 Frankfurt a. M.
Tel.: +49 69 264 8485 11
E-Mail: th.reuter@cogia.de

Über den Autor

Thomas Reuter leitet bei der Cogia Intelligence GmbH die Redaktion und ist u. a. verantwortlich für die Qualitätskontrolle der Analyse-Ergebnisse. Er arbeitet eng zusammen mit der hauseigenen F&E-Abteilung in Hinsicht auf die Entwicklung neuer Analyse-Verfahren.

Weitere Informationen
www.cogia.de

Pascal de Buren, Caplena GmbH

Offenen Nennungen gekonnt analysieren

Offene Fragen: mächtiges Werkzeug - viel Arbeit

Offene Angaben in Umfragen bringen einen hohen Erkenntnisgewinn und tragen unvoreingenommene Meinungen zu Tage. Die Analyse dieser offenen Angaben erfordert allerdings einen hohen personellen Aufwand.

Um den Mehrwert offener Angaben vollständig ausschöpfen zu können und gleichzeitig die manuelle Arbeitsleistung zu reduzieren, ermöglicht codit.co, schnell große Mengen an offenen Nennungen zu analysieren – mit einer Genauigkeit, die der händischen Analyse sehr nahe oder gleichkommt. Das Tool lernt anhand von wenigen Beispielen, wie der Nutzer kodiert, und wendet diese Kodierung mittels künstlicher Intelligenz (KI) auf die restlichen Daten an.

KI-gestützte Kodierung für praktizierende Marktforscher

Die Anwendung richtet sich an betriebliche Marktforscher und Institute, die nach einem effizienten Weg suchen, offene Nennungen zu analysieren. Das Tool ist intuitiv zu bedienen und sowohl für dezidierte Kodierkräfte als auch Projektleiter geeignet. Die größten Einsparungen ergeben sich bei der Verwendung der Applikation für Tracking-Studien, bei denen diese mit jeder Welle besser kodieren lernt. Auch für größere Ad-Hoc-Projekte ab ca. 400 Nennungen oder Listen-Fragen (z. B. Brand-Tracker) eignet sich das Tool. Voraussetzung ist, dass die Mehrzahl der Antworten vergleichsweise kurz ist (im Schnitt max. zwei Tweets lang, was ca. 500 Zeichen entspricht).

Textanalyse – Der Kontext ist entscheidend

Die Kategorisierung von Texten ist wohl eine der ältesten Anwendungsfälle der künstlichen Intelligenz. Mit einfachen Schemata zur Wortzählung oder Ansätzen, die auf Wörterbüchern beruhen, wurden bereits in den 1980er Jahren erste Erfolge vermeldet. Jedoch war bis zum Aufkommen moderner Deep Learning-Verfahren vor rund einem halben Jahrzehnt die Qualität der Resultate oft doch sehr überschaubar.

Obwohl das Verstehen von Text für Maschinen auf den ersten Blick als einfacheres Problem als beispielsweise die Kategorisierung von Bildern erscheint, ist dies in Realität nicht so.

Die Bedeutung von Worten ist stark vom Kontext abhängig. Ein Beispiel hierfür: Die Logik „Falls die Worte ‚nicht' und ‚teuer' vorkommen, gehört das Statement zur Kategorie ‚Preis positiv'", wird oft falsch sein, wie bei der Nennung „Das Netz funktioniert oft nicht und das Abo ist zu teuer".

Es mag offensichtlich erscheinen, dass diese Art der Logik nicht in der Lage ist, die Bedeutung von Texten zu verstehen. Jedoch ist dies bei vielen Software-Anbietern noch immer die Standard-Lösung zur Textanalyse (z. B. SVM Modelle). Dies wird durch ein mehrstufiges Lernverfahren erreicht: In einem ersten Schritt wird das neuronale Netzwerk auf öffentlich verfügbaren Daten, wie Wikipedia, trainiert.

Damit erlernt dieses ein grundsätzliches Verständnis von Worten, beispielsweise dass „teuer" und „nicht günstig" eine ähnliche Bedeutung haben.

„codit.co setzt auf eine Technologie, die Wörter nicht nur einzeln, sondern im Kontext betrachtet."

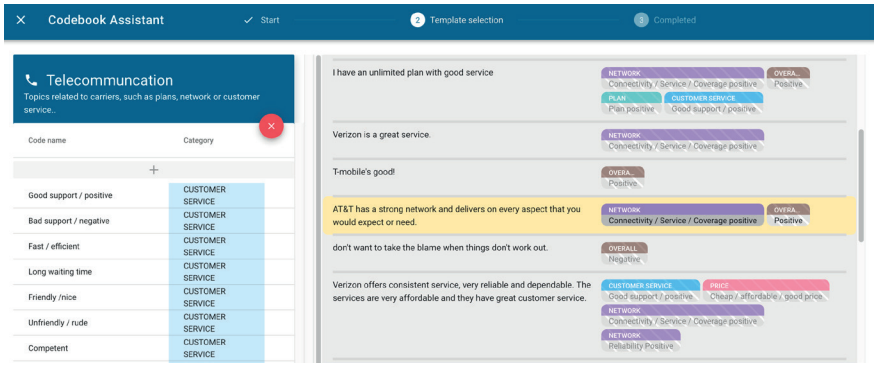
Daraufhin wird es auf allen auf codit.co händisch überprüften Umfragen geschult, was unterdessen einem Korpus von mehr als 3 Millionen Nennungen entspricht. Hier geht es darum, dem Algorithmus das System des Kodierens, was auch als Abstraktionsproblem gesehen werden kann, näherzubringen. Außerdem sind generelle Konzepte wie Preis, Service, Qualität oder das Sentiment über einen Großteil der Studien hinweg vergleichbar – diese sollten nicht jedes Mal von Grund auf neu gelernt werden müssen.

Letztendlich hilft der Nutzer dem System, die Eigenheiten der eigenen Umfrage aufzuzeigen. Dazu annotiert er einige Beispiele seiner Nennungen von Hand. Der Algorithmus ist vollständig mehrsprachig und kann die gelernte Klassifizierung auf über 36 Sprachen anwenden. Zudem wird von allen Kodierungen gelernt, die über die Plattform erfolgen, womit die KI mit jeder (neu)kodierten Nennung dazu lernt.

Eines der wichtigsten Ziele ist es, dem Nutzer eine KI-Lösung anzubieten, ohne ihn dabei einzugrenzen oder die Transparenz zu vernachlässigen. Um den vielseitigen und diversen Bedürfnissen der Marktforschung nachzukommen, wurde von Anfang an in enger Zusammenarbeit mit Instituten und betrieblichen Marktforschern (beispielsweise bei Telekommunikationsanbietern) kooperativ entwickelt. Im Folgenden werden drei unterschiedliche Anwendungsfälle in verschiedenen Organisationen aufgezeigt.

Kodier-Software für den Alltag. Factworks ist eine internationale Marktforschungsagentur. Zu ihren Projekten gehört häufig auch die Analyse von offenen Fragen. Mit der Anwendung konnten sie einerseits die Zeit, welche Mitarbeiter mit der Kodierung verbrachten, deutlich verringern, sodass diese mehr Zeit für die Auswertung der Resultate hatten.

Abbildung: Codebook Assistant



Andererseits haben indes alle Mitarbeiter einen zentralen „Place of truth" für die aktuellste Version ihrer kodierten Projekte. Dies hilft nicht nur dazu, dem Datenchaos (à la „Kodierung Version 3 (letzte) neu_final") entgegenzutreten, sondern auch die Arbeit vergangener Studien besser zu nutzen.

Anstatt die Daten auf der Festplatte eines Mitarbeiters zu lagern, können bei neuen Studien sowohl für die Entwicklung des Codebuchs, als auch für die Kodierung vorherige Wellen zur Hilfe genommen werden. Kommen beispielsweise Antworten zu einer Studie über die „Zufriedenheit der Nutzer von Website A" hinzu, kann die Projektleitung das Codebuch und damit auch die automatische Kodierung der vorjährigen Studie „Zufriedenheit der Nutzer von Website B" mit einem Klick auf die neue Studie anwenden. Anschließend können die automatische Kodierung und das Codebuch spezifisch auf die neue Studie angepasst werden.

Zu diesem Zeitpunkt ist ein Großteil der Arbeit (75 % des Codebuchs und 75 % der Kodierung) jedoch bereits erledigt. Neue Themen und unklare Verbatims werden dann dem Nutzer zur Kontrolle vorgelegt. Um die Daten unter verschieden Teams bzw. Personen zu teilen, ist es zudem möglich, Rechte individueller Nutzer auf Projektbasis zuzuweisen (z. B. nur Leserechte oder nur Rechte für den Upload neuer Daten).

NPS Tracker. Ein mittelgroßer deutscher Telekommunikationsanbieter führt monatlich NPS-Befragungen bei spezifischen Zielgruppen,

wie Wechslern oder Neukunden durch. Die Nennungen werden in ein sehr umfangreiches Codebuch mit über 100 unterschiedlichen Kategorien eingeteilt. Die Befragung dient dem Management einerseits als Früherkennungssystem von Trends und Problemen. Andererseits können spezifische Rückmeldungen direkt an die entsprechenden Abteilungen weitergeleitet werden.

> „Der vollständig mehrsprachige Algorithmus beherrscht über 36 Sprachen."

Indem neue Wellen jeweils mit allen vorhergehenden verknüpft werden, weiß das System bereits, wie die Kodierung den Vorgaben des Kunden entsprechend vorgenommen werden soll. Es werden dem Nutzer daher zuerst Nennungen präsentiert, bei welchen sich der Algorithmus in der Zuweisung unsicher ist. Dies sind oft Antworten, die entweder nicht in das bisherige Codebuch passen (z. B. wenn neue Themen auftreten) oder Nennungen, die keine Relevanz haben.

Der Nutzer kann die Kodierung dieser schwierigen Nennungen dann anpassen, sodass der Algorithmus von Monat zu Monat an Qualität hinzugewinnt und immer weniger von Hand feinjustiert werden muss.

Big Data-Projekte für Großkunden. Die Link Marketing Services AG ist eine Schweizer Marktforschungsagentur. Zu ihren Kunden zählt das Who-is-Who der Schweizer Wirtschaft – vom Mittelständler bis zum Weltkonzern. Einige dieser Kunden häufen aufgrund ihrer Größe regelmäßig umfangreiche Mengen an Kundenfeedback an, so beispielsweise auch ein bekannter Detailhändler. Quartalsweise überstellt dieser gegen 50.000 Nennungen an Link, deren händische Auswertung selbst einen geübten Kodierer an die zwei Monate Arbeit kosten würde.

Um dem Kunden eine hochwertige aber trotzdem kosteneffiziente Lösung anbieten zu können, trainiert Link das codit.co System mit einigen hundert bis tausend Nennungen pro Iteration und lässt den Rest anschließend automatisch auswerten. Dem Kunden wird damit ein großer Mehrwert geboten, da diese Nennungen bisher aus Kostengründen überhaupt nicht quantitativ ausgewertet wurden. Für das Link-Institut ist es darüber hinaus ein weiteres Merkmal, mit dem es sich von der Konkurrenz abheben und die Kundenbindung stärken kann.

Fazit

Offene Fragen können mittels innovativer Ansätze wirkungsvoll und kosteneffizient in Umfragen eingesetzt werden. Die vorgestellte Anwendung verbindet die verschiedenen Qualitäten der Marktforschung, wie das Verständnis für den Kunden und die Fragestellung mit der Konsistenz und Genauigkeit einer KI. Damit kann einerseits die operationelle Effizienz erhöht und andererseits auch neuartige Projekte im Bereich Big Data akquiriert werden.

CAPLENA GmbH

Zweierstrasse 165
8003 Zürich

Tel.: +41 79 518 91 75
E-Mail: pascal@caplena.com

Über den Autor

Pascal de Buren ist Mitgründer von codit.co. Er entwickelt KI-Systeme, die unstrukturierte Daten wie Text und Bild verarbeiten. Davor hat er an der ETH Zürich KI-Technologien auf Probleme in der Chemie und Physik angewendet.

Weitere Informationen
https://codit.co