Datenbasierte Entscheidungshilfe im Bereich Social Media

Transparente Analyse geschäftsrelevanter Daten

Ein Beitrag
von Alexander Gusser

Mit Sentiment-Analysen lassen sich Social-Media-Postings und -Diskussionen auf Meinungen, Trends und Stimmungen analysieren. Unternehmen können mit ihrer Hilfe herausfinden, welche Einstellung das Netz zu ihren Produkten hat oder welches Image über sie vorherrscht. Auch Persönlichkeiten des öffentlichen Lebens oder politische Parteien können mit Hilfe von Sentimentanalysen herausfiltern, wie sie im Netz eingeschätzt werden. Hier ein Überblick über die verschiedenen Ansätze.

Das Verstehen von Aktivitäten und Textbeiträgen in den Social Media ist heutzutage wichtiger denn je und gewinnt immer stärker an Bedeutung. Innerhalb von Twitter, Facebook, Xing, LinkedIn etc. existieren große Mengen an bisher teils ungenutzten, geschäftsrelevanten Daten [Sch15]. Vor diesem Hintergrund erscheint eine detaillierte Analyse über die Postings und Diskussionen in Foren, Blogs und sozialen Netzwerken hilfreich – und hier setzt die Sentimentanalyse an. Meinungen, Trends und Stimmungen können sichtbar und quantitativ messbar gemacht werden. In diesem Zusammenhang ist es wichtig, bestehende Ansätze zu kennen und deren Abläufe nachzuvollziehen, um die Nützlichkeit dieser Analyseform bestimmen zu können. Im Folgenden werden diese Aspekte näher beleuchtet.

Sentimentanalyse untersucht unstrukturierte und semistrukturierte Texte

Die Sentimentanalyse beziehungsweise die Polaritätsanalyse ist eine Teildisziplin des Text Mining und lässt sich speziell dem Themengebiet des Natural Language Processing (NLP) zuordnen. Diese Begriffe beschreiben eine Vielzahl von Methoden zur Analyse von unstrukturierten und semistrukturierten Texten [Kel12]. Eines der Ziele besteht darin, die im Text enthaltenen Informationen, soweit möglich, in messbare Textmerkmale zu übertragen und damit zu quantifizieren. Im Anschluss werden Algorithmen auf die Merkmale angewendet, mit deren Hilfe sich die inhaltliche Wertung bzw. Polarität (positiv, neutral oder negativ) eines Textes bestimmen lässt.

In der Rubrik TDWI INSIDE veröffentlichen Mitglieder der neu gegründeten TDWI Young Guns ihre Beiträge. Die TDWI Young Guns sind ein selbstorganisierter Kreis im TDWI e.V. Sie sind eine Anlaufstelle für Studenten und Young Professionals, die sich für die TDWI Community begeistern. Infos und Kontakt unter: **Young-Guns@tdwi.eu**

Der sogenannte *Lexicon-based*-Ansatz benötigt zur Festlegung der Orientierung entsprechende Bibliotheken mit vordefinierten Begriffen und den dazugehörigen numerischen Bewertungen der Polarität. Die Wörter sind somit unterschiedlich stark der jeweiligen Polarität zugeordnet. Begriffe, die nicht in der Bibliothek existent sind, sind per Definition als *neutral* oder *unbekannt* gekennzeichnet. Ein Text kann daher in seiner vollständigen Bewertung theoretisch einen beliebigen positiven, neutralen oder negativen Wert annehmen. Allerdings wird dieser Ansatz oft als unflexibel angesehen, da sich die meisten Wörter oder Begriffe lediglich themenbezogen als positiv oder negativ betrachten lassen [Gez15].

Zudem hat der Ansatz Schwierigkeiten, „Linguistische […] Begriffe, Entitäten, Fakten und Kernaussagen aus den Texten“ [Kel12] zu erkennen. Eine Zuordnung einzelner Wörter wird durch den variierenden Kontextbezug, durch Negation sowie den entstehenden Interpretationsspielraum erschwert, den die deutsche Sprache mit sich bringt. Zum Beispiel wird die Aussage „Er will sie nicht“ anders verstanden als: „Er will, sie nicht.“ Dabei erfüllt die Zeichensetzung eine bedeutende Funktion. Diese wird jedoch in den meisten Lexicon-based-Ansätzen nicht beachtet, was wiederum das Verstehen der Semantik zu einer komplexen und manuell intensiven Angelegenheit macht [Min12].

Genau hier setzt der *Supervised-Learning*-Ansatz (überwachtes Lernen) an. Zur Klassifikation ist keine Bibliothek mit bewerteten Wörtern erforderlich, sondern eine Sammlung von Texten mit bereits vordefinierter Polaritätszuordnung als Trainings- und Testdatenmenge. Die Einordnung lässt sich durch automatische oder manuelle Verfahren vornehmen. Das bedeutet, dass das Modell zum Beispiel anhand der vom Fachpersonal gelesenen und bewerteten Texte trainiert wird.

Dabei liegt der Vorteil insbesondere in der Anpassung des Verwendungszwecks, da sich der daraus resultierende Kontextbezug schneller und effizienter herstellen lässt. Dem Risiko eines überangepassten Modells (Overfitting) kann durch den Einsatz gängiger Kreuzvalidierungsverfahren, Pruning oder Early-Stopping-Verfahren entgegengewirkt werden. Somit steigt die Qualität der messbaren Ergebnisse, was die wahrgenommene Nützlichkeit des Ansatzes stei-

gert. Die im Folgenden beispielhaft skizzierte Sentiment-Anwendung, kurz SENA, soll diesen Ansatz der Sentimentanalyse greifbarer machen.

SENA

Im Bereich der Textanalyse existieren durchaus zahlreiche Anwendungen, die eine automatisierte Sentimentanalyse sowie inhaltliche Monitoringsysteme anbieten. Jedoch halten die Hersteller ihre Algorithmen in der Regel im Verborgenen, sodass die Herleitung der Polarität dadurch schwer bis unmöglich nachvollziehbar ist [Kel12]. Eine Eigenentwicklung überwindet die genannten Nachteile. Die flexible und zeitnahe Erweiterbarkeit spricht trotz des deutlich höheren Entwicklungsaufwands für die eigene Umsetzung. Im Folgenden wird das Vorgehen schrittweise erläutert.

Social Media als Datenbasis

Social Media ist eine wichtige Grundlage der empirischen Analyse und dabei insbesondere der qualitativen Sozialforschung. Dabei enthalten Postings und Diskussionen beispielsweise in sozialen Netzwerken große Mengen geschäftsrelevanter Daten wie Lob, Kritik, Meinungen und Bewertungen über Produkte und Dienstleistungen von Unternehmen [Sch15]. Um mögliche datenbasierte Entscheidungen treffen zu können, wird die Datenbasis für SENA hier exemplarisch auf den Microblogging-Dienst Twitter eingeschränkt. Die Tweets repräsentieren authentische und unverfälschte Aussagen, die als Datengrundlage für die Sentimentanalyse dienen [Kel12]. Die Daten werden zunächst mit Hilfe einer Informationsextraktion über vorab festgelegte Suchparameter wie beispielsweise Suchwörter, Zeitraum und/oder Sprache gefiltert und in einer Datenbank gespeichert.

Trainings- und Testdaten

Die gezielte Integration des Fachwissens der Mitarbeiter wird, wie immer bei überwachten Lernverfahren im Text Mining, auch in SENA als wichtiger Faktor eingesetzt. Dazu werden unbehandelte Tweets aus der Datenbasis bereitgestellt. Neben den Wörtern enthalten die Rohdaten Informationen, die der Mensch intuitiv verarbeitet, wie Zeichensetzung, Sarkasmus oder Zynismus. Die Mitarbeiter stufen die Texte in die jeweiligen Kategorien positiv, negativ oder neutral ein. Anschließend werden die Texte in Trainings- und Testdaten unterteilt. Auf diesem Trainingsdatenbestand wird im nächsten Schritt der Klassifikator trainiert und überprüft.

ALEXANDER GUSSER ist Consultant bei der gmc² gerhards multhaupt consulting GmbH. Dort liegt sein Fokus in der Entwicklung von Business-Intelligence-Lösungen.
E-Mail: A.Gusser@gmc2.de

Klassifikation mit bedingten Wahrscheinlichkeitswerten

Die Auswahl der Verfahren, mit denen mögliche Klassifikatoren trainierbar sind, ist groß. Eine kleine Aufzählung der gängigsten und bekanntesten Vertreter umfasst Entscheidungsbäume, Regelwerke, graphenbasierte Verfahren, Support Vector Machines und Künstliche Neuronale Netzwerke. SENA verwendet den Naive-Bayes-Klassifikator, der eine Klassifikation mit bedingten Wahrscheinlichkeitswerten ausführt. Die mathematisch transparente Herleitung, warum ein Text der jeweiligen Polarität zugeordnet wurde, war ausschlaggebend für die Wahl dieses Verfahrens.

Zu Beginn müssen die Texte in messbare Merkmale überführt werden. Das Bag-of-Words-Modell [Ram15] hilft dabei, die Frequenz, also die Häufigkeit jedes Wortes innerhalb des Textes zu bestimmen.

Danach erfolgt die Berechnung der bedingten Wahrscheinlichkeiten für jedes Wort mit der zuvor festgelegten Polarität basierend auf der Trainingsmenge [Wag12]. Die Polaritätsberechnung setzt sich aus der Multiplikation jeder auftretenden bedingten Wahrscheinlichkeit der Wörter im Text zusammen. Anschließend wird die Polarität mit dem höchsten Ergebnis, dass der Text dieser zugehörig ist, ausgewählt. Die Testdaten dienen abschließend zur Evaluation des erzeugten Klassifikators. Dieser lässt sich ab sofort zur Bestimmung der Polarität neuer Texte nutzen. Das Training ist in regelmäßigen Abschnitten zu wiederholen, um die Qualität sicherzustellen bzw. zu steigern.

Fazit

Die Sentimentanalyse ermöglicht einen Überblick über die Stimmung innerhalb von Social-Media-Daten. Dabei bestehen grundsätzlich zwei Um-

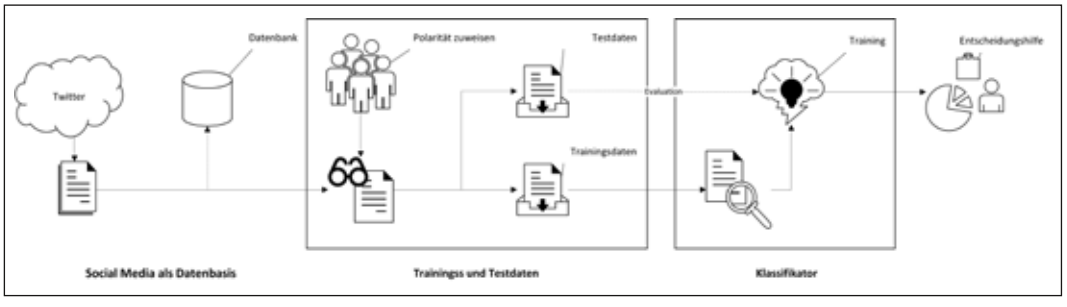

Abb. 1: Verarbeitungsschritte der Sentiment-Anwendung (SENA)

setzungsmöglichkeiten, mit denen eine systematische Klassifizierung der Texte zu den jeweiligen Polaritäten, bedingt durch die Einbeziehung von Begriffsgewichtungen, realisierbar ist. Der wesentliche Unterschied zwischen den beiden Ansätzen ist der Verwendungszweck. Der Lexicon-based-Ansatz kann die generelle Stimmung ohne großen Trainingsaufwand des Modells wiedergeben. Der Bezug zum Unternehmen und/oder zu seinen Produkten, Kampagnen, vielleicht auch zu seinen Konkurrenten lässt sich entschieden präziser durch den Supervised-Learning-Ansatz herstellen.

Der hier vorgestellte SENA-Prototyp zeigt, dass es sich lohnt, eine Eigenentwicklung zu betreiben: Im Vergleich zu den undurchsichtigen Blackbox-Verfahren der Hersteller gilt für die hier vorgestellten Ansätze, dass das Unternehmen damit fundierte, datenbasierte Entscheidungen auf Grundlage einer transparenten, nachvollziehbaren und differenzierteren Herleitung der Algorithmen treffen und daraus Handlungsempfehlungen ableiten kann.

Es zeigt sich schon jetzt, dass die hier vorgestellten Verfahren sowie Analysen für die inhaltliche Wertung von Texten zukünftig deutlich mehr an Bedeutung gewinnen werden. Die daraus resultierenden Erkenntnisse bieten Unternehmen das Potenzial, sich eine bessere Positionierung im Wettbewerb zu verschaffen.

Literatur

[Gez15] Gezici, G. et al.: Sentiment Analysis. Using Domain-Adaptation and Sentence-Based Analysis. In: Studies in Computational Intelligence. Springer Science Business Media 2015

[HiR06] Hippner, H. / Rentzmann, R.: Text Mining. Informatik-Spektrum 2006

[Kel12] Keller, B. et al.: Zukunft der Marktforschung – Entwicklungschancen in Zeiten von Social Media und Big Data. Springer-Verlag 2012

[Min12] Miner, G. et al.: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press 2012

[Ram15] Ramesh, B. et al.: Shape classification using invariant features and contextual information in the bag-of-words model. Pattern Recognition. 2015

[Sch12] Schulten, M. et al.: Social Branding: Strategien – Praxisbeispiele – Perspektiven. Gabler Verlag 2012

[Sch15] Schirmer, D. et al.: Die qualitative Analyse internetbasierter Daten. Methodische Herausforderungen und Potenziale von Online-Medien (Soziologische Entdeckungen). Springer-Verlag 2015

[Wag12] Wagenführer, D.: Konsumenteneinstellungen im Social Web – Neuartige Ansätze im internetbezogenen Kontext. Springer-Verlag 2012

Die nächste Ausgabe von BI-Spektrum erscheint am 26.06.2017