Präsentation: 24.06. 16:00 Uhr

Motivation(10pts.):

Sie stellen anhand sinnvoller Vorarbeiten die Relevanz des Themas dar und formulieren klare Forschungsfragen

Data Set(5pts.):

Sie haben den Datensatz analysiert und zur späteren Benutzung aufbereitet. Sie kennen die Hintergründe zum Datensatz und zeigen Querbezüge zu den Forschungsfragen und später verwendeten Methoden auf

Methoden(5pts.):

Die gewählten Methoden ergeben sich schlüssig aus der Problemstellung, den Forschungsfragen sowie dem gewählten Datensatz. Es wird sinnvoll und gewissenhaft mit Referenzen und Verweisen auf bestehende Arbeiten gearbeitet und das geplante methodische Vorgehen ist so vollkommen nachvollziehbar.

NLP Pipeline(5pts.):

Die Bestandteile der NLP Pipeline sind sinnvoll und nachvollziehbar gewählt und ergeben sich schlüssig aus den Anforderungen aus der Wahl des Datensatzes und dem Methodensatz

Prototyp und erste Ergebnisse(5pts.):

Der vorgestellte Prototyp lässt bereits gut erkennen, wie die methodische Ausarbeitung in der finalen Arbeit erfolgen wird. Die gezeigten Ergebnisse demonstrieren, dass die gewählte Methodik in der Lage ist, sinnvolle Ergebnisse zu produzieren.

Präsentation(10pts):

Die Präsentation ist ausgefeilt, verständlich und man kann dem Sprecher gut folgen.

Forschungsfrage Ideen:

Lässt sich eine klare Haltung der Washington Post zu Waffen erkennen (Gewalt/Gesetzen), was anhand einer Stance-Analyse des erhaltenen Datensatzes über einen Zeitraum von 10+ Jahren analysiert werden kann?

These

Die Haltung der Washington Post ist im Verlauf der letzten Jahre immer negativer

Erstes Abstract ohne wirklichen Inhalt

Medien haben ob gewollt oder ungewollt immer einen Einfluss auf Personen, die diese konsumieren. Auf Grundlage dessen wollen wir herausfinden, welche Ideologie die Washington Post zum Thema Waffen besitzt und ob sich diese im Verlauf der Jahre verändert hat, bzw. zu bestimmten Ereignissen die Haltung in den Artikeln der Washington Post deutlicher zu erkennen ist.

Um dies zu erreichen haben wir einen Datensatz von randomisierten Washington Post Artikeln erstellt und auf Grundlage dieses Datensatzes versucht zu analysieren, ob sich eine bestimmte Haltung für oder gegen das Thema Waffen herauskristallisiert.

Themen:

Waffen, Waffengewalt, Waffengesetze, NRA, Krieg, Armee, (Freiheit? Selbstverteidigung?)

**Motivation:**

In den vergangenen Jahren ist es immer wieder dazu gekommen, dass durch verschiedene Medien versucht wurde Einfluss auf ein bestimmtes Thema zu nehmen. Auf Grund dessen wollen wir mit Hilfe des zur Verfügung gestellten Datensatzes herausfinden, ob sich bei Artikeln der Washington Post im Verlauf der letzten 10 Jahre eine bestimmte Haltung zum Thema Waffen erkennen lässt.

Hier Forschungsfrage einfügen

Hierzu wollen wir verschiedene im Verlauf des Moduls NLP erlernte Methoden nutzen, um so eine Stance Analyse zu entwickeln welche uns Aufschluss darauf bietet ob und wenn ja welche Haltung die Washington Post zum Thema Waffen allgemein besitzt.

**Data Set:**

Der zur Verfügung gestellte Datensatz enthält mehr als 14GB 728k

Der zur Verfügung gestellte Datensatz hat eine Größe von ca. 14 GB und enthält insgesamt mehr als 780 tausend Artikel, welche zur Analyse genutzt werden können.

Um mit diesen Daten arbeiten zu können haben wir aus dieser Menge an Daten einen zufälligen Datensatz generiert, welcher 12 Tausend Artikel enthält.

Aus diesem Datensatz haben wir schließlich alles entfernt bis auf „contents“, da in diesem alle für uns relevanten Daten enthalten waren.

Diese Daten enthielten:

Thema

Titel

Bild?

Autor

Datum

Artikelinhalt (content)

**Methoden:**

**NLP Pipeline:**

Importieren des Datensatzes

Preprocessing der Daten:

Zufällige Auswahl an Daten für die Analyse (da schlicht zu viel)

Bereinigung der vorhandenen Daten ( z.B. war das Datum nicht klar erkennbar?)

Tokenisierung des Datensatzes unter zu Hilfenahme von spacy?

Enfernen von Stoppwörtern durch spacy

POS Tagging

**Prototyp und erste Ergebnisse**