Explainable Authorship Verification

AI Studienprojekt, WiSe 19/20

Benedikt Boenninghoff

July 3, 2019
Forschungsgruppe "Sprachliche Imitations- und Verschleierungsstrategien"

- **Motivation:**
  - Erpresserschreiben, Bekennerschreiben
  - Hatespeech in sozialen Medien
  - Chef-Masche: Emails mit Zahlungsaufforderungen
  - Foren für Kinder und Jugendliche

- **Ziel:** Entwicklung von Methoden für die automatisierte Analyse der Autorenschaft
  - Können Verschleierungsstrategien detektiert werden?
Authorship verification for social media data [Boenninghoff et al. 2019]

Objective of authorship verification

Determine if two documents were written by the same person

Challenges:

- Social media texts are relatively short (less than 1000 words)
- Large variety of different topics
- Writing styles differ from those of news texts
Visualizing the *decision-making* process

- **Blue:** denotes the sentence-based importance
- **Red:** implies the relative word importance

**Review excerpt 1:**

Tony Starks is one of the most reconivable voices and styles in hip hop. On this cd he reminds us why he was the voice of wu-tang. This is another hard Ghost face cd with very good production and great ironman lyrics and flow. This is a cd that can easily be listen to start to finish. Has good features on it. Most of the wu-tang clan is featured on one song including the late dirt migart. Not to mention a suprise for me but biggie and rekwon both on the same track, and that track is def tight [...]
Layer-wise Relevance Propagation (LRP) [Arras et al. 2017]

- A method that identifies important words by running a backward pass in the neural network. ¹

<table>
<thead>
<tr>
<th>true</th>
<th>predicted</th>
<th>N° Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>do n't waste your money .</td>
<td>1. ++</td>
</tr>
<tr>
<td></td>
<td>neither funny nor suspenseful nor particularly well-drawn .</td>
<td>2. +</td>
</tr>
<tr>
<td></td>
<td>it 's not terrible , just terribly mediocre .</td>
<td>3. -</td>
</tr>
<tr>
<td></td>
<td>... too slow , too boring , and occasionally annoying .</td>
<td>4. --</td>
</tr>
<tr>
<td></td>
<td>it 's neither as romantic nor as thrilling as it should be .</td>
<td>5.</td>
</tr>
<tr>
<td></td>
<td>the master of disaster - it 's a piece of dreck disguised as comedy .</td>
<td>6.</td>
</tr>
<tr>
<td></td>
<td>so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .</td>
<td>7.</td>
</tr>
<tr>
<td></td>
<td>a film so tedious that it is impossible to care whether that boast is true or not .</td>
<td>8.</td>
</tr>
<tr>
<td></td>
<td>choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .</td>
<td>9.</td>
</tr>
<tr>
<td></td>
<td>this idea has lost its originality ... and neither star appears very excited at rehealing what was basically a one-joke picture .</td>
<td>10. --</td>
</tr>
<tr>
<td></td>
<td>ecks this one off your must-see list ;</td>
<td>11. ++</td>
</tr>
<tr>
<td></td>
<td>this is not a Friday 's worth waiting for ;</td>
<td>12.</td>
</tr>
<tr>
<td></td>
<td>there is not an ounce of honesty in the entire production ;</td>
<td>13.</td>
</tr>
<tr>
<td></td>
<td>do n't expect any surprises in this checklist of teamwork cliches ...</td>
<td>14.</td>
</tr>
<tr>
<td></td>
<td>he has not learnt that storytelling is what the movies are about ;</td>
<td>15.</td>
</tr>
<tr>
<td></td>
<td>but here 's the real damn : it is not funny , either .</td>
<td>16.</td>
</tr>
<tr>
<td></td>
<td>these are names to remember ; in order to avoid them in the future ;</td>
<td>17.</td>
</tr>
<tr>
<td></td>
<td>the cartoon that is not really good enough to be on afternoon tv is now a movie that is not really good enough to be in theaters .</td>
<td>18.</td>
</tr>
<tr>
<td></td>
<td>a worthy entry into a very difficult genre .</td>
<td>19. ++</td>
</tr>
<tr>
<td></td>
<td>it 's a good film -- not a classic , but odd , entertaining and authentic .</td>
<td>20.</td>
</tr>
<tr>
<td></td>
<td>it never fails to engage us .</td>
<td>21. --</td>
</tr>
</tbody>
</table>

¹http://heatmapping.org/
Aufgabenstellung

(1) **Einarbeitung und Recherche**
   - Authorship Verification
   - Tensorflow
   - Explainable Neural Networks

(2) **Programmierung**
   - Implementierung eines Frameworks für die Autorschafts-Verifikation
   - Integration von LRP

(3) **Evaluierung und Visualisierung**

(4) **Dokumentation der Ergebnisse**
Organisation und Kontakt

Eckdaten:
▶ Beginn: 01.10.2019 (erstes Treffen vorher)
▶ Ende: 07.02.2020
▶ Wöchentliche (kurze) Treffen
▶ Teilnehmerzahl: 2 (Master)

Wünschenswerte Kenntnisse bzw. Fähigkeiten:

- Python / Tensorflow
- Machine Learning / Deep Learning
- Linux (Ubuntu), Git, Latex

Sehr gut geeignet für:
▶ Studierende mit Schwerpunkt Computerlinguistik
  oder Machine Learning

Ansprechpartner/Kontakt bei Fragen:
▶ Email: benedikt.boenninghoff[at]rub.de
▶ Raum: ID 2-334