DGI-Praxistage 2015

Martin Grothe*

Computerlinguistisches Echtzeitmonitoring im Internet: Digital Listening

Systematisch unknown Unknowns entdecken

DOI 10.1515/iwp-2016-0026

Digitalisierung und Vernetzung stellen Unternehmen vor vielfältige Herausforderungen. Allerdings entsteht durch die massive öffentliche digitale Kommunikation auch ein großes Potenzial: So können Unternehmen ihren Stakeholder-Gruppen und Umfeldern nun sehr direkt und zeitnah „zuhören“: Die Unstrukturiertheit und Dynamik der entstehenden digitalen (Big Data-)Textbeiträge kann methodisch und visuell für verschiedene Fachseiten erschlossen werden. Durch innovative Technologien können Unternehmen und Organisationen ihre Umfeldern – dank Digitalisierung – frühzeitiger und besser erfassen:

Computerlinguistische Algorithmen erschließen die öffentliche digitale Kommunikation von unterschiedlichen Stakeholder-Gruppen in Echtzeit und geben Auskunft über Fragen wie

- Was beschäftigt Ingenieur-Talente aktuell?
- Was bewegt die Produkt-Nutzer?
- Welche Bedrohungen gibt es?
- Welche Themen tauchen gerade neu auf?

Durch die eingesetzten Algorithmen lässt sich strukturell „Neues“ entdecken.

Die Antworten sind für verschiedene Funktionen wichtig: Marketing kann Anwendern zuhören, Recruiting die Fragen von Talenten besser verstehen, Kommunikation eigene Kampagnen passgenau platzieren und die Unternehmenssicherheit Bedrohungen früher erkennen. Die Fähigkeit zur Früherkennung reduziert Überraschungen und ermöglicht einen Vorsprung im Wettbewerb. Durch die eingesetzten Visualisierungen (insb. Netzwerk, Ranking, Kontext) können diese Informationen in neue und bestehende analytische Wertschöpfungsprozesse der jeweiligen Fachbereiche einfließen.

*Kontaktpersonen: Prof. Dr. Martin Grothe, complexium GmbH,
Neue Schönhauser Straße 20, 10178 Berlin,
E-Mail: grothe@complexium.de, www.complexium.de,

1 Digitale Transformation in Früherkennungsfunktionen

„Wir können jemandem Pardon gewähren, der besiegt wurde, aber niemanden, der sich hat *überraschen* lassen.“

Friedrich der Große

Dieser Beitrag soll dazu ermuntern, den digitalen Raum für das frühzeitige Erkennen von relevanten Informationen einzusetzen, insbesondere solchen, die a priori nicht benannt werden können, also überraschend sind. Bei näherer Betrachtung stellt sich heraus, dass eine solche Früherkennung durchaus für unterschiedliche Unternehmensfunktionen von zunehmender Bedeutung ist, seien es Produktentwicklung, Marketing und Vertrieb, aber auch Unternehmenskommunikation, Unternehmenssicherheit, Risk und Credit Management bis hin zur Personalbeschaffung.

Diese Möglichkeit, die die Digitalisierung der öffentlichen Kommunikation eröffnet, erfordert jedoch in vielen Fällen eine Weiterentwicklung der gewohnten internen Abläufe und Entscheidungswege. Der Weg in diese Richtung lässt sich als Digitale Transformation bezeichnen. Dieser Beitrag soll darauf hinweisen, dass die entsprechenden Technologien bereits erprobt vorliegen – die Herausforderung besteht nun vielmehr darin, diese Fortentwicklung nachhaltig einzuleiten.

Wenn sich aktuell formulieren lässt, dass die Digitalisierung ganze Branchen und quasi jede Funktion mit neuartigen – mitunter disruptiven – Lösungen konfrontiert, dann sollte dies doch auch für die Aufgabe der Früherkennung im Unternehmen gelten. Was zeichnet sich also ab? Welche Technologie kann helfen, diesen Sprung zu meistern? Auf dieser Vermutung und zwei wesentlichen Feststellungen beruht der hier skizzierte Lösungsansatz:

- Es gibt neue Daten, aber diese sind bizarr und „big“.
- Es gibt Algorithmen, die Erwartetes und Unerwartetes im digitalen Beitragsraum erkennen können.

Der neue Datenreichtum ist bereits hinlänglich behandelt worden. Es sind nicht nur die besonders populären Kanäle wie Facebook und Twitter, sondern nach wie vor auch die unzähligen Foren und Blogs, die neben anderen Formaten einen potenziellen Reichtum an unbekannten Informationen bereitstellen. Dieser schnelllebige Fundus wird mitunter als Social Big Data bezeichnet.

Mehr als in anderen Funktionen gilt für Früherkennungsfunktionen, z. B. Corporate Security, Credit Management und Risk Management, dass die entscheidende Information im Vorhinein nicht unbedingt beschrieben werden kann: Man weiß nicht, wonach gesucht werden soll, außer dem Umstand, dass es bedrohlich für das eigene Unternehmen, einen exponierten Unternehmensvertreter, einen Kunden oder Lieferanten sein könnte.

Donald Rumsfeld hatte hierfür den Begriff der unknown Unknowns geprägt:

*“As we know, there are **known knowns**. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also **unknown unknowns**, the ones we don’t know we don’t know.”*

Donald Rumsfeld, 2002

Durch die Digitalisierung sind damit – neben gleichzeitig auch gewachsenen Bedrohungen, die hier nicht verharmlost werden sollen – neue oder zumindest ergänzende Lösungsbeiträge als Chancen für diese Unternehmensfunktionen entstanden:

- So kann einerseits das digitale Rauschen als echtzeitnahes Früherkennungssystem von Bedrohungen eingesetzt werden,
- andererseits gilt es zu prüfen, welches Bild sich böswillige Dritte über die eigenen Schutzpersonen machen können, um entsprechend gegenzusteuern.

Hier wird nun eine Technologie vorgestellt, die auf die Nutzung dieses potenziellen digitalen Assets abzielt und bereits mit Anwendungserfahrungen unterlegt werden kann.

2 Algorithmen erschließen unknown Unknowns

Als wertschöpfende Technologien können die Computerlinguistik und soziale Netzwerkanalyse herausgestellt werden. So lassen sich Algorithmen definieren, die auch aus großen Textmengen die signifikanten Themen und Begrifflichkeiten destillieren. Nicht die absolute Häufigkeit eines Wortes zählt, sondern die relative: Das Ungewöhnliche, nicht das Übliche soll identifiziert werden.

So existiert für jede Sprache ein linguistischer Corpus, in dem feststeht, wie häufig ein Wort normalerweise vorkommt. Wird ein Wort häufiger als in dieser Normalverteilung aufgegriffen, so steigt auch seine Wichtigkeit bzw. Signifikanz. Weitere Hinweise gibt etwa die Analyse der Frequenzverteilung von Begriffen in Beitragsmengen.

Damit erschließen diese computerlinguistische Algorithmen durch Signifikanz- und Frequenzanalysen relevante Auffälligkeiten in einem umfangreichen Kontext ohne Vorgabe.

Gleichwohl kann eine – über das Banale hinausgehende – inhaltliche Bewertung nur durch einen menschlichen Verstand erfolgen. Die vorbereitenden Prozesse wachsen jedoch in der Reichweite und verkürzen sich fast auf Echtzeit. So sollte man zudem menschliche Geister nicht mit Aufgaben belegen, die Algorithmen viel klagloser übernehmen können.

Ein solchermaßen unterstützter Entdeckungsprozess ist der verbreiteten Nutzung von Social Media-Monitoring-Dashboards überlegen, weil dort lediglich Beiträge nach a priori vordefinierten Themenkarrieren erfasst und damit Überraschungen per Definition ausgeblendet werden. Klassische Social Media-Monitoring-Dashboards beschränken sich auf das Auszählen von absoluten Funden, eine inhaltliche Erschließung erfolgt nicht. Die Trefferdarstellung erfolgt dort in der Regel durch eine rein quantitative Visualisierung, eine wackelige Tonalisierung und lange Trefferlisten. Damit sind solche Werkzeuge eher zur Abbildung längerer Zeitreihen hilfreich, weniger aber, um frühzeitig schwache Signale zu erkennen.

Die digitalen Informationsräume erlauben eine deutlich intelligentere Auswertung als dies mit den Methoden der Presseauschnittsdienste möglich ist. Unknown Unknowns sind nun gerade solche Aspekte, von denen der Anwender nicht unbedingt im Vorhinein wissen kann, dass sie relevant sind, und nach denen er dementsprechend auch nicht mit einem klassischen Werkzeug suchen würde.

Fünf wesentliche **Nutzerfunktionen** können auf Basis computerlinguistischer Algorithmen – beispielhaft dargestellt durch das cloud-basierte Werkzeug GALAXY von complexium – bereitgestellt werden:

2.1 Entdeckung

Wie bereits beschrieben, können Crawler (Computerprogramme, die automatisch Internetquellen durchsuchen) und Algorithmen hypothesenfrei in digitalen Texten entstehende Auffälligkeiten identifizieren. Erkannten Begriffen wird ein Signifikanzgrad zuordnet.

Abbildung 1: Definierte Suchkategorien (Beispiel: GALAXY). Das System ermittelt die signifikanten Schlüsselwörter (Terms) auf den Hotspots einer Zielgruppe oder zu einem Thema. Zusätzlich lassen sich vordefinierte Suchkategorien verfolgen. Die Abbildung zeigt eine solche Kategorienübersicht aus einer Sicherheitsanwendung. Hinter jeder Kategorie, z. B. Aktivistengruppen, Bedrohungen, Demonstrationen oder einzelnen Standorten, steht eine Begriffs- und Floskelwolke, mit der die identifizierten Beiträge in nahezu Echtzeit untersucht werden. Die erkannten Hits lassen sich im semantischen Netz oder in der Textstellenliste DeepDive sehr effizient näher prüfen, um mögliche eigene Aktivitäten anzustoßen. Die blau unterlegte Rankingposition zeigt an, dass ein Begriff der jeweiligen Suchkategorie einen besonders hohen Signifikanzgrad in der Gesamtdiskussion des Themenfeldes, hier Sicherheit, erreicht hat. Die Liste dieser besonders signifikanten Terme wird durch computerlinguistische Algorithmen emergent und ohne Vorgabe aus den identifizierten Beiträgen ermittelt. Sie bildet das – hier nicht angezeigte – Ranking der „Diskussionsthemen“. In der Praxis kann somit ein kontinuierlicher Abgleich der Diskussionsentwicklung mit der eigenen Sicht erfolgen. Schwache Signale werden frühzeitig erkannt.

Neben solchen Diskussionsthemen, die emergent aus den Beiträgen erschlossen werden, gibt es in der Praxis für jeden Beobachtungsbereich auch bereits definierte Suchkategorien. Erst dieses Zusammenspiel dieser beiden Sichtungsklassen, known Unknowns und unknown Unknowns, verhindert „Betriebsblindheit“ und lässt auch Neues systematisch erkennen.

2.2 Ranking

Auf Basis dieser Metrik kann ein Ranking erstellt werden: Etwa ein tägliches Themen-Ranking. So wird im Sicherheitsbereich transparent, was Aktivistengruppen, im Marketingbereich, was Kundengruppen, im Personalmarketing, was Engpasszielgruppen im Tagestakt bewegt. Im Zeitablauf wird über die Tagesaktualität hinaus die Entwicklung der „emergent“ erschlossenen Begriffe deutlich:

- Was wird diskutiert? Was kommt neu hinzu?
- Wie entwickeln sich diese Themen?
- Gibt es Auf- oder Absteiger?

Hier ist es noch einmal wichtig zu betonen, dass zunächst etwa für das Risiko-Management bekannte kritische Begriffe und Parameter im Umfeld von Unternehmen und Marken kontinuierlich überprüft werden können. Bereits ein kurzes Nachdenken über solche „Alarmworte“ ergibt aber, dass diese Wortliste kaum erschöpfend implementiert werden kann: Es wird immer noch weitere Signalworte geben, die jedoch nur sehr selten auftauchen. Folglich werden Algorithmen eingesetzt, die gerade die Verwendung von üblicherweise eher seltenen Termen identifizieren. Dies ist das Feld der Computerlinguistik.

Der jeweilige funktionale Nutzer wird durch diese einfache Metrik auf schwache Signale hingewiesen, die sich als langsame Aufsteiger darstellen. Durch dieses frühzeitige Erkennen können je nach Kontext eigene Maßnahmen

Abbildung 2: Themenranking auf Ingenieur-Hotspots bei Auftreten des VW-Skandals (Beispiel: GALAXY). Nicht nur sicherheitsrelevante Aktivitäten haben Hotspots, sondern auch Talentzielgruppen für Personalmarketing und Recruiting. Die Abbildung zeigt einen täglichen „Thementicker“ für wichtige Ingenieur-Hotspots: die Top-Themen sind zumeist nicht interessant, aber auch hier zeigen sich frühzeitig ein Gewicht, d. h. Interesse zunehmende Themen: Im Recruiting-Kontext können dann etwa rechtzeitig eigene Inhalte zu auftauchenden Fragen eingebracht werden: Ein Vorsprung im War for Talents. Die Abbildung zeigt, dass das Bekanntwerden der Abgasmanipulationen das Unternehmen Volkswagen an die Spitze der digitalen Diskussionsthemen dieser Zielgruppe katapultiert hat. Ein Klick auf „VW“ führt zum semantischen Netzwerk, ein zweiter Klick zu der Textstellenliste, denn es ist keineswegs klar, aber für einen Arbeitgeber essenziell zu wissen, wie genau die Zielgruppe reagiert.

anlaufen. Es kann etwa zwischen Signalen unterschieden werden, die aufgenommen und positiv befördert werden sollten, weil sie förderlich auf die eigenen Ziele wirken, und solchen, die eher eine Abwehrreaktion erfordern.

2.3 Clustering

Ein Themenranking bildet aber nur den Einstieg. So kommen die identifizierten Begriffe in einem Beobachtungsbereich natürlich in unterschiedlicher Stärke in gemeinsamen Kontexten vor: Je enger und häufiger die gemeinsame Verwendung, desto stärker scheint die Beziehung. Algorithmen der Social Network Analysis (SNA) lassen sich nun aus der Soziologie auf diesen Anwendungsbereich übertragen: Sie erschließen, welche Gruppen von Begriffen stärker untereinander als mit dem Rest der Begriffe verbunden sind, und sind damit in der Lage, automatisch unterschiedliche **Begriffscluster** abzugrenzen. Zusammenhänge werden quellenübergreifend deutlich.

2.4 Mapping

Jede schematische Abbildung reduziert die Komplexität des Ursprungsbereiches. Einen facettenreichen Zugang zu den beobachteten digitalen Hotspots bieten **semantische Netze**. Auf Basis der beschriebenen Metriken lassen sich solche Netzvisualisierungen emergent für beliebige Zeiträume generieren. Dem Anwender bietet sich eine interaktive **Echtzeitlandkarte**, um die Kontexte unterschiedlicher Themen zu explorieren, abstrakt gesehen also über die jeweiligen Diskussionsräume zu fliegen. Die in der Gesamttextmenge verborgenen Zusammenhänge werden – in Echtzeit – aufgezeigt.

2.5 Qualifizierung

Mit solchen Umgebungsanalysen können zudem neue Attribute oder zunehmende Verbindungen auf Begriffsebene erkannt werden. Essenziell für die Qualifizierung eines möglichen schwachen Signals oder auch nur eines interessanten Aspekts ist dann der **Durchgriff auf die im Diskussionsraum verteilten Textbeiträge**. Der Nutzer, etwa

Abbildung 3: Themenlandkarte (GALAXY). Nähern sich die Term-Listen und Rankings der aktuellen Diskussion im selektierten Digitalraum der Zielgruppen an, so zeigt die Map die Zusammenhänge und Cluster. Begriffe, die durch gemeinsame Kontexte verbunden sind, werden durch eine (graphentheoretische) Kante vernetzt: Es entsteht ein beweglicher und interaktiver Zugang zu der Quellen- und Beitragsbasis als übergeordnete Orientierungsebene. Der Nutzer kann durch Klick auf einen Term die jeweilige Textstellenliste dynamisch durchgehen und sinnvolle Aktionen oder Einschätzungen markieren. Auf diese Weise kann tatsächlich eine große Beitragsmenge effektiv inhaltlich erfasst und bearbeitet werden.

ein Sicherheits- oder Kommunikationsmanager, kann sich aufgrund der Unterstützung durch die Algorithmen auf die wertschöpfenden Aufgaben konzentrieren.

So kann diese Technologie zur Früherkennung von Bedrohungen, Risiken, Issues oder Talenten aus Sicht von Früherkennungsfunktionen eingesetzt werden, da hier besonderes Augenmerk darauf gelegt werden muss, Überraschungen zu vermeiden und gerade das Unerwartete zu finden und schnell zu qualifizieren.

3 Ausblick: Mehr Sicherheit durch Premonitoring

Digitalisierung bedeutet – neben einigen anderen Dingen – dass ganz neue Kategorien und Mengen von Daten verfügbar sind. Diese potenziellen Informationswerte können für zahlreiche Unternehmensfunktionen genutzt werden. Es bedeutet auch, dass etablierte Prozesse wie die Früherkennung einer digitalen Transformation unterzogen werden sollten, um anschlussfähig zu bleiben. Neue Methoden und Werkzeuge hierzu sind bereits verfügbar.

Diese Werkzeuge – das cloud-basierte GALAXY-System wurde exemplarisch vorstehend vorgestellt – sollen hier als **Premonitor** bezeichnet werden. In der konkreten Anwendung, etwa für den Sicherheitsbereich, lassen sich mehrere Grundtypen unterscheiden (s. Tabelle 1).

Tabelle 1: Grundtypen von Premonitor-Werkzeugen für den Sicherheitsbereich.

Typ	Ausrichtung
Base Premonitor	Es wird durch übergreifende Quellen eine Übersicht zur allgemeinen Lage aufgebaut.
Corporate Premonitor	Unternehmensspezifische Besonderheiten werden in der Quellenauswahl, den Eingangsfiltern und Kategorien abgebildet.
Event Premonitor	Vorlauf und Akutphase zu einem Event, z. B. Hauptversammlung, Weltwirtschaftsforum, werden hochfrequent durchleuchtet.
Country Premonitor	Quellen werden aufgrund von regionalen Kriterien zusammengestellt.
Issue Premonitor	Quellen werden thematisch gebündelt.

Innovative Verfahren der Computerlinguistik und Netzwerkanalyse können im Aufgabenbereich Früherkennung die unstrukturierten Online-Diskussionen und Meldungen metrisch fassen: So wird nach bereits bekannten Themen und Ereignissen automatisiert gesucht, aber auch Unerwartetes gefunden. Es gelingt, schwache Signale im digitalen Rauschen früher zu erkennen, Kontexte von Themen, Debitoren, Marken, Talentgruppen oder Personen besser zu verstehen und eigene Maßnahmen besser zu steuern. Dies basiert auf zwei Pfeilern:

- **Effektivität:** Die vorgestellte Technologie erschließt beinahe in Echtzeit alle Beiträge aus relevanten

The screenshot displays the GALAXY 'Deep Dive' interface. At the top, there's a navigation bar with 'Security', a date selector 'Tue Dec 08 2015', and tabs for 'Overview', 'Map', 'Clusters', 'Terms', 'Rankings', and 'Statistics'. A 'COLLABORATION' section includes 'Deep Dive' and 'Set up GALAXIES'. A search bar contains 'stuttgart' with 'Search', 'uncheck all', and 'check all' buttons. Below the search bar, a filter bar shows 'Choose...', 'Filter...', 'List...', 'No status assigned', 'Show Category', and 'Apply Category'. The results section shows two hits. The first hit is dated '2015-12-06 23:30:00' and contains text about a video-kundgebung in Stuttgart, Köln, Nürnberg, and Oberhausen. It includes a 'View source' link and a headline from 'linksunten.indymedia.org'. The second hit is dated '2015-12-06 15:28:00' and contains text about the Turkish Consulate in Stuttgart. Both hits have an 'Assign status' dropdown set to 'No status assigned' with options to 'Add more categories' or 'Select for bulk categorisation'.

Abbildung 4: Vertiefung der Beitragsinhalte (Beispiel: GALAXY). Das System generiert durch Algorithmen verschiedene Sichten, z. B. Rankings und Themennetze, auf das digitale Rauschen der relevanten Hotspots. Diese aufgebauten Sichten machen die Beitragsmenge inhaltlich zugänglich: Der Deep Dive (Textstellenliste) zeigt zu jedem ausgewählten Term die passenden Fundstellen im Quellenbereich. Der Anwender erkennt damit thematische Auf- und Absteiger, Top-Themen oder kann zu beliebigen Themen effizient recherchieren. Mögliche Ableitungen, Sicherheitseinstufungen oder Weiterleitungen an Kollegen können direkt durch eine „Statusvergabe“ zugeordnet werden, um aus den Insights zu Actions zu kommen. In der Abbildung wird etwa ein Bekennerhinweis gefunden.

Blogs, Foren, Nachrichten- und Bewertungsportalen. Als lernendes System wird die Quellenbasis permanent fortentwickelt und der Wissensvorsprung ausgebaut.

- **Effizienz:** Durch die umfassende Abnahme nicht nur des einfachen Monitorings, sondern auch der Inhaltserschließung durch das Werkzeug, haben die Nutzer eine zeitliche Ersparnis und damit Ressourcen frei, um zu planen, zu agieren und Ergebnisse zu generieren.

Prof. Dr. Martin Grothe
complexium GmbH
Neue Schönhauser Straße 20
10178 Berlin
grothe@complexium.de
www.complexium.de

So kann eine Früherkennung von Risiken (und Chancen) unternehmensspezifisch als gewichtige Fähigkeit entwickelt werden. Die Technologien hierzu sind verfügbar. So darf man die entsprechenden Verantwortlichen getrost ermuntern, sich mit diesen Lösungen vertraut zu machen und diese digitale Transformation voranzutreiben.

Deskriptoren: Computerlinguistik, Textanalyse, Visualisierung, Prognose, Soziales Netz, Digital

Prof. Dr. Martin Grothe ist geschäftsführender Gesellschafter der complexium GmbH und Honorarprofessor an der Universität der Künste UdK in Berlin. complexium unterstützt seit 2004 renommierte Klienten mit Auswertungen und Lösungen zu Kommunikations-, Marketing-, Arbeitgeber- und Sicherheitsfragen. Erfolgsfaktor ist eine innovative Technologie, die die digitale Kommunikation inhaltlich erschließen, schwache Signale erkennen und Zielgruppen pragmatisch entschlüsseln kann. Als Honorarprofessor an der UdK für das Fach „Digitale Kommunikation/Leadership, Social Media Management“ steht das berufsbegleitende Master-Programm „Leadership in digitaler Kommunikation“ im Mittelpunkt. Grothe ist zudem Vorstandsvorsitzender des Deutschen Competitive Intelligence Forums dcif e. V. und Beirat von Quality Employer Branding queb e. V.