Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Interaktive visuelle Korrelation von Ereignissen in Sozialen Medien und Nachrichten

Fabian Danner

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dr. Steffen Koch, Franziska Huth, M.Sc., Dr. Harald Bosch, Dr. Dennis Thom
Beginn am:	31. Januar 2020
Beendet am:	25. September 2020

Kurzfassung

Im Laufe der letzten Jahre wurden soziale Netzwerke wie Twitter immer beliebter und stärker genutzt. Durch die zunehmenden Nutzerzahlen, steigt auch die Masse an verbreiteten Nachrichten. Twitter verzeichnet täglich mehr als 500 Millionen Nachrichten[1]. Richtig aufbereitet und gefiltert, sind diese eine sehr ergiebige Quelle für Ereignisse des aktuellen Zeitgeschehens. Firmen wie ScatterBlogs bereiten diese Rohdaten so auf, dass die enthaltenen Informationen genutzt werden können. Im Twitter-Datenstrom befinden sich nicht nur relevante Informationen. Um eine Filterung der Ereignisse zu erhalten, wird die Wichtigkeit der einzelnen Ereignisse beurteilt. Diese Bewertung entscheidet, ob ein Ereignis gespeichert wird. Festzustellen, ob die errechnete Wichtigkeit plausibel erscheint, ist ein hoher manueller Aufwand. Diese Arbeit stellt ein Bewertungsmaß vor, welches die errechnete Wichtigkeit in Bezug zu der durch ein Event entstandenen Medienaufmerksamkeit stellt. Letztere wird durch die Anzahl gefundener Zeitungsartikel bestimmt. Das Bewertungsmaß wird durch eine interaktive Visualisierung in Form einer Web App vorgestellt. Weiterführend bietet die Visualisierung dem Nutzer durch eine interaktive Treemap und einer nach Themen sortierten Scatterplotmatrix die Möglichkeit, den Datensatz explorativ weiter zu analysieren. Ziel der Visualisierung ist es, Informationen zu den Events so aufzubereiten, dass Verbesserungsmöglichkeiten für die Wichtigkeitsbewertung der Events gefunden werden können. Die Beurteilung von Experten zeigt, dass der Vergleich von Medienecho und errechneter Wichtigkeit, unter bestimmten Voraussetzungen, ein beschränkt aussagekräftiges Bewertungsmaß darstellt. Das Bewertungsmaß und die Visualisierung bilden eine gut geeignete Basis für weitere Forschungen.

Abstract

Over the past few years, social networks like Twitter have become more popular and used. Due to the increasing number of users, the amount of news spread also increased. Twitter records more than 500 million messages every day[1]. Properly prepared and filtered, these are a very rich source for current events. Companies like ScatterBlogs prepare this raw data in such a way that the information it contains can be used. The Twitter data stream is not just about relevant information. In order to obtain a filtering of the events, the importance of the individual events is assessed. This assessment leads to the decision whether an event candidate is saved as an actual event. Determining whether the calculated importance appears plausible is a lot of manual effort. This thesis presents an evaluation measure that places the calculated importance in relation to the media attention generated by an event. This is determined by the number of newspaper articles found. The evaluation measure is presented as an interactive visualization, packed as a web app. In addition, the visualization offered the user the opportunity to further analyze the data set through an interactive tree map and a scatter plot matrix sorted by topic. The aim of the visualization is to prepare information about the events in such a way that opportunities for improvement can be found for calculating the importance of the events. The assessment by experts shows that the comparison of media response and the calculated importance is a limited meaningful assessment measure under certain conditions. The evaluation measure and the visualization are well suited as a basis for further research.

[1] https://www.internetlivestats.com/twitter-statistics/

Inhaltsverzeichnis

Abbildungsverzeichnis

Verzeichnis der Algorithmen

Abkürzungsverzeichnis

AF Abweichungsfaktor. 15, 16, 17, 18, 19, 22, 24, 25, 26, 31, 33, 40, 41, 46

1 Einleitung

1.1 Motivation der Arbeit

Heutzutage werden wichtige Ereignisse des Zeitgeschehens vermehrt durch soziale Netzwerke geteilt. Diese schnelle und einfache Art Nachrichten zu verbreiten, ermöglicht es aktuelle Ereignisse fast ohne zeitliche Verzögerung zu veröffentlichen. Daher wird diesen Medien eine stetig wachsende Bedeutung zugeschrieben. Ein Beispiel dieser Plattformen ist Twitter. Twitter hat aktuell (Stand Mitte 2020) monatlich 330 Millionen aktive Nutzer[1]. Das macht diese Plattform zu einer der größten und erfolgreichsten ihrer Branche. Benutzer[2] von Twitter können sogenannte Tweets posten. Tweets sind kurze Textnachrichten mit maximal 280 Zeichen. Das führt dazu, dass in diesen Nachrichten viele Inhalte gekürzt dargestellt werden oder im Allgemeinen auf Abkürzungen zurückgegriffen wird. Dies erschwert die maschinelle Kontexterfassung der Tweets durch Text-Mining-Verfahren. Zusätzlich können Tweets mit Schlagworten versehen werden. Bei gut gewählten Schlagworten, lässt sich vom Nutzer, als auch maschinell sehr gut und zeitlich effizient zuordnen, um was es in dem Tweet geht. Diese Schlagworte werden mit einem voranstehenden „#“ gekennzeichnet und als *Hashtags* bezeichnet. Tweets erscheinen zu wichtigen Ereignissen sehr zeitnah, wohingegen herkömmliche Nachrichtenmedien trotz eigener Internetauftritte erst deutlich später reagieren können. Um also frühzeitig wichtige Ereignisse weltweit erkennen zu können, lohnt es sich, die Tweets auszuwerten.

Twitter bietet eine Schnittstelle an, mit der es möglich ist, die Tweets samt ihrer Metadaten abzurufen. Danach kann eine Analyse der Tweets angeschlossen werden. Aus dieser Analyse werden die benötigten Informationen gewonnen, mit deren Hilfe eine Beurteilung der Tweets vorgenommen werden kann, ob ein Tweet wichtige Informationen zum Weltgeschehen enthält.

1.2 Eventerkennung durch ScatterBlogs

Diese Arbeit entstand in enger Zusammenarbeit mit der Firma ScatterBlogs. Diese macht es sich zur Aufgabe, aus Twitternachrichten, Ereignisse zu extrahieren und diese für Analysten, Entscheidungsträger, Journalisten und für die Unterstützung des Zivilschutzes, zur Verfügung zu stellen. Der Aufbau der ScatterBlogs-Plattform wird von Thom et al. [Tho15] genau beschrieben. Weitere Informationen zur Eventerkennung in Sozial Media Diensten wie Twitter sind in dem von ScatterBlogs veröffentlichen Paper zu finden [BTH+13]. ScatterBlogs analysiert in Echtzeit den

[1] https://de.statista.com/statistik/daten/studie/232401/umfrage/monatlich-aktive-nutzer-von-twitter-weltweit-zeitreihe/

[2] Es sind stets Personen männlichen, weiblichen und diversen Geschlechts gleichermaßen gemeint; aus Gründen der einfacheren Lesbarkeit wird im Folgenden nur die männliche Form verwendet.

Twitter-Datenstrom. Dazu werden alle Tweets mittels der von Twitter bereitgestellten Schnittstelle gesammelt. Anschließend wird mit Hilfe von Text-Mining-Verfahren der Kontext der Tweets ermittelt. Dabei wird festgestellt, ob es sich um ein „interessantes“ Ereignis handelt. Interessante Ereignisse werden als Events in der Datenbank gespeichert. Für jedes Event wird die „Wichtigkeit“ vorhergesagt. Diese gibt an, wie wahrscheinlich es ist, dass auch die klassischen Medien über dieses Ereignis berichten werden. Zusätzlich werden weiter alle zu dem Event gehörenden Tweets gesammelt und gespeichert. Aus diesen Tweets werden einige Terme extrahiert, um diese als Suchparameter für die Google News Suche zu verwenden. So wird ermittelt, ob zu dem Event schon Artikel aus den klassischen Medien existieren.

1.3 Ziel der Arbeit und Forschungsfrage

Das Ziel dieser Arbeit ist die Erstellung einer interaktiven Visualisierung des Datensatzes von ScatterBlogs. Dabei wird dargestellt, wie wichtig die einzelnen Events vom Bewertungsalgorithmus eingestuft wurden. Die Visualisierung wird es ermöglichen, Events zu identifizieren, die nicht so bewertet wurden, wie ein Domänenexperte dies getan hätte. Das Tool wird die Suche nach Gründen für die abweichende Bewertung unterstützen. Es wird außerdem ein Maß für die Abweichung der Bewertung vom Erwartungswert vorgeschlagen. Durch dieses Maß soll die Bewertungsfunktion gegen folgende Hypothese getestet werden: Je wichtiger ein Event eingestuft wird, desto mehr Artikel in den klassischen Medien sind zu dem Event erschienen. So sollen mit der prototypartigen Implementierung des entstehenden Konzepts diese beiden Forschungsfragen beantwortet werden.

1. Ist die entwickelte Visualisierung geeignet, statistische Ausreißer effizient zu identifizieren und kann mit entsprechendem Vorwissen über die Bewertungsfunktion eine Hypothese aufgestellt werden, wie die Ausreißer entstanden sind?
2. Ist das vorgeschlagene Maß für die Abweichung vom Erwartungswert dazu geeignet, visuell Aussagen über die Qualität des Bewertungsalgorithmus der Events zu machen?

Diese Arbeit ist in neun Kapitel unterteilt. Ausgehend von dieser Einleitung werden in Kapitel 2 die Grundlagen vorgestellt, die für das Verständnis dieser Arbeit benötigt werden. Kapitel 3 stellt verwandte Arbeiten vor und zeigt auf, weshalb sich diese von der hier vorgestellten Arbeit unterscheiden. Darauf wird in Kapitel 4 das Konzept dieser Arbeit im Detail erläutert. Die aus dem Konzept entstandene Implementierung wird im fünften Kapitel aus technischer Sicht beschrieben. Das prototypisch implementierte Konzept wurde von Domänenexperten getestet und bewertet. Der Aufbau und die Ergebnisse aus dem geführten Experteninterview werden in Kapitel 6 behandelt. Das siebte Kapitel diskutiert diese Ergebnisse. Schließlich fasst Kapitel 8 im Rahmen eines Fazits, die Arbeit, sowie die entstandenen Ergebnisse zusammen. In Kapitel 9 werden mögliche Erweiterungen der Arbeit vorgestellt.

2 Hintergrund

Grundlagen der Arbeit

In diesem Kapitel werden die benötigten Grundlagen aus der Forschung zu den Visualisierungen und der Statistik, sowie die für das Verständnis der Arbeit erforderlichen Begrifflichkeiten erläutert. Dieses Wissen bildet die Basis der Konzeptentwicklung. Daraufhin werden die im Konzept verwendeten Inteaktionstaxonomien und Designprinzipien vorgestellt. Darauf wird auf die Entstehung des Datensatzes und dessen Aufbau eingegangen. Die für die Visualisierung verwendeten Visualisierungsarten Treemap und Scatterplot werden kurz vorgestellt. Abschließend werden die statistischen Grundlagen näher beleuchtet.

2.1 Interaktionstaxonomien und Designprinzipien

Interaktion mit Systemen oder auch die Mensch-Computer-Interaktion im Allgemeinen muss bei der Konzeption einer interaktiven Visualisierung bedacht werden. Es gibt verschiedene Arten, wie ein Mensch mit einem System interagieren kann. Yi et al. stellen sieben Kategorien vor, in die die Interaktionen unterteilt werden können [YKSJ07]. Im folgenden Abschnitt werden die in der Arbeit verwendeten Kategorien kurz erläutert.

2.1.1 Encode

Encode-Interaktionstechniken werden zum Ändern des gesamten Erscheinungsbildes einer Visualisierung genutzt. Diese Änderung kann einen maßgeblichen Einfluss auf die Qualität der Visualisierung haben, da die Interpretation durch den Benutzer stark vom Erscheinungsbild beeinflusst wird. Das Ändern der Visualisierungstechnik stellt eine Interaktionstechnik dieser Kategorie dar, die in vielen Visualisierungswerkzeugen realisiert ist. Das kann unter anderem das Ändern eines Diagrammtyps sein, um so Zusammenhänge erkennen zu können, die vor der Änderung nicht oder nur schlecht erkennbar waren. Eine weitere Interaktionstechnik dieser Kategorie ist das Einfärben von Elementen. Dazu gehört auch das Ändern des Schemas für die farbliche Kodierung. So können Variablen oder Kennzahlen auf einen Farbverlauf abgebildet werden. Eine gängige Anwendung dafür, ist das Einfärben einer topologischen Landkarte, wobei die flachen Zonen durch eine eher grünliche Färbung zu erkennen sind und die höher gelegenen Gebiete eine bräunliche Färbung annehmen. Die Interaktionstechnik „Encode“ wird bei der Bewertung des Algorithmus verwendet.

2.1.2 Abstract / Elaborate

Mit Abstract / Elaborate Interaktionstechniken erhält der Benutzer die Möglichkeit, die Anzahl der angezeigten Details zu manipulieren. Neben dem Anzeigen zusätzlicher Informationen kann so auch das Level der Abstraktion einer Visualisierung beeinflusst werden. Eine verbreitete Technik dies zu erreichen, ist der Tooltip. Durch einen Tooltip kann ein Nutzer zusätzliche Informationen temporär einblenden, in dem er den Cursor auf ein Element legt. Eine weitere Anwendung für diese Kategorie ist das „Zoomen“. Dabei kann sich der Benutzer durch das Ändern des Maßstabs entweder einen Überblick über eine große Anzahl an Daten verschaffen (zoom out) oder weniger Datenpunkte, aber dafür detaillierter betrachten (zoom in). Das Konzept kommt auch beim Einsatz von ausklappbaren Elementen zum Einsatz. Dabei ist auf den ersten Blick eine Überschrift zu sehen. Erst bei einem Klick auf diese Überschrift wird der Inhalt des Feldes ausgeklappt und eine detailliertere Ansicht freigegeben.

2.1.3 Filter

Soll nur eine Teilmenge der Datenpunkte aus dem Datensatz betrachtet werden, so kann der Benutzer durch das Einsetzen von Filter-Interaktionstechniken die betrachteten Daten eingrenzen. Dafür müssen vom Benutzer Kriterien definiert werden, nach denen gefiltert werden soll. Die Visualisierung stellt darauf hin nur noch die Daten dar, welche die Kriterien erfüllen. Die übrigen Elemente werden ausgeblendet oder treten visuell in den Hintergrund. Änderungen, die durch die Anwendung eines Filters gemacht wurden, können jederzeit rückgängig gemacht werden, da nur die Visualisierung und nicht der Datensatz angepasst wird. Eine Anwendung für diese Kategorie sind die „Dynamic Query Controls“. Mit diesen können die in einer Visualisierung dargestellten Werte begrenzt werden. Dies kann durch das Anhaken einer Checkbox oder durch einen definierten Bereich, der durch einen Slider mit zwei veränderlichen Punkten gegeben ist, erfolgen. Darauf hin wird die Darstellung an die Einstellungen angepasst, und es wird nur noch eine Teilmenge der ursprünglichen Daten dargestellt.

2.1.4 Overview first, zoom & filter, then details-on-demand

Neben den Taxonomien gilt es auch, das grundlegende Design passend zum Zweck der Visualisierung zu wählen. Dazu gibt es verschiedene Muster, die befolgt werden können. Das in dieser Arbeit genutzte Muster wird 1996 von Ben Schneiderman [Shn96] beschrieben und nennt sich „Overview first, zoom & filter, then details-on-demand“. Aus diesem Muster lässt sich eine Struktur der Visualisierung ableiten. Der Einstieg in die Visualisierung präsentiert dem Nutzer eine Darstellung, bei der ein Überblick und ein grundsätzliches Verständnis für den vorliegenden Datensatz entsteht. Die dargestellte Menge an Datenpunkten ist dabei hoch, wohin gegen der Detailgrad der Informationen gering ausfällt. Dadurch wird die Übersichtlichkeit der Ansicht deutlich verbessert. Anschließend kann der Benutzer durch Interaktionstechniken wie, „Filter“ oder „Zoom“, die dargestellten Inhalte näher eingrenzen oder explorativ tiefer in den Datensatz eintauchen. Wird ein sehr hoher Detailgrad benötigt, so muss der Benutzer ebenso durch aktive Interaktion die detaillierten Informationen beim System anfordern. Dies kann durch das Bewegen des Cursors auf ein Element geschehen, wodurch

ein Tooltip eingeblendet wird oder durch das Klicken auf eine Schaltfläche, welche den Benutzer zu einer anderen Ebene der Darstellung führt. Dieser Ansatz zoomt nicht auf eine optische sondern viel mehr auf eine inhaltliche Art. Diese Technik wird als „Semantic Zoom“ bezeichnet.

2.2 Datensatz

2.2.1 JavaScript Object Notation

JavaScript Object Notation oder kurz JSON ist ein Datenformat, welches für den Austausch von Daten zwischen Anwendungen und Systemen Verwendung findet. Es ist unabhängig von der verwendeten Programmiersprache und somit standardisiert. Viele Programmiersprachen bieten auch Bibliotheken an, die es ermöglichen Objekte automatisiert als JSON zu serialisieren oder zu deserialisieren. JSON-Dokumente sind aus Schlüssel-Wert-Paaren zusammengesetzt. Der Schlüssel ist immer ein String. Der Wert kann *null*, eine Zahl, eine Zeichenkette, ein Array oder ein Objekt sein. Objekte sind durch geschweifte Klammern eingeschlossen und bestehen ebenfalls aus Schlüssel-Wert-Paaren. So ist es möglich eine komplexe, verschachtelte oder hierarchische Struktur aufzubauen, welche sowohl für Menschen als auch für Maschinen leicht lesbar und gut verständlich ist.

2.2.2 Datensatzbeschreibung

Die mittels der Twitter-API gesammelten Tweets durchlaufen einige Verarbeitungsschritte [CTB+12]. Dabei wird festgestellt, ob die Tweets wichtige Ereignisse repräsentieren. Ist dies der Fall, wird aus den Twitterdaten ein Event erzeugt. Events bestehen immer mindestens aus zwei Tweets. Tweets, die das selbe Event beschreiben, werden als „related_tweet“ mit zu diesem Event abgelegt. Zu jedem Tweet werden auch Informationen, wie ein Zeitstempel und weitere Metainformationen, gespeichert. Danach werden unter Zuhilfenahme von Text-Mining-Verfahren, Schlüsselwörter aus den Tweets extrahiert, welche das Ereignis möglichst gut beschreiben. Diese werden dann als Suchparameter für die Google News Suche verwendet. Gesucht wird nach Artikeln der klassischen Medien zu dem Ereignis. Die Ergebnisse dieser Suche werden dem Event ebenfalls als „related_news“ beigelegt. Dadurch entsteht zu jedem Event ein Katalog von zugeordneten Nachrichtenartikeln. Die Events werden in der Datenbank tageweise gebündelt. So wird für jeden Tag ein separates Dokument in der Datenbank angelegt. Dieses lässt sich als Sammlung weiterer Dokumente betrachten und nennt sich „Collection“. In der Collection werden alle Events des Tages gesammelt.

Der verwendete Datensatz enthält alle Informationen zu Events vom 08.03.2020 - 15.03.2020, die von ScatterBlogs gesammelt wurden. Ein Event besteht aus einer Reihe von Attributen, welche als JSON-Dokument in einer Datenbank liegen. Zu den gespeicherten Eigenschaften jedes Events gehören unter anderem eine ID, Report-Score, Sprache, drei Terme als Überschrift, verwandte Events, so wie die vom Algorithmus zugeordneten Tweets. Im folgenden Abschnitt werden die für diese Arbeit wichtigen Attribute vorgestellt.

- **event_id:** Zu jedem Event wird ein eindeutiger Primärschlüssel benötigt. Daher wird für jedes Event eine eindeutige ID vergeben. So können Events gesucht und eindeutig zugeordnet werden.

- **report_score:** Der Report-Score beschreibt die errechnete Wichtigkeit eines Events. Berechnet wird der Wert aus der Abschätzung wie ungewöhnlich die Wörter innerhalb der Tweets sind, die zu dem Event gehören. Zusätzlich beeinflusst die Frequenz, mit welcher, die zu dem Event gehörenden Tweets gesammelt werden, den Wert. Werden in kurzer Zeit viele Tweets zu dem Event entdeckt, erhöht sich der Report-Score. Dieser Wert kann alle Zahlen in einem Wertebereich zwischen 0 und 1 annehmen. Events, deren Report-Score geringer als 0.3 berechnet wird, werden nicht in der Datenbank persistiert.
- **entries:** Das Attribut „entries“ beschreibt eine Sammlung aller Tweets, die zu dem Event gehören. Für jeden Tweet werden eine ID, Zeitstempel, Koordinaten, Anzahl der Retweets und der Text gespeichert.
- **related_news:** Unter diesem Abschnitt werden alle, durch die Google News Suche ermittelten Artikel zu dem Event gesammelt. Für jeden Artikel werden unter anderem eine Überschrift, ein Zeitstempel, der Beschreibungstext und ein Link gespeichert, der den Pfad zum originalen Artikel enthält.
- **first_notification_time:** first_notification_time gibt den Zeitpunkt an, zu dem das Event erstmals gefunden wurde. Als Unixzeit codiert ist sowohl das Datum, als auch die Uhrzeit.
- **prime_tags:** Primetags sind Terme, welche mittels Text-Mining-Verfahren aus den Tweets extrahiert wurden. Diesen Termen wird zusätzlich eine Zahl zugewiesen. Sie gibt die Anzahl der Vorkommen des Terms in der Sammlung der Tweets an.
- **headline:** Die Headline ist eine aus drei Worten zusammengesetzte Überschrift. Die drei Worte sind drei Terme, die das Event bestmöglich zusammenfassen. Dazu werden die drei Terme, die am häufigsten in den Tweets des Events vorkommen, aus der Liste der Primetags verwendet.
- **term:** Der Term ist ein Wort oder Hashtag, der das Event am besten beschreibt. Auch dieser wird aus der Liste der Primetags entnommen. Dafür wird der Eintrag mit der höchsten zugewiesenen Zahl gewählt.
- **related_events:** Unter diesem Punkt werden alle Events gesammelt, die eine thematische und/oder zeitliche Verbindung mit dem Event haben.
- **topic_channel:** Der Topic-Channel bezeichnet das Themengebiet, in welchem sich das Event verorten lässt.

2.2.3 Treemap

Treemaps eignen sich um hierarchische Beziehungen zwischen Daten übersichtlich abzubilden [SW01]. Durch diese Visualisierungsform wird der zur Verfügung stehende Platz bestmöglich ausgenutzt. Dabei wird eine visuell räumliche Darstellung eines gewurzelten Baumes erstellt. Als Datensatz zur Befüllung der Treemap dient eine Baumstruktur, bestehend aus einem Eltern-Knoten und davon abgehenden Kind-Knoten.

Derartige Baumstrukturen werden oft als Node-Link-Diagramme repräsentiert. Dabei wird ein Baumknoten als Kreis abgebildet. Die Verbindung zwischen den Knoten übernehmen teilweise gewichtete Linien (Kanten). Ein Treemap-Element wird jedoch abweichend von einem Baumknoten durch eine gewichtete Fläche dargestellt. Die Gewichtung wird als Flächeninhalt der Kachel abgebildet und muss im Datensatz definiert werden oder zur Laufzeit aus Attributen des Datensatzes berechnet werden. Zudem können die Kacheln farblich kodiert werden, wodurch schnell ein Überblick über die Daten erreicht werden kann.

Mit einer Interaktiven und zoombaren Implementierung einer Treemap, lässt sich das das Konzept „Overview first, zoom & filter, then details-on-demand“ gut umsetzen. So lassen sich sehr große Datensätze auf einer detailarmen, aber dafür übersichtlichen Ebene bündeln, und durch einen Klick in eine Kachel, kann mit steigender Tiefe der Detailgrad stetig erhöht und die Gruppierung nach und nach aufgehoben werden, bis man schlussendlich bei der detaillierten Ansicht eines Datenpunktes angelangt ist. In dieser Arbeit wird ein interaktiv veränderliches „Squarified Treemap“ Layout verwendet.

Squarified Treemap

Der „Squarified Treemap Algorithmus“ unterteilt die Zeichenfläche in rechteckige Flächen. Dabei wird versucht, jedem Rechteck ein möglichst gleiches Seitenverhältnis zuzuweisen. Die so entstehenden Rechtecke sind ähnlich der Form eines Quadrates. Der Algorithmus teilt rekursiv jede Hierarchieebene in näherungsweise quadratische Flächen. So wird der vorhandene Platz optimal genutzt, und es entsteht eine übersichtliche und benutzerfreundliche Visualisierung [BHW00].

2.2.4 Scatterplot

In einem Scatterplot, oder auch Streudiagramm genannt, werden grafisch Wertepaare statistischer Merkmale abgebildet. Jedes Wertepaar wird durch einen Punkt im Diagramm repräsentiert. Ein drittes Merkmal kann beispielsweise durch das Einfärben der Punkte visualisiert werden oder durch die Größe des Flächeninhalts. Mit Hilfe dieser Darstellung lassen sich Abhängigkeitsstrukturen wie Korrelationen und Cluster innerhalb eines Datensatzes abbilden und so auch erkennen.

2.3 Statistische Grundlagen

Im folgenden Abschnitt, werden die für das Verständnis dieser Arbeit benötigten statistischen Grundlagen geklärt. Diese werden für die Beantwortung der zweiten Forschungsfrage verwendet.

2.3.1 Pearson Korrelation

Für die Korrelationsanalyse des Datensatzes wird der Korrelationskoeffizient nach Pearson verwendet. Korrelation im Allgemeinen bietet eine Möglichkeit, Abhängigkeiten zwischen zwei Variablen zu beschreiben. Der Pearson Korrelationskoeffizient r, ist eine statistische Methode um den linearen Zusammenhang zweier Variablen zu ermitteln. Der Er wird durch folgende Formel bestimmt. [HHS19]

Sei:

$$\begin{aligned} x, y &: \text{Zufallsvariablen,} \\ \bar{x}, \bar{y} &: \text{Arithmetisches Mittel,} \\ n &: \text{Anzahl der Datenpunkte} \end{aligned}$$

So gilt für r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \tag{2.1}$$

Der Pearson Korrelationskoeffizient r kann Werte zwischen -1 und +1 annehmen und wird folgendermaßen interpretiert:

- $r \approx 0$: Wenn die zu prüfenden Variablen einen Koeffizienten nahe Null haben, so lässt sich keine statische Korrelation feststellen.
- $r > 0$: Wenn sich für r ein positiver Wert ermitteln lässt, so weist das auf eine positive Korrelation hin. Gemeint ist damit, dass für steigende x-Werte auch steigende y-Werte zu erwarten sind.
- $r < 0$: Wenn sich für r ein negativer Wert ermitteln lässt, so weist das auf eine negative Korrelation hin. Für steigende x-Werte werden die y-Werte kleiner.

2.3.2 z-Score

Für die spätere Visualisierung wird ein Indikator benötigt, um zu bestimmen, wie zutreffend die Bewertung der Wichtigkeit eines Events ist. Als Basis dafür dient der z-Score. Der z-Score oder auch z-Wert genannt, bemisst in der Statistik die Distanz zwischen dem Erwartungswert (Durchschnitt) und dem Rohwert. Die Einheit dieses Wertes ist eine Standardabweichung $[z * \sigma]$. Wird ein positiver z-Score ermittelt, so ist die Zufallsvariable X um $z * \sigma$ größer als der Erwartungswert. Ist der z-Score negativ, so ist X um $z * \sigma$ kleiner als der Erwartungswert.

Der z-Score berechnet sich aus der Zufallsvariable *X*, ihrem Erwartungswert *E(X)* und der Standardabweichung σ.

Sei:

$$\begin{aligned} X &: \text{Zufallsvariable,} \\ E(X) &: \text{Erwartungswert von } X, \\ \sigma &: \text{Standardabweichung} \end{aligned}$$

So gilt für z:

$$z = \frac{X - E(X)}{\sigma} \tag{2.2}$$

3 Related Work

3.1 Visualisieren hierarchischer Daten mit Treemaps

Eine Referenzimplementierung von Treemaps hat das HCI Lab der Universität in Maryland vorgestellt[1]. Diese Implementierung erlaubt es, durch einen Doppelklick auf ein Element eine Ebene hineinzuzoomen. Herausgezoomt wird durch einen Rechtsklick. Diese Interaktionstechnik ist jedoch für den Einsatz im Web nicht gut geeignet. Ein Klick mit der rechten Maustaste im Browser löst standardmäßig das Öffnen eines Kontextmenüs aus. Diese Interaktion zu verändern ist zwar möglich, fördert jedoch nicht das Benutzererlebnis. Der Benutzer erwartet im Browser keine steuernde Interaktion durch Klicken mit der rechten Maustaste. Ein Eingriff in die standardisierten Interaktionsmechanismen mit Browsern erfüllt nicht die Erwartung des Nutzers und führt zu Nichterkennen von vorhandenen Steuerungsmöglichkeiten. Dieses Problem existiert bei einer Vielzahl von Umsetzungen. Einige ähnliche Arbeiten, auf die dieselbe Problematik zutrifft, sind durch Schneiderman et al. [SP98] zusammengeführt. Die genannten Implementierungen fokussieren sich hauptsächlich auf die Darstellung der hierarchischen Struktur. Die Visualisierung detaillierter Informationen einzelner Datenpunkte rückt dabei in den Hintergrund. Durch das Darstellen des gesamten Baumes in einer Ebene wird die Übersichtlichkeit stark beeinflusst. Es sind zu viele grafische Elemente auf dem Bildschirm zu sehen. Eine klare Trennung der Äste des dargestellten Baumes zeigt weniger Elemente pro Ebene der Treemap an und sorgt so für eine bessere Übersicht. Eine Implementierung der Interaktion mit Zoomable Treemaps wird von Blanch et al. [BL07] vorgestellt. Diese legt die Grundsteine für die Übergänge zwischen den Ebenen und definiert erste Algorithmen für das „Zooming“. Elmqvist et al. [EF09] stellen einen Ansatz vor, der die Treemap-Felder mit weiteren Visualisierungsformen füllt. In die Kacheln werden die Inhalte durch eine visualisierte Aggregation (Pie-Chart) dargestellt. In meiner Arbeit wird eine Visualisierungsvariante vorgestellt, die das Konzept der „Zoomable Treemap“ und das Befüllen der Treemap-Kacheln mit anderen Visualisierungen kombiniert, um aus dem Twitter-Datenstrom extrahierte Events darzustellen.

3.2 Visualisieren von Ereignissen

Im Laufe der letzten Jahre brachte die Forschung einige Visualisierungsmethoden hervor, die es ermöglichen, die durch Sozial Media Dienste angefallenen Daten visuell zu analysieren und so besser zu verstehen. Diakopoulos et al. [DNK10] stellten eine Visualisierung vor, die durch verlinkte Ansichten dem Nutzer eine Analyse der Sozial Media Daten ermöglicht. In Eddi, [BSH+10] wird dann eine der ersten Visualisierungen von Twitterevents vorgestellt. Diese Implementierung stellt

[1] http://www.cs.umd.edu/hcil/treemap/

die Events in Form einer Wordcloud dar. Mathioudakis und Koudas stellten 2010 den „*TwitterMonitor*“ [MK10] vor. Mit diesem System lässt sich die genauere Beschaffenheit der Events besser erkennen. Angezeigt, wird eine Liste aller Events, wobei jedes Event durch eine Menge von Worten beschrieben wird. Zusätzlich wird ein Plot ausgegeben, welcher die Volumenentwicklung der Tweets pro Event zeigt. Weiter verbessert wurde dieser Ansatz dann von Lee et al. 2013[LLM13]. Das vorgestellte Tool umspannt einen ähnlichen Funktionsumfang wie der „*TwitterMonitor*“. Jedoch sind die Visualisierungsmethoden verbessert worden. So wird keine Liste mit Termen angezeigt um ein Event zu beschreiben, sondern die Terme werden mittels einer Wordcloud visualisiert. Marcus et al. [MBB+11] stellt *TwitInfo* vor. Ein Analysewerkzeug, welches automatisiert erkennt und visualisiert, wenn besonders viele Tweets gesendet werden. Eine zusammenfassende Visualisierung des Twitter-Datenstroms stellt Dörk et al. [DGWC10] vor. In dieser webbasierten Visualisierung kann der gesamte Twitter-Datenstrom durch Filterung der Tweets und verschiedene Visualisierungsformen analysiert werden. So wird die Entwicklung der diskutierten Themen zeitlich, in Form eines gestapelten Line-Charts dargestellt. Zu der zeitlichen Darstellung werden die beteiligten Personen sowie die vorhandenen Tweets visualisiert. Eine Sammlung der passenden geposteten Bilder wird dem Nutzer ebenfalls präsentiert. Diese Visualisierung zeigt den zeitlichen Verlauf der Themen im Twitter-Datenstrom. Eine weitere Visualisierung mit dem Ziel den Verlauf der Themen zu betrachten, ist die „ThemeCrowd“ von Archambault et al. [AGCH11]. Darin werden die Themen-Trends von Twitter, als „multilevel tag clouds“ präsentiert. Dazu werden die Twitter-User hierarchisch geclustert. Basierend auf den Themen, die sie behandeln, werden sie in hierarchisch unterteilten TagClouds visualisiert.

Guille et al. [GF15] stellen eine Herangehensweise vor, mit der Events durch eine Anomalierkennung aus dem Twitter-Datenstrom extrahiert werden. Das Paper stellt weiter einen Visualisierungsansatz der Ereignisse vor, welcher ein ähnliches Vorgehen bei der Form der Visualisierung wählt, wie diese Arbeit. Guille et al. gliedern die Visualisierung in drei Ansichten.
Die Erste visualisiert die Events entlang einer Zeitachse. Dabei wird die Dauer der Ereignisse sowie ihre zeitliche Abfolge dargestellt. Im Gegensatz zu dem in dieser Arbeit vorgestellten Konzept, werde thematischen Beziehungen der Events in dieser Ansicht nicht beachtet. Diese Ansicht teilt sich weiter in zwei Bereiche. Im unteren Bereich wird die Zeitachse gezeigt. Diese stellt Events und einen beschreibenden Term dar. Der obere Teil zeigt Details, zu einem im unteren Bereich ausgewählten Event. Angezeigt wird eine aus den Tweets extrahierte Beschreibung, ein zu dem Event passendes Bild und ein *Hypertext*.
Die zweite Visualisierungsvariante zeigt die Events anhand ihrer Auswirkungen auf die Anzahl Tweets zu diesem Ereignis. Gezeigt wird, wie sich die Anzahl gefundener Tweets zu einem Event über Zeit verändert. Visualisiert wird dies durch ein Line-Chart.
Die dritte Variante zeigt die Events thematisch sortiert. Dabei werden die Events in Knoten eines Baumes sortiert. Daraus wird ein Graph gezeichnet. Auch bei dieser Variante wird jedem Event ein Hauptterm zugewiesen. Er wird als grau gefärbter Knoten dargestellt. Der Durchmesser beschreibt das Maß der Auswirkung des Events. Zusätzlich werden verwandte Wörter als blaue Knoten dargestellt, welche über eine gewichtete Kante mit den Haupttermen verbunden sind. Die Dicke der Kante gibt an, wie stark die Verwandtschaft ist. Die beschriebene Visualisierung hilft dabei, verwandte Events anhand der thematischen Verwandtschaft zu identifizieren. Dennoch wird der Bezug, den die Events haben, nicht ganz klar. Die Informationen, die der Nutzer in dieser Ansicht über die Events gewinnen kann, ist begrenzt.

Die Events werden zeitlich und thematisch sortiert angezeigt. Je nach gewählter Ansicht wird auch die Wichtigkeit hergeleitet, indem die Auswirkung der Events auf die Anzahl der veröffentlichten Tweets abgebildet wird. ScatterBlogs hingegen erkennt die Events durch einen LDA[2]-Ansatz und kombiniert dies, mit dem Erkennen von Anomalien [CTB+12]. Die Wichtigkeit (siehe Kapitel 2.2) der Events wird unter anderem durch das Auswerten der gefundenen Anomalien und der Frequenz der eingehenden Tweets bestimmt. Die erkannten Events werden zeitlich, nach Themengebiet und Veröffentlichungsort gefiltert, angezeigt. Die Auswertung und Darstellung erfolgt in Echtzeit. Die vielen Filtermöglichkeiten erlauben es, Ereignisse in einer bestimmten Region und zu einem bestimmten Themenbereich einzusehen [BTH+13]. Das in meiner Arbeit vorgestellte System nutzt den von ScatterBlogs erstellten Datensatz mit Events. Basierend auf historischen Daten wird das Verhältnis von zugewiesenem Report-Score und der Anzahl, der aus einem Event entstandenen Nachrichtenartikel, dargestellt. So kann visuell ermittelt werden, ob der Report-Score plausibel ist. Um nach den Ursachen für unplausibel bewertete Events suchen zu können, werden viele Informationen zu dem Event benötigt. Um bei der Analyse nicht zwischen mehreren Systemen wechseln zu müssen, werden die bekannten Informationen zu den Events in der Visualisierung ebenfalls dargestellt.

[2]Latent Dirichlet Allocation

4 Konzept

In diesem Abschnitt wird das Konzept der Arbeit und die entwickelte Visualisierung vorgestellt. Des weiteren wird ein Bewertungsmaß eingeführt, welches es erlaubt, die Qualität des Report-Scores zu bemessen. Beschrieben wird der Prozess der Datenvorverarbeitung, wie die Erwartungswerte für das Bewertungsmaß bestimmt werden sowie der detaillierte Aufbau der Visualisierung. Letztere stellt die Relation von Report-Score und der Resonanz der klassischen Medien auf ein Event, beschrieben durch das Bewertungsmaß, dar.

4.1 Vorverarbeitung der Daten

Der in der Datenbank befindliche Datensatz, enthält alle Events und deren Beziehungen, welche visualisiert werden sollen. Um später ein besseres Benutzererlebnis bieten zu können und die Berechnung der Attribute für die Treemap aus dem Browser fern zu halten sowie die Ladezeiten der App zu verkürzen, werden die Rohdaten vorverarbeitet. Die verarbeiteten Daten werden in einem seperaten Dokument in der Datenbank abgelegt. So muss die Datenaggregation nicht bei jeder Anfrage des Nutzers erneut durchgeführt werden. Jedes Event wird hierzu auf die für diese Arbeit relevanten Attribute gekürzt. Somit sind in dem an den Browser übertragenen Datensatz nur noch die Daten enthalten, die auch tatsächlich für die Visualisierung benötigt werden.

4.1.1 Ermittlung der Erwartungswerte

Für diese Arbeit stellt die erwartete Anzahl zugeordneter Artikel je Event den Erwartungswert dar. Die Annahme ist, dass die Anzahl gefundener Newsartikel sich proportional zum Report-Score verhält. Daher sollte mit steigendem Report-Score auch eine höhere Anzahl an Nachrichtenartikeln für jedes Event gefunden werden. Zum Nachweis wird der Datensatz auf die Korrelation dieser Variablen getestet. Zu Gunsten der Übersichtlichkeit ist in Abbildung 4.1 nur der Plot aller aufgezeichneten Events vom 10.03.2020 dargestellt. Für die Datenanalyse wird jedoch der gesamte zur Verfügung stehende Datensatz genutzt. Auf der X-Achse ist der Report-Score aufgetragen. Die Y-Achse beschreibt die Anzahl gefundener Artikel. Die Visualisierung als Scatterplot, lässt auf den ersten Blick keinen linearen Zusammenhang zwischen Newsanzahl und Report-Score erkennen. Daher wird der Datensatz statistisch auf Korrelation getestet. Als Maß für die Stärke der Korrelation wird der Korrelationskoeffizient nach Pearson herangezogen. Die Berechnung des Pearson Korrelationskoeffizienten ist in Kapitel 2.3.1 detailliert beschrieben. Für die konkrete Berechnung wird die Python-Bibliothek *scipy.stats.stats* genutzt.

Da die zur Verfügung stehenden historischen Daten keine oder eine nur sehr schwache Korrelation von Report-Score und Artikelanzahl aufweisen, ist es nicht möglich, den Erwartungswert auf Grundlage der Regression zu prognostizieren [YS09]. Daher wird ein anderes Verfahren entwickelt,

um dennoch näherungsweise einen Erwartungswert bestimmen zu können. Dafür wird die Stichprobe und damit die in die Berechnung der Erwartungswerte einfließenden Datenpunkte verkleinert, so dass für kleinere Teilmengen des Datensatzes eine separate Berechnung des Erwartungswertes erfolgen kann. Der Datensatz wird dafür in sieben Cluster unterteilt. Jedes Cluster enthält die Events, die einen Report-Score-Bereich von 0, 1 umspannen. Der Datensatz enthält kein Event mit einem Report-Score, der geringer als 0, 3 ist. So beinhaltet das erste Cluster alle Events, deren Report-Score zwischen 0, 3 und 0, 4 liegt. Das Nächste startet bei einem Report-Score von 0, 4 und endet bei 0, 5. In Abbildung 4.1 sind die entsprechende Bereiche im Diagramm farblich gekennzeichnet. Dieses Verfahren setzt sich für den gesamten Wertebereich des Report-Scores fort. So wird jedes Element einem Cluster zugeordnet. Die Cluster werden im weiteren Verlauf der Arbeit ”Bins”genannt.

Sei:

$$\begin{aligned} 0,3 < i < 0,9, \\ j = i + 0,1, \\ e : \text{ein Event} \end{aligned}$$

wobei gilt:

$$e \in Cluster_{\text{i,j}} \Longleftrightarrow report\_score(e) \in [i, j] | j = i + 0,1 \tag{4.1}$$

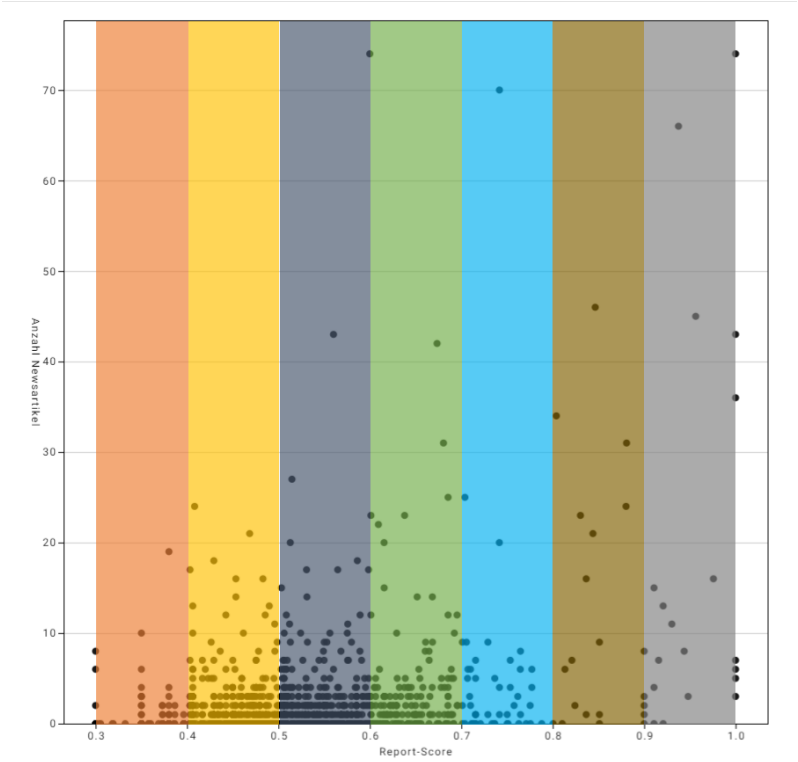

Abbildung 4.1: Clusterbildung

Für jeden Bin, wird das arithmetische Mittel der Artikelanzahl berechnet und jedem Event, welches sich innerhalb des Bins befindet, als „Erwartete Anzahl“ zugeordnet. So können für kleinere Bereiche des Report-Scores die Mittelwerte berechnet werden. Diese werden dann als erwartete Größe definiert. Jedem Event wird auch die für den Bin geltende Standardabweichung hinterlegt. Mit Hilfe der binspezifischen Zuordnung der Standardabweichung lassen sich die Events über den gesamten Datensatz vergleichen, obwohl keine (signifikante) Korrelation von Report-Score und der Anzahl gefundener Newsartikel über den gesamten Datensatz nachgewiesen werden kann.

Um für einzelne Events bestimmen zu können, ob das Verhältnis von Report-Score und Reaktion der klassischen Medien übereinstimmt, muss das Verhältnis von Report-Score und der Anzahl gefundener Artikel auf eine visuell gut darstellbare Kennzahl abgebildet werden. Dieses Verhältnis lässt sich nicht durch ein lineares Maß darstellen. Daher muss es für jeden Bin einzeln berechnet werden. Zusätzlich soll die dabei entstehende Kennzahl visualisiert werden. Daher soll das angelegte Maß leicht verständlich und somit gut interpretierbar sein. Das Verhältnis lässt sich mittels des z-Scores jedes Events leicht verständlich und gut virtualisierbar abbilden. Deshalb wurde dieser als Grundlage für das Bewertungsmaß gewählt. Entgegen der herkömmlichen Berechnung des z-Score, bei dem die Standardabweichung des gesamten Datensatzes genutzt wird, wird für die Berechnung in dieser Arbeit die binspezifische Standardabweichung als Grundlage für die Berechnung genutzt. Der so entstehende Wert bildet das Bewertungsmaß der Qualität des Report-Scores numerisch ab. Als grundlegende Annahme wird hierbei vorausgesetzt, dass bei einem hohen Report-Score auch eine große Resonanz zu diesem Event in den klassischen Medien existiert. Daher wird für das Bewertungsmaß die Relation von Report-Score und der Anzahl gefundener News eines Events untersucht. Der entstehende Wert wird im weiteren Verlauf der Arbeit Abweichungsfaktor (AF) genannt und berechnet sich je Event wie folgt:

Sei:

$$
\begin{aligned}
X &: \text{Anzahl gefundener News,} \\
\mu_{\text{Bin}} = E(X) &: \text{Erwartungswert (Durchschnittliche Newsanzahl des Bins),} \\
\sigma_{\text{Bin}} &: \text{binspezifische Standardabweichung}
\end{aligned}
$$

So gilt für den Abweichungsfaktor:

$$AF = \frac{X - \mu_{\text{Bin}}}{\sigma_{\text{Bin}}} \tag{4.2}$$

Dadurch wird ein standardisiertes Maß geschaffen, welches dennoch die unterschiedlichen Verteilungen innerhalb der Bins beachtet und somit eine Vergleichbarkeit über den gesamten Datensatz schafft. Der AF kann beliebig große, sowohl positive als auch negative Werte annehmen. Der Benutzer kann in der Visualisierung einstellen, wie die Events, abhängig von diesem Faktor, eingefärbt werden. Daher kann der Faktor benutzt werden, um die Schwelle derjenigen Events zu bestimmen, die als plausibel oder richtig zugeordnet gelten. Wie auch beim klassischen z-Score ist die Einheit dieses Faktors eine Standardabweichung.

Die Werte des AF werden wie folgt interpretiert:

- $AF < 0$: Das Event, hat einen hohen Report-Score relativ zu den restlichen Events in diesem Bin. Die Anzahl gefundener Nachrichtenartikel ist relativ zum Report-Score gering. Die Wahrscheinlichkeit, dass dieses Event als „zu wichtig“ eingestuft wurde, ist hoch.
- $AF \approx 0$: Liegt der AF nahe null, so liegt das Event im Durchschnitt des Bins. Somit ist die gefundene Anzahl der Nachrichtenartikel nahe dem Erwartungswert (durchschnittlicher Wert des Bins). Das deutet darauf hin, dass der Algorithmus die Wichtigkeit des Events korrekt eingestuft hat. Wie groß der Bereich für „als richtig eingestuft“ ist, kann der Benutzer des Visualisierungstools bestimmen.
- $AF > 0$: Liegt der AF im positiven Bereich, so wurden mehr Nachrichtenartikel zu dem Event gefunden, als durchschnittlich bei anderen Events des selben Bins. Daher liegt die Vermutung nahe, dass das Event nicht als wichtig genug eingestuft wurde.

4.1.2 Gruppierung der Daten nach topic_channel

Schon bei der Vorverarbeitung der Daten im Backend wird der Datensatz thematisch gruppiert. Die Gruppierung erfolgt nach „Topic-Channel“. Der Algorithmus von ScatterBlogs weist jedem Event einen Topic-Channel zu (siehe Kapitel 2.2). Er zeigt an, zu welchem Themengebiet das jeweilige Event am wahrscheinlichsten gehört. Die Topic-Channel stellen die erste Ebene des Baumes, unterhalb des Wurzelknotens für die Treemap dar siehe Abbildung 4.2. Die nächst tiefer liegende Ebene, also die Kindknoten der Topic-Channel, sind Terme. Jedes Event hat als Attribut einen Term besc, der das Event mit einem Wort oder „Hashtag“ beschreibt. Da mehreren Events der gleiche Term zugewiesen wird, sind die Kindknoten der Terme die Events, die diesen Term als Attribut tragen. Jede Ebene des Baumes wird in eine separate Ebene der Treemap überführt.

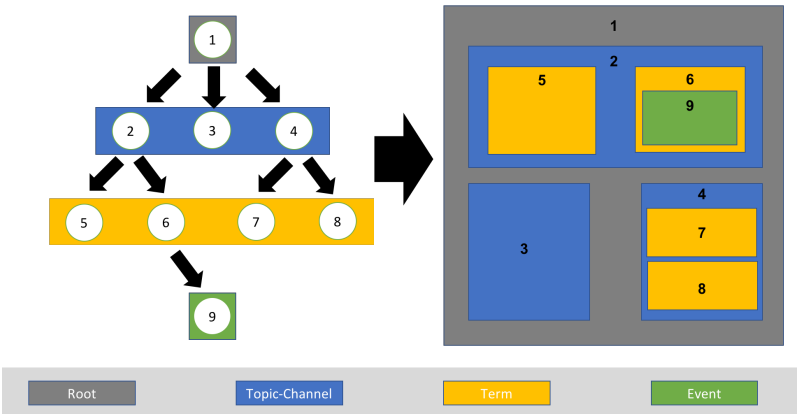

Abbildung 4.2: Abbildung der Baumstruktur als Treemap-Visualisierung

4.2 Konzept Zoomable Treemap

Die statische Treemap-Visualisierung wird für diese Arbeit um zahlreiche Interaktionsmöglichkeiten erweitert. Entgegen der herkömmlichen Darstellung einer Treemap aus Kapitel 2.2.3 wird für die Implementierung eine interaktiv veränderliche Form gewählt. Um die Übersicht eines großen

Datensatzes noch weiter zu verbessern, werden die in der Treemap dargestellten Informationen nicht alle auf einer grafischen Ebene und daher geschachtelt dargestellt, sondern in mehrere untereinander liegende Ebenen aufgeteilt. Der Benutzer sieht in jeder Ebene nur die Elemente, welche in der gleichen Ebene liegen. Der Flächeninhalt der einzelnen Kacheln spiegelt den aufsummierten Report-Score der enthaltenen Events wieder. Damit bildet er die aggregierte Gesamtrelevanz der Kachel ab. So ist es möglich, direkt zu erkennen in welchem thematischen Bereich sich die meisten relevanten Events befinden. Die Hintergrundfarbe der Kacheln zeigt an, ob sich in dem jeweiligen Themengebiet Events befinden, bei denen der Report-Score nicht zu der gefundenen Anzahl an Nachrichtenartikeln passt. Je dunkler die Färbung ist, desto weiter ist das „schlechteste“ Element vom Erwartungswert entfernt. Das Event, bei dem der Erwartungswert absolut am weitesten von der erwarteten Anzahl an Newsartikeln für diesen Report-Score abweicht (absolut höchster AF), wird in diesem Zusammenhang als das „schlechteste“ Element bezeichnet. In Abbildung 4.11 wird das Themengebiet „sb_sports“ blau eingefärbt, da mindestens ein Event enthalten ist, welches einen sehr hohen AF hat.

4.3 Beschreibung der GUI

In diesem Abschnitt wird der grafische Aufbau der Implementierung inklusive aller Funktionen vorgestellt. Grundlegend ist die Visualisierung in zwei Ansichten unterteilt. Der Benutzer kann über ein Einstellungsmenü zwischen einer Ansicht als Zoomable Treemap und einer Ansicht als interaktiver Scatterplot wechseln. Standardmäßig steigt der Benutzer über die Plotansicht ein. Sollte keine Verbindung zum Backend-Service vorhanden sein, so wird eine Fehlermeldung angezeigt, und die Visualisierung kann nicht abgerufen werden.

4.3.1 Plotansicht

In Abbildung 4.3 sind die Events eines einzelnen Tages visualisiert. Die Ansicht ist unterteilt in unterschiedlich groß dargestellte Kacheln, welche die Topic-Channel repräsentieren. Der Flächeninhalt jeder Kachel wird bestimmt, durch die Anzahl der Events und deren Report-Score, die dem Topic-Channel zugehörig sind. Je größer eine Kachel dargestellt ist, desto mehr Events mit hohem Report-Score beinhaltet das dazugehörige Themengebiet. Die Kacheln werden von links oben nach rechts unten kleiner. Demnach sind die Themengebiete von groß und wichtig nach klein sortiert. Dadurch wird die Aufmerksamkeit des Nutzers zuerst auf die Kacheln gelenkt, welche die meisten Events enthalten. Die Annahme hierbei ist, dass ein Themengebiet, zu dem es viele Events gibt, interessanter ist, als eines zu dem nur wenige Events existieren. Da die natürliche Lese- und Schreibrichtung in Europa von links nach rechts ist, wird die Darstellung auch hier so gewählt. Dies bietet dem Nutzer ein möglichst natürliches Benutzererlebnis, da sich die gewohnte Leserichtung auch beim Analysieren der Daten wiederfindet. Die Anordnung der Themengebiete gleicht der einer Treemap-Visualisierung. So wird die erste Ebene einer Treemap dargestellt. Jedoch wird nicht nur ein Label angezeigt, sondern der Inhalt der Kachel wird unter Zuhilfenahme einer weiteren Visualisierungsmethode detailliert dargestellt. Jede Kachel ist daher neben der Überschrift des Themengebiets mit einem Scatterplot der zugehörigen Events gefüllt. Die X-Achse beschreibt den Report-Score und auf der Y-Achse ist die Anzahl der zugeordneten Artikel (related_news) aufgetragen. Der Wertebereich und die Beschriftung der Achsen wird dynamisch an die darzustellenden Daten angepasst, so dass der verfügbare Platz innerhalb einer Kachel, bestmöglich genutzt

wird. Durch das Verknüpfen der zwei Visualisierungsmethoden (Scatterplot und Treemap) kann der verfügbare Platz auf dem Bildschirm optimal ausgenutzt werden, um eine detailreiche aber übersichtliche Visualisierung zu schaffen. So wird der Vorteil der optimalen Nutzung des Platzes durch die Treemap und der Vorteil der Scatterplots, die kompakte Darstellung vieler Events, kombiniert.

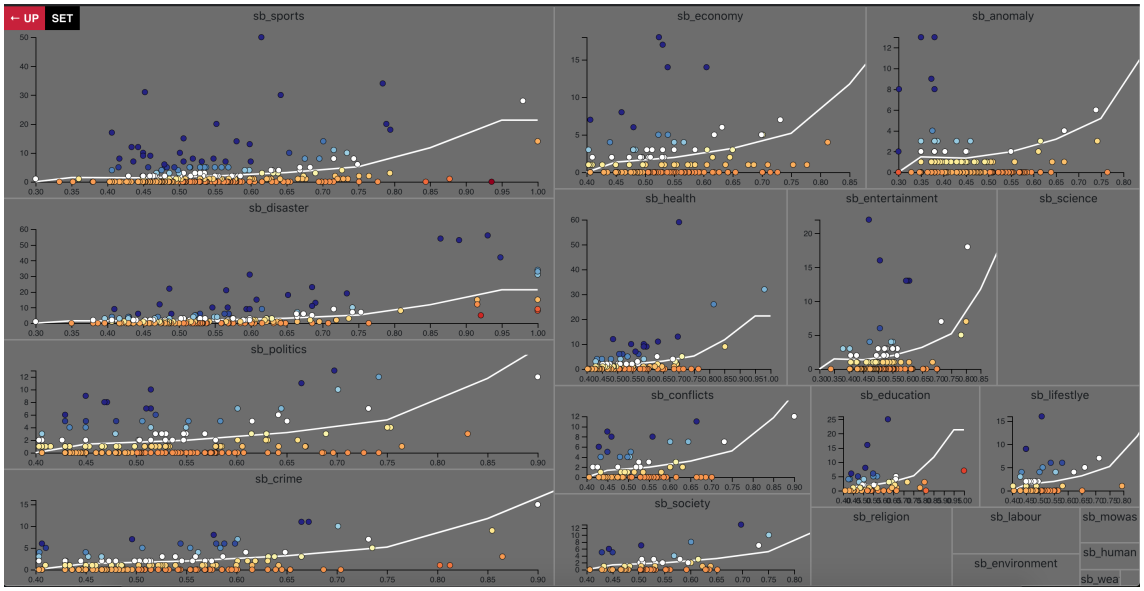

Abbildung 4.3: Plotansicht

Die Events sind durch gefärbte Punkte dargestellt. Die Färbung der Punkte gibt an, wie weit die tatsächliche Artikelanzahl von der erwarteten Anzahl abweicht. Visualisiert wird nicht die absolute Differenz, sondern der vorher berechnete AF des Events. Ereignisse, denen weniger Artikel zugeordnet wurden als der Erwartungswert vermuten lässt, werden durch eine rötliche Färbung dargestellt. Je dunkler der Rotton ist, desto weiter ist der tatsächliche Wert von der Erwartung entfernt. Events, zu denen mehr Artikel gefunden werden als der Erwartungswert vorgibt, sind blau eingefärbt. Je größer die Differenz zum Erwartungswert ist, desto dunkler wird der Blauton. Weiß dargestellt werden jene Events, deren zugeordnete Anzahl an Artikeln dem Erwartungswert entsprechen. Der Benutzer kann in den Einstellungen festlegen, wie weit ein Event abweichen darf, um noch als weiß gekennzeichnet zu werden. So entsteht ein Farbverlauf von rot zu weiß zu blau. In jeder Kachel wird der Verlauf des Erwartungswertes durch eine weiße Linie beschrieben. Diese dient dazu, dem Nutzer die absolute Abweichung vom Erwartungswert anzuzeigen. So wird die relative Abweichung innerhalb des Bins durch die farbliche Kodierung und die absolute Differenz zwischen Erwartungswert und tatsächlichem Wert durch den Abstand zwischen dem Punkt und der weißen Linie angegeben. Wird eine Kachel in der Breite schmaler als ca. 200 Pixel dargestellt, so ist der Inhalt auf gängigen Desktop-Monitoren häufig nicht mehr gut erkennbar. Achsenbeschriftungen sind nicht mehr lesbar oder die dargestellten Datenpunkte sind nicht mehr klar differenzierbar. Daher wird der Inhalt in diesem Fall nicht mehr in der Übersicht dargestellt (siehe Abbildung 4.3 unten rechts). Der Nutzer kann jedoch durch Klicken in die graue Fläche dieser Kachel oder auf die Überschrift, sich den Inhalt und somit den fehlenden Plot anzeigen lassen. Der Plot öffnet sich dann vergrößert und zeigt den gesamten Inhalt der geklickten Kachel an. Dies ist in Abbildung4.4 zu sehen. Ein solches Problem wäre auch durch das Einfügen einer aggregierenden Visualisierungsform

in die Kachel lösbar gewesen. Den Plot auszublenden und nach dem Klicken vergrößert darzustellen, bringt jedoch den Vorteil, dass so keine zusätzliche Visualisierungsform zum Einsatz kommt. Die Handhabung und die Verständlichkeit der Gesamtvisualisierung wird so unterstützt.

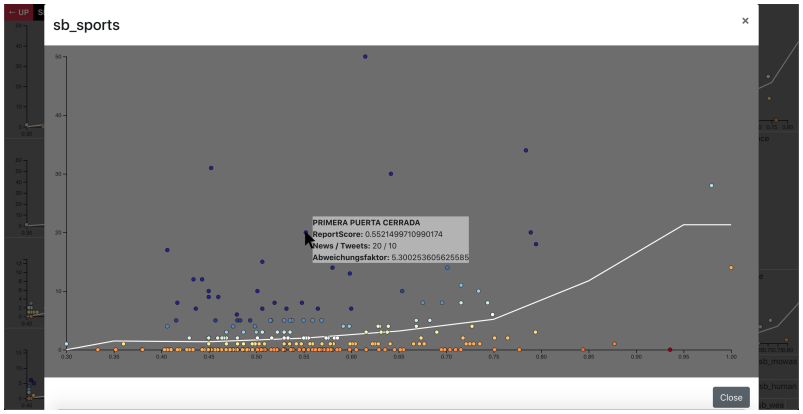

Abbildung 4.4: Vergrößerter Plot mit angezeigtem Tooltip

Legt der Nutzer den Cursor über einen Datenpunkt, so wird ein Tooltip eingeblendet, der die Überschrift des Events, den Report-Score, die Anzahl zugeordneter News und Tweets sowie den AF anzeigt. (Abbildung 4.4). Dadurch ist schon auf der obersten Übersichtsebene ein Überblick über die Events sowie ein Einblick in ausgewählte Details möglich. Dies ist insbesondere deshalb wichtig, da auf die Rasterung der Plots verzichtet wurde. Ohne eingezeichnete Gitternetzlinien, sind die Absolutwerte der Events nicht mehr so leicht abzulesen, jedoch wird die Übersichtlichkeit deutlich verbessert. Jeder Datenpunkt (ein Event) kann zusätzlich angeklickt werden. Sowohl in der Übersicht als auch in den vergrößerten Plots. Wird ein Event angeklickt, so werden alle bekannten Details zu dem Event grafisch aufbereitet in der Detailansicht angezeigt.

4.3.2 Eventdetails

Die Detailansicht der Events öffnet sich in einem Modal[1]. So bleibt die Visualisierung im Hintergrund erhalten. Der Nutzer kann die Detailansicht jederzeit durch Klicken auf den „Close-Button“ oder durch Klicken neben das Modal verlassen. Er gelangt direkt wieder zu der zuvor dargestellten Ansicht. So wird zeitaufwendiges und störendes Neuladen der gesamten Seite vermieden. Die Detailansicht beinhaltet grafisch aufbereitet alle verfügbaren Informationen zu einem Event. Diese Informationen dienen dazu, als Nutzer zu verstehen, um was für ein Event es sich handelt und wie der Report-Score zustande gekommen ist. Diese Ansicht teilt sich in fünf Abschnitte. Wie in Abbildung 4.5 ersichtlich ist, werden im obersten Abschnitt allgemeine Informationen zu dem Event angezeigt. Darunter zählen die Überschrift, Report-Score, Anzahl Artikel und Tweets, Erwartungswert, AF sowie drei weitere beschreibende Felder der Datenbank. Mit Hilfe dieser Informationen ist noch kein vollständiger Einblick in das Event möglich. Daher kann der Nutzer weiter unten in dieser Ansicht vier weitere Felder aufklappen. Um dem Nutzer auf einen Blick signalisieren zu können,

[1] Überlappendes UI-Element. https://getbootstrap.com/docs/4.0/components/modal/

welche Informationen für die Analyse zur Verfügung stehen, sind die aufklappbaren Felder beim Öffnen der Ansicht geschlossen. Nur so können alle Überschriften der Felder gleichzeitig dargestellt werden.

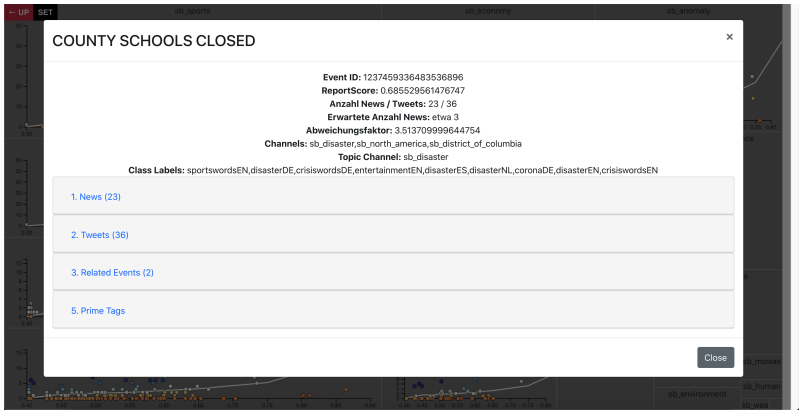

Abbildung 4.5: Detailansicht eines Events

Das erste dieser Felder beinhaltet alle, dem Event zugeordneten Artikel chronologisch sortiert. In der Überschrift des Feldes lässt sich erkennen, wie viele Nachrichtenartikel zugeordnet wurden. Existieren keine Artikel, so wird in der Überschrift eine 0 angezeigt und in dem aufgeklappten Feld ist der Schriftzug „keine News gefunden“ eingeblendet. Zu jedem Artikel im Datensatz wird angezeigt, sowohl wann er erschienen ist, als auch welche Überschrift er trägt. Zusätzlich wird, falls verfügbar, eine kurze Beschreibung angezeigt, damit der Nutzer einen Überblick über die thematische Einordnung des Artikels bekommt. Über einen bereitgestellten Link, kann zu dem originalen Online-Artikel gesprungen werden.

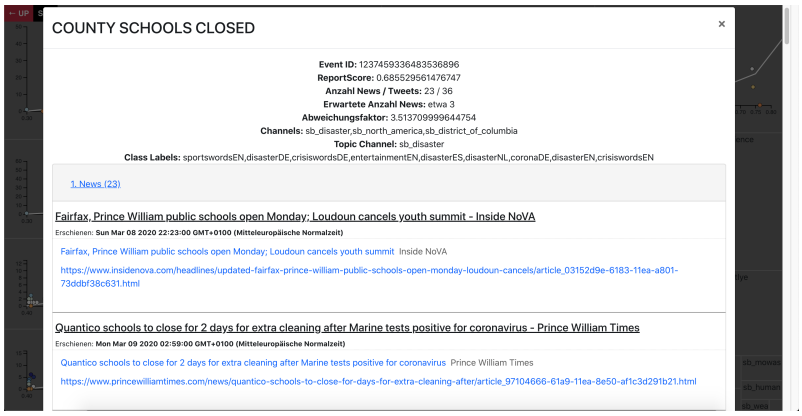

Abbildung 4.6: Detailansicht eines Events mit aufgeklapptem Feld News

Klappt der Nutzer das zweite Feld auf, so wird das erste automatisch wieder eingefahren, um unnötig langes Scrollen beim Wechseln der Felder zu vermeiden. In diesem Feld werden ähnlich der Darstellung der Artikel, die zu dem Event gehörenden Tweets aufgelistet. Auch zu jedem Tweet wird der Zeitstempel und der Tweetinhalt ausgegeben. Als zusätzliche Information und um die Relevanz eines Tweets besser einschätzen zu können, wird die Anzahl der „Retweets“ mit ausgegeben. Ein

Retweet ist ein erneut geposteter Tweet. Dies kann sowohl ein eigener oder der eines anderen Nutzers sein. Wurde ein Tweet oft retweeted, so ist davon auszugehen, dass die enthaltene Nachricht eine höhere Relevanz besitzt.

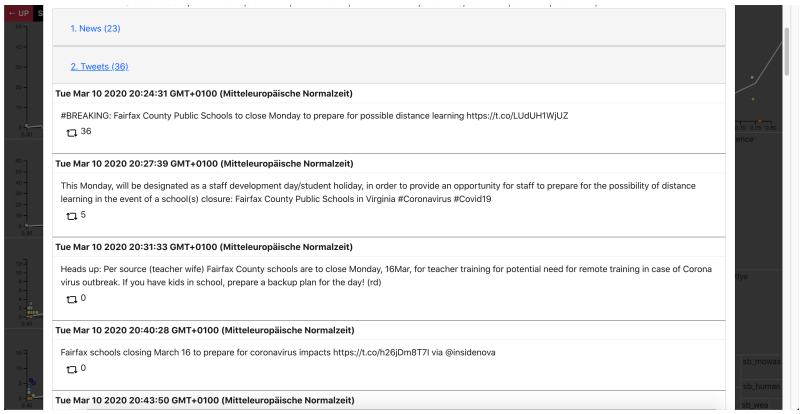

Abbildung 4.7: Detailansicht eines Events mit aufgeklapptem Feld Tweets

Bei der Speicherung der Events werden zu jedem Ereignis auch verwandte Events abgelegt. siehe Kapitel 2.2 Diese werden zeitlich sortiert und in Relation zu dem aktuell betrachteten Event im dritten Feld dargestellt. Die Ereignisse werden alternierend links und rechts neben einer Zeitachse positioniert. Das oberste Event ist dabei das zeitlich am weitesten zurückliegende. Das Event, welches gerade in der Detailansicht geöffnet ist, wird farblich gekennzeichnet. So können die verwandten Events mit dem aktuell betrachteten leichter verglichen werden. Für jedes Event wird die Überschrift und der Zeitstempel dargestellt. Des weiteren wird der zeitliche Verlauf der Events durch Verbindungslinien an die Zeitachse modelliert. Dabei werden die Abstände zwischen den Events auf die Zeitspanne zwischen dem frühesten und dem spätesten Event linear transformiert. Damit wird ersichtlich, wenn Ereignisse sehr nah beieinander liegen und somit wenig zeitlicher Abstand zwischen ihnen liegt. So lässt sich herleiten, ob die dargestellten Events zu einem größeren Ereignis gehören können.

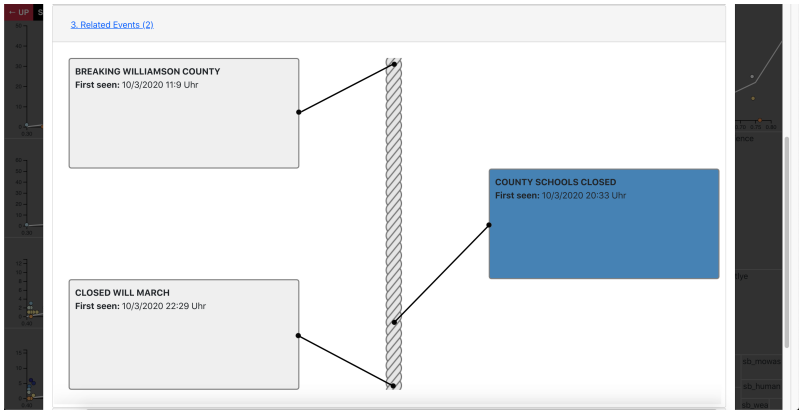

Abbildung 4.8: Detailansicht eines Events mit aufgeklapptem Feld Related Events

Klappt der Nutzer das vierte Feld auf, so wird eine Wordcloud angezeigt, welche die „Primetags“, siehe Kapitel 2.2, aus der Datenbank visualisiert darstellt. Dabei werden die Worte, denen in der Vorverarbeitung eine größere Bedeutung zugesprochen wurden, größer dargestellt. Mit Hilfe der

Tagcloud lässt sich abhängig von der Qualität der extrahierten Primetags erahnen, um was für ein Event es sich handelt, ohne die Artikel oder Tweets komplett lesen zu müssen. Diese Ansicht dient als zusätzliches Hilfsmittel zum schnelleren und besseren Verständnis des Events. Die Visualisierung eines Textes mit Hilfe der Wordcloud, ist ein bewährtes Mittel, um schnell den Kontext eines Textes zu erkennen [HLLE14]. Dieser Vorteil wird auch hier genutzt. Der Nutzer kann aus den oben genannten Feldern die benötigten Informationen herauslesen, um selbst zu bewerten, wie relevant das Event eingestuft werden sollte. Die daraus gewonnene Einschätzung lässt sich dann mit dem Errechneten „report_score“ und der automatisierten Auswertung durch den AF vergleichen.

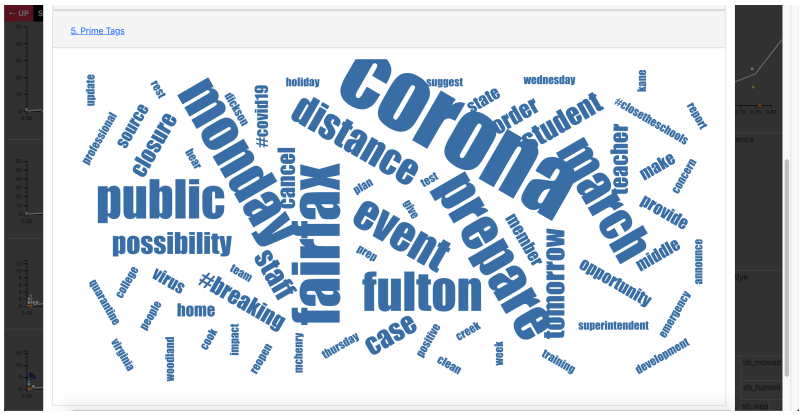

Abbildung 4.9: Detailansicht eines Events mit aufgeklapptem Feld Primetags

4.3.3 Einstellungsmenü

Die Schaltfläche, um das Einstellungsmenü zu öffnen, befindet sich in der linken oberen Ecke direkt neben dem „up-Button“. Diese Schaltfläche ist unabhängig der gewählten Ansicht immer sichtbar. Auf das dauerhafte Einblenden der Einstellungen in einer Sidebar wurde zu Gunsten des Platzes verzichtet. Mit Klick auf diese Schaltfläche wird das Einstellungsmenü (Abbildung 4.10) eingeblendet. Über ein Dropdown-Menü kann die angezeigte Visualisierungsart gewechselt werden. So kann zwischen der Plotansicht und Hierarchiedarstellung umgestellt werden. Die sonstigen Einstellungen sowie die Daten bleiben dabei erhalten. Es wird ausschließlich die visuelle Gestaltung der Ansicht gewechselt.

Darunter kann die farbliche Darstellung der Events sowohl für die Plotansicht als auch für die hierarchische Darstellung angepasst werden. Es kann eingestellt werden, ab welchem AF die Events farblich kodiert werden sollen. Da jeder Datensatz eine andere Varianz und somit auch eine andere Standardabweichung aufweist, kann so der Nutzer die Visualisierung auf jeden Datensatz individuell anpassen. Zudem kann durch das Verstellen der Schranke der Fokus des Anwenders geleitet werden. Wird die Schranke sehr hoch eingestellt, so werden alle Events weiß (farblich nicht kodiert) dargestellt, bis auf die mit der größten Abweichung. Events die farblich nicht kodiert sind, können für die Plotansicht ausgeblendet werden (Haken entfernen bei „show good events“). So werden die großen statistischen Ausreißer sehr gut und leicht ersichtlich. Der Nutzer kann dann schrittweise die Schranke senken, und so systematisch den Datensatz analysieren.

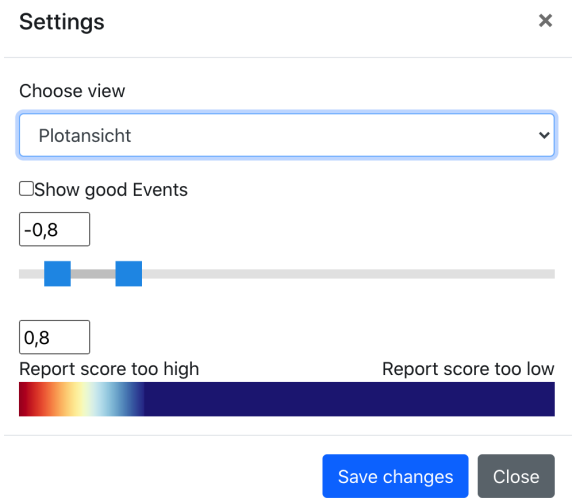

Abbildung 4.10: Einstellungsmenü

Nach dem Klick auf „Save changes“, werden alle Einstellungen gespeichert. So können nach wiederholtem Öffnen oder Neuladen der Seite, die zuvor getätigten Einstellungen wiederhergestellt werden.

4.3.4 Treemap-Ansicht

Hat der Nutzer im Einstellungsmenü die Ansicht auf die hierarchische Darstellung gewechselt, so wird eine Anzeigevariante gewählt, die von der Treemap abgeleitet wurde. Die Treemap wird in drei Ebenen unterteilt. Jede Ebene zeigt einen höheren Detailgrad. Der Nutzer kann durch Klicken der Elemente in eine tiefere Ebene wechseln und sich die Inhalte detaillierter anzeigen lassen. Durch einen Klick auf die sich links oben im Eck befindliche Schaltfläche „<–“, gelangt der Nutzer wieder eine Ebene nach oben. In dieser Schaltfläche wird dem Nutzer neben dem Pfeil die Bezeichnung der aktuellen Ebene angezeigt. Dadurch wird sichergestellt, dass beim Wechseln der Ebenen der Überblick stets erhalten bleibt.

Abbildung 4.11: Hierarchische Darstellung der Ebene 1. Dargestellt werden die Topic-Channel.

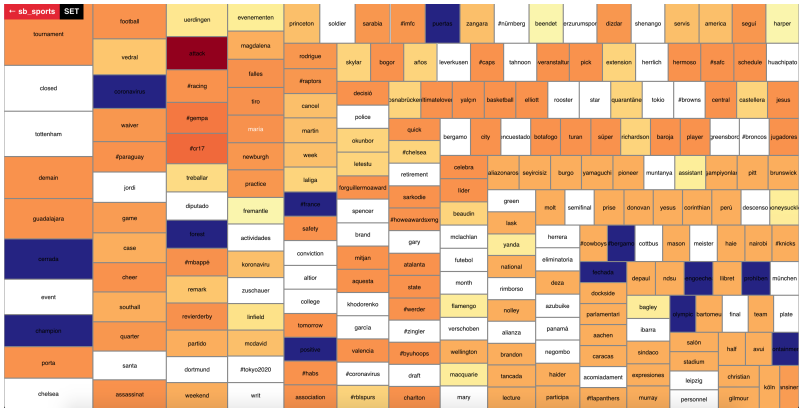

Abbildung 4.12: Hierarchische Darstellung der Ebene 2. Dargestellt werden die Terme innerhalb des Topic-Channel „sb_sports“.

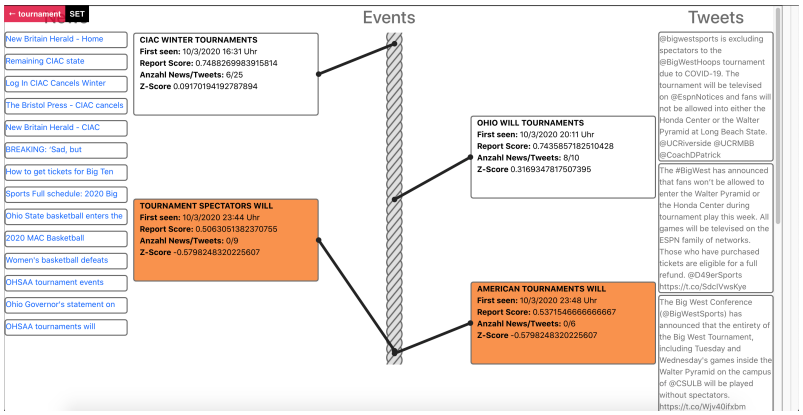

Abbildung 4.13: Hierarchische Darstellung Ebene 3. Dargestellt werden die Events die dem Term „tournament“ im Topic-Channel „sb_sports“ zugewiesen wurden.

Die Gruppierung der Events auf der ersten Ebene ist analog zur Plotansicht gewählt (siehe Abbildung 4.3 und Abbildung 4.11). So werden auch bei dieser Ansicht die Events nach „topic_channel“ gruppiert. Jede dargestellte Kachel repräsentiert einen Topic-Channel. Der Benutzer erhält eine thematische Übersicht über die im Datensatz enthaltenen Events. Die Größe der Flächeninhalte ist auch hier proportional zu der Anzahl Events und deren Report-Score, die von der jeweiligen Kachel repräsentiert werden. Die farbliche Kodierung der Kachel spiegelt die größte Abweichung vom Erwartungswert wieder. Dabei wird das Event berücksichtigt, welches den absolut größten AF hat. Dieser wird dann entsprechend dem in Kapitel 5.5.1 beschriebenen Verfahren kodiert. Dadurch erhält der Nutzer auf einen Blick die Übersicht, in welcher Kategorie sich mindestens ein Event befindet, welches einen prüfenden Blick erfordert. Der Fokus der Visualisierung ist darauf ausgerichtet, schnell und effizient die größten Ausreißer zu identifizieren. Daher wird die Kachel nach dem „schlechtesten“ Event gefärbt. Wie auch bei der Plotansicht, kann die Treemap mittels Interaktionen manipuliert werden. Durch Klicken auf ein Themengebiet, werden die anderen Topic-Channel ausgeblendet und eine feiner aufgelöste, gruppierte Übersicht des gewählten Themas erscheint (siehe Abbildung 4.12). Die Label der Kacheln repräsentieren jetzt sogenannte Terme. Jedes Event im Datensatz hat mehrere zugewiesene Terme. Sie sind als ”Primetags”hinterlegt. Der

erste Primetag in der Liste mit dem höchsten Wert, wird als Term gesetzt. Da mehreren Events die gleichen Terme zugewiesen werden, bilden sie die nächste Stufe der Gruppierung. Für eine genauere Erklärung der Terme siehe Kapitel 2.2. Die Kacheln werden auch in dieser Ebene abhängig vom höchsten vorkommenden AF eingefärbt.

Klickt der Nutzer auf eine dieser Kacheln, so gelangt er zu einer Übersicht, der in dem Term der Kachel enthaltenen Events, dargestellt in Abbildung 4.13. Diese Ansicht ist in drei Spalten unterteilt. Die linke Spalte zeigt chronologisch sortiert die Überschriften der Artikel zu allen Events. Legt der Benutzer den Cursor auf eine Überschrift, so wird ein Tooltip eingeblendet, welcher die in der Datenbank hinterlegte Beschreibung des Artikels anzeigt (siehe Abbildung 4.14).

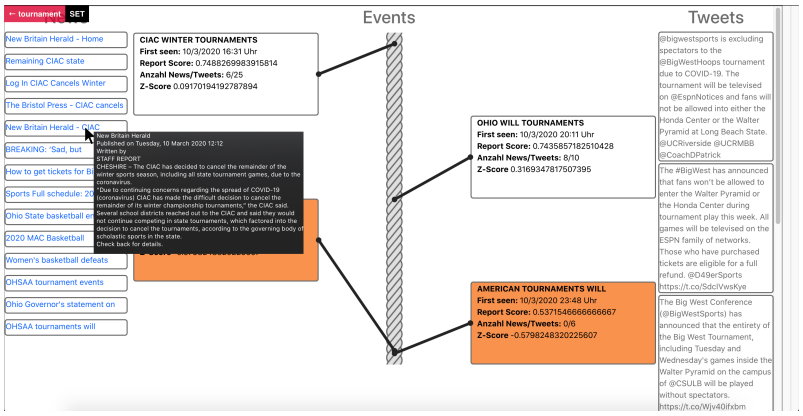

Abbildung 4.14: Hierarchische Darstellung Ebene 3. Nach dem Überfahren eines Artikels wird ein Tooltip mit einer Kurzbeschreibung des Artikels eingeblendet.

Klickt der Nutzer auf einen der Artikel, so wird er zu dem Online-Artikel weitergeleitet und kann sich dort den gesamten Artikel durchlesen. Die mittlere Spalte zeigt alle Events der Gruppierung zeitlich sortiert an. Das Event, welches am frühesten erfasst wurde, steht oben. Wie schon aus der Detailansicht bekannt, findet sich hier die Zeitachse wieder, mit deren Hilfe die Abstände zwischen den Events modelliert werden. Die einzelnen Events werden als rechteckige Kästchen dargestellt und sind mit den Basisinformationen zu den Events beschriftet. Dazu zählen eine Überschrift, der Zeitstempel, der Report-Score, die Anzahl der Artikel und Tweets sowie der AF. Die Hintergrundfarbe der Events orientiert sich an der farblichen Kodierung durch den AF. Wird ein Event geklickt, so wird die Detailansicht zu dem geklickten Event angezeigt. Es erscheint die gleiche Ansicht, wie bei dem Klick auf einen Datenpunkt in der Plotansicht. Die dritte und damit rechte Spalte, dient zum Anzeigen aller Tweets der in der mittleren Spalte dargestellten Events. Auch diese ist chronologisch sortiert. Der Tweet mit dem frühesten Zeitstempel wird als oberstes Element angezeigt. Diese Ansicht zeigt die thematische Verbindung der Events sehr gut.

Eine Möglichkeit das Erscheinungsbild der gesamten Treemap-Ansicht zu ändern, ist das Anpassen der Farbkodierung der Kacheln. Diese Anpassung kann im Einstellungsmenü vorgenommen werden und wirkt sich auf jede Ebene der Treemap aus. Die Checkbox „show good events“ hat in der Treemap-Ansicht keine Auswirkungen. Es werden immer alle Events in allen Gruppen beachtet und dargestellt. Würden die Events gänzlich ausgeblendet werden, so würden sich auch die Flächeninhalte der Kacheln und damit einhergehend auch die Sortierung der Größe nach ändern. Da dies nicht erwünscht ist, werden als „gut“ markierte Events lediglich farblich gekennzeichnet.

4.4 Exemplarischer Ablauf der Analyse

Ein typischer Analyseablauf mit dem Ziel, die statistischen Ausreißer zu identifizieren und eine Erklärung für die Bewertung zu finden kann wie in Abbildung 4.15 beschrieben gestaltet werden. Zuerst entscheidet der Nutzer sich für eine der Darstellungsvarianten (Plotansicht oder Treemapansicht). Danach wird in den Einstellungen der Slider so lange angepasst, bis das gewünschte Bild entsteht (Abb. 4.15.a1,b1).

Wurde die Plotansicht(Abb. 4.15.a) gewählt, können zusätzlich Events, die in der aktuellen Analyse nicht betrachtet werden sollen, ganz ausgeblendet werden. Sind die Events in den kleinen Plots schlecht erkennbar, kann der Nutzer jeden Plot vergrößert darstellen lassen (Abb. 4.15.a2). Hat der Nutzer ein Event gefunden, zu dem mehr Informationen benötigt werden, kann dieses angeklickt werden. So gelangt der Nutzer direkt zur Detailansicht (Abb. 4.15.a3,b4).

Hat der Nutzer sich für die hierarchische Darstellung entschieden (Abb. 4.15.b), so muss ein Topic-Channel (Kapitel 2.2) ausgewählt werden. Die Treemap *zoomt* in die nächst tiefere Ebene (Abb. 4.15.b2). Dort werden, die im Topic-Channel enthaltenen Events nach Termen gruppiert dargestellt. Der Benutzer wählt einen Term. Die Treemap *zoomt* wieder eine Ebene weiter. Es werden die Events, denen der gewählte Term zugewiesen wurde, chronologisch sortiert angezeigt (Abb. 4.15.b3). Zudem werden die Tweets und Zeitungsartikel der Events dargestellt. Diese sind ebenfalls chronologisch sortiert. Der Nutzer kann ein Event durch Klicken auswählen. So gelangt er zur Detailansicht des Events (Abb. 4.15.a3,b4).

Die Detailansicht (siehe Kapitel 4.3.2) dient dazu, die Informationen über die Events zu visualisieren. Mit diesen Informationen kann der Nutzer Rückschlüsse ziehen, wie ein Event den Report-Score zugewiesen bekam und warum es durch das Bewertungsmaß als Ausreißer identifiziert wurde.

4.5 Vom Abweichungsfaktor zur Clusterbildung

Durch die farbliche Kennzeichnung der Events entstehen zwei disjunkte Mengen. Diese sind in der Plotansicht als optische Cluster differenzierbar. Ein Cluster enthält alle Events, die ungefärbt sind. Das Zweite beinhaltet alle Events, die farblich markiert sind. Die ungefärbten Events, lassen sich als Events interpretieren, die einen plausiblen Report-Score zugewiesen bekommen. Damit entsteht ein Datensatz, der optisch gelabelte Elemente enthält. Da der Bereich der ungefärbten Events durch die Einstellung des Bereichs des AF durch den Benutzer vorgenommen werden kann, lässt sich das so erreichte Clustering auf jeden visualisierten Datensatz individuell anpassen. Durch Anpassen der Einstellungen kann optisch ein geeigneter Bereich gewählt werden. Durch die Größe der Cluster kann abgeschätz werden, wie viele Events als Ausreißer gelten. Diese Teilung ist in der Plotansicht sehr gut sichtbar.

Sei also der Bereich B mit b_b als minimale Grenze und b_e als maximale Grenze der vom Nutzer eingestellte Wertebereich für AF.

Sei also:

B =$[b_b, b_e]$,
e : ein Event,
E :$\{e_1, e_2, ..., e_n\}$ die Menge aller Events

wobei gilt:

$$Menge_{\text{in}} = \{e \in E | AF(e) \in [b_b, b_e]\} \tag{4.3}$$

und

$$Menge_{\text{out}} = \{e \in E | AF(e) \notin [b_b, b_e]\} \tag{4.4}$$

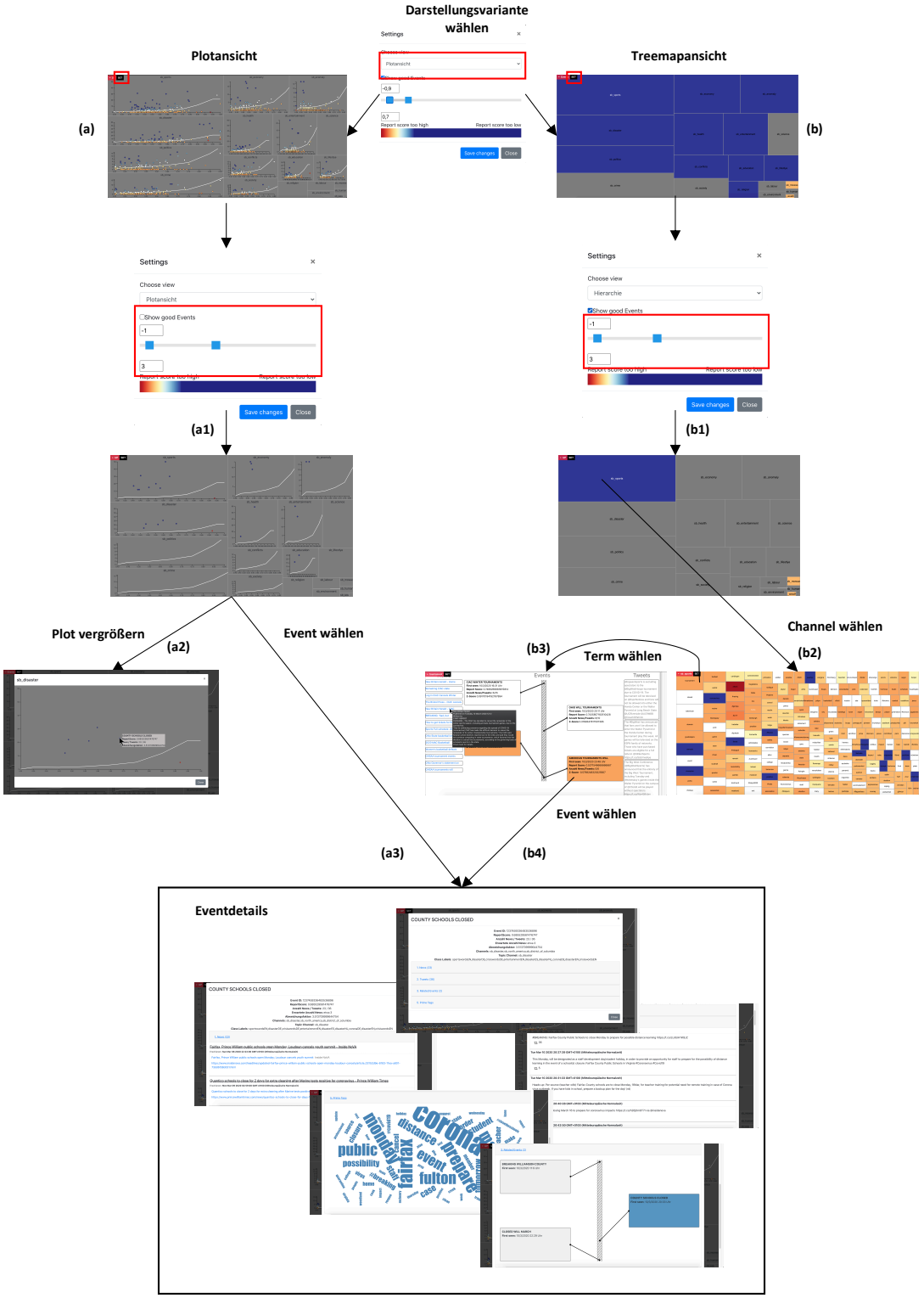

Abbildung 4.15: Exemplarischer Ablauf der Suche nach einem Event mit unplausiblem Report-Score.

5 Implementierung

Zur Evaluation der Forschungsfragen (siehe Kapitel 1.3) wird das in Kapitel 4 vorgestellte Konzept implementiert. Das nun folgende Kapitel beschreibt die Implementierungsarbeit mithilfe der Komponenten und deren Zusammenspiel. Es werden die genutzten Technologien und die Algorithmen des erstellten Systems aus technischer Sicht vorgestellt.

5.1 Architektur

5.2 Front-End - zur Visualisierung der Daten

Um eine hohe Flexibilität und eine damit einhergehende Portabilität zu ermöglichen, ist das Front-End als Webanwendung implementiert. Ein weiterer Vorteil dieser Umsetzung, ist die gute Verfügbarkeit von Drittanbieter-Software-Bibliotheken wie die Visualisierungsbibliothek von D3. Das Front-End ist in der Skriptsprache JavaScript implementiert. Die grafischen Elemente sind mit *HTML* und *CSS* erstellt.

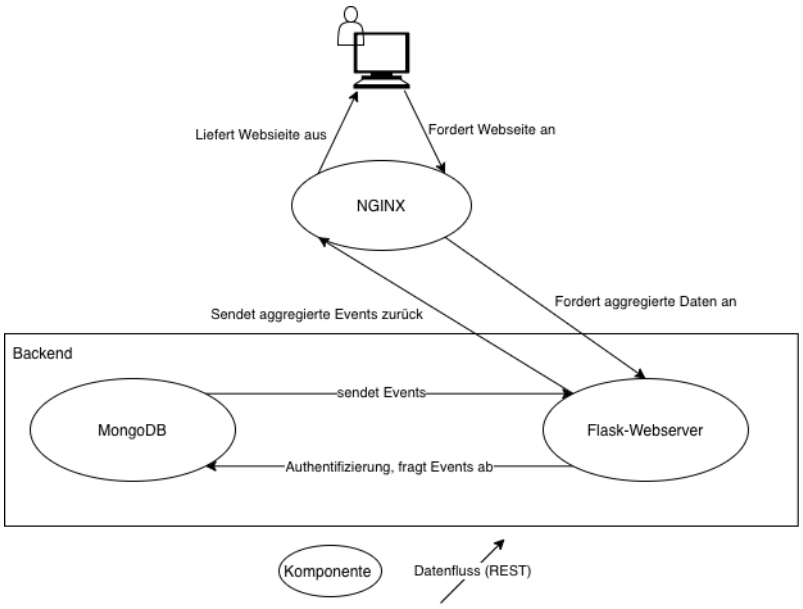

Abbildung 5.1: Architekturschaubild

5.2.1 D3.js - als Visualisierungsframework

Das JavaScript Framework „Data-Driven-Documents“[1] („D3“) ist eine webbasierte Software-Bibliothek für das interaktive Visualisieren großer Datenmengen. Mit D3 ist es sowohl möglich Daten in verschiedener Art und Weise anzuzeigen als auch interaktiv veränderbare Schaubilder zu erzeugen. Damit können bestimmte Bereiche hervorgehoben oder Datenbereiche gefiltert werden, ohne ein neues Diagramm erzeugen zu müssen. Diese Arbeit verwendet als Visualisierungsgrundlage das Treemap Layout und Scatterplots von D3. Dies sind nur zwei von vielen Visualisierungsmöglichkeiten, die D3 bietet. Des weiteren kommt D3 beim Manipulieren des „Document Object Model“ (DOM) zum Einsatz. Für diesen Zweck besitzt das Framework eine vereinfachte syntaktische Umsetzung im Code.

5.2.2 NGINX - als Webserver zur Auslieferung des Front-Ends

NGINX[2] ist eine modular aufgebaute Webserver-Software. Neben dem Ausliefern von statischen Webseiten bietet NGINX eine Reihe weiterer Module an. Die Hauptaufgaben dieser Module sind Techniken wie Load Balancing, Reverse Proxying, SSL und FastCGI. Für diese Arbeit wird NGINX für das statische Ausliefern des Frond-Ends verwendet.

5.3 Python Back-End - zur Verarbeitung der Daten

Der Back-End-Service ist in Python implementiert. Als Programmiersprache wurde Python gewählt, da diese sich sehr gut eignet um eine große Datenmenge effizient zu verarbeiten. Dies kommt insbesondere bei der in Kapitel 4.1 beschriebenen Vorverarbeitung der Daten zum Einsatz. Um die vorverarbeiteten Daten bereitzustellen, ohne die Programmiersprache ändern zu müssen, kommt der Python-Webserver *Flask*[3] zum Einsatz. Dieses, in Python geschriebene Webframework, bietet eine leichtgewichtige Implementierung eines Webservers. Eine REST-API (Representational State Transfer - Application Programming Interface) ist eine Programmierschnittstelle zum standardisierten Austausch von Daten in verteilten Systemen und wird in dieser Arbeit zur Kommunikation zwischen Back- und Front-End verwendet.

Das mit den vorgestellten Technologien umgesetzte Back-End ist zuständig für:

1. Das Entgegennehmen der Anfragen des Front-Ends
2. Das Abrufen der angeforderten Daten aus der Datenbank
3. Das Verarbeiten und Aggregieren der Daten zu einer hierarchischen Struktur
4. Das Ausliefern der Daten durch die REST-API an das Front-End

[1] https://d3js.org/
[2] https://www.nginx.com/
[3] https://flask.palletsprojects.com/en/1.1.x/

5.4 MongoDB - Datenbank für die Datenhaltung

Als Datenbank kommt die nicht relationale, dokumentbasierte Datenbank *MongoDB*[4] zum Einsatz. Diese eignet sich sehr gut, wenn sich die Struktur und die Felder des Datensatzes über Zeit ändern. So können bei einzelnen Datenpunkten Felder hinzufügt oder entfernt werden, ohne dass es einen Einfluss auf andere Datenpunkte hat. Das ermöglicht ein hohes Maß an Flexibilität. Der Quelldatensatz von ScatterBlogs ist ebenfalls in einer MongoDB gespeichert. Das verwendete Dokumentenmodell von MongoDB speichert die Daten in einem JSON-ähnlichen Format.

5.5 Verwendete Algorithmen

5.5.1 Berechnung der Farbe eines Events

Für die Berechnung der Farbe eines Events wird die divergierende Farbscala *d3.interpolateRdBu* von D3.js verwendet. Mittels der Funktion *d3.scaleDiverging(d3.interpolateRdBu)* wird eine Farbe innerhalb eines festgelegten Bereichs ermittelt. Der Bereich wird durch eine obere und eine untere Schranke definiert und berechnet sich wie folgt:

$$\text{Bereich} = [-Schranke, Schranke]$$

Wobei:

$$Schranke = \begin{cases} |min(AF)|, & \text{falls } |max(AF)| > |min(AF)| \\ |max(AF)|, & \text{sonst} \end{cases} \tag{5.1}$$

Sei:

$$min(AF) = \text{Der kleinste Wert des AF im gesamten Datensatz}$$
$$max(AF) = \text{Der größte Wert des AF im Datensatz}$$

Die Funktion *d3.scaleDiverging(d3.interpolateRdBu).domain([-Schranke,0,Schranke])* wandelt übergebene Werte zwischen *-Schranke* und *Schranke* in eine Farbe um. Als Basis dafür dient ein Farbverlauf von rot zu weiß zu blau. Rot repräsentiert dabei den kleinsten (*-Schranke*) und blau den größten übergebenen Wert (*Schranke*). Weiß wird zurückgegeben, wenn eine *0* an die Funktion übergeben wird. Wird ein Wert übergeben, der nicht innerhalb der Schranken liegt, wird die jeweils dunkelste Farbe übergeben. Blau steht für Werte größer als *Schranke* und rot für Werte kleiner als *-Schranke*. Liegt der AF des Events innerhalb des durch den Benutzer festgelegten „guten“ Bereichs (Slider) siehe Kapitel 4.3.3, so wird ebenfalls die Farbe Weiß zurückgegeben.

[4] https://www.mongodb.com/de

Algorithmus 5.1 getColor

procedure GETCOLOR(Event:e,Float:Schranke)
$color \leftarrow d3.scaleDiverging(d3.interpolateRdBu).domain([-Schranke, 0, Schranke]$
return color(e.AF)
end procedure

Algorithmus 5.2 computeMinMaxAf

procedure COMPUTEMINMAXAF
$E \leftarrow AllEventsasanArray$
$min \leftarrow highestPositiveNumber$
$max \leftarrow highestNegativeNumber$
for each $e \in E$ **do**
if $e.AF < min$ **then**
$min \leftarrow e.AF$
end if
if $e.AF > max$ **then**
$max \leftarrow e.AF$
end if
end for
return min, max
end procedure

Algorithmus 5.3 computeEventColor

procedure COMPUTEEVENTCOLOR(Event:e)
$minAF, maxAF \leftarrow computeMinMaxAf$
$schranke \leftarrow min(|minAF|, |maxAF|)$
$loverLimitSlider, upperLimitSlider \leftarrow setinsettinsmenuviaslider$
if $lowerLimitSlider < e.AF < upperLimitSlider$ **then**
$eventColor \leftarrow Weiß$
else
$eventColor \leftarrow getColor(e, schranke)$
end if
return $eventColor$
end procedure

6 Ergebnisse

Die prototypische Implementierung der in dieser Arbeit vorgestellten Konzepte wurde von zwei Domänenexperten im Rahmen eines Experteninterviews getestet. Die bewertenden Experten sind zwei der ScatterBlogs-Gründer. Diese sind besonders gut für die Einschätzung qualifiziert, da sie den visualisierten Datensatz bereits kennen und durch ihre mehrjährige Mitarbeit am Institut für Visualisierung an der Universität Stuttgart Erfahrung mit dem Thema Visualisierung haben. Zudem waren sie bei der Erstellung des Konzeptes beteiligt. Der Ablauf des Interviews, die Aufgabenstellung der Experten und das gegebene Feedback, werden in diesem Kapitel vorgestellt.

6.1 Aufbau und Ablauf des Experteninterviews

Die Experten waren anfangs via Screenshare zugeschaltet. Sie bekamen eine detaillierte Einführung in die Interaktionsmöglichkeiten mit dem Prototypen. Dabei konnten sie die Funktionen des Tools kennenlernen. Anschließend wurde das Screenshare getauscht und die Probanden konnten das Tool in ihrem eigenen Browser testen. Zum Testen wurde Google Chrome verwendet. Der verwendete Datensatz ist derselbe, der für die Entwicklung des Systems verwendet wurde. Ziel war es, schnellstmöglich die größten statistischen Ausreißer im Datensatz zu finden und daraufhin einzuschätzen, ob gefundene Events trotz der statistischen Abweichung zutreffend bewertet wurden. Dabei wurde erst das Event betrachtet, welches den höchsten positiven AF hat und anschließend das mit dem geringsten (höchsten negativen) AF. Bei nicht zutreffender Bewertung sollte durch die Visualisierung und die durch das Tool präsentierten Daten ermittelt werden, warum die Events abweichend von der persönlichen Experteneinschätzung bewertet wurden. Anschließend sollten zufällige Events aufgerufen werden, die durch die Visualisierung als „zutreffend bewertet“ markiert wurden. Für diese Events sollte auch bestimmt werden, wie die persönliche Wichtigkeit eingeschätzt wird. Der dabei entstandene Eindruck über die Aussagekraft des Bewertungsmaßes sollte wiedergegeben werden. Danach konnten die Experten ohne konkrete Vorgaben die Visualisierung weiter testen. Sie wurden gebeten ihre Eindrücke und Gedanken während des gesamten Interviews laut auszusprechen. Die Beobachtungen und Aussagen wurden mittels der Methode „Lautes Denken“ und einem Mitschrieb der ausgesprochenen Gedanken festgehalten. Zusätzlich wurden den Probanden folgende vorher definierte Fragen gestellt:

1. Zu der visuellen Gestaltung
    1.1 Ist das Konzept und die Funktionsweise des Tools leicht verständlich?
    1.2 Sind die Interaktionsmöglichkeiten und Funktionalitäten auch ohne eine Einführung intuitiv ersichtlich?
    1.3 Sind die Interaktionsmöglichkeiten mit der Visualisierung ausreichend? Wenn nein, was wäre noch wünschenswert?

1.4 Ist die Visualisierung geeignet, statistische Ausreißer effizient zu identifizieren und kann mit entsprechendem Vorwissen über die Bewertungsfunktion eine Hypothese aufgestellt werden, wie die Ausreißer entstanden sind? (Beantwortung der Forschungsfrage 1)

1.5 Fällt Ihnen spontan eine Funktionalität ein, die für die Zukunft in dieser Visualisierung wünschenswert wäre?

2. Zum Bewertungsmaß

2.1 Empfinden Sie das vorgestellte Bewertungsmaß als gut geeignet, um die Qualität des Report-Scores zu bewerten?

2.2 Haben Sie einen konkreten Vorschlag, wie die Bewertung verbessert werden könnte?

6.2 Ergebnisse des Experteninterviews

In diesem Abschnitt werden die Ergebnisse aus dem Expertenfeedback vorgestellt. Durch das Gespräch mit den Experten konnten Bereiche des Konzeptes sowie der Implementierung identifiziert werden, die verbessert oder erweitert werden können, um ein noch besseres Benutzererlebnis ermöglichen zu können. Dazu gehören sowohl die gegebenen Antworten auf die oben stehenden Fragen, als auch die während des Interviews eingefangenen Eindrücke der Experten.

Ist das Konzept und die Funktionsweise des Tools leicht verständlich?

Beide Experten fanden das Konzept leicht verständlich. Nach einer kurzen Einführung waren beide in der Lage, sicher und selbständig durch das Tool zu navigieren. Als besonders positiv wurde hervorgehoben, dass viele intuitive Interaktionsformen eingesetzt werden. Darunter zählen Hovereffekte so wie viele direkt klickbare Elemente, wie die Datenpunkte in den Scatterplots. Als ungünstig wurde erwähnt, dass bei einem Klick in die Treemap eine Navigation in die nächste Ebene ausgelöst wird. Auf einen Klick in die Plotansicht hingegen wird ein „Zoom In“ ausgeführt, da der Plot vergrößert in einem überlappenden Fenster dargestellt wird. An dieser Stelle wurde der Wunsch nach einer einheitlichen Vorgehensweise der Interaktionen geäußert.

Sind die Interaktionsmöglichkeiten und Funktionalitäten auch ohne eine Einführung intuitiv ersichtlich?

Beide Befragte beschrieben die Interaktion mit der Visualisierung als intuitiv und selbsterklärend. Kleine vorgeschlagene Änderungen waren das Hinzufügen von Tooltips und Beschriftungen im Einstellungsmenü. Dort ist nicht direkt ersichtlich, was durch das Verstellen des Sliders eingestellt wird. Ein beschreibendes Label wäre für eine produktiv eingesetzte Lösung hilfreich.

Sind die Interaktionsmöglichkeiten mit der Visualisierung ausreichend? Wenn nein, was wäre noch wünschenswert?

Die Interaktionsmöglichkeiten wurden allgemein als sehr gut beschrieben. Die Aufgaben, die mittels der Visualisierung erledigt werden sollen, können gut und effizient gelöst werden. Es steht immer eine, zum Use-Case passende Interaktion bereit. Eine, als besonders positiv wahrgenommene Interaktion, wurde in der Plotansicht identifiziert. Die Möglichkeit, direkt auf ein Event im Plot klicken zu können und so zu der Detailansicht des Events zu gelangen, wurde als sehr nützlich beschrieben. Des weiteren fiel positiv auf, dass die Events durch das Verschieben eines Sliders dynamisch gefärbt werden können. Es wurde dabei jedoch angemerkt, dass die gute Funktionalität noch um ein Begeisterungsmerkmal ergänzt werden könnte. Ein direktes Feedback während des Verstellens des Sliders wäre eine schöne Erweiterung (Brushing and Linking). So entfällt das Speichern und erneut Öffnen des Einstellungsmenüs. Ebenso wurde die Zoomfunktion der Plots positiv erwähnt. Werden Plots wegen fehlendem Platz auf dem Bildschirm nur noch sehr klein dargestellt, wird der Plot nicht mehr in die Kachel gezeichnet. Ein Klick in die Kachel öffnet den Plot vergrößert. Dies wurde als gut Idee hervorgehoben. Als sehr angenehm wurde empfunden, dass beim Überfahren der klickbaren Elemente, der Cursor eine Interaktionsmöglichkeit signalisiert. Daher ist die Bedienung des Tools auch ohne vorherige Einführung leicht erlernbar.

Ist die Visualisierung geeignet, statistische Ausreißer effizient zu identifizieren und kann mit entsprechendem Vorwissen über die Bewertungsfunktion eine Hypothese aufgestellt werden, wie die Ausreißer entstanden sind?

Die Befragten waren sich einig, dass die Visualisierung diese Frage klar mit „Ja“ beantworten kann. Die Visualisierung stellt die benötigten Informationen zur Verfügung, um die statistischen Ausreißer zu erkennen und zu beurteilen, warum die Events mit dem entsprechenden Report-Score ausgezeichnet wurden. Während des Feedbacks wurde eine Verbesserungsmöglichkeit der Hierarchiedarstellung sichtbar. Es wurde angemerkt, dass die Kacheln der Treemap auf der obersten Ebene entweder blau oder grau eingefärbt waren. Daraus lässt sich schließen, dass in keiner der Kategorien Events existieren, deren Report-Score zu hoch eingeschätzt wurde obwohl derartige Events vorhanden waren. Den Befragten zufolge ist die Visualisierung besonders gut geeignet, die Extrembereiche zu untersuchen. Aus der Visualisierung lassen sich Probleme mit verwendeten Metriken ableiten. Die Plotansicht ist besonders gut für diese Anwendung geeignet. Trotz der in diesem Kapitel genannten kritischen Anmerkungen waren sich die Experten einig, dass die Visualisierung für den vorgesehen Verwendungszweck sehr gut geeignet ist.

Fällt Ihnen spontan eine Funktionalität ein, die für die Zukunft in dieser Visualisierung wünschenswert wäre?

Als Erweiterungsmöglichkeiten wurden folgende Punkte genannt. In der Detailansicht der Events könnte der von ScatterBlogs erstellte Report zu dem betreffenden Event angezeigt werden. Dieser bietet eine weitere, schon aggregierte Sammlung der Daten über das Event. So lässt sich auch dort nach Informationen zu dem Event suchen. Zusätzlich wäre eine direkt aus der Detailansicht verfügbare, Übersetzungsfunktion der Tweets und Newsartikel eine Erleichterung der Benutzung. So müssen Tweets nicht erst kopiert werden, um sie übersetzen zu lassen. Diese Funktion ist besonders

praktisch bei Tweets, die weder in Deutsch noch Englisch verfasst sind. Auch für die Plotansicht wurden zwei Erweiterungen vorgeschlagen. Die Plotansicht zeigt alle, im Datensatz verfügbaren, Topic-Channel. Eine Filtermöglichkeit nach bestimmten Topic-Channel wäre daher wünschenswert. Wenn nur noch die den ausgewählten Topic-Channel entsprechenden Plots angezeigt werden, so können die besonders interessanten Themengebiete leichter verglichen werden.

Empfinden Sie das vorgestellte Bewertungsmaß als gut geeignet, um die Qualität des Report-Scores zu bewerten?

Die Befragten waren sich auch in der Beantwortung dieser Frage einig. Das Bewertungsmaß ist allein keinesfalls vollständig aussagekräftig. Sehr gut geeignet ist es jedoch für die Aufmerksamkeitssteuerung. Das in dieser Arbeit vorgestellte und visualisierte Bewertungsmaß ist sehr gut geeignet um die Extremfälle zu erkennen. Die Events, die einen sehr niedrigen Report-Score haben aber eine vergleichsweise hohe Anzahl an zugeordneten News, sind außergewöhnlich. So auch die Events, deren Report-Score signifikant zu hoch erscheint, im Vergleich zu der Anzahl Newsartikel, die zu diesem Event gefunden wurden. Derartige Events werden sehr zuverlässig durch das Bewertungsmaß aufgezeigt. Das Bewertungsmaß lässt jedoch keine direkte Aussage über die Qualität des Report-Scores zu. Ein Problem, welches als Grund für die beschränkte Aussagefähigkeit genannt wurde, war die unzuverlässige Kontrollgröße „Anzahl zugeordneter News“.

Haben Sie einen konkreten Vorschlag, wie die Bewertung verbessert werden könnte?

Als Verbesserung der Bewertungsmetrik wurde genannt, dass das Maß genauer ist, wenn die Anzahl der gefundenen News verlässlicher ist. Um die Aussagekraft zu steigern, müssen weitere Attribute als Kenngrößen mit in die Bewertung einbezogen werden. Zusätzlich soll durch weitere Filtermöglichkeiten in der Visualisierung eine präzisere Eingrenzung der Daten erfolgen. ScatterBlogs bietet den Kunden die Möglichkeit, erhaltene Reports zu bewerten. Daraus kann ein weiteres Bewertungsmaß erstellt werden.

Weitere Anmerkungen

- Der Verlauf der Erwartungswerte in der Plotansicht wirkt je nach betrachtetem Topic-Channel unterschiedlich. Dieser Eindruck entsteht, da die Achseneinteilungen sich dynamisch an den darzustellenden Daten orientieren. Dazu wurde angemerkt, dass die Vergleichbarkeit der Topic-Channel besser wäre, wenn die Skalierung und die Achseneinteilung der Plots einheitlich gewählt wird.
- Als allgemeine Schwäche der Visualisierung wurde die Skalierbarkeit genannt. Für das Visualisieren großer Datensätze wurden Bedenken bezüglich der Übersichtlichkeit geäußert. Die Plotansicht sei dann gut geeignet, wenn eine zeitliche Eingrenzung des Datensatzes vorgenommen wird.
- Als störend bei der Benutzung fiel auf, dass der Slider manchmal bei erneutem Aufrufen des Einstellungsdialogs zurückgesetzt wird und nicht mehr die vorher eingestellten Werte anzeigt.

- Die Checkbox zum Ausblenden ungefärbter Events, wurde als nicht selbsterklärend empfunden. Daher wurde das Hinzufügen eines Tooltips vorgeschlagen, welcher die Funktion der Schaltfläche näher erläutert.
- Die Umsetzung der dritten Ebene der Treemap wurde als sehr positiv hervorgehoben. In dieser Ansicht werden die zeitlichen Abstände der Events, welche den gleichen Term als beschreibendes Wort zugeordnet bekamen, sehr gut sichtbar.
- Ebenfalls positiv hervorgehoben wurde die leicht verständliche Navigation durch die Visualisierung. Besonders aufgefallen ist hierbei die Detailansicht der Events. Das Verwenden einklappbarer Elemente für die Bereitstellung der detaillierten Eventinformationen wurde als sehr benutzerfreundlich bezeichnet.
- Informationen zu Events und die Entstehung des Report-Scores konnten von beiden Experten bei deutsch- oder englischsprachigen Events sehr leicht nachvollzogen werden.

Allgemein war die Rückmeldung der Experten sehr positiv. Die Visualisierung wurde einstimmig als sehr hilfreich und intuitiv benutzbar bezeichnet. Es wurde der Wunsch geäußert, das Tool in einer produktiven Umgebung einzusetzen.

7 Diskussion der Ergebnisse

Das Feedback zum Prototypen, durch die Experten von ScatterBlogs, wird in dem nun folgenden Kapitel diskutiert. Entscheidungen werden begründet und Lösungsvorschläge für identifizierte Verbesserungsmöglichkeiten vorgestellt. Das Kapitel gliedert sich in zwei Abschnitte. Im Ersten wird das Feedback zum Visualisierungsteil der Arbeit diskutiert. Im zweiten Abschnitt geht es um das in der Arbeit vorgestellte und durch die Experten erprobte Bewertungsmaß für den Report-Score.

7.1 Visualisierung

Während des Expertenfeedback zeichnete sich das klare Bild ab, dass die Visualisierung sehr gut geeignet ist, die Ausreißer des Datensatzes zu identifizieren. Weiter wurde festgestellt, dass auch die Begründung der Ausreißer mittels der visualisierten Informationen leicht möglich ist. Für die Suche der extremen Ausreißer, hat sich die Plotansicht als besser geeignet herausgestellt. Gegenüber der hierarchischen Ansicht sind die Ausreißer direkt ersichtlich, da alle Events in einer Übersicht dargestellt sind.

Durch die Funktion, Events innerhalb eines AF-Bereichs ausblenden zu können, wird die Übersichtlichkeit weiter verbessert. Denn dadurch wird verhindert, zu viele für die aktuelle Analyse unwesentliche Informationen anzuzeigen. Eine gute Möglichkeit die Benutzerfreundlichkeit der Anwendung weiter zu erhöhen ist das Einstellungsmenü nicht mittels eines Dialogfeldes zu lösen. Eine Eingliederung der Einstellungen in die immer sichtbare Fläche der Visualisierung ist daher empfehlenswert. So lässt sich auch ein weiteres Begeisterungsmerkmal umsetzen. Aktuell muss eine getätigte Einstellung am Slider gespeichert werden, bevor das Ergebnis sichtbar wird. An dieser Stelle ist ein direktes Feedback wünschenswert. Ist der Slider immer sichtbar, so kann während des Verstellens schon auf die Eingabe reagiert werden. So passt sich die Visualisierung in Echtzeit den Eingaben an. Das Menü kann dafür in der ersten Zeile der Visualisierung positioniert werden. Abbildung 7.1 zeigt skizziert eine mögliche Variante. Geöffnet wird das Menüband jedoch erst nach dem Überfahren des „SET-Buttons“. Es schließt sich automatisch wieder, wenn der Cursor das Menüband verlässt. Bei geschlossenem Menü ist nur der „SET-Button“ sichtbar.

Die Achsen der Scatterplots werden individuell an die darzustellenden Daten angepasst. Ein solches Vorgehen führt dazu, dass die Themenbereiche schwerer vergleichbar sind. Der Effekt wird durch eine Stauchung oder Streckung der Linie, welche die Erwartungswerte repräsentiert, weiter verstärkt. Der Vorteil der dynamischen Skalierung ist die effizientere Nutzung des Platzes. Die Plots werden so nicht unnötig gestaucht. Der Vorteil einer einheitlichen Skalierung jedoch ist eine leichte Vergleichbarkeit der Plots. Damit steigt die Gefahr, dass die Plots deren Themenbereiche vergleichsweise sehr geringe Newszahlen beinhalten, schlechter differenzierbar werden, da diese sehr stark vertikal gestaucht werden. Die Datenpunkte werden vermehrt überlappend dargestellt, da dieselbe Anzahl an Punkten auf einer kleineren Fläche dargestellt werden muss. Die Punkte kleiner

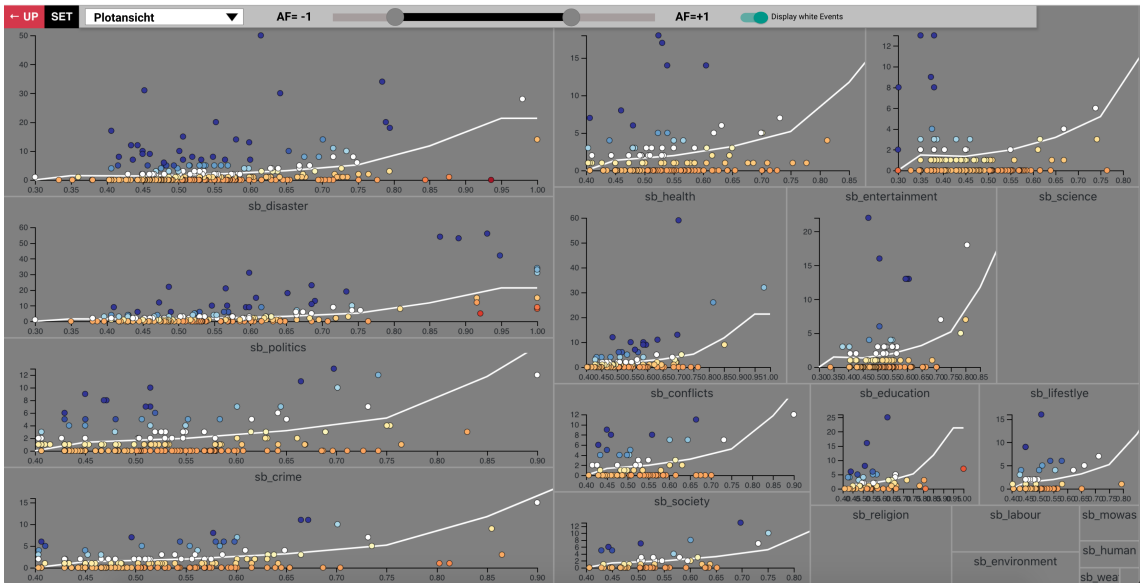

Abbildung 7.1: Plotansicht mit einem Verbesserungsvorschlag für das Einstellungsmenü

zu zeichnen, scheidet als Option aus, da sie für eine Interaktion mit dem Eingabegerät Maus nicht zu klein sein dürfen.
Um beide Varianten abzubilden, kann die dynamische Skalierung der Plots im Einstellungsmenü aktiviert oder deaktiviert werden. So kann je nach Anwendungszweck entschieden werden, welche Ansicht besser geeignet ist. Um die Übersichtlichkeit noch weiter zu verbessern, kann ein Konzept entwickelt werden, welches die Achseneinteilung so wählt, dass die Auswirkung großer Ausreißer auf die Lesbarkeit des Plots, minimal ist.

Die hierarchische Darstellungsvariante zeigt ihre Stärken, wenn themenbasierte Analysen gefordert sind. Innerhalb eines Themenbereichs zu navigieren, ist mit der Visualisierungsform Treemap deutlich leichter. Beziehungen zwischen Events sowie die zeitliche Abfolge verwandter Ereignisse, lassen sich in dieser Darstellung gut einsehen. Aus der zeitlichen Verteilung thematisch ähnlicher Events kann auf Zusammenhänge selbiger geschlossen werden. So kann hergeleitet werden, ob ein Event tatsächlich ein eigenständiges Ereignis abbildet, oder ein „Unterevent“ darstellt. Dieser Fall ist dann wahrscheinlich, wenn Events thematisch in die gleiche Kategorie einsortiert sind und ihre Erfassung zeitlich sehr nahe beieinander liegt. Aus diesen Erkenntnissen kann der Algorithmus für die Bewertung der Events weiter verbessert werden.

In der obersten Ebene der Treemap werden die Kacheln (Themengebiete) entsprechend dem Event, welches den größten AF hat, eingefärbt. In der Implementierung des Prototyps wird zuerst das Event mit dem größten AF innerhalb der Kachel bestimmt. Danach wird für dieses Event die Farbe, wie in Kapitel 5.5.1 beschrieben, berechnet. Diese Funktion liefert als Ergebnis die Farbe Weiß zurück, wenn das Event in den, mittels des Sliders, definierten Bereich fällt. Ist das der Fall, so wird die entsprechende Kachel der Treemap nicht eingefärbt. Ein unerwartetes Verhalten der Visualisierung tritt dann auf, wenn das Event mit dem absolut größten AF innerhalb des Slider-Bereichs liegt. Ein Event, welches einen absolut geringeren, aber außerhalb des Slider-Bereichs liegenden AF aufweist, wird für die Farbgebung der Kachel nicht beachtet. Da die Kachel nicht eingefärbt wird, entsteht der Eindruck, dass kein Event enthalten ist, welches einen AF aufweist, der außerhalb des akzeptierten

AF-Bereichs liegt. Die Farbe der Kacheln sollte sich daher stets nach dem „schlechtesten“ Event außerhalb des definierten Bereichs richten. Daher muss die Prüfung, ob das Event mit dem höchsten AF innerhalb des definierten Bereichs liegt, so lange erfolgen, bis ein Event gefunden wurde, das den größten Abweichungsfaktor innerhalb des Bereichs aufweist.

7.2 Bewertungsmaß

Das in dieser Arbeit vorgestellte Bewertungsmaß bemisst die Qualität des Report-Scores der Events anhand des Verhältnisses von Report-Score und Anzahl der zu einem Event gefundenen Newsartikel. Wie die Auswertung des Expertenfeedbacks zeigt, lässt sich die Qualität des Report-Scores nicht auf ein einzelnes Attribut abbilden. Um tatsächlich aufzuzeigen, wie gut die Bewertung zutrifft, müssen weitere Attribute der Events berücksichtigt werden.

Der errechnete Abweichungsfaktor setzt voraus, dass die Anzahl der gefundenen Artikel immer die vollständige Menge, der vorhandenen Artikel umfasst. In stichprobenartigen Tests wurde jedoch festgestellt, dass die Anzahl der gefundenen Artikel stark von der Sprache der Events und der Qualität der extrahierten Terme (Suchparameter) abhängig ist. Bei der Verarbeitung von Tweets, die weder in deutscher noch englischer Sprache verfasst sind, ist die Extraktion geeigneter Suchparameter erschwert. Nicht nur die Sprache allein erschwert die Google News Suche. Das in Tweets verwendete Vokabular weicht teilweise stark von der Ausdrucksweise der Presse ab. Suchparameter die aus den Tweets extrahiert werden, ergeben bei abweichend verwendetem Vokabular weniger oder andere Suchergebnisse. So wird die Basisgröße des Bewertungsmaßes stark beeinflusst. Ist die Ergebnismenge der Google News Suche unvollständig, so vergrößert sich die Distanz zum Erwartungswert. Das erzeugt fälschlicherweise den Eindruck, das Event, sei als zu wichtig eingestuft. Gleiches gilt für Events mit niedrigem Report-Score und vielen gefundenen News.

Der AF liefert jedoch sehr zuverlässig Aufschluss darüber, ob ein Event einen gänzlich unplausiblen Report-Score hat. Ein Event, mit einem sehr hohen Report-Score aber keinen gefunden News oder ein Event mit niedrigem Report-Score und vielen News, sollte in jedem Fall überprüft werden. Daher ist das Bewertungsmaß bestens für die Aufmerksamkeitssteuerung geeignet. Es zeigt an, wo das Verhältnis von Report-Score und die Aufmerksamkeit der Presse gänzlich unplausibel erscheinen.

8 Fazit

Diese Arbeit entstand in enger Zusammenarbeit mit zwei der Gründern von ScatterBlogs. ScatterBlogs macht es sich zur Aufgabe, aus Twitternachrichten, Ereignisse zu generieren und diese für die Unterstützung des Zivilschutzes zur Verfügung zu stellen. Da die Plattform Twitter einen sehr großen Datenstrom bereitstellt, werden Ereignisse aus aller Welt und von unterschiedlicher Wichtigkeit gesammelt. Um jene Ereignisse zu erkennen, die für die Arbeit diverser Hilfsorganisationen benötigt werden, müssen die „wichtigen Ereignisse“ herausgefiltert werden. ScatterBlogs vergibt daher für jedes gesammelte Ereignis eine Score. Dieser repräsentiert die Wichtigkeit des Ereignisses. Die Ereignisse werden zusammen mit dem Score und einigen Metainformationen gespeichert. Das Ziel dieser Arbeit war es eine Visualisierung zu entwickeln, die diesen Datensatz darstellt. Visualisiert werden sollte nicht nur der Datensatz an sich, sondern auch die Beziehungen zwischen Ereignissen sowie der Score, welcher die Wichtigkeit der Events beschreibt. Um die Qualität des Scores einschätzen zu können, wurde ein Bewertungsmaß entwickelt, welches den vergebenen Score und die entstandene Resonanz der Presse zu vergangenen Ereignissen, in Relation stellt. Dieses Bewertungsmaß wurde verwendet, um die Daten hervorzuheben, bei denen der Score nicht plausibel erscheint. So werden die Ereignisse farblich hervorgehoben, bei denen mittels des in der Arbeit vorgestellten Bewertungsmaßes eine Abweichung vom Erwartungswert festgestellt wird. Verglichen wird dabei die Anzahl der Nachrichtenartikel in klassischen Medien und der Sore der Wichtigkeit.

Ziel der Visualisierung war es schnell und einfach Ereignisse herauszufiltern, die weit vom Erwartungswert abweichen. Dafür werden zwei Implementierungen der Visualisierung vorgeschlagen. Ein Ansatz verfolgt das Konzept, die Ereignisse nach Themengebieten zu gruppieren. Diese gruppierten Daten werden in einer Treemap dargestellt. So kann der Datensatz explorativ analysiert werden. Der zweite Ansatz implementiert eine Übersichtsdarstellung, die für jedes Themengebiet einen Scatterplot der enthaltenen Ereignisse darstellt. Die Datenpunkte in den Scatterplots sind mit Interaktionsmöglichkeiten ausgestattet. Zum einen wird durch das überfahren der Punkte ein Tooltip angezeigt, der die Basisinformationen zu den Ereignissen enthält. Zum Anderen, werden durch Anklicken der Punkte alle bekannten Informationen des Ereignisses grafisch aufbereitet eingeblendet.

Eine große Stärke diese Konzeptes ist, die übersichtliche Darstellung eines ganzen Datensatzes mit der detaillierten Darstellung der Informationen zu einem einzelnen Datenpunkt, interaktiv so zu verbinden, dass der Nutzer leicht verständlich und intuitiv durch den Datensatz navigieren kann. Die Verbindung der beiden verwendeten Visualisierungsformen, schafft einen sehr breiten Einsatzbereich des Tools. Als besonders praktisch erwies sich das Konzept für das Finden derer Events, die durch einen unplausiblen Report-Score auffallen. Aus den durch die Visualisierung aufbereiteten Informationen zu diesen Events und deren Abweichungsfaktor kann auf Verbesserungsmöglichkeiten der Berechnungsverfahren des Report-Scores geschlossen werden. Der Abweichungsfaktor allein ist nicht aussagekräftig genug, um sich auf dessen Bewertung verlassen zu können. So ist es notwendig, dieses Bewertungsmaß weiter zu verbessern. Die prototypische Implementierung dieser

Visualisierung(en) wurde durch Domänenexperten getestet. Das Konzept und die Visualisierung kamen bei allen Testern sehr gut an. Der Prototyp soll die Basis für eine produktiv eingesetzte Umsetzung bilden.

9 Ausblick

9.1 Ausblick auf mögliche Verbesserungen und weiterführende Arbeiten

In dem abschließenden Abschnitt werden Anregungen für weiterführende Arbeiten gegeben. Der Ausblick besteht aus logischen Folgerungen, Verbesserungsmöglichkeiten und dem Blick auf weitere Forschungen, bei denen diese Arbeit als Basis dient.

9.1.1 Verbessern der Google News Suche mit anderen Parametern

Eine mit diesem Bewertungsmaß (Abweichungsfaktor) erkannte Abweichung bedeutet nicht zwingend einen nicht zutreffenden Report-Score. Während des Entwicklungsprozesses fiel auf, dass die gefundenen Abweichungen der Bewertung mancher Events nicht auf die Vergabe eines falschen Report-Scores zurückzuführen ist. Vielmehr, ist die gefundene Newsanzahl nicht so hoch, wie man es erwarten würde. Daher hat die Wahl der Suchparameter für die Google News Suche einen erheblichen Einfluss auf die Anzahl der gefundenen News und somit auch auf die Bewertung durch das eingeführte Bewertungsmaß. Werden durch das Ändern der Suchparameter mehr Artikel gefunden, so verringert sich die Abweichung vom Erwartungswert. Daher bietet es sich an, für eine weiterführende Arbeit diesen Einfluss genauer zu untersuchen. Es kann ein alternativer Ansatz entwickelt werden, mit dem die Suchparameter aus den Tweets extrahiert werden. Die Wahl eines anderen Werkzeugs für die Extraktion der Suchparameter und die Evaluation gegenüber dem aktuell genutzten System ist eine denkbare Vorgehensweise.

9.1.2 Finden eines anderen Bewertungsmaßes und Training eines Klassifikators

Das in dieser Arbeit vorgestellte Bewertungsmaß für die Qualität des Algorithmus, welcher den Report-Score der Events berechnet, basiert auf historischen Daten. Eine weiterführende Arbeit kann ein Bewertungsmaß vorschlagen, welches sich nicht an den Mittelwerten des vorhandenen Datensatzes orientiert, sondern andere Attribute der Events für die Berechnung heranzieht. Wird ein geeignetes Bewertungsmaß gefunden, so kann dieses als Basis für das Training eines binären Klassifikators verwendet werden. Dieser klassifiziert die Events abhängig von der Plausibilität des Report-Scores. Um Trainingsdaten für den Klassifikator zu erhalten, muss der bestehende Datensatz unter Verwendung des neuen Bewertungsmaßes manuell gelabelt werden. Ein daraus entstandener Klassifikator kann für automatisiertes Labelling eines anderen Datensatzes verwendet werden.

9.1.3 Trainieren eines Maschine Learning Modells mit verschiedenen Werten für den Abweichungsfaktor

In den Einstellungen kann eine Schranke für den AF definiert werden, bis zu welcher ein Event als zutreffend bewertet markiert wird. Durch Anpassen dieser Schranke, lässt sich der Datensatz in zwei Cluster einteilen. Das eine Cluster beinhaltet alle Events, deren AF innerhalb der Schranke liegt. Solche Events sind als „plausibel bewertet“ eingestuft. Das zweite Cluster enthält alle Events, deren AF außerhalb der Schranke liegt. Diese Events gelten als „nicht plausibel bewertet“. Nun wird eine Einstellung für die Schranke gesucht, bei der nur noch die gewünschten Events im Cluster der „plausibel bewerteten“ sind. Eine weiterführende Arbeit kann für besagte Events eine Exportfunktion implementieren, welche die entsprechenden Events in einen neuen Datensatz überführt. Der neue Datensatz kann für das Training eines maschine learning Ansatzes dienen, welcher lernt, den Report-Score zu vergeben. Spannend hierfür wäre es, mehrere Trainingsdatensätze mit unterschiedlichen Schranken zu erzeugen und die daraus resultierenden Ergebnisse des trainierten Modells miteinander zu vergleichen.

9.1.4 Verbessern der Skalierbarkeit

Die aktuelle Visualisierung ist darauf ausgerichtet, den Datensatz eines Tages darzustellen. Um einen größeren Datensatz analysieren zu können, müssen Änderungen vorgenommen werden. Die Anzahl der simultan darstellbaren Events ist durch den auf dem Bildschirm zur Verfügung stehenden Platz begrenzt. Daher muss für die Analyse eines größeren Datensatzes ein Konzept entwickelt werden, welches die Menge der gleichzeitig dargestellten Datenpunkte so eingrenzt, dass die Übersichtlichkeit erhalten bleibt. Der Einsatz einer geeigneten Filterfunktion ist eine Möglichkeit, dieses Problem zu adressieren. Eine weiterführende Arbeit kann daher eine andere Form der Visualisierung vorstellen, die es ermöglicht mehr Events gleichzeitig darzustellen. Alternativ kann auch ein Vergleich verschiedener Herangehensweisen vorgenommen werden. Es können verschiedene Filtermöglichkeiten oder Visualisierungsformen untereinander verglichen werden. Ziel dabei ist es, den Ansatz zu finden, welcher am besten geeignet ist, einen größeren Datensatz so abzubilden, dass die Funktionalität dieser Arbeit auch bei einem großen Datensatz zur Verfügung gestellt wird.

Literaturverzeichnis

[AGCH11] D. Archambault, D. Greene, P. Cunningham, N. Hurley. „ThemeCrowds: Multiresolution summaries of twitter usage“. In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. 2011, S. 77–84 (zitiert auf S. 10).

[BHW00] M. Bruls, K. Huizing, J. J. van Wijk. „Squarified Treemaps“. In: *Data Visualization 2000*. Hrsg. von W. C. de Leeuw, R. van Liere. Vienna: Springer Vienna, 2000, S. 33–42. ISBN: 978-3-7091-6783-0 (zitiert auf S. 7).

[BL07] R. Blanch, E. Lecolinet. „Browsing zoomable treemaps: Structure-aware multi-scale navigation techniques“. In: *IEEE transactions on visualization and computer graphics* 13.6 (2007), S. 1248–1253 (zitiert auf S. 9).

[BSH+10] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, E. H. Chi. „Eddi: interactive topic-based browsing of social status streams“. In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 2010, S. 303–312 (zitiert auf S. 9).

[BTH+13] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, T. Ertl. „ScatterBlogs2: Real-Time Monitoring of Microblog Messages through User-Guided Filtering“. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), S. 2022–2031 (zitiert auf S. 1, 11).

[CTB+12] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, T. Ertl. „Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition“. In: *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2012, S. 143–152 (zitiert auf S. 5, 11).

[DGWC10] M. Dörk, D. Gruen, C. Williamson, S. Carpendale. „A visual backchannel for large-scale events“. In: *IEEE transactions on visualization and computer graphics* 16.6 (2010), S. 1129–1138 (zitiert auf S. 10).

[DNK10] N. Diakopoulos, M. Naaman, F. Kivran-Swaine. „Diamonds in the rough: Social media visual analytics for journalistic inquiry“. In: *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE. 2010, S. 115–122 (zitiert auf S. 9).

[EF09] N. Elmqvist, J.-D. Fekete. „Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines“. In: *IEEE Transactions on Visualization and Computer Graphics* 16.3 (2009), S. 439–454 (zitiert auf S. 9).

[GF15] A. Guille, C. Favre. „Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach“. In: *Social Network Analysis and Mining* 5.1 (2015), S. 18 (zitiert auf S. 10).

[HHS19] R.-D. Hilgers, N. Heussen, S. Stanzel. „Korrelationskoeffizient nach Pearson“. In: *Lexikon der Medizinischen Laboratoriumsdiagnostik*. Hrsg. von A. M. Gressner, T. Arndt. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, S. 1389–1389. ISBN: 978-3-662-48986-4. DOI: 10.1007/978-3-662-48986-4_1763. URL: https://doi.org/10.1007/978-3-662-48986-4_1763 (zitiert auf S. 7).

[HLLE14] F. Heimerl, S. Lohmann, S. Lange, T. Ertl. „Word cloud explorer: Text analytics based on word clouds“. In: *2014 47th Hawaii International Conference on System Sciences*. IEEE. 2014, S. 1833–1842 (zitiert auf S. 22).

[LLM13] P. Lee, L. V. Lakshmanan, E. Milios. „Keysee: Supporting keyword search on evolving events in social streams“. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, S. 1478–1481 (zitiert auf S. 10).

[MBB+11] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, R. C. Miller. „Twitinfo: aggregating and visualizing microblogs for event exploration“. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2011, S. 227–236 (zitiert auf S. 10).

[MK10] M. Mathioudakis, N. Koudas. „Twittermonitor: trend detection over the twitter stream“. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010, S. 1155–1158 (zitiert auf S. 10).

[Shn96] B. Shneiderman. „The eyes have it: A task by data type taxonomy for information visualizations“. In: *Proceedings 1996 IEEE symposium on visual languages*. IEEE. 1996, S. 336–343 (zitiert auf S. 4).

[SP98] B. Shneiderman, C. Plaisant. „Treemaps for space-constrained visualization of hierarchies“. In: (1998) (zitiert auf S. 9).

[SW01] B. Shneiderman, M. Wattenberg. „Ordered treemap layouts“. In: *Information Visualization, IEEE Symposium on*. 2001, S. 73–73 (zitiert auf S. 6).

[Tho15] D. Thom. „Visual analytics of social media for situation awareness“. In: (2015) (zitiert auf S. 1).

[YKSJ07] J. S. Yi, Y. Kang, J. Stasko, J. Jacko. „Toward a Deeper Understanding of the Role of Interaction in Information Visualization“. In: *Visualization and Computer Graphics, IEEE Transactions on* 13 (Dez. 2007), S. 1224–1231. DOI: 10.1109/TVCG.2007.70515 (zitiert auf S. 3).

[YS09] X. Yan, X. Su. *Linear regression analysis: theory and computing*. World Scientific, 2009 (zitiert auf S. 13).

Alle URLs wurden zuletzt am 20. 09. 2020 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift