Richtlinie zum „Software-Sprint“

HASS_FILTERN

Hasskommentare automatisiert filtern – Eine interaktive Erklärung

Schlussbericht

Zuwendungsempfänger:

Johannes Filter

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18S65 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? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

In den Kommentarspalten des Internet tobt der Hass. Um diesen zu einzudämmen, müssen Kommentare manuell geprüft werden. Weil dies viele Redaktionen überfordert, schließen sie die Kommentarfunktion gleich ganz ab. In diesem Kontext gelobt Machine Learning Besserung, indem automatisiert Kommentare ausgewertet werden. Die aktuellen Verfahren sind jedoch unausgegoren. So können sie leicht „ausgetrickst“ werden oder sind rassistisch.

Aktuell ist jedoch noch nicht klar, welche Aufgaben wir den Maschinen überlassen wollen. Damit es darüber eine gesellschaftliche Debatte gibt, müssen zuerst breite Teile der Bevölkerung die dahinter liegenden technischen Mechanismen verstehen. Weil die zugrundeliegenden Mechanismen für Laien schwer zu durchdringen sind, bereiten wir diese pädagogisch auf, damit sich mehr Menschen in die Debatte um Machine Learning einbringen können.

In diesem Projekt werden Techniken des Natural Language Processing am Beispiel der automatisierten Identifikation von Hasskommentaren erklärt. Neben Grafiken und Text gibt es einen

interaktiven Teil, mit dem die Nutzer*innen live im Browser ein Machine-Learning-Verfahren nutzen können. Durch die spielerische Aufbereitung werden Möglichkeiten aber auch Risiken erläutert.

Zudem war es das Ziel, dass das Softwaresystem mit innovativen Methoden umgesetzt wird. Hier gab es einige Änderungen zu dem ursprünglichen Plan. Schlussendlich entstand ein Softwarepacket, dass sich an NLP-Expert:innen richtet. Es benutzt ein Verfahren, dass sich in wissenschaftlichen Veröffentlichungen bewährt habt. Bis jetzt gab dazu keine effiziente Implementierung.

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?

Die Webseite richtet sich an Menschen, die den aktuellen Hype um Machine Learning nachvollziehen wollen. Diese müssen explizit nicht “in der Materie stecken”, daher wird kein Wissen vorausgesetzt um die Seite nutzen zu können. Die erste Zielgruppe profitiert von dem Bildungsangebot und somit wird deren Data Literacy gestärkt. Der Umgang mit modernen Technologien ist heute essentiell. Die Zweite Zielgruppe umfasst die Natural Language Community.

Das Projekt war an den Themenschwerpunkt “Maschinelles Lernen” orientiert. Zu einem wurde notwendige Bildungsarbeit geleistet aber zum anderen auch eine Arbeit die dem technischen und wissenschaftlichen Fortschritt dient.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Das Ziel des Projektes war eine Webseite auf der erklärt wird, wie Natural Language Processing am Beispiel von Kommentaren funktioniert. Die meisten Meilensteine wurde erreicht, jedoch ist die finale Veröffentlichung noch offen. Die Webseite besteht aus Text, Grafiken und einem interaktiven Teil. Die Idee der “Word Embedding” werden auf der Seite erklärt. Dabei handelt es sich um ein Verfahren des Natural Language Processing. In diesem wird Bedeutung in Wörtern ‘erzeugt’. Der Computer kann Text nicht den Sinn von Wörtern erfassen. Word Embedding generiert ein Model, bei dem ich z. B. zu einem gegebenen Wort ähnlich Wörter oder Synonyme erhalte. Die Ähnlichkeit wird darüber erzeugt, in welchem ähnlichem Kontexten sie vorkommen.

Die technische Umsetzung, sprich das Erstellen der Word Embeddings, war auch Teil des Projekts. Es gibt in diesem Bereich schon viel Forschung, aber wenig einfach zu benutzende Open-Source-Software. Und oft verlangen bestehende Softwarelösung große Datenmengen. In vielen Fällen, wie in unserem, gibt es diese jedoch nicht. Daher ist unsere Lösung für Bereiche, in denen es wenig Daten gibt. Es wurde das Paper von Omer Levy et al. “Improving Distributional Similarity with Lessons Learned from Word Embeddings” implementiert. Die Autoren der Studie hatten mit ihrem Paper Code veröffentlicht, dieses war jedoch nicht für den normalen Gebrauch geeignet. In diesem

Forschungsumfeld spielen die Qualität keine Rolle und die meiste Software ist sehr prototypisch. Die Method ist jedoch weiterhin State-of-the-Art und so entschied ich mich die Software weiter zu entwickeln. Dazu habe ich Teile wiederverwendet, aber viele Teile neu geschrieben. Mit der Software habe ich mehrere Word Embeddings trainiert und auch zu weiteren Verwendung zum download angeboten.

Die Mitarbeiter und Mitarbeiterinnen der Open Knowledge Foundation Deutschland haben mir bei Fragen zur Förderung stets zurate. Ich konnte mit ihnen auch inhaltliche Fragen in Bezug zur Öffentlichkeitsarbeit klären. Zudem habe ich mich mit mehreren Akteuren ausgetauscht, die sich intensiv mit der Kommentarkultur im Internet beschäftigen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung? Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Durch die Open-Source-Lizenz kann eine Weiterarbeit problemlos erfolgen. Ich bin bereits in Gesprächen mit Forschern, die den Code für eine Veröffentlichung benutzen wollen.

In dem Projekt konnte ich vor allem meine technischen Fähigkeiten weiterentwickelt. Es war wichtig, dass der Code möglichst effizient läuft.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

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

Zunächst gab es die Planung direkt im Browser Machine Learning durchzuführen. Erste Tests mit Nutzer:innen ergaben jedoch, dass es für sei irrelevant ist, wie die Idee genau technisch umgesetzt werden sollte. So kam es durch eine Überarbeitung der Projektplanung und der finalen Ergebnisse. Am Anfang gab es den Plan richtiges Hass auf der Webseite darzustellen, jedoch wurde die Idee verworfen. Dieser sollte nicht noch mehr Aufmerksamkeit erhalten.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

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

Die Ergebnisse werde final auf der Webseite https://kommentare.vis.one präsentiert werden. Die Software ist unter https://github.com/jfilter/ptf zu finden. Da die Arbeit in mehrere kleinere Softwarepakete aufgeteilt it, gibt es Auflistung mit Links und Erläuterungen.

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?

Es gab einen Mehraufwand zu Ende des Projektes, sodass die finale Veröffentlichung verschoben werden musste. Wahrscheinlich wird sie im Oktober 2019 erfolgen wird.

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?

Nein.