Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Interaktive Kennzeichnung großer multimedialer Nachrichten-Korpora

Johannes Messner

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Qi Han, Markus John, Kuno Kurzhals
Beginn am:	15. November 2017
Beendet am:	15. Mai 2018

Kurzfassung

Heutzutage werden in immer mehr Bereichen große Mengen an Informationen gesammelt und veröffentlicht. Nachrichtenagenturen weltweit veröffentlichen Nachrichtensendungen zu jeglichen Ereignissen. Das Auftreten vieler Ereignisse und Themen erstreckt sich dabei über einen längeren Zeitraum und somit bietet sich die Möglichkeit an, durch eine entsprechende Visualisierung den Lebenszyklus und die Dynamik der Themen zu Untersuchen. Allerdings ist hierfür eine Unterteilung der Nachrichtensendungen in die einzelnen Berichte und einer Kategorisierung dieser von Vorteil. Der Prozess diese zu kategorisieren ist jedoch sehr langsam und mühsam. Diese Arbeit befasst sich nun mit einem Visuellen Ansatz dem Benutzer eine schnelle und effektivere Lösung zur Kategorisierung und Kennzeichnung von solchen Nachrichten-Korpora bereitzustellen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ARD Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland. 43

Pegida Patriotische Europäer gegen die Islamisierung des Abendlandes. 43

SVM Support Vector Machine. 11

1 Einleitung

Heutzutage lässt sich eine eine große und stetig wachsende Ansammlung von Mitschnitten von Nachrichtensendungen bestimmter Nachrichtenagenturen weltweit beobachten. Diese Menge an Mitschnitten enthält bestimmte Ereignisse und Themen welche ein höheres Interesse mancher Agenturen wecken und sogleich auch über einen längeren Zeitraum auftreten können. Daher sind diese auch weit verstreut in verschiedenen Mitschnitten zu finden.

Eine entsprechende Visualisierung ermöglicht es den Lebenszyklus jener Ereignisse besser zu beobachten und zu untersuchen, um diesen besser über einen längeren Zeitraum zu verstehen. Zusätzlich könnte eine Analyse der Dynamik der Themen für bestimmte Gruppen von Nutzern von größerem Interesse sein.

Um dies zu ermöglichen ist es essentiell, dass jene Mitschnitte in bestimmte Kategorien eingeteilt werden können, welche den Bedürfnissen des Benutzers am besten entsprechen. Dies ermöglicht nun eine effektivere Analyse durch die verbesserte Struktur der Sammlung von Informationen.

Die manuelle Zuweisung diese Kategorien ist jedoch langsam und mühsam für den Benutzer, während einer automatisierte Zuweisung das Subjektive wissen fehlt um dies zuverlässig zu tun und erfordert somit eine ständige Kontrolle eines Benutzers.

Diese Arbeit befasst sich nun mit einem visuellen Ansatz eine schnelle und effektive Plattform zur Kennzeichnung großer Datenmengen zu bieten. Hierbei wird es dem Benutzer ermöglicht die Kennzeichnungen iterativ zu verfeinern während die Plattform selbst, anhand des Kennzeichnungsmusters des Benutzers, bereits Kennzeichnungen vorschlägt. Zusätzlich erhält der Benutzer eine Rückmeldung über die Auswirkungen seiner Aktionen auf den Prozess.

Indes wird ein besonders großer Wert darauf gelegt, dem Benutzer die wichtigsten Informationen zur Kennzeichnung jener Mitschnitte klar visuell vorzulegen, damit dieser möglichst schnell, die richtige Entscheidung treffen kann. Des weiteren soll auch die Möglichkeit, alle gewünschten Informationen zu einer Kategorie schnell zu finden, geboten werden.

Durch eine Analyse und Gewichtung der Worte innerhalb jeden Mitschnittes ist es möglich mit Hilfe einer Support Vector Machine (SVM) anhand der bereits kategorisierten Mitschnitte durch den Benutzer, eine Vorhersage über die Kategorie noch unbearbeiteter Berichte, mit einer bestimmten Wahrscheinlichkeit zu tätigen, welche dem Benutzer ebenfalls als Information bereitgestellt wird.

Gliederung

Die Arbeit ist in folgender Weise gegliedert:

Kapitel 3 – Grundlagen: Hier werden die Grundlagen der Funktionen, welche in dieser Arbeit verwendet werden, beschrieben.

Kapitel 2 – Verwandte Arbeiten befasst sich mit ähnlichen Arbeiten im dem Bereich der Visualisierung von Informationen.

Kapitel 4 – Konzept: Hier werden mehrere Konzepte vorgestellt, welche die Grundlage für eine Implementierung bilden und es wird erläutert warum, sich das entsprechende Konzept durchgesetzt hat.

Kapitel 5 – Implementierung zeigt die Umsetzung von der Theorie zur Praxis.

Kapitel 6 – Anwendungsbeispiel: Hier wird gezeigt mit welchen Schritten der Benutzer des Programms, eine Aufgabe Schritt für Schritt lösen kann.

Kapitel 7 – Diskussion erörtert die Lösung der initialen Problemstellung und dessen Qualität.

Kapitel 8 – Zusammenfassung und Ausblick fasst die Ergebnisse der Arbeit zusammen und stellt Anknüpfungspunkte vor.

2 Verwandte Arbeiten

In diesem Kapitel werden wir nun verwandte Arbeiten näher betrachten, die bereits ähnliche Aufgabenstellungen im Bereich der Visualisierung bearbeitet haben. Dies ermöglicht uns in dem Themengebiet einen besseren Einblick zu erhalten und dient zusätzlich als ein Orientierungspunkt für diese Arbeit in diesem. Zu Beginn wird die jeweilige Arbeit betrachtet und dann in Bezug zu unserer Arbeit gesetzt. Anschließend wird beschrieben wie das erarbeitete Wissen angewandt und für diesen Zweck angepasst wurde.

MediaTable [RWW10] ist ein Tool, welches einem Benutzer ein schnelles Betrachten und Kategorisieren von großen multimedialer Sammlungen ermöglicht. Hierzu wird eine Technik zur automatischen Analyse des Inhaltes mit einem, zur visuellen Kategorisierung großer Ergebnismengen optimierten, Interface verbunden.

In der Hauptansicht (Table Interface) werden die Videoaufnahmen, repräsentiert durch einen Frame und Namen, untereinander angeordnet. Zusätzlich werden die Metadaten, sowie die vom Benutzer zusätzlich hinzugefügten Daten in der entsprechenden Reihe, Links des Frames angeordnet. So entsteht die Form einer Tabelle und die Videoaufnahmen können durch das Auswählen spezifischer Kategorien angeordnet und sortiert werden. Metadaten sind die mit den Bildern gespeicherte Informationen, wie zum Beispiel Aufnahmedatum, Name, vom Benutzer festgelegte Label oder auch der Speicherort des Bildes. Zusätzlich wird links neben der Hauptansicht eine Gesamtübersicht aller Daten und deren Kategorien bereitgestellt und unter der Tabelle, in Form von 'buckets', werden die Kategorien dargestellt.

Um die Aufgabe, den Videoaufnahmen Kategorien zuzuteilen, effektiver lösen zu können, wird eine zusätzliche Ansicht bereitgestellt. In dieser Ansicht, dem 'grid preview', werden nur die Frames der einzelnen Segmente dargestellt. Diese können so schneller erkannt und von einander getrennt, den 'buckets' zugeteilt werden. Bereits zugeteilte Frames erhalten einen Rahmen in der Farbe des zugehörigen 'buckets'.

Somit ergibt sich folgender Ablaufplan, dargestellt in Abbildung 2.1, für die Kategorisierung:

In einem ersten Schritt werden durch visuelle Analyse der Daten, oder durch entsprechende Vorkenntnisse die Kategorien ('buckets') festgelegt. Hierbei ist es wichtig, dass für jedes Element mindestens eine Kategorie zutrifft. Diese kann durch eine zusätzliche Kategorie wie zum Beispiel 'irrelevante Elemente' garantiert werden. Nachdem die Kategorien festgelegt wurden, exploriert der Benutzer, die bereits von den Metadaten extrahierten Informationen, welche in den Spalten der Hauptansicht enthalten sind. Die dafür bereitgestellten Werkzeuge ermöglichen es dem Benutzer die Aufnahmen zu sortieren oder irrelevante Aufnahmen herauszufiltern.

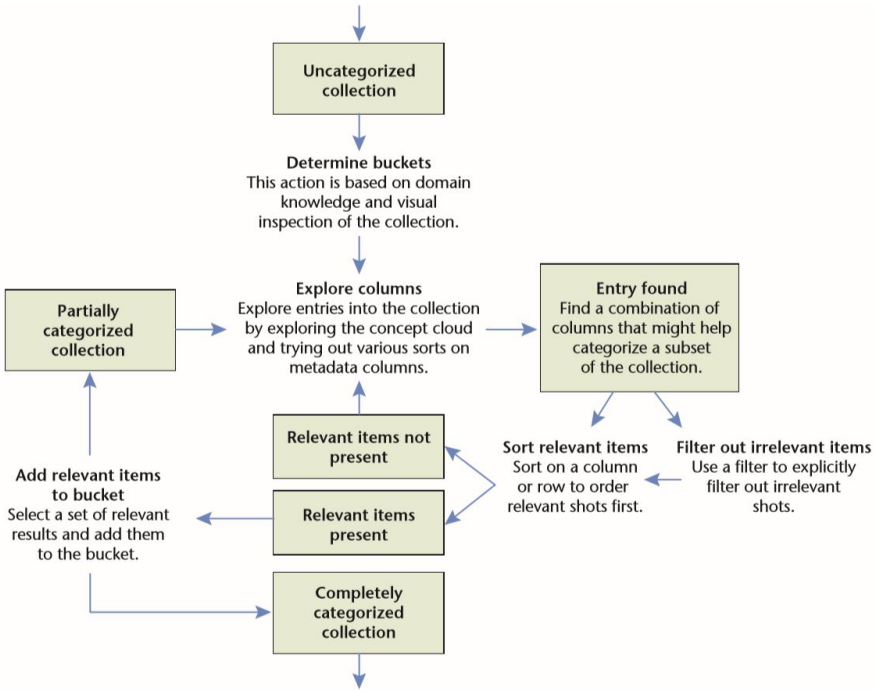


Abbildung 2.1: Typischer ablauf einer Kategorisierung mit MediaTable. Zuerst wird die Sammlung für mögliche Kategorien durchsucht. Durch sortieren der Listen können die Aufnahmen für die selben Kategorien gruppiert werden. Diese werden dann vom Benutzer den Kategorien zugeordnet. Dieser Schritt wiederholt sich [RWW10].

Nach einer entsprechenden Anwendung der Werkzeuge, kann der Benutzer somit eine größerer Menge an Aufnahmen, für eine Kategorie in der Ansicht heranholen und diese in eine Kategorie einteilen. Während nun weitere Aufnahmen exploriert werden, können die bereits gefüllten Kategorien die Daten zusätzlich filtern. Der Prozess wiederholt sich nun solange bis alle Aufnahmen in eine oder mehrere Kategorien eingeteilt ist, wobei jeder weitere Schritt, durch die bereits zugewiesenen Elemente deutlich verkleinert ist.

Ein weiterer Ansatz zur Kategorisierung von großen Sammlungen von Bildern ist die 'Interactive Categorizaition of Large Image Collecions' kurz ICLIC von van der Corput und van Wijk [CW16]. Ähnlich wie bei MediaTable ist das Ziel Bilder zu Analyseren und zu Kategorisieren. Im Vordergrund steht hier die Information aus den Metadaten der Bilder, sowie die Kategorisierung als Zusatzinformation zu den Metadaten. Diese Informationen werden dann so eng wie möglich in den Prozess eingearbeitet, damit der Benutzer diese Verbindungen durch Interaktion erkennen kann.

Der Arbeitsablauf von ICLIC (Abbildung 2.2) für die Kategorisierung wird in vier wichtige Teile unterteilt, diese stehen dem Benutzer in einer eigenen Ansicht zur Verfügung.

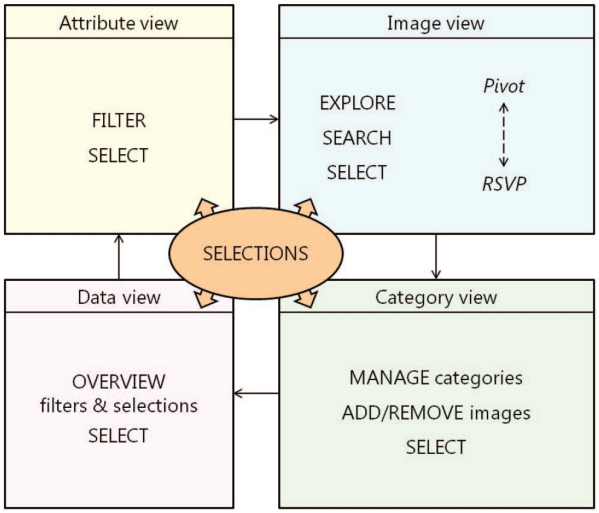


Abbildung 2.2: Schematische Darstellung des Arbeitsablaufs von ICLIC, mit dem Interface bestehend aus vier Ansichten. Pfeile zeigen den Ablauf, großgeschriebene Worte sind Teil des wichtigsten Ablauf der jeweiligen Ansicht [CW16].

- Anhand der Informationen aus den Metadaten der Bilder kann man Suchanfragen stellen, Filter anwenden oder bestimmte Attribute auswählen.
- Um die Bilder untersuchen zu können, müssen diese zum einen mit den Metadaten zusammen Strukturiert werden, wie auch lediglich durch ihren visuellen Inhalt durchsucht werden können.
- Weiterhin muss die Möglichkeit gegeben sein, die Bilder zu Kategorisieren, um diese letztendlich als Resultat abzuspeichern.
- Zusätzlich ist es wichtig, den Überblick des ganzen Prozesses zu behalten, um bestimmte Fragen, wie zum Beispiel 'Wie sind meine Kategorien über die Daten verteilt?' oder 'Wie groß ist der momentan Ausgewählte Bereich?'

Die Kategorien zählen bei diesem Prozess zu den Attributen. Jede neue Kategorisierung von Elementen verändert somit den Bereich der Attribute. Dies kann dann zu neuen Erkenntnisse führen, welche wiederum eine neue Kategorisierung erzeugen können. Das Resultat hierbei ist ein Zyklus der immer mehr Erkenntnisse bereitet.

Eine etwas andere Anwendung der Visualisierung bei der Analyse von Bildern wurde für PICTuReVis von van der Corput und van Wijk [CW17] verwendet. Sie haben eine Methode entwickelt Menschen innerhalb einer Masse an Bildern wiederzufinden, indem dem Benutzer Muster gezeigt werden um diese zu Vergleichen. Die Metrik für die Ähnlichkeit zwischen den Menschen basiert dabei auf Zeitlichen und räumlichen Aspekten der Bilder und interaktiver Visualisierung, um Analysten die Möglichkeit zu bieten, die gefundenen Muster im Detail zu betrachten und auszuwählen.

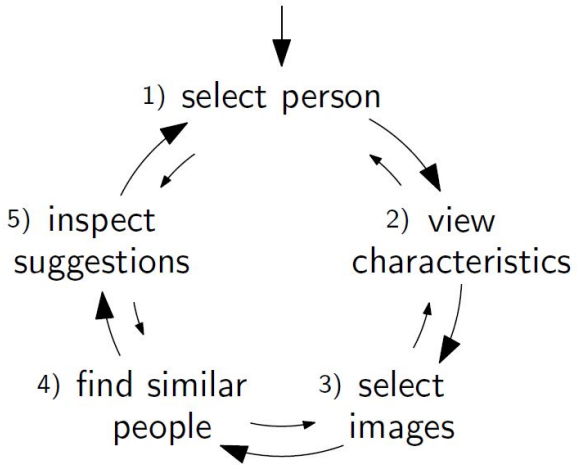


Abbildung 2.3: Arbeitsablauf eines Anwendungsfall. Große Pfeile zeigen die Richtung und kleine mögliche Korrekturen [CW17].

Der Ablauf einer solchen Arbeitsstellung (Abbildung 2.3) wurde hierfür in fünf Schritte unterteilt [CW17]:

1. Person auswählen: Von einem Analysten wird eine oder mehrere Personen ausgewählt.
2. Charakteristik betrachten: Das Auswählen einer Person impliziert eine Auswahl von Bildern die zu dieser Person gehören, wie auch die zeitlichen und räumlichen Daten, wie auch Konzepten. Durch eine Inspektion dieser werden dann die Relevanten Daten festgelegt, um einen Wert für die Ähnlichkeit zu bestimmen.
3. Bilder auswählen: Durch eine Suche anhand festgelegter Eigenschaften, wird nun nach Bildern gesucht die diesen Faktoren entsprechen.
4. Ähnliche Menschen suchen: In dem ersten Schritt wurde die Referenzperson bestimmt und eine Untermenge dessen in Schritt drei gefunden. Diese werden nun anhand ihrer Eigenschaften verglichen. Das Resultat kann dann nach den besten Treffen sortiert werden.
5. Vorschläge überprüfen: Anhand der zeitlichen Mustern, die mit den der Referenzperson übereinstimmen, kann der Analyst die Bilder auswählen und vergleichen, ob diese ein Wahrer Treffer sind oder nicht.

Ein Ansatz, Eye-Tracking Daten mit der Hilfe von automatischem Clustering zu Annotieren und in ein interaktives Kategorisierungs- und Analysesystem einzubinden, wurde von Kurzhals et al. [KHSW17] entwickelt.

Hierfür wurden die Videodaten in einem Vorverarbeitungsschritt in einzelne Bilder um den Blickpunkt extrahiert und gleichbleibende Sequenzen entfernt. So konnten einzelne unterschiedliche Daten aus den Videodateien extrahiert werden. Diese Bilder werden nach Ähnlichkeit automatisch zu Clustern zusammengesetzt.

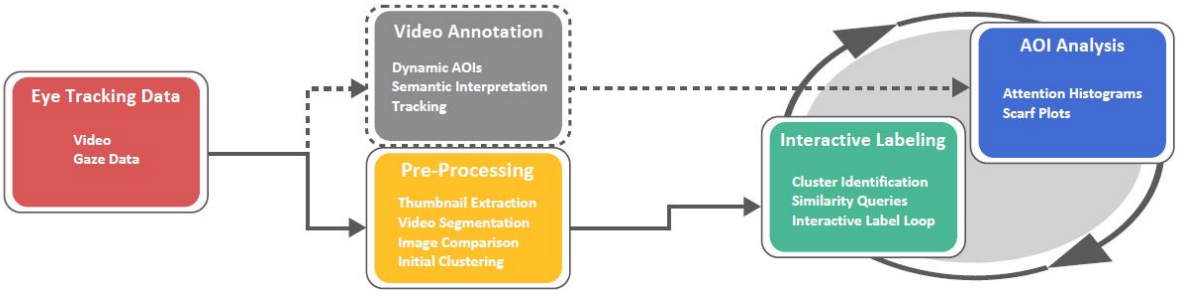


Abbildung 2.4: Analyseprozess von mobile Eye-Tracking Daten. Die Daten (rot) werden automatisch verarbeitet (gelb), und müssen so nur noch beschriftet werden (grün) um diese zu Analysieren (blau) im Gegensatz dazu der herkömmliche Ansatz (grau) [KHSW17].

Um die Daten zu analysieren und fehlerhafte Klassifizierungen zu korrigieren, wurde eine visuelle Analytik Umgebung erstellt. Hierfür wurden mehrere Ansichten bereitgestellt. Die Hauptansicht ermöglicht es dem Benutzer seine Analyse und eine Beschriftung durchzuführen. Eine zweite Ansicht für das Modifizieren, Erstellen, Inspizieren und Löschen von den Clustern, sowie ein Videospieler um die Videodaten direkt einsehen zu können. Der Prozessablauf wird in Abbildung 2.4 verdeutlicht. Die Daten werden in einem ersten Schritt verarbeitet, sodass der Benutzer lediglich eine leichte Korrektur und Verfeinerung der Daten durchführen muss, bevor diese analysiert werden können.

Das Ergebnis war eine deutlich effektivere Beschriftung im Vergleich zu den herkömmlichen Methoden dieser Art.

Der unterschied dieser Arbeiten zu dem Ansatz, welcher hier vorgestellt wird, ist die geringe Menge an Metadaten, die den einzelnen Elementen angeheftet sind. Für die Arbeit von Kurzhals et al. waren abgesehen von den Bilddaten keine Metadaten für eine Sortierung vorhanden, weshalb hier auf ein Clustering der Daten zurückgegriffen wurde. Die Möglichkeit auf eine große Anzahl von Daten in einem solchen visuellen Analytik Ansatz zurückzugreifen um den automatisierten Prozess iterativ zu verfeinern wird in den folgenden Kapiteln erläutert.

Maschinelles Lernen für eine Klassifizierung von Dokumenten, in eine relevante und nicht-relevante Menge zu teilen, um in einer Menge von Dokumenten eine bestimmte Informationen zu einer Suchanfrage zu finden wurde von Heimerl et al. [HKBE12] vorgeschlagen. Ziel hierbei war es die Qualität einer Suchanfrage auf eine Menge von Dokumenten mit der Unterstützung von maschinellem Lernen zu verbessern. Das Problem ist hier jedoch, dass ein entsprechender Klassifikator erst einmal trainiert werden muss.

Hierfür stellten Heimerl et al. mehrere Methoden vor, die es einem Benutzer erlaubt einen solchen Klassifikator zu trainieren. Durch einen Iterativen Prozess werden von dem Benutzer Suchanfragen gestellt und die vorgeschlagenen Suchergebnisse von diesem in relevant oder nicht relevant eingeteilt. Nachdem ein Dokument einer Kategorie zugeordnet wurde, wird

der Klassifikator neu trainiert und anschließend wird mit der Überprüfung und Zuteilung der Dokumente fortgesetzt.

Dieser Arbeit unterscheidet sich von Heimerl et al. insofern, dass ein solcher Prozess zur Kategorisierung vieler Unterschiedlicher Kategorien verwendet wird, um letzten Endes den Benutzer bei seiner Aufgabe zu unterstützen.

3 Grundlagen

Neben Visualisierung werden in dieser Arbeit noch andere Werkzeuge aus unterschiedlichen Themengebieten verwendet. Daher werden in diesem Kapitel die Grundlagen jener Themengebiete erläutert um somit das Verständnis der Arbeit zu erleichtern.

3.1 Maschinelles Lernen

Maschinelles Lernen befasst sich mit Systemen, die ohne explizite Anweisungen wie ein Problem zu lösen ist, automatisch lernen und sich durch Sammeln von Erfahrung ständig verbessern. Das System beginnt von einem bereitgestellten Grundwissen (Daten) zu lernen. Durch das Erkennen und Anwenden von Mustern auf neue Daten, soll das System immer bessere Entscheidungen treffen.

Maschinelles Lernen wird Heutzutage in vielen Bereichen angewandt, unter Anderem auch bei Klassifizierung von Texten und der Natürlichen Sprachverarbeitung. Die prominentesten Probleme auf die Maschinelles Lernen angewandt wird, sind nach Mohri, Rostamizadeh und Talwalkar [MRT12]:

- *Klassifikation*: Jedem Element eine Kategorie zuweisen, wie zum Beispiel die Klassifizierung von Bildern in Kategorien wie Landschafte, Tiere, oder Portraits und auch die Klassifizierung von Texten zu bestimmten Themen wie Politik, Wirtschaft, Sport oder Wetter.
- *Regression*: Die Zuweisung von Werten zu jedem Element. Ein Beispiel hierfür wäre die vorhersage von Börsenwerten.
- *Ranking*: Anhand von bestimmten Kriterien, Elemente anordnen. Beispiel hierfür sind Suchmaschinen die Ergebnisse meistens nach Relevanz anordnen.
- *Clustering*: Einteilung von Elemente in homogene Bereiche.

Das wichtigste praktische Ziel von Maschinellem Lernen besteht darin, eine genaue Vorhersage über Elemente zu treffen und effiziente sowie robuste Algorithmen zu entwickeln um solch eine Vorhersage zu generieren [MRT12].

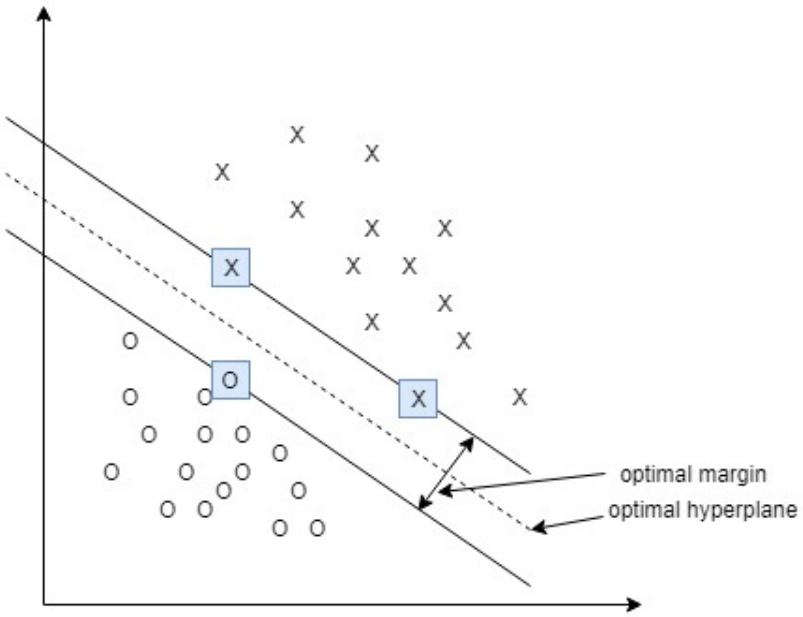


Abbildung 3.1: Beispiel für ein Separierungsproblem in einem 2 dimensionalen Raum. Der support vector (Blaue Kästchen) definiert die Grenze mit dem größten Abstand zwischen beiden Klassen [CV95].

Support-Vector Machine

Die SVM oder auch Support-Vector Network [CV95] ist ein Algorithmus des maschinellen Lernens für binäre Klassifizierungsprobleme.

Ziel des Algorithmus ist es eine Hyperebene zu finden, welche die Datenpunkte am besten in zwei Klassen unterteilt. Die optimale Hyperebene ist hierbei die Ebene mit dem maximalen Abstand zwischen den Vektoren beider Klassen (Abbildung 3.1). Nach Cortes und Vapnik [CV95] reicht es aus mit einer kleinen Menge an Trainingsdaten, den Support Vektoren, diese Grenzen zu bestimmen.

Um den Algorithmus auch anwenden zu können, wenn in einem zweidimensionalen Raum keine eindeutige Hyperebene gefunden werden kann, wird zu Beginn eine kleine Menge an Trainingsdaten in Form von Vektoren bereitgestellt. Diese werden in einem Mehrdimensionalen Feature Raum projiziert, in dem dann eine lineare Entscheidungsfläche konstruiert wird. Durch das hervorheben bestimmter Features eines Vektors kann dann die Genauigkeit noch weiter verbessert werden.

Der Vorteil der SVM ist hierbei dessen Flexibilität, der Algorithmus ist sehr robust und benötigt daher kein menschliches Eingreifen, selbst mit einem kleineren Trainingsset ist ein gutes Ergebnis erzielbar und es existiert immer genau eine Lösung. Der Nachteil der SVM ist, dass

Abbildung 3.2: Schritte um Dokument in seine normalisierten Einzelteile zu trennen.

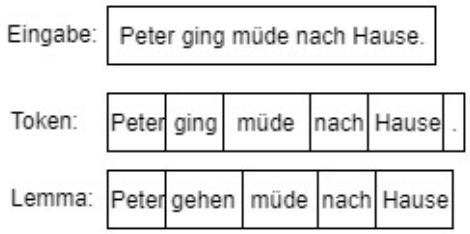


Abbildung 3.3: Beispiel einer Normalisierung eines Dokuments. Token werden extrahiert und anschließend normalisiert.

die Ergebnisse unterschiedlicher Elemente untereinander ungleich gewichtet werden, daher kann das Ergebnis nicht einfach als eine Parameterfunktion dargestellt werden.[AM08]

3.2 Information Retrieval

Bei dem Durchsuchen und der Verarbeitung des Textinhaltes von Dokumenten sind die Methoden der natürlichen Sprachverarbeitung von großem Vorteil.

Das initiale Problem bei dem Analysieren von Texten in natürlicher Sprache ist, dass Worte mit selben Stamm in unterschiedlichen Formen existieren können. Zum Beispiel sind die Worte gewinnen, gewann, gewonnen, gewinnst, gewinne das selbe Worte mit jeweils unterschiedlicher Form. Zusätzlich liegt das zu analysierende Dokument zuerst als eine einzelne Einheit vor, die in einem ersten Schritt in die einzelnen Bestandteile unterteilt werden muss. Um den Inhalt eines Dokuments analysieren und verarbeiten zu können, muss dieser zuerst mit den Methoden der natürlichen Sprachverarbeitung angepasst werden.

Wie in Abbildung 3.2 veranschaulicht wird das Dokument in einem ersten Schritt in einzelne Zeichenketten, in sogenannte Tokens geteilt. Dies entspricht im Allgemeinen einer Trennung nach Leerzeichen und Satzzeichen.

Bei der Lemmatisierung [JM] werden im nächsten Schritt die Worte auf eine Grundform reduziert oder normalisiert. Das bedeutet, dass Token mit der semantischen selben Bedeutung, zu einer einheitlichen Form angepasst werden, wie in dem Beispiel in Abbildung 3.3 veranschaulicht.

Tf-idf

Bei einer Suchanfrage werden die Dokumente nach dem entsprechenden Worten durchsucht. Nun kann es passieren, dass bei einer Suchanfrage in sehr vielen Dokumenten ein Treffer erzielt wird. Um den Benutzer nicht zu überfordern, ist es von Vorteil ein Ranking für die Suchergebnisse zu benutzen.

Ein solches Ranking ist das 'term frequency–inverse document frequency' (tf-idf) [SB88] Ranking. Es ist das beliebtesten Rankings weltweit[BGLB16] und setzt sich aus zwei Komponenten zusammen.

Die 'term frequency' $tf_{t,d}$ beschreibt wie oft ein bestimmter Term t in einem Dokument d vorkommt und die 'inverse document frequency'

$$idf_{t,D} = log(\frac{N}{1 + |d \in D : t \in d|}$$

mit N = # Dokumente

und $|d \in D : t \in d|$ = # Dokumente in denen der Term t vorkommt, besagt wie groß der Informationsgehalt des Terms ist.

Somit wird mit der Formel

$$tf\text{-}idf_{(t,d,D)} = tf_{t,d} * idf_{t,D}$$

das Gewicht jeden Dokuments bestimmt, welches bei der Suchanfrage einen Treffer erzielt hat. Die Dokumente können dann, entsprechend diesem Gewicht, sortiert werden. Je höher der Wert desto relevanter ist der Treffer.

3.3 Visualisierung

Visualisierung kann man als eine Projektion von rohen Daten in eine visuell verständliche Form für den Betrachter verstehen. Card, Mackinlay und Shneiderman [CMS99] veranschaulichen diesen Prozess mit ihrem Referenzmodell für Visualisierung, veranschaulicht in Abbildung 3.4.

Das Referenzmodell für Visualisierung (Abbildung 3.4) beschreibt eine Bewegung von Daten zu dem Betrachter. Die Daten werden dabei mehreren Veränderungen untergehen.

Zu Beginn existieren nur die rohen Daten ('Raw Data'). Der Begriff rohe Daten bezeichnet hierbei Daten ohne eine bestimmte Struktur. Durch eine oder mehrere Datentransformationen werden diese dann in Datentabellen ('Data Tables') umgewandelt.

Als Datentabellen liegen die Daten nun in einer strukturierten Form vor und sind somit einfacher in eine visuelle Form zu bringen. Die visuelle Projektion ('Visual Mapping') transformiert nun die Datentabellen in visuelle Strukturen, diese Strukturen kombinieren Markierungen

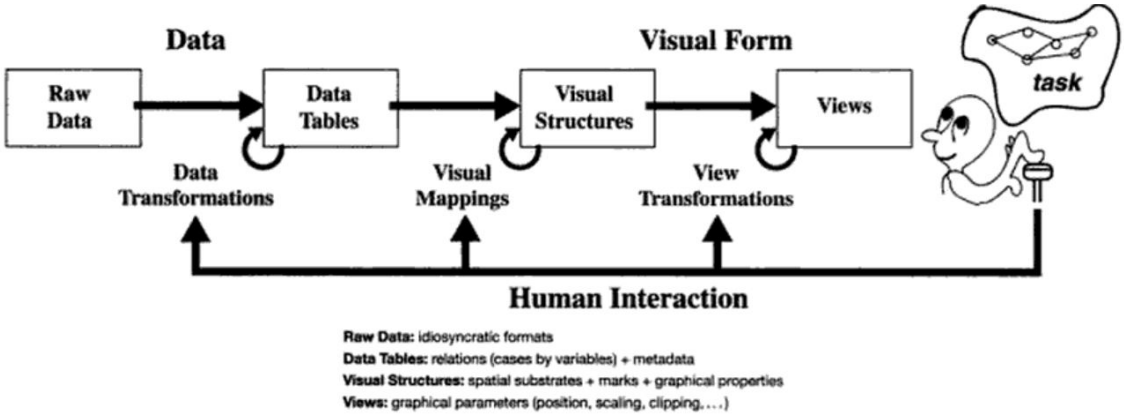


Abbildung 3.4: Referenzmodell für Visualisierung. Visualisierung als Projektion von Daten in eine visuelle Form, die menschliche Interaktion für visuelles Verständnis ermöglicht [CMS99].

und Grafiken in einem Raum. Eine ausdrucksstarke visuelle Transformation stellt dabei alle und nur die Daten, welche in der Datentabelle vorhanden sind. Es dürfen keine Beziehungen entstehen, die nicht in der Datentabelle existieren, um ausdrucksstark zu sein und für eine effektive visuelle Transformation muss diese von einem Mensch schnell verstanden werden.

In einer letzten Transformation, der Ansichtstransformation ('View Transformation'), werden nun die Ansichten der Strukturen erstellt, indem die grafischen Parameter wie Position, Skalierung und Begrenzungen spezifiziert werden. Durch eine Anpassung des Ortes der Ansicht ('Location Probing'), eine Veränderung des Fokus ('Viewpoint Controls') oder mit einer Verzerrung der Ansicht, um große Strukturen ganzheitlich zu sehen ('Disortion'), können somit noch zusätzliche Informationen aus der Datentabelle hervorgehoben werden.

Der Benutzer hat dabei Einfluss auf die Transformationen. Durch das Festlegen und Verändern der Parameter, die bei den Transformationen benutzt werden, um, zum Beispiel, die Menge der visualisierten Daten auf bestimmte Eigenschaften zu begrenzen, kann der Benutzer so die Ansicht verändern um diesen seiner Aufgabe anzupassen.

Design

Je mehr Informationen in einem Datenelement enthalten sind, desto größer ist die Gefahr den Betrachter mit Informationen zu überladen. Das Resultat davon ist, dass die Exploration von Daten kein bereicherndes Erlebnis für den Nutzer darstellt, sondern diesen überfordert. Um den Informationsgehalt eines Elements verständlich und interaktiv darzustellen hat Shneiderman [Shn03] eine Richtlinie in Form eines Mantras erstellt:

> Übersicht zuerst, zoomen und filtern, dann Details auf Anfrage [Shn03]

Weiterhin leitet Shneiderman sieben Aufgabenstellungen für eine Visualisierung von Informationen für eine Sammlung von Elementen mit mehreren Attributen ab.

Übersicht: Eine Übersicht der gesamten Sammlung.

Zoomen: Interessante Elemente näher betrachten.

Filtern: Uninteressante Elemente herausfiltern.

Details auf Anfrage: Ein Element oder eine Gruppe von Elementen auswählen und wenn nötig, mehr Details anzeigen.

Zuordnen: Verbindungen zwischen Elementen betrachten.

Historie: Eine Historie von Aktionen speichern, um diese Rückgängig zu machen oder zu verfeinern.

Extrahieren: Sammlungen und Parameter extrahieren.

Der Vorteil von visuellen Darstellungen laut Shneiderman gegenüber Textuellen, ist die menschliche Wahrnehmungsgabe visuelle Informationen schnell aufzunehmen und zu verarbeiten. Verbindungen zwischen Elementen können so in visuellen Darstellungen durch Farben, Nähe oder Linien zwischen einander dargestellt werden. Zusätzlich können Highlights verwendet werden um einzelne Elemente in einer Umgebung von Tausenden hervorzuheben.

Visual Analytics

Das Gebiet der visuellen Analytik befasst sich damit, den Vorgang der Daten- und Informationsverarbeitung ersichtlich zu machen, um über diesen Prozess analytisch reden zu können [KAF+08].

Durch die Masse an Informationen die heutzutage zur Verfügung stehen, ist das Problem, entsprechende Methoden und Modelle zu entwickeln um diese Daten in verwertbares Wissen umzuwandeln. Diese Methoden werden dann auf die Daten angewendet und das Ergebnis wird dem Nutzer vorgelegt. Oft agieren diese Methoden, indem diese sich das spezifische wissen während dem Prozess aneignen, um somit das Problem zu lösen.

Das Hauptproblem einer solchen Vorgehensweise entsteht vor allem dann, wenn das präsentierte Ergebnis falsch ist. In diesem Fall ist es besonders wichtig die Vorgehensweise der Methode zu betrachten. Dies ist allerdings nicht besonders durchschaubar, da der Prozess selbst, sein Wissen, welches er sich angeeignet hat nicht präsentieren kann.

Mit Hilfe einer Visualisierung solcher Prozesse wird es ermöglicht diese Prozesse genauer zu untersuchen, zu diskutieren und anschließend zu verbessern, anstatt nur ein Ergebnis präsentiert zu bekommen.

Die Lösung der visuellen Analytik solcher Probleme, erfolgt durch die Entwicklung von Technologien, welche die Stärke von menschlicher und elektronischer Datenverarbeitung kombinieren. Hierbei werden die distinktiven Fähigkeit beider Seiten optimal miteinander Kombiniert.

4 Konzept

In diesem Kapitel werden wir nun die Konzepte betrachtet, welche für die Lösung der Problemstellung der Arbeit entwickelt wurden. Zu beginn werden wir hierfür die Anforderungen erörtern, welche für die Umsetzung der Aufgabe benötigt werden. Dies beinhaltet sowohl die Möglichkeiten zur Informationsfindung und visuellen Repräsentation der benötigten Informationen wie auch die Werkzeuge für eine effektive und bereichernde Benutzung des Programms.

Die Aufgabe dieser Arbeit ist es eine Methode zu entwickeln, mit der eine große Masse an Nachrichtenberichten effektiv von einem Benutzer kategorisiert werden können.

4.1 Anforderungen

Das Ziel dieser Arbeit ist es, eine Plattform zur visuellen Präsentation von Daten zu erschaffen. Die präsentierten Daten werden von dem Benutzer interpretiert und anschließend in Kategorien eingeteilt. Die Kategorien können selbständig von dem Benutzer, wenn benötigt, erstellt werden.

A1: Kategorien können jederzeit erstellt werden.

Da es sich hier um eine sehr große Anzahl von Daten handeln soll, ist eine weitere Anforderung, dass dem Benutzer die relevanten Daten zur Interpretation klar präsentiert werden. Dadurch kann dieser sehr schnell die dargestellten Informationen über den entsprechenden Bericht erkennen und somit binnen kurzer Zeit eine Entscheidung über die Kategorie des jeweiligen Berichts treffen. Die präsentierten Informationen dürfen den Betrachter nicht überfordern.

A2: Übersichtliche und kompakte Darstellung für jedes Element.

Ein Nachrichtenbericht besteht immer aus einer Überschrift, welche den Inhalt kurz und direkt widerspiegelt, ein oder mehrere Bilder und den eigentlichen Inhalt. Der Inhalt ist meist gesprochen oder geschrieben je nach Art des Berichts. Während die Überschrift sehr schnell verstanden werden kann, ist das lesen des Inhaltes sehr zeitaufwändig, beinhaltet aber auch mehr Informationen, die eine entsprechende Kategorisierung eindeutiger machen als nur der Titel.

A3: Auf Anfrage muss die Möglichkeit bestehen, mehr Informationen über einem Bericht zu erhalten.

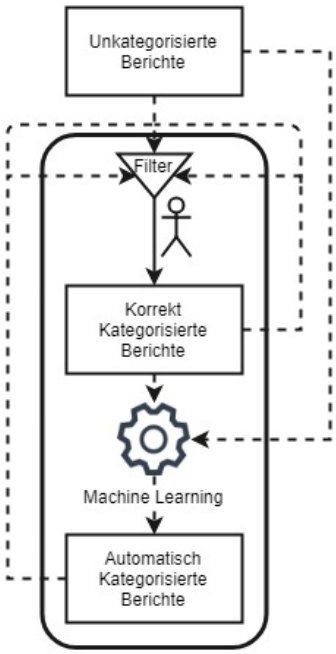


Abbildung 4.1

Neben den Bildinformationen aus Nachrichtensendungen existieren auch Metadaten zu diesen Berichten. Diese beinhalten das Datum, wie auch die Länge des Berichts. Zusätzlich dazu und im Gegensatz zu den Projekten, vorgestellt in Kapitel 2, stehen ebenfalls Untertitel oder Text zur Verfügung. Diese Untertitel enthalten sehr viel Informationen die man für eine genauere automatische Analyse verwenden kann um den Benutzer zu unterstützen.

A4: Durch Analyse und Verarbeitung der Metadaten soll der Benutzer maximal unterstützt werden.

Die nächste Anforderung an diese Arbeit ist es neben der Präsentation der einzelnen Nachrichten, dem Benutzer die Möglichkeit zu bieten die Daten zu durchsuchen um sich auf bestimmte Themengebiete, beziehungsweise Signalworte eines Themengebiets, zu beschränken. Dies ermöglicht zum einen eine erleichterte und schnelle Kategorisierung des Benutzers.

A5: Exploration der Daten anhand von Vorwissen und neuen Erkenntnisse, sowie **A6:** das Filtern von uninteressanten Informationen.

Die letzte Anforderung an diese Arbeit ist es dem Benutzer ein Feedback zu seiner Arbeit bereitzustellen. Die Aufgabe des Benutzers ist es, eine sehr große Ansammlung an Daten zu Kategorisieren. Dies erfordert einen entsprechend großen Zeitaufwand und Durchhaltevermögen. Indem dem Benutzer ein entsprechendes Feedback über seine bereits vollbrachte Arbeit, wie auch dessen Auswirkung auf die gesamte Ansammlung, gestellt wird, erhält dieser ein Gefühl des Fortschritts und der Errungenschaft. Weiterhin ist es wichtig dem Benutzer zu zeigen wie viel der Gesamtdaten ihm in diesem Moment präsentiert werden.

Abbildung 4.2

A7: Der Nutzer soll ein Feedback über seine Fortschritt bei der Kategorisierung erhalten und eine Übersicht die seinen Fortschritt verdeutlicht.

Die Idee für den Ablauf der Kategorisierung, dargestellt in Abbildung 4.1, ist dann folgendermaßen festgelegt:

Der Benutzer erhält Zugriff auf die benötigten Werkzeuge wie Filter, Suche sowie die Erstellung von Kategorien. Mit diesen Werkzeugen kategorisiert der Benutzer einen Teil der Berichte. Diese kategorisierten Berichte dienen dann einem Algorithmus des maschinellen Lernens als ein Trainingsset, der wiederum die restlichen, noch nicht kategorisierten Berichte anhand der Metadaten in Kategorien einteilt. Da Nachrichtenberichte und das Trainingsset sehr unterschiedlich ausfallen können, ist es besonders wichtig, dass der Benutzer diese Zuweisungen inspiziert und bestätigt oder verbessert, während er die Informationen des maschinellen Lernens bereits ausnutzt, um die Berichte zu Unterteilen und zu Filtern.

4.2 Visualisierung

Wie in 4.1 A2 beschrieben, ist es wichtig die Informationen eines Berichts so kurz wie möglich darzustellen. Durch den Titel eines Berichts wird dessen Inhalt bereits kurz und aussagekräftig zusammengefasst und enthält somit bereits Informationen darüber, was der Leser von diesem Bericht zu erwarten hat. Daher wird der Titel als erste Information zur Kategorisierung jeden Berichts festgelegt.

Um dem Benutzer eine Menge jener Titel zu präsentieren, um diese zu kategorisieren bieten es sich an, dies zum Einen mit einer Liste zu tun. Ein Listeneintrag enthält den Titel sowie die festgelegte Kategorie und Signalwörter die den Inhalt des Berichts wiedergeben (Abbildung 4.2). Somit erhält der Betrachter die wichtigsten Informationen auf einen Blick und kann somit schnell eine Kategorie bestimmen.

Eine Darstellung in Form einer Liste ist jedoch sehr monoton und eignet sich für diese Aufgabe nicht ausreichend genug, das Interesse des Benutzers für die Bewältigung der Aufgabe aufrecht zu halten. Eine bessere Möglichkeit hierfür ist es das Titelbild eines Berichts in den Mittelpunkt der Information zu rücken. Für eine derartige Darstellung ist eine Kachel optimal. Zum einen können so die einzelnen Berichte deutlich von einander getrennt dargestellt, wie auch die Informationen innerhalb einer Kachel interessant angeordnet werden. Bilder enthalten viel

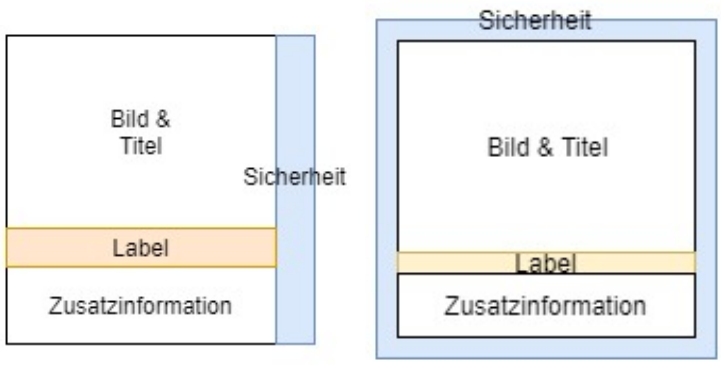


Abbildung 4.3: Layout für eine übersichtliche Darstellung der Nachrichtenberichte in Form einer Kachel.

Informationen, die von einem Betrachter sehr schnell aufgenommen und verstanden werden können, ohne diesen zu überfordern.

Da bei dieser Arbeit eine manuelle und automatische Kategorisierung der Elemente stattfindet ist es auch wichtig, zum einen, die eingeteilten Kategorien zu sehen und deren Art der Zuweisung, visuell von einander zu unterscheiden. Dafür wird der Begriff der Sicherheit eingeführt.

Die Sicherheit einer Zuweisung besagt, ob diese vom Benutzer selbst getätigt wurde oder automatisch statt fand. Im Falle einer automatischen Zuweisung sollte zusätzlich noch die Sicherheit der Zuweisung hinzugefügt werden.

In Abbildung 4.3 sind zwei mögliche Layouts für die Kacheln dargestellt. Die Sicherheit einer Zuweisung bestimmt darüber, welche Priorität dieser Bericht für eine Überprüfung des Nutzers hat. Daher soll diese Information am deutlichsten dargestellt werden und um den Nutzer nicht zu überladen mit Hilfe einer Farbcodierung verdeutlicht werden. Eine Möglichkeit (Abbildung 4.3b) wäre ein Rahmen um die gesamte Kachel, Titelbild mit der Farbe der zugewiesenen Kategorie und zusätzlichen Informationen darunter. Die zweite Option dargestellt in Abbildung 4.3a ist es die Sicherheit nur an einer Seite der Kachel anzuheften.

Kachel (a) hat mehrere Vorteile gegenüber Kachel (b). Die Fläche des Rahmen ist größer als die Fläche von einem einfachen seitlichen Balken. Somit stehen bei gleicher Dimensionierung der Kachel in (b) weniger Fläche zur Verfügung.

Wenn in einer Ansicht nun sehr viele Kacheln kompakt nebeneinander angeordnet sind (Abbildung 4.4) erkennt man, dass die Kacheln (a) deutlicher von einander unterscheidbar sind, während Kacheln (b), vor allem wenn sie die selbe Rahmenfarbe haben, schwieriger als einzelne Kachel zu identifizieren sind. Somit kostet es den Betrachter mehr Anstrengung mit einer solchen Ansicht zu arbeiten.

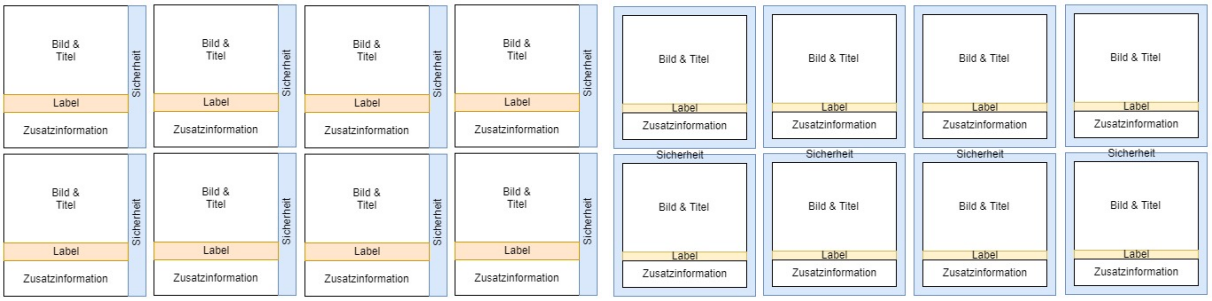


Abbildung 4.4: Nebeneinaderstellung vieler Kachel aus Abbildung 4.3 in einer Ansicht.

Entsprechend dieser Einsicht, wird für die Darstellung der Kacheln die Version von Abbildung 4.3a verwendet.

4.3 Analyse

Während die in 4.2 beschriebe Visualisierung für den Benutzer initial nur die wichtigsten Informationen bereitstellt, existiert eine sehr große Menge an Informationen des jeweiligen Berichts in dessen Inhalt. Hierfür bietet es sich an eine zusätzliche Analyse des Programms über jenen Inhalt zu verwenden (A4).

Zum einen wird daher ein Index über die Worte, welche in den Berichten existieren erstellt um diese schnell und zuverlässig durchsuchen zu können und zum Anderen, werden die Worte jeden Berichts als Token abgespeichert. Da der Benutzer selbst Kategorien erstellt und Berichte diesen Kategorien zuordnet, ist der nächste Schritt der Analyse eine Vorhersage über eine zukünftige Zuordnung anhand der in den Berichten enthaltenen Token.

Durch eine derartige Analyse ist es somit möglich zusätzliche kompakte Informationen bereitzustellen, wie zum Beispiel eine Wahrscheinlichkeit für eine vorhergesagte Kategorie eines Berichtes korrekt zu sein.

4.4 Interaktion

Die Interaktion mit den in 4.2 beschriebenen visualisierten Daten und dem Benutzer soll einfach und leicht verständlich sein. Daher findet die Interaktion des Zuteilens von Kategorien durch einen Rechtsklick mit einer folgenden Auflistung der Kategorien statt. Durch einen Linksklick der Maus auf die entsprechende Kategorie, wird diese daraufhin zugewiesen. Zusätzlich wird es dem Nutzer ermöglicht noch mehr Informationen über den ausgewählten Bericht, auf die selbe Weise wie das Zuweisen der Kategorie, in einer neuen Ansicht zu erhalten.

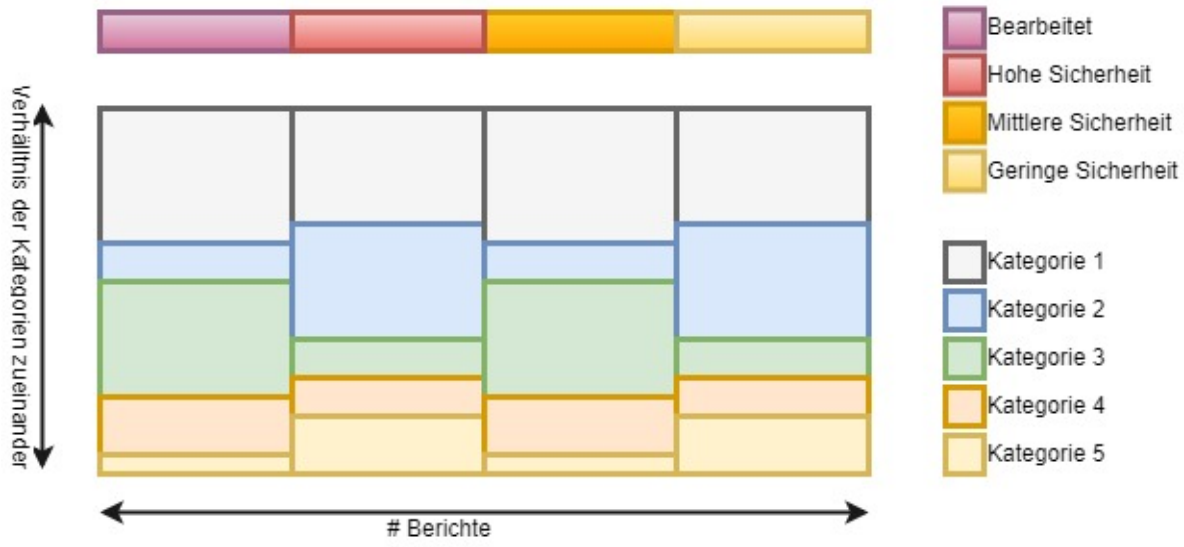


Abbildung 4.5: Die Graphik zeigt die Gesamtsituation der Sammlung an.

Neben der Zuweisung von Kategorien zu den Berichten wird es dem Benutzer zusätzlich ermöglicht, wie in 4.1 von A4 gefordert, die Berichte zu durchsuchen, die Suchergebnisse einzusehen und die Berichte nach Kategorien zu Filtern. Das Ziel dabei ist es ein systematisches Vorgehen des Benutzers bei dem Zuteilen von Kategorien zu ermöglichen.

Um fehlerhafte Daten beseitigen zu können, wird eine Löschfunktion für den Benutzer bereitgestellt.

4.5 Feedback

Um dem Benutzer ein Gefühl des Fortschritts und der Errungenschaft zu vermitteln wie in 4.1 A7 beschrieben, ist es wichtig, diesem ein sichtbares Feedback zur Verfügung zu stellen. Dies geschieht zum Einen direkt bei der Kategorisierung des entsprechenden Berichts und zum Anderen in einer Übersicht des gesamten Zustand aller Berichte.

Die Übersicht des Gesamtzustandes der Sammlung (Abbildung 4.5) zeigt dem Betrachter die Anzahl der kategorisierten Berichte im Verhältnis zu den noch zu Kategorisierenden. Die Breite des Schaubildes spiegelt die Gesamtmenge der, in der Sammlung enthaltenen, Berichte wieder. Eine Spalte repräsentiert eine Sicherheitsstufe und deren Breite das Verhältnis der Berichte mit dieser Stufe zu den Restlichen. Die Spalte ganz rechts repräsentiert dabei die höchste Sicherheitsstufe, also jene Berichte die von dem Benutzer kategorisiert wurden. Somit erhält der Betrachter ein konkretes Feedback über seinen Fortschritt bei der Kategorisierung sowie ein klares Ziel, welches es zu erreichen gilt. Des Weiteren zeigt das Schaubild innerhalb einer Spalte das Verhältnis der einzelnen Kategorien zueinander. Dies lässt dann auf eine Verteilung der Kategorien in der Sammlung schließen.

Durch die in 4.3 beschriebene Analyse der Daten und der Handlungen des Benutzers, ist es möglich die Auswirkungen auf die Analyse und eine veränderte Auffassung dessen zu erfassen. Diese Veränderung wird mit der Übersicht der Arbeit des Benutzers verknüpft. Dadurch erhält der Benutzer nicht nur ein Feedback über den stand seiner Arbeit, sondern auch über deren Auswirkung auf die Analyse des Programms.

Durch eine Interaktion mit der Grafik ist es möglich die Kategorien direkt einzusehen und noch zusätzliche Informationen zu erhalten. Diese Informationen beinhalten dann eine zusätzliche Analyse der gefilterten Berichte.

Das Ziel diesen Feedbacks ist es, das Interesse des Benutzers bei der Arbeit zu fördern und diesen gleichzeitig bei der Kategorisierung durch zusätzliche Interaktionen zu unterstützen.

5 Implementierung

In diesem Kapitel wird nun die Umsetzung der in Kapitel 4 vorgestellten Konzepte des Prototyps betrachten. Hierfür werden unter anderem die verwendeten Werkzeuge, die Verwendeten Datenstruktur und die Präsentation näher betrachten.

5.1 Verwendete Werkzeuge

Wie bereits in vorangegangenen Kapitel beschrieben, ist das Ziel dieser Arbeit eine effektive Methode für die Kategorisierung großer multimedialer Medien zu entwickeln. Hierfür wurde zur Veranschaulichung ein Prototyp entworfen.

Eines der Features der entwickelten Methode ist, dass zusätzliche Daten, neben den Metadaten der Elemente, zur automatischen Analyse zur Verfügung stehen. Hierfür wird eine Methode des maschinellen Lernens verwendet. Einige Programmiersprachen die sich hierfür eignen sind unter anderem Phyton und Java. Was Java allerdings von anderen Sprachen unterscheidet, ist ihre Plattformunabhängigkeit, weshalb diese als Programmiersprache für diesen Prototypen gewählt wurde.

Zur Erstellung der Benutzeroberfläche, die als Schnittstelle zwischen Programm und Mensch dient, wurde die Library JavaFx verwendet. Zusätzlich zu der Erstellung von Elementen wie Buttons und Textfelder, wurde JavaFX auch für die gesamte Visualisierung verwendet.

Für die Textverarbeitung, beziehungsweise die Verarbeitung der natürlichen Sprache der Dokumente wurde das Toolkit Stanford CoreNLP [MSB+14] verwendet. Diese Toolkit bietet eine einfache Umwandlung von Text zu Tokens und schließlich mit einem zusätzlichen Sprachpaket, eine Umwandlung zu den entsprechenden Lemmas in mehreren Sprachen, unter anderem auch für die deutsche Sprache.

Für die Textsuche wurde die Bibliothek Apache Lucene 7.3 [Fou] verwendet. Apache Lucene ist eine Bibliothek in Java, die mit geringem Arbeitsspeicheraufwand Indizes von Texten erstellt, welche sehr schnell durchsucht werden können, um sehr schnell ein Ergebnis zu bekommen.

Weiterhin wurde zur Implementierung des maschinellen Lernens Liblinear [FCH+08] verwendet. Liblinear ist eine Bibliothek zur linearen Klassifizierung.

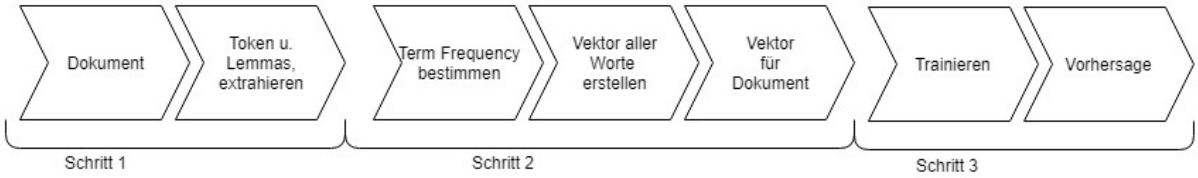


Abbildung 5.1

5.2 Analyse

A4 in Kapitel 4.1 fordert eine Analyse der Metadaten zur optimalen Unterstützung des Nutzers bei der Kategorisierung. In diesem Kapitel werden nun die einzelnen Schritte betrachtet, mit denen eine solche Analyse umgesetzt wurde.

Der Datensatz für das Testen des Prototyps besteht aus, bereits in einzelne Berichte unterteile, Tagesschauepisoden. Diese Episoden wurden mit Hilfe von Bilderkennung bei einer Änderung des Hintergrundes zerteilt. Für jeden Bericht wurde dabei der Start- und Endframe sowie das Datum, der Titel, die Framenummern und die Untertitel in einer csv-Datei gespeichert.

Beim Einlesen des Datensatzes in den Prototypen wird für jeden Bericht ein Objekt erstellt und dieses dann in einer Hashmap-Datenstruktur mit einem Schlüssel, bestehend aus einer Verknüpfung von Datum und Framenummer, gespeichert. Somit können die Objekte schnell gefunden werden.

In einem nächsten Schritt werden nun die Untertiteltexte aller Berichte gemäß Abbildung 5.1 verarbeitet. Hierfür werden diese in einem ersten Schritt durch Anwendung der Methoden der Stanford CoreNLP Bibliothek in einzelne Sätze unterteilt. In einem nächsten Schritt werden dann die zusammenhängenden Abfolgen von Zeichen in jedem Satz als einzelnes Token gespeichert. In einem letztem Schritt werden nun die Token mit der entsprechenden Methode Lemmatisiert, also in das entsprechende Lemma umgewandelt.

Nachdem ein Token in ein Lemma umgewandelt wurde, wird dieses Lemma in einer HashMap gespeichert. Der Schlüssel zum wiederfinden des Lemmas ist dabei das Lemma selbst und der Wert zu dem Schlüssel ist zu Beginn 1. Wenn nun ein Token zu einem Lemma umgewandelt wird, welches sich bereits in der HashMap befindet, wird lediglich der Wert des Eintrages um 1 inkrementiert. Somit erhält man am Ende eine HashMap, in der die 'term frequency' zu jedem Lemma des Berichts enthalten ist. Die HashMap wird dann in dem Objekt des entsprechenden Berichts gespeichert.

Da für diesen Prototypen mit einem statischen Datensatz gearbeitet wurde und um Rechenzeit bei dem Trainingsvorgang zu sparen, werden in dem zweiten Schritt von Schritt 2 (Abbildung 5.1) bereits alle Lemmas aller Berichte kombiniert. Wenn dies nicht der Fall ist, reicht es aus nur die Lemmas des Trainingssets zu kombinieren, dies muss dann allerdings bei jedem neuen Trainingsvorgang wiederholt werden.

Die Lemmas aller Berichte werden nun in ein Set eingefügt, sodass keine Duplikate entstehen. Dieses Set wird dann in einem nächsten Schritt in einen Array, den Vektor aller Worte, umgewandelt, um jedem Lemma ein statischen Index zuzuweisen. Gleichzeitig dazu wird auch die Anzahl der Dokumente gezählt in denen jedes Lemma vorhanden ist um somit einen Array zu erstellen, in dem für jeden Index eines Lemmas die Anzahl von Dokumenten gespeichert ist, welche dieses Lemma beinhalten (Feature Array).

In dem letzten Schritt von Schritt 2, werden nun die Vektoren der einzelnen Dokumente erstellt. Hierfür wird für jedes Lemma der entsprechende Index aus dem Vektor aller Worte ermittelt. Zusätzlich dazu wird dann mit der term frequency, der Anzahl aller Dokumente und dem Wert aus dem Feature Array des entsprechenden Lemmas das Tf-idf-Maß (siehe Kapitel 3.2) berechnet. Der Vektor des Dokuments beinhaltet dann für jedes Lemma dessen Index, das entsprechende Tf-idf-Maß und wenn bereits Klassifiziert, die Kategorie.

In Schritt 3 (Abbildung 5.1) werden nun alle Vektoren der Dokumente des Trainingssets an die Methode 'Train' von der Bibliothek Liblinear übergeben. Dokumente des Trainingssets sind alle Dokumente die vom Benutzer Kategorisiert wurden. Somit wird ein Modell für die Vorhersage der noch nicht Kategorisierten Dokumente erstellt, welches dann in einem letzten Schritt in der Methode 'Predict' der Bibliothek, zusammen mit den Vektoren der noch nicht Kategorisierten Dokumente, eine Vorhersage über deren Kategorien trifft.

Für die Vorhersage eines Dokumentes, wird für jede mögliche Kategorie ein Wert anhand der enthaltenen Worte und deren Gewichtung für die jeweilige Kategorie berechnet. Die Kategorie mit dem höchsten Wert entspricht dann der Vorhersage. Um noch mehr Informationen für den Benutzer bereitzustellen werden hier nun zusätzlich zu der besten Kategorie, noch die zwei nächsthöchsten Kategorien und deren Werte gespeichert um diese in die Visualisierung einzubinden.

Suche

Für die in Kapitel 3, **A5** geforderte Exploration der Daten anhand von Vorwissen bietet diese Applikation unter anderem eine einfache Stichwortsuche an.

Diese Suche wurde wie bereits in 5.1 erwähnt mit Hilfe von Apache Lucene implementiert. Hierfür wird ein Objekt Indexer aus der Bibliothek initialisiert, welchem dann die Dokumente übergeben werden. Dieses Objekt erstellt und speichert dann einen Suchindex.

Mit Hilfe eines Searcher Objekts der Lucene Bibliothek können dann Suchanfragen auf den anfangs erstellten Suchindex gestartet werden. Die Suchergebnisse werden dann nach einer Tf-idf Gewichtung geordnet und sind dann in der Form eines ScoreDoc abrufbar. Dieses Objekt beinhaltet dann den Namen beziehungsweise die Identifikation des ursprünglichen Dokuments und dessen Wertung.

Zusätzlich zu einer reinen Suchfunktion enthält Lucene eine Methode die es ermöglicht Dokumente nach Ähnlichkeit zueinander zu finden und zu sortieren. Diese Methode ist ebenfalls in

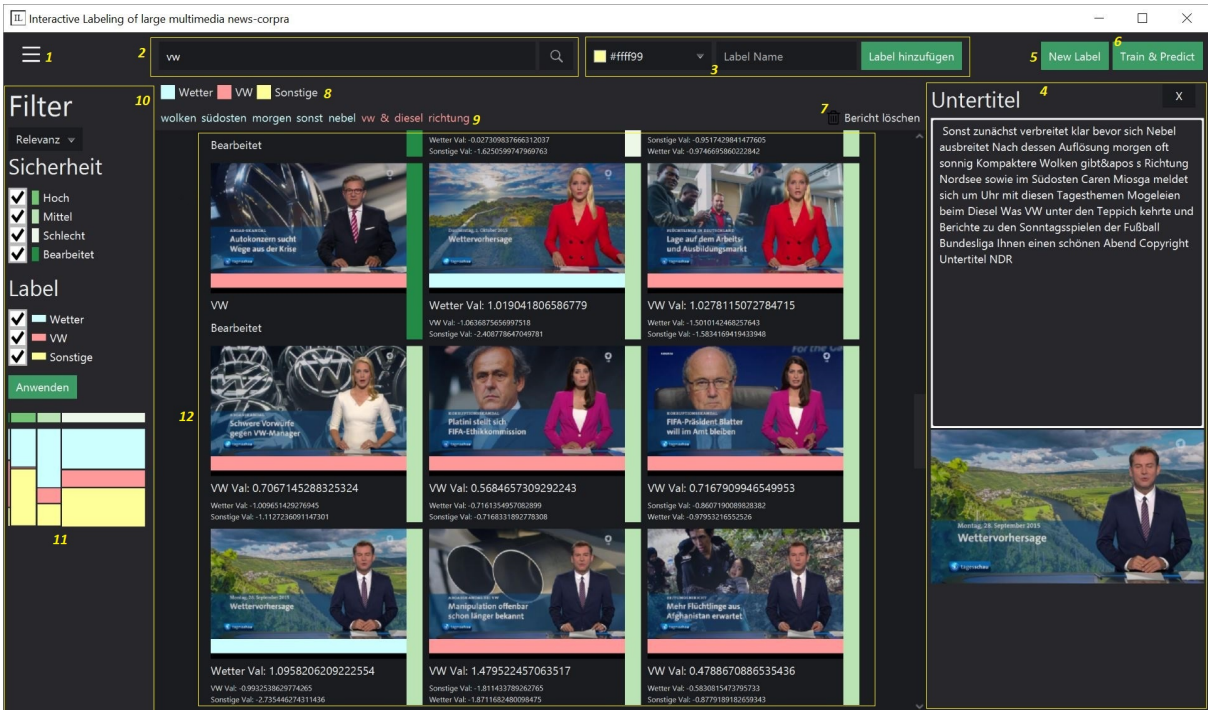


Abbildung 5.2: Das gesamte Layout der Applikation. (1) Button zum Schließen und Öffnen des Filters, (2) Stichwortsuche, (3) Eingebe um Kategorien zu erstellen, (4) Detailansicht, (5) Button zum Öffnen/Schließen von (3), (6) Button um neue Vorhersage zu starten, (7) Löschfunktion von Berichten, (8) Legende der Kategorien, (9) relevante Stichworte eines Berichts über dem sich die Maus befindet, (10) Filter, (11) Übersicht, (12) Hauptansicht mit der Visualisierung der Berichte.

dem Searcher Objekt implementiert und benötigt als Eingabe die Nummer des Dokuments im Suchindex, weshalb diese ebenfalls für die gefundenen Objekte gespeichert werden.

5.3 Benutzeroberfläche

Nachdem in Abschnitt 5.2 die Implementierung der Hintergrundaufgaben betrachtet wurde, wird sich nun dieser Teil der Benutzeroberfläche widmen.

Das Layout der Benutzeroberfläche bildet eine Border Pane. Diese Teilt die Ansicht in 6 Teile auf. Im Zentrum steht die hier die Auflistung der einzelnen Kacheln, angeordnet mit einem Gitter, der obere Teil der Border Pane bildet die Toolbar. In dem linken Teil befindet sich die Filteransicht und auf der rechten Seite die Detailansicht. Der untere Teil der Border Pane wird nicht genutzt.

Toolbar

Im oberen Teil der Applikation befindet sich die Toolbar. Hier befinden sich die wichtigsten Werkzeuge für den Kategorisierungsprozess.

In Abbildung 5.2 (1) befindet sich ein Button. Dieser öffnet oder schließt die Filteransicht. Rechts daneben befindet sich dann die Texteingabe für die Stichwortsuche (2). Die Suche selbst wird durch das Bestätigen der 'Enter'-Taste oder durch das Drücken des Buttons rechts neben dem Textfeld gestartet.

Um eine neue Kategorie zu erstellen benutzt man den Button (5) 'New Label', diese öffnet dann die Möglichkeit eine Farbe und einen Namen in (3) einzugeben. Nachdem eine Farbe und ein Name gewählt wurden, wird die Kategorie mit dem Button 'Label hinzufügen' erstellt und der Legende der Kategorien (8), mit der zugehörigen Farbe angehängt. Die Legende ist immer Sichtbar, sodass die vergebenen Kategorien, dem Benutzer schnell ersichtlich werden. Entsprechen der Anforderung **A1** aus Abschnitt 4.1 kann jederzeit eine neue Kategorie hinzugefügt werden.

Der 'Train & Predict' Button (6) startet den automatischen Kategorierungsprozess, wie in 5.2 beschrieben und in Abbildung 5.1 Schritt 3 dargestellt. Nach dem beendigen der Vorhersage wird die Ansicht der Berichte im Zentrum der Applikation aktualisiert.

Um irrelevante oder fehlerhafte Berichte zu löschen kann der Bericht durch halten der linken Maustaste zu dem Papierkorb (7) gezogen werden. Bei erfolgreicher Ausführung des Löschvorgangs wird der entsprechende Bericht aus der Ansicht und der Sammlung entfernt.

Filter

Der Filter ermöglicht es dem Nutzer die Suchergebnisse anzupassen oder die Anzeige der Berichte auf bestimmte Kriterien zu beschränken, wie in Abschnitt 4.1 **A6** gefordert.

Mit einer Dropdown-Liste kann der Benutzer auswählen, wie die Berichte angeordnet werden. Nach Relevanz werden die Berichte, wie in dem Kapitel der Suche beschrieben, mit Hilfe von dem bereitgestellten Wert von Lucene angeordnet. Weitere Auswahlmöglichkeiten sind dann noch nach auf- oder absteigendem Datum zu sortieren.

Zum Filtern der angezeigten Berichte besteht die Möglichkeit diese auf eine oder mehrere Sicherheiten und Kategorien zu beschränken. Durch das Drücken des 'Anwenden' Buttons wird der Filter dann auf die aktuelle Anzeige der Berichte angewandt. Ziel ist es hier dem Benutzer die Möglichkeit zu bieten, schnell die für ihn relevanten Kategorien zu verbessern und die Menge an simultaner Arbeit zu verringern, indem die Menge der Berichte vorübergehend beschränkt wird.

Übersicht

Im unteren Teil des Filters befindet sich eine Grafik (Abbildung 5.2 (11)). Diese dient zum einen als Übersicht und zum anderen ermöglicht diese ebenfalls noch eine zusätzliche Exploration der Berichte.

Der genaue Aufbau der Grafik wird in Abschnitt 4.5 beschrieben. Neben der Funktion dem Benutzer ein Feedback zu geben (**A7**), ermöglicht das Anklicken der Felder die Daten zusätzlich zu erforschen beziehungsweise zu explorieren (**A6**). Wenn die Maus über einen Teil der Graphik platziert wird, erscheint die genau Anzahl der Berichte dieser Kategorie, für diese Sicherheit in Form eines Tooltips. Bei dem Klicken mit der linken Maustaste wird dann automatisch ein Filter angewendet, der alle Berichte dieser Kategorie und Sicherheit anzeigt. Zusätzlich dazu wird bei dem Filtern mit Hilfe der Grafik, unter der Grafik noch sechs Stichworte ähnlich wie bei (9) aufgelistet. Drei Worte sind dabei grün und repräsentieren die besten Signalworte für die Auswahl der Berichte, drei rote Worte repräsentieren die drei Stichworte die für diese Kategorie den schlechtesten Wert haben.

Detailansicht

Um dem Benutzer noch mehr Informationen zu jedem Bericht zur Verfügung zu stellen (**A3**) werden in der Detailansicht (Abbildung 5.2 (4)) die gesamte Information des entsprechenden Berichts auf Anfrage angezeigt.

Durch Klicken des 'X' wird diese Ansicht geschlossen.

Dokumentansicht

Die Dokumentansicht (Abbildung 5.2 (12)) befindet sich im Zentrum der Applikation. Hier werden die zu kategorisierenden Elemente aufgelistet.

Jedes Element repräsentiert ein Bericht/Dokument. Der Aufbau der Kachel ist, wie in Abschnitt 4.2 beschrieben, umgesetzt. Der Großteil jeder Kachel besteht aus einem Aussagekräftigen Bild. Der Balken rechts Signalisiert mit seiner Farbe die Sicherheit der automatischen Zuweisung, beziehungsweise ob eine Bericht bereits bearbeitet wurde. Unterhalb des Bildes befindet sich ein waagrechter Balken der die Farbe der zugeteilten Kategorie enthält. Unter diesem Balken befindet sich noch ein dünnerer Balken der die Farbe der vorherigen Zuweisung enthält, um somit auf eine Veränderung durch eine bestimmte Aktion aufmerksam zu machen.

Als Zusatzinformation zu jedem Bericht wird der Wert der automatischen Zuweisung der besten drei Kategorien angezeigt. Wie dieser zustande kommt wird in Abschnitt 5.2 beschrieben.

Abbildung 5.3: Dokumentansicht. Per Rechtsklick können Kategorien zugewiesen, ähnliche Dokumente gefunden oder mehr Details angezeigt werden.

Um eine schnelle Kategorisierung zu ermöglichen wurde für die Kategorisierung die Funktion implementiert, mehrere Berichte durch das Halten der STRG-Taste zu Markieren und zusammen zu Kategorisieren (Abbildung 5.3). Die markierten Elemente sind dabei weiß umrahmt.

Unter Verwendung eines Rechtsklicks mit der Maus auf ein Element öffnet sich ein Menü. Das Menü, dargestellt in Abbildung 5.3, ist in drei Teile unterteilt. Im unteren Teil des Menüs befinden sich die Kategorien. Durch das Drücken einer solchen Kategorie wird diese dem entsprechenden Elemente oder wenn mehrere Elemente selektiert sind, den entsprechenden Elementen zugewiesen. Mit dem Auswählen der Spalte 'Mehr Info' wird die Detailansicht für das Element geöffnet, das mit dem Rechtsklick ausgewählt wurde.

Auch in dieser Ansicht kann eine Suche nach bestimmten Dokumenten erfolgen. Durch das Auswählen der Funktion 'Zeige ähnliche' in dem Menü, werden die Dokumente nach Ähnlichkeit zu dem ausgewählten Dokument angezeigt. Die Umsetzung diesen Features ist in Abschnitt 5.2 genauer Beschrieben.

Durch das Schließen des Filters und der Detailansicht, nimmt die Dokumentenansicht den Großteil der Fläche der Benutzeroberfläche ein. Dadurch können sehr viele Dokumente auf einmal eingesehen werden. Mit Hilfe der Mehrfachauswahl von Elementen, der Schnellen Zuweisung von Kategorien zu den Elementen und der klar erkennbaren und voneinander abgegrenzten Kacheln, kann der Benutzer die Kategorisierung somit schnell durchführen.

6 Anwendungsbeispiel

Um die entwickelten Funktionen besser verstehen zu können, werden diese Funktionen nun in diesem Kapitel konkret für die Lösung einer Problemstellung verwenden. Dabei geht man bei diesen Beispielen davon aus, dass eine Person ein Nachrichtenereignis näher betrachten möchte. Um jegliches Auftreten diesen Ereignisses zu finden, wird nun unser Prototyp mit den entwickelten Funktionen benutzt und das Vorgehen des Benutzers schrittweise Dokumentiert.

In den zwei folgenden Anwendungsbeispielen betrachten wird eine Sammlung der Tagesschau-Sendungen vom 01.01.2015 bis 31.12.2015, die von 20.00 Uhr bis 20.15 Uhr jeden Tag, in Das Erste ausgestrahlt wurden. Die Tagesschau ist eine täglich produzierte Nachrichtensendung des Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland (ARD). Sie befasst sich drei mal täglich mit den aktuellsten und wichtigsten Geschehnissen in Deutschland und der Welt. Die berichteten Themengebiete beinhalten Politik, Naturereignisse, Kultur, wichtige Veranstaltungen und Sport, wie auch ein aktueller Wetterbericht für Deutschland. Die Tagesschau wird in mehreren Fernsehsendern wie auch online in einem Livestream ausgestrahlt und ist ebenfalls online auf Abruf einsehbar.

Um die Tagesschausendungen für unser Projekt zu verwenden, wird jede Episode der Tagesschau und deren Untertitel in einzelne Segmente unterteilt, welche dann jeweils genau einen Bericht der Sendung enthalten. Zusätzlich wird ein Frame als Bild eines Segments gespeichert.

6.1 Auftreten der Pegida-Bewegung im Jahr 2015

In diesem Beispiel betrachten wir nun einen fiktiven Verhaltenswissenschaftler, der sich für das Auftreten von rechtspopulistischen Organisationen in der Öffentlichkeit interessiert. Hierfür möchte er vor allem das zeitliche Auftreten der Organisation Patriotische Europäer gegen die Islamisierung des Abendlandes (Pegida) in der Tagesschau im gesamten Jahr 2015 erforschen. Dabei möchte der Wissenschaftler sowohl Nachrichten über das Auftreten der Pegida, wie auch die Reaktion auf das Auftreten der Pegida von anderen Parteien untersuchen.

Um alle relevanten Berichte über die Pegida in den Tagesschausendungen schnell und zuverlässig zu identifizieren benutzt der Wissenschaftler nun den in dieser Arbeit entwickelten Ansatz

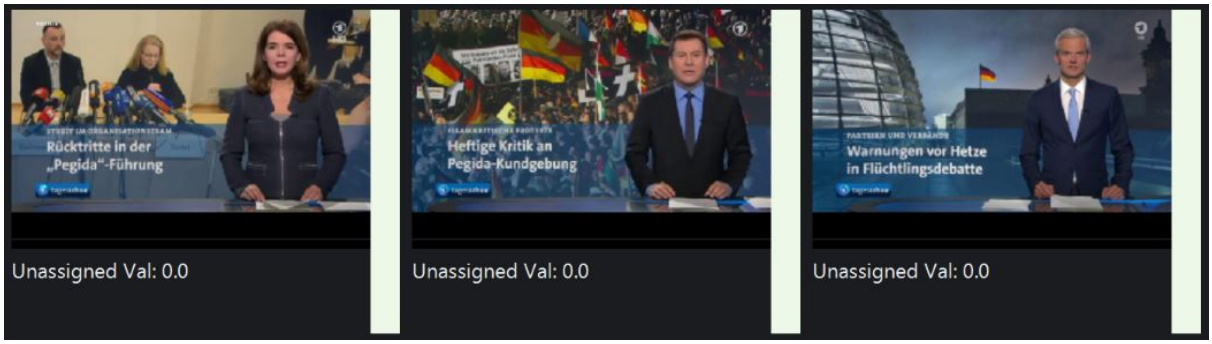


Abbildung 6.1: Die Ersten vier Ergebnisse bei der Suche nach dem Wort 'pegida'.

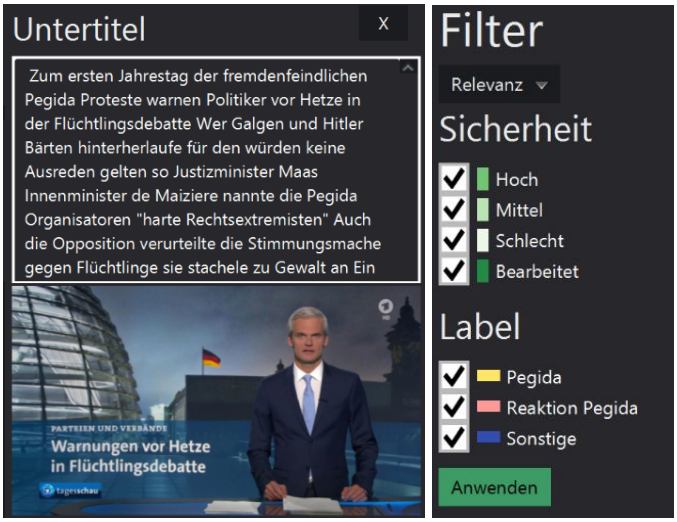
zur Kategorisierung. Der Wissenschaftler erhofft sich so einen Überblick zu erhalten, wie auch die Möglichkeit die einzelnen Vorkommnisse genauer zu betrachten.

Zu Beginn seiner Arbeit erstellt der Verhaltenswissenschaftler nach dem öffnen des Programms, zuerst drei Label mit Hilfe der 'neues Label' Funktion. Das erste Label nennt er 'Pegida' mit gelber Farbe, das Zweite 'Reaktion auf Pegida' mit roter Farbe und als Drittes erstellt der Wissenschaftler noch das Label 'Sonstige' mit blauer Farbe.

Für die Suche nach den relevanten Berichten benutzt der Wissenschaftler die Stichwortsuche. Dafür schreibt er in die Textfeld, gekennzeichnet mit dem Wort 'Suche', einen Begriff von dem er das beste Ergebnis erwartet. Der Filter (Abbildung 6.2b), der bei Programmstart geöffnet ist, zeigt an, dass er die Suchergebnisse nach Relevanz sortiert. Dies ist für den Beginn seiner Arbeit am nützlichsten, somit verändert der Wissenschaftler vorerst nichts an den Filtereinstellungen. Nun gibt er in das Textfeld das Wort 'pegida' ein und bestätigt durch das Drücken der Enter-Taste die Suche.

Die Suchergebnisse werden nun aufgelistet. In Abbildung 6.1 werden die drei besten Ergebnisse dargestellt. Durch den Titel des ersten gelisteten Berichts erkennt der Wissenschaftler bereits, dass dieser Bericht in die Kategorie 'Pegida' fällt. Mit Hilfe eines Rechtsklicks mit der Maus auf die entsprechende Kachel erhält der Wissenschaftler nun eine Liste mit mehreren Auswahlmöglichkeiten. Er hat nun die Option nach ähnlichen Berichten zu Suchen, mehr Informationen über den Bericht zu erhalten oder den Bericht in eine Kategorie einzuteilen. Er entscheidet sich sofort diesen, wie auch den nächsten Bericht der Kategorie 'Pegida' zuzuordnen. Bei dem Dritten Ergebnis wird Ihm allerdings nicht sofort ersichtlich, ob dies direkt mit Pegida zusammenhängt oder nur eine allgemeinere Stellungsnahme einer anderen Partei zu einem Thema ist, welches mit dem der Pegida zusammenhängt.

Unsicher über die Kategorisierung des dritten Berichts, entscheidet sich der Verhaltenswissenschaftler nun mehr Informationen über diesen Bericht zu erfahren, indem dieser per Rechtsklick mit der Maus auf die Kachel und dann in der erschienenen Liste 'Mehr Info' anklickt. Wie in Abbildung 6.2a dargestellt erhält dieser nun den kompletten Untertiteltext des ausgewählten Nachrichtenberichts. Durch schnelles überfliegen des Textes erhält man somit den gesamten Inhalt und der Wissenschaftler kann den Bericht so der zweiten Kategorie zuordnen, da es



(a) Untertitelanzeige **(b)** Filtereinstellungen

Abbildung 6.2: Abbildung 6.2a zeigt die zusätzlichen Informationen, Abbildung 6.2b stellt die möglichen Filteroptionen dar.

eine Reaktion auf das fremdenfeindliche Verhalten der Pegida Anhänger ist und es sich nicht um eine spezifische, aktuelle Handlung derer handelt.

Dieses Vorgehen wird nun von dem Wissenschaftler auf die folgenden Berichte ebenfalls angewandt, bis alle angezeigten Berichte kategorisiert sind.

Der nächste Schritte des Wissenschaftlers ist es jetzt noch nach relevanten Berichten zu suchen, welche nicht von der Stichwortsuche mit dem Suchbegriff ’pegida’ gefunden wurden. Um dies möglichst schnell zu erreichen, ohne alle Berichte manuell zu betrachten, benutzt der Wissenschaftler als nächstes nun die ’Train & Predict’ Funktion des Programms. Diese Funktion verwendet nun die bereits kategorisierten Berichte des Wissenschaftlers um den Machinelearning-Algorithmus zu trainieren. Als nächstes wendet dieser das gelernte Wissen auf die noch nicht kategorisierten Berichte an. Diese Berichte werden so nun automatisch den verschiedenen Kategorien zugeordnet.

Durch die Verwendung des Filters aus Abbildung 6.2b, kann der Wissenschaftler nun alle Berichte Filtern, die von dem Algorithmus nur mit einer sehr schlechten Sicherheit zugeordnet wurden und sich zusätzlich nur auf die Kategorien von ’Pegida’ und ’Pegida Reaktion’ beschränken. Dadurch wird die Menge an Relevanten Berichten deutlich reduziert.

In einem letzten manuellen Schritt kategorisiert und kontrolliert der Verhaltenswissenschaftler nun die gefilterten Berichte und weißt sie den Kategorien wie zu Beginn zu. Da der Wissenschaftler nun alle Berichte die für seine Arbeit relevant sind gefunden hat, kann dieser mit seiner eigentlichen Arbeit beginnen.

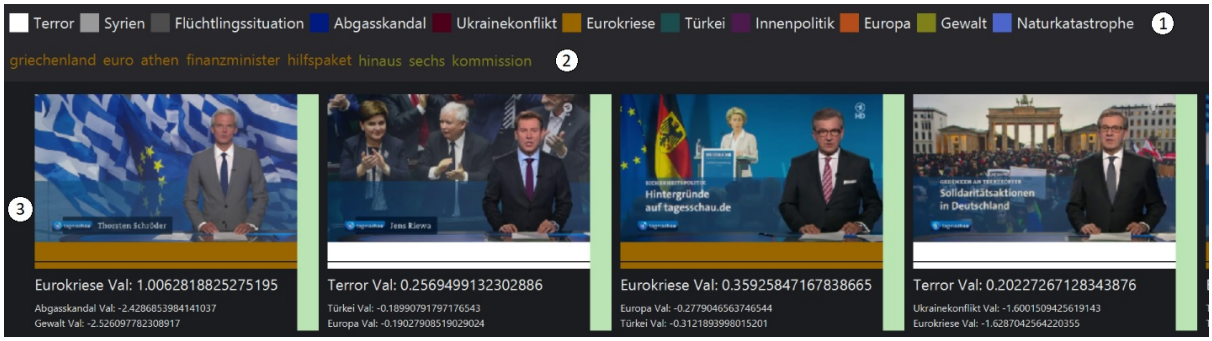


Abbildung 6.3: (1) Legende der Kategorien (2) Hervorgehobene Stichworte des Berichts links in (3) Nachrichtenberichte von Algorithmus in Kategorien eingeteilt

6.2 Übersicht der Nachrichtenthemen in Jahre 2015

In diesem zweiten Beispiel für eine Anwendung der Applikation betrachten wir die Arbeit eines Analysten. Die Aufgabe des Analysten ist es die Tagesschau von dem Jahr 2015 zu analysieren. Dabei ist es von besonderem Interesse wie mächtig das jeweilige Thema in Relation zu den anderen Themen war. Hierfür sind bestimmte Teile der Tagesschau, wie das Wetter und die Sportergebnisse, von keinem Interesse. Weiterhin möchte der Analyst aber wenig Zeit damit verbringen, jede Tagesschaufolge manuell durchzusehen. Daher ist sein Ziel mit einem geringen Aufwand trotzdem einen relativ guten und korrekten Überblick zu bekommen. Auch in diesem Fall wird er versuchen, dies mit dem Prototyp dieser Arbeit zu erreichen.

Der Analyst startet die Applikation und beginnt nun einige Kategorien zu erstellen, von denen er weiß dass diese unter den Berichten vorkommen. Dafür benutzt er wie in Abschnitt 6.1 den ’neues Label’ Button in der Toolbar und ordnet jeder Kategorie einen Namen und eine Farbe zu, wie in Abbildung 6.3 (1) zu sehen ist. Nachdem ein Teil der Label angelegt wurden beginnt der Analyst nun mit der Kategorisierung einiger Berichte. Hierfür stellt er den Filter so ein, dass die Berichte nach dem Datum sortiert werden. Der Analyst teilt nun einige Berichte in die entsprechende Kategorien ein, indem er diese aus der Liste auswählt, welche sich mit einem Rechtsklick auf den Bericht öffnet. Nachdem der Analyst einen Bericht einer neuen Kategorie zugeteilt hat, will dieser noch mehr Berichte zu dem selben Thema finden, um dadurch möglichst schnell für jede seiner Kategorien eine mittelgroße Menge an Stichproben zu finden. Indem er nach einem Rechtsklick auf den Bericht, in der Liste die Funktion ’zeige Ähnliche’ wählt, werden ihm entsprechende Berichte angezeigt. Nachdem er genug Stichproben für eine Kategorie gefunden hat, stellt er den Filter so ein, dass Berichte die bereits von ihm bearbeitet wurden nicht mehr angezeigt werden. Das erreicht er indem er den Haken aus dem entsprechenden Kästchen entfernt. Nun bestätigt er eine Suche ohne eine Suchwort einzugeben und sieht wieder eine Chronologische Auflistung aller Berichte, welche noch nicht bearbeitet wurden.

Abbildung 6.4: Oben eine Grafik, welche die Verteilung der Kategorien in Relation zueinander und der Sicherheit der Zuweisung aufzeigt, unten, Stichworte einer Kategorie, die für (grün) und gegen (rot) die Kategorie sprechen.

Nachdem der Analyst nun einige Stichproben in die Kategorien eingeteilt hat, drückt er den 'Train & Predict' Button in der Toolbar. Den unbearbeiteten Berichten werden nun wie in Abbildung 6.3 (3) Kategorien zugeordnet. Der Analyst erkennt nun in dem Bericht links, dass bei diesem Bericht der Titel fehlt. Daraufhin bewegt der Analyst seine Maus über diesen Bericht und erhält somit eine Anzahl an Stichworten, wie in Abbildung 6.3 (2) dargestellt. Die Stichworte Griechenland, Euro, Athen, Finanzminister und Hilfspaket sind hier Indizien, dass der Bericht in die Kategorie Eurokrise fällt. Der Analyst stimmt daher der Zuweisung zu und weist nun zur Bestätigung dem Bericht ebenfalls die Kategorie Eurokrise zu.

Um die Genauigkeit der automatischen Zuweisung noch zu verbessern, entscheidet sich der Analyst nun dazu Stichprobenartig die Zuweisungen zu überprüfen und falls nötig, noch mehr Kategorien hinzuzufügen.

Um nun die von dem Algorithmus des maschinellen Lernens eingeteilten Berichte zu überprüfen, verwendet der Analyst nun die interaktive Grafik die in Abbildung 6.4 dargestellt ist. Im oberen Teil der Grafik sieht man die Verteilung der vergebenen Sicherheiten der Zuweisung, im unteren Teil die Verteilung der Kategorien innerhalb der jeweiligen Sicherheit.

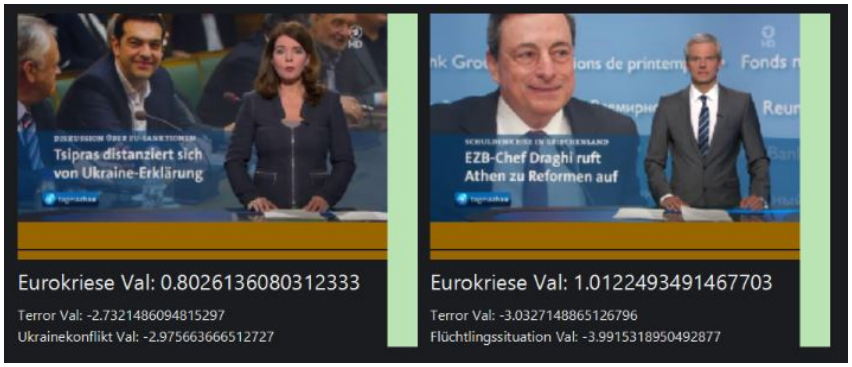


Abbildung 6.5: Darstellung von zwei Berichten nach mehreren Lernprozessen.

Der Analyst entscheidet sich nun dafür die Kategorie der Eurokrise (Orange) bei mittlerer Sicherheit (Hellgrün) genauer zu betrachten, da dieser Spalte deutlich mehr Berichte zugeteilt wurden. Dafür klickt er mit einem Mausklick auf die entsprechende Spalte in der Grafik. Dadurch wird der Filter automatisch angepasst und die Berichte angezeigt. Zusätzlich erhält der Analyst unter der Grafik drei grüne und rote Stichworte (Abbildung 6.4 unten). Dem Analyst fallen vor allem die roten Worte auf. Diese Worte sind Stichworte, welche den nun angezeigten Berichten entnommen wurden. Grün sind dabei die Worte die in der Kategorie einen hohen Wert haben und Rot, Stichworte die einen hohen Wert in anderen Kategorien halten.

Um das Vorkommen eines eher untypischen Stichwortes der Eurokrise-Kategorie in den Berichten zu überprüfen, klickt der Analyst nun auf das Stichwort Russland und die Anzeige filtert nun alle Berichte aus, die dieses Wort nicht enthalten. Der Analyst überprüft nun die angezeigten Berichte (Abbildung 6.5). Nach einer Überprüfung der Stichworte tendiert der Analyst bereits dazu, dass die Berichte Korrekt eingeordnet sind. Um auf Nummer sicher zu gehen wirft der Analyst doch noch einen Blick in den Untertitel beider Berichte indem er mit Rechtsklick auf den Bericht das Menü aufruft und 'Mehr Info' auswählt. Er kommt somit zu dem Schluss dass beide Berichte richtig eingeordnet wurden, da beide Berichte zu dieser Kategorie gehören.

Den letzten Schritt wiederholt der Analyst nun für die meisten Kategorien mit sehr schwach zugeteilter Sicherheit mehrmals und erhält so nach recht kurzer Zeit einen guten Überblick über die Menge an Berichten in den jeweiligen Kategorien.

7 Diskussion

In diesem Kapitel werden wir nun auf die Vor-und Nachteile des entwickelten Ansatzes eingehen um zu erörtern wie effektiv dieser die Problemstellung bearbeitet und wie die bestehenden Nachteile noch behoben oder verbessert werden können.

Der Arbeitsbereich des Benutzers in der Benutzeroberfläche ist in eindeutige Bereiche unterteilt. Optional können Features wie der Filter und die Detailansicht an den Seiten ein- oder ausgeblendet werden, im Mittelpunkt stehen allerdings immer die Dokumente. Der gesamte Prozess findet also innerhalb eines einzelnen Fensters statt. Dies ist für eine effektive Kategorisierung klar von Vorteil. Der Benutzer wird nicht ständig durch das Öffnen und Schließen unterschiedlicher Ansichten gestört oder abgelenkt und erlernt den Bearbeitungsprozess unmittelbar. Durch das klare Design versteht der Benutzer die Funktionen intuitiv.

Die Suchfunkion ermöglicht dem Benutzer die Dokumente zu durchsuchen. Der Nutzer kann so sein eigenes Vorwissen über vergangene Geschehnisse aus Nachrichten und Themen verwenden um einen Anfangspunkt für seine Aufgabe zu finden.

Nach einer initialen Kategorisierung einzelner Elemente und Erstellung dieser Kategorien, zu Beginn der Arbeit, kann der Benutzter den automatischen Kategorisierungsprozess starten. Dieser ermöglicht es, schnell neue Kategorien zu identifizieren, indem besonders niedrige Sicherheiten untersucht werden. Dies ist besonders von Vorteil, wenn der Datensatz aus sehr vielen und unterschiedlichen Themengebieten stammt. Weiterhin ermöglicht es dem Nutzer, zwischen mehreren Wegen bei der Kategorisierung zu wählen.

Die Exploration der Daten ist ein wichtiger Teil bei der Kategorisierung von Dokumenten. Der Nutzer kann einen Weg wählen bei dem er eine Kategorie nach der anderen abarbeitet, er kann sich darauf konzentrieren alle Kategorien zu identifizieren oder einen Mittelweg entsprechend seiner Stimmung wählen.

Ein weiterer Vorteil ist es, dass bei einer ständigen Verbesserung des Trainingsset der Algorithmus die vielen Dokumente ebenfalls besser einordnet. Somit nimmt der Arbeitsaufwand des Benutzers immer mehr ab, bis dieser nur noch eine Kontrollfunktion bei der Kategorisierung einnimmt, anstatt die Kategorisierung selbst zu leiten.

Im Großen und Ganzen hat der Benutzer also die Wahl wie er die Kategorisierung handhaben will. Egal welchen Weg dieser wählt, er wird immer maximal gut von den Features unterstützt.

Der Nachteil diesen Ansatzes ist allerdings, dass mit einer sehr großen und diversen Datenmenge, die automatische Kategorisierung anhand des maschinellen Lernens zu Beginn, sehr große Fehler macht. Es fehlt zu Beginn ein großer Teil der Kategorien, somit ist es besonders schwer ein entsprechend erfolgreiches Muster zu finden. Dies kann den Eindruck erwecken, dass dieses Feature den einfacheren Features wie der Stichwortsuche und der Suche nach ähnlichen Dokumenten deutlich unterlegen ist. Somit wirkt die Darstellung der schlecht Kategorisierten Dokumente eher störend und überflüssig.

Die Veränderung des Datensets wird zwar durch das Speichern alter Kategorien eines Elements und der Veränderung des Überblicks etwas ersichtlich, es besteht allerdings trotzdem noch der Wunsch nach mehr Informationen in diese Richtung. Dies könnte allerdings zu einer Störung des Arbeitsprozesses führen, da der Benutzer dann zu leicht abgelenkt wird. Hierfür wären noch einige Fragen zu erörtern und zu lösen.

8 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war es einen visuellen Ansatz zu entwickeln der eine effektive und interaktive Kennzeichnung von großen Multimediadatenmengen ermöglicht. Dabei sollten die Auswirkung einer initialen Kennzeichnung auf den Lernprozess ersichtlich werden und es ermöglichen den Kennzeichnungsprozess iterativ zu verfeinern, um somit ein durch den Benutzter festgelegtes Kennzeichnungsverfahren der Daten zu erreichen und ein leichteres Verständnis dessen zu erreichen.

Zunächst wurden hierfür andere Arbeiten mit ähnlichen Fragestellung in Kapitel 2 betrachtet um an deren Vorarbeit anzuknüpfen, um diese dann für unsere Aufgabe später zu erweitern. In dem Kapitel der Grundlagen wurde dann auf bestimmte Themengebiete näher eingegangen, die für die Lösung dieser Aufgabe relevant waren. Dazu gehörte unter anderem das maschinelle Lernen und das Gebiet des 'Information Retival'.

Als nächstes wurden die Konzepte veranschaulicht, welche für die Lösung der Problemstellung ausgearbeitet wurden. Hierfür wurde in einem ersten Schritt die Aufgabe im Detail beschrieben und dann Anforderungen gestellt, welche für die Lösung der Problemstellung nötig waren. Daraufhin wurden die Konzepte vorgestellt, die für diese Anforderungen erarbeitet wurden. Diese befassten sich mit dem Arbeitsablauf des Systems, für die Bewältigung der Aufgabe wie auch mit allgemeinen Designfragen um die Informationen der Berichte optimal und möglichst kurz darzustellen. Weiterhin wurde ein Konzept vorgestellt um das Interesse und die Stimmung des Benutzers mit entsprechendem Feedback auf einem hohen Niveau zu halten.

Die Konzepte wurden dann in einem Prototypen umgesetzt. Die Verwirklichung wurde in Kapitel 5 erläutert. Dies umfasste die Gestaltung der Benutzeroberfläche sowie die Funktionen für die Interaktion zwischen dem Prototypen und einem Benutzer. Zusätzlich dazu wurde ebenfalls die Funktionsweise der Informationsfindung und Informationsverarbeitung erläutert.

Um einen Einblick in eine praktische Anwendung zu erhalten, wurden zwei mögliche Anwendungsfälle betrachtet. Hierfür wurden zwei Szenarien vorgestellt in denen die implementierten Funktionen praktisch verwendet wurden.

Als letzten Teil dieser Arbeit, wurden die Vor- und Nachteile in einer kurzen Diskussion angesprochen.

Ausblick

Die Funktion der entwickelten Methode ist es dem Benutzer eine interaktive Möglichkeit für eine Kategorisierungsaufgabe bereitzustellen und ihn anhand maschinellen Lernens zu unterstützen.

Der Ansatz kann allerdings noch erweitert werden. Es ist denkbar die Funktionsweise des maschinellen Lernens noch stärker in den Kategorisierungsprozess einzubinden. Um so noch eine genauere und zuverlässigere Zuweisung zu erzielen. Der Benutzer könnte die Möglichkeit erhalten Worte manuell für die Kategorisierung zu Klassifizieren, beziehungsweise hervorheben oder in den Hintergrund zu rücken, um den Algorithmus zu beginn eine bessere Wissensgrundlage zu schaffen.

Weiterhin könnte dem Benutzer mehr Wahl bei der Darstellung gegeben werden, sodass dieser die Zusatzinformationen die in den Kacheln dargestellt werden, selbst festlegen kann.

Wie bereits bei der Diskussion erwähnt, könnten auch noch mehr Informationen über die Veränderungen des Datensatzes gespeichert und Visualisiert werden, um auch hier zusätzliche Informationen zur Verfügung zu stellen.

Literaturverzeichnis

[AM08] L. Auria, R. A. Moro. „Support vector machines (SVM) as a technique for solvency analysis“. In: (2008) (zitiert auf S. 21).

[BGLB16] J. Beel, B. Gipp, S. Langer, C. Breitinger. „paper recommender systems: a literature survey“. In: *International Journal on Digital Libraries* 17.4 (2016), S. 305–338 (zitiert auf S. 22).

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999 (zitiert auf S. 22, 23).

[CV95] C. Cortes, V. Vapnik. „Support-vector networks“. In: *Machine learning* 20.3 (1995), S. 273–297 (zitiert auf S. 20).

[CW16] P. van der Corput, J. J. van Wijk. „ICLIC: Interactive categorization of large image collections“. In: *Pacific Visualization Symposium (PacificVis), 2016 IEEE*. IEEE. 2016, S. 152–159 (zitiert auf S. 14, 15).

[CW17] P. van der Corput, J. J. van Wijk. „Comparing personal image collections with PICTuReVis“. In: *Computer Graphics Forum*. Bd. 36. 3. Wiley Online Library. 2017, S. 295–304 (zitiert auf S. 15, 16).

[FCH+08] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin. „LIBLINEAR: A Library for Large Linear Classification“. In: *Journal of Machine Learning Research* 9 (2008), S. 1871–1874 (zitiert auf S. 35).

[Fou] T. A. S. Foundation. *Welcome to Apache Lucene*. URL: lucene.apache.org/ (besucht am 12. 05. 2018) (zitiert auf S. 35).

[HKBE12] F. Heimerl, S. Koch, H. Bosch, T. Ertl. „Visual classifier training for text document retrieval“. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), S. 2839–2848 (zitiert auf S. 17, 18).

[JM] D. Jurafsky, J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (zitiert auf S. 21).

[KAF+08] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon. „Visual analytics: Definition, process, and challenges“. In: *Information visualization*. Springer, 2008, S. 154–175 (zitiert auf S. 24).

[KHSW17] K. Kurzhals, M. Hlawatsch, C. Seeger, D. Weiskopf. „Visual analytics for mobile eye tracking“. In: *IEEE transactions on visualization and computer graphics* 23.1 (2017), S. 301–310 (zitiert auf S. 16, 17).

[MRT12] M. Mohri, A. Rostamizadeh, A. Talwalkar. *Foundations of machine learning*. MIT press, 2012 (zitiert auf S. 19).

[MSB+14] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky. „The Stanford CoreNLP Natural Language Processing Toolkit“. In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, S. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010 (zitiert auf S. 35).

[RWW10] O. Rooij, J. van Wijk, M. Worring. „Mediatable: Interactive categorization of multimedia collections“. In: *IEEE Computer Graphics and Applications* 30.5 (2010), S. 42–51 (zitiert auf S. 13, 14).

[SB88] G. Salton, C. Buckley. „Term-weighting approaches in automatic text retrieval“. In: *Information processing & management* 24.5 (1988), S. 513–523 (zitiert auf S. 22).

[Shn03] B. Shneiderman. „The eyes have it: A task by data type taxonomy for information visualizations“. In: *The Craft of Information Visualization*. Elsevier, 2003, S. 364–371 (zitiert auf S. 23, 24).

Alle URLs wurden zuletzt am 13. 05. 2018 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