LECTURE 5: NATURAL LANGUAGE PROCESSING APPLICATIONS OF LLMS

Creating Business Value with Generative AI
Fall 2025

AARHUS BSS
DEPARTMENT OF MANAGEMENT
AARHUS UNIVERSITY

01. October 2025

Magnus Bender
Assistant Professor

WHY THIS LECTURE?

- This lecture focuses on practical applications of classification, a core task in Natural Language Processing (NLP).
- Almost anything you do with an LLM uses – to some extent – classification.
  - Asking an LLM to evaluate whether something is a good idea; quality of idea is a „class"
  - Asking an LLM to generate text in a particular style; style is a "class"
- Today's class will introduce some core concepts of classification, and we'll take a deeper dive into the specifics of how the two empirical papers do classification.

THIS LECTURE RELATIVE TO YOUR PROJECT

This lecture is intended to give you:

1. A better understanding of breaking down a process from the technical, NLP oriented, side
2. A better sense of how you can classify (or more general: analyze) things:
   - customer emails,
   - product reviews,
   - …
3. An overview of NLP tasks across disciplines that you can try to match to your use-case.

AGENDA FOR TODAY

- Natural Language Processing (NLP) tasks
  - What do we often want to do with large collections of documents?
  - What are the different kinds of tasks?
  - How do they work?
  - Which kind of task requires which type of model and/ or available data
  - Which kind of task may not be suitable for an LLM
- NLP tasks with LLMs (readings)
  - How well do they perform – at least according to the readings
  - How do we know, and what can we do with
- A very brief overview of NLP tasks across different disciplines, for your inspiration

CLASSIFICATION

... in a machine leaning context

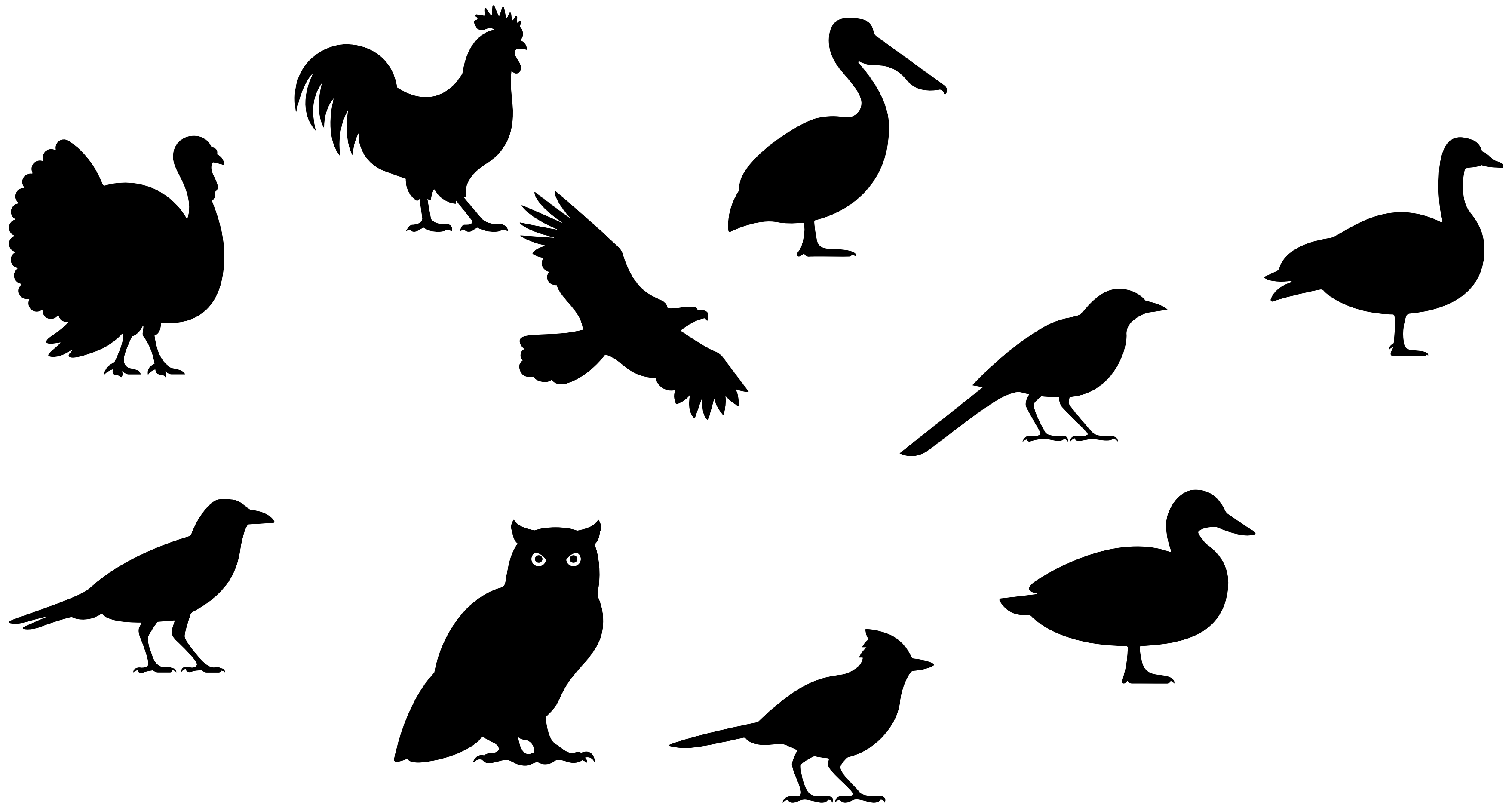
01. October 2025

Magnus Bender
Assistant Professor

CORE CONCEPTS: FEATURES AND CLASSES

What are these things?!?

- **Features**: Specific (processed data) that defines a thing both in its own right, and in contrast to other things
- **Classes**: Collections of things with shared features that we want to think of as a "something"



Bird **features**:

- Length of beak
- Shape of beak
- Color of beak
- Body posture
- Color of feathers
- Size
- …

Bird **classes**:

- Hen
- Duck
- Goose
- …

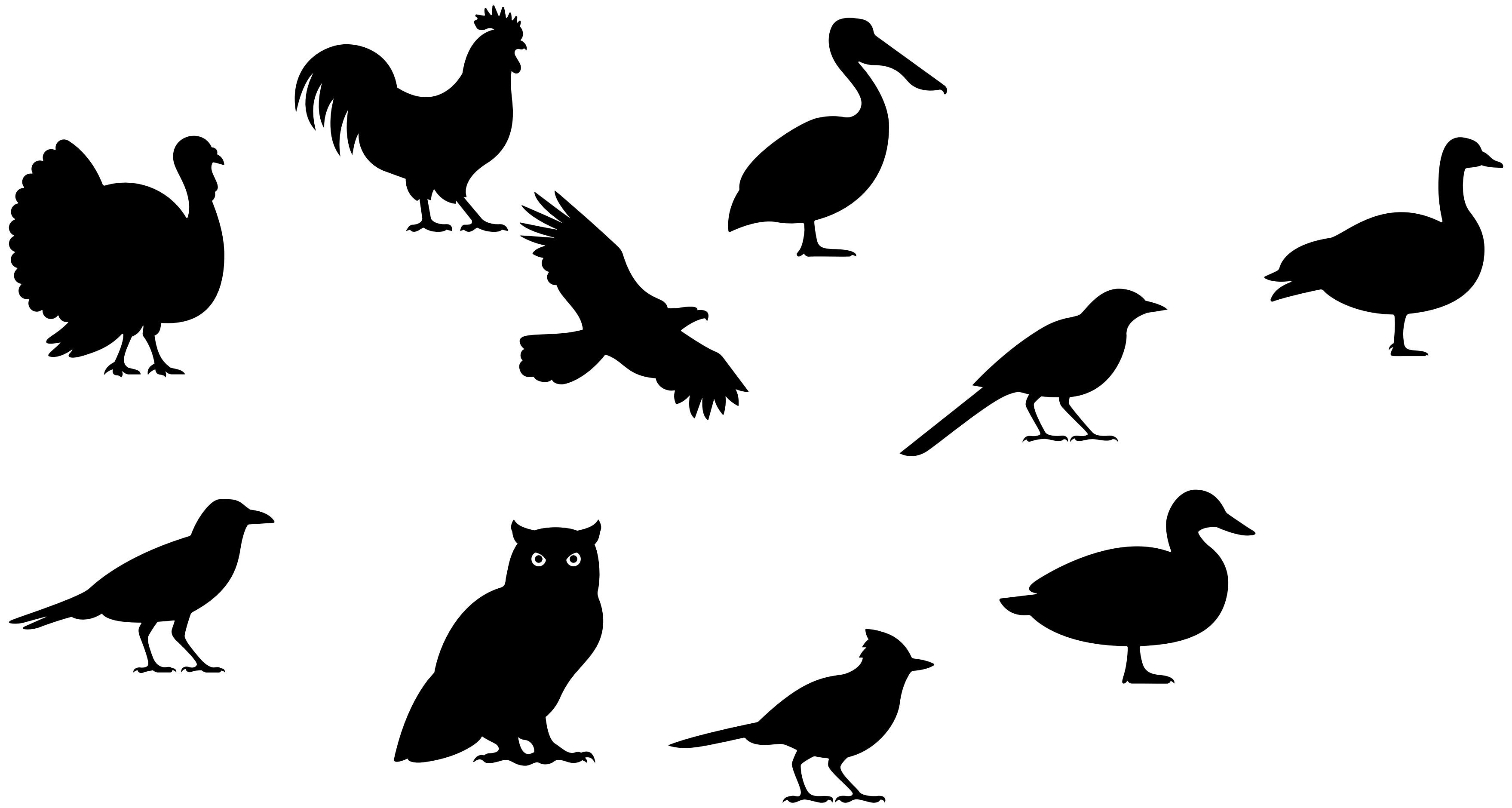
WHAT'S HARD ABOUT THAT?

Features and classes seem easy enough, right?

For birds, biologists have already done most of the work for us and defined features, like birds. They have structured the data for us.

When we work with unstructured text data, <u>we often need to identify and define these ourselves.</u>

<u>And tell, e.g., ChatGPT about them!</u>



Bird **features**:
- Length of beak
- Shape of beak
- Color of beak
- Body posture
- Color of feathers
- Size
- …

Bird **classes**:
- Hen
- Duck
- Goose
- …

HOW?

- The process of
  1. Defining
  2. Identifying, and
  3. Extracting

  features and classes, and articulating them in ways that an LLM will understand, relative to your data.
- In today's readings, these were all "pre-defined". But you might let your classes and features emerge from your data, rather than pre-defining them.
- Keep in mind:
  <u>**Classification features should not just define a class in its own right. They should define what makes the class different from other classes.**</u>

NLP TASK: SENTIMENT ANALYSIS

AARHUS BSS

DEPARTMENT OF MANAGEMENT
AARHUS UNIVERSITY

01. October 2025

Magnus Bender
Assistant Professor

WHAT CAN WE DO WITH LLMS THAT ARE NOT JUST CHATBOTS?

- **Customer Feedback Analysis:** Automatically process and categorize large volumes of customer reviews, surveys, and support tickets to uncover trends, sentiments, and key concerns.
- **Market and Competitive Intelligence:** Analyze vast amounts of news articles, social media posts, and industry reports to identify emerging market trends, competitor strategies, and shifts in consumer preferences.
- **Content Classification and Organization:** Efficiently classify, tag, and organize large collections of business documents, emails, or contracts for easier retrieval and compliance.
- **Sentiment and Opinion Mining:** Detect customer sentiment and opinions across social media, forums, and online platforms to inform product development, marketing strategies, and brand reputation management.
- **Risk and Compliance Monitoring:** Leverage NLP to scan and analyze regulatory documents, legal contracts, and financial reports to ensure compliance and identify potential risks or breaches.

WHAT CAN WE DO WITH LLMS THAT ARE NOT JUST CHATBOTS?

- Natural Language Processing has been around for a long time (since 1950s)

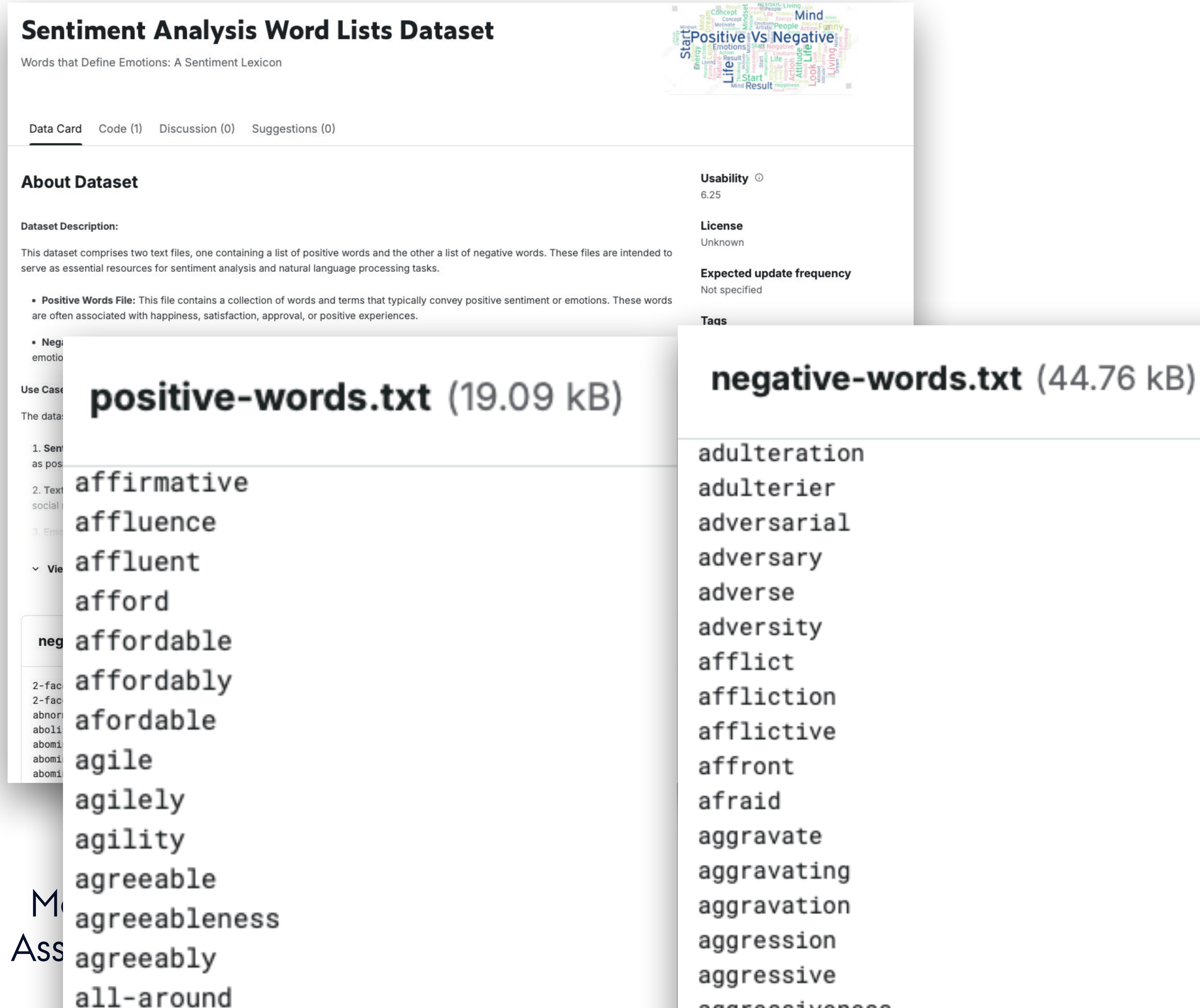
  - Automatic translation
  - **Sentiment analysis**
  - Text generation
  - Text classification

- The techniques used for these varied, but are all, by now, completely outdated
- However: **Outdated does not imply useless!**
  - *We still have horses after we invented cars or trains after we invented planes.*

SENTIMENT ANALYSIS, PRE-2017

To measure the sentiment of a text, we used to:

1. Load a dictionary of words:
   - *positive*
   - *negative*
2. Computationally read through a document
3. Count the number of positive and negative words:
   - Add them all up (+1 for positive words, -1 for negative words)
4. Obtain a result:
   - if the sum is > 0 → positive sentiment
   - if the sum is < 0 → negative sentiment



Sentiment Analysis Word Lists Dataset

Words that Define Emotions: A Sentiment Lexicon

About Dataset

Dataset Description:

This dataset comprises two text files, one containing a list of positive words and the other a list of negative words. These files are intended to serve as essential resources for sentiment analysis and natural language processing tasks.

- **Positive Words File:** This file contains a collection of words and terms that typically convey positive sentiment or emotions. These words are often associated with happiness, satisfaction, approval, or positive experiences.
- **Nega**
  emotio

Use Case

The data

1. **Sen**
as pos

2. **Text**
social

3. Emo

Vie

neg

```
2-fac
2-fac
abnor
aboli
abomi
abomi
abomi
```

Usability 6.25

License Unknown

Expected update frequency Not specified

Tags

positive-words.txt (19.09 kB)

```
affirmative
affluence
affluent
afford
affordable
affordably
afordable
agile
agilely
agility
agreeable
agreeableness
agreeably
all-around
```

negative-words.txt (44.76 kB)

```
adulteration
adulterier
adversarial
adversary
adverse
adversity
afflict
affliction
afflictive
affront
afraid
aggravate
aggravating
aggravation
aggression
aggressive
```

PROBLEMS WITH THIS OLD-SCHOOL METHOD

- Context matters, word count does not:
  - „In spite the horrifyingly bad weather, things worked out well.“ → negative
- Synonym phrases don't count:
  - „He cheated the system“ → negative
  - „He dodged the rules“ → neutral
- Large Language Models can avoid many of the traps that these older *dictionary-approaches* fell in.
  - No dependency on language and use-case specific list
  - ...

LLMS TO ASSIST WITH PROBLEMS & PROCESSES

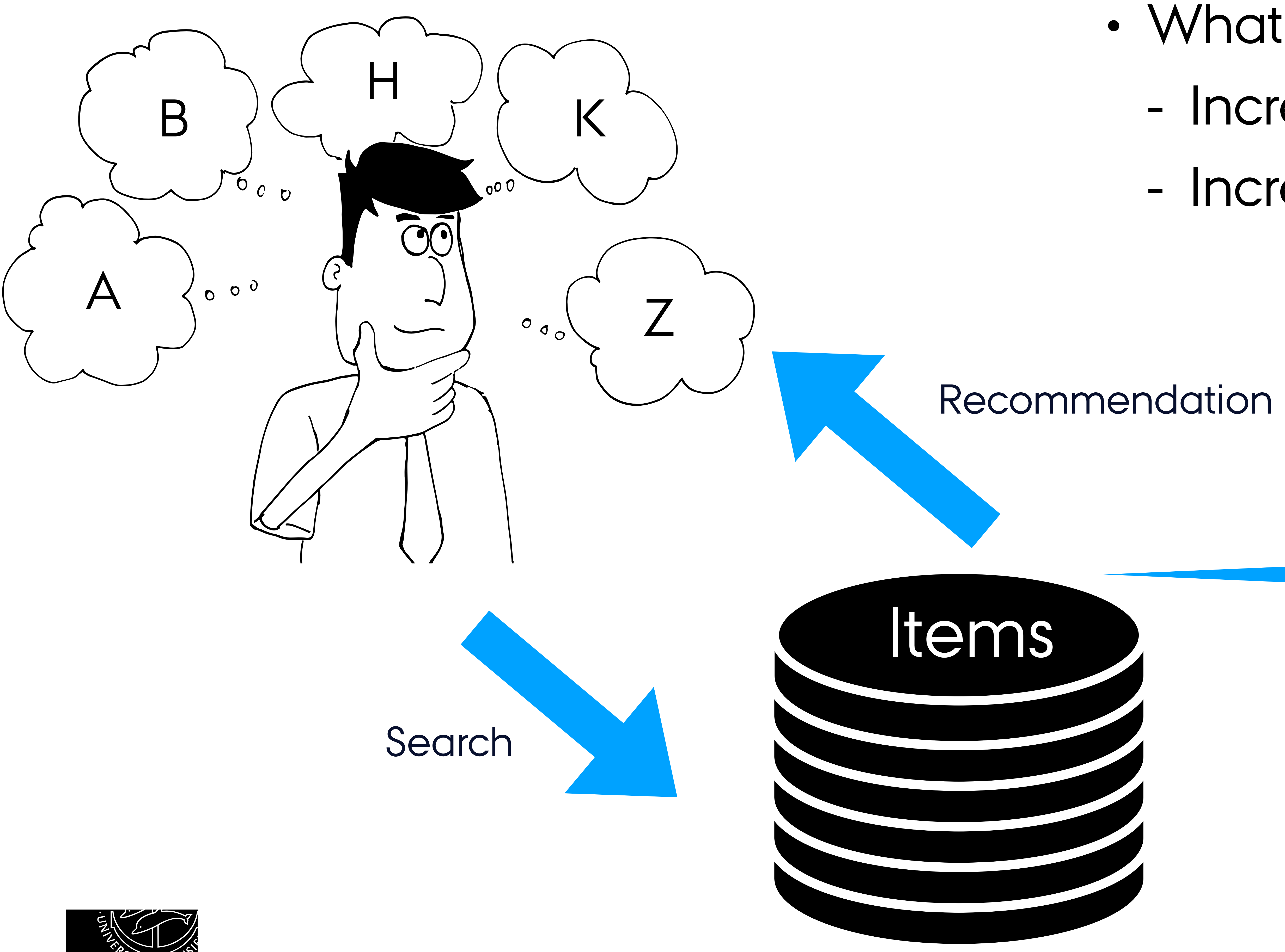
- Today, focus on the *technical* site
- Or: „What NLP tasks do LLMs offer and how can I use them to solve my problem?“

EXAMPLES FROM LECTURE 2

1. Personalized recommendations of items in an online shop ➡ Some relevant item out of all available items
2. Automatic forwarding of customer’s e-mails to correct department ➡ The best matching department to handle the case
3. Inspection of CVs of applicants for a position ➡ „Complete“ or „Incomplete“ for each CV
4. Selection of the applicant to hire for a position ➡ The best applicant out of all applicants

1. RECOMMENDATIONS

- What are good recommendations?
  - Increased customer satisfaction
  - Increased sales for the provider



- Predict how strong a “customer's” interest in an object is
- Recommend to “customers” precisely those objects from the set of all existing objects that are of most interest.

amazon
Deliver to Denmark
All
☰ All
Today's Deals
Prime Video
Registry
Gift Cards
Customer Service
Sell
books
Categories
New & Trending
Deals & Rewards
Best Sellers
AI Agents in Action
amazon.com

artificial intelligence a modern approach

Books › Computers & Technology › Computer Science › AI & Machine Learning › Intelligence & Semantics

Artificial Intelligence: A Moc

Edition 4th Edition

by Peter Norvig (Author), Stuart Russell (Autho

4.7 ★★★★☆ 516 ratings

Thelong-anticipated revision of *ArtificialIntelli* breadth and depth of the field of artificialintelli to date on the latest technologies,presents con new or expanded coverageof machine learning, systems,robotics, natural language processing, fairness, and safe AI.

Report an issue with this product or seller

ISBN-10	ISBN-13
1292401133	978-1292401133

See all details

Read sample

Follow the author

Peter Norvig Follow

The AI Engineering Bible: T Build, Develop and Scale Pr

by Thomas R. Caldwell

★★★★★ 620 | $48.19

Shop now ›

Sponsored

Frequently bought together

Sponsored

Frequently bought together

This item: Artificial Intelligence: A Modern Approach, Global Edition

$60⁰⁰

Deep Learning (Adaptive Computation and Machine Learning series)

$61⁶⁰

More items to explore

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts,...
› Aurélien Géron
★★★★☆ 748
Paperback
$48.25
$19.65 shipping

Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter
› Wes McKinney
★★★★☆ 444
Paperback
$43.99
$17.86 shipping

Videos

Help others learn more about this product by uploading a video!

Upload your video

About the author

Follow authors to get new release updates, plus improved recommendations.

Peter Norvig

I live in Palo Alto, CA with my wife and two children. I am currently the Director of Research for Google, and I am teaching an In the world. You can buy some of my books here at Amazon.

Follow

Related products with free delivery on eligible orders

Sponsored ⓘ Try Prime for unlimited fast, free shipping

The AI Job Shift: How to Transition Into AI-Supported Careers
Bridgeway36 Inc
★★★☆☆ 2
Paperback
$9⁹⁹
✓prime

San Francisco: The AI Capital of the World (Intelligence Amplifier)
Arvin Lioanag
Hardcover
$38³⁸
✓prime

AI Agents Millionaire: [3 in 1] 1000+ Automation Secrets to Skyrocket Profits, Rein...
Darren Pascall
★★★★★ 35
Paperback
$24⁹⁷
✓prime

Co-Intelligence: Living and Working with AI
Ethan Mollick
★★★★☆ 3,321
#1 Best Seller
Hardcover
-41% $17⁷⁹
List: ~~$30.00~~
✓prime

Customer Success In Th World Of Artificial Intelligence
Vivek S Nair
Paperback
$34⁹⁵
✓prime

https://www.amazon.com/Artificial-Intelligence-Modern-Approach-Global/dp/1292401133/

SHOPPING CART TRANSACTIONS

- Given transactions in the form of shopping carts of different people
- Looking for recommendations of additional items for each person or shopping cart

Transaction ID	Person ID	Items
1	1	Bread, Oatmeal, Potatoes
2	2	Apples, Bread, Sugar
3	3	Apples, Bread, Potatoes, Sugar
4	2	Bread, Potatoes, Sugar
5	4	Bread, Oatmeal, Potatoes, Sugar

RECOMMENDATION GENERATION

- Content-based
  - Requires knowledge about the items themselves
  - Requires no knowledge about previous orders

  ➡Provides recommendations of content-wise similar items

- Collaborative
  - Requires knowledge previous orders and transactions
  - Requires **no knowledge about the items** themselves

  ➡Provides recommendations of items other persons chose in this situation

    i. Identify similar persons

    ii. Identify similar transactions, i.e., item combination in shopping cart

Person ID	Items
1	Bread, Oatmeal, Potatoes
2	Apples, 2x Bread, 2x Sugar, Potatoes
3	Apples, Bread, Potatoes, Sugar
4	Bread, Oatmeal, Potatoes, Sugar

ASSOCIATION RULE LEARNING

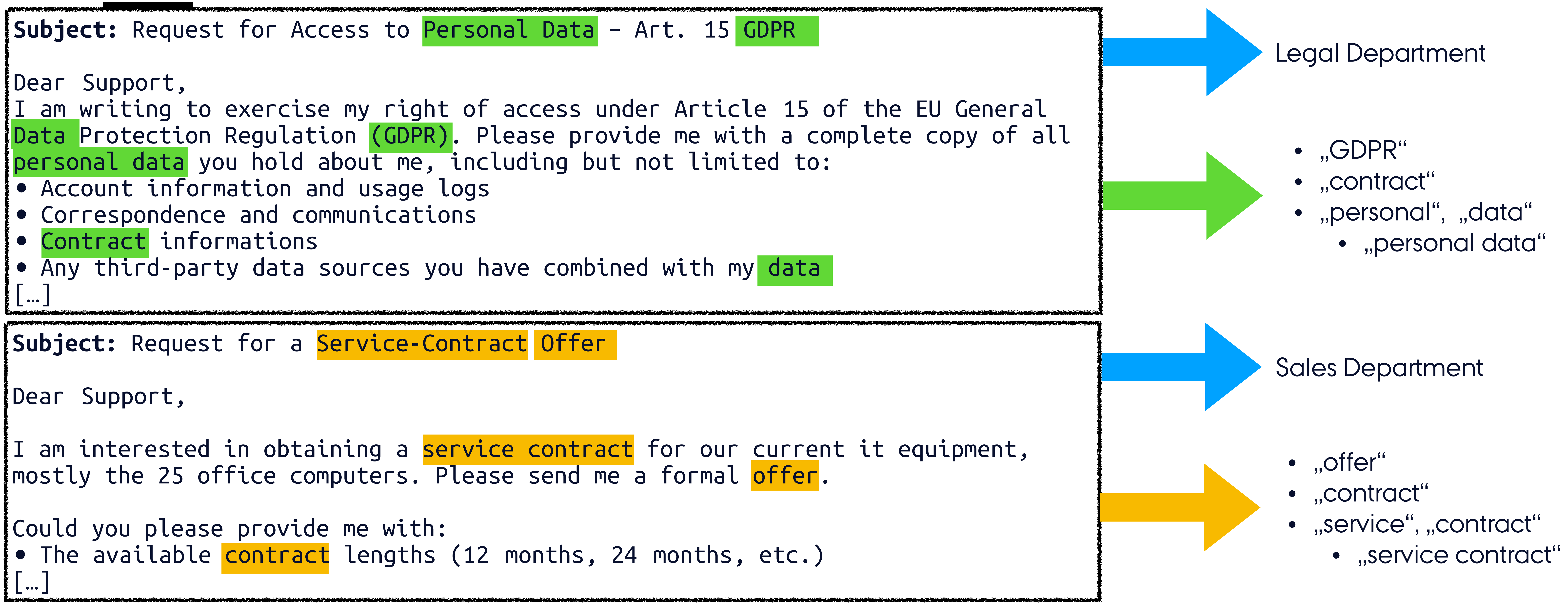
- Collaborative approach
- Algorithmic approach to mine association rules
  - { *Item in cart* } → { *Implies further items* }
- Applied to our example transactions:
  - { Apples } → { Bread, Sugar }
    - Correct in all transactions containing { Apples }
  - { Bread } → { Apples }
    - Correct only in 2 out of 5 transactions containing { Bread }
- Recommendation generation:
  - Create those rules, only keep the ones sufficiently often correct
  - Use rules to suggest new items to customers

Transaction ID	Items
1	Bread, Oatmeal, Potatoes
2	Apples, Bread, Sugar
3	Apples, Bread, Potatoes, Sugar
4	Bread, Potatoes, Sugar
5	Bread, Oatmeal, Potatoes, Sugar

2. FORWARD E-MAILS

- More technical rewrite of problem:
  - Problem:
    Classification of e-mails to a department for handling based on content
  - Class information:
    Name and description of each available department
  - Training data:
    Large amount of previously received e-mails labeled with responsible department
  - Input:
    A newly received e-mail
  - Output:
    The most suitable department for handling
  - Risk:
    Low (e-mail forwarded to wrong department, just manually forward to correct)

EXAMPLE: WORDS IN E-MAILS



Subject: Request for Access to Personal Data - Art. 15 GDPR

Dear Support,
I am writing to exercise my right of access under Article 15 of the EU General Data Protection Regulation (GDPR). Please provide me with a complete copy of all personal data you hold about me, including but not limited to:

- Account information and usage logs
- Correspondence and communications
- Contract informations
- Any third-party data sources you have combined with my data

[…]

Legal Department

- „GDPR“
- „contract“
- „personal“, „data“
  - „personal data“

Subject: Request for a Service-Contract Offer

Dear Support,

I am interested in obtaining a service contract for our current it equipment, mostly the 25 office computers. Please send me a formal offer.

Could you please provide me with:

- The available contract lengths (12 months, 24 months, etc.)

[…]

Sales Department

- „offer“
- „contract“
- „service“, „contract“
  - „service contract“

LOOKING AT THE WORDS

- How to identify relevant words? (automatically)
  - Sets of e-mails, one per department for handling
- Identify pivotal words for assigning to each department
  - Generally less relevant words:
    - „a“, „the“, „is“, „and“ → so-called *stopwords*
  - Topic specific words
    - „contract“, „customer“, „support“
  - Department specific words
    - „GDPR“, „offer“
- Word combinations
  - Use bi-grams or tri-grams, i.e., „service contract“, „personal data“

Legal Department:
- „GDPR“
- „contract“
- „personal“, „data“
  - „personal data“

Sales Department:
- „offer“
- „contract“
- „service“, „contract“
  - „service contract“

TEXT CLASSIFICATION USING TF-IDF

- Term-Frequency (TF)
  - Count the occurrence of each word per e-mail
  - Divide by number of words in e-mail to take length of e-mail into account
- Inverse-Document-Frequency (IDF)
  - Identify *rare*, thus *pivotal* words:
    - Word present in all sets for all departments *not specific*
    - Word often present in one set for one department *very specific*
- Combine together:

  Term-Frequency & Inverse-Document-Frequency → TF-IDF

➡ Relevance value for each word

Compare relevant words extracted from a new mail to the relevant words per department → Assign mail to this department

SIDE NOTE: TF-IDF FORMAL

- Defined for each word (term) and each e-mail (document)
- Produces a per-word score representing the relevance of each word present in any of the e-mails (corpus)

➡ Get the pivotal words in e-mail for assigning to departments

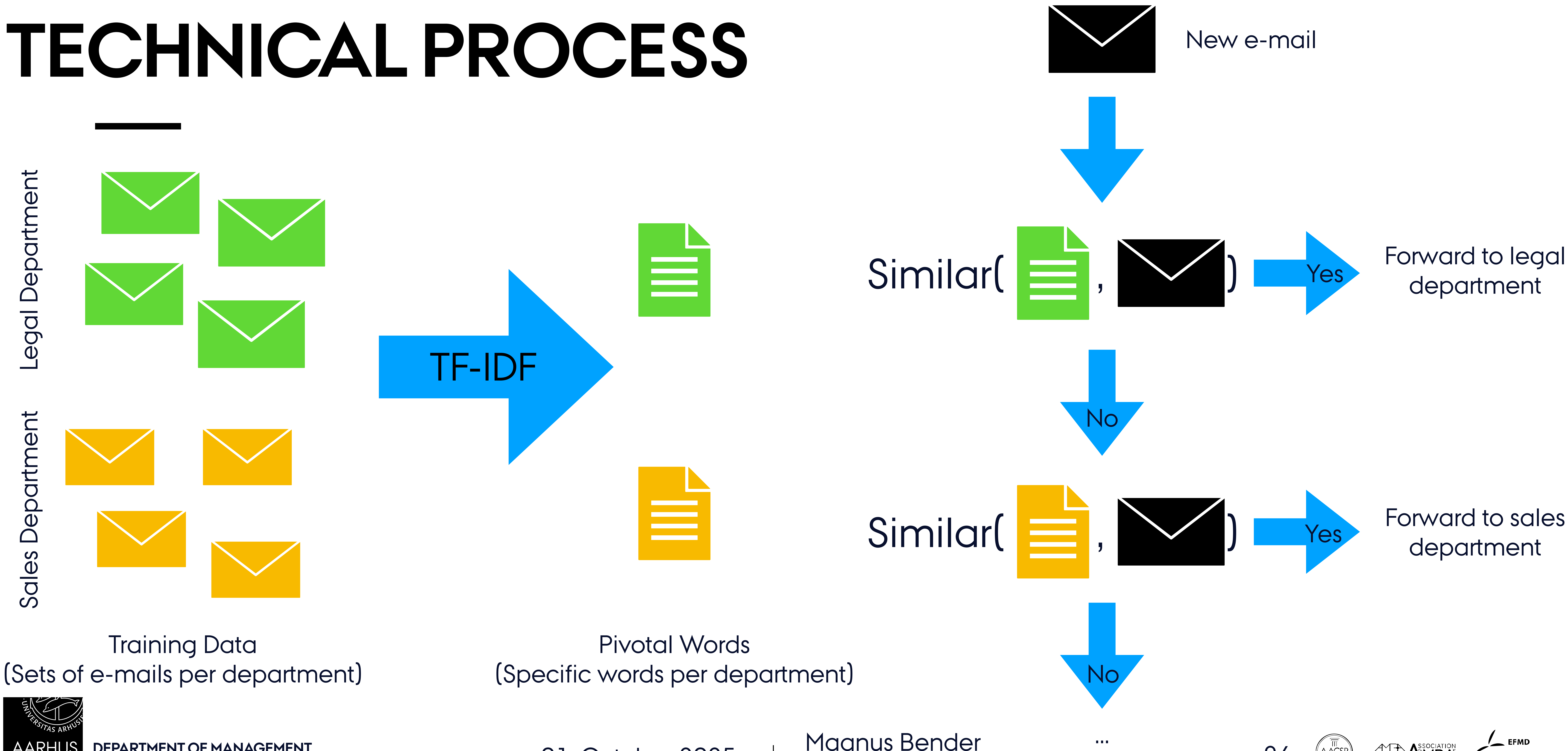
➡ Use these words

$$tf.idf_{t,d} = tf_{t,d} \cdot idf_t = \frac{t_d}{|d|} \cdot \log\left(\frac{N}{df_t}\right)$$

using

$$\begin{aligned} t &= \text{Word} \\ d &= \text{E-Mail} \\ |d| &= \text{Number of words in e-mail } d \\ N &= \text{Number of overall e-mails} \\ df_t &= \text{Number of documents containg word } t \end{aligned}$$

TECHNICAL PROCESS



Training Data
(Sets of e-mails per department)

Pivotal Words
(Specific words per department)

TEXT CLASSIFICATION USING LLMS

a) Train an LLM on the previously received e-mails and labels
   - Requires huge amount of data, training hardware, programming skills
   - May require less hardware, resources after training
   - Trained model is fixed to the departments available in training data

b) List and describe the available departments as part of the prompt and ask LLM to select the most appropriate
   - No training or training data required, less programming
   - Changes in departments can be easily implemented by updating prompts
   - Constant cost and resource requirements
   - Providers like OpenAI may update their models → may require changing prompt

3. INSPECTION OF CVS

- More technical rewrite of problem:
  - Problem:
    Classification of CVs as „complete“ or „incomplete“ based on content
  - Class information:
    Name and description of „complete“ or „incomplete“, i.e., rules for required information in CV
  - Training data:
    Not necessarily available, as required information in CVs changes from case to case
  - Input:
    A PDF file of an CV
  - Output:
    „complete“ or „incomplete“
  - Risk:
    Medium to high (rejection of applicant, even though CV is complete; unable to observe *error*)

TEXT CLASSIFICATION USING LLM OR PYTHON

a) Implement a rule-based classifier in Python
   - Requires programming skills, time for implementation
   - Requires less hardware and resources after implementing
   - No errors if all rules are correctly and fully implemented, but difficult to archive

b) List and describe the required information in a CV as part of the prompt and ask LLM to check if everything is included
   - No training or training data required, less programming
   - Changes in required information can be easily implemented by updating prompt
   - Constant cost and resources requirements
   - Providers like OpenAI may update their models → may require changing prompt

4. SELECTION OF APPLICANT

Not compatible with AI-Act!

- More technical rewrite of problem:
  - Problem:
    Selection of the best fitting applicant based on CVs of all applicants
  - Available Information:
    Requirement for the position based on job offer
  - Input:
    The PDF files of the CVs of all applicants
  - Output:
    The best fitting applicant
  - Risk:
    High to unacceptable (rejection because of system's misunderstanding, …; unable to observe *error*)

Very vague rules (from job offer) used for a very explicit decision (hire). Thereby, hardly any supervision or transparency.

BACK TO THE EXAMPLES

	Task Description	Input	Output	Formalized Problem	(Training) Data	LLM
1	**Personalized recommendations of items in an online shop**	User-selected item	Relevant items	Top-*k* choice	Previous transactions	not suitable
2	**Automatic forwarding of customer's e-mails to correct department**	E-mail text	Suitable department	Classification	Previous e-mails assigned to departments (or rules for LLM)	suitable, but not required
3	**Inspection of CVs of applicants for a position**	PDF file of CV	„Complete" or „Incomplete"	Classification	Rules about necessary information, examples	suitable and helpful
4	**Selection of the applicant to hire for a position**	PDF files of CVs	*Best* applicant	Top-*1* choice	Requirements of position, (prob. examples)	required*

INTERIM SUMMARY

- The use-case heavily influences the tools to use
  - You need to understand your problem from both sides:
    - From the outer workflow site: The problem located in its overall (business) process/ environment
    - From the inner technical site:
      - The inputs and expected outputs of the system and model
      - The information the model uses for making decisions, i.e., data available for training or the rules to provide
- There are typical NLP task and more Data Science (none text-focused) task
  - The latter may be solved without any NLP techniques or LLMs
- There are old-school and new (LLM-based) NLP techniques
  - Old-school techniques require specific data formats or do lossy pre-processing
  - LLMs are de-facto the only solution to process text of unknown structure
- LLMs are quite easy to get started with
  - The task description is given in natural language, the output is natural language again
  - But it may be difficult to get reliable and steady results
  - It may become expensive
  - There might be less computational intensive alternatives

READINGS

READINGS TODAY

- Show LLMs being used to solve many of these older tasks, focusing on *classification*.
    - Sentiment analysis
        - Is the text fundamentally positive or negative?
    - Emotion analysis
        - Which emotions are present in the text?
    - Offensiveness
        - Is this text intended to offend someone?
    - Stance detection
        - Given an issue, does the text agree with or disagree with this issue?
    - Frame detection
        - What is the framing of a particular news story

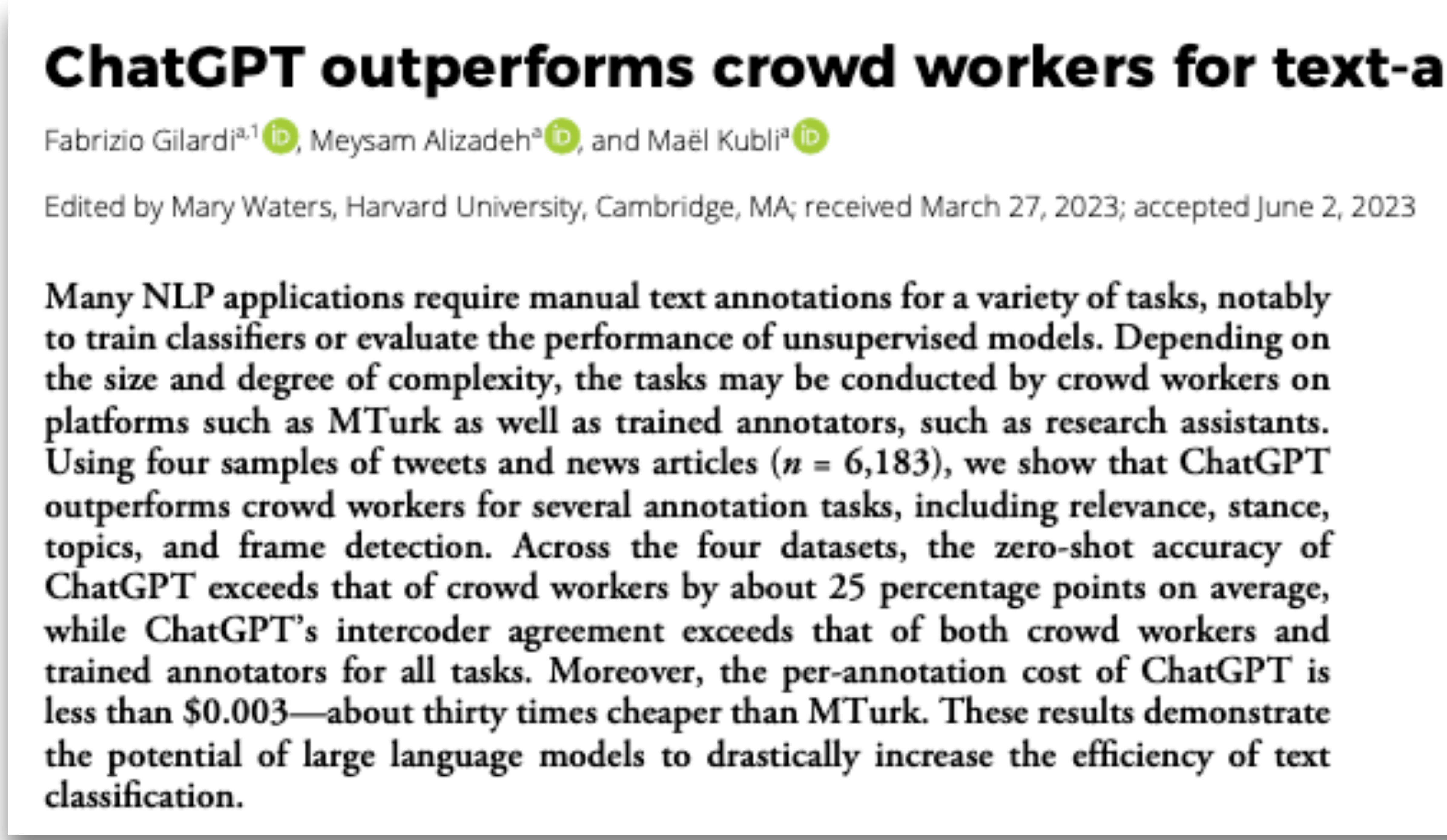
GILARDI ET AL.

- The study uses GPT to "annotate" data.

- Text-annotation is the catch-all term for adding analytical indicators to data.

- Their data consist of four different datasets, comprising a total of 6183 documents, distributed across tweets and newspaper articles.

- Specifically, their study compares GPT with
  - human non-expert annotators, and
  - trained human annotators (e.g. expert coders)



ChatGPT outperforms crowd workers for text-a

Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles (n = 6,183), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

GILARDI ET AL.

- Five different classification tasks:
  - relevance (does this relate to politics, or to content moderation?),
  - stance (is this in favor of, against, or neutral to a specific piece of legislation?)
  - topics (six different classes),
  - and two kinds of frame detection (16 different classes).

- They use the same codebook (i.e. instructions) as they gave to their trained annotators as prompt to GPT and to the MTurk annotators.

- Finally, they compare how well GPT performs.

THEIR APPROACH IN DETAILS

- So, they have some text in a big list that they iterate over

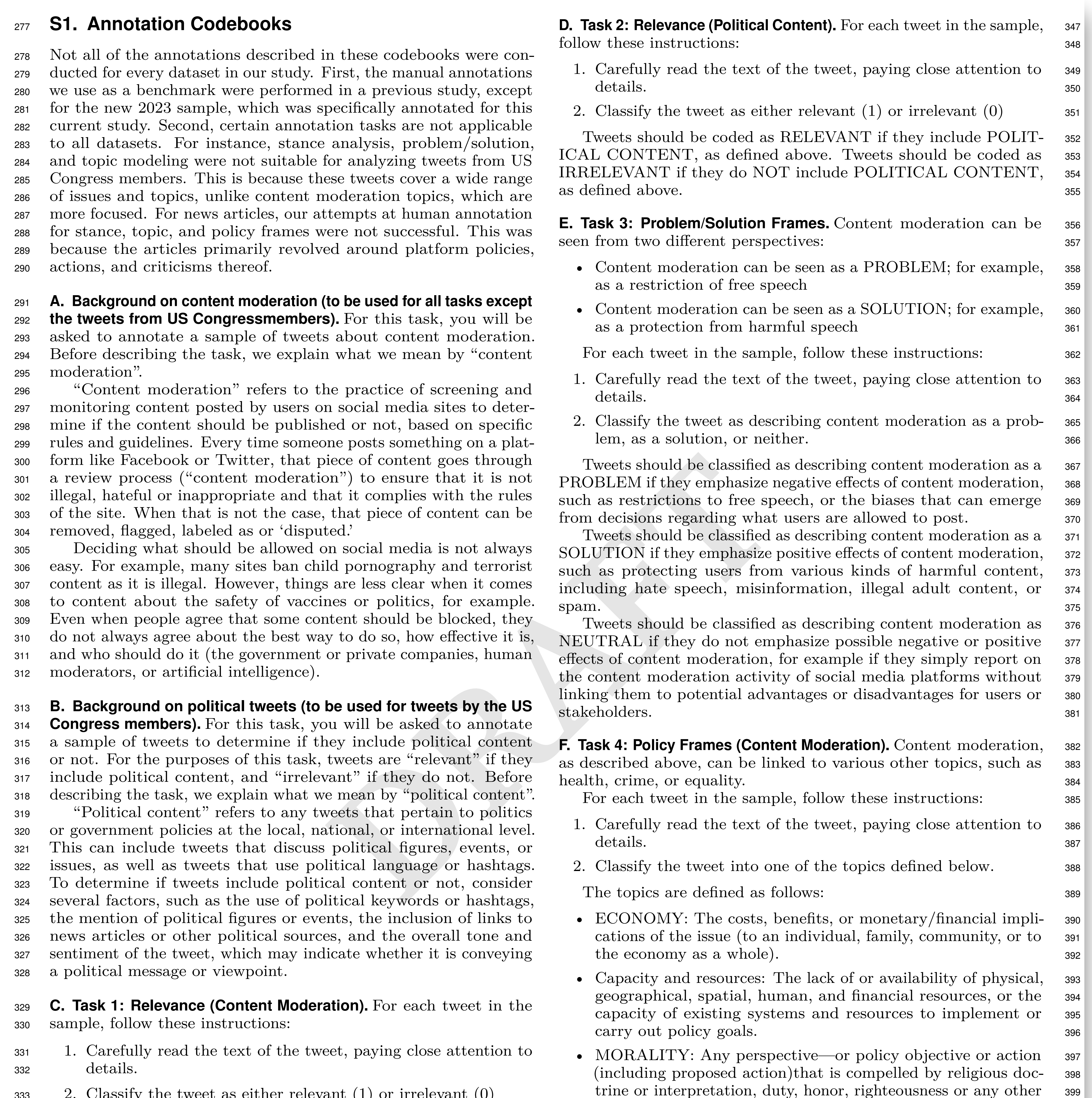
- Write an instruction that outlines the task, provides a definition of all the possible classes, and instruct the LLM to only respond with the name of the relevant class.

- Sends the instruction as system instructions + "here's the tweet I picked, please label it as 'Relevant' or 'Irrelevant':"+ text "

- Have the LLM respond with "Relevant" or "Irrelevant"

- Add the response to a list, and finally compare with "human coders" later.

SUPPLEMENTARY MATERIAL



S1. Annotation Codebooks

Not all of the annotations described in these codebooks were conducted for every dataset in our study. First, the manual annotations we use as a benchmark were performed in a previous study, except for the new 2023 sample, which was specifically annotated for this current study. Second, certain annotation tasks are not applicable to all datasets. For instance, stance analysis, problem/solution, and topic modeling were not suitable for analyzing tweets from US Congress members. This is because these tweets cover a wide range of issues and topics, unlike content moderation topics, which are more focused. For news articles, our attempts at human annotation for stance, topic, and policy frames were not successful. This was because the articles primarily revolved around platform policies, actions, and criticisms thereof.

A. Background on content moderation (to be used for all tasks except the tweets from US Congressmembers). For this task, you will be asked to annotate a sample of tweets about content moderation. Before describing the task, we explain what we mean by "content moderation".

"Content moderation" refers to the practice of screening and monitoring content posted by users on social media sites to determine if the content should be published or not, based on specific rules and guidelines. Every time someone posts something on a platform like Facebook or Twitter, that piece of content goes through a review process ("content moderation") to ensure that it is not illegal, hateful or inappropriate and that it complies with the rules of the site. When that is not the case, that piece of content can be removed, flagged, labeled as or 'disputed.'

Deciding what should be allowed on social media is not always easy. For example, many sites ban child pornography and terrorist content as it is illegal. However, things are less clear when it comes to content about the safety of vaccines or politics, for example. Even when people agree that some content should be blocked, they do not always agree about the best way to do so, how effective it is, and who should do it (the government or private companies, human moderators, or artificial intelligence).

B. Background on political tweets (to be used for tweets by the US Congress members). For this task, you will be asked to annotate a sample of tweets to determine if they include political content or not. For the purposes of this task, tweets are "relevant" if they include political content, and "irrelevant" if they do not. Before describing the task, we explain what we mean by "political content".

"Political content" refers to any tweets that pertain to politics or government policies at the local, national, or international level. This can include tweets that discuss political figures, events, or issues, as well as tweets that use political language or hashtags. To determine if tweets include political content or not, consider several factors, such as the use of political keywords or hashtags, the mention of political figures or events, the inclusion of links to news articles or other political sources, and the overall tone and sentiment of the tweet, which may indicate whether it is conveying a political message or viewpoint.

C. Task 1: Relevance (Content Moderation). For each tweet in the sample, follow these instructions:

1. Carefully read the text of the tweet, paying close attention to details.
2. Classify the tweet as either relevant (1) or irrelevant (0)

D. Task 2: Relevance (Political Content). For each tweet in the sample, follow these instructions:

1. Carefully read the text of the tweet, paying close attention to details.
2. Classify the tweet as either relevant (1) or irrelevant (0)

Tweets should be coded as RELEVANT if they include POLITICAL CONTENT, as defined above. Tweets should be coded as IRRELEVANT if they do NOT include POLITICAL CONTENT, as defined above.

E. Task 3: Problem/Solution Frames. Content moderation can be seen from two different perspectives:

- Content moderation can be seen as a PROBLEM; for example, as a restriction of free speech
- Content moderation can be seen as a SOLUTION; for example, as a protection from harmful speech

For each tweet in the sample, follow these instructions:

1. Carefully read the text of the tweet, paying close attention to details.
2. Classify the tweet as describing content moderation as a problem, as a solution, or neither.

Tweets should be classified as describing content moderation as a PROBLEM if they emphasize negative effects of content moderation, such as restrictions to free speech, or the biases that can emerge from decisions regarding what users are allowed to post.

Tweets should be classified as describing content moderation as a SOLUTION if they emphasize positive effects of content moderation, such as protecting users from various kinds of harmful content, including hate speech, misinformation, illegal adult content, or spam.

Tweets should be classified as describing content moderation as NEUTRAL if they do not emphasize possible negative or positive effects of content moderation, for example if they simply report on the content moderation activity of social media platforms without linking them to potential advantages or disadvantages for users or stakeholders.

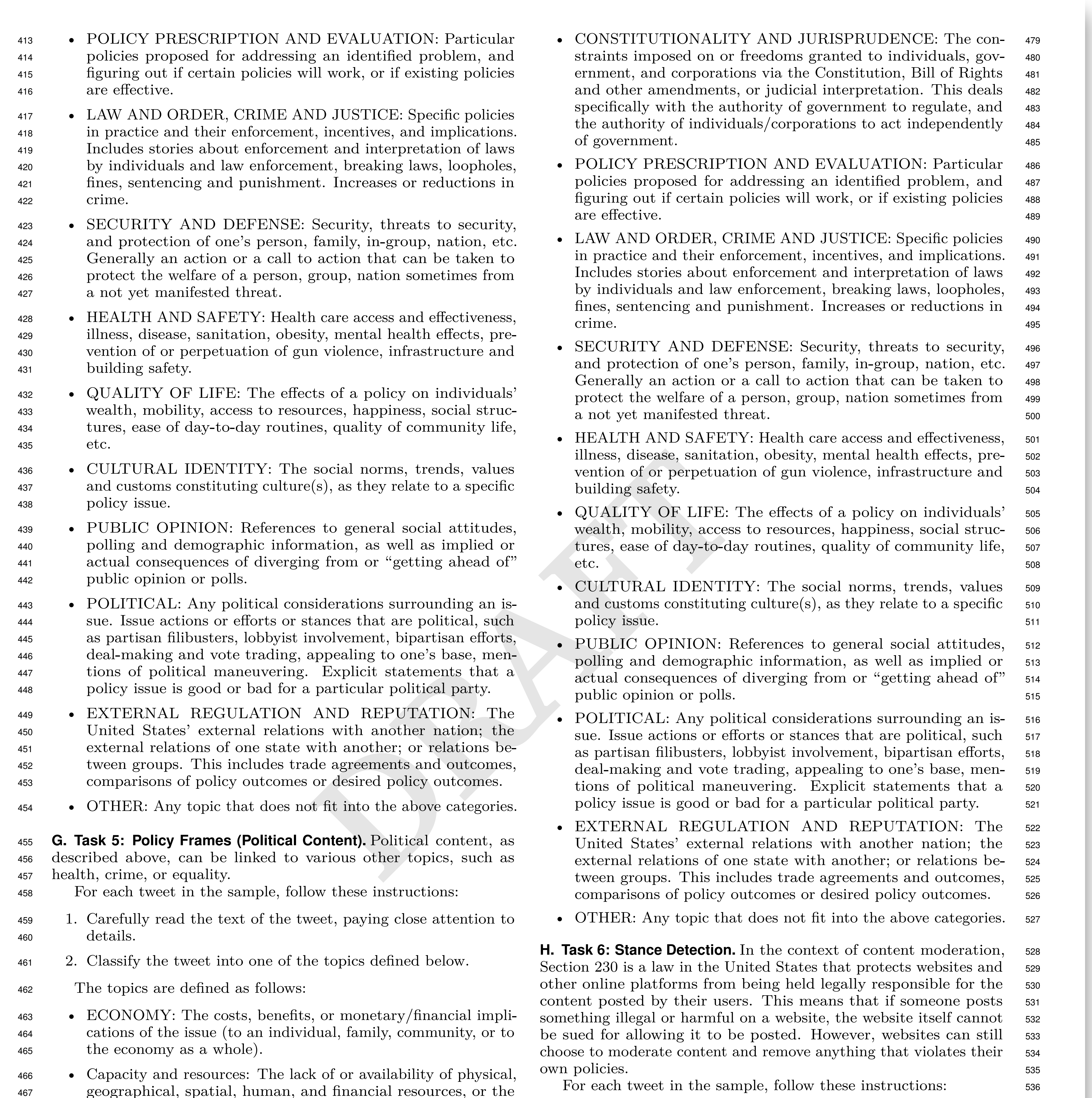
F. Task 4: Policy Frames (Content Moderation). Content moderation, as described above, can be linked to various other topics, such as health, crime, or equality.

For each tweet in the sample, follow these instructions:

1. Carefully read the text of the tweet, paying close attention to details.
2. Classify the tweet into one of the topics defined below.

The topics are defined as follows:

- ECONOMY: The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community, or to the economy as a whole).
- Capacity and resources: The lack of or availability of physical, geographical, spatial, human, and financial resources, or the capacity of existing systems and resources to implement or carry out policy goals.
- MORALITY: Any perspective—or policy objective or action (including proposed action)that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other



- POLICY PRESCRIPTION AND EVALUATION: Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.
- LAW AND ORDER, CRIME AND JUSTICE: Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.
- SECURITY AND DEFENSE: Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.
- HEALTH AND SAFETY: Health care access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.
- QUALITY OF LIFE: The effects of a policy on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.
- CULTURAL IDENTITY: The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.
- PUBLIC OPINION: References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or "getting ahead of" public opinion or polls.
- POLITICAL: Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.
- EXTERNAL REGULATION AND REPUTATION: The United States' external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.
- OTHER: Any topic that does not fit into the above categories.

G. Task 5: Policy Frames (Political Content). Political content, as described above, can be linked to various other topics, such as health, crime, or equality.

For each tweet in the sample, follow these instructions:

1. Carefully read the text of the tweet, paying close attention to details.
2. Classify the tweet into one of the topics defined below.

The topics are defined as follows:

- ECONOMY: The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community, or to the economy as a whole).
- Capacity and resources: The lack of or availability of physical, geographical, spatial, human, and financial resources, or the

- CONSTITUTIONALITY AND JURISPRUDENCE: The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government.
- POLICY PRESCRIPTION AND EVALUATION: Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.
- LAW AND ORDER, CRIME AND JUSTICE: Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.
- SECURITY AND DEFENSE: Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.
- HEALTH AND SAFETY: Health care access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.
- QUALITY OF LIFE: The effects of a policy on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.
- CULTURAL IDENTITY: The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.
- PUBLIC OPINION: References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or "getting ahead of" public opinion or polls.
- POLITICAL: Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.
- EXTERNAL REGULATION AND REPUTATION: The United States' external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.
- OTHER: Any topic that does not fit into the above categories.

H. Task 6: Stance Detection. In the context of content moderation, Section 230 is a law in the United States that protects websites and other online platforms from being held legally responsible for the content posted by their users. This means that if someone posts something illegal or harmful on a website, the website itself cannot be sued for allowing it to be posted. However, websites can still choose to moderate content and remove anything that violates their own policies.

For each tweet in the sample, follow these instructions:

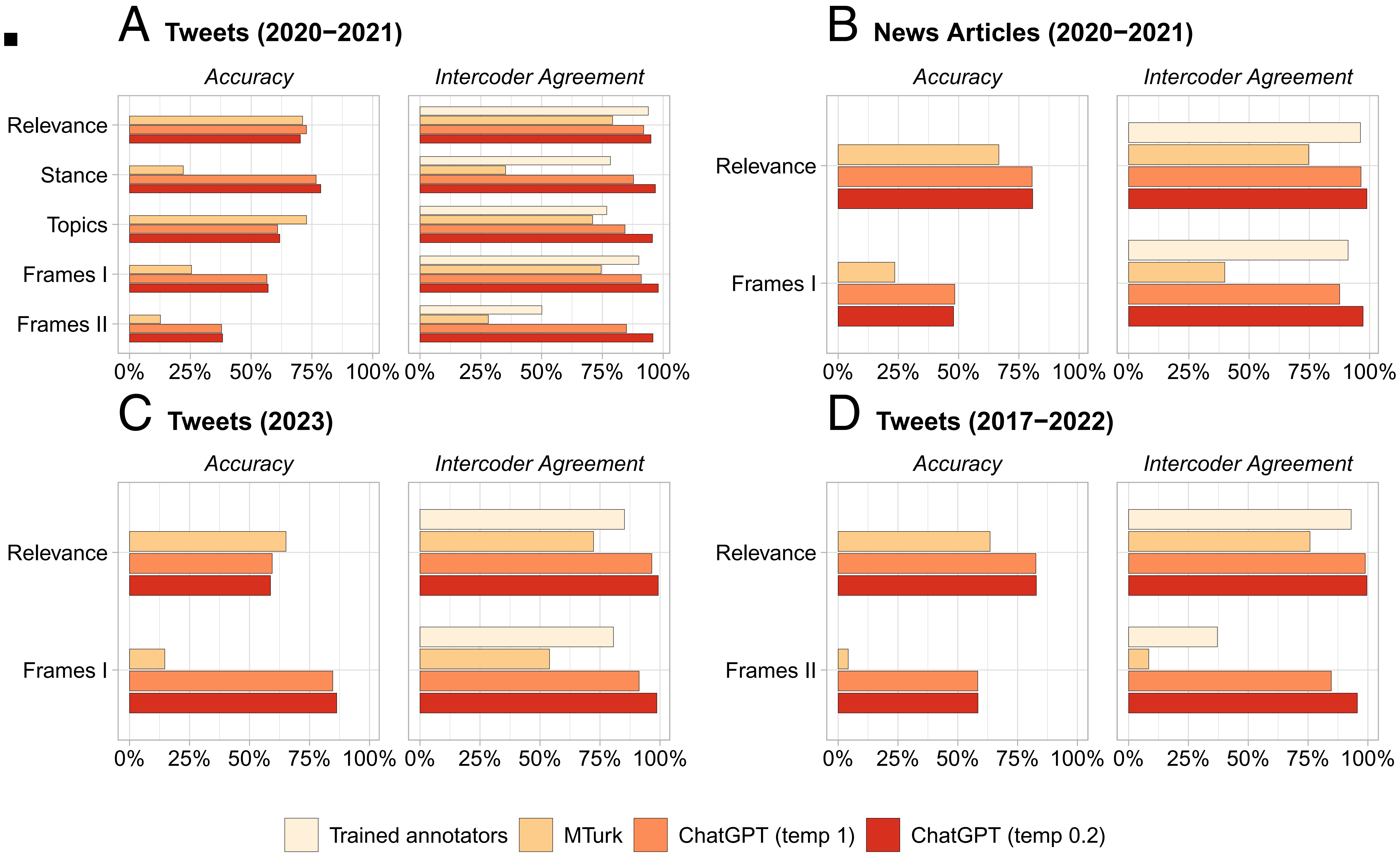
GILARDI ET AL.

So what are we looking at? Let's take some time.

Differences in accuracy between the classification tasks, both for tweets and news articles

High intercoder reliability between trained annotators and GPT.

What does all this tell us?

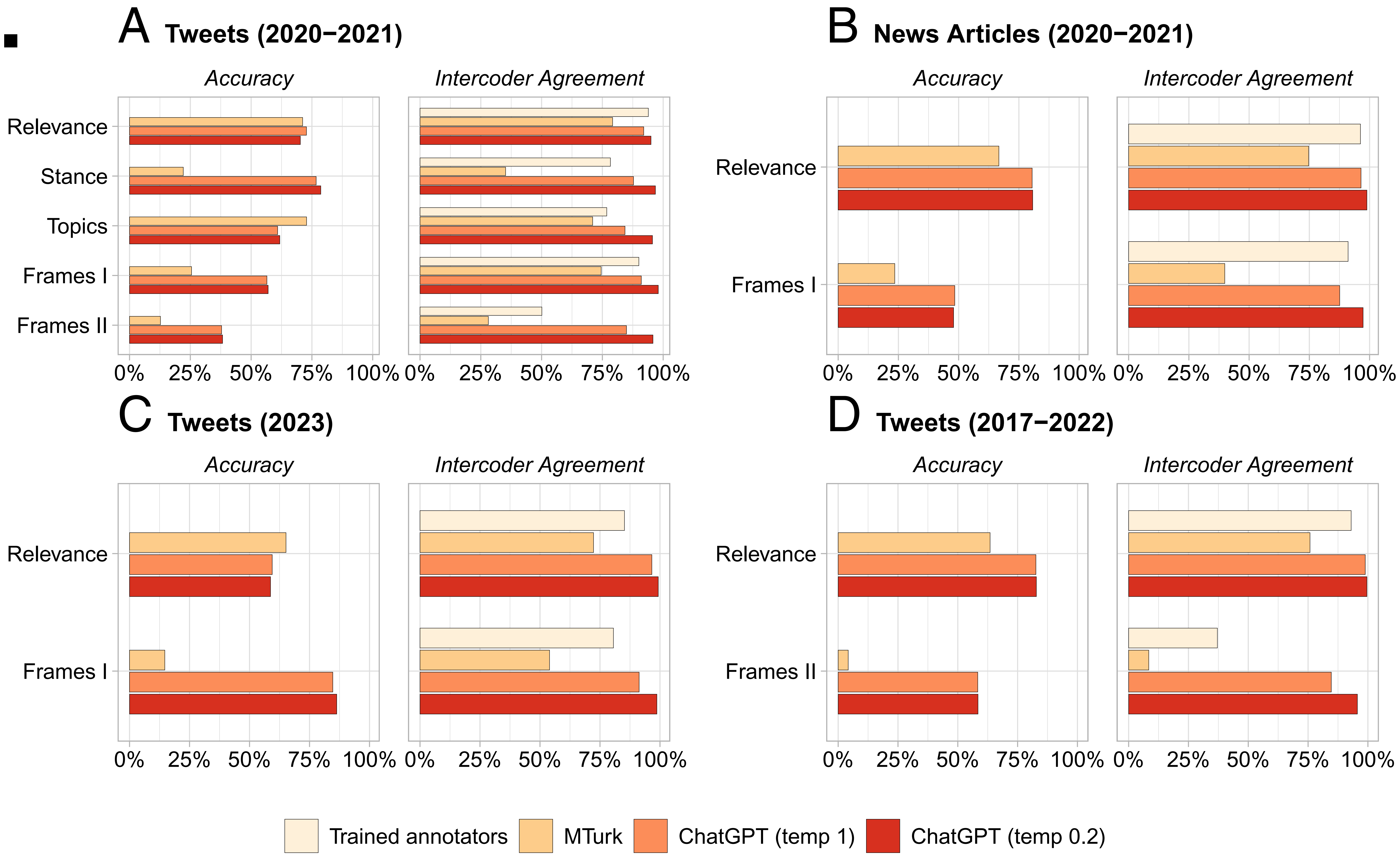


GILARDI ET AL.

I wonder why the accuracy for MTurk is so low on stance, and frames

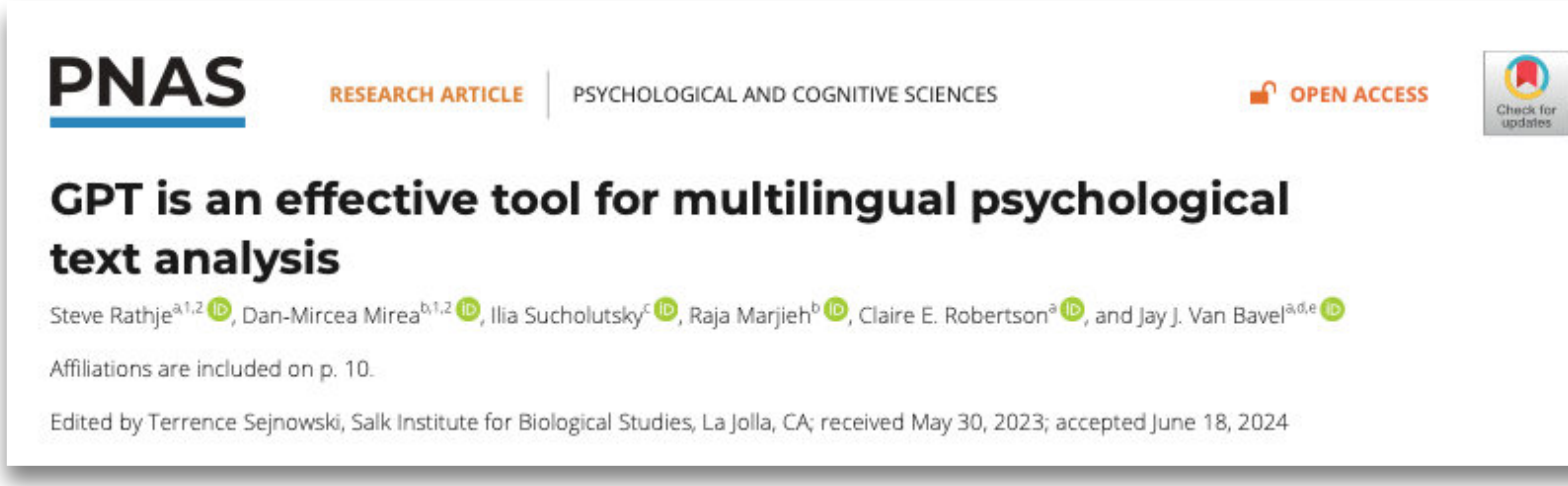
I wonder why intercoder agreement is important for GPT, especially at low temperatures

I wonder how many tweets or news articles fell into each category within their datasets.



RATHJE ET AL.

- Rathje et al. do a similar study but focusing on detecting/ classifying psychological constructs
  - Sentiment: positive/negative
  - Emotions: Anger/joy/Sadness/Optimism
  - Offensiveness: yes/no
  - Discrete sentiment: 1 (very negative) – 7 (very positive)
  - Moral foundations



PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Check for updates

GPT is an effective tool for multilingual psychological text analysis

Steve Rathje[a,1,2], Dan-Mircea Mirea[b,1,2], Ilia Sucholutsky[c], Raja Marjieh[b], Claire E. Robertson[a], and Jay J. Van Bavel[a,d,e]

Affiliations are included on p. 10.

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received May 30, 2023; accepted June 18, 2024

RATHJE ET AL. DATA

Rely entirely on existing datasets

→ This is a good thing – something you should consider too!

SemEval 2017 dataset:

https://huggingface.co/datasets/midas/semeval2017

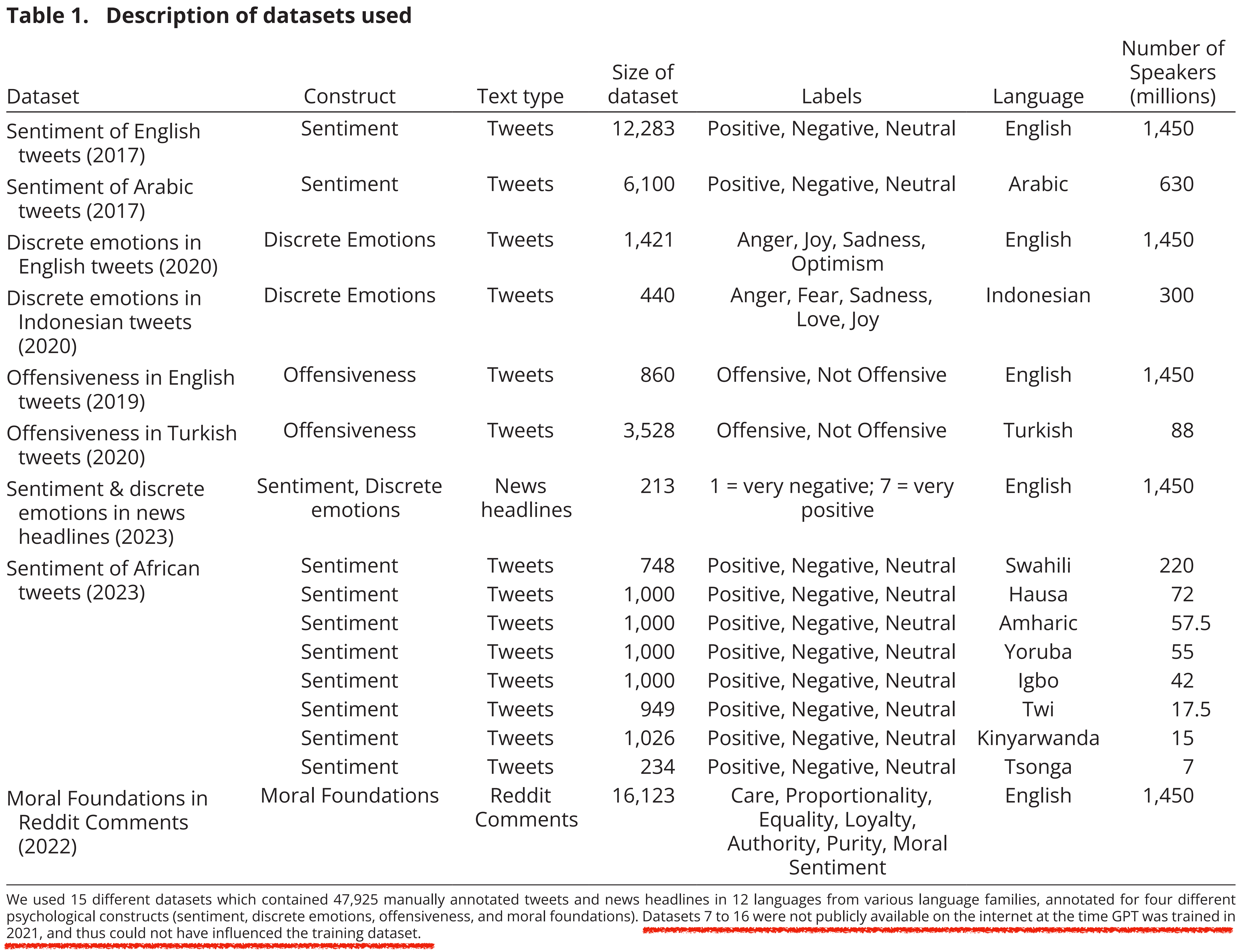


Table 1. Description of datasets used

Dataset	Construct	Text type	Size of dataset	Labels	Language	Number of Speakers (millions)
Sentiment of English tweets (2017)	Sentiment	Tweets	12,283	Positive, Negative, Neutral	English	1,450
Sentiment of Arabic tweets (2017)	Sentiment	Tweets	6,100	Positive, Negative, Neutral	Arabic	630
Discrete emotions in English tweets (2020)	Discrete Emotions	Tweets	1,421	Anger, Joy, Sadness, Optimism	English	1,450
Discrete emotions in Indonesian tweets (2020)	Discrete Emotions	Tweets	440	Anger, Fear, Sadness, Love, Joy	Indonesian	300
Offensiveness in English tweets (2019)	Offensiveness	Tweets	860	Offensive, Not Offensive	English	1,450
Offensiveness in Turkish tweets (2020)	Offensiveness	Tweets	3,528	Offensive, Not Offensive	Turkish	88
Sentiment & discrete emotions in news headlines (2023)	Sentiment, Discrete emotions	News headlines	213	1 = very negative; 7 = very positive	English	1,450
Sentiment of African tweets (2023)	Sentiment	Tweets	748	Positive, Negative, Neutral	Swahili	220
	Sentiment	Tweets	1,000	Positive, Negative, Neutral	Hausa	72
	Sentiment	Tweets	1,000	Positive, Negative, Neutral	Amharic	57.5
	Sentiment	Tweets	1,000	Positive, Negative, Neutral	Yoruba	55
	Sentiment	Tweets	1,000	Positive, Negative, Neutral	Igbo	42
	Sentiment	Tweets	949	Positive, Negative, Neutral	Twi	17.5
	Sentiment	Tweets	1,026	Positive, Negative, Neutral	Kinyarwanda	15
	Sentiment	Tweets	234	Positive, Negative, Neutral	Tsonga	7
Moral Foundations in Reddit Comments (2022)	Moral Foundations	Reddit Comments	16,123	Care, Proportionality, Equality, Loyalty, Authority, Purity, Moral Sentiment	English	1,450

We used 15 different datasets which contained 47,925 manually annotated tweets and news headlines in 12 languages from various language families, annotated for four different psychological constructs (sentiment, discrete emotions, offensiveness, and moral foundations). Datasets 7 to 16 were not publicly available on the internet at the time GPT was trained in 2021, and thus could not have influenced the training dataset.

THEIR PROMPTS

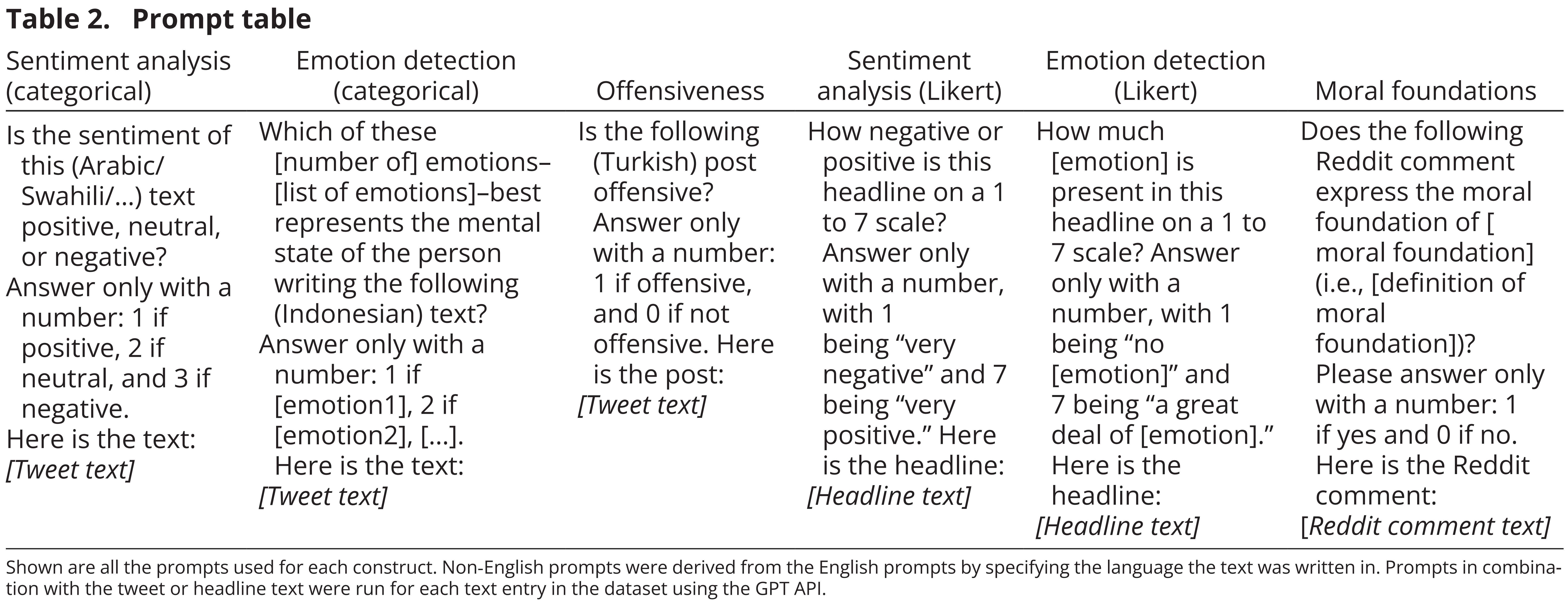


Table 2. Prompt table

Sentiment analysis (categorical)	Emotion detection (categorical)	Offensiveness	Sentiment analysis (Likert)	Emotion detection (Likert)	Moral foundations
Is the sentiment of this (Arabic/ Swahili/...) text positive, neutral, or negative? Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative. Here is the text: *[Tweet text]*	Which of these [number of] emotions–[list of emotions]–best represents the mental state of the person writing the following (Indonesian) text? Answer only with a number: 1 if [emotion1], 2 if [emotion2], [...]. Here is the text: *[Tweet text]*	Is the following (Turkish) post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post: *[Tweet text]*	How negative or positive is this headline on a 1 to 7 scale? Answer only with a number, with 1 being “very negative” and 7 being “very positive.” Here is the headline: *[Headline text]*	How much [emotion] is present in this headline on a 1 to 7 scale? Answer only with a number, with 1 being “no [emotion]” and 7 being “a great deal of [emotion].” Here is the headline: *[Headline text]*	Does the following Reddit comment express the moral foundation of [moral foundation] (i.e., [definition of moral foundation])? Please answer only with a number: 1 if yes and 0 if no. Here is the Reddit comment: [*Reddit comment text]*

Shown are all the prompts used for each construct. Non-English prompts were derived from the English prompts by specifying the language the text was written in. Prompts in combination with the tweet or headline text were run for each text entry in the dataset using the GPT API.

Note

As reported in the paper, they do not use the system instructions at all. They send this as a "user" message.
They inject which language the text is in.
They ask GPT to respond in numbers.

OVERALL RESULTS

- What do we notice here?

- **Note**: We will cover F1 score in the lecture on validating LLMs. But for now, briefly, it is a way to report performance that takes into account the frequencies of codes.

- Smaller languages generally perform worse than English

- **But overall, GPT performs pretty okay**

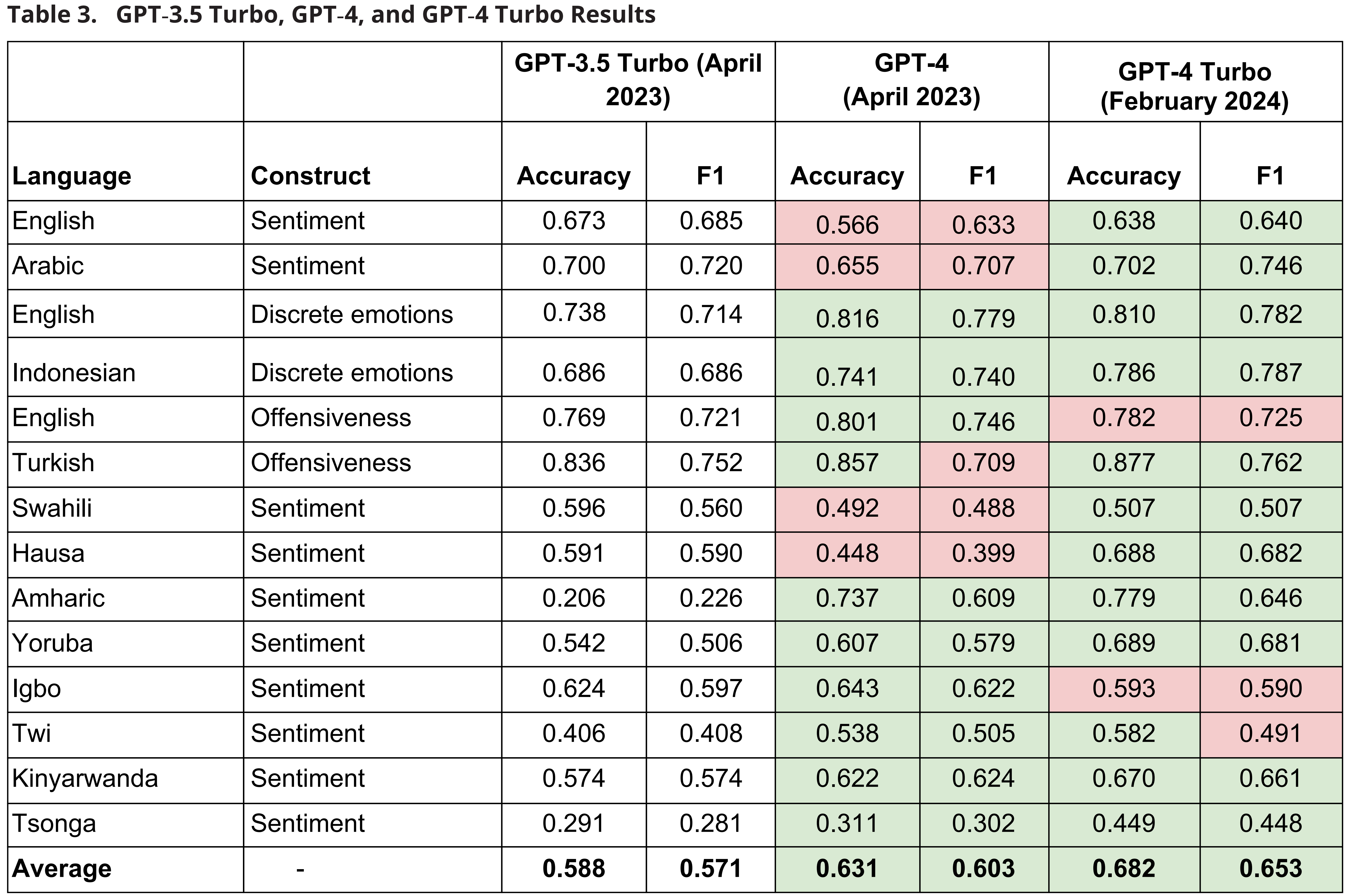


Table 3. GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo Results

		GPT-3.5 Turbo (April 2023)		GPT-4 (April 2023)		GPT-4 Turbo (February 2024)	
Language	**Construct**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
English	Sentiment	0.673	0.685	0.566	0.633	0.638	0.640
Arabic	Sentiment	0.700	0.720	0.655	0.707	0.702	0.746
English	Discrete emotions	0.738	0.714	0.816	0.779	0.810	0.782
Indonesian	Discrete emotions	0.686	0.686	0.741	0.740	0.786	0.787
English	Offensiveness	0.769	0.721	0.801	0.746	0.782	0.725
Turkish	Offensiveness	0.836	0.752	0.857	0.709	0.877	0.762
Swahili	Sentiment	0.596	0.560	0.492	0.488	0.507	0.507
Hausa	Sentiment	0.591	0.590	0.448	0.399	0.688	0.682
Amharic	Sentiment	0.206	0.226	0.737	0.609	0.779	0.646
Yoruba	Sentiment	0.542	0.506	0.607	0.579	0.689	0.681
Igbo	Sentiment	0.624	0.597	0.643	0.622	0.593	0.590
Twi	Sentiment	0.406	0.408	0.538	0.505	0.582	0.491
Kinyarwanda	Sentiment	0.574	0.574	0.622	0.624	0.670	0.661
Tsonga	Sentiment	0.291	0.281	0.311	0.302	0.449	0.448
Average	-	**0.588**	**0.571**	**0.631**	**0.603**	**0.682**	**0.653**

OVERALL RESULTS

- Things that I wonder about:

1. Accuracy for sentiment isn't great.
2. It's higher for Arabic than English which is surprising.
3. We don't know the distributions of the scores, so it's hard to tell if the model just predicts the same thing over and over

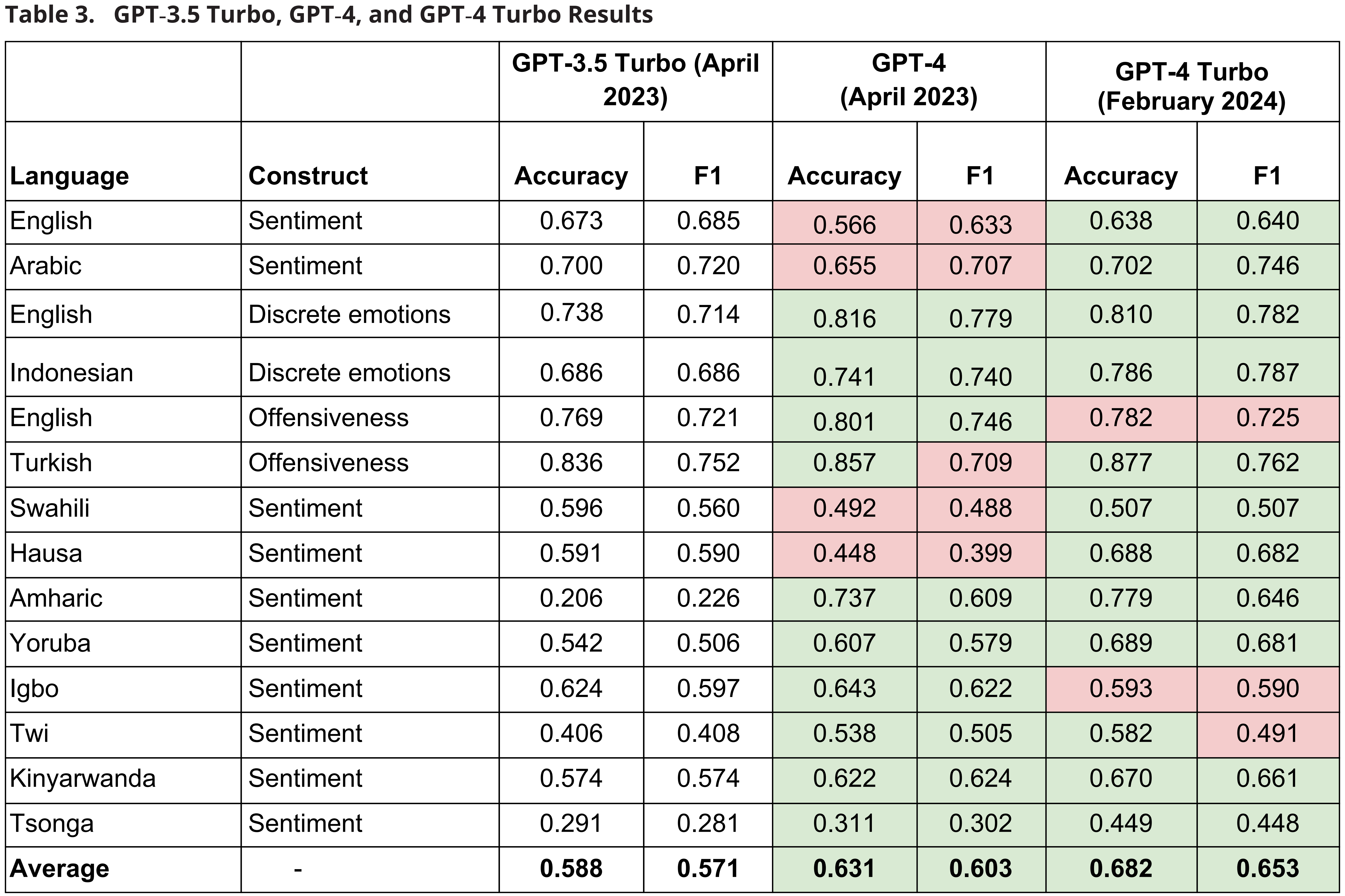
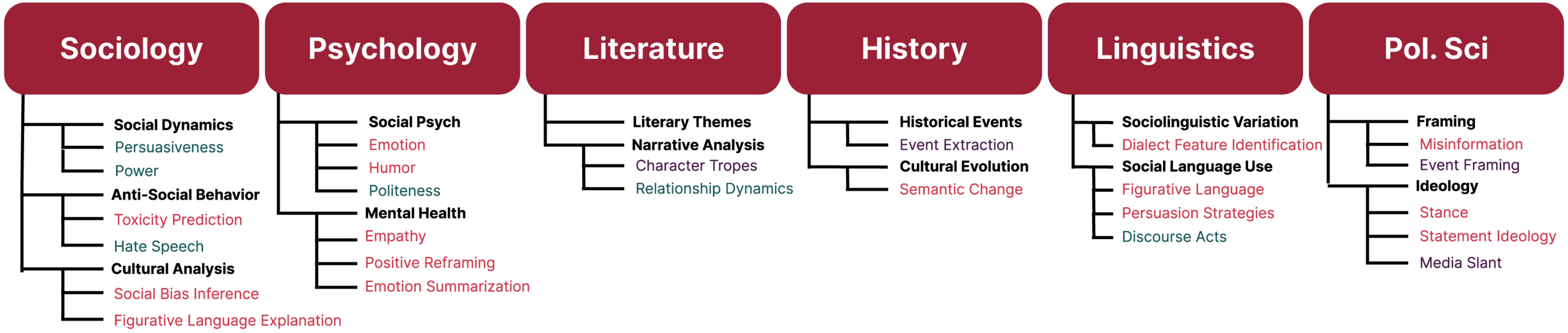


Table 3. GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo Results

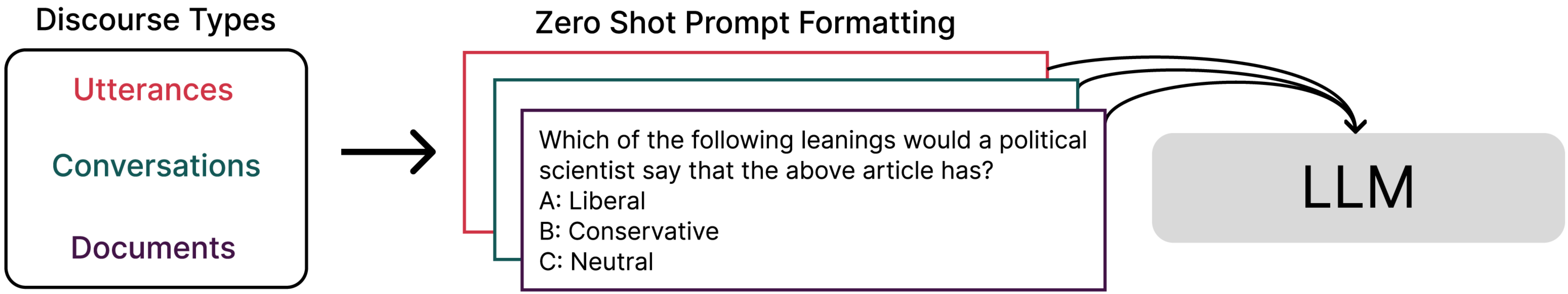
		GPT-3.5 Turbo (April 2023)		GPT-4 (April 2023)		GPT-4 Turbo (February 2024)	
Language	**Construct**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
English	Sentiment	0.673	0.685	0.566	0.633	0.638	0.640
Arabic	Sentiment	0.700	0.720	0.655	0.707	0.702	0.746
English	Discrete emotions	0.738	0.714	0.816	0.779	0.810	0.782
Indonesian	Discrete emotions	0.686	0.686	0.741	0.740	0.786	0.787
English	Offensiveness	0.769	0.721	0.801	0.746	0.782	0.725
Turkish	Offensiveness	0.836	0.752	0.857	0.709	0.877	0.762
Swahili	Sentiment	0.596	0.560	0.492	0.488	0.507	0.507
Hausa	Sentiment	0.591	0.590	0.448	0.399	0.688	0.682
Amharic	Sentiment	0.206	0.226	0.737	0.609	0.779	0.646
Yoruba	Sentiment	0.542	0.506	0.607	0.579	0.689	0.681
Igbo	Sentiment	0.624	0.597	0.643	0.622	0.593	0.590
Twi	Sentiment	0.406	0.408	0.538	0.505	0.582	0.491
Kinyarwanda	Sentiment	0.574	0.574	0.622	0.624	0.670	0.661
Tsonga	Sentiment	0.291	0.281	0.311	0.302	0.449	0.448
Average	-	**0.588**	**0.571**	**0.631**	**0.603**	**0.682**	**0.653**

red = worse than previous version
green = better than previous version

FOR INSPIRATION: WHAT KINDS OF QUESTIONS CAN YOU ANSWER WITH NLP-TASKS



- **Sociology**
  - **Social Dynamics**
    - Persuasiveness
    - Power
  - **Anti-Social Behavior**
    - Toxicity Prediction
    - Hate Speech
  - **Cultural Analysis**
    - Social Bias Inference
    - Figurative Language Explanation
- **Psychology**
  - **Social Psych**
    - Emotion
    - Humor
    - Politeness
  - **Mental Health**
    - Empathy
    - Positive Reframing
    - Emotion Summarization
- **Literature**
  - **Literary Themes**
  - **Narrative Analysis**
    - Character Tropes
    - Relationship Dynamics
- **History**
  - **Historical Events**
    - Event Extraction
  - **Cultural Evolution**
    - Semantic Change
- **Linguistics**
  - **Sociolinguistic Variation**
    - Dialect Feature Identification
  - **Social Language Use**
    - Figurative Language
    - Persuasion Strategies
    - Discourse Acts
- **Pol. Sci**
  - **Framing**
    - Misinformation
    - Event Framing
  - **Ideology**
    - Stance
    - Statement Ideology
    - Media Slant



Discourse Types

Utterances

Conversations

Documents

Zero Shot Prompt Formatting

Which of the following leanings would a political scientist say that the above article has?
A: Liberal
B: Conservative
C: Neutral

LLM

https://aclanthology.org/2024.cl-1.8.pdf

WHEN READING PAPERS ON LLMS, IF YOU WANT TO ACTUALLY KNOW WHAT THEY DID

Read the supplementary materials (appendix in pdf, source code repositories on GitHub) and look at:

- What do their data look like?
- What did their code look like? Do I understand how it works?
- Which figures and tables didn't make it into the paper, and what do they add to the story?
- How did they prompt?
- And how did they instruct the model to respond?

PRACTICAL TIPS FOR YOUR PROJECT WORK

- The articles from today give you
  - Specific things that you can do with LLMs, and for each of those a source of data and a source of code + inspiration
  - At least a dozen available datasets that you could build on, e.g.,
    - semeval = pre-labelled sentiment evaluations
    - CovidET = pre-labelled emotions in reddit posts about COVID
    - Misinfo Reaction Frames = pre-labelled 'intention' dataset
- Think about how you can use either of them or search for further datasets.

SUMMARIZE TODAY

- Take home messages.

WHY SO MANY NON-LLM EXAMPLES

- You need to get a feeling about different NLP tasks
- For each problem to solve using LLMs, take a look at both sides
  - The outside (use-case) and the inside (technical)
- LLMs are not suitable for every problem, but for many
  - LLMs are good in processing unstructured (textual) content
    - If you do not need to look at content, LLMs will not help
    - If you have quite strict rules, it may be more reliable to implement those rules in code
  - There are old-school NLP tasks, which can be carried out in a fast and convenient way without LLMs
    - LLMs are still a super to get startet, but at some point a different implementation may be much more efficient

TUTORIAL ON FRIDAY

- Outlook this week
  - Data type annotations in Python and custom data types using pydantic
  - Use the OpenAI API for analyzing and retrieving structured data
- Preparation (or do it on Friday)
  - Think about a (small) use-case where LLMs can help to extract structured data from (unstructured) text.
  - Have one or two short example input texts ready (may use ChatGPT to create them!)
  - Example
    - Each employee reports their daily time recordings and time spend per project via e-mail
    - Transfer all the information from the e-mails to a structured Excel sheet

EXAMPLE:
TIME TRACKING VIA E-MAILS

Subject: Time Tracking Report – Monday, 2 September 2025

Dear Alex,

I began my workday at eight fifteen in the morning and wrapped up at five twenty-five in the afternoon. During that time I dedicated three hours and ten minutes to the Customer Portal redesign, starting at eight thirty and concluding at eleven forty. After a short break I shifted focus to the API integration for the new payment gateway, working from twelve fifteen until two twenty-five, which amounts to two hours and ten minutes. The remainder of the afternoon, from two thirty until five twenty-five, was spent on the quarterly performance analytics dashboard, where I logged exactly two hours and fifty minutes of development and testing.

Please let me know if you need any further details or clarification on any of the tasks.

Best regards,
Jordan

gpt-oss:120b

QUESTIONS & FEEDBACK: LECTURE(S) & TUTORIAL(S)

Menti Q&A

UNIVERSITAS ARHUSIENSIS

AARHUS
BSS

DEPARTMENT OF MANAGEMENT
AARHUS UNIVERSITY