End-to-end Data-Warehouse-Szenario von Social-Media-Daten mithilfe moderner SAP-Technologien

Frederic Wall

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
frederic.a.w@web.de

Oliver Meier

SAP SE
Dietmar-Hopp-Allee 16
69190 Walldorf

Frank Morelli

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
frank.morelli@hs-pforzheim.de

Schlüsselwörter

Sentimentanalyse, Data Warehousing, Social-Media-Daten, Twitter, Qualtrics

Problemstellung und Zielsetzung

Weltweit versuchen Unternehmen Social-Media-Daten auszuwerten um wertvolle Informationen zu erhalten. Sie nutzen diese, um bspw. Kosten durch optimierte Werbekampagnen einzusparen, neue Erkenntnisse über die Kunden zu erhalten oder um eine optimierte Informationslage für zukünftige Unternehmens-entscheidungen zu schaffen. Ein Problem bei der Verarbeitung und Auswertung von Daten aus sozialen Netzwerken besteht darin, dass ein großer Teil der Daten unstrukturiert und in heterogenen Formaten vorliegt, was eine automatisierte Verarbeitung erschwert. Im Rahmen der Masterarbeit wird ein End-to-End Data Warehouse Szenario auf Basis mehrerer SAP Tools (HANA, BW/4HANA, SAP Analytics Cloud) entwickelt.
Als Use Case fungiert eine Sentimentanalyse. Diese ermöglicht die automatisierte Auswertung von Texten aus sozialen Netzwerken, bspw. Twitter. Mithilfe diverser Softwarelösungen kann die Extraktion, Speicherung und Auswertung von Tweets automatisiert erfolgen. Die Ergebnisse der Sentimentanalyse können anschließend quantifiziert und im Kontext interner Unternehmensdaten sowie externer Umfrageergebnisse von Qualtrics visualisiert werden.

Methodisches Vorgehen

Um ein anwendungsrelevantes Ergebnis zu erzielen erfolgt im Rahmen der Thesis die Konzeption eines Use Case mit betriebswirtschaftlicher und technischer Ausgangssituation. Anhand des Anwendungsfalls werden zwei Systemarchitekturalternativen aus einer vorgegebenen Anzahl an Softwaretools abgeleitet, welche die für das Erreichen der Zielsetzung geeignet sind. Um die vergleichsweise optimale Systemarchitektur für den gewählten Use Case zu bestimmen, erfolgt eine Evaluation der beiden Szenarien anhand von fünf Kriterien.

Sentimentanalyse

Bei einer Sentimentanalyse handelt es sich um eine Reihe von Methoden, Techniken und Werkzeugen zur Erkennung und Extraktion subjektiver Informationen, bspw. Meinungen und Haltungen, aus natürlicher Sprache. Sentimentanalysen können in drei Ebenen erfolgen: Dokumentenebene, Satzebene und Aspektebene. Während die Dokumentenebene klassifiziert, ob ein Dokument als Ganzes ein positives oder negatives Sentiment zum Ausdruck bringt, werden bei der Satzebene einzelne Sätze ausgewertet. Auf Aspektebene erfolgt die Auswertung, welchen Aspekt die Ersteller des Dokuments negativ bzw. positiv bewerten.
In der Wissenschaft haben sich diverse technische Ansätze etabliert, um die beschriebene Auswertung natürlicher Sprache durchzuführen. Grundsätzlich lassen sich diese Vorgehensweisen in zwei Kategorien einteilen: technische Ansätze basierend auf Machine Learning vs. lexikalische Herangehensweisen. Die Technik der Sentimentanalyse auf SAP HANA ist dem wörterbuchbasierten Ansatz zuzuordnen. Hiebei werden Texte nach Meinungsworten durchsucht, um ein hinterlegtes Lexikon nach Synonymen und Antonymen abzugleichen. Den Einträgen im Wörterbuch wird jeweils ein positives bzw. negatives Sentiment zugeordnet. Findet das System beim Durchsuchen des Texts zugehörige Einträge aus dem Lexikon, kann anhand des hinterlegten Sentiments auf die Stimmung des Texts geschlossen werden.

Architekturvergleich

Für die Extraktion der Tweets, die Durchführung der Sentimentanalyse und die Visualisierung der Daten erweist sich die Kombination unterschiedlicher Softwarewerkzeuge als notwendig. Ausgearbeitet werden zwei Systemarchitekturen, wobei Variante A im Gegensatz zu Variante B kein Data Warehouse für die zentrale Datenkonsolidierung verwendet.
Bei Architektur A erfolgt der Import der Daten von Twitter mithilfe des Data Provisioning Agents in SAP HANA, wo die Datenaufbereitung und die Sentimentanalyse durchgeführt wird. Die Bereitstellung der Ergebnisse der Sentimentanalyse gemeinsam mit internen Finanzdaten wird in SAP Analytics Cloud (SAC) umgesetzt. Auf dieser Ebene werden ebenfalls

Umfrageergebnisse von Qualtrics mit den übrigen Daten harmonisiert und visualisiert. Abbildung 1 stellt den schmatischen Aufbau der Architekturvariante A dar.

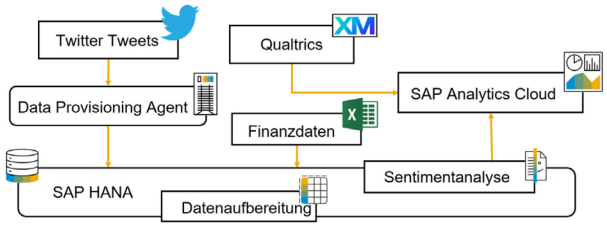

Abbildung 1: Systemarchitektur A

Bei Architektur B wird neben den Tools von Architektur A zusätzlich ein Data Warehouse eingebunden. Dieses ermöglicht die zentrale Harmonisierung und Konsolidierung aller Datenquellen und die anschließende Bereitstellung.

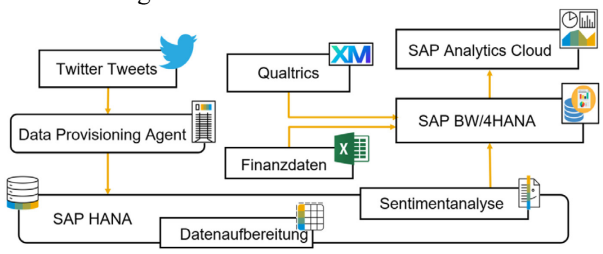

Abbildung 2: Systemarchitektur B

Um eine Entscheidung bzgl. der Architekturalternativen zu treffen, werden beide Ansätze anhand von fünf Bewertungskriterien qualitativ verglichen.

Einfachheit (bewertet die Vermeidung von Komplexität bei Implementierung, Modifizierung und Wartung der Architekturmodelle)

Datenintegration (evaluiert die Möglichkeit der Integration diverser Datenquellen und deren Modellierung)

Flexibilität (bewertet die Möglichkeit der Anpassung und Erweiterung der Architektur bzw. des Use Case)

Stabilität (beurteilt die Robustheit der einzelnen Architekturelemente sowie deren Integration)

Compliance: (evaluiert die Möglichkeit der Einhaltung von Gesetzen sowie interner und externer Richtlinien)

Die Evaluation erfolgt mithilfe einer 5-stufigen Harvey-Ball-Skala (○: nicht vorhanden, ◔: schwach ausgeprägt, ◑: mäßig ausgeprägt, ◕: deutlich ausgeprägt, ●: erheblich ausgeprägt). In Tabelle 1 ist das Ergebnis des Architekturvergleichs dargestellt.

Kriterien	Architektur A	Architektur B
Einfachheit	◑	◔
Datenintegration	◑	●
Flexibilität	◔	◕
Stabilität	◕	◕
Compliance	◑	●

Tabelle 1: Architekturvergleich

Die Evaluation beider Systemarchitekturen weist auf die Vorteilhaftigkeit von Variante B hin: Zwar weist diese Alternative eine geringfügig höhere Komplexität aus, sie ist Variante A jedoch in den anderen Bewertungskriterien überlegen.

Kritische Würdigung

Während die Sentimentanalyse mithilfe des wörterbuchbasierten Ansatzes viele Sentimente richtig klassifizieren kann, stößt die Technologie teilweise auch an ihre Grenzen. Wie die meisten Technologien für Sentimentanalysen enthält SAP HANA keine Möglichkeit, ironische oder sarkastische Aussagen von Twitter-Nutzern zu erkennen, entsprechende Tweets werden daher in der Regel falsch zugeordnet. Darüber hinaus können Twitter-Bots, d.h. automatisierte Accounts, ebenfalls Einfluss auf die Ergebnisse der Sentimentanalyse nehmen.

Trotz vereinzelt inkorrekt klassifizierter Tweets lässt sich nach umfangreichen Stichproben die Annahme treffen, dass die Analyse meist einen korrekten Eindruck der Stimmung auf Twitter zu den entsprechenden Themen vermittelt.

Fazit

Mithilfe der vorgestellten Architektur lassen sich Daten von Twitter automatisiert extrahieren, auswerten und eine Sentimentanalyse durchführen. Des Weiteren können Umfragewerte von Qualtrics direkt in ein Data Warehouse importiert und gemeinsam mit weiteren Unternehmensdaten konsolidiert werden. Mithilfe von SAC ist es beispielweise möglich, die Ergebnisse zu visualisieren und Endnutzern in Form von dynamischen Dashboards zur Verfügung zu stellen. Die erarbeitete Lösung stellt einen ersten Prototypen für die technische Umsetzung der Zielsetzung dar und kann weiter optimiert und erweitert werden. Die Umsetzung der vorgestellten Lösung vermag es, Unternehmen in die Lage zu versetzen, Daten sozialer Netzwerke automatisiert zu extrahieren, zu verarbeiten und zu interpretieren, um die Meinungen der Nutzer dieser Netzwerke auszuwerten und darauf zu reagieren.