FernUniversität in Hagen
Fakultät für Mathematik und Informatik

Artificial Intelligence Group

Web Science - Die Qualität von Wikipedia-Artikeln

Fachpraktikumsbericht

vorgelegt von
Robin Suxdorf, Sebastian Bunge, Johannes Krämer, Emmanuelle Steenhof, Alexander Kunze

Zusammenfassung

Erstmal nur ein Draft. Inhalte werden weiter abgestimmt.

1 Einleitung

Das Ziel dieses Projektpraktikums ist die praktische Anwendung von Methoden des maschinellen Lernens auf einen vorgegebenen Datensatz aus dem Bereich der Web Science. Wir haben uns für das Thema **Qualität von Wikipedia-Artikeln** entschieden und nutzen dafür den Datensatz von Kaggle: `https://www.kaggle.com/datasets/urbanbricks/wikipedia-promotional-articles`
Im Rahmen dieses Projekts bearbeiten wir folgende Teilaufgaben:

1. **Analyse des Datensatzes und Identifizierung einer geeigneten Problemstellung**: Wir untersuchen den bereitgestellten Datensatz eingehend, um ein maschinelles Lernproblem zu formulieren, das mit den vorhandenen Daten gelöst werden kann.

2. **Aufbereitung und Vorverarbeitung des Datensatzes**: Wir bereinigen und transformieren die Daten, um sie für die Modellierung vorzubereiten.

3. **Anwendung von drei klassischen Methoden des maschinellen Lernens**: Basierend auf den Inhalten der Kapitel 2 und 3 des Kurses „Einführung in Maschinelles Lernen“ implementieren wir drei klassische Algorithmen, um die identifizierte Problemstellung zu adressieren.

4. **Anwendung eines Deep-Learning-Ansatzes**: Wir recherchieren einen geeigneten Deep-Learning-Ansatz, setzen diesen um und wenden diesen auf die Problemstellung an

5. **Entwickeln eines eigenen Ansatzes**: Im Rahmen dieser Ausarbeitung wird ein eigener Ansatz für die Problemstellung entwickelt und beschrieben.
6. **Interpretation und Diskussion der Ergebnisse**: Basierend auf den bisherigen Resultaten entwickeln wir eine neue Idee für einen passenden Ansatz, beispielsweise eine neue Architektur für ein neuronales Netzwerk, die wir implementieren und anwenden.

2 Aufgabenverteilung

Im Kick-Off Meeting wurde Robin Suxdorf als Teamleiter und Kommunikationskanal zu den Praktikumsbetreuern gewählt. Für die Umsetzung der Teilaufgaben wurden jeweils verantwortliche bestimmt:

1. Klassische Methode 1: Ansatz: Bayes Leiter: Sebastian Bunge
2. Klassische Methode 2: Ansatz: SVM Leiter: Johannes Krämer
3. Klassische Methode 3: Ansatz: Logistische Regression Leiter: Alexander Kunte
4. Deep-Learning Methode: LSTM Transformer oder Ansatz über Embeddings; Zuerst einmal werden Embeddings angeschaut Leiter Robin
5. Eigener Ansatz: wird noch genauer angeschaut Leiter: Emmanuelle

Während des gesamten Praktikums schreibt Alexander Kunze fortlaufend den Praktikumsbericht weiter. -Weitere Themen: Präsentation, Vorträge, usw.

3 Teaminterne Organisation

Im Rahmen des Kick-Offs wurde beschlossen, dass Discord (bereitgestellt über Alexander Kunze und Github (bereitgestellt von Robin Suxdorf) als Kollaborationsplattformen dienen. Ein wöchentlicher Jour-Fixe sichert den regelmäßigen Austausch. Jeder Teilnehmer verantwortet die Weiterentwicklung seiner Methode. Das bedeutet, er entwickelt die Methode weiter, gibt zum Jour-Fixe ein Update zum Stand und teilt mit, wenn es Herausforderungen gibt. Das Team unterstützt dabei jeden Leiter und gibt Feedback bei jeder Statusvorstellung.

4 Datensatz und Problemstellung

4.1 Datensatz

Die Datensätze stammen von Kaggle: `https://www.kaggle.com/datasets/urbanbricks/wikipedia-promotional-articles`. Ein Datensatz enthält Wikipedia-Artikel, die als *promotional* (also werbend) klassifiziert sind. Dabei sind folgende Label vergeben:

- advert – „Dieser Artikel enthält Inhalte, die wie eine Werbeanzeige verfasst sind.“
- coi – „Ein Hauptautor dieses Artikels scheint eine enge Verbindung zu seinem Thema zu haben.“
- fanpov – „Dieser Artikel ist möglicherweise aus der Sicht eines Fans geschrieben, statt aus einer neutralen Perspektive.“
- pr – „Dieser Artikel liest sich wie eine Pressemitteilung oder ein Nachrichtenartikel oder basiert weitgehend auf routinemäßiger Berichterstattung oder Sensationslust.“
- resume – „Dieser biografische Artikel ist wie ein Lebenslauf geschrieben.“

Der zweite Datensatz enthält Wikipedia Artikel die *nicht-promotional* klassifiziert sind.

4.2 Problemdefinition

Das Ziel dieses Projekts ist die Entwicklung von Modellen zur automatisierten Klassifikation von Wikipedia-Artikeln als *promotional* (werblich) oder *nicht-promotional*. Wikipedia strebt nach objektiven und neutralen Inhalten; daher ist die Identifizierung von Artikeln mit werbenden Charakter von großer Bedeutung, um die sachliche Qualität der Plattform zu gewährleisten.

4.3 Zielsetzung

Die Hauptziele des Projekts sind:

- Entwicklung von drei klassischen maschinellen Lernmodellen und einem Deep-Learning-Modell zur Klassifikation von Wikipedia-Artikeln.
- Vergleich der Modelle anhand von Leistungsmetriken wie Genauigkeit, Präzision, Recall und F1-Score.
- Identifikation des Modells mit der besten Leistung für die gegebene Aufgabe.

5 Ansätze

Blubb

6 Experimente

Blubb

6.1 Evaluationsmetriken

1. Sei $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ ein Datensatz und $clf : \mathbb{R}^n \to \{0, 1\}$ ein (binärer) Klassifikator. Das **Genauigkeitsmaß** acc von clf bezüglich D ist definiert durch

$$acc(D, clf) = \frac{1}{m} \sum_{i=1}^{m} \left(1 - \left|y^{(i)} - clf(x^{(i)})\right|\right) \tag{1}$$

2. Wir definieren

$$TP(D, clf) = |\{i \mid y^{(i)} = 1, clf(x^{(i)}) = 1\}| \tag{2}$$
$$TN(D, clf) = |\{i \mid y^{(i)} = 0, clf(x^{(i)}) = 0\}| \tag{3}$$
$$FP(D, clf) = |\{i \mid y^{(i)} = 0, clf(x^{(i)}) = 1\}| \tag{4}$$
$$FN(D, clf) = |\{i \mid y^{(i)} = 1, clf(x^{(i)}) = 0\}| \tag{5}$$

   Die *Konfusionsmatrix* von clf bzgl. D stellt die vier oben genannten Werte tabellarisch wie folgt dar:

	$y = 1$	$y = 0$
$clf = 1$	$TP(D, clf)$	$FP(D, clf)$
$clf = 0$	$FN(D, clf)$	$TN(D, clf)$

3. Sei $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ ein Datensatz und $clf : \mathbb{R}^n \to \{0, 1\}$ ein (binärer) Klassifikator. Definiere

   - **Präzision:**
$$\text{prec}(D, clf) = \frac{TP(D, clf)}{TP(D, clf) + FP(D, clf)} \tag{6}$$
   - **Recall:**
$$\text{rec}(D, clf) = \frac{TP(D, clf)}{TP(D, clf) + FN(D, clf)} \tag{7}$$
   - **F1:**
$$\text{F1}(D, clf) = \frac{2 \cdot \text{prec}(D, clf) \cdot \text{rec}(D, clf)}{\text{prec}(D, clf) + \text{rec}(D, clf)} \tag{8}$$

7 Ausblick

Blubb

8 Zusammenfassung und Fazit

Blubb

Literatur

Erklärung

Ich erkläre, dass ich die schriftliche Ausarbeitung zum Fachpraktikum selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe. Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht. Die Versicherung selbstständiger Arbeit gilt auch für enthaltene Zeichnungen, Skizzen oder graphische Darstellungen. Die Ausarbeitung wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Mit der Abgabe der elektronischen Fassung der endgültigen Version der Ausarbeitung nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate geprüft werden kann und ausschließlich für Prüfungszwecke gespeichert wird.

.....
(Ort, Datum) (Unterschrift)

.....
(Ort, Datum) (Unterschrift)

.....
(Ort, Datum) (Unterschrift)

.....
(Ort, Datum) (Unterschrift)

.....
(Ort, Datum) (Unterschrift)