TEXT MINING

L05. DATA COLLECTION AND ANNOTATION

SUZAN VERBERNE 2022

Universiteit Leiden

TODAY'S LECTURE

- Quiz about week 4
- Evaluation exercise
- How to get example data?
- Challenges of manual annotation
- Inter-rater agreement

QUIZ ABOUT WEEK 4

- The classification task for the 20 newsgroups text dataset is...
  a. Binary
  b. Multi-class
  c. Multi-label
  d. I don't know what the 20 newsgroups text dataset is

QUIZ ABOUT WEEK 4

- The classification task for the 20 newsgroups text dataset is...
  a. Binary
  b. Multi-class
  c. Multi-label
  d. I don't know what the 20 newsgroups text dataset is

https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

QUIZ ABOUT WEEK 4

- We have a collection of 10,000 documents. The term shark occurs in 10 documents. What is the idf for shark?

a. 3
b. 4
c. 5
d. 10

QUIZ ABOUT WEEK 4

- We have a collection of 10,000 documents. The term shark occurs in 10 documents. What is the idf for shark?

a. 3
b. 4
c. 5
d. 10

log(10000/10) = 3

QUIZ ABOUT WEEK 4

- We have a document with length 100 in which the term shark occurs once. According to the log-variant of term frequency, what is the tf of shark for this document?

a. -2
b. 0.01
c. 0
d. 1

QUIZ ABOUT WEEK 4

- We have a document with length 100 in which the term shark occurs once. According to the log-variant of term frequency, what is the tf of shark for this document?

  a. -2
  b. 0.01
  c. 0
  d. 1

1+log(1) = 1

QUIZ ABOUT WEEK 4

- Have you worked on the exercise of week 4 (text categorization)?
  a. I have completed it
  b. I have completed at least half of it
  c. I have started
  d. No

EXERCISE: TEXT CATEGORIZATION

- https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- Some possible challenges you might have encountered:
  - compatibility warnings (Python 3/Python 2)
  - version errors?

```
#from sklearn.model_selection import GridSearchCV
from sklearn.grid_search import GridSearchCV
```

- Please contact the TAs if you have any questions: tmcourse@liacs.leidenuniv.nl

EXERCISE (WEEK 4)

- We have evaluated a classifier for spam on 2000 messages

	True (reference): spam	True (reference): no spam
Assigned: spam	600	400
Assigned: no spam	200	800

EXERCISE (WEEK 4)

- We have evaluated a classifier for spam on 2000 messages

	True (reference): spam	True (reference): no spam
Assigned: spam	600	400
Assigned: no spam	200	800

- What is the precision for the ‘spam’ class?
- What is the recall for the ‘spam’ class?
- What is the precision for the ‘no spam’ class?
- What is the recall for the ‘no spam’ class?

EXERCISE (WEEK 4)

- We have evaluated a classifier for spam on 2000 messages

	True (reference): spam	True (reference): no spam
Assigned: spam	600	400
Assigned: no spam	200	800

- What is the precision for the ‘spam’ class?
- What is the recall for the ‘spam’ class?
- What is the precision for the ‘no spam’ class?
- What is the recall for the ‘no spam’ class?

- 600/1000 = 0.60
- 600/800=0.75
- 800/1000=0.80
- 800/1200≈0.67

CLASSIFIER EVALUATION

Je hebt geretweet

François Chollet
@fchollet

If your classifier is "99% accurate", either you're using the wrong metric (a metric this high is not informative), or you have an overfitting or leakage problem.

Metrics are feedback points on the way towards better models. Not trophies to show off. They should be actionable.

EXAMPLE DATA

WHY DO WE NEED EXAMPLE DATA?

- In supervised learning we need example data for training and evaluation
  - labelled data, reference data, gold-standard data, ground truth data

Doc id	Content	Class
1	request urgent interest urgent	Spam
2	assistance low interest deposit	Spam
3	symposium defence june	No spam
4	siks symposium deadline june	No spam
5	registration assistance symposium deadline	?

HOW TO GET EXAMPLE DATA

1. Use existing labelled data
2. Create new labelled data

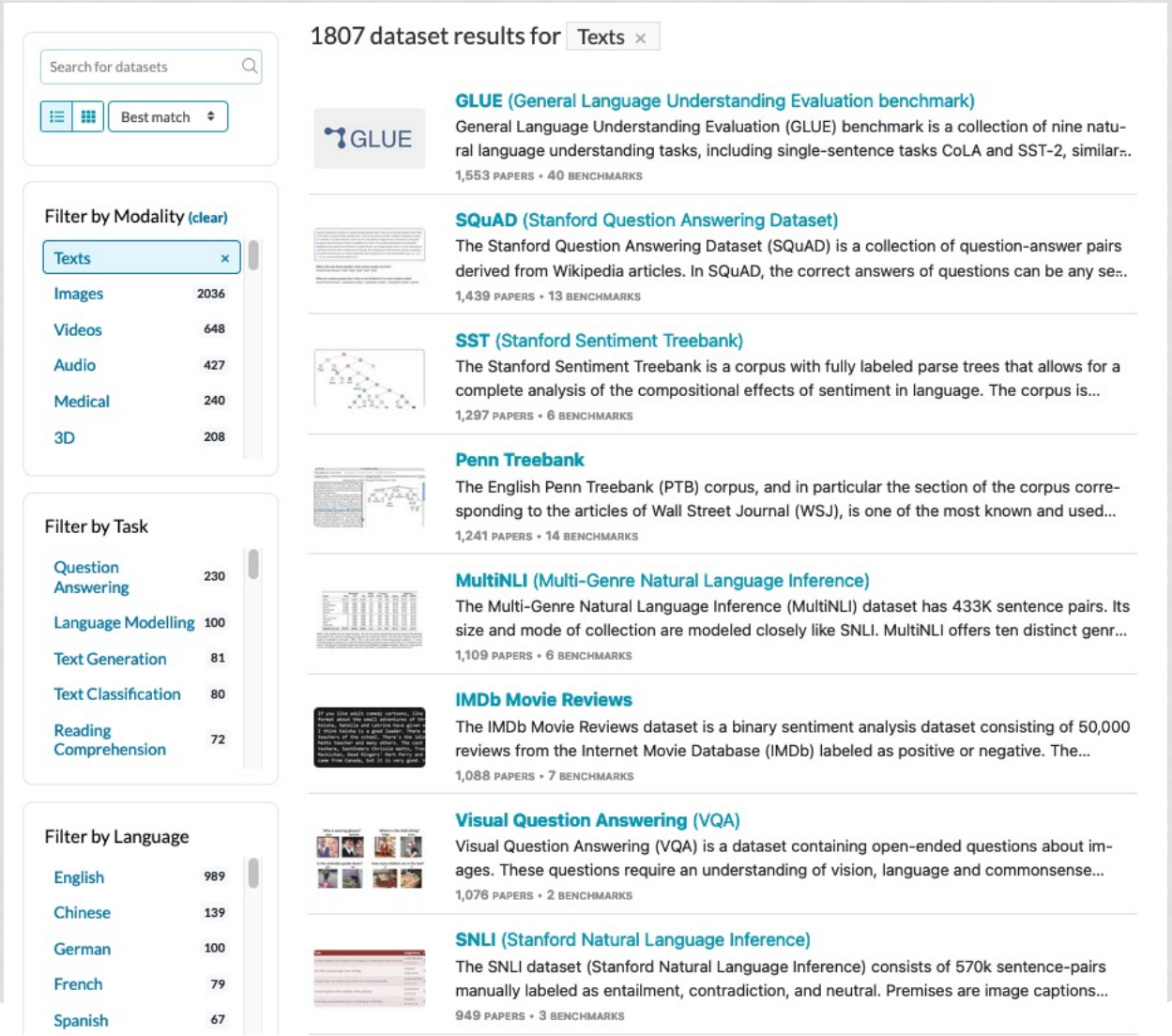
EXISTING LABELLED DATA

HOW TO GET EXAMPLE DATA

1. Existing labelled data
   - A benchmark dataset
   - Existing human labels
   - Labelled user-generated content

BENCHMARK DATA

- Benchmark datasets are used to evaluate and compare methods
  - https://paperswithcode.com/datasets?mod=texts&page=1



BENCHMARK DATA

- Classic text classification benchmark: RCV1 (Reuters Corpus Volume 1)
  - https://paperswithcode.com/dataset/rcv1

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2330" id="root" date="1996-08-20" xml:lang="en">
<title>USA: Tylan stock jumps; weighs sale of company.</title>
<headline>Tylan stock jumps; weighs sale of company.</headline>
<dateline>SAN DIEGO</dateline>
<text>
<p>The stock of Tylan General Inc. jumped Tuesday after the maker of
process-management equipment said it is exploring the sale of the
company and added that it has already received some inquiries from
potential buyers.</p>
<p>Tylan was up $2.50 to $12.75 in early trading on the Nasdaq market.</p>
<p>The company said it has set up a committee of directors to oversee
the sale and that Goldman, Sachs &amp;amp; Co. has been retained as its
financial adviser.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
  <code code="USA"> </code>
</codes>
<codes class="bip:industries:1.0">
  <code code="I34420"> </code>
</codes>
<codes class="bip:topics:1.0">
  <code code="C15">  </code>
  <code code="C152"> </code>
  <code code="C18">  </code>
  <code code="C181"> </code>
  <code code="CCAT"> </code>
</codes>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-08-20"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="SAN DIEGO"/>
<dc element="dc.creator.location.country.name" value="USA"/>
<dc element="dc.source" value="Reuters"/>
</metadata>
</newsitem>
```

Figure 1: An example Reuters Corpus Volume 1 document.

BENCHMARK DATA

- Benchmark data is often created in the context of shared tasks
- Classic Named Entity Recognition (NER) benchmark: CoNLL-2003 shared task
  - https://paperswithcode.com/dataset/conll-2003

Language-Independent Named Entity Recognition (II)

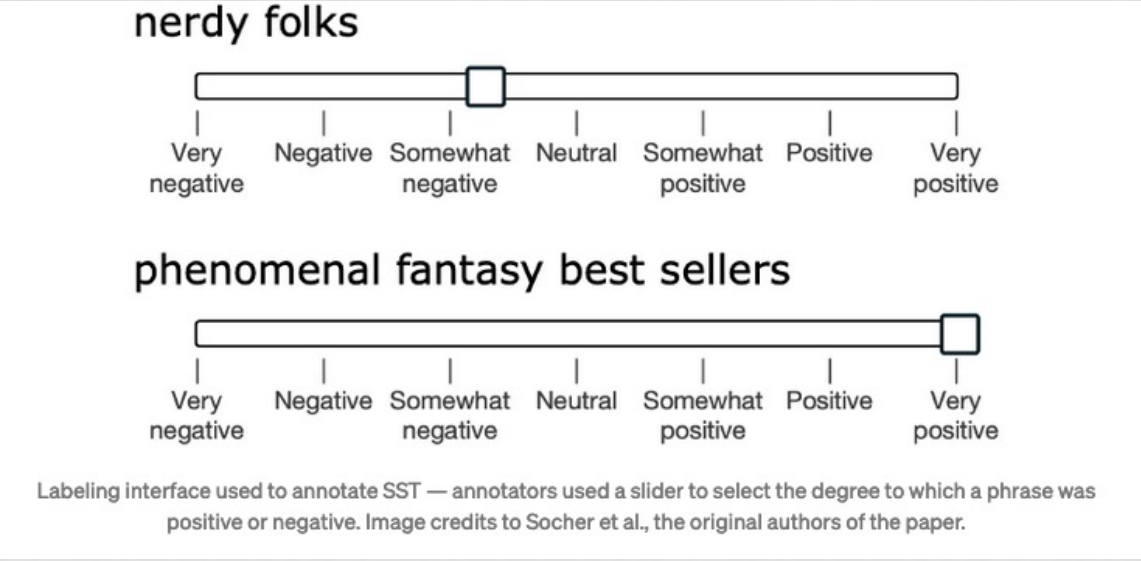
Named entities are phrases that contain the names of persons, organizations, locations, times and quantities. Example:

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

BENCHMARK DATA

Sentiment analysis benchmark: SST (Stanford Sentiment Treebank)

- Movie reviews (Rotten Tomatoes)
- 10,662 sentences, half of which negative and the other half positive
- Sentences split into phrases: 215,154 phrases
- Phrase annotation



Labeling interface used to annotate SST — annotators used a slider to select the degree to which a phrase was positive or negative. Image credits to Socher et al., the original authors of the paper.

BENCHMARK DATA

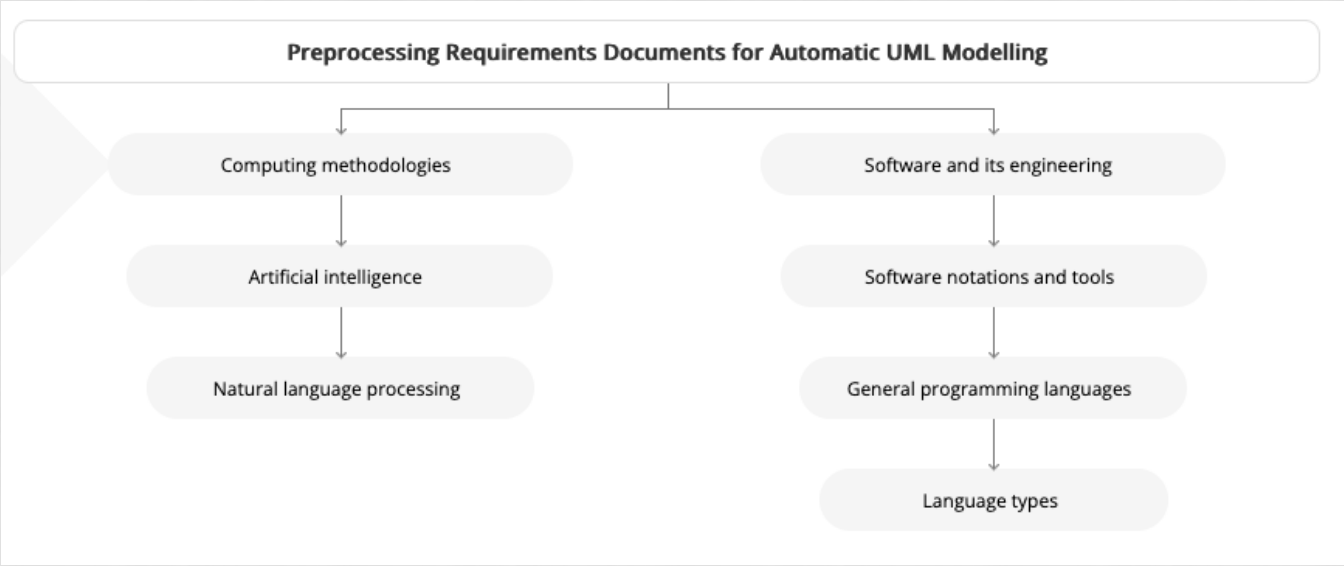
- Advantages
  - High-quality
  - Re-usable
  - Compare results to others
- Disadvantages
  - Not available for every specific problem and data type
    - (e.g. suppose we want to extract medications and side effects mentioned in patient support groups)

HOW TO GET EXAMPLE DATA

1. Existing labelled data
   - A benchmark dataset
   - Existing human labels
   - Labelled user-generated content

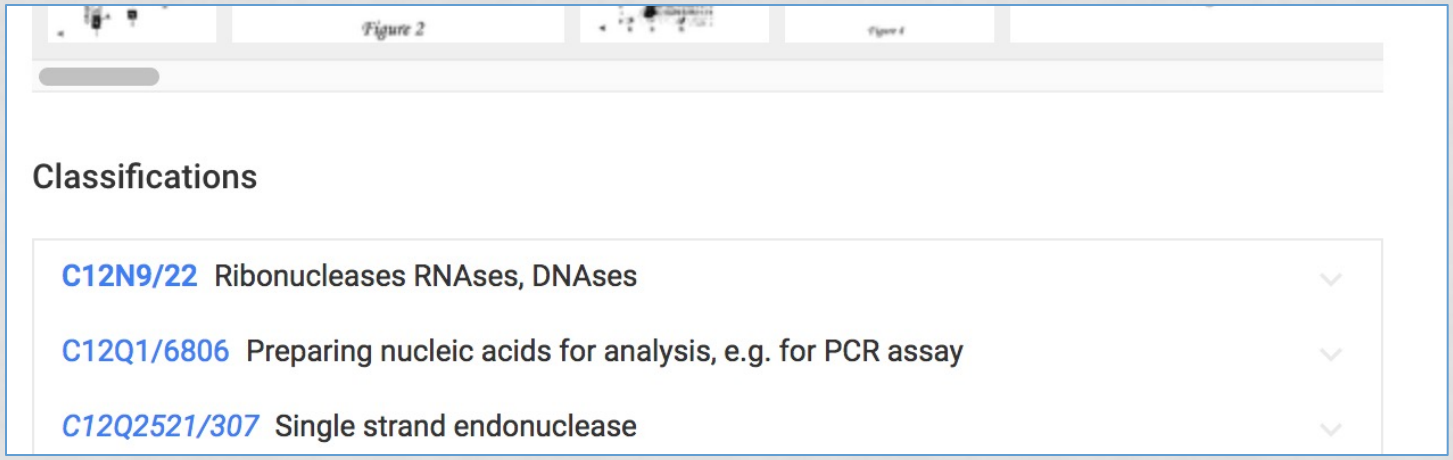
EXISTING HUMAN LABELS

- Labels that were added to items by humans, but not originally created for training machine learning models, e.g.
  - Keywords in digital libraries (e.g. http://dl.acm.org)



EXISTING HUMAN LABELS

- Labels that were added to items by humans, but not originally created for training machine learning models, e.g.
  - The international patent classification (IPC) system:
  - Millions of patents manually classified in a hierarchical classification system by patent experts



EXISTING HUMAN LABELS

- Advantages:
  - High-quality
  - Potentially large
  - Often freely available
- Disadvantages:
  - Not available for every specific problem and data type
  - Not always directly suitable for training classifiers

HOW TO GET EXAMPLE DATA

1. Existing labelled data
   - A benchmark dataset
   - Existing human labels
   - Labelled user-generated content

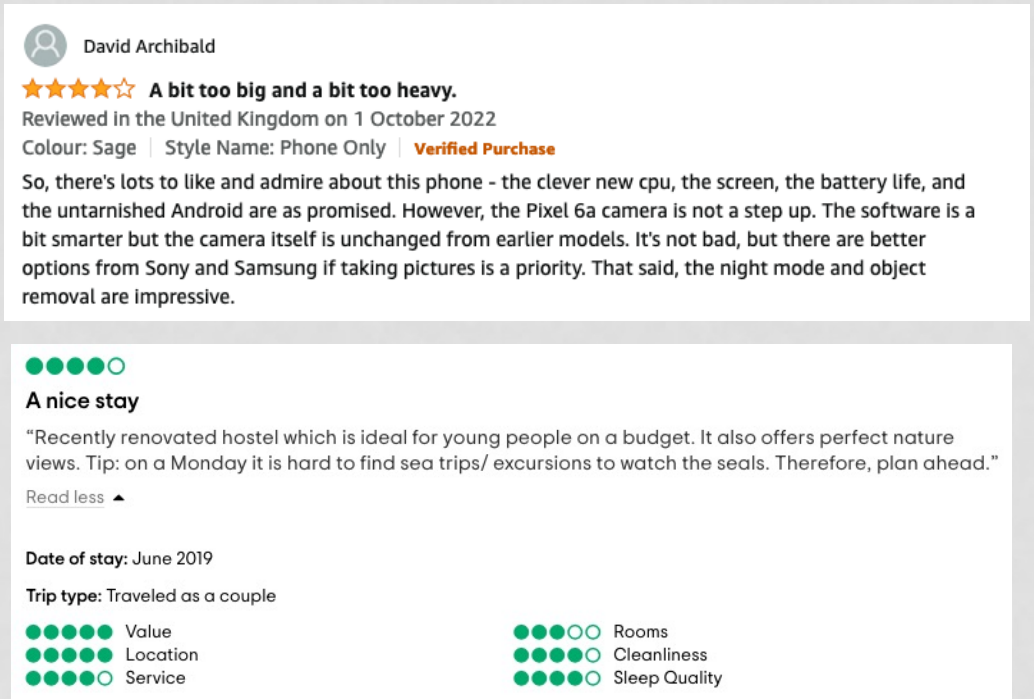
LABELLED USER-GENERATED CONTENT

- Hashtags on Twitter, e.g.
  - Use the hashtag #not to learn to detect sarcasm in text

  - Use network information (e.g. who is connected to whom) to learn user representations

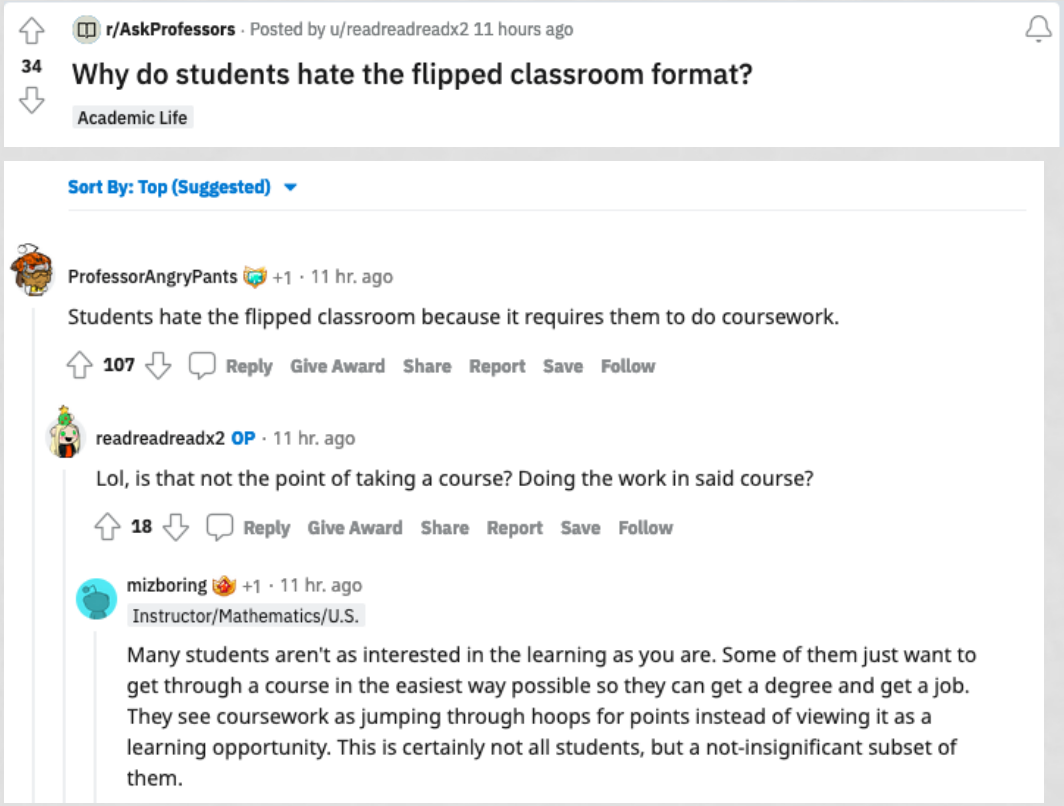
LABELLED USER-GENERATED CONTENT

- Scores and aspects in customer reviews
  - to learn sentiment and opinion



LABELLED USER-GENERATED CONTENT

- Likes of posts to learn which comments are the most interesting



LABELLED USER-GENERATED CONTENT

- Advantages
  - Potentially large
  - Human-created
  - (Freely available, depending on the platform)
- Disadvantages
  - Noisy: often inconsistent
  - May be low-quality
  - Indirect signal

HOW TO GET EXAMPLE DATA

1. Use existing labelled data
2. Create new labelled data

CREATE LABELLED DATA

CREATE LABELLED DATA

1. Make a sample of items
2. Define a set of categories
3. Write annotation guidelines version 1
4. Test and revise the guidelines with new annotators until the guidelines are sufficiently clear
   - The task should be clearly defined
   - But not trivial (‘mark all numbers in the text’)

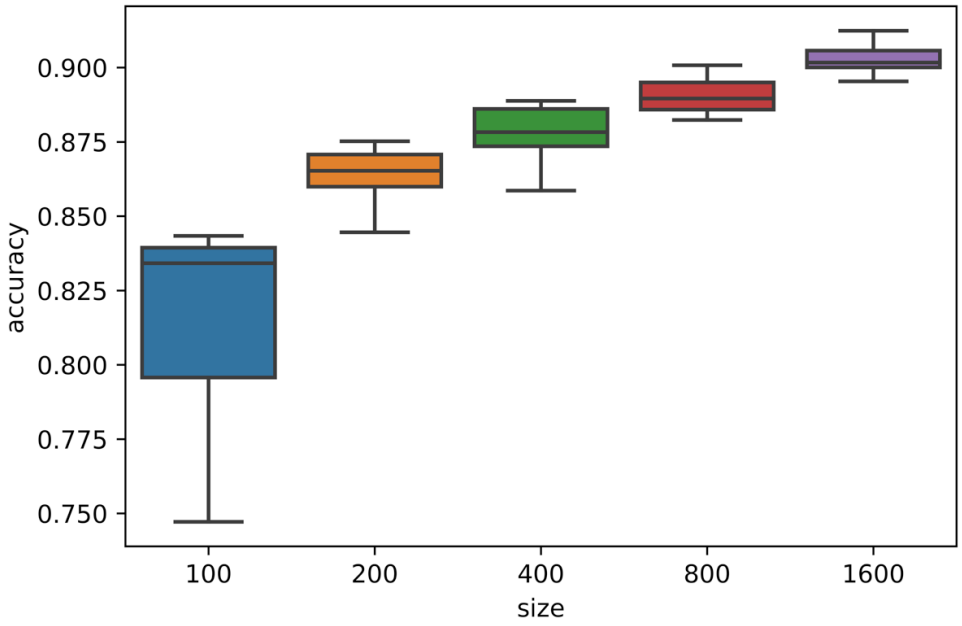
CREATE LABELLED DATA

5. Human annotation
   - Experts
   - Crowdsourcing (Amazon Mechanical Turk, Crowdflower)
6. Compare the labels by different annotators to estimate the reliability of the data (inter-rater agreement)

CREATE LABELLED DATA

How many examples do you need?

- At least dozens/hundreds per category
- The more, the better
- The more difficult the problem, the more examples needed



Example results for a binary classification task with increasing number of training examples

Universiteit Leiden

Van der Burgh, B., & Verberne, S. (2019). The merits of Universal Language Model Fine-tuning for Small Datasets--a case with Dutch book reviews. *arXiv preprint arXiv:1910.00896*.

CROWDSOURCING

- “Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call” -- Jeff Howe, 2006
  - Useful for tasks that humans are typically good at while computers need a lot of examples to do it properly
  - and where no experts are needed (no domain-specific knowledge needed)
  - e.g. object detection in images, name detection in texts

CROWDSOURCING

- Main challenge: quality control
  - Don't pay too little
  - Have a check in the task set-up (e.g. workers need to answer one dummy question to make sure they pay attention)
  - Say that their work is compared to expert annotations (also if you don't)
  - Evaluate the reliability of the data by measuring the inter-rater agreement (this is discussed next)

DATA ANNOTATION

EXERCISES

EXAMPLE TASK

1. Make a sample of items
2. Define a set of categories
3. Write annotation guidelines version 1
4. Test and revise the guidelines with new annotators until the guidelines are sufficiently clear

- Example goal: we want to train a classifier that can recognize the sharing of personal experiences online.
- Example task: label a set of forum messages according to the question "Does the post contain a personal experience? (y/n)"
- Exercise: create a single ordered list of 41 items [y,n] (sample in pdf also on Brightspace, week 5)

1	May your surgery go really well, and your recovery even better -NAME-!!!
2	I was only on Gleevec for 1 year so I guess about 3 years later.
3	I'm not sure what you mean by severe but while taking gleevec I was prone to eye bleeds, more than just blood shot but bright red . It was only in my left eye and at the same time I was being seen by the eye hospital about something else . They new about my medication and gave my eyes a good examination but could find no reason for the bleeding and I had no cause to worry . It has stopped since I stopped taking gleevec .
4	Absolutely brilliant news - Enjoy the journey now!
5	I live in roswell ga 10 Minutes away....i have local oncologists but i take direction from Dr -NAME- in miami . Feel free to PM me
6	Most important: Consult an oncologist who is knowledgeable and experienced with GIST, specifically . Ask lots of questions, and keep asking until you get the answers you understand.
7	yes, bad chills at the beginning of my treatment that I end up at emergency, this happened about three times . After two years taken Gleevec still very cold, but not too bad.
8	Someone asked me yesterday how cancer has changed my life . You said it much better than I did! Thanks.
9	:)
10	Those cysts are quite common in just about everyone . However us GIST warriors it makes us worry . Always good to check and ask for an opinion.
11	The water retention is a pain in the ass . Can you do sports? It's what helps me most.
12	Votrient is another cancer drug that has been used off label for GIST with mixed results . I was on Stivarga also, it made my BP skyrocket . Fornunately, my primary care physican is a good friend and has taken over the managing of BP meds for me . It is great to have a doctor who gives you their personal cell phone number and will take your call or text even on the weekend! I am blessed.
13	I hope you feel better soon ...
14	great news
15	The pharmaceutical companies have assistance programs . Inquire to see if you qualify.
16	Hi this is -NAME-, doing much better each day . Still in alot pain but every thing is going smooth . Cant believe i did it . Thats for all the prayers
17	Good luck & many prayers for you
18	Hello -NAME- and no ma'am not at all!!! I have been very very blessed!
19	Yeah nothing beats that feeling of just awfulness . It's the combination of all side effects I also don't sleep much . There's that and the fact I'm on pain killers a lot . I'm glad I posted on here as not many ppl in Australia have gist well not that I've met and no one really understands what I'm going through.

20	Sending you healing thoughts and prayers that you start to feel better soon . Sorry to hear it's been a rough recovery ...
21	Always nice to hear.
22	A few folks were talking about insomnia and ways to help sleep better . Check this article out :
23	Thank you all for your input! I don't feel so alone :) I will sure try her book! Wish you all the best! Thanks again!
24	Thanks, my main concern was going to a third world country, and risk of infection etc...Think I best be asking the doctors..
25	That would be lovely ..good luck x
26	Wasn't in great shape before surgery--and then eased very slowly into a very active lifestyle (e.g., did my first half marathon 18 mos after surgery). Do the drugs make it more difficult? Absolutely . But don't let the side effects stop you from taking one more step each day....
27	Thank you -NAME-.
28	I hope it's like that for me!!
29	Thank you so much for your support! My mother was diagnosed back in March, and we want to support her and help her through out this process . So I am so thankful for being here for us! God bless you all
30	Wonderful news!
31	Being on gleevec left me in a cloud also and I have been off it for almost 2 years...I'm still in a cloud!
32	my copay was $1600 a month for a year!!! its unbelievable!!!!!!: (
33	I am so sorry for you both . Thoughts and prayers go out to you!!!
34	It is wonderful . I travel 8 hours just to go there . Let me know if you have any questions anytime . I have been on Gleevec 400 mg, then Sutent and back on Gleevec but at 800 mg.
35	THANKS for sharing this, Mar!!! Our heart and prays with you!!!
36	My Husband has Gist, but we travel to Oregon...Dr -NAME-, he's the best ...!
37	All I can say is make really good lists . I couldn't remember ANYTHING to the point of tears sometimes . I've been much better off of Gleevec but still not 100% after 1 year without it.
38	Love the way you put that into words!
39	There is a loophole . The generic is really only approved for CML patients right now . If your doctor looks it up, he can probably TELL them``no substitutions, not approved for GIST''and possibly the pharmacist could look it up as well . In a year though, I think just about everyone will be on a generic, and there is, so far, no reason to think it won't work fine . However, if you are concerned, there is your loophole.
40	Wow -NAME-! I can't wait until we get to the surgery point! We had our second oncology visit Friday and the doctor thinks his are already shrinking due to his increase of appetite and normal blood work.
41	Hahaha I never watched it, but I might give it a try now!

DISCUSSION

- What difficulties did you encounter?

DISCUSSION

- What difficulties did you encounter?
  - Task definition
  - The option to answer '?'
  - The need for clear guidelines
- What about the reliability of your annotations?

1. Make a sample of items
2. Define a set of categories
3. Write annotation guidelines version 1
4. Test and revise the guidelines with new annotators until the guidelines are sufficiently clear
5. Human annotation
6. Compare the labels by different annotators to estimate the reliability of the data

INTER-RATER AGREEMENT

INTER-RATER AGREEMENT

- Human labelled example data = 'the truth' for the classifier
  - both training and evaluation
- But 2 human classifiers do never fully agree
- We therefore always have part of the example data labelled by 2 or 3 raters and then compute the inter-rater agreement
  - to know the reliability of the example data
  - also a measure for the difficulty of the task
- Measure for inter-rater agreement: Cohen's Kappa

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- Pr(a) = actual (measured) agreement: percentage agreed
- Pr(e) = expected (chance) agreement (based on the probabilities of occurrence of each of the values)

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(a) = actual (measured) agreement: percentage agreed
- Pr(e) = expected (chance) agreement

- Pr(a) =

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(a) = actual (measured) agreement: percentage agreed
- Pr(e) = expected (chance) agreement

- Pr(a) = (20+15)/(20+5+10+15) = 35/50 = 0.70

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(e):
  - A1 says 'yes' to 25 items → 50% of all items
  - A2 says 'yes' to 30 items → 60% all items
  - Pr(e,yes) = 0.50*0.60 = 0.30
  - Pr(e,no) =

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(e):
  - A1 says 'yes' to 25 items → 50% of all items
  - A2 says 'yes' to 30 items → 60% all items
  - Pr(e,yes) = 0.50*0.60 = 0.30
  - Pr(e,no) = 0.50*0.40 = 0.20
  - Pr(e) = Pr(e,yes) + Pr(e,no) = 0.50

COHEN'S KAPPA

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(e):
  - A1 says 'yes' to 25 items → 50% of all items
  - A2 says 'yes' to 30 items → 60% all items
  - Pr(e,yes) = 0.50*0.60 = 0.30
  - Pr(e,no) = 0.50*0.40 = 0.20
  - Pr(e) = Pr(e,yes) + Pr(e,no) = 0.50
- K = (0.70 – 0.50)/(1 – 0.50) = 0.20/0.50 = 0.40

INTERPRETATION OF KAPPA

- < 0: no agreement
- 0–0.20: slight agreement
- 0.21–0.40: fair agreement
- 0.41–0.60: moderate agreement
- 0.61–0.80: substantial agreement
- 0.81–1: almost perfect agreement

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

EXERCISE

- Compare your y/n labels to the ones of your neighbour
- Make the agreement table
- Compute the inter-rater agreement between your annotations in terms of Cohen's Kappa
- Example repeated on the next slide

COHEN'S KAPPA EXAMPLE

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(e):
  - A1 says 'yes' to 25 items → 50% of all items
  - A2 says 'yes' to 30 items → 60% all items
  - Pr(e,yes) = 0.50*0.60 = 0.30
  - Pr(e,no) = 0.50*0.40 = 0.20
  - Pr(e) = Pr(e,yes) + Pr(e,no) = 0.50
- K = (0.70 – 0.50)/(1 – 0.50) = 0.20/0.50 = 0.40

CONCLUSIONS

HOMEWORK

- Read:
  - Finin et al. (2010) “Annotating named entities in Twitter data with crowdsourcing”.
  - McHugh (2012). “Interrater reliability: the kappa statistic” (reference paper for the explanation of Kappa)
- Complete assignment 1: text categorization
  - See Brightspace: Assignments -> Assignment 1
  - **Deadline: October 17**
  - Submit your report as PDF and your python code as separate file.
  - Your report should not be longer than 3 pages

REMINDER ABOUT COURSE GRADING

- The assessment of the course consists of
  - a written exam (50% of course grade)
  - practical assignments (50% of course grade)
    - two smaller assignments (10% each) during the course
    - one more substantial assignment (30%) at the end of the course
- Passing the course:
  - The average grade for the written exam and the practical assignments should be 5.5 or higher in order to complete the course.
  - If a task is not submitted the grade for that task is 0
- Re-sit deadline for assignment 1 and 2: January 8.
  Maximum grade at re-sit is 6.

AFTER THIS LECTURE…

- You can describe the advantages and disadvantages of using
  - benchmark data,
  - existing human-labelled data,
  - user-generated content,
  - and crowdsourcing
- You can describe the challenges of manual annotations
- You can calculate inter-rater agreement between two human annotators in terms of Cohen's Kappa