This document is publicly editable, as of this time. By contributing to this document, you agree that your contributions are given to the public domain for posterity. The tutorial presenters may further organize or edit the document as we see fit before finalizing the document for general public dissemination. Please ensure any materials that you place in this document are suitable for public use: images should be credited for fair use, and be explicitly ok for use in public materials, documents quoted should retain a note to their source.

Understanding Ethics in NLP Authoring and Reviewing

Videoconferencing Link: Join Live Session

QR Code links

Activity Slides
http://example.com/ethics-comp-slides

Presentation Slides
http://example.com/ethics-slides

Scribing document
http://example.com/ethics-doc
(you are here)

Activity Links:

0) Tell us about you
1) What else?
2) Final Call to Action
3) Your Feedback on this Tutorial

Program

1. Introduction and Foundations for Ethics by Presenters
2. Case Studies: Problematic Ethical Research by all
3. Structured Interaction / Dialogue Presenters
4. Case studies — Second reading by all
5. Group Presentations by Group Leads
6. Summary and Common Issues by Presenters
7. Discussing and Troubleshooting Ethics and Further Resources by Presenters

This document is publicly editable, as of this time. By contributing to this document, you agree that your contributions are given to the public domain for posterity. The tutorial presenters may further organize or edit the document as we see fit before finalizing the document for general public dissemination. Please ensure any materials that you place in this document are suitable for public use: images should be credited for fair use, and be explicitly ok for use in public materials, documents quoted should retain a note to their source.

---

0) Tell us about you

Duration: 3 minutes

Add your name and where you are visiting from

- E.g., Karën Fort from France

1a) Case Studies in Problematic Abstracts

Use this document to co-construct your thoughts on these abstracts.
As groups, construct a slide on the respective slide deck http://bit.ly/eacl23-ethics-comp-slides to present.

Abstract 1: Faces contain more information about sexual orientation than can be perceived by the human brain. We used deep neural networks to extract features from over 35 thousand facial images. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 80% of cases, and in 70% of cases for women. Accuracy increased to 90% and 80%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone detected with 55% and 53% accuracy for gay males and gay females, respectively. Such findings advance our understanding of the origins of sexual orientation and the limits of human perception. Given that organizations are using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Abstract 2: Do people feel better when on vacation? We study the social media sentiment of families when on vacation and during normal periods. Using a snowball sampling method, we solicited participants active on various social media platforms (N = 1,337) to voluntarily disclose when and where they went on vacation. We then crawled our participants' social media posts to determine whether specific time periods and destinations for vacations were correlated with higher levels of enjoyment. Initial analysis showed that such features do yield significant prediction performance improvement: for example, vacations taken close-by tended to have markedly higher satisfaction. Our model, Vacay-ok, accounts for such factors in novel gating network architecture, improving joint vacation period and sentiment detection by over 10% F1.

Our analysis validates patterns of interest: vacations with immediate family exhibit a bimodal distribution, and that ones without in-laws have a significantly higher satisfaction and detection rate.

For robustness in identification, we collect and study the pattern and text of social media posts by parents, children, extended family members and friends of vacation-goers.
To ensure both the accessibility and the reproducibility of our experiments, we have gathered the raw social media data, culled with permission from the original posters, and released it as a zip archive, available without moderation, in the footnote.

Abstract 3: In this paper, we present the largest existing language model as of today (989 trillion parameters, dataset of 968 Pb), BigBlue. We trained it using not only freely available Web content, but also Web archives and publicly-accessible EPub versions of all the books of the commercial Amazonia bookshop web storefront. The best results were obtained using reinforcement learning with human feedback (RLHF) coming from crowdworkers. The model improves the state-of-the-art in Natural Language Processing (NLP) in most tasks, including sentiment analysis (+0.02 F-measure), dependency syntax (+0.015 UAS), named-entity recognition (+0.05 F-measure), etc. BigBlue is available through an API on our company's website.

Abstract 4: In this paper, we present the first language model developed for Fridonian, a regional language spoken in the South-East of Austrafrancia by more than 500,000 people. We detail the collection of the corpus, the creation of the model and its evaluation on basic NLP tasks. The corpus creation itself is an achievement, as it is well-known and documented (Birdoff, 2020) that Fridonian speakers are reluctant to allow the recording of their language, for fear it will be used against them by the country's Sidonian majority. However, we managed to record 100 hours of speech in different settings: family discussions, local authority meetings, ceremonies and traditional storytelling contests. We used this dataset to train a language model which we evaluated on sentiment analysis and named-entity recognition tasks, with impressive results (resp. 0.56 and 0.75 in F-measure). The created language model FridoBERT and the dataset FridoSPEECH are freely available on GitHub.

Abstract 5: Is sentiment analysis dependent on culture or only on language? In this paper we describe the collection and annotation of a billion-word dataset over 7 different Spanish dialects spoken in Latin America, annotated with fine-grained sentiment analysis cues. To the best of our knowledge, this is the largest Latin American dataset currently available. We used data from public groups in Telegram where locals talk about products and services. We evaluate the state-of-the-art sentiment analysis models on it, showing that their performance is significantly lower than the state-of-the-art results on well-known benchmark for Spanish. The dataset annotation took three months. We selected crowdworkers after a careful exam to ensure they were fluent in the target dialect. We paid them 4 US dollars per hour, a fair wage considering it was two times the minimum wage in these countries. Our sentiment classifier fine-tuned per dialect outperforms previous models on Spanish by large margins. Our model has been used for 12 months by hundreds of companies in Latin America. To serve the model, we implemented a costing model to provide this service at a uniform price to all regions of the world. We conducted three usability studies in collaboration with three companies, which demonstrated that the integration of our sentiment analysis detector markedly increased users' engagement with their marketing advertisements.

Abstract 6: With the rise of large language models (LLMs), they have become useful in critical settings such as healthcare support, helping reduce administrative burden and improve predictive analytics. However, there is an emerging concern that LLMs encode gender and racial biases, which propagate to healthcare decisions. In this work, we propose a simple and effective approach to controlling for biases in LLMs through data anonymization, specifically focusing on patient names that often reveal gender and race. We introduce a novel and highly granular dataset of over 9.9 million patients' electronic health records annotated with patient demographics, including gender and race. We first fine-tune state-of-the-art LLMs on the raw healthcare data, and establish the presence of harmful biases in standard NLP benchmarks such as coreference resolution (10% lower F1 score for women compared to men) and named entity recognition (7% lower for White compared to Black patients). We also show that naively fine-tuned language models can be used to predict gender and race of the patients in held-out health records with high accuracy. Then, we show that a simple anonymization of the health data, by replacing patient names with generic placeholders, reduces the prediction gap in fine-tuned LLMs by up to 4% for gender, and up to 2.5% absolute for race. Our findings address important questions for fairness in NLP and algorithmic decision-making. Our code and data are publicly available to facilitate reproducibility.

1b) Ethical Consideration Sections

Ethical consideration section 1:

This research was approved by our IRB.

[This paper is anonymized and submitted for blind review]

Ethical consideration section 2:

The work in this paper raises no ethical issues not already raised by established NLP applications, and it mitigates some existing risks of LLMs.

Ethical consideration section 3 (generated by ChatGPT):

The use of large language models (LLMs) in critical settings such as healthcare is an area of significant interest. However, there is an emerging concern that these models encode gender and racial biases, which propagate to healthcare decisions. Such biases can have significant ethical implications, including the exacerbation of existing health disparities and discrimination against marginalized groups.

In this work, the authors propose a simple and effective approach to controlling for biases in LLMs through data anonymization. Specifically, the authors focus on patient names that often reveal gender and race. While data anonymization is a promising approach, it is important to consider the ethical implications of anonymizing healthcare data. For instance, anonymization may not be sufficient to prevent re-identification of individuals in certain cases, which could pose risks to their privacy and security. It is therefore important for the authors to carefully consider the potential risks and benefits of data anonymization in the context of their study.

Additionally, the authors note the presence of harmful biases in standard NLP benchmarks, such as coreference resolution and named entity recognition. Such biases can lead to inaccurate predictions and decisions that could have significant consequences for patient care. As such, the authors should consider the ethical implications of using LLMs that have not been adequately controlled for biases, particularly in critical settings such as healthcare.

Finally, the authors have made their code and data publicly available to facilitate reproducibility. While data sharing is important for scientific progress, it is crucial to consider the potential ethical implications of sharing sensitive healthcare data, including risks to patient privacy and confidentiality. The authors should therefore ensure that they have obtained appropriate consent and taken necessary measures to protect patient privacy before making their data publicly available.

In summary, while the proposed approach is promising for controlling biases in LLMs, it is important to carefully consider the potential risks and benefits of data anonymization and data sharing in the context of this study. Moreover, the authors should ensure that their work adheres to ethical principles such as respect for patient autonomy, beneficence, non-maleficence, and justice.

This document is publicly editable, as of this time. By contributing to this document, you agree that your contributions are given to the public domain for posterity. The tutorial presenters may further organize or edit the document as we see fit before finalizing the document for general public dissemination. Please ensure any materials that you place in this document are suitable for public use: images should be credited for fair use, and be explicitly ok for use in public materials, documents quoted should retain a note to their source.

2) What else is missing?

Duration: 5 minutes

Write your name and two or more sentences to write briefly about a potential harm that we did not discuss today, or link to a news article you want to share with the rest.

- e.g., Luciana: https://www.reuters.com/world/americas/brazil-lawmakers-vote-controversial-bill-clean-up-social-media-2023-05-02/
- Karën:
  - Book: Automating Inequalities – Virginia Eubanks- St. Martin’s Press
  - Book: Ruined by design - Mike Monteiro
- *Add your link or harm here!*

This document is publicly editable, as of this time. By contributing to this document, you agree that your contributions are given to the public domain for posterity. The tutorial presenters may further organize or edit the document as we see fit before finalizing the document for general public dissemination. Please ensure any materials that you place in this document are suitable for public use: images should be credited for fair use, and be explicitly ok for use in public materials, documents quoted should retain a note to their source.

3) Final Call to Action

Would you like us to help add your name for consideration to ethics reviewing committees in upcoming conferences in the upcoming cycles? Please let us know here.

Add your name, affiliation, homepage and email and let us know your availability for serving (if you have restricted times)

- e.g., Min-Yen Kan, National University of Singapore, http://www.comp.nus.edu.sg/~kanmy, kanmy@comp.nus.edu.sg , for the next year.
- *Add your row here!*

4) Your Feedback on this Tutorial

We want your help to further improve this tutorial! Can you please help give your critical and constructive feedback? Feel free to keep your comments anonymous or attribute them. Thank you!

How to give good feedback:

- Be positive
- Be specific
- Suggest the next step!

[Please copy this block and customize it with your feedback]

- A new rater
  - Please rate your overall experience [dreadful 1 – 10 exceeding expectations]
  - Please describe what you liked about the practical.
  - Please describe any challenges you might have had / what you didn't like.
  - Please provide any suggestions that might help us improve the experience.
- *Add your feedback items here!*