Bias in Datensätzen und ML-Modellen: Erkennung und Umgang in den DH

Lassner, David
lassner@tu-berlin.de
TU Berlin

Brandl, Stephanie
stephanie.brandl@tu-berlin.de
TU Berlin

Guy, Louisa
louisa.guy.etu@univ-lemans.fr
Le Mans Université

Baillot, Anne
anne.baillot@univ-lemans.fr
Le Mans Université

Vortragende

David Lassner

Master Informatik David Lassner, Doktorand an der TU Berlin im Bereich Maschinelles Lernen für Digital Humanities, insbesondere für quantitative Literaturanalyse.

Stephanie Brandl

Dipl. Math. Stephanie Brandl, Technische Universität Berlin. Forschungsschwerpunkte: Maschinelles Lernen, Natural Language Processing.

Louisa Guy

Louisa Guy, Doktorandin, Le Mans Université. Forschungsinteressen: Digitale Textanalyse, Anwendung von Methoden der Computerlinguistik auf sozialwissenschaftliche Kontexte.

Anne Baillot

Prof. Dr. Anne Baillot, Le Mans Université. Forschungsschwerpunkte: Digitale Philologie, Digital Humanities, Translation Studies.

Anforderungen

Maximalanzahl Teilnehmender: 25

Räumliche Anforderungen:

- Beamer
- Whiteboard/Tafel
- Stromversorgung für Laptops der Teilnehmenden
- Wifi

Anforderungen an die Teilnehmenden:

Wir erwarten, dass die Teilnehmenden ihre eigenen Laptops mitbringen, die bestenfalls schon die nötige Software vorinstalliert haben. Wir werden kurz vor der Konferenz eine Willkommens-E-Mail mit den Softwareanforderungen verschicken. Die praktischen Sitzungen werden mithilfe von Jupyter Notebooks (Python3, Jupyter) abgehalten. Wir planen zusätzlich als Absicherung einen Online-Zugang zu einem JupyterHub Server mit vorinstallierten Paketen für Teilnehmende, bei denen die Installation Schwierigkeiten macht. Die praktischen Sitzungen sind so konzipiert, dass nur sehr geringe, bis gar keine Programmierkenntnisse notwendig sind. Im Wesentlichen sollen die Teilnehmenden die Parameter und Eingabedaten der vorgegebenen Programme modifizieren, Teilnehmende mit mehr Programmierkenntnissen ermutigen wir natürlich tiefer in die Programme einzusteigen und auch diese zu modifizieren.

Beschreibung

Der Workshop besteht aus einem allgemeineren Teil zu Bias im Maschinellen Lernen, in dem grundlegend in die Thematik eingeführt wird, und einem spezifischeren Teil, in dem ML-Biases im Kontext von DH behandelt werden. Beide Teile beinhalten Vortrags- sowie Mitmachsessions. Ziel des Workshops ist es, dass die Teilnehmenden sich des Problems von Bias in Machine Learning Modellen bewusst werden und die grundlegenden Techniken zur Erkennung und zur Unterdrückung von Biases kennenlernen. Es soll außerdem gemeinsam erarbeitet werden, auf welche Weise DH-ForscherInnen mit den Biases umgehen können - denn in vielen Anwendungen sind diese nicht gewünscht: Ein System zur Vorauswahl von Bewerbern sollte Männer nicht bevorzugen,[1] ein Modell zur Gesichtserkennung sollte keinen Unterschied in der Genauigkeit haben, weil sich die Hautfarbe der Personen auf den Bildern ändert (Buolamwini et al. 2018), und ein Modell zur Erkennung von Hate-Speech im Internet sollte nicht kontextfrei bspw. Begriffe wie “homosexuell” als toxisch einstufen.[2]

Gleichzeitig können Biases in ML Modellen erwünscht sein, wenn man beispielsweise die Veränderung von Biases in Sprache analysiert.

Teilnehmende werden im Vorfeld ermutigt eigene Daten mitzubringen, mit denen sie im zweiten praktischen Teil experimentieren können.

Das Workshopprogramm wird online unter bias-ml-dh.davidlassner.com öffentlich zur Verfügung gestellt und die Kursmaterialien auf Github unter github.com/millawell/bias-ml-dh veröffentlicht. Dort sollen die Teilnehmenden auch schon im Vorfeld einen Eindruck bekommen, welche ihrer eigenen Daten möglicherweise zum Workshop mitgebracht werden könnten.

Zeittafel

Zeit	Titel	Vortragende
Halbtag 1.1	Einleitung, Motivation	Anne Baillot, David Lassner
Halbtag 1.2	Erkennung von Biases in ML	David Lassner
Kaffepause		
Halbtag 1.3	Verhinderung von Biases in ML	Stephanie Brandl
Halbtag 1.4	Praktische Sitzung 1	
Halbtag 2.1	Autorinnen um 1800	Anne Baillot
Halbtag 2.2	Revolte auf Twitter	Louisa Guy
Kaffepause		
Halbtag 2.3	Praktische Sitzung 2	
Halbtag 2.4	Abschlussdiskussion	

Erkennung von Biases

Zu Beginn steht die Begriffsklärung (Datenbias, Modellbias, etc.) und konkrete Beispiele zur Erkundung verschiedener Biases in verschiedenen Datensätzen, sowie Modellarchitekturen. Beispielsweise anhand konkreter Architekturen neuronaler Netze zur Textklassifikation, deren erster Layer ein Embedding-Layer auf Word2Vec-Basis ist (Mikolov et al. 2013).

Es werden verschiedene Methoden vorgestellt, wie Biases in Modellen und Daten erkannt werden können (Caliskan et al. 2017, May et al. 2019, Garg et al. 2018, Bolukbasi et al. 2016, Swinger et al. 2019).

Wie lassen sich Biases verhindern?

Innerhalb der letzten 3 Jahre wurden zahlreiche Methoden veröffentlicht, die darauf abzielen Biases in Word Embeddings und anderen NLP Anwendungen zu verringern. In diesem Teil wollen wir einen Überblick über die wichtigsten Methoden verschaffen, ihre Stärken und Schwächen aufzeigen und diskutieren.

Aktuell können diese Methoden in 3 Kategorien eingeteilt werden:

Manipulation von Datensätzen

Datensätze werden so verändert, beispielsweise durch Datenanreicherung, dass Biases im Datensatz nicht mehr zu finden sind und so auch nicht mitgelernt werden. Zum Beispiel schlagen Zhao et al (2018) vor, jeden Satz in einem Datensatz zu kopieren, sodass dieser in mehreren Varianten vorkommt: eine für jedes grammatikalische Geschlecht. So wird eine balancierte Repräsentation zwischen den (binären) Geschlechtern garantiert. Bestehende ML-Methoden die ansonsten biased Ergebnisse erzeugen, können so faire Modelle lernen.

Anpassung der Methode

Zhang et al (2018) schlagen vor den Einfluss geschützter demografischer Informationen wie Geschlecht oder Postleitzahl auf das Klassifikationsergebnis mit Adversarial Learning zu verringern. Drei verschiedene Definitionen von „equality“ und Parität werden analysiert und für jeden Definition wird eine entsprechende Strategie vorgestellt um demografische Parität zu sichern.

Zusätzlicher Analyseschritt

Bolukbasi et al (2016) zeigen, dass mit Hilfe von Wortlisten ein Unterraum errechnet werden kann, der die geschlechtsbezogene Information in Word Embeddings beinhaltet. Wörter werden mit Hilfe dieser Wortlisten in geschlechtsneutral (z.B. doctor) und geschlechtsspezifisch (z.B. grandmother) eingeteilt. In dem entsprechenden Unterraum werden dann alle Wörter, die grammatikalisch geschlechtsneutral sind, auch neutralisiert, so dass beispielsweise *doctor* zentriert zwischen den Word Embeddings für “Mann” und “Frau” liegt.

Allerdings zeigen auch einige dieser Methoden Schwächen und es wurde bereits gezeigt, dass in vielen Fällen Biases weiterhin rekonstruiert werden können (Gonen & Goldberg, 2019).

Praktische Sitzung 1

Im ersten praktischen Teil sollen dann ML Modelle selbst ausprobiert und werden und, anhand von verschiedenen Analysemethoden, Biases explorativ erkundet werden.

Wir stellen eine fertige ML-Pipeline zur Textklassifizierung zur Verfügung, die mit vortrainierten Word Embeddings arbeitet. Die Klassifizierung soll dahingehend analysiert werden, welche Biases sie enthält. Dann sollen die vortrainierten Word Embeddings mithilfe von Tensorflow Projector erkundet werden und es sollen Richtungen identifiziert werden, die für die Biases

in den Ergebnissen verantwortlich sein könnten. Die Teilnehmenden sollen die vortrainierten Word Embeddings auf Grundlage ihrer Erkenntnisse modifizieren und untersuchen, wie sich das Klassifikationsergebnis dadurch ändert.

Des Weiteren sollen die Biases dieser Pipeline mithilfe von standardisierten Wort-list Tests (SEAT, May et al. 2019 / WEAT, Caliskan et al. 2017) analysiert werden.

Zuletzt soll den Teilnehmenden auch die Möglichkeit gegeben werden, die Korpuszusammensetzung für das Training der Word Embeddings zu modifizieren und selber trainierte Word Embeddings anstelle der vortrainierten zu verwenden, beispielsweise mithilfe von Sampling, Vereinigung, Mitteln.

Erkenntnisgewinn für DH durch Untersuchung von Biases

Biases in historischen Textdatensätzen können auf Biases in den Gesellschaften ihrer Entstehung sowie in ihrer Aufbewahrungs- und Tradierungsgeschichte aufdecken. Mit Blick auf die wachsende Wichtigkeit von Cultural Heritage Studies in den Digital Humanities sind diese Art von Biases ein hochaktuelles Forschungsfeld (Garg et al 2018). Der Korpuskonstruktion muss in diesen Fällen allerdings besondere Sorgfalt beigemessen werden, da nur bei einem für die jeweilige Forschungsfrage möglichst ausgewogenen Korpus auch tatsächlich durch die Biases im Korpus auch auf die Biases in der Gesellschaft Rückschlüsse gezogen werden können (Underwood 2019, Bode 2020). Kurz gesagt birgt jeder Schritt in der Geschichte der zu untersuchenden Objekte die Gefahr eines unbewusst und ungewollt induzierten Bias, die der bewussten und gewollten Analyse von Biases im Wege stehen können.

Autorinnen um 1800

Digitale Methoden machen es möglich, das traditionelle Narrativ der Literaturgeschichte zu überdenken und damit Literatur in den Vordergrund zu rücken, die etwa aus Gendergründen im Kanon als zweitrangig überliefert worden war. Zumindest machen sie es theoretisch möglich: Es soll nämlich gezeigt werden, dass digitale Korpora und Methoden die Biases der traditionellen Historiographie auch im literarischen Bereich nur zu leicht reproduzieren und dass die Korpusbildung und der Trainingsprozess einer besonderen Zuspitzung brauchen, um z.B. die Rolle von schreibenden Frauen deutlich machen zu können. Argumentiert wird hier am Beispiel von Autorinnen aus der Zeit um 1800 – der Phase nämlich, wo der (wohl männliche) Autor sich als literarischer, wirtschaftlich tragfähiger Wert etabliert.

Tweetanalyse von #aufschrei und #blacklivesmatter

Auf dem sozialen Netzwerk Twitter führten die Hashtags „aufschrei“ und „blacklivesmatter“ 2013 zu kollektiven Revolten, die online begannen, sich dann aber auch auf den Alltagsdiskurs ausweiteten. Unter #aufschrei berichteten Frauen über ihre Erfahrungen mit Sexismus und unter #blacklivematters ging es um Erlebnisse mit Rassismus. An diesem Beispiel werden Methoden zur Quellenanalyse vorgestellt. Ziel ist es, die Dynamik der digitalen Bewegungen von #aufschrei und #blacklivesmatter anschaulich zu machen.

Praktische Sitzung 2 und Abschlussdiskussion

Im zweiten praktischen Teil sollen die Teilnehmenden ihre eigene Expertise einbringen und in Gruppen individuelle Fragestellungen formulieren, die mithilfe der zuvor kennengelernten Modelle untersucht werden können. Wenn möglich, sollen sofort erste Prototypen entwickelt werden.

Falls Teilnehmende keine eigenen Korpora bzw. Fragestellungen mitbringen, stellen wir eine ML-Pipeline zur Verfügung, die existierende Systeme zur Erkennung von Hatespeech im Internet auf Tweets mit dem Hashtag #aufschrei bzw. #blacklivesmatter sowie einer Kontrollgruppe aus zufälligen anderen Tweets anwendet. Mithilfe dieser Pipeline sollen Teilnehmende untersuchen, wie Sprache einer neu entstehenden Bewegung, die nicht dem Mainstream entspricht, möglicherweise automatisch als Hatespeech erkannt wird.

Fußnoten

1. Was tatsächlich bei Amazon passiert ist: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
2. Beispiel von https://twitter.com/jessamyn/status/900867154412699649 bezüglich des www.perspectiveapi.com Interfaces, außerdem Davidson et al. (2019)

Bibliographie

Bode, Katherine (Forthcoming 2020): Why You Can’t Model Away Bias, Modern Language Quarterly 81.1. preprint: katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf [letzter Zugriff 27. September 2019].

Bolukbasi, Tolga / Kai-Wei Chang / James Y Zou / Venkatesh Saligrama / Adam T Kalai (2016): Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Conference of NIPS.

Buolamwini, Joy / Timnit Gebru (2018): Gender shades: Intersectional accuracy disparities in

commercial gender classification. Conference on fairness, accountability and transparency.

Caliskan, Aylin / Joanna J. Bryson / Arvind Narayanan. (2017): Semantics derived automatically from language corpora contain human-like biases. Science 356.

Davidson, Thomas / Debasmita Bhattacharya / Ingmar Weber (2019): Racial Bias in Hate Speech and Abusive Language Detection Datasets. arXiv preprint arXiv:1905.12516.

Garg, Nikhil / Londa Schiebinger / Dan Jurafsky / James Zou (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences.

Gonen, H. / Yoav Goldberg (2019): Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. Conference of the NAACL.

May, Chandler / Alex Wang / Shikha Bordia / Samuel R. Bowman / Rachel Rudinger (2019): On Measuring Social Biases in Sentence Encoders. Conference of the NAACL.

Mikolov, T. / Chen, K. / Corrado, G. / Dean, J. (2013): Efficient estimation of word representations in vector space. In ICLR.

Sap, Maarten / Dallas Card / Saadia Gabriel / Yejin Choi / Noah A. Smith (2019): The Risk of Racial Bias in Hate Speech Detection. Conference of the ACL.

Swinger, Nathaniel / Maria De-Arteaga / Neil Heffernan IV / Mark Leiserson / Adam Kalai (2019): What are the biases in my word embedding?. Conference on Artificial Intelligence, Ethics, and Society (AIES).

Underwood, Ted (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Zhang, B. H. / Lemoine, B. / Mitchell, M. (2018): Mitigating unwanted biases with adversarial learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Zhao, J. / Wang, T. / Yatskar, M. / Ordonez, V. / Chang, K. W. (2018): Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876.