EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Wirtschafts- und Sozialwissenschaftliche Fakultät

LEHRSTUHL FÜR MARKETING
Prof. Dr. Dominik Papies

Universität Tübingen · LS für Marketing · Nauklerstr. 47 · 72074 Tübingen

Telefon +49 7071 29-76977
Telefax +49 7071 29-5078
dominik.papies@uni-tuebingen.de
www.uni-tuebingen.de/wiwi/marketing

Seminar "Ethical Aspects of Machine Learning Methods"

I. Type of seminar

"10 Ways Machine Learning Is Revolutionizing Marketing" – Headlines like this example from forbes.com[1] are omnipresent in the business press today. Machine Learning methods are now widely applied in many companies across different industry sectors and business functions as well as in the public sector. Some managers even seem to believe that Machine Learning and Artificial Intelligence can solve important business problems, and many analysts believe that Machine Learning (ML) will redefine how business is conducted. On top of that, Machine Learning has evolved as a major topic in the domain of law enforcement.

At the same time, concerns have been mounting over the appropriate use of Machine Learning methods. While, e.g., face recognition software has – without doubt – several advantages, its unregulated use in the public sphere has many disadvantages, in particular if combined with detection errors that bias against minorities. Other examples include algorithmic biases, which may lead to situations in which an algorithm recommends that a black defendant goes to jail, while a white defendant is free on probation.

The surge in the use of Machine Learning methods and the demand of policy makers and business leaders for Machine Learning and "AI" solutions makes it crucial that we understand sources of biases and other ethical aspects of Machine Learning methods. It is therefore the key goal of this seminar to explore these biases and problems, and develop a deep understanding of the opportunities and – in particular – the ethical challenges of Machine Learning and "AI". To this end, students will apply the methods on appropriate data, but will also devote sufficient space in their thesis to a reflection of ethical aspects of their empirical work. In addition to working with data and writing a thesis, we are planning a **virtual screening** of the movie "Coded Bias" for all seminar participants with subsequent discussion.

In this seminar, students will also acquire relevant tools to be prepared for writing a research-based master thesis. This will be supported by an obligatory workshop on academic research as well as an obligatory workshop on presentation skills, which includes a short presentation of each student's current state of the thesis ("research plan presentation"). On top of that, we expect and encourage active participation and interaction between students.

It is expected that students have **very solid skills in statistical software (preferably R or Python)**, equivalent to, e.g., a successful completion of DS400 Data Science Project Management and DS404 Data Science with Python. In addition, we expect that students are willing to **familiarize themselves** with new methods and approaches as well as new tools in R or Python. The respective supervisor will support students in this.

[1] https://www.forbes.com/sites/louiscolumbus/2018/02/25/10-ways-machine-learning-is-revolutionizing-marketing/

II. Topics and introductory reading material

Topic 1	**Algorithmic Bias in Image Classification** Neural networks are powerful tools that can make use of visual data. They showed extraordinary performance in image classification, recognition and even image generation. However, ethical considerations emerge due to potential algorithmic biases. Algorithmic biases are systematic errors that discriminate against certain characteristics and result in unfair outcomes. Underrepresentation of a certain group in the training data can lead to bad classification rates for this class like, for example, worse detection of people of color compared to white skinned people in face recognition. Using neural networks and a topic of their choice, students will explore ethical aspects and algorithmic biases in the context of image data.
Literature	Van Norden, Richard (2020): The ethical questions that haunt facial recognition research (https://www.nature.com/articles/d41586-020-03187-3) Buolamwini, Joy (2016): How I'm fighting bias in algorithms. TED Talk (https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare) AlgorithmWatch (2020): Google apologizes after its Vision AI produced racist results (https://algorithmwatch.org/en/story/google-vision-racism/) Chollet, F. & Allaire, J.J. (2018). *Deep Learning with R*. Manning Publications Company. Chollet, F. (2018). *Deep Learning with Python*. Manning Publications Company.
Data	There are various freely available datasets (e.g. https://www.face-rec.org/databases/, http://web.mit.edu/emeyers/www/face_databases.html, https://visionlab.is/stimuli-humans/). Students can analyze a dataset of their choice.

Topic 2 **Hate Speech Detection in Social Media**

Social media platforms, such as Twitter or Facebook, connect people worldwide from different religions, ethnicities, and nationalities. With social media comes the possibility for users to easily reach an incredibly large audience. While this possibility is used in many cases for good purposes, we are also increasingly seeing social media being used to spread hate speech. Despite the enormous technical capabilities of companies like Google, Facebook & Co., the detection and blocking of offensive posts still relies to a large extent on human judgment. There is an ongoing political debate as to how the dissemination of hate speech can be prevented, while maintaining an appropriate trade-off between censorship and freedom of expression. The recent ban of several social media accounts of then U.S. President Donald Trump has further fueled the debate about the responsibilities and power of social media platforms and the need for governmental regulation.

This thesis will develop and evaluate an algorithm to automatically detect offensive language in social media postings. It is an integral part of the thesis to identify and discuss potential ethical problems that may come along with such an automated approach.

Literature

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology* 22**,** 69–80. https://doi.org/10.1007/s10676-019-09516-z

MacAvaney S., Yao H.-R., Yang E., Russell K., Goharian N., & Frieder O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE* 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152

Bazelon, E. (2021, Jan. 26). Why Is Big Tech Policing Speech? Because the Government Isn't. *The New York Times Magazine.* https://www.nytimes.com/2021/01/26/magazine/free-speech-tech.html

Twitter API. https://developer.twitter.com/en/solutions/academic-research

Topic 3 **Algorithmic Bias against Minorities**

Machine learning is often used to aid human decision making, in particular for repetitive tasks. One popular example is use of algorithms to make decisions on credit loan approvals or rejections. However, with the rise of these automatic approaches, concerns have been growing that these algorithms may be biased to the extent that they discriminate against minorities or underprivileged groups in society, and this may even hold if problematic features (e.g., race) are omitted from prediction models. This can lead to, e.g., the rejection of credit applications to persons of color, whereas white persons would have received the loan, or higher probability of probation being negated for colored inmates.

Therefore, it is the goal of this thesis to study the problem of algorithmic bias and conflicting fairness definitions on a data set (e.g., the COMPAS data or data on credit loan decisions) that is prone to these algorithmic biases. One component of this thesis should be the trade-off between predictive accuracy and minimal bias against underprivileged groups.

Literature

Lee, M. S. A., & Floridi, L. (2020). Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs. Minds and Machines.

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science, 65(7), 2966–2981.

Mattu, J. A., Jeff Larson,Lauren Kirchner,Surya. (2016). *Machine Bias*. ProPublica. Retrieved 2 February 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Data

Publicly available data sets, e.g., from Kaggle, or other data sets that can be used to study this aspect of algorithmic bias (e.g., https://github.com/propublica/compas-analysis)

Topic 4 **Advances and Challenges in Face Detection and Recognition**

Algorithms to detect faces in images and video have been around since well before the 90s. However, due to their tremendous performance increase and the possibility to deploy such algorithms on internet browsers, smartphones, and smart home devices, face recognition is omnipresent in today's world (e.g., when unlocking your smartphone, entering a country, or finding people on social media).
Still, there are several challenges. Predictions based on faces are not always accurate and can be cheated, and face detection and recognition algorithms are often biased against certain groups of people (e.g., people of color). Face recognition in public areas, therefore, is now banned in some cities (e.g. Boston).

Using an algorithm and a specific topic of their choice, students should explore the advances and challenges in face detection and recognition. Examples may include but are not limited to topics such as mood and emotion detection, age and gender prediction, deceiving face recognition systems, limitations of face recognition (e.g. when wearing face masks), accuracy-speed-tradeoffs of face detection algorithms, or deepfakes.

Literature

Van Norden, Richard (2020): The ethical questions that haunt facial recognition research, https://www.nature.com/articles/d41586-020-03187-3

Rosebrock, Adrian (2018). OpenCV Face Recognition, https://www.pyimagesearch.com/2018/09/24/opencv-face-recognition/

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Conference on Computer Vision and Pattern Recognition (CVPR). https://research.fb.com/wp-content/uploads/2016/11/deepface-closing-the-gap-to-human-level-performance-in-face-verification.pdf

Serengil, S. I. & Alper, O. (2020). LightFace: A Hybrid Deep Face Recognition Framework. Innovations in Intelligent Systems and Applications Conference (ASYU), https://github.com/serengil/deepface

Data

There are various freely available datasets (e.g. https://www.face-rec.org/databases/, http://web.mit.edu/emeyers/www/face_databases.html, https://visionlab.is/stimuli-humans/, https://datasetsearch.research.google.com/). Students can analyze a dataset of their choice that fits well with their research question.

- III. Dates

9th February – 11th April 2021	Online Application – please see our website for further information.
April 16, 2021	5.45 pm – 7.15 pm Kick-off and topic assignment Workshop video „Academic Writing" available online
April 22, 2021	7.30 pm - 9.30 pm Virtual Screening of the Movie "Coded Bias"
April 30, 2021	9 am – 10 am: 1st Virtual Q&A Session on Programming with R (send us your questions via email in advance)
May 3, 2021	Video "Presentation Skills" online available
May 10, 2021	9 am – 10 am: Virtual Q&A Session on Presentation Skills 10 am – 11 am: 2nd Virtual Q&A Session on Programming (send us your questions via email in advance)
May 20, 2021	All day Research Plan Presentation
June 25, 2021	12 noon s.t. Term paper is due (you can drop your term paper in the letterbox outside the faculty (addressed to Chair of Marketing - Nauklerstr. 47) or send it by post (postmark date is relevant).) Containing 2 versions of the term paper with a filing clip (https://de.wikipedia.org/wiki/Heftstreifen) The electronic version (pdf) of the term paper incl. analyses as file upload in ILIAS.
July 9, 2021	All day (dates will be coordinated individually) Feedback Session
July 25, 2021	8 pm s.t. Upload Presentation in ILIAS
July 26, 2021	Seminar (all day)

IV. Course credits

Students can obtain course credit (9 ECTS). To obtain course credit students must meet the following criteria:

- Students participate in all meetings listed above
- Students submit their 12-page thesis on time
- Students present their thesis during the seminar
- Students actively participate during the seminar

Tübingen, April 2021