Richtlinie zum „Software-Sprint“ Media Uncovered -

Schlussbericht

Zuwendungsempfänger:

Natalie Widmann

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18S04 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen?

Eine unabhängige und vielfältige Medienlandschaft bildet die Grundlage für eine informierte Zivilgesellschaft, die aktiv demokratische Prozesse mitgestalten kann. Allerdings hat jedes Nachrichtenportal eine Voreingenommenheit. Diese kommt unbewusst - etwa durch kulturelle Prägung, Teams von ähnlich denkenden Menschen - oder bewusst - durch die Teilhabe von Parteien und Firmen oder in manchen Ländern durch Zensur und staatliche Propaganda - zustande.

Diese Voreingenommenheit beeinflusst Leser*innen in ihrer Wortwahl, ihrem Blick auf die Welt und natürlich auch in ihrem Verhalten - ob im alltäglichen Umgang mit Menschen oder an der Wahlurne.
Diese Verzerrungen sind umso bedenklicher, je weniger sie uns bewusst sind.

Mit Media Uncovered ist es möglich den Sprachgebrauch von Medien zu untersuchen und politische, ideologische oder gesellschaftliche Tendenzen aufzuzeigen. Zum Thema Flüchtingskrise gab es beispielsweise große Diskussionen welche Wörter verwendet werden und welche Emotionen diese auslösen. Mit Media Uncovered kann der Sprachgebrauch zweier Medienportale analysiert und verglichen werden.
Dabei schauen wir nicht auf einzelne Artikel sondern versuchen über mehrere tausend Texte einer Zeitung hinweg ein übergeordnetes Bild von deren Vokabular und semantischen Assoziationen zu bekommen.

Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Das Projekt Media Uncovered gliedert sich in drei Hauptbestandteile:

- das Erstellen einer Datenbank mit möglichst vielen Artikeln von verschiedenen Online-Zeitungen. Mithilfe eines Web Crawler wird die Datenbank automatisch um neuveröffentlichte Artikel erweitert.

- das Erkennen von Korrelation und Sprachmustern, die für das menschliche Auge nicht greifbar sind. Dafür verwenden wir Methoden aus dem Bereich des maschinellen Lernens. Word Embeddings (word2vec, fastText) bilden den Inhalt einer großen Kollektion an Texten in einem Vektorraum ab, so dass Wortassoziationen erhalten bleiben.

- die Darstellung dieser Modelle. Eine Web Applikation ermöglicht Nutzern eine intuitive und interaktive Exploration und Analyse der Themen.

Beitrag des Projektes zu den Zielen der Förderinitiative „Software-Sprint“

Wer ist die Zielgruppe für Deine Lösung? Wie profitiert sie von den Ergebnissen Deines Projekts? Welche Bezüge gibt es zu den Themenfeldern und Zielen des Software Sprints?

Media Uncovered setzt sich für mehr Transparenz und einen bewussten Umgang mit Medien ein.

Unser Analysetool ermöglicht es den Sprachgebrauch von Zeitungen zu analysieren und konsistente politische, ideologische oder gesellschaftliche Tendenzen aufzudecken. Dadurch können sich interessierte Leser kritisch mit dem eigenen Medienkonsum auseinandersetzen und unterschiedliche Informationsquellen miteinander vergleichen.
Journalisten und Organisationen erhalten Einblicke in den gesamten Inhalt von Medien und können Belege zu Themen wie Medienkonsolidierung, Sprachgebrauch und Abhängigkeit von Medienportalen sammeln.
Dies schafft Transparenz und stärkt die Zivilgesellschaft.

Diversität ist der Fokus dieser Förderrunde des Prototype Funds. Unser Projekt unterstreicht die Notwendigkeit einer diversen Medienlandschaft, in dem es die Informationsblasen von Nachrichtenportalen aufzeigt. Diese Transparenz schafft auch Anreize für Zeitungen durch die Zusammenstellung diverser Teams und die Einnahme unterschiedlicher Perspektiven und Positionen bei der Berichterstattung, unterbewussten Tendenzen im Sprachgebrauch entgegenzusteuern.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden?

Während des Prototype Fund Förderzeitraums konnten wir sehr viele unserer Ideen umsetzen. Wir haben über 40 000 Artikel von 6 sechs unterschiedlichen Zeitungen verarbeiten und darauf Word Embedding Modelle trainiert.
In der Web Applikation können Nutzer nun ein Modell auswählen und den Sprachgebrauch dieser Zeitung interaktiv explorieren und mit anderen Medienportalen vergleichen.

Uns war es wichtig, dass die Nutzer das grundlegende Konzept dieser Modelle verstehen und sich somit kritisch eine eigene Meinung über deren Interpretier- und Vergleichbarkeit bilden können. Dazu gehört einerseits eine Übersicht der verwendeten Daten: Wie viele Artikel von einer Zeitung sind vorhanden und in welchem Zeitraum sind sie erschienen?
Um Modelle replizieren zu können, geben wir die genaue Word Embedding Methode (fasttext, word2vec), sowie deren spezifische Parameter an.
Darüber hinaus zeigen die Ergebnisse eines Benchmarktests mit 20 000 Analogien in 14 semantischen und morphologischen Kategorien, wie gut sich das Modell an die Realität angepasst hat. Eine korrekte Analogie im Bereich Familie wäre beispielsweise *Mutter ist zu Vater wie Tochter zu Sohn*, bei grammatikalischen Kategorien wie der Vergangenheitsform ist *laufen zu lief wie hören zu hörte.*

Nachdem der Nutzer einen Überblick über die Daten und das Modell hat, stehen unterschiedliche Analysemethoden zur Verfügung:

- Sprachgebrauch explorieren
Bei Eingabe eines Wortes wird eine Wortwolke mit den 50 ähnlichsten Begriffen angezeigt. Je größer die Begriffe desto stärker ist die Assoziation mit dem Suchbegriff.

- Bias detektieren
Um selbständig Annahmen über bestimmte Tendenzen in einer Zeitung zu untersuchen, können relevante Wörter auf eine Achse projeziert werden. Zum Beispiel kann man herausfinden ob Wörter wie Technologie, Mathematik oder Psychologie eher mit Frauen oder mit Männern assoziiert werden.

- Analogien generieren
Anhand zweier Wörter, wie beispielsweise *Merkel:Deutschland* oder *konservativ:progressiv* werden Analogien generiert die in einem ähnlichen Verhältnis zueinander stehen.

Infoboxen geben dem Nutzer Hintergrundinformationen zu den verwendeten Methoden und Beispieleingaben machen sie intuitiv nutzbar.
Damit haben wir die zentralen Meilensteine unseres Projektes erreicht.

Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Ein Hauptbestandteil unserer Arbeit ist das Kommunizieren der verwendeten Methoden mit deren Möglichkeiten und Einschränkungen. Unterschiedliche Algorithmen berechnen bei denselben Daten unterschiedliche Modelle in denen andere Wörter stärker miteinander assoziiert sind. Auch zu wenig Artikel oder Wörter die nur selten Vorkommen führen zu unzuverlässigen Modellen oder fehlinterpretierbaren Ergebnissen.
Deswegen haben wir unterschiedliche Methoden wie fasttext, word2vec und glove ausprobiert, sie miteinander verglichen und versucht Verlässlichkeitsindikatoren zu identifizieren.

Eine wichtige Erkenntnis, die auch von der Open Knowledge Foundation immer wieder unterstrichen wird, ist die Bedeutung von Nutzertests. Auch wenn wir immer das Gefühl hatten, dass die Applikation noch nicht weit gut genug entwickelt ist um ein Nutzertesting durchzuführen, haben wir sehr wichtiges und hilfreiches Feedback erhalten. Oft stellt sich dabei heraus, dass das geplante Feature gar nicht den erhofften Mehrwert liefert oder dass eine Darstellung viel zu komplex ist.

Des weiteren ist das Zusammenkommen mit anderen Projekten sehr motivierend und hilfreich. Auch wenn wir teilweise an sehr unterschiedlichen Problemen arbeiten, lohnt sich ein gemeinsamer Austausch, da es große Überschneidungen bei der Arbeitsweise und den Methoden und Technologien, die wir nutzen, gibt. Vor allem aber ist es sehr inspirierend und lässt Raum für neue Ideen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weiter-gehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Unser Tool gibt Lesern und interessierten Bürgen Einblicke in den übergeordneten Sprachgebrauch von Zeitungen.
Um es auch für Journalisten und Organisationen, die sich für freie und unabhängige Medien einsetzen, nutzbar zu machen, müssen Rückschlüsse auf die Originalartikel und deren Kontext möglich sein. Zum Beispiel kann eine starke Assoziation der Wörter *Flüchtling* und *Terrorist* auch die Folge eines positiven Kontexts, wie *Ein Flüchtling ist KEIN Terrorist* sein. Deshalb ist eine wichtige Weiterentwicklung von Media Uncovered das Hinzufügen von Originalartikeln um den genauen Kontext untersuchen zu können.

Weitere Pläne beinhalten den direkte Vergleich zweier Zeitungen in der Web Applikation, sowie die Möglichkeit durch das Kombinieren von vielen unterschiedlichen Nachrichtenportalen Rückschlüsse auf die Diversität der Medienlandschaft zu ziehen.

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Die Zusammenarbeit im Team war auf sehr unterschiedliche Art und Weise lehrreich und motivierend.

Im Team passiert es oft, dass man die grundlegenden Annahmen und Ziele des Projektes reflektiert und kritisch hinterfragt. Dieser Prozess, der zwar mühsam und zeitaufwendig ist, hat Media Uncovered stark definiert und die Ergebnisse geprägt. Aber auch Kompromisse finden und sich schnell in komplexe Themengebiete einzuarbeiten, um Entscheidung über grundlegende Infrastruktur treffen zu können, sind Fähigkeiten, die nicht nur auf fachlicher Ebene hilfreich sind.

Ich persönlich habe sehr viel über Frontend Entwicklung mit Angular, sowie Datenbanken und `Best Coding Practices`gelernt. Außerdem haben wir uns viel mit unterschiedlichen Aspekten des agilen Projektmanagement und des Nutzertestings beschäftigt.
Auch wenn diese Methoden mir nicht unbekannt waren, so sind sie doch mit Verhaltensänderungen verbunden, die vor allem unter Zeitdruck oft schwierig einzuhalten sind. Im Team lernt man deren Bedeutung jedoch besonders schätzen und kann sich gegenseitig dafür motivieren.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hinter-gründe, und wie bist Du alternativ vorgegangen?

Für die Analyse von Online-Zeitungen müssen tausende Artikel eines Nachrichtenportals gesammelt und gespeichert werden. Da jede Online-Zeitung eine unterschiedliche Daten Struktur für ihre Webseite nutzt, muss der Crawler manuell daran angepasst werden und über einen möglichst langen Zeitraum viele Artikel sammeln.
Darüber hinaus sind Zeitungsartikel urheberrechtlich geschützt und dürfen nur für den persönlichen Gebrauch gespeichert und verarbeitet werden. Um möglichen Konflikten vorzubeugen haben wir deswegen den öffentlich zugänglichen *All the News* Datensatz von Kaggle (https://www.kaggle.com/snapcrack/all-the-news/home) zum Training der Word Embedding Modelle verwendet. Er enthält circa 15 verschiedene englischsprachige Zeitungsportale, darunter CNN, New York Times, Washington Post, Breitbart, etc. mit mindestens 8000 Artikeln pro Zeitung.

Eine weitere Idee war die interaktive Darstellung der berechneten Modelle in einem drei-dimensionalen Vektorraum. Der Tensorflow Embedding Projector (https://projector.tensorflow.org/) ermöglicht genau dies. Wir wollten die Hauptfunktionalitäten in unserem Code wiederverwerten und neue Darstellungsweisen hinzufügen. Leider veröffentlicht Tensorflow nur die verarbeitete html Datei aber nicht den kompletten Code, so dass es nicht möglich ist kleine Anpassungen vorzunehmen oder Funktionen hinzuzufügen.

Darüber hinaus zeigten die Ergebnisse der Nutzertests, dass die drei-dimensionalen Darstellung und die vielen verschiedenen Interaktionsmöglichkeiten zu kompliziert sind und die Nutzer überfordert. Deswegen haben wir uns für eine alternative Darstellung entschieden in der die Hauptfunktionalitäten, den Sprachgebrauch erkunden, Bias entdecken und Analogien generieren, in separaten Graphen dargestellt wird. Durch die Eingabe von für den Nutzer relevanter Begriffe kann dieser unterschiedliche Aspekte einer Online Zeitung interaktiv und intuitiv erkunden.
Desweiteren können wir flexibel Änderungen vornehmen und neue Analysemethoden hinzufügen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GHitHub, Veröffentlichungen)?

Das Projekt Media Uncovered veröffentlicht allen Code, sowie Besprechungsprotokolle und wichtige Entscheidungen auf Github (https://github.com/MediaUncovered).

Eine aktuelle Demoversion wird noch veröffentlicht.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Media Uncovered war geplant als Teilzeitprojekt an dem kontinuierlich jede Woche ca. 10-15 Stunden gearbeitet wird. Aufgrund eines Stipendiums im Rahmen eines anderen Projektes hat sich dieser Arbeitsplan jedoch kurzfristig geändert: in den ersten drei Monaten des Förderzeitraums wurde weniger Zeit für Media Uncovered aufgebracht, sodass sich die hauptsächliche Arbeit in die zweite Hälfte verschoben hat.

Auch bei der Bearbeitung der Teilaufgaben, haben wir weniger Zeit in das Erstellen von Beispielanalysen und die Erforschung von weiteren Vergleichsmethoden gesteckt als erwartet, da das Erstellen und Testen einer intuitiven Web Applikation mehr Zeit in Anspruch nahm.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Das Open-Source-Projekt *ethically (http://docs.ethically.ai/)* eines Freundes beschäftigt sich damit maschinelle Lernmodelle auf ihre moralische Vertretbarkeit zu überprüfen. Im Rahmen dessen beschäftigt er sich auch mit Word Embeddings und implementierte eine aus der Fachliteratur bekannte Methode zur Generierung von Analogien.
Die für Media Uncovered ebenfalls geplante Implementierung haben wir daraufhin eingestellt und *ethically* in unser Projekt integriert.