Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Fabian Sawatzki

Analyse und Visualisierung von Daten im Kontext der sozialen Medien

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Fabian Sawatzki

Analyse und Visualisierung von Daten im Kontext der sozialen Medien

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Olaf Zukunft
Zweitgutachter: Prof. Dr. Stefan Sarstedt

Eingereicht am: 15. September 2015

Fabian Sawatzki

Thema der Arbeit
Analyse und Visualisierung von Daten im Kontext der sozialen Medien

Stichworte
Twitter, Social-Media, Visualisierung, Analyse

Kurzzusammenfassung
Diese Bachelorarbeit befasst sich mit dem Themengebiet des Monitoring von sozialen Medien. Zum einen wird anhand des aktuellen Standes der Technik aufgezeigt, welche Möglichkeiten und Probleme es in diesem Zusammenhang gibt. Zum anderen besteht ein Aspekt dieser Arbeit darin, die möglichen Visualisierungen von analysierten Datenbeständen zu behandeln. Im Rahmen dessen wird ein eigenes Social-Media-Monitoring-Tool namens Sawatzki-Werkzeugkasten entwickelt und mit etablierten Produkten verglichen

Fabian Sawatzki

Title of the paper
Analysis and visualization of data regarding social media

Keywords
Twitter, social-media, visualization, analysis

Abstract
This bachelor thesis concerns the aspect of social-media-monitoring. On the one hand the possibilities and problems of different monitoring-techniques will be shown. On the other hand the visualization of social-media-data will be discussed. Furthermore an own social-media-monitoring-tool called Sawatzki-Werkzeugkasten has been developed and will be compared with established products in this particular sector.

Inhaltsverzeichnis

1 Einleitung

Nahezu jeder Informationsprozess bezieht heutzutage Online-Quellen ein. Sei es die Suche nach dem besten Autohändler in der unmittelbaren Umgebung oder die Abwägung, einen bestimmten Kinofilm zur Gestaltung des persönlichen Abendprogrammes zu besuchen. Eine besondere Rolle spielen dabei die sozialen Medien, deren Akzeptanz in den vergangenen Jahren enorm gestiegen ist. Dies lässt sich vor allem an den exponentiell wachsenden Nutzerzahlen ablesen, welche beispielhaft für das soziale Netzwerk *Facebook* in der Abbildung 1.1 dargestellt werden.

Allein Facebook bedient mit 1,42 Milliarden Nutzern[1] im März 2015 mehr als ein Drittel der gesamten Internetnutzer. Dieser Trend sorgt für eine hohe Informationstransparenz quer durch die verschiedenen Alters- und Interessensgruppen im Internet (vgl. Ceyp und Scupin (2013)).

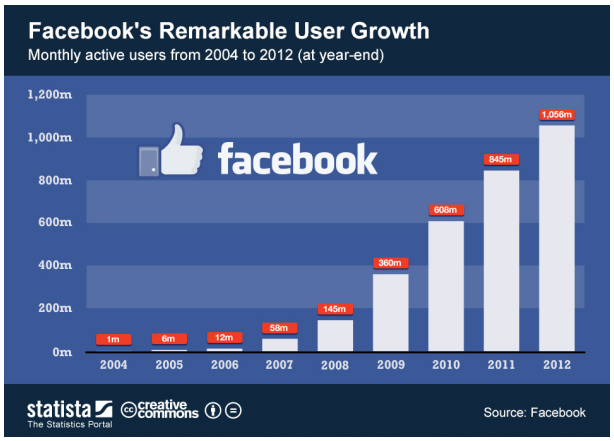

Abbildung 1.1: Wachstum der Facebook-Userzahlen seit Dezember 2004 - Quelle: statista.com (a)

[1] statista.com (b)

Die sozialen Medien erschaffen eine öffentliche Meinung wie sie zuvor nur durch die von Journalisten erstellten Massenmedien verbreitet wurde (vgl. König u. a.). Neben den persönlichen Informationen der Nutzer stellen die sozialen Medien vorrangig Meinungen, Bewertungen und Interessen der Nutzer bereit. Da die Beiträge in den sozialen Medien zudem als vertrauenswürdig eingestuft werden, handelt es sich darüber hinaus um potenziell repräsentative Daten (vgl. Cyganski und Hass (2011)). Damit wird es für jeden versierten Nutzer theoretisch möglich, eine ausgeprägte Markttransparenz zu schaffen.

Entsprechend interessant sind die sozialen Medien auch für Unternehmen geworden. Durch das sogenannte Monitoring der sozialen Medien (engl. *Social-Media-Monitoring*) ergibt sich die Möglichkeit, eine umfassende Reputation zu den angebotenen Produkten bzw. Dienstleistungen zu erhalten, eine wesentlich intensivere Kundenbeziehung zu entwickeln und Marketingziele effizienter zu stecken (vgl. König u. a.). Das Konzept ist vielversprechend; sowohl die Wirtschaft als auch die Endkunden könnten von dem Monitoring der sozialen Medien profitieren.

Definition 1 (Social-Media-Monitoring) *Der Begriff des Social-Media-Monitoring bezeichnet nach Ceyp und Scupin (2013) die Identifikation, Beobachtung und im weitesten Sinne auch die Analyse von benutzergenerierten Inhalten in sozialen Medien und Netzwerken.*

Jedoch stellen sich in der Praxis viele Fragen nach einer passenden Umsetzung. Welche Quellen sollen hinzugezogen werden? Wie werden die erfassten Daten optimal aufbereitet? Wann sind die Analyseergebnisse repräsentativ? (siehe Aßmann und Pleil)

Diese Bachelorarbeit befasst sich mit dem Themengebiet des Monitoring von sozialen Medien. Zum einen wird anhand des aktuellen Standes der Technik aufgezeigt, welche Möglichkeiten und Probleme es in diesem Zusammenhang gibt. Zum anderen besteht ein Aspekt dieser Arbeit darin, die möglichen Visualisierungen von analysierten Datenbeständen zu behandeln. Im Rahmen dessen wird ein eigenes Social-Media-Monitoring-Tool namens Sawatzki-Toolbox (im Folgenden als *STB* bezeichnet) entwickelt und mit etablierten Produkten verglichen.

1.1 Motivation

Im Rahmen des Projektes *Lernende Agenten* an der HAW Hamburg hat sich eine Gruppe von Studenten mit dem Themengebiet *Datamining* beschäftigt. Speziell wurde sich mit dem Aspekt

des Textmining befasst und versucht, eine Stimmungsanalyse in Bezug auf Beiträge im Social-News-Aggregator Reddit durchzuführen. Dabei haben die Teilnehmer neben kleinen Erfolgen vor allem zahlreiche Probleme im Zusammenhang von Textmining und sozialen Medien aufgedeckt. Dies ist Anlass, sich in dieser Bachelorarbeit tiefgreifender mit der Analyse von sozialen Medien zu beschäftigen und dabei die aus der Projektarbeit gewonnenen Erkenntnisse einfließen zu lassen.

1.2 Ziele

Ein Ziel dieser Bachelorarbeit besteht darin, das Potenzial eines eigens programmiertem Open-Source-Social-Media-Monitoring-Tool aufzuzeigen. Damit einhergehend soll auch das aktuelle Angebot an Tools kritisch betrachtet werden. Im Grunde geht es um die Fragestellung: Inwiefern wurden die sozialen Netzwerke bereits erfasst und erforscht, sodass repräsentative Aussagen auf Basis derer Inhalte getroffen werden können? Und weiterführend: Wurden die entsprechenden Anforderungen an die Social-Media-Monitoring-Tools in aktuell am Markt etablierten Systemen berücksichtigt oder gibt es dort noch einen großen Verbesserungsbedarf?

1.3 Gliederung dieser Bachelorarbeit

Neben der vorangegangen Einleitung dringt diese Bachelorarbeit zunächst noch detaillierter in die Thematik des Social-Media-Monitoring ein. Es wird differenziert, welche Chancen und Risiken sich durch den Einsatz entsprechender Software ergeben. Jedoch werden ebenfalls Probleme und technische Beschränkungen im Kapitel Social-Media-Monitoring-Tools herausgestellt. Besonders deutlich gemacht werden diese technischen Unzulänglichkeiten in einem Experiment zum Thema Textmining für Social-Media-Monitoring.

Anschließend steht die Visualisierung von analysierten Datensätzen aus dem Bereich der sozialen Medien im Fokus. Zum einen werden die verschiedenen Möglichkeiten samt Vor-und Nachteilen näher behandelt. Zum anderen wurde ein eigenes Social-Media-Monitoring-Tool namens Sawatzki-Toolbox sowie eine spezielle Graph-Visualisierung für die Zielgruppenanalyse in sozialen Medien namens Sawatzki-Ansatz entwickelt.

2 Social-Media-Monitoring-Tools

Dieses Kapitel beschäftigt sich damit, eine Einführung in die Thematik des Social-Media-Monitoring zu geben. Im Zuge dessen werden die Anforderungen für Social-Media-Monitoring-Tools skizziert, um anhand von aktuellen Systemen zu überprüfen, inwiefern diese bereits erfüllt worden sind.

2.1 Social-Media-Monitoring

> *Allein im Jahre 2014 wurden mehr Daten innerhalb von 10 Minuten generiert als vom Beginn der Menschheit bis ins Jahre 2002.*[1]

Dieses Zitat verdeutlicht die rasante technologische Entwicklung von Datenerzeugung, Datenverwaltung und Datenspeicherung in den vergangenen Jahren. Aufgrund des Trends *Ubiquitous Computing* sowie immer günstiger verfügbaren Datenspeichers war diese Evolution möglich (vgl. Kohlhammer u. a. (2013)). Plötzlich haben Datenanalysten Zugriff auf Datenmengen, die sie mit den klassischen Methoden der Datenverarbeitung nicht mehr effizient angehen können. In diesem Fall spricht man von Big Data.

Definition 2 (Big Data) *Big Data bezeichnet nach den Wikipedia-Autoren (01.05.2015) abstrakt gesagt eine neue Form von Daten, die sich nicht mehr mit den klassischen Datenverarbeitungsmethoden handhaben lässt. Charakterisiert wird Big Data anhand der Attribute Volume, Velocity sowie Variety. Es handelt sich folglich um Daten, die sehr groß sind (Dimensionen von Petabytes und aufwärts) und extrem schnell erzeugt werden. Daher muss in entsprechend kürzester Zeit darauf reagiert werden. Zum anderen weist Big Data eine große Varietät an Datentypen auf. Dazu gehören vor allem auch Multimedia-Daten wie Bilder und Videos.*

[1] Wikipedia-Autoren (01.05.2015)

Maßgeblich beteiligt an dieser ständig wachsenden Datenmasse- und Vielfalt sind die sozialen Netzwerke. Der in Abbildung 1.1 erkennbare Trend ist kein Einzelfall. Viele soziale Netzwerke hatten in den vergangenen Jahren ein exponentielles Userwachstum zu verzeichnen. Doch nicht nur die Betreiber sozialer Netzwerke profitieren von dieser Entwicklung. Auch für Unternehmen werden neben den internen betriebswirtschaftlichen Zahlen auch Texte aus den sozialen Medien oder geographische Daten immer interessanter (vgl. Kohlhammer u. a. (2013)). Diese lassen sich nutzen, um beispielsweise neue Zielgruppen-Merkmale für ein beworbenes Produkt aufzuspüren. Die Hoffnung besteht darin, ein facettenreiches und authentisches Bild vom potenziellen Kunden zu erstellen (vgl. Ceyp und Scupin (2013)).

Um dies zu ermöglichen, kann ein onlinebasiertes-Reputationsmanagement-System (kurz: ORM) herangezogen werden. Die Aufgaben eines ORM bestehen darin, den Ruf einer Person, einer Organisation oder eines Produkts in den digitalen Medien zu überwachen und zu beeinflussen. Arbeitet das ORM mit sozialen Medien, so betreibt es Social-Media-Monitoring in reinster Ausprägung. Das ORM ermöglicht es, relevante Informationen aus den gewünschten Online-Medien zu beobachten sowie zu beeinflussen. Damit sind in den meisten Fällen Äußerungen und Meinungen von Nutzern der sozialen Medien in Bezug auf ein bestimmtes Thema oder Produkt gemeint, für die sich das jeweilige Unternehmen interessiert. Das ORM ist in der Lage, die entsprechenden Informationen aufzubereiten und passende Handlungsmaßnahmen abzuleiten (vgl. Elgün und Karla (2013)).

Ein häufig in der Praxis genutztes Beispiel für ein ORM ist Radian6 von dem internationalen Cloudanbieter Salesforce. Dieses ist in der Lage, Online-Konversationen über Marken, Produkte und Themen zu beobachten, bewerten und analysieren. Es werden Online-Quellen jeder Art in insgesamt 17 verschiedenen Sprachen unterstützt. Der Nutzer von Radian6 ist auch in der Lage, in Echtzeit auf den jeweiligen Quellen zu interagieren, um beispielsweise direkt auf das Feedback der Nutzer zu reagieren.

Ein besonders interessantes Feature von Radian6, welches in diese Kerbe schlägt, ist das Reputationsfrühwarnsystem. Dieses durchsucht die entsprechenden Online-Quellen nach Diskussionen mit möglichem negativen Einfluss auf die Intentionen des nutzenden Unternehmens und warnt den Nutzer. Dabei wird sich die Tatsache zunutze gemacht, dass brisante Themen in den sozialen Medien in der Regel wesentlich zeitnaher diskutiert werden als in den klassischen Massenmedien.

Des Weiteren besteht auch die Möglichkeit, einen Meinungsführer zu identifizieren. Damit sind Nutzer sozialer Medien gemeint, die eine besonders hohe Reichweite besitzen und daher die Bewertung eines bestimmten Themas oder Produkts maßgeblich beeinflussen können. In der Fachliteratur wird deshalb empfohlen, diese in das Marketing einzubinden (vgl. Michelis und Schildhauer (2012))

Natürlich gibt es auch einige Probleme, die im Zusammenhang mit dem Social-Media-Monitoring bestehen. So ist im Beispiel von Radian6 die Erkennung von Sarkasmus in den verschiedenen Beiträgen der Online-Quellen kaum auszumachen. Und dies ist längst kein Einzelfall: Im Abschnitt Betrachtung verschiedener Systeme wird auf diese Problematik anhand von weiteren Beispielen eingegangen.

Generell steht ein Unternehmen bei der Entscheidung für das Social-Media-Monitoring vor der Wahl: Zum einen kann ein Software-Tool eingekauft werden, welches neben fachlicher Kompetenz auch Personaleinsatz erfordert. Dabei sollten die dadurch aggregierten Informationen in den Geschäftsprozessen des Unternehmens anerkannt sein und systematisch in die Aufgabenfelder integriert werden, damit sich die Investition rentiert (vgl. Bernhard Steimel | Christian Halemba | Tanya Dimitrova). Zum anderen besteht die Möglichkeit, das Social-Media-Monitoring auszulagern, was wesentlich kostspieliger ist als es im eigenen Unternehmen durchzuführen und sich dadurch negativ in den Finanzen widerspiegelt.

2.2 Anforderungen an ein Social-Media-Monitoring-Tool

Aufgrund der aufstrebenden sozialen Netzwerke sowie dem Potenzial von Social-Media-Monitoring befinden sich viele entsprechende Produkte auf dem Markt. Diese stellen häufig erst im Rahmen eines kostenpflichtigen Abonnements ihren vollen Funktionsumfang bereit. Die Anforderungen an ein solches Social-Media-Monitoring-System wurden von Seth Grimes in seinem Kommentar (siehe Seth Grimes) sinngemäß wie folgt zusammengefasst:

- Metadaten sind entscheidend: Nicht nur der betrachtete Beitrag sollte beachtet werden, sondern wie sich dieser Beitrag in die Social-Media-Landschaft einfügt. Interessante Informationen sind unter anderem, ob es sich bei dem Autor um einen Meinungsführer handelt, von welchem Standort der Beitrag gesendet wurde oder ob der Beitrag sich auf einen vorangegangenen Beitrag bezieht und eventuell Teil einer größeren Diskussion ist.

- Resolution bezeichnet die Fähigkeit, Daten aus dem betrachteten Beitrag sowie weiteren Quellen zu beziehen. Twitter stellt beispielsweise für jedes Profil weiterführende Informationen wie den echten Namen des Autors, seine Tätigkeit oder einen Link zu seiner Webseite bereit. Diese Informationen sollten miteinbezogen werden. In Bezug auf den Beitrag an sich sind anspruchsvolle Methoden des Natural Language Processing sowie der Stimmungserkennung erforderlich, um weitere Informationen aus dem Text zu extrahieren.

- Integration von Daten, auch über die Grenzen eines sozialen Netzwerkes hinweg: Möglicherweise entstammt ein auf Twitter veröffentlichter Beitrag ursprünglich aus einem Forum mit einem voneinander abweichenden Autoren. Beiträge sollten nicht isoliert betrachtet werden.

- Ein Social-Media-Monitoring-Tool sollte Messungen und Vorhersagen treffen und diese miteinander abgleichen.

- Ein Interface sollte neben einem Dashboard und Reporting-Funktionalität auch umfangreiche Filterungsfunktionen für die Auswahl der Daten und Generierung der Visualisierungen bereitstellen.

Diese Punkte decken sich größtenteils mit denen im wissenschaftlichen Artikel Elgün und Karla (2013) genannten Punkten. Seth Grimes erwähnt weiterhin, dass er von einem Social-Media-Monitoring-System nicht erwartet, alle Punkte in Gänze zu erfüllen. Jedoch sollte jedes System jene Punkte umsetzen, die entscheidend für den zu bietenden Mehrwert sind.

2.3 Betrachtung verschiedener Systeme

Ein wiederkehrendes Muster bei der Begutachtung verschiedener Social-Media-Monitoring-Systeme ist ein simples Formular, das für die Eingabe des zu beobachtenden Schlagwortes genutzt wird: Getreu dem Motto *keep it simple*. Dies erinnert stark an die Aufmachung des Suchmaschinen-Marktführers Google. Damit wird dem Nutzer folglich eine ähnlich intelligente Stichwortsuche suggeriert. In Bezug auf die sozialen Medien gestaltet sich dies jedoch schwierig wie folgendes Beispiel erläutert:

Beispiel 1 (Firma Wienek testet Social-Media-Monitoring-Tool) *Die Firma Wienek ist seit mehreren Jahren auf verschiedenen Social-Media-Plattformen aktiv. Besonders viel Interaktion mit den Kunden wird auf Twitter erreicht. Über die Jahre hat sich dort das Schlagwort #WienekQandA für die direkte Kommunikation zwischen Kunde und Unternehmen entwickelt. Da die Betreuung der Social-Media-Kanäle der Firma Wienek mittlerweile viele Arbeitsstunden erfordert und niemand den gesamten Überblick über die Daten besitzt, interessiert sich Wienek für die Nutzung eines Social-Media-Monitoring-Tools.*

Während der Recherche für ein passendes System stößt ein Mitarbeiter auf das Tool Topsy[2], welches verspricht, Daten aus der Vergangenheit bezüglich eines bestimmten Schlagwortes zu sammeln und zu analysieren. Das Formular zum Erstellen der Resultate ist dabei denkbar einfach: Es handelt sich um ein einfaches Texteingabefeld, welches ein oder mehrere Schlagworte erwartet. Nun war der Mitarbeiter der Firma allerdings verwundert als er das Tool anhand des Schlagwortes ***Wienek*** *getestet hat, denn es wurden längst nicht alle Beiträge bezüglich seiner Firma gefunden. Viele Tweets wurden beispielsweise erst berücksichtigt, wenn das Schlagwort #WienekQandA genutzt wurde.*

Das im Beispiel genannte Problem trifft nicht nur auf das Tool Topsy, sondern eine Reihe weiterer Social-Media-Monitoring-Tools zu. Im Rückgriff auf die im vorigen Kapitel genannten Anforderungen an ein Social-Media-Monitoring-Tool bedeutet dies, dass das Tool Topsy vor allem bei der Integration der Daten ein Defizit besitzt.

Dementsprechend schwierig ist es für den Anwender, seinen gesamten Auftritt in den sozialen Medien im Blick zu behalten. Möchte er die Ergebnisse verschiedener Schlagworte zusammenführen, müsste er diese im Fall von Topsy manuell berechnen. Ein großer Mehrwert von Social-Media-Monitoring-Systemen sollte jedoch dadurch gegeben sein, alle gewünschten Informationen auf einen Blick zu erhalten.

Ein ähnliches Problem ergibt sich bei den angebotenen Visualisierungen der Systeme. In der aktuellen Variante von Quintly[3] hat der Nutzer die Möglichkeit, sich auf Basis verschiedener Social-Media-Accounts Visualisierungen zu erstellen, die bestimmte Metriken wie die Anzahl der Fans, die Interaktionsrate oder die Page Impressions visualisieren. Allerdings ist die Auswahl der Metriken im Fall von Quintly sehr unübersichtlich. Viele Metriken werden nach Zeiteinheit unterschieden und in dem Auswahlmenü gibt es entsprechend viele Einträge pro Metrik -

[2] http://topsy.com/
[3] https://www.quintly.com/

beispielsweise *Own Posts By Hour*, *Own Posts by Weekday* usw. Im Sinne von Seth Grimes und Elgün und Karla (2013) wäre es sinnvoller, eine einzige Metrik namens *Own Posts* anzubieten und innerhalb der Visualisierung entsprechende Filterfunktionen anwählbar zu machen.

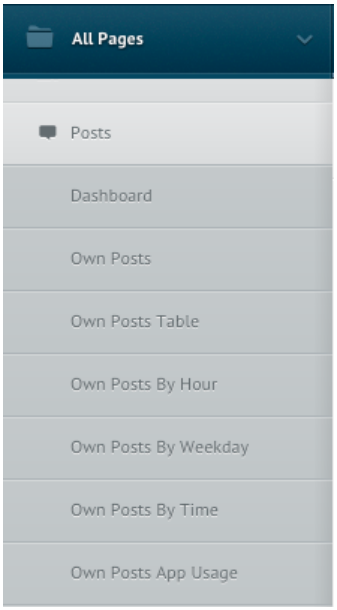

Abbildung 2.1: Die Aufspaltung der Metriken in verschiedene Zeitbereiche könnte in Bezug auf die Übersichtlichkeit besser in ein Dropdown-Menü überführt werden

2.4 Fazit

In der beispielhaften Betrachtung verschiedener Social-Media-Monitoring-Tools zeigen sich bereits kleine Schwächen. In dem folgenden Kapitel des Textmining für Social-Media-Monitoring wird ein Aspekt des Monitoring genauer untersucht. Im Abschnitt Benchmark im Kapitel der Sawatzki-Toolbox werden viele Tools im Vergleich unter Berücksichtigung der bereits definierten Anforderungen an ein Social-Media-Monitoring-Tool gemessen.

Zusammenfassend lässt sich sagen, dass die sozialen Medien und das Monitoring der Kommunikation für das Unternehmen zukünftig immer mehr an Bedeutung gewinnen wird. Das Monitoring der sozialen Medien bietet viele Chancen für die Unternehmenskommunikation, birgt jedoch auch gewisse Risiken. Dabei steht derzeit noch nicht das *Return on Investment*, sondern vielmehr das *Risk of Ignorance* im Fokus der Aktivitäten (vgl. Elgün und Karla (2013)).

3 Textmining für Social-Media-Monitoring

Das folgende Kapitel beschäftigt sich mit dem Textmining, einem besonders im Zusammenhang mit den sozialen Medien interessanten Aspekt der Datenanalyse. Ziel von Textmining ist es, mithilfe von linguistischen und statistischen Methoden neues und potenziell nützliches Wissen aus Textdokumenten zu extrahieren (vgl. Hippner und Rentzmann (2006)). Die Wikipedia-Autoren (siehe Wikipedia-Autoren (27.07.2015)) formulieren es wie folgt: *Textmining ist ein Bündel von Algorithmus-basierten Analyseverfahren zur Entdeckung von Bedeutungsstrukturen aus un- oder schwachstrukturierten Textdaten.* Im Rahmen von vielen Social-Media-Monitoring-Tools wird eine sogenannte Stimmungserfassung angeboten, einer Teildisziplin des Textmining-Bereiches. Dabei geht es darum, Texte zu kategorisieren, wobei es sich meist um eine Postiv-Negativ-Kategorisierung handelt. Vereinzelt wird noch die neutrale Kategorie als Graustufe genutzt. Da dies abhängig von den Inputdaten einige Problematiken mit sich bringen kann, wird im Folgenden eine Umfrage durchgeführt, die sich dessen annimmt.

3.1 Aufbau

Die Aufgabe der Teilnehmer bestand darin, anhand verschiedener Texte aus sozialen Netzwerken eine geäußerte Haltung als positiv oder negativ zu erkennen. Die Texte wurden dabei aus dem sozialen Netzwerk Twitter entnommen. Es wurde darauf geachtet, nicht nur Texte mit einer eindeutigen Haltung des Verfassers zu nutzen, sondern auch Texte, die Ironie, einen Link oder ein Zitat beinhalten. Daher war es in einigen Fällen nicht ohne weiteres Wissen möglich, den Text entsprechend der Intention des Autors zu klassifizieren.

Um ein repräsentatives Ergebnis zu erhalten, wurden hauptsächlich junge Menschen mit einem Interesse für neuartige Medien befragt. Um dies sicherzustellen wurde der Link lediglich auf dem Social-News-Aggregator Reddit veröffentlicht, der entsprechendes Publikum anzieht. Ge-

nauer gesagt wurden sogar ausschließlich spezielle *Subreddits* wie https://www.reddit.com/r/rocketbeans/ angesprochen, um eine möglichst homogene Zielgruppe zu gewährleisten. Insgesamt wurden den Teilnehmern 15 Texte zur Verfügung gestellt, welche jeweils in die Kategorie positive Haltung oder negative Haltung sortiert werden mussten.

3.2 Beobachtung

Das Gesamtergebnis der Umfrage nach sechs Tagen Laufzeit und 56 Teilnahmen (Stand 10.08.2015) befindet sich graphisch aufbereitet im Anhang unter dem Punkt: Ergebnisse der Textmining-Umfrage

Eindeutige Antworten mit mehr als 85 Prozent Übereinstimmung der Teilnehmer hat es in Bezug auf folgende Tweets gegeben:

- *seit heute nachmittag kein internet, telefon und tv mehr. danke #kabeldeutschland - nicht! - von @der_Ben83* wurde mit 100 % Übereinstimmung als negativ bewertet
- *Das Klacken der Kaffeemaschine wenn sie fertig ist ist das beste Geräusch der Welt. - von @extraktiv* wurde mit 94,64 % Übereinstimmung als positiv bewertet
- *Guten Morgen ihr Lieben... Ist noch #Kaffee da? - von @iZerf* wurde mit 92,86 % Übereinstimmung als positiv bewertet
- *Mitglieder Atlantik-Brücke: Sollte immer wieder mal erwähnt werden, damit man sich nicht wundert #TTIP [...] - von @tauss* wurde mit 87,5 % Übereinstimmung als negativ bewertet
- *Bleibt doch mal sitzen, bis die Ansage für den Bahnhof kommt, Herrgott! - von @HerrLevin_* wurde mit 94,64 % Übereinstimmung als negativ bewertet
- *Grade Urlaub für Fallout 4 im November beantragt. Lustiger Smiley #Fallout4 - von @GuyLikesGames* wurde mit 96,43 % Übereinstimmung als positiv bewertet
- *Nebeneinkünfte: Das sind die Topverdiener im Bundestag... [...] - von (@SPIEGEL_Politik)* wurde mit 85,71 % Übereinstimmung als negativ bewertet

- *Ich hab jetzt keinen Bock mehr zu arbeiten. Es geht raus in die #Sonne, an die #elbe. Wer ist dabei? - von @stevengaetjen* wurde mit 87,5 % Übereinstimmung als positiv bewertet

Ein Trend mit mehr als 65 Prozent Übereinstimmung ließ sich bei folgenden Tweets erkennen:

- *Deutschland, Deutschland, du tüchtiges Land! #berlin #bundestag - von @julmaxpaul* wurde mit 66,07 % Übereinstimmung als negativ bewertet
- *#Essen #Zeuge nach Verkehrsunfall gesucht 19.01.2015 [...] - von @MiloFornazzo* wurde mit 76,79 % Übereinstimmung als negativ bewertet
- *Aus der aktiven #politik hat sich #sarah #palin zurückgezogen - von @chrispillennews* wurde mit 76,79 % Übereinstimmung als positiv bewertet

Keinerlei Trend ließ sich bei folgenden Tweets mit weniger als 65 Prozent Übereinstimmung herauslesen:

- *Und im Himmel legt Bob #Marley den Joint kurz bei Seite und ballt die Faust. #Wimbledon2015 @DreddyTennis - von @HeikoOldoerp*
- *#Hoax = #Wasser trinken hilft gegen Kopfschmerzen - von @MartinKaindel*
- *Freitag Abend. Ich schaue den Krimi auf @ZDF, trinke Tee und stricke. So fühlt sich also dieses Erwachsenwerden an. #dontgrowup #itsatrap - von @lisarossel*
- *Das wird eine anstrengende Woche #gamescom #videodays - von @_pleasestandby*

Daher wurden mit acht der fünfzehn Tweets lediglich 53,3 Prozent der Tweets eindeutig von den Teilnehmern kategorisiert. Bei 20 Prozent der Tweets war immerhin ein Trend zu erkennen. Bei den restlichen 26,7 Prozent der Tweets herrschte große Uneinigkeit zwischen den Teilnehmern, sodass die Kategorisierung beinahe einem Münzwurf gleichkommt.

3.3 Deutung

Nehmen wir uns einmal folgenden Tweet zum Beispiel: *Und im Himmel legt Bob #Marley den Joint kurz bei Seite und ballt die Faust. #Wimbledon2015 @DreddyTennis.* Um diesen Beitrag korrekt beurteilen zu können, benötigt der Beurteilende unterschiedliches Vorwissen. Wer ist Bob Marley, was ist Wimbledon und wer verbirgt sich hinter dem Pseudonym *@DreddyTennis*? Selbst wenn der Beurteiler darüber informiert ist, dass es sich bei Wimbledon um das älteste und prestigeträchtigste Tennisturnier der Welt, bei @DreddyTennis um Dustin Brown, einen deutschen Tennisspieler jamaikanischer Herkunft und bei Bob Marley um einen sehr bekannten, jamaikanischen Sänger, Gitarristen und Songwriter handelte, reichen die Informationen noch nicht aus.

Der in den 80er-Jahren verstorbene Bob Marley könnte schließlich auch seine Faust aus Wut ballen, weil Dustin Brown einen entscheidenden Fehler gemacht hat. Eine automatisierte Stimmungserfassung, die lediglich auf einer Bewertung der einzelnen Worte basiert und das Wort *ballen* als negativ erachtet, würde den Beitrag vorschnell als negativ erachten. Beachtet man jedoch, dass der Tweet am Tage des überraschenden Sieges von Dustin Brown über Raffael Nadal erschienen ist, wird die Aussage klar: Der Autor des Tweets verbreitet seine Freude über den Sieg von Dustin Brown in dieser kreativen Form. Jedoch herrschte bei den Teilnehmern deutliche Uneinigkeit. Das zeigt, wie umfangreich es sein kann, die Aussage eines Tweets zu entschlüsseln.

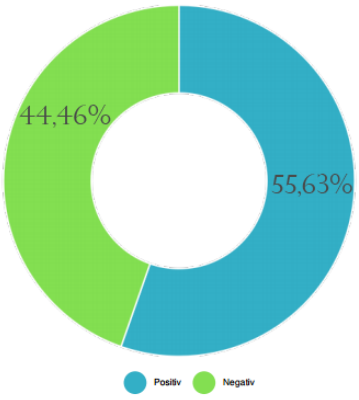

Abbildung 3.1: Die Verteilung der Stimmen bezüglich des Tweets: *Und im Himmel legt Bob #Marley den Joint kurz bei Seite und ballt die Faust. #Wimbledon2015 @Dreddy-Tennis*

Ein weiteres sehr interessantes Beispiel ist folgender Tweet: *Grade Urlaub für Fallout 4 im November beantragt. Lustiger Smiley #Fallout4 - von @GuyLikesGames.* Die Bedeutung des Textes steht und fällt in diesem Fall mit dem Begriff *Fallout 4.* Direkt übersetzt steht Fallout für Atomstaub, was ein überaus negativ behaftetes Wort ist. Man könnte nun daraus ableiten, dass der Urlaub des Autors deswegen beantragt wurde, weil er eine Apokalypse erwartet und sein Statement *Lustiger Smiley* sarkastisch gemeint ist.

Jedoch handelt es sich bei Fallout 4 um ein Computerspiel, welches von vielen Fans sehnsüchtig erwartet wird. Bei dem Autor des Textes wiederum handelt es sich um einen Videospielredakteur, der außerordentlich auf dieses Spiel gespannt ist und sich deshalb sogar Urlaub genommen hat. Die Teilnehmer dieser Umfrage wurden größtenteils aus dem Subreddit https://www.reddit.com/r/rocketbeans/ gebildet, welcher sich hauptsächlich mit der Materie der Computerspiele auseinandersetzt und dessen Nutzer mit dem Autor des Textes vertraut sind. Daher war das Ergebnis mit vielen korrekt positiven Antworten durchaus zu erwarten. Wäre die Umfrage nicht in dieser Zielgruppe durchgeführt worden, hätte dies womöglich ein deutlich weniger eindeutiges Ergebnis zur Folge.

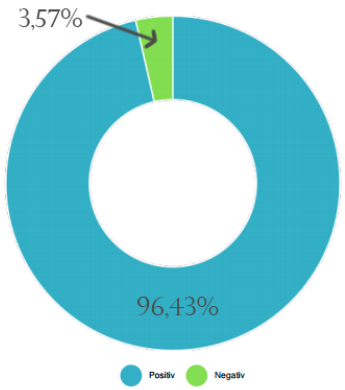

Abbildung 3.2: Die Verteilung der Stimmen bezüglich des Tweets: *Grade Urlaub für Fallout 4 im November beantragt. Lustiger Smiley #Fallout4 - von @GuyLikesGames*

Auch folgendes Ergebnis ist entscheidend für die Deutung: *Nebeneinkünfte: Das sind die Topverdiener im Bundestag... [...] - von (@SPIEGEL_Politik).* Dieser Beitrag ist besonders deshalb interessant, weil der Text an sich keine Wertung beinhaltet. Es ist lediglich ein Hinweis auf die Nebeneinkünfte einiger Politiker im Bundestag mit einem entsprechend weiterführenden Link, den die Teilnehmer bewusst nicht zur Verfügung gestellt bekommen haben.

Dennoch wurde der Beitrag mehrheitlich als negativ erachtet. Im Gespräch mit einigen Teilnehmern hat sich herausgestellt, dass die Einkünfte von Politikern in der befragten Zielgruppe ohnehin als kritisch betrachtet werden. Besonders das Wort Nebeneinkünfte würde sinnbildlich für Politikverdrossenheit stehen. Daher wurde dieser Beitrag mehrheitlich negativ bewertet.

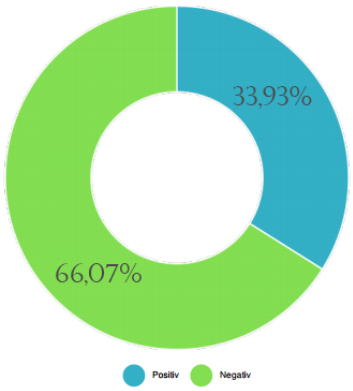

Abbildung 3.3: Die Verteilung der Stimmen bezüglich des Tweets: *Nebeneinkünfte: Das sind die Topverdiener im Bundestag... [...] - von (@SPIEGEL_Politik)*

Die Umfrage hat deutlich gezeigt, dass die Bewertung von Beiträgen aus den sozialen Netzwerken zum Teil sehr diffizil ist. Einige Beiträge offenbaren ihre wahre Haltung erst mit notwendigem Vorwissen und bestimmte Formulierungen haben je nach Sichtweise des Lesenden eine deutlich unterschiedliche Wahrnehmung. Der Mensch tätigt Entscheidungen aufgrund seiner eigenen Erfahrung, Moral und Einstellung zu bestimmten Themen (vgl. Wikipedia contributors (19.08.2015)).

Eine künstliche Intelligenz kann sicherlich Entscheidungen treffen, die auf bestimmten Kriterien beruhen. Auch ist es möglich, dass die künstliche Intelligenz versucht, die Semantik aus bestimmten Begriffen, angegebenen Links oder Bildern zu erfassen. Jedoch führt dies auch zu beliebig komplexen Algorithmen und bisher nicht optimal gelösten Problemen wie Sarkasmus (in Bildern), fehlerhaften Beiträgen oder Übersetzungen (siehe Reyes und Rosso (2014)). Niemals wird es jedoch möglich sein, einen Beitrag dahingehend zu bewerten, dass jeder Mensch damit übereinstimmen würde.

Das zeigte die Umfrage, in der im Bezug auf viele Beiträge Uneinigkeit herrschte. Dabei waren die Teilnehmer aus der gleichen Altersgruppe mit überschneidenden Interessen. Bezogen auf alle Menschen wäre die Kluft vermutlich deutlich größer als es diese Umfrage ohnehin schon andeutet.

4 Visualisierung

Im Themenkomplex des Social-Media-Monitoring dreht sich vieles um die geschickte Aufbereitung der gewonnen Informationen. Die Entscheidung ist dabei keinesfalls trivial. Schon bei der Akquisition der Daten können entscheidende Fehler passieren. Im folgenden Kapitel wird daher der Prozess von der Datenakquisition bis zur Entstehung einer aussagekräftigen Visualisierung betrachtet. Dabei steht vor allem die Entscheidung einer geeigneten Visualisierung auf Basis der vorhandenen Daten im Fokus.

4.1 Grundlagen

Heutige Unternehmen sammeln, speichern und verwerten Daten in einem nie da gewesenen Ausmaß. Dem möglichen Informationsgewinn aus der Masse an Daten stehen jedoch ungelöste Probleme entgegen, die verhindern, die Daten effektiv zu nutzen. Im Zuge dessen hat sich der Begriff des Data Overload herausgebildet.

Definition 3 (Data Overload) *Data Overload bezeichnet nach Kohlhammer u. a. (2013) die Gefahr, sich in Daten zu verlieren, die aktuell nicht relevant sind, in ungeeigneter Weise aufbereitet wurden oder ineffektiv dargestellt werden.*

Entsprechend werden laut Experten durchschnittlich im Unternehmen lediglich acht Prozent des Informationsangebotes wahrgenommen (vgl.Bassler (2010)). Der Gedanke hat sich verlagert. Die Frage lautet nicht mehr: Wie bekomme ich die Daten und Informationen? Sondern: Welche Informationen gebe ich weiter und wie bereite ich diese auf? Da die Entscheider im Unternehmen jedoch oftmals gezwungen werden, sich die gewünschten Informationen selbst zu erarbeiten, werden die bereitgestellten Informationen häufig ignoriert.

Grund dafür ist neben dem stressigen Alltagsgeschäft vor allem die mangelhafte Aufbereitung der Informationen (vgl. Kohlhammer u. a. (2013)). Eine Möglichkeit dies zu ändern besteht darin, die Informationen entsprechend zu visualisieren. In einer diesbezüglichen Studie des *The Data Warehousing Institute* wurden Visualisierungen von den Teilnehmern ebenfalls als sehr wichtig empfunden, dennoch werden laut der Studie 65 % der Unternehmensinformationen als Tabellen übermittelt (vgl. Wayne Eckerson and Mark Hammond). Die Möglichkeiten von adäquaten Visualisierungen werden schlichtweg nicht genutzt.

Um die Bedeutung von Daten zu ergründen bedarf es einem Kontext. Erst in diesem Kontext wird aus einem Datum eine Information mit einem möglichem Mehrwert für ein Unternehmen (vgl. Kohlhammer u. a. (2013)). So besitzt beispielsweise die Zahl 103154760 keinerlei Aussagekraft. Erst in dem Kontext Konto-, Rechnungs- oder Sendungsverfolgungsnummer wird sie zu einer Information, die sich effizient visuell aufbereiten lässt. Man unterscheidet folglich zwischen verschiedenen Datentypen, welche sich für unterschiedliche Visualisierungstechniken anbieten.

Die wesentlichen Darstellungsformen haben ihren Ursprung bereits Ende des 19. Jahrhunderts im sogenannten *golden age of statistical graphics*, während dynamische und interaktive Diagramme seit den 1950er Jahren bekannt sind (vgl. Friendly). Es haben sich allerdings auch Visualisierungsmethoden entwickelt, die sich speziell mit den sozialen Netzwerken oder im weiteren Sinne auf Big Data spezialisieren (vgl. Kohlhammer u. a. (2013)).

4.1.1 Datenklassen

Da sich ein Balkendiagramm genauso wenig für die Darstellung des unternehmensinternen Netzwerkes eignet wie eine Netzwerkdarstellung zur Visualisierung der jährlichen Verkaufszahlen, empfiehlt es sich, den Rohdaten verschiedene Klassen zuzuordnen. Generell wird zwischen quantitativen, ordinalen und nominalen Daten unterschieden. Im Folgenden, die entsprechenden Definitionen laut Kohlhammer u. a. (2013):

- Quantitative Daten enthalten numerische Werte, mit denen sich Berechnungen ausführen lassen. Dazu zählen alle diskreten und kontinuierlichen Werte. Beispielsweise zählt das Alter der Mitarbeiter oder die Umsatzdaten eines Unternehmens zu diesen Werten.

- Ordinale Daten haben einen feste, vorgegebene Ordnung. Es spielt dabei keine Rolle, ob es sich um numerische oder nicht numerische Werte handelt. Ein gutes Beispiel zur Verdeutlichung dieses Zusammenhangs sind Monatsnamen. Diese haben eine feste Ordnung von Januar bis Dezember und lassen sich sowohl anhand des Namens referenzieren, jedoch auch durch Zahlen repräsentieren.

- Nominale Daten wiederum enthalten beliebige nicht numerische Werte, die sich in keiner Ordnung befinden. Nachnamen passen in dieses Schema. Üblicherweise arbeitet man zwar mit alphabetisch sortierten Listen von Nachnamen; allerdings handelt es sich dabei nicht um eine feste, vorgegebene Ordnung.

Im Vorgriff auf das Kapitel Sawatzki-Toolbox bedeutet dies, dass die angebotenen Visualisierungen von der Art des jeweiligen Datums abhängig sind.

4.1.2 Akquisition und Qualität

Bei der Datenakquisition wird grob zwischen internen und externen Datenquellen unterschieden. Die internen Datenquellen sind fest in der Hand des Datenanalysten. Dementsprechend lassen sich Parameter wie die Häufigkeit der Datenerhebung oder die Aggregation der Daten nach Belieben einstellen. Dadurch kann die Datenqualität maßgeblich beeinflusst werden.

Das entsprechende Gegenstück bilden die frei verfügbaren, bzw. vom Unternehmen erworbenen externen Daten. Dazu zählen sowohl Texte aus sozialen Netzwerken als auch der Finanzkurs einer Aktie (vgl. Kohlhammer u. a. (2013)). Diese befinden sich außerhalb der eigenen Kontrolle und können daher beliebige Qualitätsmängel aufweisen. Die Daten sollten daher keinesfalls ohne Weiteres als repräsentativ betrachtet werden. Es handelt sich möglicherweise um vielversprechend aussehende Analysedaten, welche jedoch keinerlei Aussagekraft besitzen (vgl. Gaffney und Puschmann (2014)).

Besonders bei der Verarbeitung von Big Data ergeben sich häufig Trugschlüsse in der Analyse. Nehmen wir folgendes Beispiel:

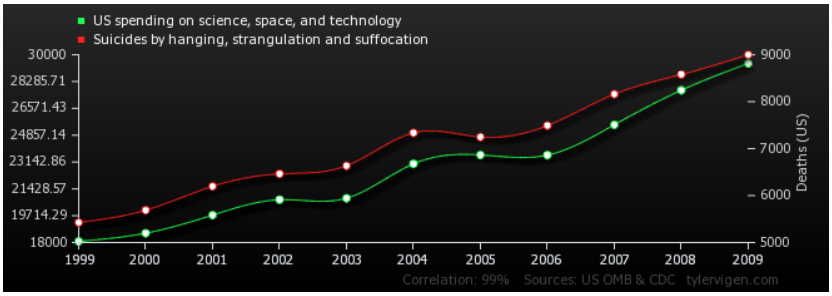

Abbildung 4.1: Investitionen der USA in den Bereichen Wissenschaft, Raumfahrt und Technologie in Relation zu Selbstmorden durch Erhängen, Strangulation und Ersticken - Quelle: tylervigen.com

In diesem Graphen korrelieren die Ausgangsdaten stark miteinander. Jedoch stehen die Investitionen der USA in den Bereichen Wissenschaft, Raumfahrt und Technologie offensichtlich nicht im Zusammenhang mit Selbstmorden durch Erhängen, Strangulation und Ersticken, sodass diese Visualisierung sogar einen humoristischen Charakter erhält. Ein automatisierter Algorithmus ist im Gegensatz zum Menschen jedoch nicht in der Lage, diesen Zusammenhang zu erkennen.

In der englischen Sprache hat sich in diesem Zusammenhang ein passendes Sprichwort entwickelt: *garbage in, garbage out*. Entsprechend lassen sich fehlende Daten, Ausreißer oder falsche Daten auch nicht im Nachhinein durch eine dahingehend korrigierte Visualisierung ausgleichen. Sollte dies dennoch versucht werden, kann die entstehende Visualisierung eine Qualität suggerieren, die im Vorfeld nie vorhanden war (vgl. Kohlhammer u. a. (2013)).

In der Literatur wird von zwei maßgeblichen Faktoren in Bezug auf eine erfolgreiche Visualisierung gesprochen. Diese lauten Expressivität und Effektivität. Während die Expressivität verlangt, dass alle - und ausschließlich die zugrunde liegenden - Daten verwendet werden, erfordert die Effektivität, dass die Visualisierung schneller verstanden wird oder mehr Einzelheiten darstellt als eine andere Visualisierung. Expressive Daten lassen sich grundsätzlich leichter ohne umfangreiche Vorverarbeitung der Daten erreichen. Wenngleich eine Datenreduktion die Gefahr birgt, interessante Aspekte aus den Eingangsdaten zu entfernen, kann sie dennoch sinnvoll sein (vgl. Kohlhammer u. a. (2013)). Im Zuge einer Stimmungsanalyse ist es beispielsweise üblich, sogenannte Stoppwörter aus den Texten zu entfernen. Das sind

Wörter wie zum Beispiel bestimmte und unbestimmte Artikel, die bei der Klassifizierung nicht benötigt werden (vgl. Dalal). Die Begriffe Expressivität und Effektivität sind entscheidend für die folgenden Kapitel. Sie werden im Folgenden häufig genutzt, um Visualisierungen zu charakterisieren und zu bewerten.

4.2 Datentypen

Nachdem im vorherigen Abschnitt die Grundlagen für eine expressive und effektive Visualisierung erklärt wurden, findet sich im Folgenden ein genauerer Überblick darüber, welche Visualisierungen sich für verschiedene Ausgangsdaten anbieten. Beschrieben werden alle Datentypen, die in der Fachliteratur als essenziell im Zusammenhang mit der systematischen Analyse von Daten (aus sozialen Medien) betrachtet werden.

4.2.1 Texte

Texte begegnen uns besonders im Zusammenhang mit den sozialen Medien gehäuft. Diese sind je nach Medium sehr unterschiedlich ausgerichtet. Bei einem Blick auf den *Social-News-Aggregator* Reddit finden sich zu beliebigen Fragestellungen umfangreiche, sauber strukturierte Kommentare mit einem Umfang von bis zu 10.000 Zeichen. Im Gegensatz dazu steht das *soziale Netzwerk* Twitter, welches die Länge eines Beitrags auf 140 Zeichen beschränkt und aufgrund der Schnelllebigkeit viele Trendworte und Rechtschreibfehler beinhaltet.

Sowohl Reddit als auch Twitter speichern für ihre Beiträge eine Fülle an Metadaten, welche sich sehr gut dafür eignen, Visualisierungen aufzubauen. So lässt sich beispielsweise auf Twitter anhand der Kommentare zu einem bestimmten Beitrag ein Netzwerk von Usern aufbauen, die sich allesamt für das Thema des jeweiligen Beitrags interessieren.

Im Fall von gänzlich unstrukturiertem Text spielt die Vorverarbeitung eine wichtige Rolle. Es existiert eine Fülle an Algorithmen, die Informationen aus einem Text aggregieren (vgl. Kohlhammer u. a. (2013)). Dazu zählt auch die Stimmungsanalyse, die einen Text anhand der geäußerten Haltung als positiv oder negativ klassifiziert.

Um einen ersten visuellen Eindruck eines Textes zu erhalten, bietet sich eine sogenannte Wordcloud an. Prinzipiell wird die Schriftgröße eines Wortes in der Wordcloud durch dessen Häufigkeit bestimmt, wobei häufig vorkommende Wörter größer dargestellt werden. Es sind jedoch auch andere Gewichtungen denkbar. Es wäre beispielsweise möglich, die Wörter eines Textes anhand dessen Stimmung zu bewerten. Besonders kritische Worte wie *abgeneigt* oder *enttäuscht* könnten entsprechend groß geschrieben werden.

Abbildung 4.2: In Deutschland beliebte Twitter-Hashtags um Weihnachten 2014 in einer nach dem Kriterium der Worthäufigkeit erstellten Wordcloud

4.2.2 Listen und Tabellen

Wie wir bereits im Abschnitt Grundlagen festgestellt haben, werden laut einer TDWI-Studie 65 % der unternehmensinternen Informationen als Tabellen festgehalten und weitergereicht. Typisch sind zweidimensionale Tabellen, um zum Beispiel den Umsatz pro Kunde darzustellen. Wenn zusätzlich der Faktor Zeit berücksichtigt wird, weil diese Umsatzdaten monatlich erfasst werden, kommt eine dritte Dimension hinzu. Diese Zusammenhänge lassen sich schwerlich in einer Tabelle erfassen, jedoch gibt es für beide Beispiele geeignete Visualisierungen.

Eine häufig zur Verwaltung von Listen und Tabellen eingesetzte Applikation ist Microsoft Excel. Diese bietet bereits viele Möglichkeiten zur Visualisierung von Listen und Tabellen. Wenn man herausfinden möchte, welcher Kunde in welchem Monat den meisten Umsatz erwirtschaftet hat, besteht die Wahl zwischen einem Säulen-, Balken-, Linien- oder einem Kreisdiagramm. Im nachfolgenden Abschnitt Diagramme werden die Vor-und Nachteile der jeweiligen Diagramme im Detail und anhand von Beispielen erläutert.

4.2.3 Hierarchien und Bäume

Bei Bäumen handelt es sich um spezielle Graphen; Hierarchien hingegen sind spezielle Bäume. In vielen Systemen existieren explizite Hierarchien. Im Bereich des Maschinenbaus ist beispielsweise jedes Produkt hierarchisch bis auf die einzelne Schraube heruntergebrochen. Um einen schnellen Gesamtüberblick über das zu fertigende Produkt zu erhalten, bietet es sich an, diese Hierarchien auf eine Visualisierung abzubilden. Im Beispiel des Maschinenbaus würde die Wurzel des Baumes das zu fertigende Produkt darstellen. In den darunter folgenden Ebenen würde sich das Produkt in verschiedene Teilprodukte aufspalten. Dieser Prozess führt sich solange fort, bis wir auf der untersten Ebene bei einem Bauteil angekommen sind, welches sich nicht weiter unterteilen lässt: Einem Blatt des Baumes (siehe Kohlhammer u. a. (2013)). In dem Beispiel könnte dieses Blatt durch eine Schraube repräsentiert werden. In der folgenden Grafik ist eine entsprechende Visualisierung zu sehen:

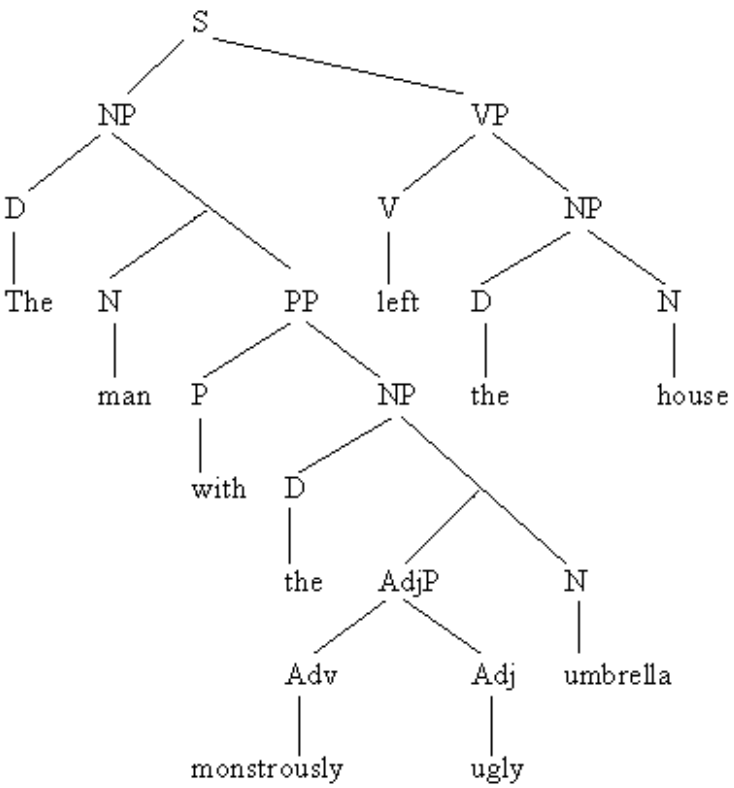

Abbildung 4.3: Eine Baumdarstellung, welche die Bestandteile eines Satzes aufschlüsselt - Quelle: Arizona State University (23.06.2000)

4.2.4 Netzwerke

Eine Visualisierungsform, welche besonders naheliegend im Zusammenhang mit den sozialen Medien erscheint, ist das Netzwerk. So ließen sich zum Beispiel die bidirektionalen Freundschaftsbeziehungen, wie sie aus sozialen Netzwerken wie *Facebook*, *LinkedIn* oder *Xing* bekannt sind, in ein graphisches Netzwerk übertragen. Jedoch scheitert diese Darstellung bereits im kleinen Rahmen an einem Überangebot an Information. Das Netzwerk ist zu komplex, um anschaulich Aufschluss über die zu vermittelnde Information zu geben. Daher muss im Vorfeld klar definiert werden, welche Informationen aus den Daten extrahiert werden sollen.

Ein weiteres denkbares Szenario wäre die Darstellung von Kommunikation. Dies lässt sich hervorragend mit Daten aus dem sozialen Netzwerk Twitter realisieren. Dort kann der Grad der Kommunikation festgestellt werden, indem gezählt wird wie stark die Nutzer über Retweets und @-Annotationen miteinander verbunden sind. Durch die Beschränkung auf ein Hashtag kann dabei sogar der Themenbereich eingegrenzt werden, beispielsweise auf ein aktuelles Fußballspiel. In beiden Fällen werden die resultierenden Visualisierungen jedoch erst interessant, sobald zusätzliche Informationen manuell hinzugefügt worden sind. Andernfalls leiden die komplexen Graphen an der mangelnden Übersicht. Zudem muss beachtet werden, dass ein Graph stets nur eine Momentaufnahme eines bestimmten Zeitabschnittes darstellt. Dabei befinden sich besonders die Beiträge aus sozialen Medien in einem ständigen und rasanten Wandel (vgl. König u. a.).

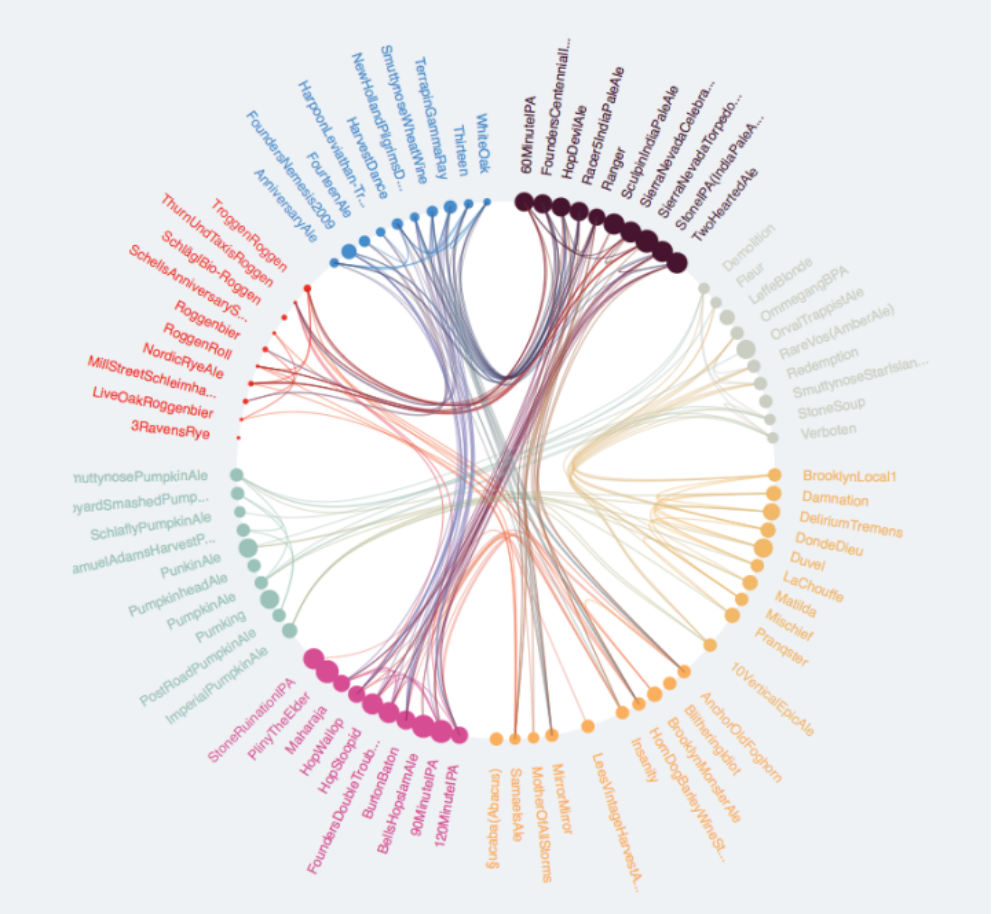

Abbildung 4.4: Ein Netzwerk von Bieren, die sich im Hinblick auf Aussehen, Geschmack und Aroma ähneln - Quelle: Yau

4.2.5 Zeitreihen

Ein Impuls in Form einer Pressemitteilung oder einer Live-Übertragung ist dazu in der Lage, die Meinung von Menschen in Bezug auf ein bestimmtes Thema schlagartig zu verändern (siehe Nicole Perlroth and Michael D. Shear). Umso wichtiger ist es, auch in puncto Visualisierung den Faktor Zeit zu beachten. So lassen sich Trends und mögliche Korrelationen über einen längeren Zeitraum beobachten. Im Fall von wiederkehrenden Mustern lassen sich entsprechend aussagekräftige Schlussfolgerungen ziehen. Besonders wichtig im Zusammenhang mit der Zeit ist der gewählte Zeitraum, der für eine Messreihe in Anspruch genommen wird. Im Fall von unterschiedlichen Zeiträumen ist die Umrechnung zwischen Stunden, Monaten oder Jahren

keinesfalls trivial und auch die zu wählenden Visualisierungsarten unterscheiden sich. Im Extremfall lassen sich die Daten aus verschiedenen Messreihen nicht mehr vergleichen (vgl. Kohlhammer u. a. (2013)).

Es bieten sich vor allem Säulen- und Liniendiagramme zur Visualisierung von Zeitreihen an. Je nachdem wieviele Werte dargestellt werden, ist eines der beiden vorzuziehen. Aber auch ein Punktdiagramm wäre denkbar, wenn die Differenz zweier Zeitreihen im Vordergrund der Visualisierung stehen soll. Zwei denkbare Beispiele für die Zeitreihen-Visualisierung sind die *Suchanfragen-Trends* von Google und die *trending Hashtags* von Twitter. Durch diese lassen sich Themen ermitteln und verfolgen, die aktuell im öffentlichen Interesse stehen. Generell bieten Zeitreihen die Möglichkeit, die Dynamik und den Fluss von sozialen Medien zu veranschaulichen, was einen wichtigen Aspekt von Visualisierungen darstellt (vgl. König u. a.).

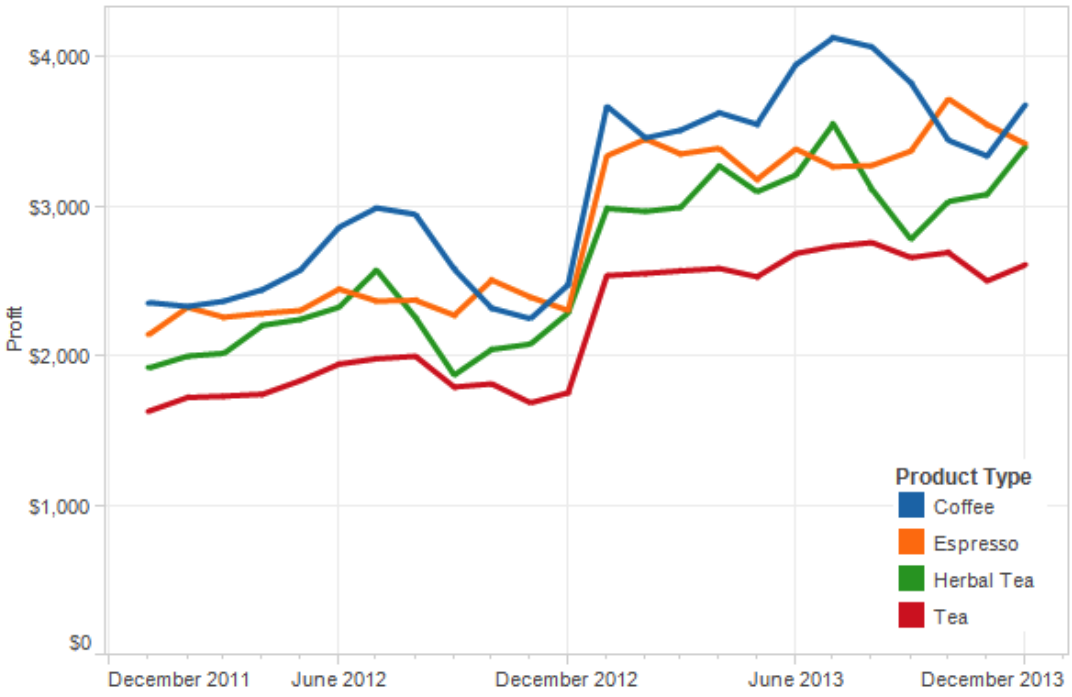

Abbildung 4.5: Dieses Punktdiagramm stellt den Profit verschiedener Produkte eines fiktiven Unternehmens in Abhängigkeit von der Zeit dar - Quelle: eagereyes.org (08.09.2015)

4.2.6 Geographische Daten

Die Meinungen und Interessen von Nutzern sozialer Medien variieren je nachdem, welchem kulturellen Kreis und welcher gesellschaftlichen Gruppe diese angehören. Zudem weichen die Lokalzeiten der verschiedenen Länder zu einem Zeitpunkt X bis zu zwölf Stunden von der koordinierten Tageszeit ab. Entsprechend interessant ist der geographische Standort, der einem Nutzer zum Zeitpunkt seiner Aussage zugeordnet ist. Um diesen adäquat darzustellen, eignen sich Karten. Diese sind in der Lage, komplexe Zusammenhänge leicht erfassbar zu machen (vgl. Kohlhammer u. a. (2013)). Eine erfolgreiches Beispiel wäre folgendes:

Abbildung 4.6: Diese Weltkarte visualisiert rund zehn Millionen Facebook-Freundschaftsbeziehungen - Quelle: Facebook

Diese Karte besteht aus einer Visualisierung von Facebook-Freundschaftsbeziehungen in Form von Verbindungslinien zwischen den Wohnorten der Nutzer. Jene Bereiche, die durch die Verbindungslinien besonders stark herausgearbeitet werden, besitzen dementsprechend besonders viele Freundschaftsverbindungen. Ein großes Problem bei der Erstellung von Karten auf Basis von öffentlich zugänglichen Daten ist jedoch der fehlende Zugriff auf genügend Geo-Informationen. In Bezug auf Twitter bedeutet dies zum Beispiel, dass lediglich 0,77 % der Tweets mit Geo-Tags versehen sind.[1] Allerdings ist diese Tatsache unabhängig vom sozialen Medium, da es meist freiwillig ist, diese Daten preiszugeben. Interessante und aussagekräftige Karten-Visualisierungen sind daher rar gesät (vgl. König u. a.).

[1] Semiocast

4.3 Diagramme

Im Kapitel Datentypen wurden je nach Art und Struktur der Daten verschiedene Diagramme und andere Visualisierungen zur Auswahl gestellt. Im Folgenden werden einige elementare Diagrammtypen im Detail vorgestellt. Dabei werden insbesondere die Vor- und Nachteile in den Fokus gestellt, damit im folgenden Kapitel der Sawatzki-Toolbox einige Bewertungskriterien zur Klassifikation der Visualisierungen bekannt sind. Zur Verdeutlichung wurden pro Diagrammart ein oder mehrere Beispielgraphen mithilfe von Microsoft Excel erstellt. Die zugrundeliegenden Daten entsprechen dabei dem Monatsumsatz eines fiktiven Unternehmens aufgeteilt auf die verschiedenen Kunden.

4.3.1 Säulen- und Balkendiagramm

In der Fachliteratur werden diese beiden Diagramme meistens in einem Atemzug genannt. Sie unterscheiden sich unter anderem an der Ausrichtung. Die Säulen des Säulendiagramms sind vertikal an der X-Achse ausgerichtet, während die Balken des Balkendiagramms horizontal an der Y-Achse ausgerichtet sind. Das Säulendiagramm eignet sich besonders gut für Zeitreihenvergleiche, weil sich an der Höhe der Säule gut erkennen lässt wie sich die entsprechende Entwicklung über die Zeit vollzogen hat. Jedoch wird das Säulendiagramm mit steigender Anzahl der Säulen schnell unübersichtlich. Sollte dieser Fall eintreten, ist das Liniendiagramm dem Säulendiagramm vorzuziehen.

Das Balkendiagramm hingegen eignet sich speziell für Strukturvergleiche von Regionen oder Produkten, weil die Bezeichnungen an den Balken deutlich länger werden können als beim Säulendiagramm, ohne die Übersicht zu zerstören. Auch Rangfolgenvergleiche bieten sich an, weil Balkendiagramme sich leicht gedanklich ordnen lassen (vgl. Kohlhammer u. a. (2013)). Im Folgenden wird entsprechend der Basisdaten ein Vergleich der beiden Diagramme in Bezug auf die Darstellung von Umsatzdaten verschiedener Kunden eines fiktiven Unternehmens angestellt. Da bei dieser Fragestellung klar der Vergleich der Kunden im Vordergrund steht und die Kundennamen vergleichsweise lang ausfallen, spielt das Balkendiagramm seine Vorteile im Vergleich zu dem Säulendiagramm aus:

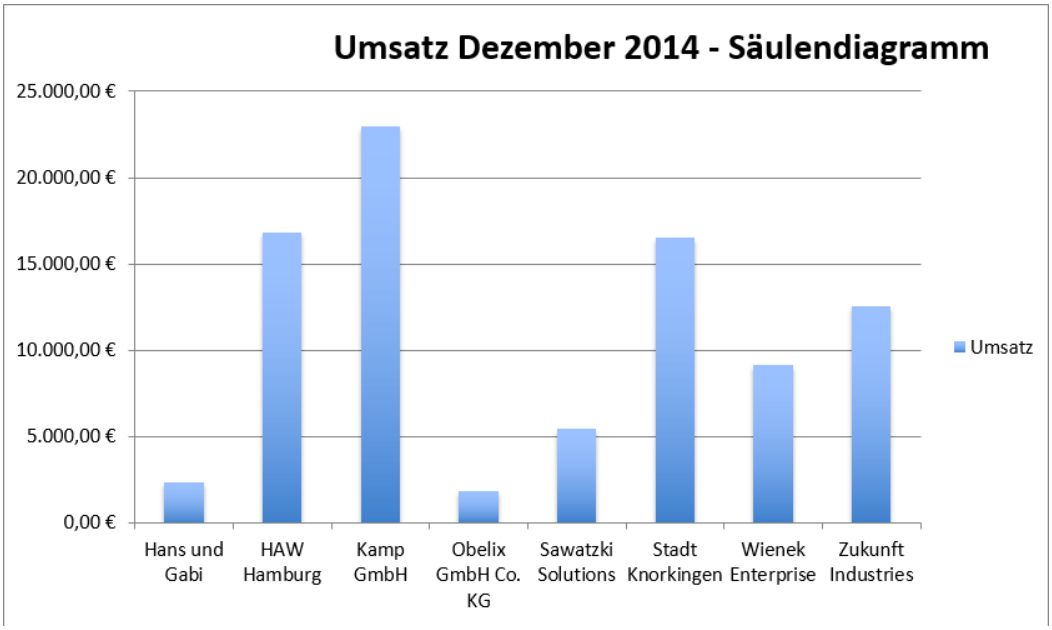

Abbildung 4.7: Umsatz der verschiedenen Kunden im Dezember 2014 als Säulendiagramm dargestellt

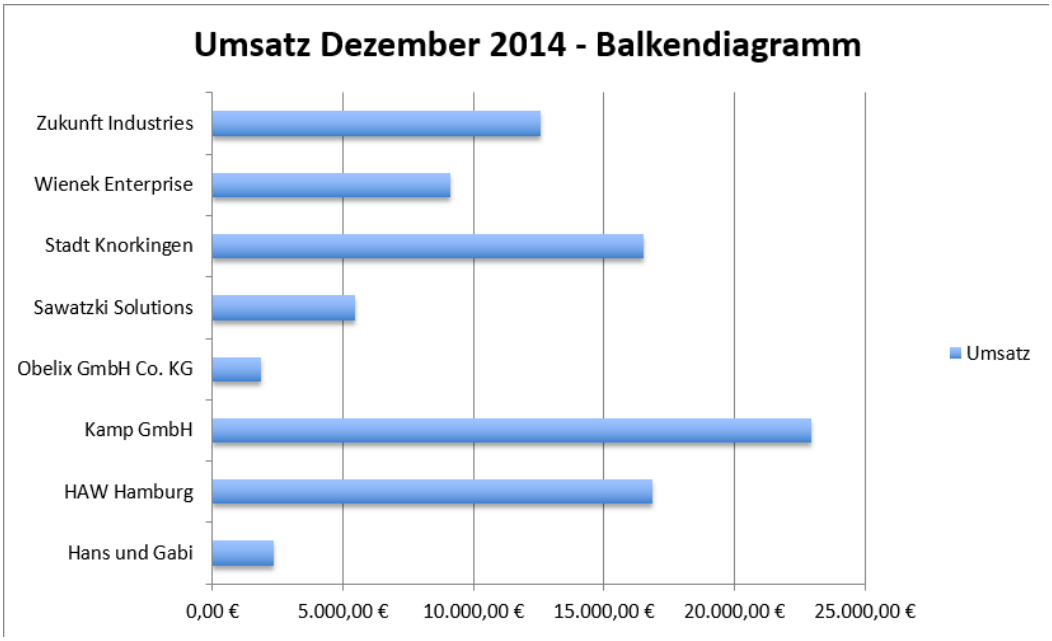

Abbildung 4.8: Umsatz der verschiedenen Kunden im Dezember 2014 als Balkendiagramm dargestellt

4.3.2 Kreisdiagramm

Sobald im Fokus der Visualisierung die Relation von etwas in Beziehung zum Gesamten dargestellt werden soll, bietet sich ein Kreisdiagramm an. In diesem Fall wäre es der Anteil von einzelnen Kundenumsätzen am Gesamtumsatz des Unternehmens. Wenn dieser Sachverhalt mithilfe eines Kreisdiagramms dargestellt wird, ist direkt ersichtlich, wie groß der Anteil des jeweiligen Kunden am Gesamtumsatz ist.

Allerdings wird ein Kreisdiagramm mit steigender Anzahl von Anteilen unübersichtlich. In der Fachliteratur wird von einer unzureichenden Übersichtlichkeit bei mehr als sechs Anteilen gesprochen (vgl. Kohlhammer u. a. (2013)). Es besteht zwar die Möglichkeit, die geringen Anteile unter Sonstige zusammenzufassen, jedoch gehen dadurch Informationen verloren. In diesem Fall wäre es die bessere Entscheidung, auf ein Säulen- oder Liniendiagramm zurückzugreifen. Dies lässt sich auch am folgenden Beispieldiagramm erkennen, denn unsere Basisdaten umfassen die Umsätze von acht Kunden:

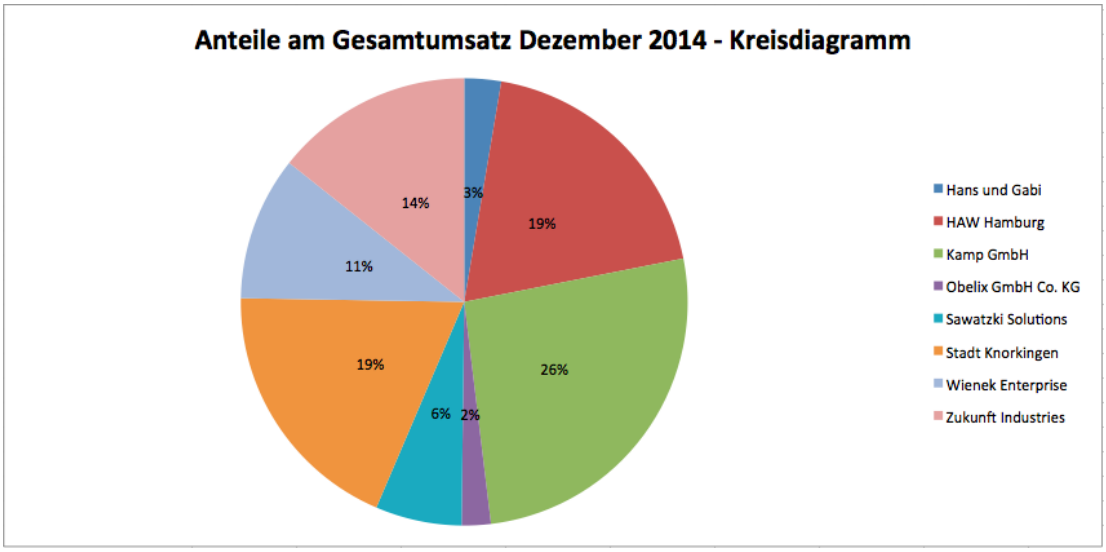

Abbildung 4.9: Anteile der verschiedenen Kunden am Gesamtumsatz im Dezember 2014

4.3.3 Liniendiagramm

Das Liniendiagramm eignet sich besonders für Zeitreihenvergleiche über einen langen Zeitraum. Die Skalierung der X-Achse ist variabel und lässt auch bei Millionen von Werten eine

übersichtliche Darstellung zu. Zudem eignet sich ein Liniendiagramm besonders, wenn es gilt einen Trend aufzuzeigen. Dieser kann wahlweise durch eine Trendlinie eingeblendet werden, ist jedoch auch gedanklich leicht nachzuvollziehen. Im nachfolgenden Beispiel wurde in Anlehnung des vorangegangenen Kapitels Zeitreihen auch der Faktor Zeit mit einbezogen. Entsprechend resultiert eine dreidimensionale Darstellung, welche zusätzlich einen Vergleich der Umsatzdaten der verschiedenen Monate zulässt.

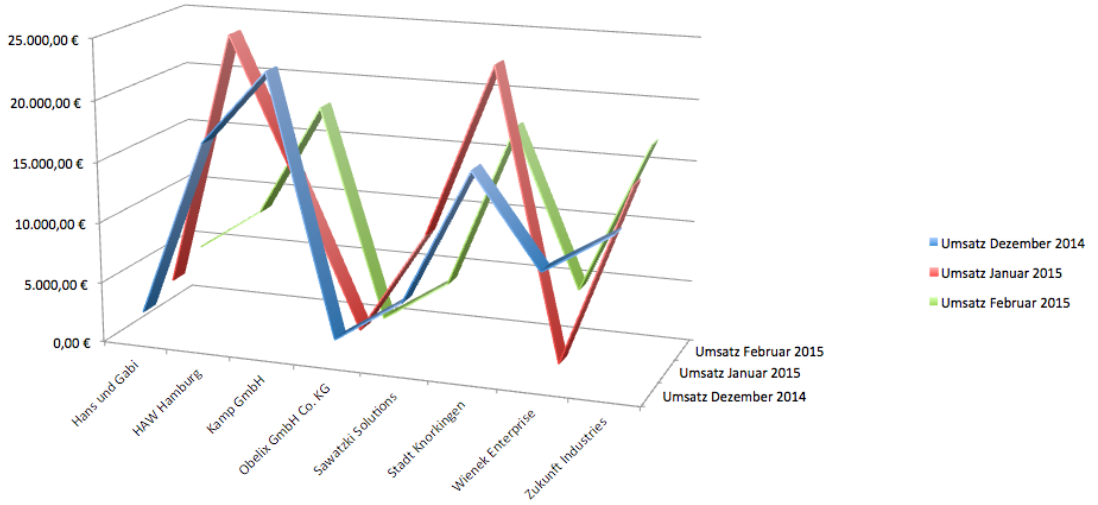

Abbildung 4.10: Umsatz in Abhängigkeit der verschiedenen Kunden und des Geschäftsmonats

4.3.4 Punktdiagramm

Ein Punktdiagramm ist in Betracht zu ziehen, wenn das Verhältnis zweier Variablen dargestellt werden soll. In dem Beispiel bezüglich des Umsatzes pro Kunde ließen sich somit zwei Kunden X und Y explizit miteinander vergleichen. Zu jedem gegebenen Wert ließe sich die Differenz der jeweiligen Werte direkt am Diagramm ablesen. Da dies in Anbetracht der hohen Kundenanzahl jedoch keine effektive Visualisierung mehr darstellt, wird zusätzlich ein zweites, wesentlich effektiveres Beispiel angeführt: Der Vergleich dreier Fußballvereine der ersten Bundesliga in Bezug auf die Tabellenposition. Drei unterschiedliche Werte pro Spieltag lassen sich schnell erfassen:

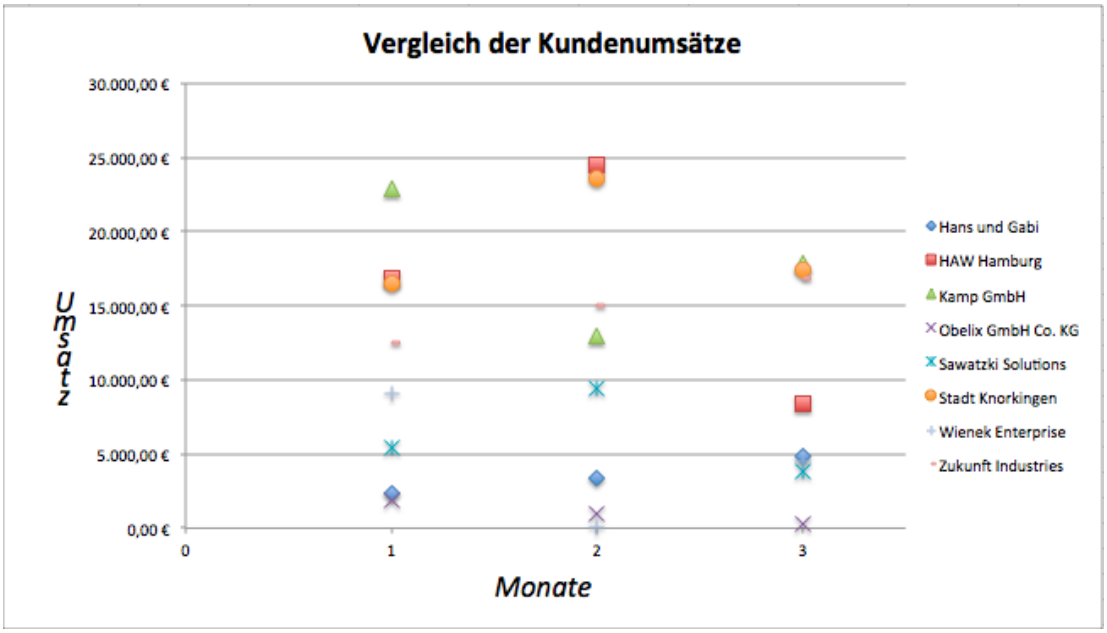

Abbildung 4.11: Umsatz der verschiedenen Kunden in Abhängigkeit des Geschäftsmonats

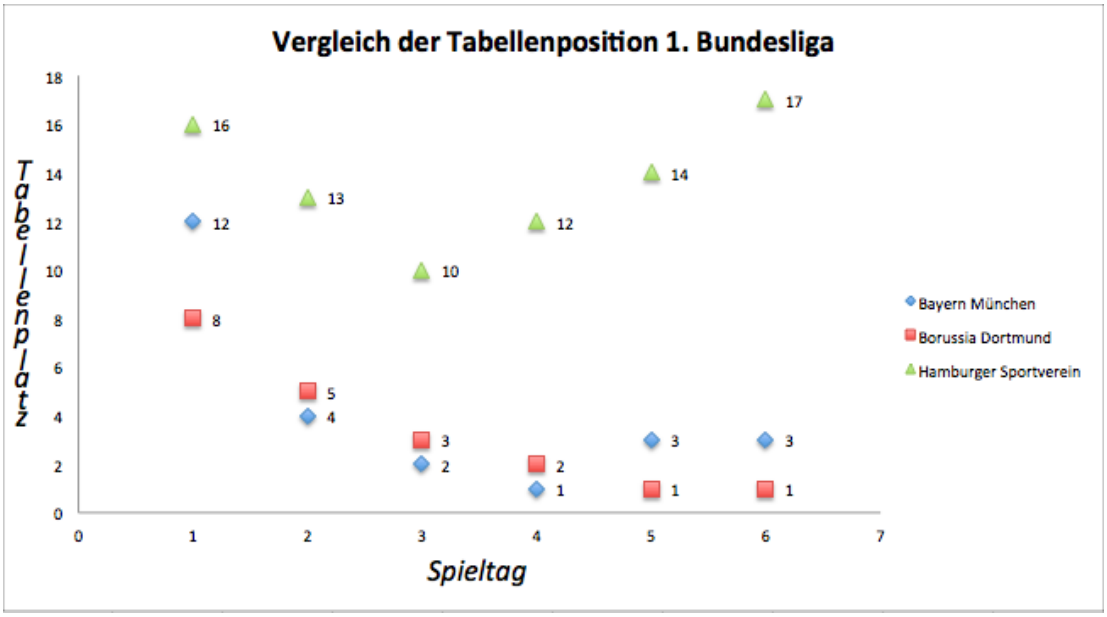

Abbildung 4.12: Vergleich des Tabellenplatzes verschiedener Fußballvereine der ersten Bundesliga in Abhängigkeit des Spieltages

5 Sawatzki-Ansatz

Der Sawatzki-Ansatz ist eine Visualisierung für die Marketing-Abteilung eines Unternehmens. Er soll hauptsächlich dazu dienen, einen Bewertungstrend zu erkennen und eine Zielgruppenanalyse zu ermöglichen. Vordergründig wurde diese Möglichkeit der Visualisierung für das soziale Netzwerk Twitter entworfen.

5.1 Sicht eines Social-Media-Analysten

Ein zentrales Element bei der Kommunikation auf Twitter ist das Hashtag. Dieses wird von den Usern als Schlagwort genutzt, um das Thema des Tweets, bzw. die Aussage des Tweets zu untermauern. Dies kann im Rahmen der Social-Media-Analyse genutzt werden, indem zunächst alle Tweets bezüglich eines bestimmten Hashtags über einen Zeitraum bestimmter Länge gesammelt werden. Zeitgleich oder im Anschluss daran werden innerhalb der gesammelten Tweets jene Hashtags gesucht, die besonders oft im Zusammenhang mit dem Ursprungs-Hashtag genannt werden. Dieses Vorgehen wiederholt sich anschließend mit den jeweils neu akquirierten Hashtags beliebig oft. Der Sawatzki-Ansatz muss sich dabei keinesfalls auf das soziale Netzwerk Twitter beschränken. Jegliches soziale Netzwerk, welches mit Tags, beziehungsweise Schlagworten arbeitet (*Instagram*, *Facebook*, *Tumblr* usw.) würde sich dazu eignen. Um den Vorgang transparenter zu gestalten, im Folgenden ein Beispiel:

Beispiel 2 (Firma Wienek bewirbt Teleporter) *Die auf Zukunftstechnologien spezialisierte Firma Wienek hat mit dem Teleporter ein neues Produkt entwickelt und dieses über Twitter beworben. Die Reichweite der Firma Wienek ist dank einigen populären Produkten, die sich bereits erfolgreich am Markt etabliert haben, recht groß. Folglich tauschen unter dem Hashtag #Teleporter nun etliche User ihre Meinungen zu dem Produkt aus. Ein hoher Prozentsatz dieser Tweets wird entsprechend der Twitter-Subkultur neben #Teleporter mit weiteren Hashtags versehen. Jene*

Hashtags, die besonders oft in Verbindung mit #Teleporter auftauchen, werden nun von der Firma Wienek gesammelt und gespeichert. Anhand dieser gesammelten Hashtags können nun weitere Hashtags akquiriert werden, die im weiteren Sinne mit #Teleporter assoziiert werden. Firma Wienek kann diesen Vorgang nun beliebig oft wiederholen, um immer mehr Hashtags zu erhalten, die allerdings mit jeder Iteration eine geringere Korrespondenz mit dem Ursprungs-Hashtag #Teleporter aufweisen. Anhand aller gesammelten Hashtags ist Firma Wienek schlussendlich in der Lage, verschiedene Rückschlüsse in Bezug auf das neue Teleporter-Produkt zu ziehen.

Die gesammelten Hashtags könnten zum einen dazu genutzt werden, die weitläufige Meinung der User bezogen auf das Ursprungs-Hashtag zu evaluieren. Dies wäre beispielsweise mit einer Sentiment-Analyse möglich, die in der Regel mit einem speziellen Wörterbuch arbeitet, welches Wörtern bestimmte Stimmungswerte zuweist (siehe König u. a.). Zudem wird deutlich, welche Zielgruppen mit dem Thema des Ursprungs-Hashtags in Berührung kommen. Angenommen, die Daten sind repräsentativ, wäre - rückblickend auf das vorangegangene Beispiel - die Firma Wienek in der Lage, auf die Bewertungen und Kritiken der User zu reagieren. Dadurch bestünde die Möglichkeit, die Hauptzielgruppen besser zu identifizieren und gezielter anzusprechen.

5.2 Sicht eines Informatikers

Formal beschrieben erzeugt der Sawatzki-Ansatz eine Menge von Hashtags, die miteinander in Korrespondenz stehen. Diese Menge wird gebildet, indem ausgehend von dem Ursprungs-Hashtag pro Iteration eine bestimmte Anzahl an weiteren Hashtags gefunden wird, die am meisten mit dem Ursprungs-Hashtag korrespondieren. In den Beispielen dieser Bachelorarbeit wurde diese Menge auf vier beschränkt. Nach der ersten Iteration stehen folglich vier neue Hashtags zur Verfügung, die in aktuellen Beiträgen des Twitter-Netzwerks direkt in Verbindung mit dem Ursprungs-Hashtag genannt wurden. In der nächsten Iteration würden für jedes dieser vier neuen Hashtags ebenfalls vier weitere Hashtags gefunden werden, die zwar nicht direkt mit dem Ursprungs-Hashtag genannt wurden, jedoch im Zusammenhang mit den vier Hashtags, die direkt mit dem Ursprungs-Hashtag in Verbindung stehen. Da jedes Hashtag einen Knoten im resultierendem Graphen repräsentiert, lässt sich die Gesamtanzahl der Knoten durch 4^k berechnen, wobei k die Anzahl der Iterationen darstellt. Die Anzahl der Knoten des Graphen wächst folglich mit der Anzahl der Iterationen exponentiell, deshalb wurde diese im Rahmen dieser Arbeit auf drei beschränkt, damit die entsprechenden Graphen mit maximal 64

Knoten nicht zu unübersichtlich werden. In der Praxis lässt sich die Anzahl der Iterationen per Parameter beliebig variieren, je nachdem wie breit der Nutzer die Analyse des Ursprungs-Hashtags wünscht. Wurden alle Iterationen ausgeführt, so wird der Graph rekursiv - ausgehend vom Ursrpungs-Hashtag - aufgebaut.

Aus der Sicht eines Informatikers ist der Sawatzki-Ansatz besonders deshalb interessant, weil er in einem Bereich angesiedelt ist, der am Markt hauptsächlich durch kostenpflichtige Lösungen abgedeckt wird. Ohne erhebliche Investitionen ist es kaum möglich, repräsentative Daten aus den sozialen Medien zu erhalten. Umso interessanter ist es, innovative Visualisierungen auf Basis der von den sozialen Medien zur Verfügung gestellten Metadaten zu kreieren.

5.3 Deutung am Beispielgraphen

Das Ergebnis könnte dabei wie folgt aussehen:

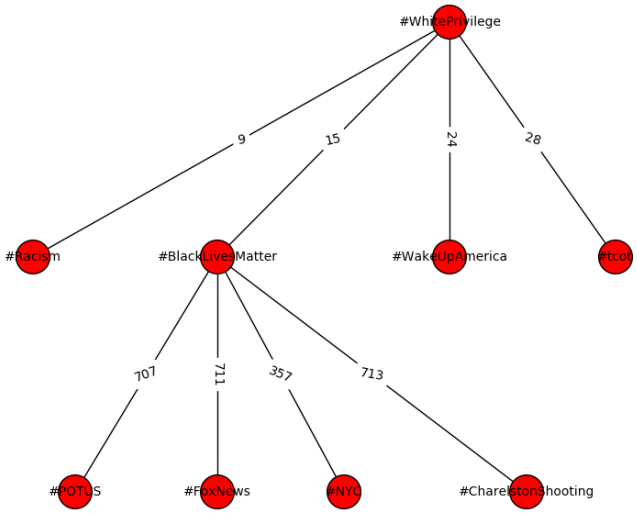

Abbildung 5.1: Mit dem Hashtag *WhitePrivilege* korrespondierende Hashtags als Baum visualisiert

Die Wurzel des Baumes bezeichnet das Ursprungs-Hashtag, das im Fokus des Interesses steht. In diesem Fall ist es das Hashtag *#WhitePrivilige*. Es handelt sich folglich um das sehr kontroverse Themengebiet der Gleichberechtigung. Es wurde 2 Stunden auf entsprechende Tweets gehorcht und das Ergebnis ist in der zweiten Ebene des Graphen zu erkennen: Hashtags die am meisten mit *#WhitePrivilige* in Verbindung gebracht werden sind *#Racism*, *#BlackLivesMatter*, *#WakeUpAmerica* und *#tcot*). Im Fall von *#BlackLivesMatter* hat die Datenstream-Komponente des Werkzeugkastens bereits weitere 2 Stunden nach entsprechenden Tweets gesucht und entsprechend in die Datenbank geschrieben. Daher war die Visualisierungs-Komponente in der Lage, diese Ergebnisse ebenfalls zu berücksichtigen. Dabei fällt auf, dass im Fall von *#BlackLivesMatter* ein wesentlich höheres Kantengewicht an den Kanten zu den Kindknoten angelegt ist als beim Ursprungs-Hashtag *#WhitePrivilige*. Das Kantengewicht beschreibt, wie oft zwei Hashtags innerhalb eines Beitrages genannt werden. Demzufolge wurden sehr viel mehr Tweets zum Hashtag *#BlackLivesMatter* verfasst als zum Ursprungs-Hashtag *#WhitePrivilige*. Dies liegt daran, dass zum Zeitpunkt als entsprechende Tweets von der STB gesammelt wurden, neun Afroamerikaner während einer Bibelstunde in einer Kirche in Charleston (South

Carolina) erschossen wurden. Aufgrund dieser Tragödie wurden entsprechend viele Tweets unter dem Hashtag *#BlackLivesMatter* erstellt.

Eine weitere Art der Darstellung des Sawatzki-Ansatzes wäre in Form eines Netzwerkes anstatt eines Baumes. Dazu folgende Beispielvisualisierung:

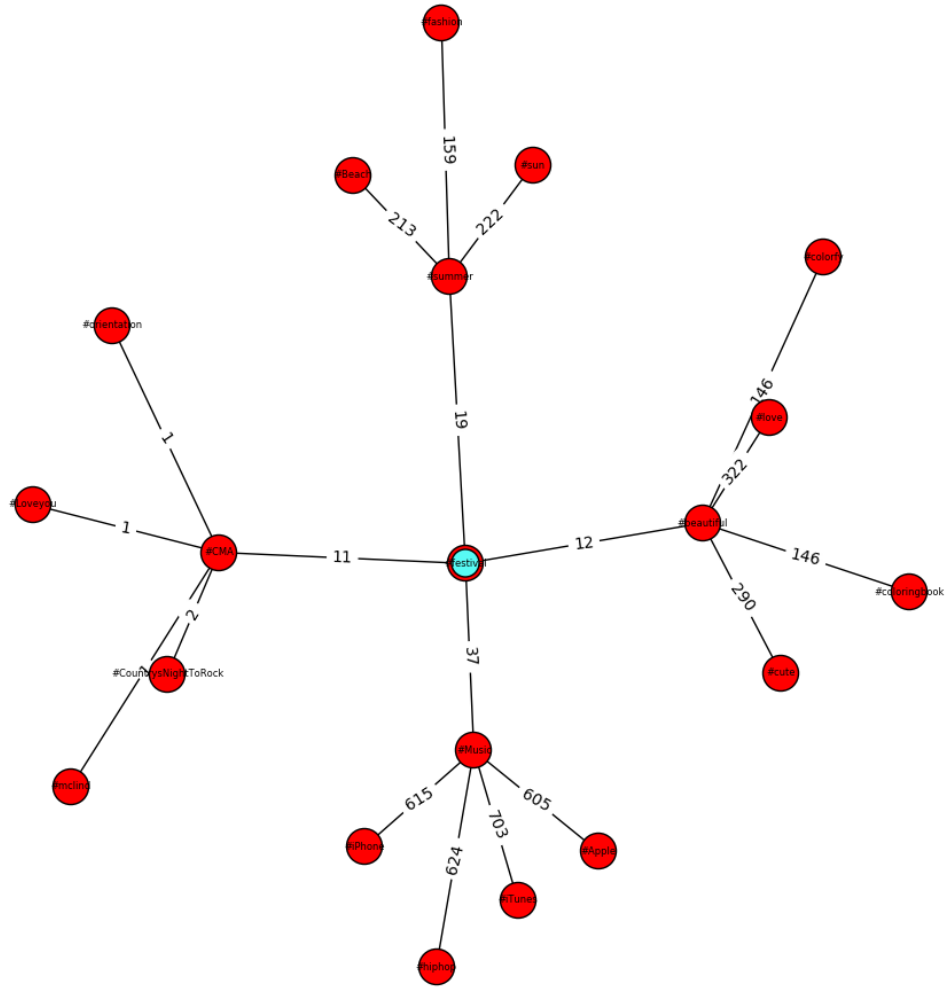

Abbildung 5.2: Mit dem Hashtag *Blatter* korrespondierende Hashtags als Netzwerk visualisiert

In diesem Fall bildet das Türkis gefärbte Ursprungs-Hashtag den Kernknoten des Netzwerkes, von dem alle Verzweigungen ausgehen. In dieser Form der Darstellung wird die Korrespondenz

der Hashtags nicht durch unterschiedliche Ebenen visualisiert, sondern an der Distanz im Netzwerk. Die Darstellung im Netzwerk erlaubt es, die Kantenlänge zu variieren, sodass ausgehend von einem bestimmten Hashtag, die Kantenlänge zu einem anderen Hashtag anzeigt, wie stark diese miteinander in Beziehung stehen. Wird ein Hashtag besonders oft im Zusammenhang mit einem anderen Hashtag gebraucht, ist die Kante zwischen diesen beiden Hashtag-Knoten entsprechend kurz. Je weniger ein Hashtag mit einem anderen Hashtag in Beziehung steht, desto länger wird folglich die Kante zwischen den entsprechenden Knoten.

Letztendlich bieten beide Darstellungen den gleichen Informationsgehalt, allerdings ist die Darstellungsform als Netzwerk zumindest im kleinen Rahmen effektiver, da der Grad an Korrespondenz zwischen zwei Hashtags aufgrund der dynamischen Kantenlänge deutlich schneller erkennbar ist. In der Baumdarstellung sind alle Hashtag-Knoten der ersten Ebene visuell gleich weit voneinander entfernt.

Erst beim Vergleich der Kantengewichtungen wird der Unterschied deutlich. Diese Abstraktion wird im Fall des Netzwerkes von der dynamischen Kantenlänge abgenommen, sodass der Zusammenhang schneller erkannt werden kann. Mit steigender Knotenanzahl spielt jedoch die Baumdarstellung aufgrund der Übersichtlichkeit seine Vorteile aus. Beim Betrachten eines Teilabschnittes ist bei dem Baumgraphen aufgrund der Ebenenstrukturierung klar, wie weit man sich vom Ursprungsknoten entfernt hat, während bei der Netzwerkdarstellung die Übersichtlichkeit schnell verloren geht.

6 Sawatzki-Toolbox

Dieses Kapitel beleuchtet die Entstehung der STB; ein Social-Media-Monitoring-Tool auf Basis der in dem Kapitel Social-Media-Monitoring-Tools gewonnenen Erkenntnisse. Die STB vereint unterschiedlichste Möglichkeiten der Visualisierung auf Basis eines einheitlichen Datenbestandes. Ziel der STB ist es unter anderem, verschiedene Visualisierungen gezielt miteinander zu vergleichen, um die jeweiligen Vor-und Nachteile präzise herauszustellen. Die Methoden zur Basis des Vergleichs von Visualisierungen wurden im Kapitel Visualisierung vermittelt.

Zudem soll sich die STB als Open-Source-Social-Media-Monitoring-Tool im Vergleich mit mehreren kommerziellen Produkten messen (siehe Benchmark). Vor allem, weil es kaum kommerzielle Anwendungen in diesem Bereich gibt, die sich effektiv kostenlos nutzen lassen, ist dieser Vergleich interessant.

6.1 Anforderungen

Der grobe Rahmen der STB wurde bereits durch die Anforderungen an ein Social-Media-Monitoring-Tool im Kapitel Social-Media-Monitoring-Tools festgelegt. Diesbezüglich soll sich die STB vor allem auf den Aspekt der Metadaten spezialisieren. Besonders der Sawatzki-Ansatz soll als innovative Visualisierung herausgearbeitet werden. Im Folgenden befinden sich die weiteren Anforderungen an die STB:

A1 Die Eingabequelle der STB hat keinen Einfluss auf die Funktionalität

A2 Die Nutzer der STB sollen nach einmaliger Reservierung Zugriff auf den gesamten (von allen Nutzern generierten) Inhalt haben

A3 Registrierte Nutzer können anhand eines Formulars Datensätze aus der Eingabequelle anfordern

A4 Generierte Datensätze stehen permanent zur Verfügung und können nicht gelöscht werden

A5 Registrierte Nutzer können auf Basis der verfügbaren Datensätze Visualisierungen erstellen

A6 Die Berechnung einer Visualisierung darf nicht mehr als 60 Sekunden in Anspruch nehmen

A7 Registrierte Nutzer können auf Basis der verfügbaren Visualisierungen Vergleiche von Visualisierungen erstellen

A8 Registrierte Nutzer sind in der Lage, eigens erstellte Visualierungen und Kommentare zu bearbeiten sowie zu löschen

A9 Von anderen Nutzern erstellte Kommentare und Visualisierungen lassen sich nur betrachten

A10 Administratoren haben die Möglichkeit, jegliche Inhalte (insbesondere Blogeinträge) zu erstellen, zu bearbeiten und zu löschen

A11 Die STB muss unter Google Chrome Version *45.0.2454.85* m arbeiten

A12 Die entscheidenden Bestandteile der Geschäftslogik der STB müssen anhand von deutschen Kommentaren erläutert werden

.

Im Kapitel Fazit der Sawatzki-Toolbox wird diese Anforderungsliste anhand des aktuell bestehenden Systems überprüft.

6.2 Fachliche Architektur

Der Datenbestand wird in folgenden Beispielen von Twitter-Tweets gebildet, um den Aspekt der Einheitlichkeit zu gewährleisten. Es sind entsprechend der Anforderung *A1* jedoch auch weitere Eingabequellen denkbar. Um die genaue Vorgehensweise der STB zu beschreiben, folgendes Datenmodell:

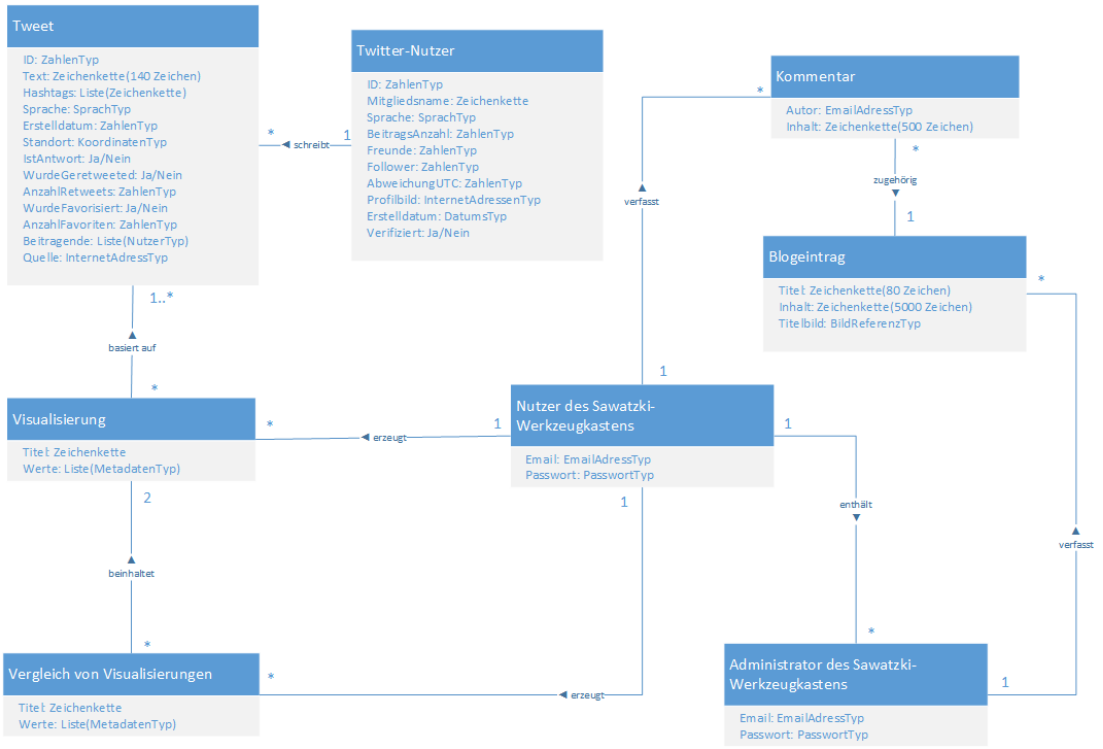

Abbildung 6.1: Fachliches Datenmodell der STB

Vor der Registrierung soll es dem Nutzer lediglich möglich sein, den Entwicklerblog zu lesen, um sich über die aktuellen Features der STB informieren zu können. Nach der Registrierung steht ihm beinahe die gesamte Funktionalität des Werkzeugkastens zur Verfügung. Der Nutzer kann alle bisher erstellten Visualisierungen betrachten und Kommentare zu Blogeinträgen erstellen. Löschen und bearbeiten kann er allerdings nur von ihm persönlich erstellte Visualisierungen und Kommentare. Der Administrator hingegen nimmt eine Sonderrolle ein. Er hat alle Möglichkeiten des registrierten Nutzers in Bezug auf alle Inhalte (nicht nur die von ihm selbst erstellten) und ist zudem in der Lage, Blogeinträge zu erstellen.

Um eine Visualisierung zu erstellen, hat der registrierte Nutzer der STB die Möglichkeit, sich ein beliebiges Attribut aus der Entität *Tweet* herauszusuchen. Anschließend lassen sich beliebig viele Visualisierungen erstellen, sofern diese in der Lage sind, das entsprechende Attribut sinnvoll zu illustrieren. Ein Säulendiagramm beispielsweise könnte keine nominalen Daten darstellen, während eine Wordcloud mit quantitativen Daten den ursprünglichen Sinn verfehlt (siehe Abschnitt Datentypen). Wurden mehrere Visualisierungen erstellt, lassen sich diese gegenüberstellen, um festzustellen, welche Informationen bei welcher Art der Visualisierung

besonders schnell ersichtlich werden; welche Informationen durch die Wahl der Darstellung erzeugt werden oder möglicherweise sogar verloren gehen.

6.3 Technische Architektur

Die STB basiert auf drei wesentlichen Komponenten, die in der folgenden Bausteinsicht dargestellt werden:

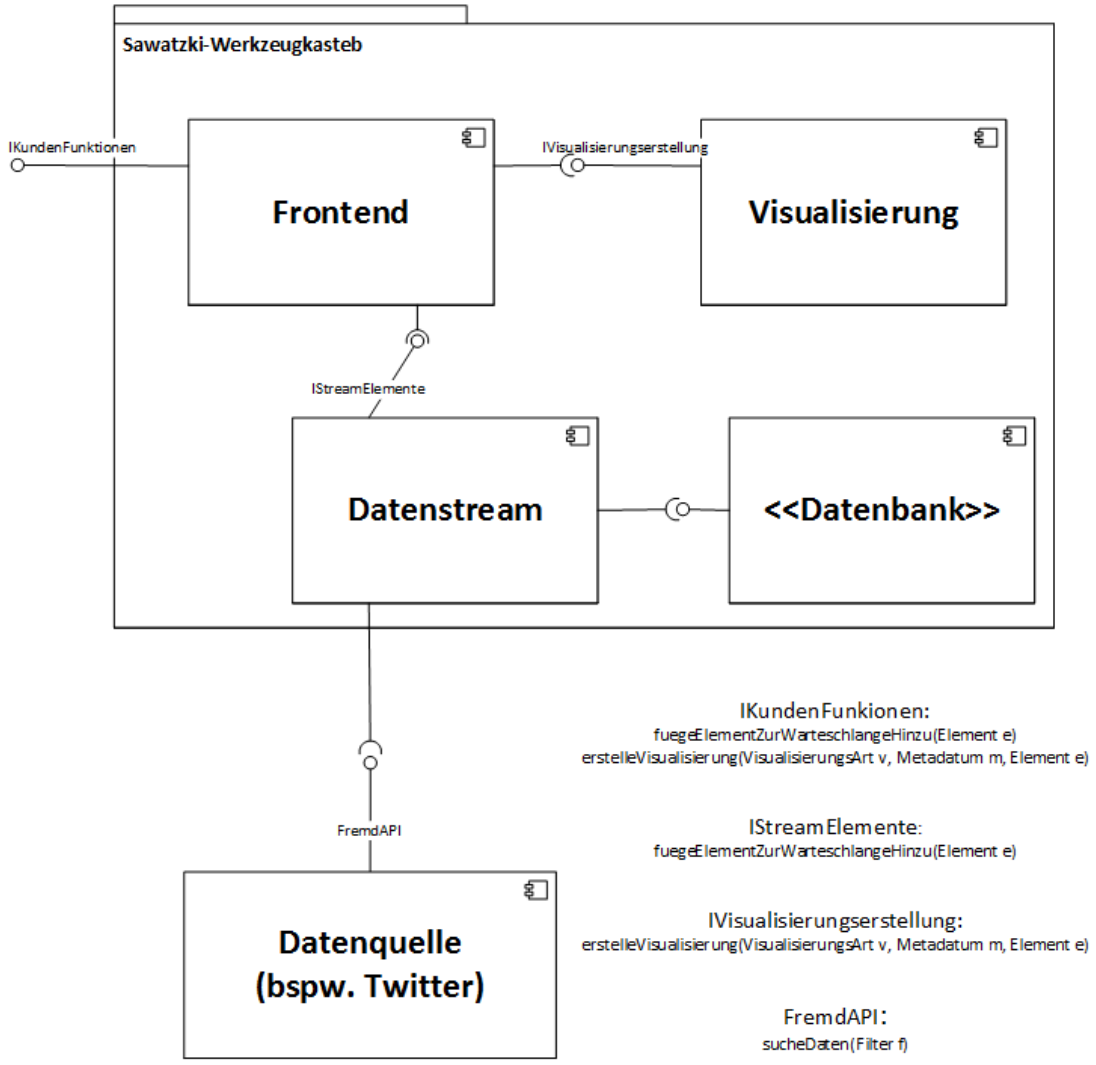

Abbildung 6.2: Bausteinsicht der STB

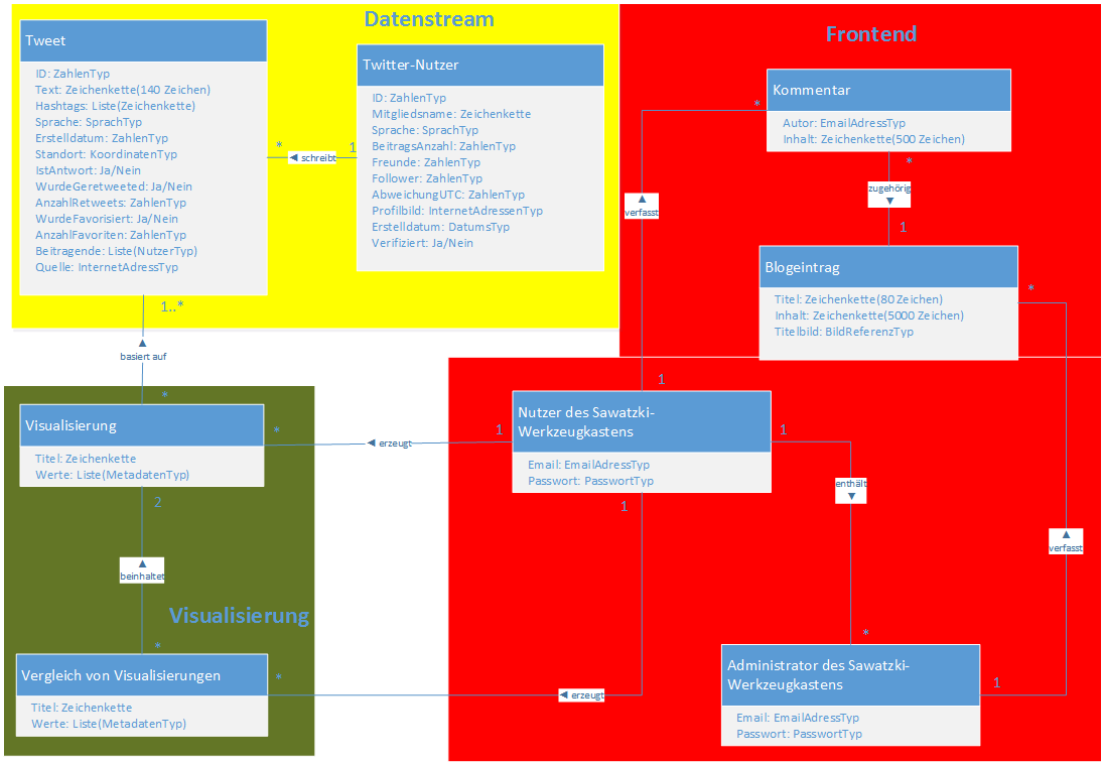

Abbildung 6.3: Zugehörigkeit der Entitäten des fachlichen Datenmodells zu den verschiedenen Komponenten

Dreh- und Angelpunkt des Systems bildet die Datenstream-Komponente. Diese nutzt beispielsweise die Twitter-Streaming-API, um Twitter-Tweets nach bestimmten Kriterien zu filtern. Generell arbeitet der Datenstream mit einer Queue, die verschiedenen Filter enthält, anhand dessen der Stream Daten aus der Datenquelle extrahieren soll. Die Queue kann dabei zum einen durch den Nutzer des Frontends beeinflusst werden: Dieser ist in der Lage einen Filter mit Kriterien zur Queue hinzuzufügen, nach denen die Datenquelle durchsucht werden soll. Zum anderen sucht sich der Datenstream auch selbst jene Elemente heraus, die im Interesse des Nutzers sein könnten. In diesem Fall wären dies Tweets, die Hashtags beinhalten, welche besonders oft in bereits gestreamten Tweets genannt wurden (siehe Sawatzki-Ansatz). Dieser Vorgang entspräche einer wesentlich kleiner gefassten Version der global gültigen *Trending Hashtags* von Twitter.

Die Daten werden während der Beschaffung laufend in eine Datenbank geschrieben, die neben dem Text auch alle weiteren Metadaten des jeweiligen Tweets erfasst. Je nachdem wie relevant

das Thema zur Zeit der Datenakquirierung ist und wie limitierend die Twitter-API wirkt, würde dieser Schritt unterschiedlich viel Zeit benötigen, um genügend Daten für eine repräsentative Analyse festzuschreiben. Anstatt darauf zu warten, bis eine bestimmte Anzahl an Tweets gefunden wurde, sammelt der Datenstream in einem vom Anwender bestimmten Zeitraum Tweets zu einem Filter. Dies ist die sichere Variante, denn in seltenen Fällen werden bspw. Hashtags nach denen gefiltert werden soll plötzlich nicht mehr genutzt und es würde beliebig lange dauern, eine bestimmte Anzahl der entsprechenden Tweets zu sammeln. Zudem würde ein Account für die Twitter-Streaming-API für die Dauer der Suchzeit blockiert werden. Da Twitter nur eine begrenzte Anzahl an Zugriffen pro IP-Adresse auf die Streaming-API erlaubt, wäre dies fatal für die Performance und Verfügbarkeit der Datenstream-Komponente, die mit diesen Accounts arbeitet.[1] Ähnliches gilt sicherlich für weitere Fremdsystem-APIs. Daher sollte die Datenstream-Komponente generell zeitbasiert arbeiten, um diesem Problem aus dem Weg zu gehen.

Da Twitter vollständige Daten aus der Vergangenheit nicht kostenlos anbietet, ist an dieser Stelle entscheidend, ob die genutzte Streaming-API das Kriterium des Vollständigkeit erfülllt. Da die Twitter-Streaming-API anhand eines Hashtags durchsucht wird, handelt es sich jeweils um einen kleinen Teilausschnitt des Twitter-Netzwerkes. Jedoch war anhand verschiedener Stichproben auf Hashtags mit hoher Frequentierung keinerlei Einschränkung feststellbar. Die gesammelten Tweets der Datenstream-Komponente waren in der Menge sogar deutlich mehr als die von der Twitter-Webseite bereitgestellten Live-Tweets. Dabei wurden in der Spitze sogar mehr als 60 Tweets pro Sekunde von der Datenstream-Komponente gesammelt.

Die Visualisierungs-Komponente ist dafür zuständig, aus den von der Datenstream-Komponente gesammelten Daten Visualisierungen zu erstellen, wobei ein Großteil der laut Fachliteratur relevanten Visualisierungen unterstützt werden. Als Innovation im Sinne der Anforderungen an ein Social-Media-Monitoring-Tool kommt der Sawatzki-Ansatz hinzu. Demzufolge hat der Nutzer die Wahl zwischen Säulendiagramm, Balkendiagramm, Liniendiagramm, Punktdiagramm und dem Sawatzki-Ansatz. Sobald eine Anfrage für eine Visualisierung vom Anwender im Rahmen der Kundenfunktionen im Frontend gestellt wird, wird der gewählte Datenbestand, das gewünschte Metadatum (Tweet-Attribut) sowie die Visualisierungsart an die Visualisierungs-Komponente übermittelt. Anhand eines Pythonskriptes wird anschließend die gewünschte Visualisierung berechnet und in einem für das Frontend-zugänglichem Ordner gespeichert. Im Fall von Ruby on Rails ist dies üblicherweise der *Assets-Ordner*. Der Dateiname wird durch

[1] https://dev.twitter.com/streaming/overview

den Unix-Zeitstempel in Millisekunden zum Zeitpunkt des Skriptaufrufs bestimmt. Dieser Zeitstempel wird der Frontend-Komponente auch als Rückgabewert übermittelt, sodass diese den Pfad zur entsprechenden Visualisierung nachvollziehen kann. Die Grapherstellung des Pythonskriptes basiert auf unterschiedlichen Frameworks. Die beiden elementaren Frameworks bilden dabei *Networkx*[2] und *pygal*[3].

Bezüglich der Entwicklung der Frontend-Komponente wurde eine Ruby-on-Rails-Webanwendung gewählt. Der Gedanke hinter dieser Entscheidung ist es, die STB einer breiteren Öffentlichkeit zur Verfügung zu stellen, sodass die Nutzer von den unterschiedlichsten Datensätzen, die ein beliebiger Nutzer generiert, profitieren können. Bei der Erstellung von Visualisierungen stehen dem Nutzer folglich nicht nur die eigens kreierten Datensätze zur Verfügung, sondern auch alle Datensätze, die von anderen Usern generiert wurden. Dies ermöglicht dem individuellen Nutzer vor allem, ein wesentlich breiteres Bild von den sozialen Medien zu erhalten als er es bekommen würde, wenn er lediglich seine eigenen Datensätze isoliert betrachten würde. Inwiefern sich dieser Ansatz weiter verfolgen ließe, wird im Kapitel Möglichkeiten und zukünftige Entwicklung geklärt.

Um das Frontend zudem übersichtlich und strukturiert zu halten, wird Bootstrap eingesetzt. Eine Bilderreihe mit entsprechenden Mockups samt Erklärungen befindet sich im Anhang unter Funktion und Mockup der STB.

Eine beispielhafte Benutzung der STB auf Basis der beschriebenen Architektur sieht dabei wie in folgendem Sequenzdiagramm dargestellt aus:

[2] https://networkx.github.io/
[3] http://pygal.org/

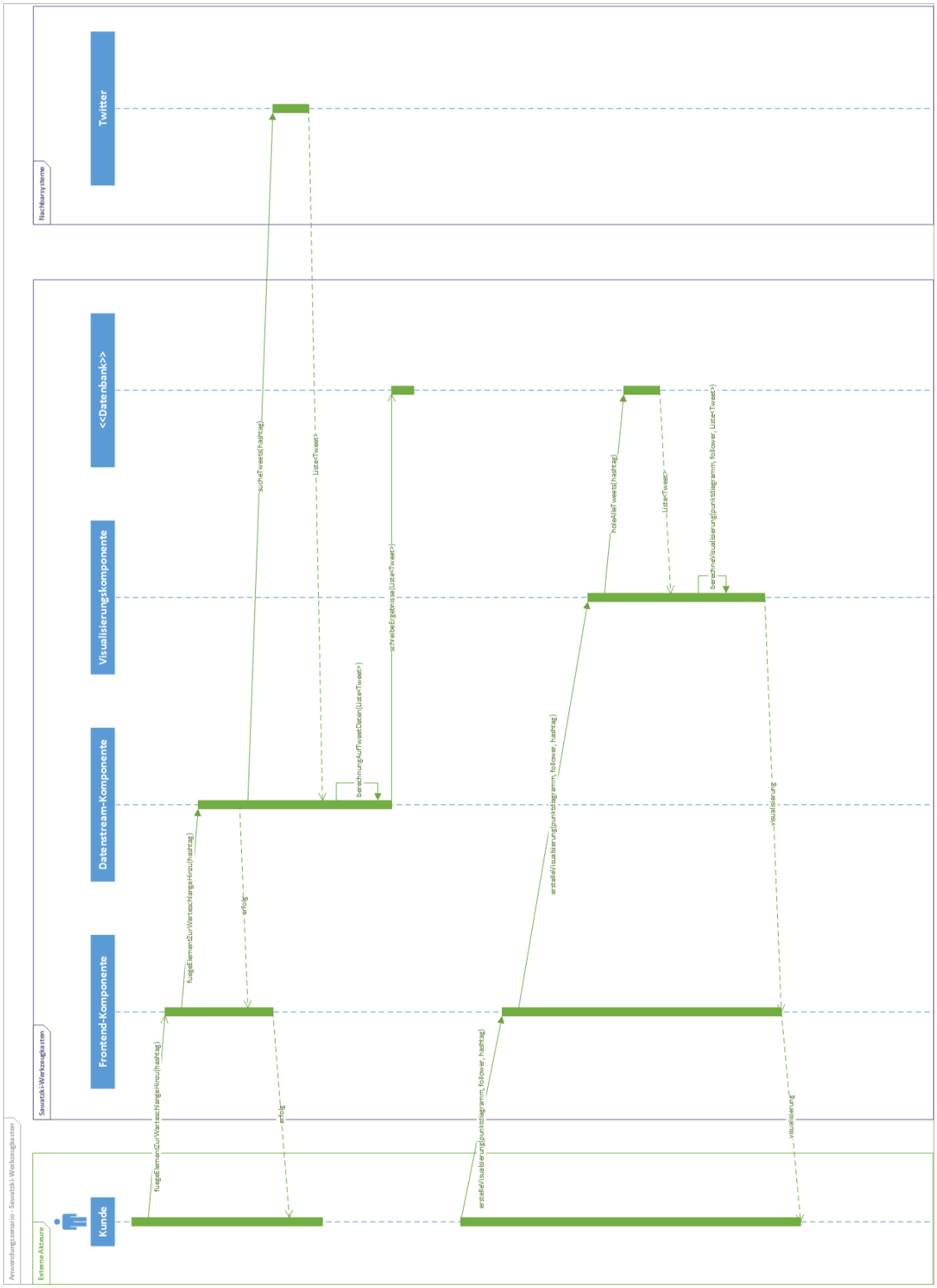

Abbildung 6.4: Sequenzdiagramm mit einem beispielhaftem Anwendungsszenario (Soziales Netzwerk Twitter)

Dabei fügt der Kunde zunächst per Frontend-Formular einen Filter in die Warteschlange des Datenstreams ein. Die Frontend-Komponente leitet diesen Aufruf an die Datenstream-Komponente weiter, welche die Eintragung des Filters bestätigt. Sollte der Filter unzulässig sein, könnte an dieser Stelle auch ein Fehlercode an die Frontend-Komponente weitergeleitet werden. Das ist beispielsweise der Fall, wenn sich ein identischer Filter bereits in der Warteschlange des Datenstreams befindet. Im Folgenden wird jedoch von einem Erfolgsszenario ausgegangen. Die Frontend-Komponente leitet entsprechend die Erfolgsmeldung an den Nutzer weiter. Sobald der Filter von der Datenstream-Komponente aus der Warteschlange entfernt wurde, wird das entsprechende soziale Netzwerk für einen Zeitraum X auf den entsprechenden Filter gescannt und laufend Beiträge mitsamt aller verfügbaren Metadaten gesammelt. Damit Daten wie die korrespondierenden Hashtags direkt vorliegen, führt die Datenstream-Komponente noch einige Berechnungen auf den gesammelten Daten aus und speichert anschließend die Rohbeiträge sowie die berechneten Daten in die Datenbank.

Nachdem sich einige Datensätze in der Datenbank befinden, hat der Kunde die Möglichkeit eine Visualisierung zu erstellen. Dies macht er ebenfalls über ein entsprechendes Formular des Frontends. Die Frontend-Komponente leitet den Aufruf mit Informationen über die gewählte Visualisierungsart, das gewünschte Metadatum sowie dem gewünschten Element das visualisiert werden soll an die Visualisierungs-Komponente weiter. Die Visualisierungs-Komponente stellt anschließend eine Anfrage an die Datenbank, um sich alle Daten bezüglich des gewünschten Elementes und Metadatums zu besorgen. Mithilfe dieser gesammelten Daten kann die Visualisierungs-Komponente anschließend die gewünschte Visualisierung berechnen. Das Ergebnis wird an die Frontend-Komponente weitergeleitet, die es wiederum dem Nutzer zukommen lässt.

6.4 Probleme bei der Umsetzung

Neben mehrerer geringerer Schwierigkeiten haben sich besonders drei Problematiken für die spezifische Umsetzung der STB auf Basis der vorangegangenen Architektur ergeben:

Eine Schwierigkeit bestand in der Auswahl des Frameworks zur Erstellung der Graphen im Sinne des Sawatzki-Ansatzes. Für eine effektive Darstellung sollten die Wurzel färbbar, die Kantenlänge variabel und eine Kanten- sowie Knotenbeschriftung möglich sein. Jedoch unterstützen auch etablierte Frameworks nicht alle gewünschten Features. Selbst bei dem

gewählten Framework *networkx* haben sich einige Schwierigkeiten ergeben: Ordnet man den Graphen als Baum an, so ist die Kantenlänge nicht mehr variabel, weil ansonsten Knoten der gleichen Ebenen in unterschiedlicher vertikaler Lage dargestellt werden müssten, was den Graphen nicht mehr klar als Baum erscheinen ließe. Außerdem war es schwierig, die entsprechenden Graphen mit einer dynamischen Größe zu exportieren, damit die resultierenden Bildformate eine passende Auflösung besitzen, sodass die Beschriftungen deutlich lesbar und die Kantenlängen unterscheidbar bleiben.

Die zweite große Schwierigkeit bestand in der Datenstream-Komponente. Bei der Nutzung der Twitter-Streaming-API ergab sich ein Problem mit der Zwangstrennung des Internets, sofern die entsprechende Komponente in einem privaten Haushalt betrieben wurde. Verschiedene Ansätze, dieses Problem durch eine automatische Wiederherstellung der Verbindung zu beheben, scheiterten, sodass die Datenstream-Komponente sowie die Datenbank auf einen professionellen Server ohne Zwangstrennung umgezogen wurden. Anfragen an den Datenstream sowie die Datenbank erfolgten schließlich nicht mehr lokal, sondern entfernt, was den Overhead der entsprechenden Anfragen leicht erhöhte.

Verschiedenste Probleme bestanden zudem in der Konsistenz der Datenbank. Dabei führten meist Sonderfälle, die in der Planungsphase nicht bedacht wurden zu entsprechenden Inkonsistenzen. Einer dieser Sonderfälle bestand in dem Problem, dass zwei oder mehr Nutzer zu unterschiedlichen Zeiten einen Filter in die Warteschlange der Datenstream-Komponente eintrugen. Die Datenstream-Komponente sammelte nun zu unterschiedlichen Zeiten Daten zu dem gleichen Filter, was problemlos klappte. Wollten die Nutzer jedoch Visualisierungen aus ihren Datensätzen erstellen, so wurde nicht zwischen diesen verschiedenen Anfragen zugehörigen Datensätzen unterschieden, sondern die Datenstream-Komponente hat lediglich den neuesten Datensatz berücksichtigt und den alten in der Datenbank überschrieben. Demzufolge musste die Zeit der Filterung in die Datenbank mit aufgenommen werden, um diesen Sonderfall zu unterscheiden.

Wenngleich die genannten Probleme die Fertigstellung der STB in seinem aktuellen Status beeinträchtigt haben, so ließ sich die zuvor festgelegte Architektur dennoch ohne weitreichende Umstrukturierungen umsetzen. Dennoch hat dieser Rückblick gezeigt, dass sich feingranulares Denken bei der Architektur auszahlt. Denn wenn die Probleme in der Entwicklung auftreten, verursachen sie einen weitaus höheren Zeitaufwand.

6.5 Benchmark

Wie schlägt sich die STB im Vergleich mit Systemen, die sich bereits auf dem Markt befinden? Diese Frage gilt es im folgenden Abschnitt zu klären. Als Bewertungsbasis dienen die im Abschnitt Anforderungen an ein Social-Media-Monitoring-Tool vorgestellten Aspekte Metadaten, Resolution, Integration, Abgleich und Interface. Für jede Kategorie wurden verschiedene Bewertungskriterien gewählt, die in unterschiedlichem Maße in die Bewertung der verschiedenen Kategorien eingeflossen sind - je nachdem wie entscheidend sie für diese Kategorie sind. Verglichen wurde die STB mit den Social-Media-Monitoring-Tools *Quintly*[4], *SumAll*[5] und *Hootsuite*[6]. Dabei handelt es sich jeweils um kommerzielle Lösungen mit entsprechend hohen Nutzerzahlen und zum Teil namhaften Kunden. Dabei wurde im Rahmen dieser Bachelorarbeit von den Anbietern der Tools nur ein kleiner Teilbereich der vollständigen Funktionalität, beziehungsweise eine Testversion zur Verfügung gestellt. Jedoch war bei allen Systemen ersichtlich, welche zusätzlichen Funktionalitäten sich bei einem Kauf ergeben würden. Das Monitoring von zumindest einem sozialen Netzwerk wurde bei allen Tools auch in der Testversion gestattet. Aber es ist nicht auszuschließen, dass bei einem freigeschaltetem Account weitere Funktionalitäten hinzukommen, die in dieser Bachelorarbeit nicht berücksichtigt werden konnten.

6.5.1 Sawatzki-Toolbox

In Bezug auf den Aspekt der Metadaten kann der Nutzer der STB aus allen quantitativen Daten wählen, die Twitter für die Tweets bereithält und daraus eine Visualisierung erstellen. Die quantitativen Werte werden direkt den Metadaten eines Tweets entnommen; es handelt sich nicht um berechnete Werte. Innovation hingegen bietet der Sawatzki-Ansatz. Dieser stellt die Korrelation verschiedener Schlagworte dar, was aus den reinen Tweet-Metadaten nicht ersichtlich ist. Dabei nutzt er die Möglichkeiten der Netzwerk-Visualisierung, um zusätzliche Informationen bereitzustellen. Wie im Abschnitt Grundlagen des Kapitels Visualisierung bereits erwähnt wurde, ist dies besonders bei den Netzwerken ein entscheidendes Kriterium für eine effektive Darstellung. Filterungsmöglichkeiten bestehen für den Nutzer zum einen bei der Wahl des Datensatzes. Sofern der Datenstream einen Filter bereits zwei Mal oder öfters auf ein soziales Medium angewandt hat, kann der Nutzer den Zeitraum des jeweiligen

[4] https://www.quintly.com/
[5] https://sumall.com/
[6] https://hootsuite.com/de/

Datensatzes wählen. In Bezug auf die Darstellung der Visualisierungen besteht derzeit noch keine andere Möglichkeit als die Daten in Abhängigkeit von den Tagesstunden darzustellen. Im Abschnitt Möglichkeiten und zukünftige Entwicklung werden jedoch Ideen beschrieben, die Filterungsmöglichkeiten der STB zu erweitern.

Weiterführende Quellen im Sinne der Anforderungen an ein Social-Media-Monitoring-Tool werden bisher nicht berücksichtigt. Auch eine anspruchsvolle Textanalyse wurde nicht durchgeführt. Der Sawatzki-Ansatz wertet zwar die Texte aus, allerdings wird lediglich nach Worten gesucht, die mit einem #-Zeichen beginnen und diese werden entsprechend gezählt. Das ist kein Algorithmus im Sinne des *Natural Language Processing*.

Beim Vergleich von verschiedenen Visualisierungen besteht die Möglichkeit, diese explizit gegenüberzustellen, um sie genau miteinander zu vergleichen. Einzelne Visualisierungs-Metriken lassen sich dabei jedoch nicht an- oder abschalten. Verschiedene Datensätze zu aggregieren ist ebenfalls nicht möglich.

Vorhersagen über die weitere Entwicklung bestimmter Metriken werden nicht getroffen.

Auch ein Dashboard oder eine Reporting-Funktionalität ist nicht implementiert.

6.5.2 Quintly

Quintly weist in Bezug auf die Metadaten das umfangreichste Angebot des Testfeldes auf. Die Testversion bezog sich lediglich auf die Nutzungen der Facebook-Visualisierungs-Metriken, jedoch wurden nicht nur die von Facebook bereitgestellten Metadaten in unterschiedlichste Visualisierungs-Metriken aufgesplittet, sondern vereinzelt auch aus den Metadaten berechnete Visualisierungs-Metriken bereitgestellt. So wurde beispielsweise eine Visualisierungs-Metrik bereitgestellt, welche die Änderungsrate der Interaktionen in Form von *Likes*, *Comments* und *Shares* darstellt. Dabei lassen sich für beinahe alle Metriken die Zeiteinheiten wählen, in der die Daten visualisert werden sollen. Der Nutzer hat dabei die Wahl zwischen Tagesstunden, Wochentagen oder einem eigens definiertem Zeitraum. Die Art der Visualisierung ändert sich entsprechend, sodass diese maximal effektiv ist.

Auch bei Quintly werden zumindest in der Testversion keine weiteren Quellen zu Analysezwecken hinzugezogen. Bezüglich der Textanalyse gibt es lediglich vage Ansätze. Die Texte werden auf Links oder Multimedia-Dateien überprüft und entsprechend kategorisiert.

Vergleiche verschiedener Visualisierungen lassen sich über das Dashboard simulieren. Dort können verschiedene Visualisierungen nebeneinander positioniert werden, um diese anschließend zu vergleichen. Darüber hinaus lassen sich verschiedene Visualisierungs-Metriken innerhalb einer Visualisierung anwählen. Die Aggregation von Daten ist hingegen nicht möglich.

Vorhersagen über die zukünftige Entwicklung bestimmter Visualisierungs-Metriken werden nicht getroffen.

Ein Dashboard wurde im Sinne des Kapitels Anforderungen an ein Social-Media-Monitoring-Tool realisiert. Dort lassen sich die gewünschten Metriken nach Wünschen des Nutzers einfügen und anordnen. Auch ein Report von gewünschten Metriken lässt sich in vielen gängigen Datenformaten exportieren.

6.5.3 SumAll

Das Social-Media-Monitoring-Tools *SumAll* bietet bei der Auswahl der Metadaten lediglich ausgewählte Visualisierungs-Metriken, die sich allesamt auf von dem sozialen Netzwerk zur Verfügung gestellte Daten stützen. In der Testversion lässt sich beispielsweise das soziale Netzwerk Twitter beobachten. Dort stehen nur Visualisierungs-Metriken wie *Favoriten*, *Retweets* etc. zur Verfügung. Berechnete, innovative Visualisierungs-Metriken sind nicht vorhanden. Filterungsmöglichkeiten bestehen dabei zum einen in dem gewählten Zeitraum, wobei sich die Art der Visualisierung nicht anpasst, und in der Größe der Visualisierung, wobei zwischen zwei verschiedenen Größen gewählt werden kann.

Vom Hinzuziehen weiterer Quellen scheint *SumAll* keinen Gebrauch zu machen. Und auch von Textanalyse beziehungsweise *Natural Language Processing* wird anscheinend kein Gebrauch gemacht.

Der Vergleich von Visualisierungen wird direkt nicht unterstützt. Allerdings lassen sich innerhalb einer Visualisierung verschiedene Visualisierungs-Metriken auf Basis verschiedener Datensätze (Kanäle sozialer Medien) zuschalten, um diese miteinander zu vergleichen.

Vorhersagen bezüglich der zukünftigen Entwicklung werden nicht getroffen.

Ein Dashboard ist nicht vorhanden, allerdings lassen sich die Visualisierungen in einem Report exportieren.

6.5.4 Hootsuite

Im Vergleich mit dem Tool *SumAll* nutzt *Hootsuite* ebenfalls hauptsächlich Visualisierungs-Metriken, die lediglich die vom sozialen Netzwerk verfügbaren Metadaten visualisieren. Jedoch gibt es vereinzelt auch innovativ berechnete Visualisierungs-Metriken. Eine davon ähnelt dem Sawatzki-Ansatz und ermittelt die meist benutzten Schlagwörter in Bezug auf ein Thema. Die Visualisierungen lassen sich nach einem bestimmten Zeitraum filtern, wobei sich die Visualisierungsarten nicht verändern.

Weiterführende Quellen zum Zwecke der Analyse werden nicht betrachtet. Bezüglich der Textanalyse wird ähnlich dem Sawatzki-Ansatz innerhalb der Texte aus den sozialen Netzwerken nach bestimmten Stichworten gesucht, jedoch handelt es sich in diesem Fall ebenfalls nicht um eine anspruchsvolle Textanalyse.

Ein Vergleich von Visualisierungen gestaltet sich schwierig. Gewünschte Visualisierungen lassen sich innerhalb eines Reports nebeneinander anordnen, jedoch gibt es verschiedene Arten von Reports und diese unterscheiden sich in unterstützten Metriken und Visualisierungen. Die Aggregation verschiedener Daten wird nicht unterstützt.

Vorhersagen bezüglich der zukünftigen Entwicklung werden nicht getroffen.

Ein Dashboard ist nicht vorhanden, allerdings lassen sich Reports mit verschiedenen Visualisierungen erstellen und exportieren.

6.5.5 Ergebnis

Die vorangegangen im Volltext erläuterten Merkmale der verschiedenen Social-Media-Monitoring-Tools werden in folgender Tabelle sowie einem Balkendiagramm zusammengefasst:

	Sawatzki	**Quinty**	**SumAll**	**Hootsuite**
Anzahl Metriken aus den Metadaten:	Grundlegend und geringe Innovation	Breites und innovatives Angebot	Grundlegend	Grundlegend und geringe Innovation
Filterung von Metriken:	Zeitraum (Ausgangsdaten)	Zeitraum (Ausgangsdaten, Zeitachse der Visualisierung)	Zeitraum (Ausgangsdaten)	Zeitraum (Ausgangsdaten)
Innovative/berechnete Daten:	Sawatzki-Ansatz	bsp. Interaktionsrate	Keine	Schlagworte
Metadaten:	2,55	1,59	3,49	2,84
Weitere Quellen:	Nein	Nein	Nein	Nein
Textanalyse/NLP:	Vage	Vereinzelt	Nein	Vage
Resolution:	5,00	4,50	6,00	5,00
Datenbestände lassen sich aggregieren:	Nein	Nein	Nein	Nein
Datenbestände lassen sich vergleichen:	Ja	Ja	Ja	Kaum
Integration:	4,20	3,60	3,60	5,40
Trifft Vorhersagen:	Nein	Nein	Nein	Nein
Genauigkeit der Vorhersagen:	-	-	-	-
Analyse:	6,00	6,00	6,00	6,00
Dashboard:	Nein	Ja, personalisierbar	Nein	Nein
Reporting-Funktionalität:	Nein	Ja	Ja	Ja
Interface:	6,00	1,00	3,50	3,50
Gesamt:	4,48	2,94	4,23	4,29

Tabelle 6.1: Stichpunktartige Zusammenfassung der Benchmark-Ergebnisse, wobei die Kategorien Metadaten 2-fach, Resolution 1-fach, Integration 1,5-fach, Analyse 1-fach und Interface 1,5-fach in die Gesamtwertung eingeflossen sind

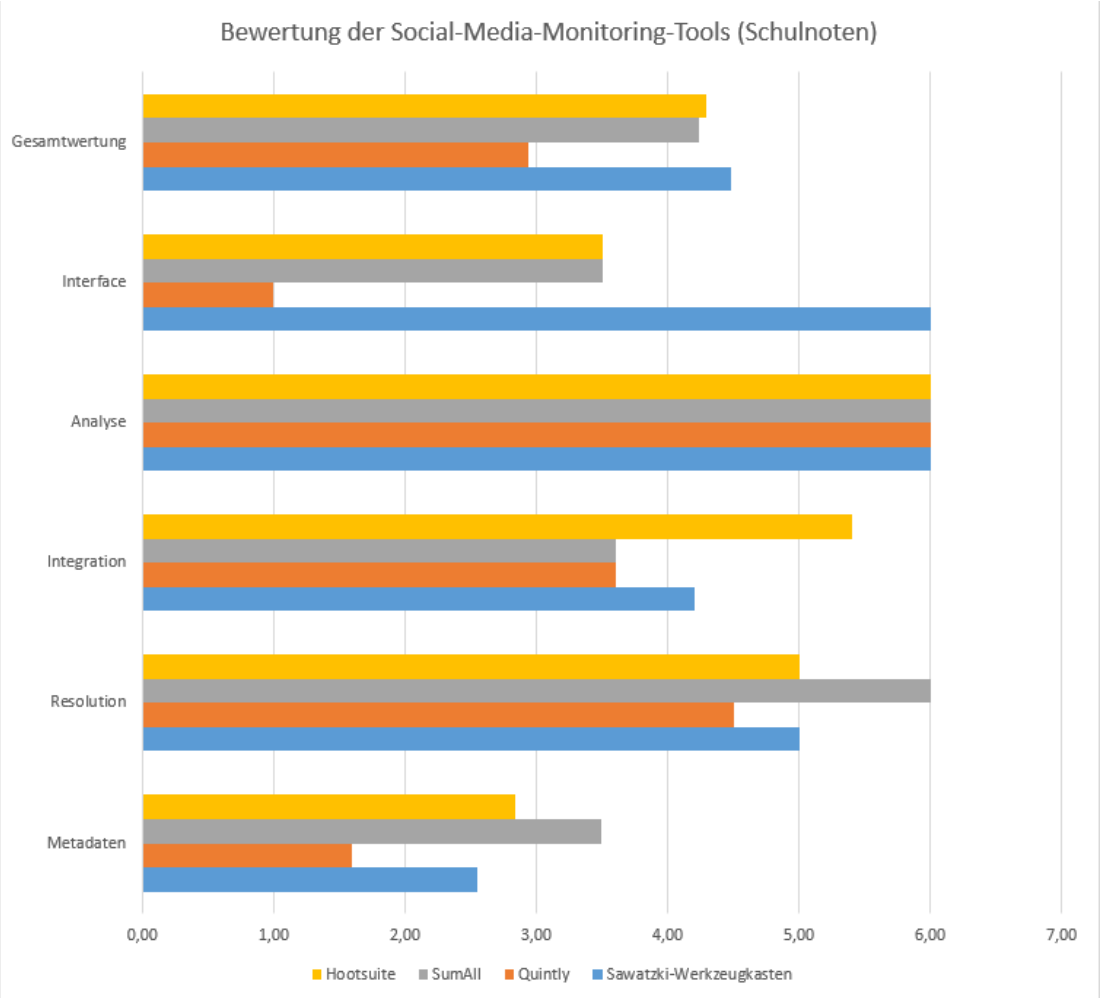

Abbildung 6.5: Die Ergebnisse des Benchmarks als Balkendiagramm dargestellt

Es fällt auf, dass in den wissenschaftlichen Artikeln weitaus mehr Möglichkeiten der Gestaltung eines Social-Media-Monitoring-Tools genannt und beschrieben werden als die bisher am Markt etablierten Systeme unterstützen. In Bezug auf die für diesen Benchmark genutzten Bewertungskriterien schneidet die STB mit einer Note von 4,48 nur unwesentlich schlechter ab als Hootsuite und SumAll. Lediglich Quintly kann sich aufgrund einiger exklusiver Features etwas von der Konkurrenz abheben. Allerdings hat Quintly mit einer Note von 2,94 ebenfalls ein hohes Verbesserungspotenzial. Seth Grimmes erwähnt zwar in seinem Kommentar(siehe Seth Grimes), dass er von einem Social-Media-Monitoring-Tool nicht erwartet, dass dieses alle genannten Aspekte in Gänze unterstützt, jedoch sollte es die wichtigen Aspekte möglichst in Perfektion umsetzen. In dem Kapitel Textmining für Social-Media-Monitoring ist beispielsweise die Bedeutsamkeit der Betrachtung externer Quellen herausgekommen. Diese wird jedoch von keinem der hier betrachteten Tools weiter thematisiert. *SumAll* beschränkt sich bei den angebotenen Visualisierungs-Metriken beispielsweise lediglich auf die vom sozialen Netzwerk ohnehin bereitgestellten Informationen. Dabei sollte es gerade der Sinn von einem Social-Media-Monitoring-Tool sein, über den Tellerrand zu schauen. So lassen sich Informationen sammeln, die exklusiv und wertvoll sind. Insofern ist es ebenfalls verwunderlich, dass sich kein einziges Tool mit dem Thema der Vorhersage beschäftigt.

6.6 Fazit der Sawatzki-Toolbox

Rückblickend auf die eingangs formulierten Anforderungen an die STB, lässt sich sagen, dass die meisten Aspekte im vollem Umfang realisiert wurden. In Bezug auf die Anforderungen an ein Social-Media-Monitoring-Tool sollte der Fokus auf dem Kriterium der *Metadaten* liegen. Im Benchmark wurde festgestellt, dass die STB mit einer Note von 2,55 immerhin den zweitbesten Wert der vier getesteten Social-Media-Monitoring-Tools erreicht, wenngleich noch Verbesserungspotenzial vorhanden ist. Hinsichtlich der Anforderungsliste muss lediglich bei der Anforderung *A1* eine Einschränkung gemacht werden. Es wird gefordert, dass die Eingabequelle keinen Einfluss auf die Funktionalität hat. Dies ist bei der entworfenen und umgesetzten Architektur auch tatsächlich der Fall. Allerdings wird bisher lediglich mit dem sozialen Netzwerk Twitter gearbeitet. Damit die STB auch in Produktion mehrere soziale Medien als Eingabequelle unterstützt, sind noch Anpassungen von Nöten, die bisher

nicht umgesetzt, jedoch im folgenden Abschnitt Möglichkeiten und zukünftige Entwicklung beschrieben werden. Ansonsten erfüllt die aktuelle Version des STB[7] alle Anforderungen.

6.7 Möglichkeiten und zukünftige Entwicklung

Für und neben den bisher vorgestellten, bereits implementierten Funktionen, sind einige Verbesserungen für die Zukunft der STB denkbar. Zum einen wäre es erstrebenswert, die Filterfunktion der Datenstream-Komponente auf einen bestimmten Zeitpunkt festzulegen. Bisher ist es so implementiert, dass der Filter zur Queue der Datenstream-Komponente hinzugefügt wird und auch in jedem Fall abgearbeitet wird, jedoch ist besonders aufgrund der Schnelllebigkeit der sozialen Medien der Zeitraum entscheidend für die gesammelten Ergebnisse. Wenn der Nutzer beispielsweise ein bestimmtes Event wie den Audi Cup 2015 verfolgen möchte, so wird er vorwiegend daran interessiert sein, Daten aus den sozialen Medien zu erhalten, die während des Events generiert wurden. Wenn die Datenstream-Komponente den entsprechenden Filter aufgrund einer ungünstigen Position in der Queue jedoch erst in der darauffolgenden Nacht des Events auswertet, könnte es passieren, dass bereits keine Daten oder für den Nutzer irrelevante Daten extrahiert werden.

Zum anderen wäre es wünschenswert, neben Twitter weitere soziale Netzwerke anzubinden. Dabei würde bestenfalls eine Lösung angestrebt werden, die es im Sinne von den Anforderungen an ein Social-Media-Monitoring-Tool erlaubt, nicht nur eine einzige Quelle auszuwählen, sondern auch mehrere soziale Medien (gleichzeitig) nach bestimmten Filtern zu durchsuchen und die Ergebnisse festzuschreiben. Dazu wäre jeweils ein Adapter notwendig, der zwischen der Datenstream-Komponente und der jeweiligen Quelle übersetzt. Zudem müsste zwischen der Datenstream-Komponente und der Datenbank eine weitere Abstraktion geschaffen werden, sodass die Daten unabhängig von ihrer Herkunft in ein genormtes Format übersetzt werden.

Im Abschnitt Fachliche Architektur wurde bereits von der Offenheit der STB gesprochen. Diese soll dem einzelnen Nutzer ermöglichen über den Tellerrand zu gucken, indem er auch auf die von anderen Nutzern generierten Visualisierungen und Datensätze zugreifen kann. Mit steigenden Nutzerzahlen würde ohne eine gute Filterfunktion jedoch die Übersichtlichkeit enorm leiden. Daher wäre eine weitere zukünftige Entwicklungsmöglichkeit, die Datensätze in bestimmte Kategorien wie Sport, Politik, Deutschland etc. einzusortieren. Dieser Gedanke

[7] http://maskenball.ddns.net:3000

ließe sich sogar soweit ausführen, dass aus der STB eine Web-3.0-Anwendung im Sinne des semantischen Webs wird.

7 Fazit und Ausblick

In Rückbezug auf die eingangs formulierten Ziele lässt sich ein Teilerfolg festmachen. Die STB kann in einigen Bereichen der Anforderungen an ein Social-Media-Monitoring-Tool überzeugen, jedoch ist noch viel ungenutztes Potenzial vorhanden. Überraschenderweise liegt die STB dennoch nicht weit abgeschlagen hinter dem im Benchmark für Social-Media-Monitoring-Tools festgelegtem Testfeld. Besonders die Aspekte Datamining und statistische Vorhersagen wurden von keinem Tool in dem Testfeld abgedeckt. Dafür boten fast alle Tools unterschiedlichste Visualisierungen, die je nach ihren Stärken oder Schwächen eingesetzt wurden. Diese wurden im Kapitel Visualisierung ausführlich geschildert.

Die Aussagekraft eines Social-Media-Monitoring-Tools ist im Hinblick auf die Ergebnisse dieser Bachelorarbeit zwiespältig zu sehen. Das Potenzial ist durchaus vorhanden, allerdings gibt es viele Randbedingungen die auf dem Weg von der Datenakquirierung bis hin zur aussagekräftigen Visualisierung beachtet werden müssen. Besonders in dem Teilbereich der anspruchsvolleren Textanalyse in Form von Stimmungserfassung, die im Kapitel Textmining für Social-Media-Monitoring anhand einer Umfrage betrachtet wurde, stach diese Problematik heraus. Die Semantik eines Textes ist nicht immer eindeutig auszumachen. Eine große Chance des Social-Media-Monitoring könnte in der Entwicklung des semantischen Webs liegen, sodass sich bestimmte Zusammenhänge leichter durch einen Computer erfassen lassen. Zudem ist zu beachten, dass die Bereiche Big data sowie soziale Medien in der Historie der Informatik vergleichsweise jung sind.

Literaturverzeichnis

[Arizona State University 23.06.2000] Arizona State University: *3_38.gif (397×412)*. 23.06.2000. – URL http://www.public.asu.edu/~gelderen/314text/images/3_38.gif. – Zugriffsdatum: 08.09.2015

[Aßmann und Pleil] Assmann, Stefanie ; Pleil, Thomas: Social Media Monitoring: Grundlagen und Zielsetzungen, S. 585–604

[Bassler 2010] Bassler, Anna: *Reihe*. Bd. 13: *Die Visualisierung von Daten im Controlling: Univ. der Bundeswehr, Diss.–München, 2010*. 1. Aufl. Lohmar : Eul, 2010. – ISBN 9783899369397

[Bernhard Steimel | Christian Halemba | Tanya Dimitrova] Bernhard Steimel | Christian Halemba | Tanya Dimitrova: Praxisleitfaden – Social Media Monitoring.

[Ceyp und Scupin 2013] Ceyp, Michael ; Scupin, Juhn-Petter: *Erfolgreiches Social Media Marketing: Konzepte, Massnahmen und Praxisbeispiele*. Wiesbaden : Springer Fachmedien Wiesbaden and Imprint: Springer Gabler, 2013 (SpringerLink : Bücher). – URL http%3A//www.worldcat.org/oclc/826894429

[Cyganski und Hass 2011] Cyganski, Petra ; Hass, BertholdH.: Potenziale sozialer Netzwerke für Unternehmen. In: Walsh, Gianfranco (Hrsg.) ; Hass, Berthold H. (Hrsg.) ; Kilian, Thomas (Hrsg.): *Web 2.0*. Springer Berlin Heidelberg, 2011, S. 81–96. – URL http://dx.doi.org/10.1007/978-3-642-13787-7_6. – ISBN 978-3-642-13786-0

[Dalal] Dalal, Mita K.: Automatic Text Classification: A Technical Review.

[eagereyes.org 08.09.2015] eagereyes.org: *Data: Continuous vs. Categorical*. 08.09.2015. – URL https://eagereyes.org/basics/

data-continuous-vs-categorical. – Zugriffsdatum: 08.09.2015

[Elgün und Karla 2013] Elgün, Levent ; Karla, Jürgen: Social Media Monitoring: Chancen und Risiken. In: *Controlling & Management Review* 57 (2013), Nr. 1, S. 50–57. – URL http://dx.doi.org/10.1365/s12176-013-0680-y. – ISSN 2195-8262

[Facebook] Facebook: *Visualizing Friendships.* – URL https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919. – Zugriffsdatum: 07.05.2015

[Friendly] Friendly, Michael: Milestones in the history of thematic cartography, statistical graphics, and data visualization.

[Gaffney und Puschmann 2014] Gaffney, Devin ; Puschmann, Cornelius: Data Collection on Twitter. In: *Twitter and society.* New York, NY [u.a.] : Lang, 2014, S. 55–68. – ISBN 978-1-4331-2169-2

[Hippner und Rentzmann 2006] Hippner, Hajo ; Rentzmann, René: Text Mining. In: *Informatik-Spektrum* 29 (2006), Nr. 4, S. 287–290. – ISSN 0170-6012

[Kohlhammer u. a. 2013] Kohlhammer, Jörn ; Proff, Dirk U. ; Wiener, Andreas: *Visual Business Analytics: Effektiver Zugang zu Daten und Informationen.* 1. Aufl. Heidelberg : dpunkt-Verl, 2013 (Edition TDWI). – ISBN 978-3-86490-044-0

[König u. a.] König, Christian ; Stahl, Matthias ; Wiegang, Erich: *Soziale medien: Gegenstand und Instrument der Forschung.* URL http%3A//www.worldcat.org/oclc/877103944 (Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute)

[Michelis und Schildhauer 2012] Michelis, Daniel (Hrsg.) ; Schildhauer, Thomas (Hrsg.): *Social-Media-Handbuch: Theorien, Methoden, Modelle und Praxis.* 2., aktualisierte und erw. Aufl. Baden-Baden : Nomos, 2012. – ISBN 3832971211

[Nicole Perlroth and Michael D. Shear] Nicole Perlroth and Michael D. Shear: *In Hacking, A.P. Twitter Feed Sends False Report of Explosions.* – URL http://thecaucus.blogs.nytimes.com/2013/04/23/

hacked-a-p-twitter-feed-sends-erroneous-message-about-explosions-at-?_r=0. – Zugriffsdatum: 27.05.2015

[Reyes und Rosso 2014] Reyes, Antonio ; Rosso, Paolo: On the difficulty of automatically detecting irony: beyond a simple case of negation. In: *Knowledge and Information Systems* 40 (2014), Nr. 3, S. 595–614. – ISSN 0219-1377

[Semiocast] Semiocast: *Semiocast — Twitter reaches half a billion accounts — More than 140 millions in the U.S.*. – URL http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. – Zugriffsdatum: 04.05.2015

[Seth Grimes] Seth Grimes: *What I Look For In A Social Analysis Tool - InformationWeek*. – URL http://www.informationweek.com/software/information-management/what-i-look-for-in-a-social-analysis-tool/d/d-id/1096654?page_number=1. – Zugriffsdatum: 09.08.2015

[statista.com a] statista.com: *Größte Social Networks nach Anzahl der monatlich aktiven Nutzer (MAU) im März 2015*. – URL http://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/

[statista.com b] statista.com: *Weltweit größte Social Networks nach User-Anzahl 2015 | Statistik*. – URL http://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/. – Zugriffsdatum: 07.05.2015

[tylervigen.com] tylervigen.com ; tylervigen.com (Hrsg.): *US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation*. – URL http://tylervigen.com/view_correlation?id=1597. – Zugriffsdatum: 07.05.2015

[Wayne Eckerson and Mark Hammond] Wayne Eckerson and Mark Hammond: Visual Reporting and Analysis: Seeing Is Knowing.

[Wikipedia-Autoren 01.05.2015] Wikipedia-Autoren ; Wikipedia (Hrsg.): *Big Data*. 01.05.2015. – URL http://de.wikipedia.org/w/index.php?oldid=141659332. – Zugriffsdatum: 04.05.2015

[Wikipedia-Autoren 27.07.2015] Wikipedia-Autoren ; Wikipedia (Hrsg.): *Text Mining*. 27.07.2015. – URL https://de.wikipedia.org/w/index.php?oldid=143107000. – Zugriffsdatum: 29.08.2015

[Wikipedia contributors 19.08.2015] Wikipedia contributors ; Wikipedia (Hrsg.): *Decision-making - Wikipedia, the free encyclopedia*. 19.08.2015. – URL https://en.wikipedia.org/w/index.php?oldid=676824664. – Zugriffsdatum: 29.08.2015

[Yau] Yau, Nathan: *Find new beers to drink*. – URL http://flowingdata.com/2014/03/05/find-new-beers-to-drink/. – Zugriffsdatum: 08.09.2015

Anhang A

1 Funktion und Mockup der STB

Die folgende Bilderreihe illustriert die Funktionen der STB anhand einer Reihe von Mockups:

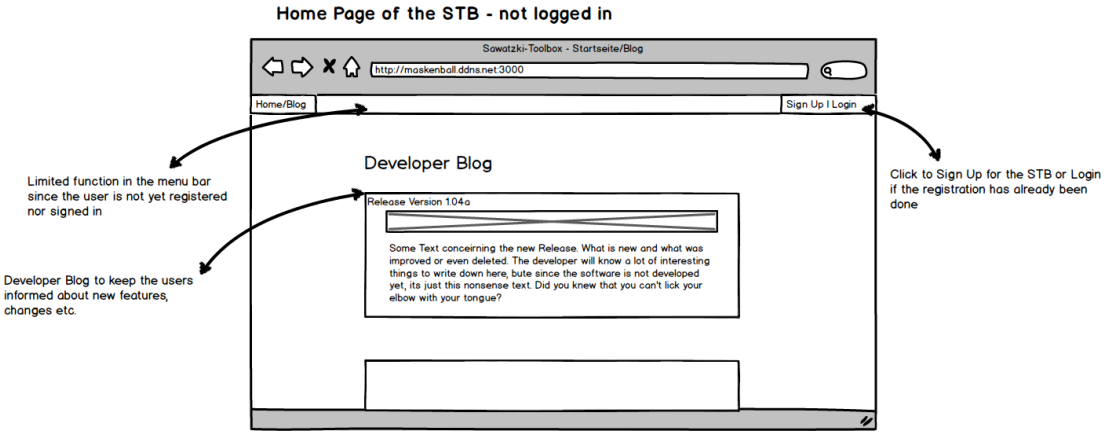

Abbildung 1: Ohne sich einzuloggen kann der Nutzer lediglich den Entwickler-Blog lesen

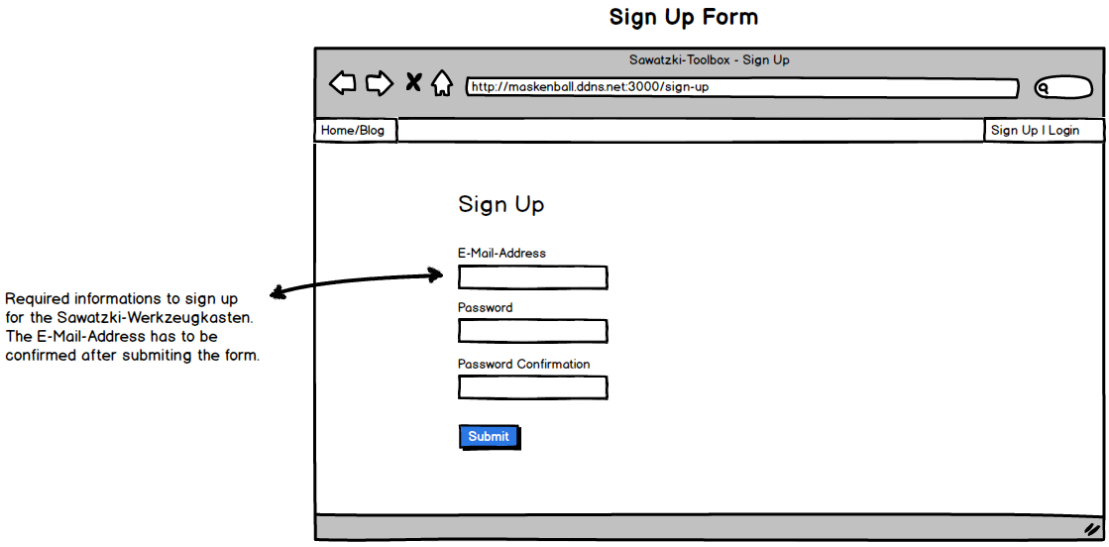

Abbildung 2: Das Registrierungs-Formular erfordert lediglich sporadische User-Informationen

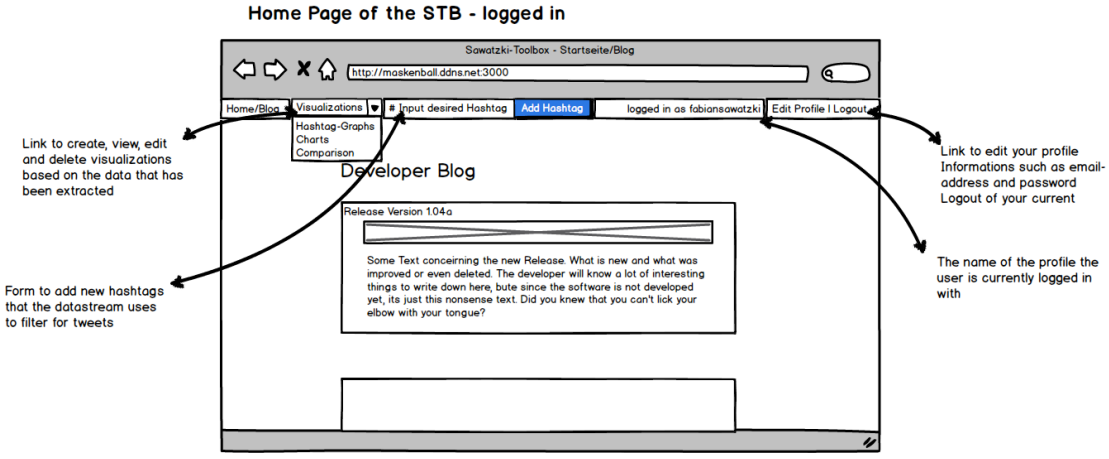

Abbildung 3: Erst nach erfolgreicher Registrierung und mit eingeloggtem Profil offenbart die STB seine Funktionalität

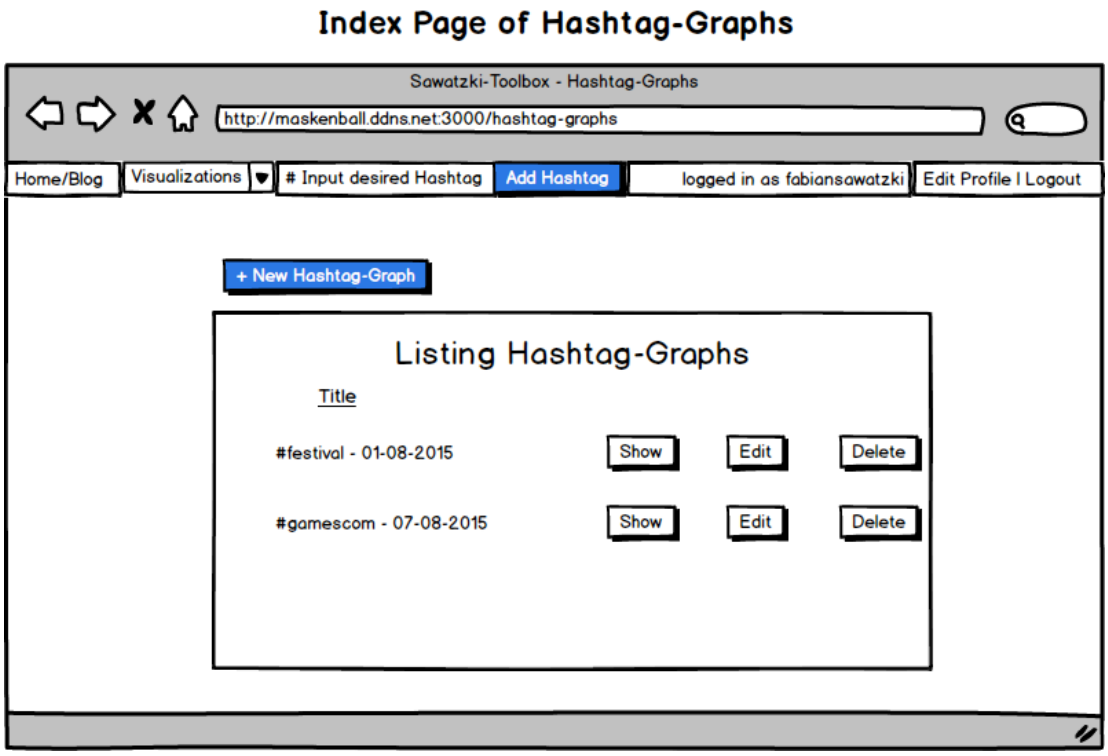

Abbildung 4: Beispielhaft die Übersichtsseite aller Hashtag-Graphen nach dem Sawatzki-Ansatz. Ähnliche Darstellungen sind für Diagramme und Vergleiche vorhanden. Der User kann nur Hashtag-Graphen bearbeiten, die er selbst erstellt hat, alle anderen lediglich ansehen.

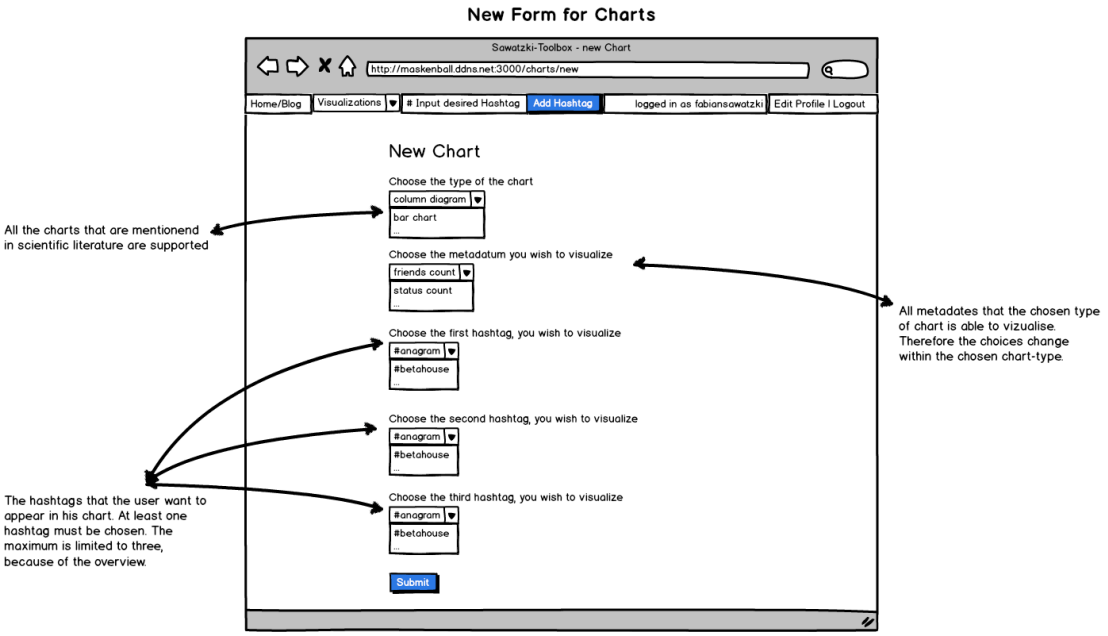

Abbildung 5: Das Formular für die Erstellung eines Diagrammes, ähnliche Formulare gibt es auch für die Hashtag-Graphen sowie die Vergleiche.

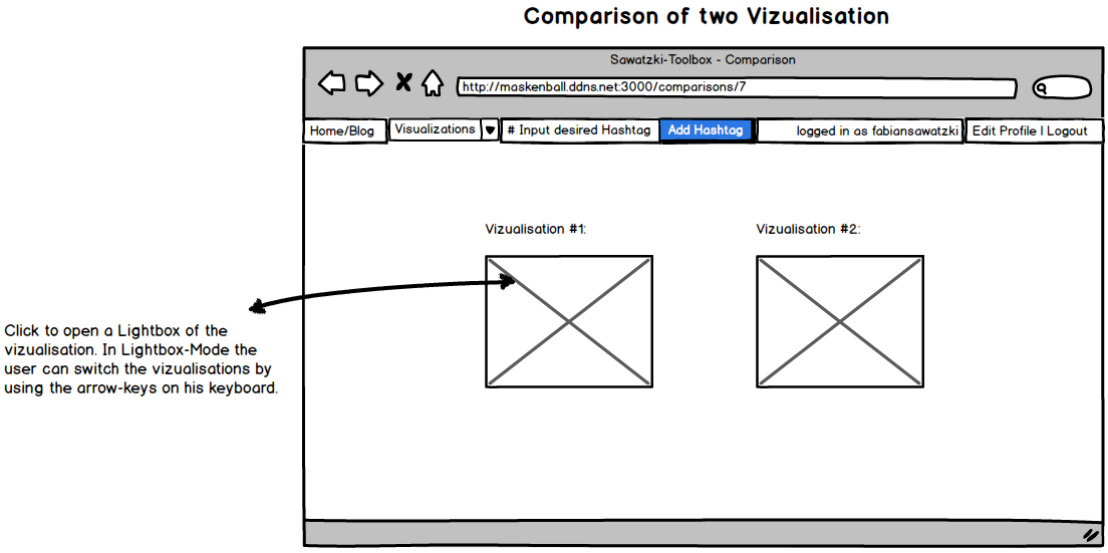

Abbildung 6: Ein bestimmter Vergleich zweier Visualisierungen, die zuvor mit der STB erstellt worden sind.

2 Ergebnisse der Textmining-Umfrage

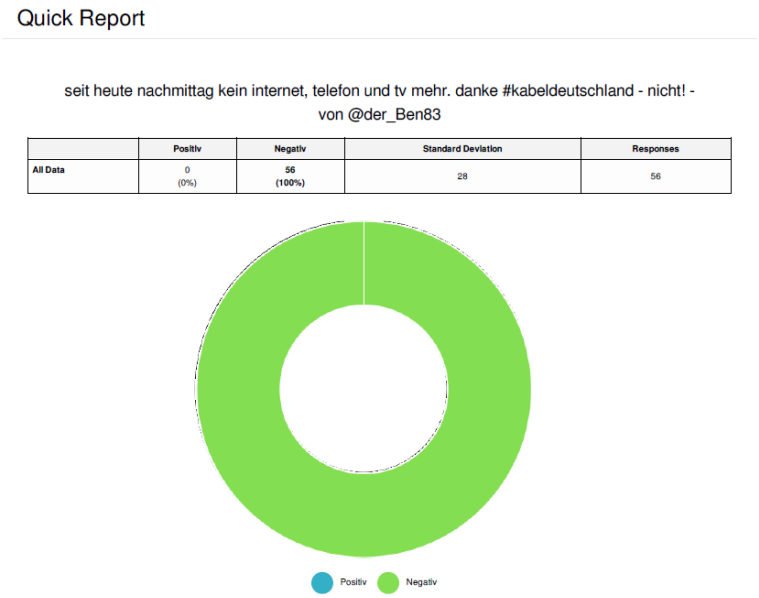

Abbildung 7: Die Verteilung der Stimmen bezüglich des Tweets: *seit heute nachmittag kein internet, telefon und tv mehr. danke #kabeldeutschland - nicht! - von @der_Ben83*

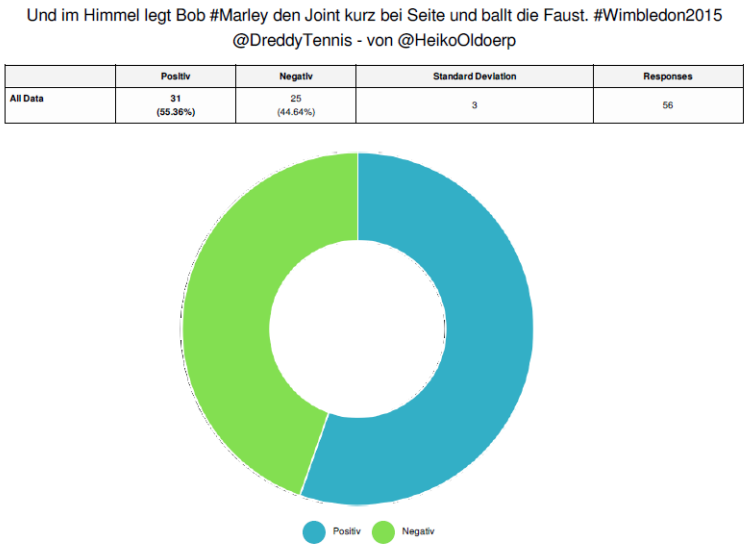

Abbildung 8: Die Verteilung der Stimmen bezüglich des Tweets: *Und im Himmel legt Bob #Marley den Joint kurz bei Seite und ballt die Faust. #Wimbledon2015 @DreddyTennis - von @HeikoOldoerp*

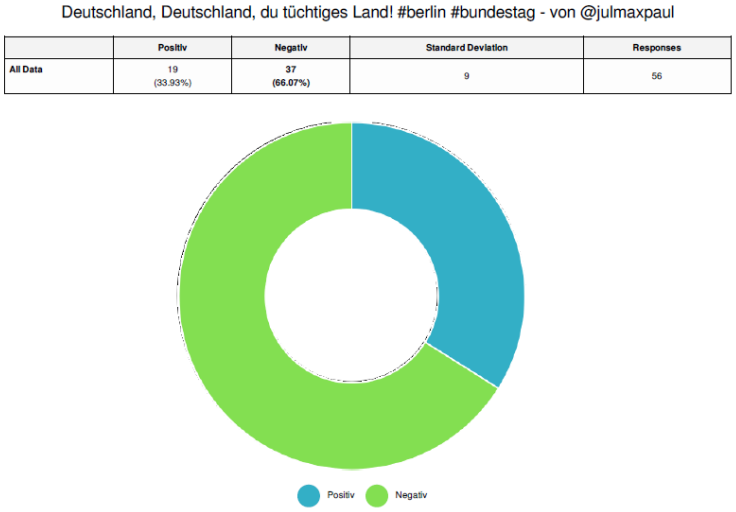

Abbildung 9: Die Verteilung der Stimmen bezüglich des Tweets: *Deutschland, Deutschland, du tüchtiges Land! #berlin #bundestag - von @julmaxpaul*

#Hoax = #Wasser trinken hilft gegen Kopfschmerzen - von @MartinKaindel

	Positiv	Negativ	Standard Deviation	Responses
All Data	21 (37.5%)	**35 (62.5%)**	7	56

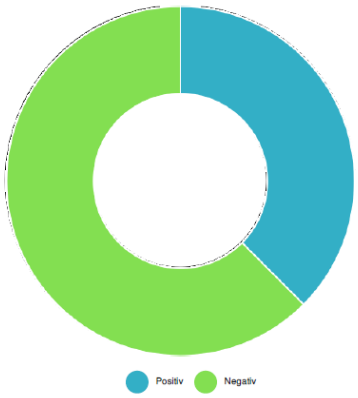

Abbildung 10: Die Verteilung der Stimmen bezüglich des Tweets: *#Hoax = #Wasser trinken hilft gegen Kopfschmerzen - von @MartinKaindel*

Das Klacken der Kaffeemaschine wenn sie fertig ist ist das beste Geräusch der Welt. - von @extraktiv

	Positiv	Negativ	Standard Deviation	Responses
All Data	**53 (94.64%)**	3 (5.36%)	25	56

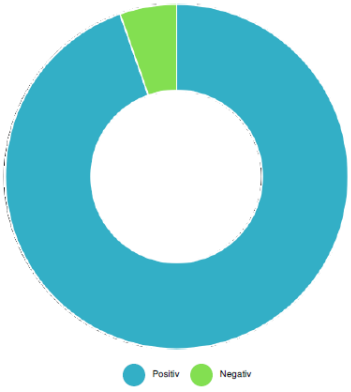

Abbildung 11: Die Verteilung der Stimmen bezüglich des Tweets: *Das Klacken der Kaffeemaschine wenn sie fertig ist ist das beste Geräusch der Welt. - von @extraktiv*

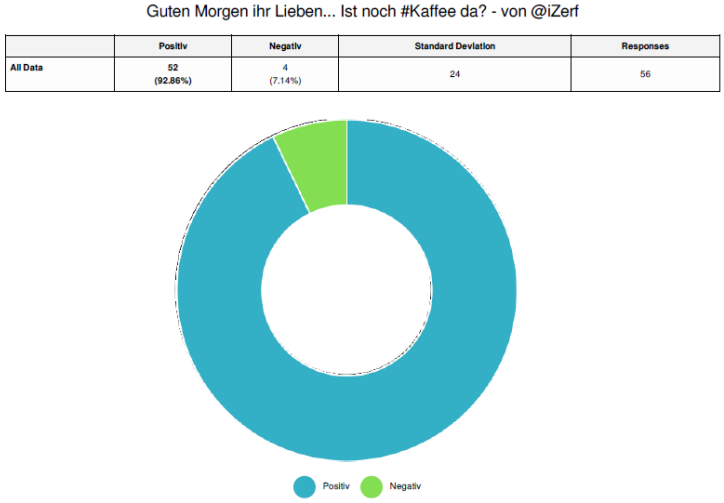

Abbildung 12: Die Verteilung der Stimmen bezüglich des Tweets: *Guten Morgen ihr Lieben... Ist noch #Kaffee da? - von @iZerf*

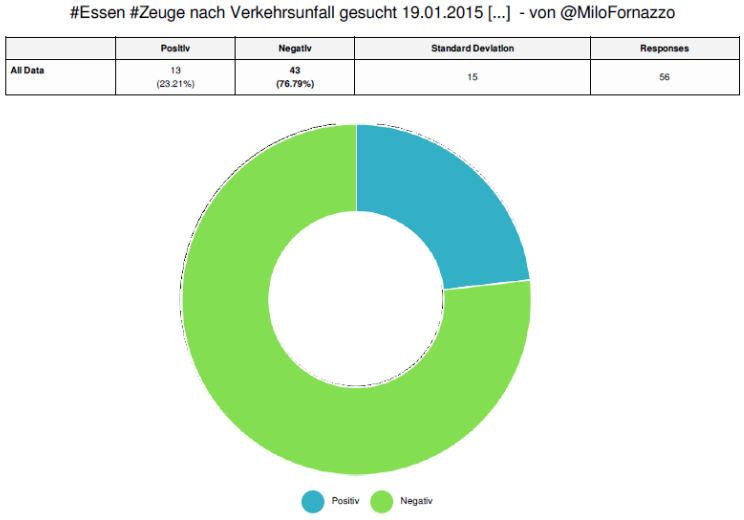

Abbildung 13: Die Verteilung der Stimmen bezüglich des Tweets: *#Essen #Zeuge nach Verkehrsunfall gesucht 19.01.2015 [...] - von @MiloFornazzo*

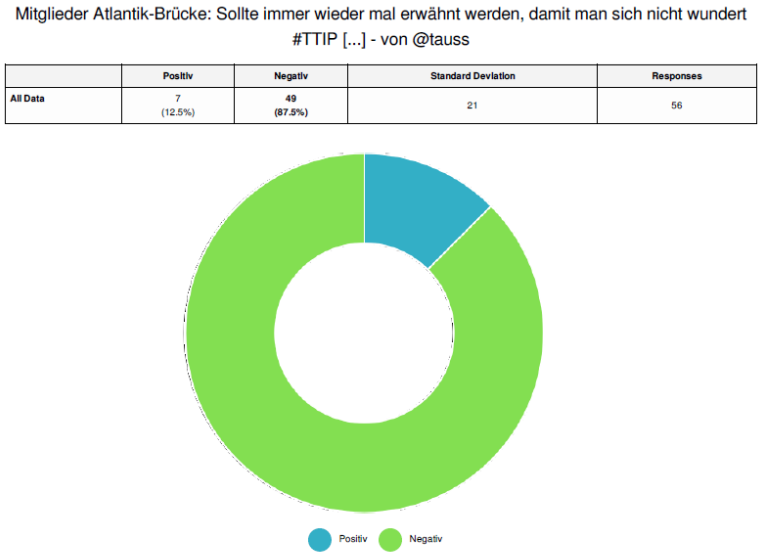

Abbildung 14: Die Verteilung der Stimmen bezüglich des Tweets: *Mitglieder Atlantik-Brücke: Sollte immer wieder mal erwähnt werden, damit man sich nicht wundert #TTIP [...] - von @tauss*

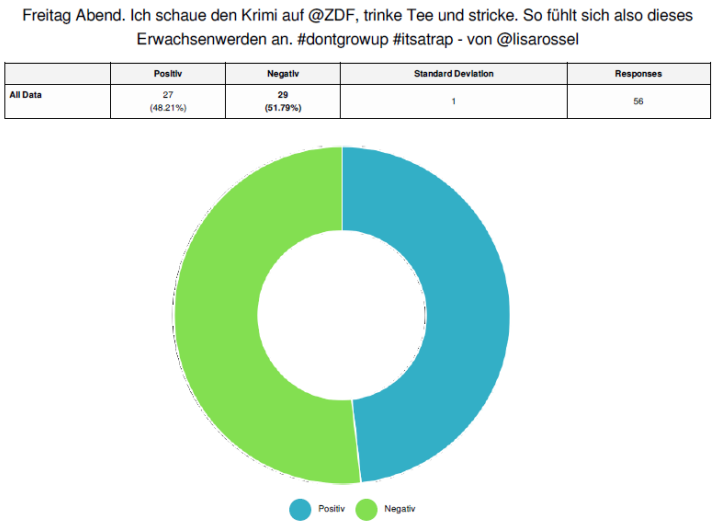

Abbildung 15: Die Verteilung der Stimmen bezüglich des Tweets: *Freitag Abend. Ich schaue den Krimi auf @ZDF, trinke Tee und stricke. So fühlt sich also dieses Erwachsenwerden an. #dontgrowup #itsatrap - von @lisarossel*

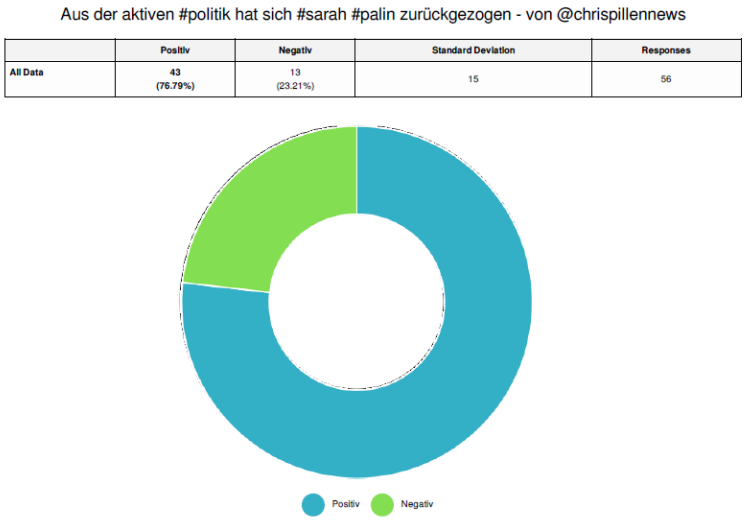

Abbildung 16: Die Verteilung der Stimmen bezüglich des Tweets: *Aus der aktiven #politik hat sich #sarah #palin zurückgezogen - von @chrispillennews*

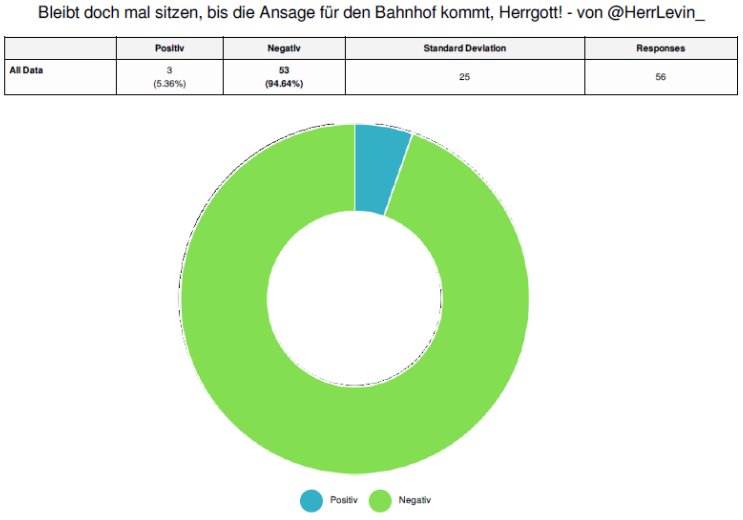

Abbildung 17: Die Verteilung der Stimmen bezüglich des Tweets: *Bleibt doch mal sitzen, bis die Ansage für den Bahnhof kommt, Herrgott! - von @HerrLevin_*

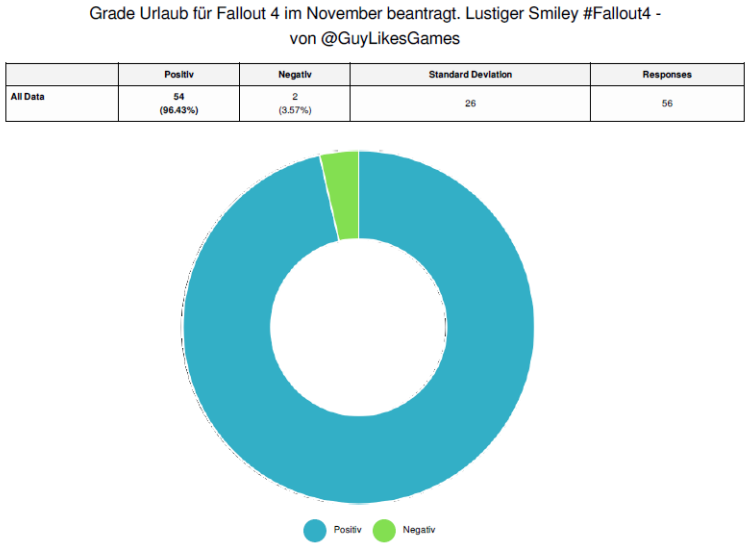

Abbildung 18: Die Verteilung der Stimmen bezüglich des Tweets: *Grade Urlaub für Fallout 4 im November beantragt. Lustiger Smiley #Fallout4 - von @GuyLikesGames*

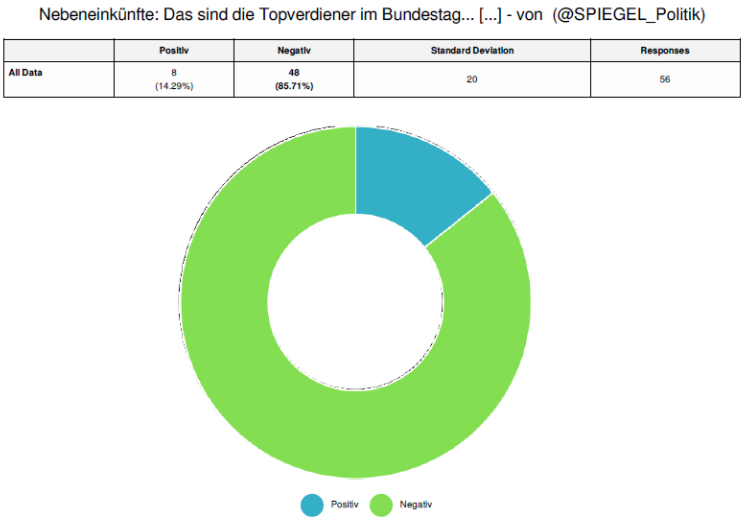

Abbildung 19: Die Verteilung der Stimmen bezüglich des Tweets: *Nebeneinkünfte: Das sind die Topverdiener im Bundestag... [...] - von (@SPIEGEL_Politik)*

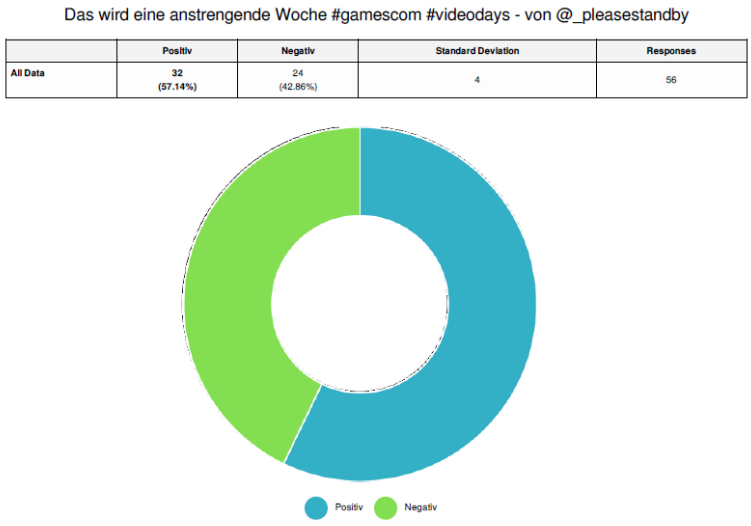

Abbildung 20: Die Verteilung der Stimmen bezüglich des Tweets: *Das wird eine anstrengende Woche #gamescom #videodays - von @_pleasestandby*

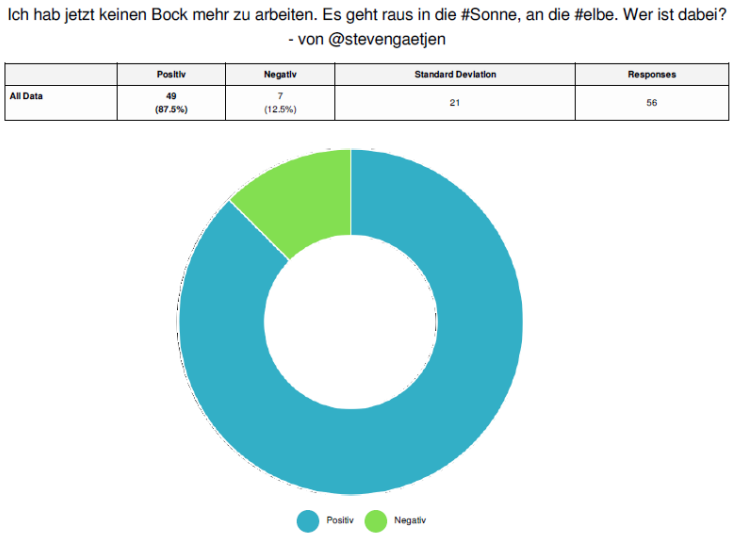

Abbildung 21: Die Verteilung der Stimmen bezüglich des Tweets: *Ich hab jetzt keinen Bock mehr zu arbeiten. Es geht raus in die #Sonne, an die #elbe. Wer ist dabei? - von @stevengaetjen*

Anhang B

1 Inhalt der CD-ROM

Der Inhalt der CD-ROM ist in durch folgende Verzeichnisstruktur ersichtlich:

Thesis: Die komplette Arbeit als PDF-Dokument

Anwendung: Der Quellcode der entwickelten Anwendung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 15. September 2015 Fabian Sawatzki