CHALLENGES OF AUTOMATED METHODS FOR ANALYSIS AND GENERATION OF SOCIO-EMOTIONAL BEHAVIORS

TOWARDS MORE TRANSPARENT SYSTEMS

Transparency here: the ability to show the user of NLP systems the process used to analyse or generate socio-emotional behaviors.

Example of polemics of NLP systems due to their lack of transparency:

Microsoft - Tay [2016]

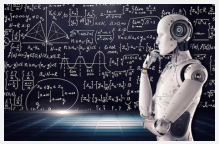
Open AI - Chat GPT [2022]

Urgent need for approaches to **better understand** what is behind these models if we want to develop **trustworthy** technologies.

TOWARDS MORE TRANSPARENT SYSTEMS

Linked concept : **Explainable AI**
"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".

D. Gunning, "Explainable Artificial Intelligence (XAI)," 2017.



TOWARDS MORE TRANSPARENT SYSTEMS

Computational models:

- rule-based,
- learned from scratch,
- **LMM fine-tuned,**
- **LLM with prompting, can be improved by reinforcement learning by human feedback**

Issue: **SOTA approaches** are very performant but also very opaque:

- It is very difficult, if not currently impossible, to explain the process by which a text is generated or analyzed.
- It is very difficult to check the generated responses

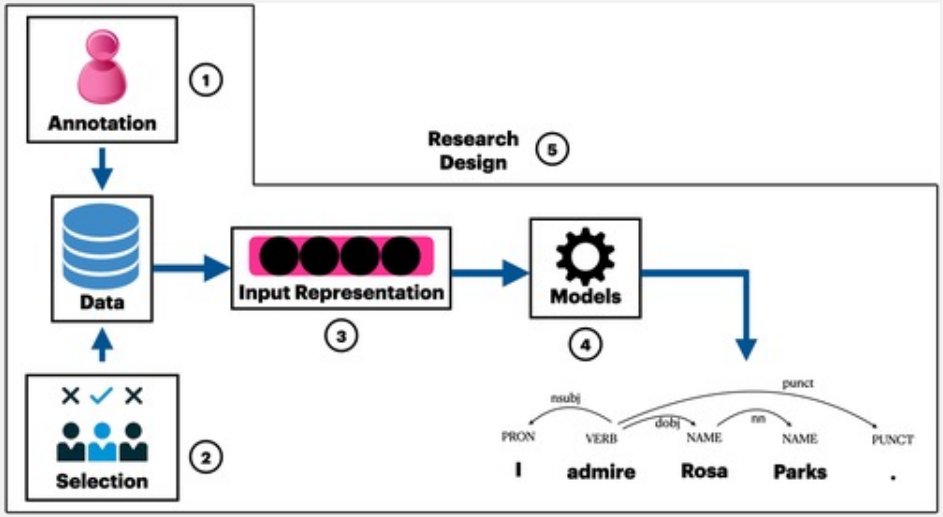
TOWARDS MORE TRANSPARENT SYSTEMS

Try to make the system used to analyse or generate socio-emotional behaviors :

- transparent with the subjectivity/the bias that they encode

Because some of them may lead to the proliferation of harmful stereotypes

Kiritchenko, S., & Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 43–53.

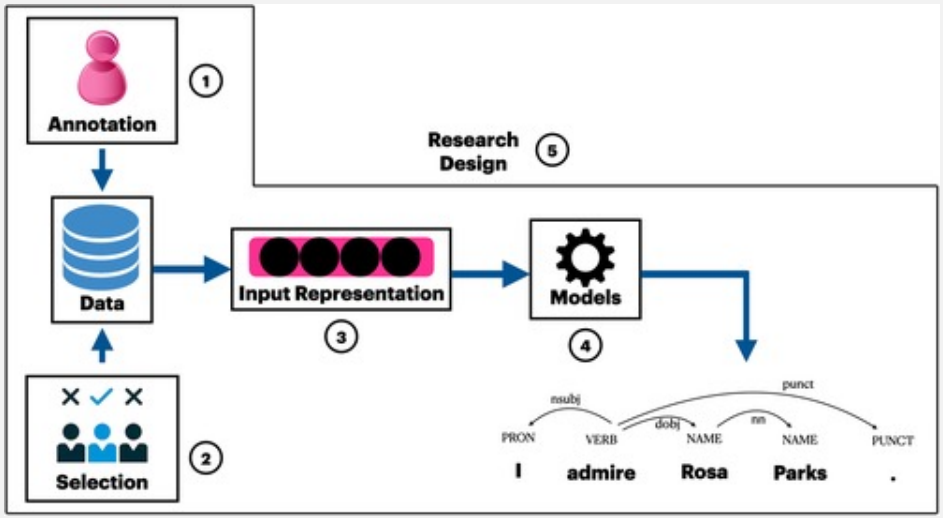


five sources where bias can occur in NLP systems:

(1) the data,
(2) the annotation process,
(3) the input representations,
(4) the models
(5) the research design (or how we conceptualize our research)

From [Hovy and Prabhumoye, 2021]

LEVERS OF TRANSPARENCY



- Lever 1: Transparent annotated data ①②
- Lever 2: Dissecting and disentangling input representations and neural models ③④
- Lever 3: Transparency by design ④⑤

LEVER 1: TRANSPARENT ANNOTATED DATA

What is generated by the models and the decision-making process behind automated analyses is strongly influenced by :

- The data used to learn them (*machine learning*)
  - Ex : GPT3 has been learned from 300 billion tokens (~words) extracted from Wikipedia, books, and crawled pages on the web
- Human annotations used to supervise (guide) the learning process, ex:
  - Annotation of opinions expressed in texts
  - Annotation of the "ethical" or "safe" character of the answer used in reinforcement learning
  - For chatGPT: redaction of standard responses by humans

LEVER 1: TRANSPARENT ANNOTATED DATA

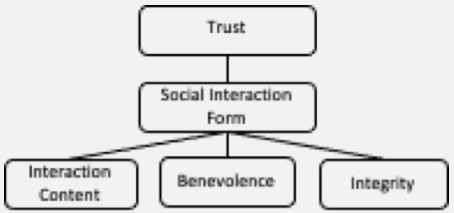
Be transparent with the data content:

- Which language?
- Which type of data? Interactions, written, spoken, etc.?
- Retrieve social variables (e.g., culture, personality) corresponding to
  - speaker/writer biases
  - labeler biases
  - Ex: POM dataset self-assessed personality of the workers (labellers) (Park et al., 2014) using (Big Five Model): Openness Conscientiousness Extraversion Agreeableness Neuroticism

LEVER 1: TRANSPARENT ANNOTATED DATA

Be transparent with the annotation process

- Rely on social science literature to define the guideline for annotating the socio-emotional phenomena
- Give details about the used annotation guide



Hulcelle et al., TURIN : A coding system for **Trust** in **hUmanRobot INteraction** ACII 2021

Rollet & Clavel. "Talk to you later" Doing social **robotics** with conversation analysis. Towards the development of an automatic system for the prediction of **disengagement**, Interaction Studies 2020

394

Theoretical models from psychology, linguistics, conversation analysis (ex. Psychological models for emotion and engagement, socio-linguistic definition of trust)

LEVER 1: TRANSPARENT ANNOTATED DATA

Points to watch:

- The reproduction by these generative AIs of:
  - stereotypes and cognitive biases of both authors and annotators of texts
    - ex: Tay, Test mid-journey with “imagine a CEO” vs. “imagine a secretary”
  - ethical principles defined in the recommendations given to annotators (raises the question of who defines these ethical principles)
- The question of copyright: generative AI can generate texts that are very close to the data present in the training data (e.g. code generation).
- Some languages and cultures are poorly represented in learning data

Microsoft - Tay [2016]

LEVER 2 : DISSECTING AND DISENTANGLING NEURAL MODELS

- DISSECTING : understand cognitive biases that are encoded in the models
  - ➡️ in pre-trained representations (cognitive biases of the authors of the texts used for the training of the embeddings)
  - ➡️ in supervised machine learning models (cognitive biases of the annotators providing labels for the supervision of the models)

- DISENTANGLING : remove sensitive information and cognitive biases from the models

DISSECTING

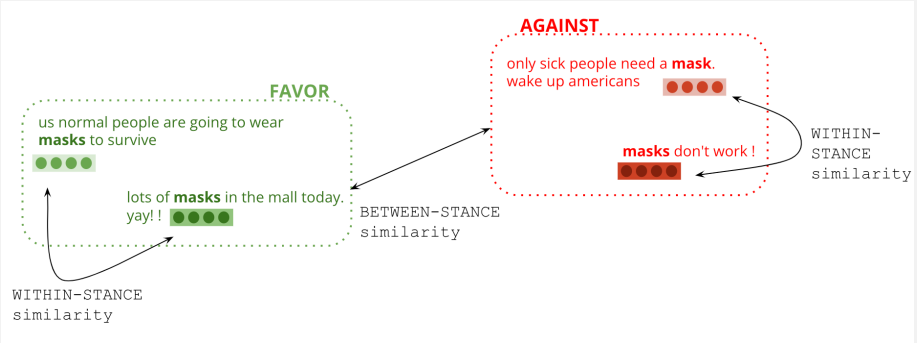
Two examples of studies:

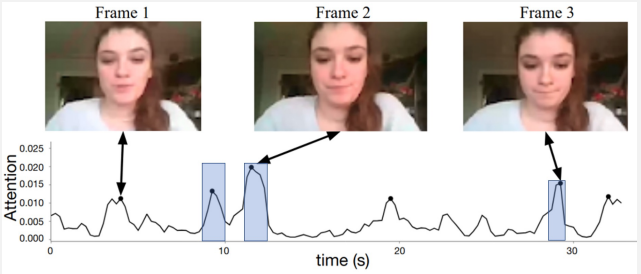
1. understand how **stances** are encoded in contextualized word representations (embedding vectors **w**)

One Word, Two Sides: Traces of Stance in Contextualized Word Representations, [Gari Soler, Labeau & Clavel COLING 2022]

2. understand **human subjectivity** dissecting supervised neural models

Multimodal Hierarchical Attention Neural Network: Looking for Candidates Behaviour which Impact Recruiter's Decision, [Hemamou, Clavel et al. IEEE Trans. of Affective Computing]





2/ DISSECTING SUPERVISED NEURAL MODELS

Take home message: just a step towards explainability

We can try to dissect a model. It gives some interesting information to try to understand the decision

BUT this is very local and we can not completely retrace the decision process such as it could be done when using reasoning models

Multimodal Hierarchical Attention Neural Network: Looking for Candidates Behaviour which Impact Recruiter's Decision, [Hemamou, Clavel et al. IEEE Trans. of Affective Computing]

LEVER 2: DISENTANGLING

- Disentangle unwelcome information
  - Subjectivity in textual representations using mutual information

    Pierre Colombo, Pablo Piantanida, Chloe Clavel. « A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations » ACL (2021)

  - Sensitive information (gender and origin) in supervised neural models using adversarial approaches

    Léo Hemamou, Arthur Guillon, Jean-Claude Martin and Chloé Clavel. Don't Judge Me by My Face: An Indirect Adversarial Approach to Remove Sensitive Information From Multimodal Neural Representation in Asynchronous Job Video Interviews, ACII 2021

DISSECTING SUPERVISED MODELS : EXISTING TOOLS

- Two popular explanation methods to explain black-box model predictions (and also used on explainable sentiment analysis specifically).
  - Local Interpretable Model-Agnostic Explanations (LIME) : local understanding of the selected black-box model
    - M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
  - SHapley Additive exPlanations(SHAP): can provide the importance of features for each prediction
    - S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. of the 31st International Conference on Neural Information Processing Sy

LEVER 3 : TRANSPARENCY BY DESIGN

Be transparent with your research design choices (task, data):

- why did you choose to address this NLP task ?
- How did you select the data?

« Because there is data available » is not a satisfactory answer -> all researchers tend to address the same tasks with the same data (ex: sentiment analysis in movie reviews in English)

- Ask yourself the following questions: 'Would I research this if the data wasn't as easily available? Would my finding still hold on another language?'

« Researchers have to be mindful of the entire research design: data sets they choose, the annotation schemes or labelling procedures they follow, how they decide to represent the data, the algorithms they choose for the task and how they evaluate the automated systems. Researchers need to be aware of the real-world applications of their work and consciously decide to choose to help marginalized communities via technology » [Hovy and Prabhumoye, 2021]

Do not forget to illustrate under-represented phenomena or languages

LEVER 3 : TRANSPARENCY BY DESIGN

Design models that are transparent by design

- We need to develop methods able to encode explicitly the socio-emotional information in the models

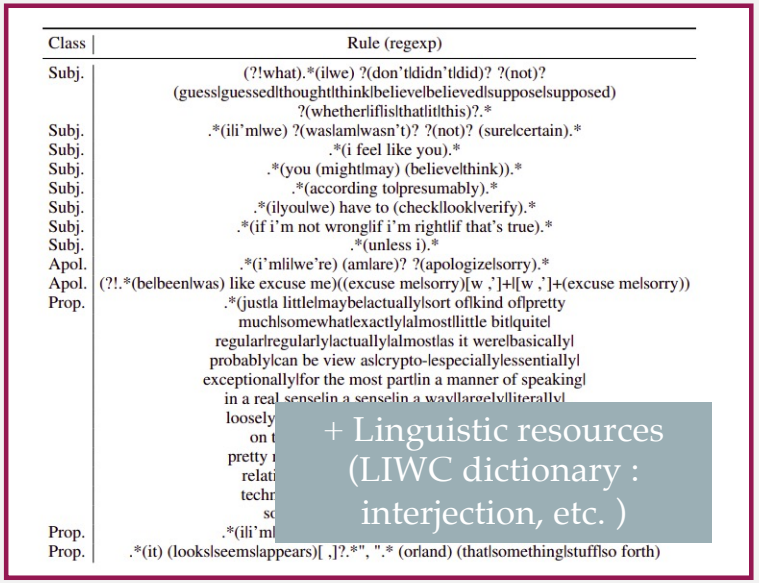
Overview of models (from the most transparent to the least transparent)

- Rule-based models
- Interpretable features + machine learning models or neural classifiers
- Neural embeddings + machine learning models or neural classifiers
- Neural generative large language models used with prompts

LEVER 3 : TRANSPARENCY BY DESIGN

- Example of more transparent models using interpretable features and machine learning models

Descriptions of different **hedges and their expressions** from linguistic theories: Rowland (2007), Fraser (2010) and Brown and Levinson (1987),



Class	Rule (regexp)
Subj.	(?!what).*(i|we) ?(don't|didn't|did)? ?(not)? (guess|guessed|thought|think|believe|believed|suppose|supposed) ?(whether|if|is|that|it|this)?.*
Subj.	.*(ili'm|we) ?(was|am|wasn't)? ?(not)? (sure|certain).*
Subj.	.*(i feel like you).*
Subj.	.*(you (might|may) (believe|think)).*
Subj.	.*(according to|presumably).*
Subj.	.*(i|you|we) have to (check|look|verify).*
Subj.	.*(if i'm not wrong|if i'm right|if that's true).*
Subj.	.*(unless i).*
Apol.	.*(i'm|i|we're) (am|are)? ?(apologize|sorry).*
Apol.	(?!.*(be|been|was) like excuse me)((excuse me|sorry)[\w ,']+|[\w ,']+(excuse me|sorry))
Prop.	.*(just|a little|maybe|actually|sort of|kind of|pretty much|somewhat|exactly|almost|little bit|quite| regular|regularly|actually|almost|as it were|basically| probably|can be view as|crypto-le|especially|essentially| exceptionally|for the most part|in a manner of speaking| in a real sense|in a sense|in a way|largely|literally| loosely on t pretty relati techn s
Prop.	.*(ili'm|
Prop.	.*(it) (looks|seems|appears)[,]?.*", ".* (or|and) (that|something|stuff|so forth)

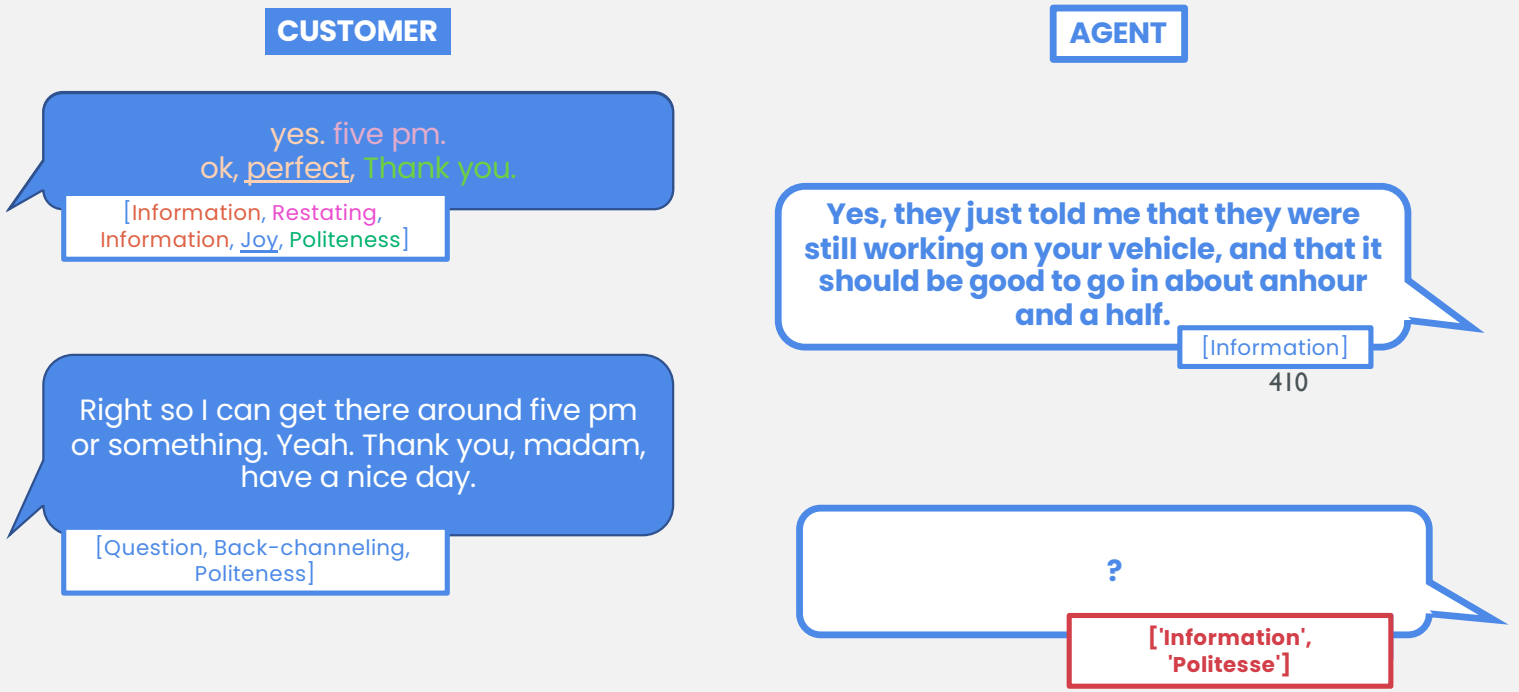
Knowledge-Driven Features (KDF)

+ Linguistic resources (LIWC dictionary : interjection, etc.)

[« You might think about slightly revising the title »: identifying hedges in peer-tutoring interactions.] (Raphalen, Clavel and Cassell, ACL 2022)

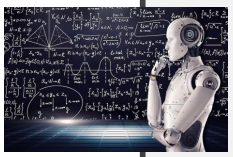
LEVER 3 : TRANSPARENCY BY DESIGN

- A step towards more transparent conversational system : generative models using explicit socio-emotional strategies



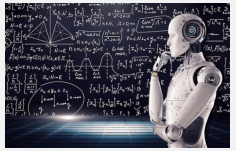
L. Vanel., A. Yacoubi., and C. Clavel. (2023), "A New Task for Predicting Emotions and Dialogue Strategies in TaskOriented Dialogue", ACII 2023

RESEARCH STATUS



- Research still at an early stage:
  - High-performance models are still very opaque and costly in terms of resources
  - Transparent and frugal models are less efficient

CONCLUSIONS - TRANSPARENCY

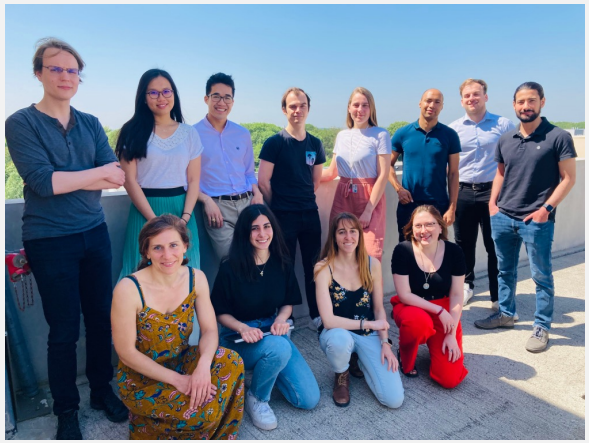


First steps presented/mentioned here:

- **Annotation guidelines** grounded on social sciences theories
- **Methods for dissecting**
- pre-trained representations (ex: traces of stance in neutral words)
- and neural models using **attention mechanisms** for local interpretations
- Approaches for **disentangling** subjective and undesirable variables
- **Explicit modelisation** through **feature design** and **planning** dialog strategies

Further work needs to be done so that tools could be trustworthy.

- a higher lever of explainability (understandable by users)
- Methods for isolating stereotypes and cognitive biases from models: we need to be able to explain what is the underlying ethical and cultural background of the models



WORK PRESENTED HERE CARRIED OUT WITH MY PHD STUDENTS, POST-DOCS AND COLLABORATORS !

HIRING NEW PHD STUDENTS OR POST-DOC!
CONTACT : CHLOE.CLAVEL@INRIA.FR

MATERIALS TO GO FURTHER

MATERIALS TO GO FURTHER

- NLP in general
  - https://nlp.stanford.edu/IR-book
  - From Miha Grcar “Text mining and Text stream mining tutorial
  - Foundations of Statistical Natural Language Processing Christopher D. Manning and Hinrich Schütze
  - Lecture from Stanford http://cs224d.stanford.edu/lecture_notes/notes1.pdf

MATERIALS TO GO FURTHER

- NLP and deep learning
  - Deep Natural Language Processing course offered in Hilary Term 2017 at the University of Oxford.
  - Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
  - Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

MATERIALS TO GO FURTHER

- Tools :
  - word2vec from Google https://code.google.com/p/word2vec/ tutorial from
  - tensorflow https://www.tensorflow.org/tutorials/word2vec
  - Other representation : Glove http://nlp.stanford.edu/projects/glove/
- Sentiment analysis
  - https://web.stanford.edu/class/cs224u/slides/cs224u-2021-sentiment-part1-handout.pdf
  - Munezero M. D., Suero Montero C., Sutinen E., Pajunen J., "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text", IEEE Transactions on Affective Computing, 2014.