TEXT MINING

L05. DATA COLLECTION AND ANNOTATION

SUZAN VERBERNE 2021

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 task of authorship attribution is...
  a. Binary
  b. Multi-class
  c. Multi-label
  d. I don't know what authorship attribution is

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 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

- What is add-one smoothing in Naive Bayes?
  a. Adding one to the occurrence of each category
  b. Adding one for each term in the vocabulary
  c. Adding one for each term in the vocabulary for each category
  d. Adding one to the posterior probability of a class given a document

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

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 created by someone else
   - Existing human labels
   - Labelled user-generated content

BENCHMARK DATA

- Benchmark datasets are used to evaluate and compare methods
- Classic text classification benchmark: Reuters-21578

```
:GISPLIT=''TRAINING-SET'' OLDID=''12981'' NEWID=''798''>
:DATE> 2-MAR-1987 16:51:43.42</DATE>
:TOPICS><D>livestock</D><D>hog</D></TOPICS>
:TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
:DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Porl
:ongress kicks off tomorrow, March 3, in Indianapolis with 160
)f the nations pork producers from 44 member states determining
.ndustry positions on a number of issues, according to the
Iational Pork Producers Council, NPPC.
)elegates to the three day Congress will be considering 26
:esolutions concerning various issues, including the future
lirection of farm policy and the tax law as it applies to the
.griculture sector. The delegates will also debate whether to
:ndorse concepts of a national PRV (pseudorabies virus) control
.nd eradication program, the NPPC said. A large
:rade show, in conjunction with the congress, will feature
:he latest in technology in all areas of the industry, the NPPC
.dded. Reuter
&\#3;</BODY></TEXT></REUTERS>
```

BENCHMARK DATA

- Benchmark data is often created in the context of shared tasks
- Classic Named Entity Recognition (NER) benchmark: CoNLL-2003 shared task

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: movie review dataset

Large Movie Review Dataset

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.

Large Movie Review Dataset v1.0

BENCHMARK DATA

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

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)

Lang Resources & Evaluation (2018) 52:461–483
https://doi.org/10.1007/s10579-017-9389-4

CrossMark

ORIGINAL PAPER

Creating a reference data set for the summarization of discussion forum threads

Suzan Verberne[1] · Emiel Krahmer[2] · Iris Hendrickx[1] · Sander Wubben[2] · Antal van den Bosch[1]

Published online: 21 April 2017

Abstract In this paper we address extractive summarization of long threads in online discussion fora. We present an elaborate user evaluation study to determine human preferences in forum summarization and to create a reference data set. We

a model can be trained that generates sensible summaries. In addition, we investigated the potential for personalized summarization. However, the results for the three raters involved in this experiment were inconclusive. We release the reference summaries as a publicly available dataset.

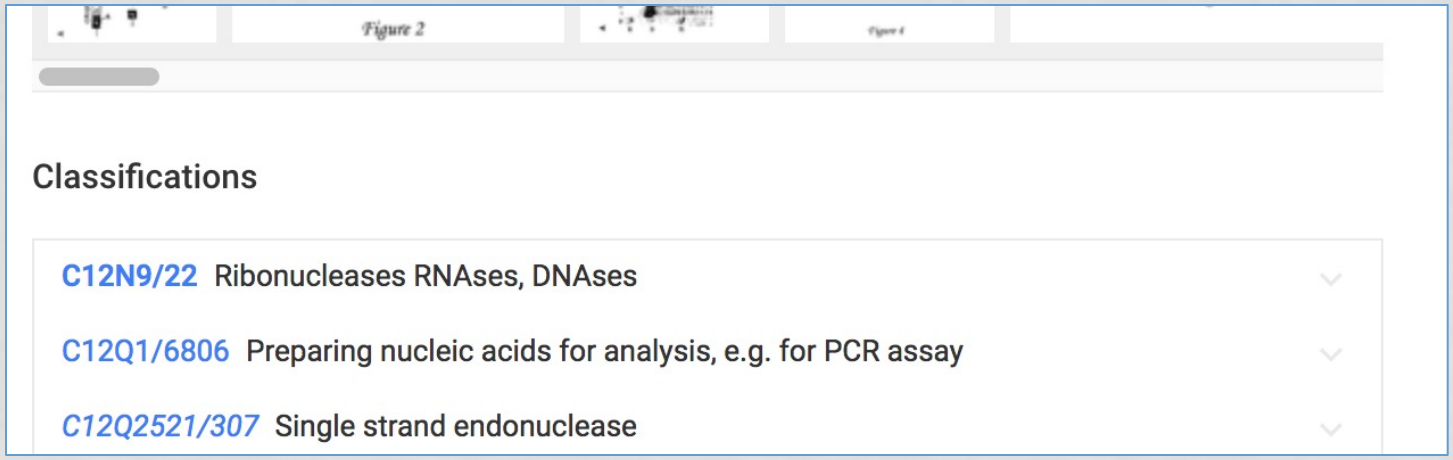
Keywords Summarization · Discussion forums · Data collection · User study · Inter-rater agreement · Evaluation

1 Introduction

Discussion forums on the web come in many flavors, each covering its own topic and having its own community. The user-generated content on web forums is a valuable source for information. In the case of question answering forums such as StackOverflow and Quora, the opening post is a question and the responses are answers to that question. In these forums, the best answer may be selected by the forum community through voting. On the other hand, in discussion forums where opinions and experiences are shared, there is generally no such thing as 'the best answer'. Moreover, discussion threads on a single topic can easily comprise dozens or hundreds of individual posts, which makes it difficult to find the relevant

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 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
  - Not always directly suitable for training classifiers

LABELLED USER-GENERATED CONTENT

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

LABELLED USER-GENERATED CONTENT

- Scores in product reviews
  - to learn sentiment and opinion

★★★★★ **Amazing phone, great buy**

March 30, 2018

Color: Razer Phone - Black | **Verified Purchase**

I don't usually purchase a phone outright but I took a leap on this one. I really enjoy consumer technology but I'm by no means a tech journalist, regardless here is my review and my opinions.

You can look up the specs on the Razer site: Snapdragon 835 w/ 8GB RAM, 64GB Internal memory with expansion slot, 120 Hz 5.7 inch 1440 x 2560 screen, dual 12MP camera, 4000 mAh battery, Android 7.1.1 (as of Mar 2018), Nova Launcher pre-installed. Pretty good specs on paper.

I consider myself a heavy user, I listen to podcasts or stream music all day at work as well as on my hour drive to and from work. I play games on the phone, like Hearthstone, Eternal and Arena of Valor, on breaks. I keep a charger with me but I have only had to use it once at work since I bought the phone a month ago. Even though I have added a 64GB micro SD card I'm just barely at the half way mark on the phones internal storage, though I don't take a lot of pictures regularly, mileage might vary on that. The dual speakers sound amazing, they're loud for phone speakers and they don't get crackly when the volume is all the way up. The equalizer is very intuitive and super customizable for all different types of listening preferences. I use Bluetooth at work for listening and when going from a Nexus 6P to this phone there is a noticeable difference in the sound quality. Not sure what tech Razer used but the sound, even through Bluetooth, is so much more clear and base hits harder. The screen is beautiful, super responsive and sharp as a tack. I've watched videos, played games and read comics on the phone and haven't run into anything that doesn't look amazing. One of my favorite parts about this phone, there is almost no pre-installed bloat wear.

LABELLED USER-GENERATED CONTENT

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

LABELLED USER-GENERATED CONTENT

- Advantages
  - Potentially large
  - Human-created
  - Freely available
- Disadvantages
  - Noisy: often inconsistent
  - May be low-quality
  - Indirect signal

CREATE LABELLED DATA

Universiteit Leiden

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’)

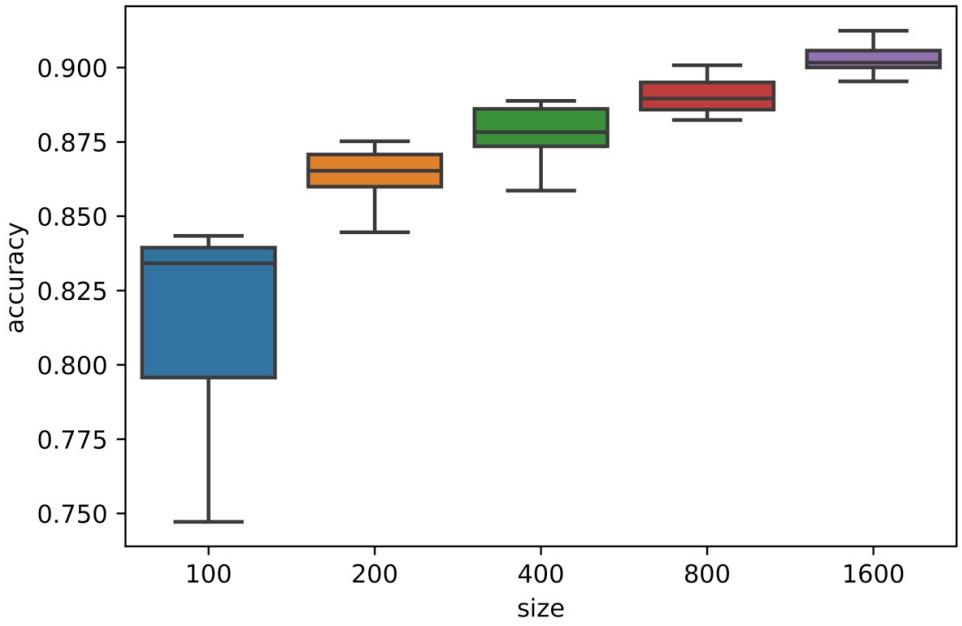
CREATE LABELLED DATA

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

Universiteit Leiden

Suzan Verberne 2021

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)

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

Universiteit Leiden

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

SUZAN VERBERNE 2021

Universiteit Leiden

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 18**
  - 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 16.
  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