Supervised Social Media Text Analysis

Max Pellert

University of Konstanz

Social Media Data Analysis

So Far

- **Block 1: Introduction**
  - Introduction to social media data analysis within social data science
  - Algorithms and digital traces: The case of Google trends
  - Ethics and privacy in social media data analysis

- **Block 2: Social dynamics**
  - Social impact theory and its application to social media
  - Social trends and the Simmel effect

- **Block 3: Text in social media**
  - Dictionary methods in social media data analysis
    - Emotion measurement
    - Basics of dictionary methods and their application to sentiment analysis
    - Application examples of dictionary methods

Outline

1. When dictionary methods go wrong
2. Supervised methods in NLP
3. Comparing methods in SMDA

The digital traces of pagers

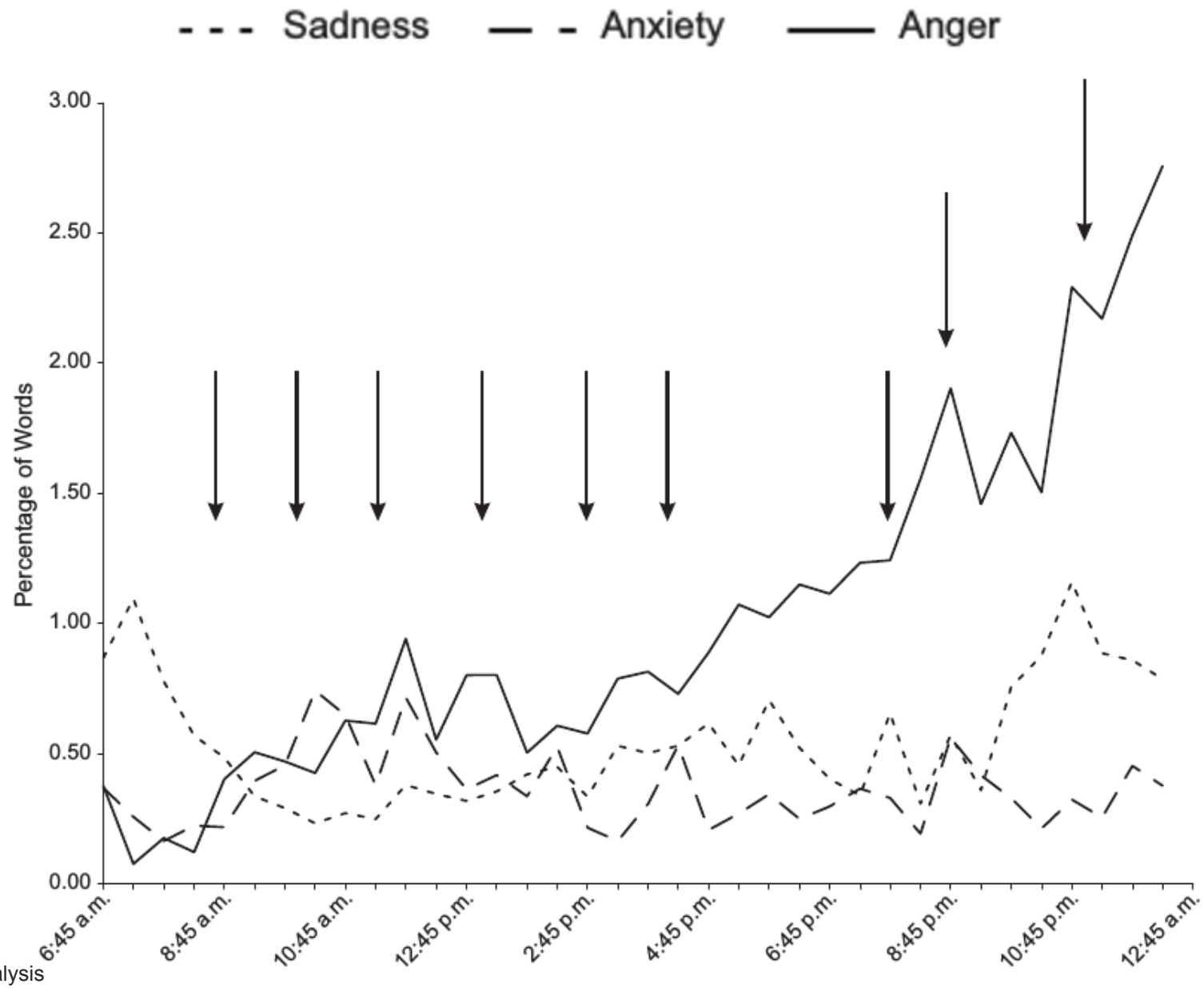
Back in the 90s, pagers were a common form of mobile communication. To send a message to a pager, you could call a special phone number, say your message, and the text of the message would appear in the screen of the pager.

Emotions in pagers after 9/11

In this study, we analyzed use of emotional words in messages sent to text pagers within the United States on September 11. The pager transcripts, published anonymously and freely available on the Internet (WikiLeaks, 2009), total more than 573,000 lines and 6.4 million words from more than 85,000 distinguishable pagers. Pager data are sorted into 5-min blocks according to the time the messages were sent. We used all 216 data blocks from 6:45 a.m. (2 hr before the first attack) to 12:44 a.m. (18 hr after the first attack).

For each of these data blocks, we computed the percentage of words related to (a) sadness (e.g., *crying*, *grief*), (b) anxiety (e.g., *worried*, *fearful*), and (c) anger (e.g., *hate*, *annoyed*) using the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001), currently the most widespread and best-validated software for automatic text analysis in psychological research (Mehl, 2006; for LIWC analyses of online diaries before and after September 11, see Cohn et al., 2004).

The Emotional Timeline of September 11, 2001. Mitja D. Back, Albrecht C.P. Küfner, and Boris Egloff. Psychological Science (2010)


Sadness
Anxiety
Anger
Percentage of Words
3.00
2.50
2.00
1.50
1.00
0.50
0.00
6:45 a.m.
8:45 a.m.
10:45 a.m.
12:45 p.m.
2:45 p.m.
4:45 p.m.
6:45 p.m.
8:45 p.m.
10:45 p.m.
12:45 a.m.

Not so angry americans

More than a third of anger words appeared in messages like these:

2001-09-12 02:25:12 Arch [0987275] C ALPHA
s0191: 09/11 13:18:30 Reboot NT machine gblnetnt05 in cabinet 311R at 13/1CMP:CRITICAL:Sep 11 13:18:30

2001-09-12 02:25:14 Arch [0987275] C ALPHA
s0191: 09/11 13:19:18 Reboot NT machine gblnetnt07 in cabinet 311R at 13/1CMP:CRITICAL:Sep 11 13:19:18

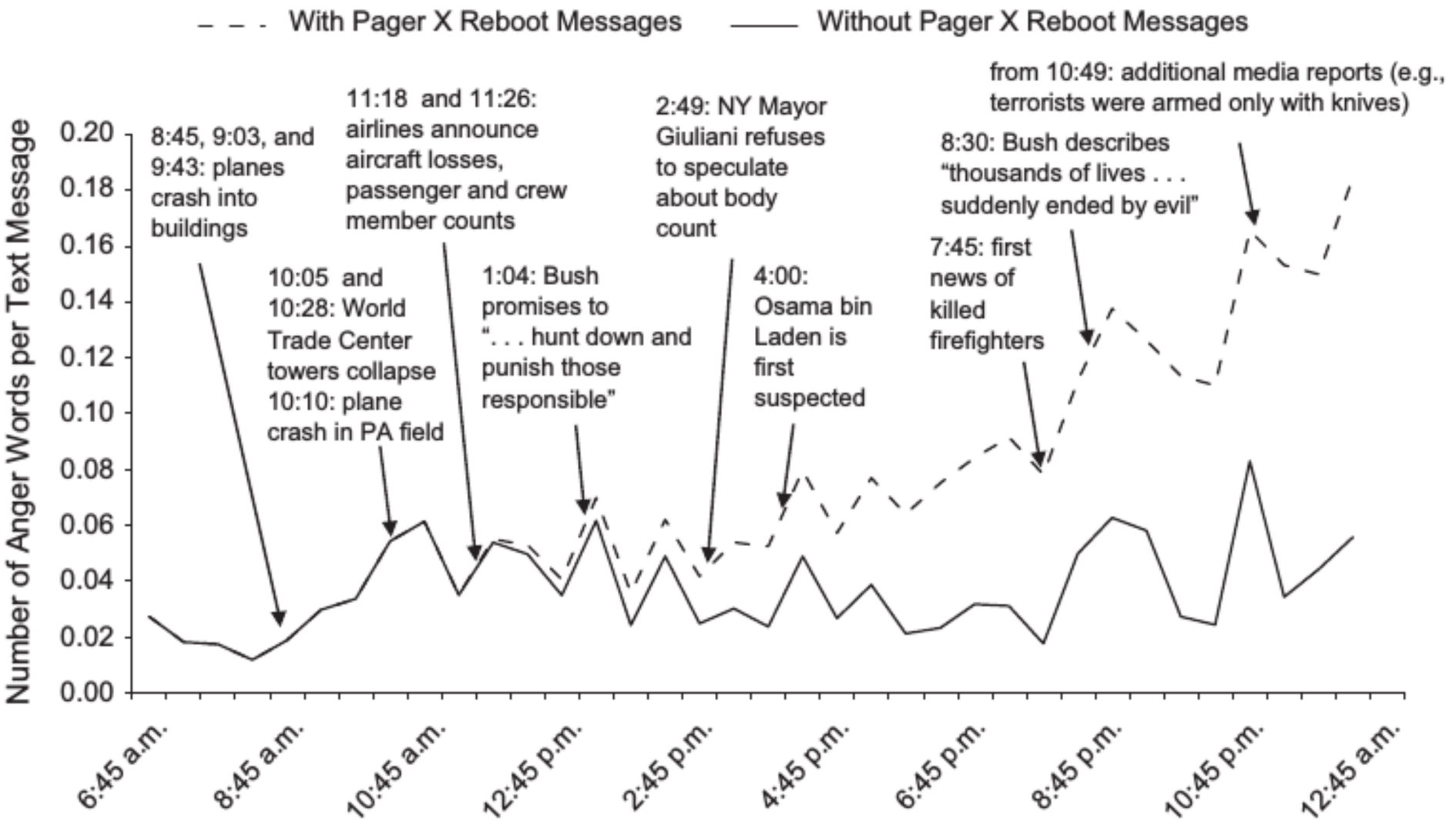
2001-09-12 02:25:16 Arch [0951146] C ALPHA
TX-013 Caddo - No answer.

2001-09-12 02:25:16 Arch [0987275] C ALPHA
s0191: 09/11 13:27:06 Reboot NT machine gblnetnt06 in cabinet 311R at 13/1CMP:CRITICAL:Sep 11 13:27:06

"Reboot NT machine [name] in cabinet [name] at [location]:CRITICAL:[date and time]."

The word "critical" is contained in the anger word list of LIWC!

Anger timeline without REBOOT messages



The issue of machine-generated traces

Not all digital traces are generated by humans, a large volume of data is generated by machines.

During the summer of 2018, Twitter made a big bot cleanse, but independent estimates before reported that between 9% and 15% of Twitter accounts were likely to be bots.

One of the most widely used methods to detect bots on Twitter was Botometer, which was for a long time in constant development by the OSoMe lab at Indiana University (now in archival mode only). Even if you manage clean bots from your data, you should always take a good look at your text. You can make word clouds, word shift graphs, or just browse through it to see if you notice anomalous patterns. To sum up:

> Take home message: Do not just analyze text, also look at it!

Supervised methods in NLP

1. When dictionary methods go wrong

2. Supervised methods in NLP

3. Comparing methods in SMDA

Representations of text

Text representation: computing a fixed-length vector from a text to train a model on the features (values) of the vector

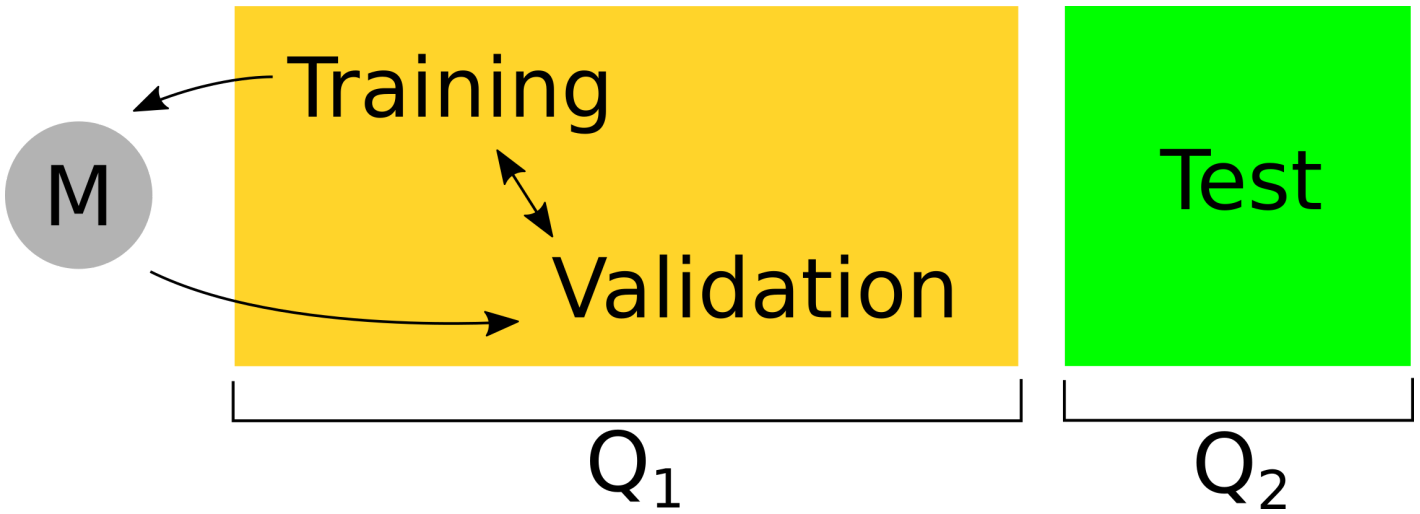
- **TF (Term Frequency) vector:** Using a bag of words assumption, represent each text as a vector of frequencies of the words in the language

- **TF-IDF (Term Frequency - Inverse Document Frequency):** Same but dividing by the total frequency of the term in all documents

    - Both TF and TF-IDF generate very long vectors (number of possible words) with many zeroes
    - Some approaches to remove rare words or uninformative stopwors
    - Requires design decisions: include punctuation? stem words?

- **Denser representations:** More on those next week

How to train your model



1. **Training:** Texts with annotations of sentiment are used to fit the model
2. **Validation:** A second set of annotated texts not used for training are used to evaluate the quality of the model: Q_1
3. **Testing:** One last evaluation run of the fitted model with a leave-out sample of annotated texts. None of these texts should have been part of the validation or training steps. Testing quality Q_2 measures predictive quality of the model.

Be extra careful with test sets

Matthew Leavitt
@leavittron

v excited to finally announce our new work that formalizes one of the most effective practices for training LLMs—something that many industry leaders have conspicuously avoided discussing

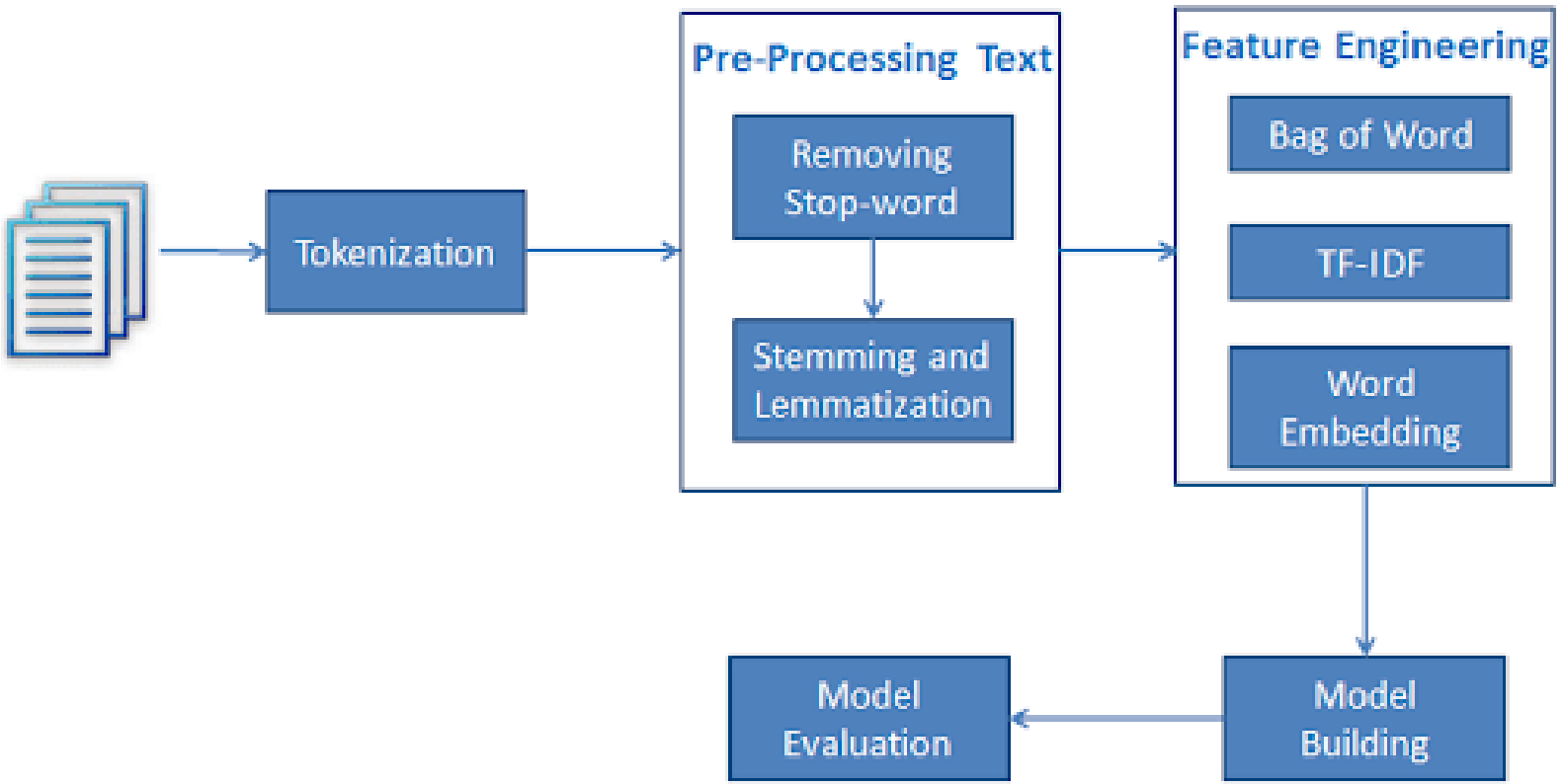
Towards Training on the Test Set

Gold Standards and Ground Truths

pos	neg	text
0	0	I wana see the vid Kyan
0	1	I cant feel my feet.. that cant be good..
1	0	10 absolutely jaw dropping concept car designs http://ow.ly/15OnX
0	0	Phil Collins- You Can’t Hurry Love

- Supervised sentiment analysis needs a set of labeled texts to learn from.
- Labels can come from the author of the text or from reading raters
- The above table is an example from a real dataset with two annotations: a positivity score and a negativity score
- Other ground truths might have numeric scores in a scale or text labels for basic emotional states.

Text preprocessing



Pre-processing from Text Analytics for Beginners using NLTK, Navlani, 2019

What model to use?

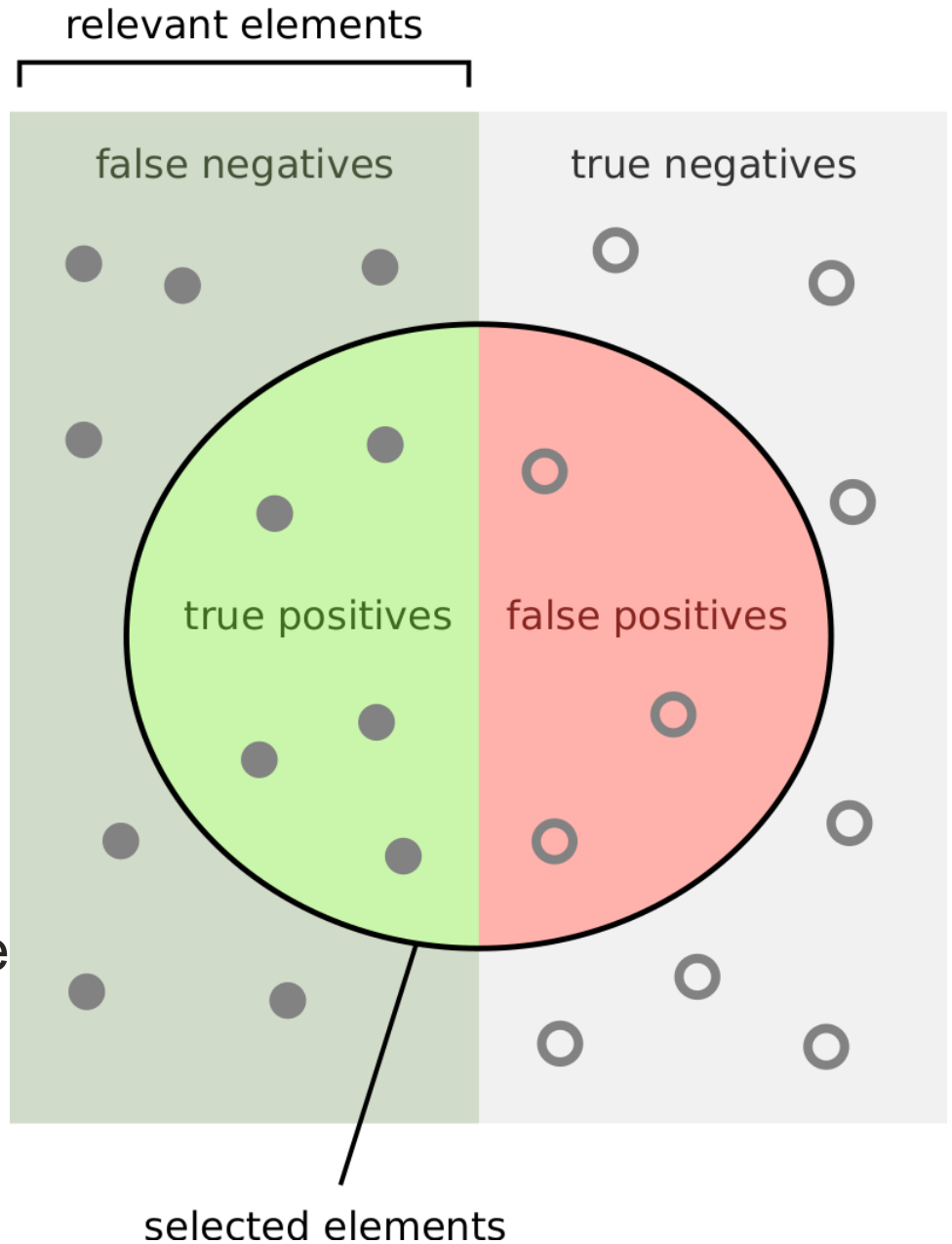
Common approaches are:

1. **Naive Bayes:** Takes features as independent signals and fits the label according to Bayes Rule
2. **Support Vector Machine:** Finds a separator given a (non-)linear combination of features
3. **Random Forest:** Finds hierarchical decision rules that divide the texts in classes

In supervised sentiment analysis, generating the ground truth data is the most critical part and is required to train the model. Producing sufficient annotations from readers or authors can be expensive. Supervised methods are usually not out-of-the-box like unsupervised tools, you would have to fit your own model to a ground truth dataset.

Evaluating classifiers

- True positives TP: All positive cases that are correctly predicted
- False positives FP: All negatives that were wrongly predicted as positive
- True negatives TN: All negative cases that are correctly predicted
- False negatives FN: All positive cases that were incorrectly predicted as negative



Accuracy, Precision, and Recall

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- The measure of precision answers the question "How sure am I of this prediction?"
- The measure of recall answers the question "How many of the things that I'm looking for are found?"

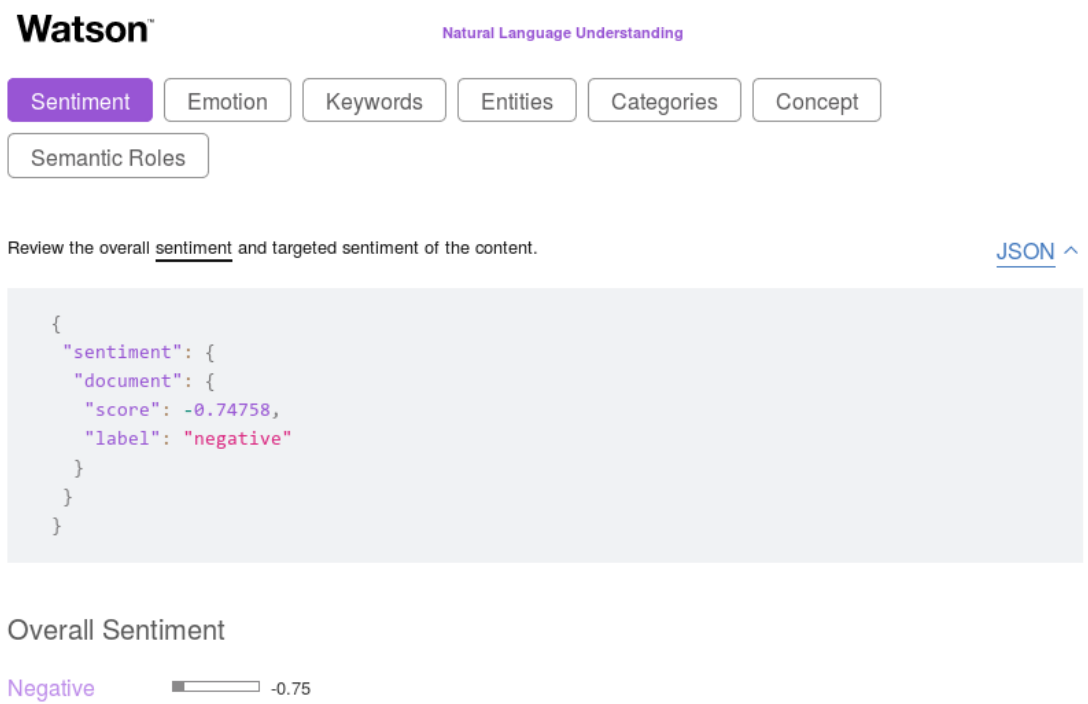
Balancing precision and recall

A way to compute a trade-off between Precision and Recall is the F_1 score, which is a harmonic mean of Precision and Recall:

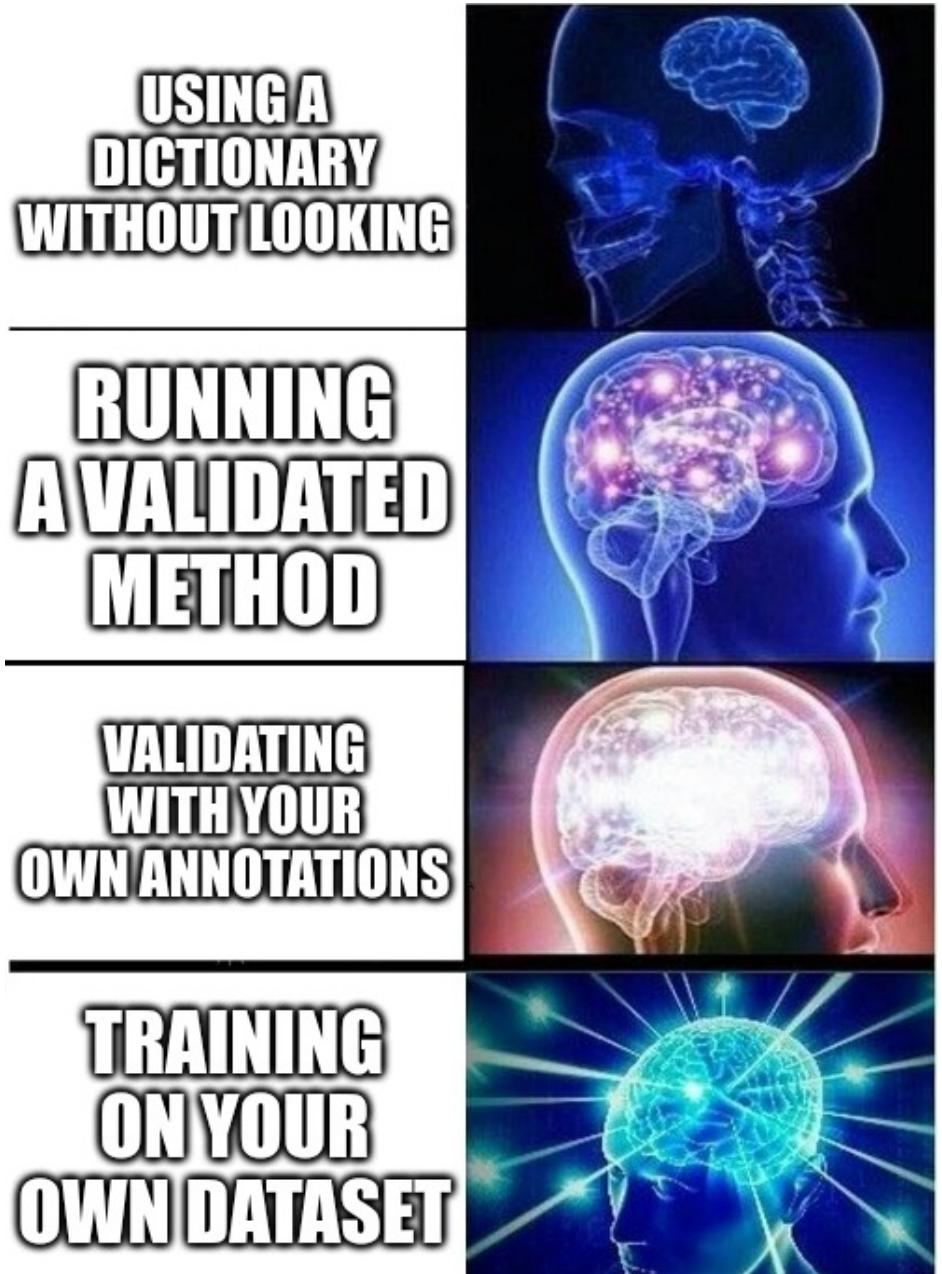
$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F_1 score is often used as a way to summarize the quality of classifiers. When more than one class is possible, you should look at the mean of F_1 over the classess or to the F_1 of each class separately. The F_1 score is often used in sentiment analysis competitions to chose the best tools, for example in the SemEval 2017 competition.

Let someone else do it: Black-box APIs



Easy to use but data and methods unknown. Do your own evaluation!


USING A DICTIONARY WITHOUT LOOKING
RUNNING A VALIDATED METHOD
VALIDATING WITH YOUR OWN ANNOTATIONS
TRAINING ON YOUR OWN DATASET

Comparing methods in SMDA

1. When dictionary methods go wrong
2. Supervised methods in NLP
3. *Comparing methods in SMDA*

Benchmarking sentiment analysis

Dataset	Nomenclature	# Msgs	# Pos	# Neg	# Neu	Average # of phrases	Average # of words	Annotators expertise	# of annotators	CK
Comments (BBC) [11]	*Comments_BBC*	1,000	99	653	248	3.98	64.39	Non expert	3	0.427
Comments (Digg) [11]	*Comments_Digg*	1,077	210	572	295	2.50	33.97	Non expert	3	0.607
Comments (NYT) [15]	*Comments_NYT*	5,190	2,204	2,742	244	1.01	17.76	AMT	20	0.628
Comments (TED) [65]	*Comments_TED*	839	318	409	112	1	16.95	Non expert	6	0.617
Comments (Youtube) [11]	*Comments_YTB*	3,407	1,665	767	975	1.78	17.68	Non expert	3	0.724
Movie Reviews [54]	*Reviews_I*	10,662	5,331	5,331	-	1.15	18.99	User rating	-	0.719
Movie Reviews [15]	*Reviews_II*	10,605	5,242	5,326	37	1.12	19.33	AMT	20	0.555
Myspace posts [11]	*Myspace*	1,041	702	132	207	2.22	21.12	Non expert	3	0.647
Product Reviews [15]	*Amazon*	3,708	2,128	1,482	98	1.03	16.59	AMT	20	0.822
Tweets (debate) [66]	*Tweets_DBT*	3,238	730	1,249	1,259	1.86	14.86	AMT+expert	Undef.	0.419
Tweets (random) [11]	*Tweets_RND_I*	4,242	1,340	949	1,953	1.77	15.81	Non expert	3	0.683
Tweets (random) [15]	*Tweets_RND_II*	4,200	2,897	1,299	4	1.87	14.10	AMT	20	0.800
Tweets (random) [67]	*Tweets_RND_III*	3,771	739	488	2,536	1.54	14.32	AMT	3	0.824
Tweets (random) [68]	*Tweets_RND_IV*	500	139	119	222	1.90	15.44	Expert	Undef.	0.643

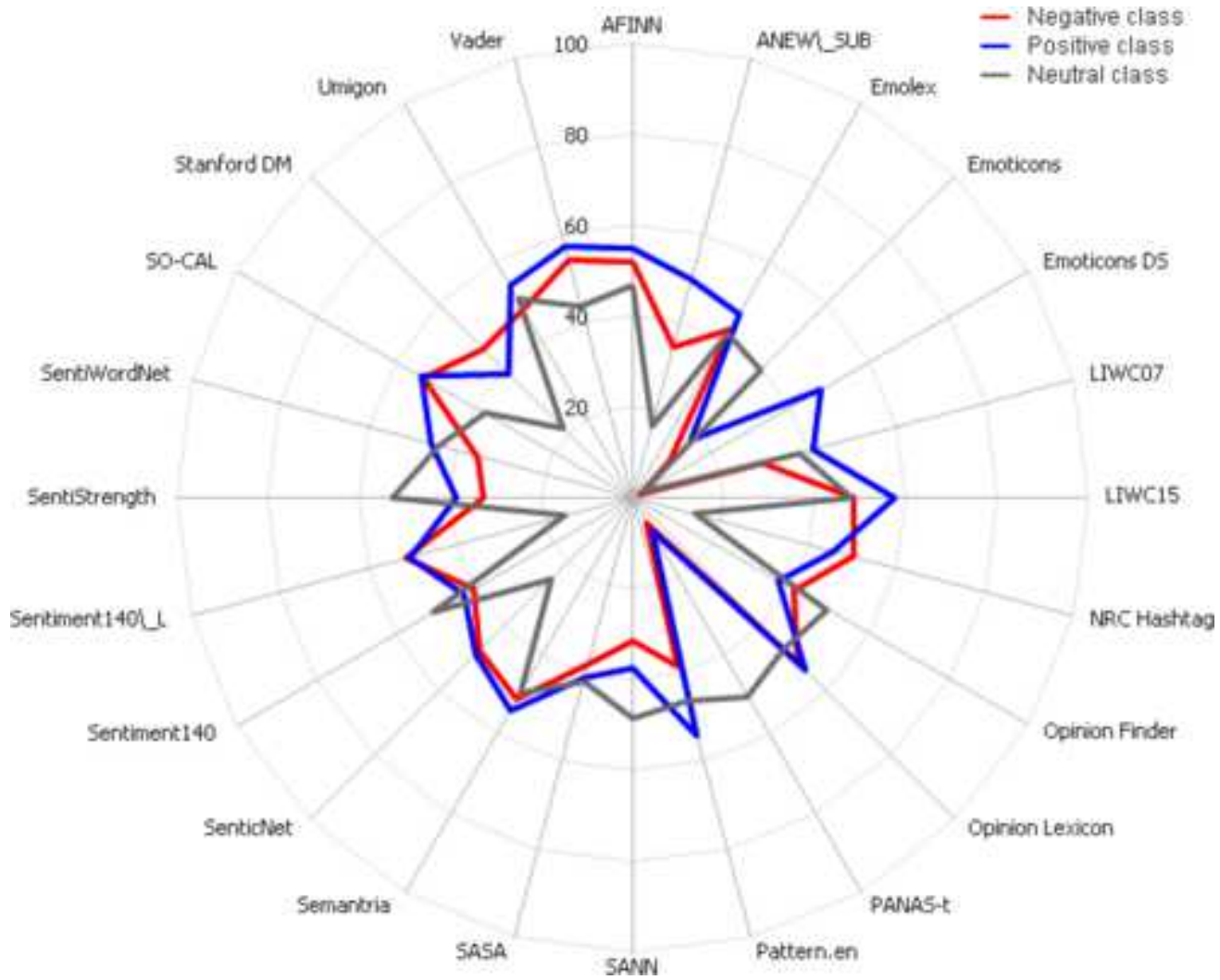
SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods F. Ribeiro, et al. EPJ Data Science (2016)

Benchmark setup

Context groups	
Social Networks	Myspace, Tweets_DBT, Tweets_RND_I, Tweets_RND_ II, Tweets_RND_III, Tweets_RND_IV, Tweets_STF, Tweets_SAN, Tweets_Semeval
Comments	Comments_BBC, Comments_DIGG, Comments_NYT, Comments_ TED, Comments_YTB, RW
Reviews	Reviews_I, Reviews_I, Amazon

- 18 labeled datasets in three groups: Social networks, coomments, reviews
- 24 out-of-the box sentiment analysis methods
  - includes dictionary-based (LIWC, NRC)
  - rule-based (VADER, SentiStrength)
  - some supervised methods (SASA)
- Two tasks:
  - 2-class (positive/negative), given that it is not neutral
  - 3-class (positive/negative/neutral)
- Evaluation based on F_1 score per class - summary as mean rank of methods

Mean F_1 across datasets

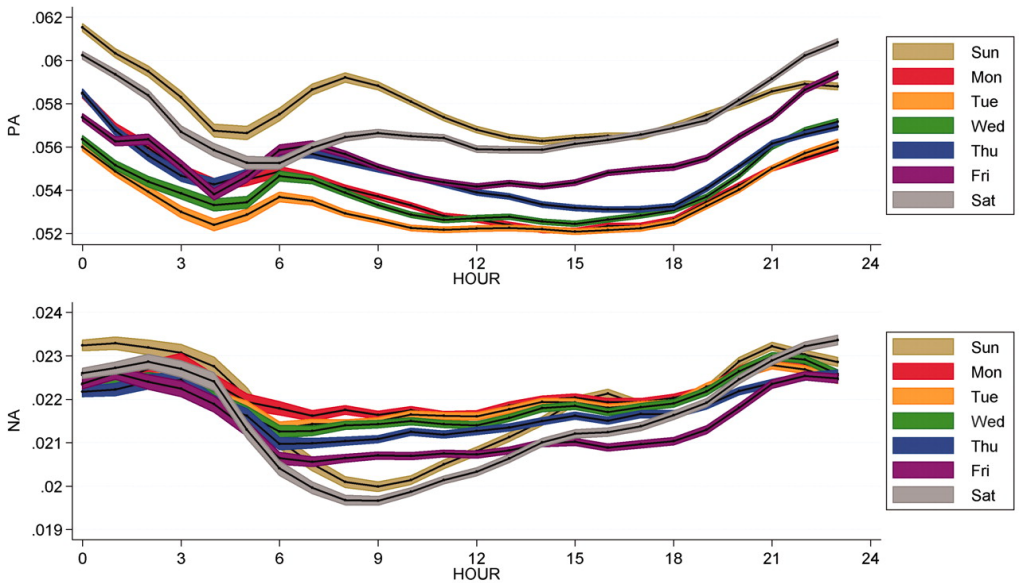


Ranking for social network data

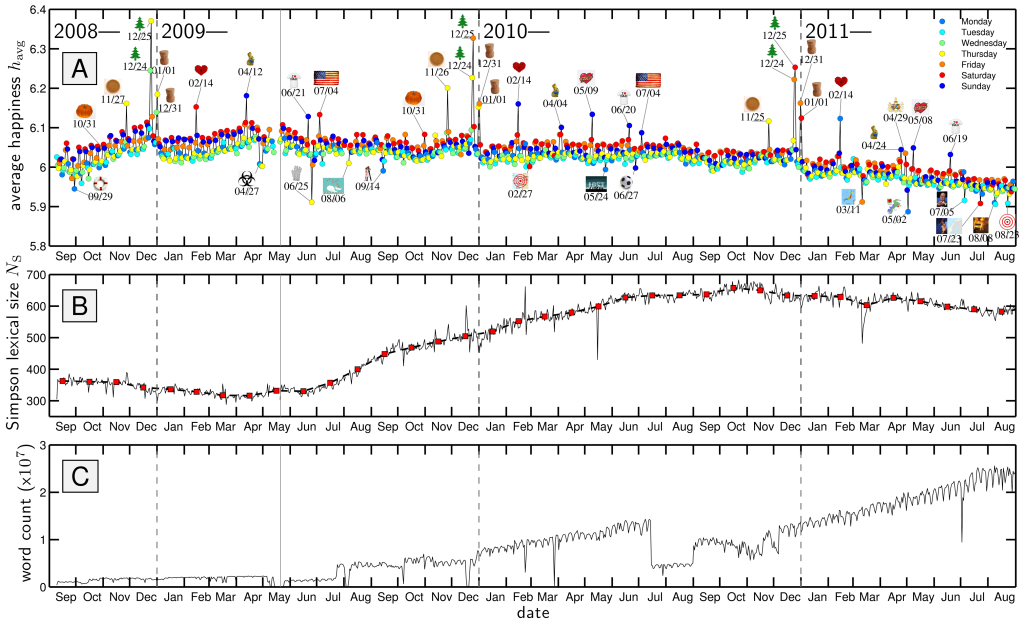
Table 13 Mean rank table for datasets of social networks

3-classes			2-classes			
Pos	**Method**	**Mean Rank**	**Pos**	**Method**	**Mean Rank**	**Coverage (%)**
1	Umigon	2.57	1	SentiStrength	2.22 (2.57)	31.54 (32.18)
2	LIWC15	3.29	2	Sentiment140	3.00	46.98
3	VADER	4.57 (4.57)	3	Emoticons	5.11	18.04
4	AFINN	5.00	4	LIWC15	5.67	71.73
5	Opinion Lexicon	5.57	5	Semantria	5.89	61.98
6	Semantria	6.00	6	PANAS-t	6.33	5.87
7	Sentiment140	7.00	7	Opinion Lexicon	7.56	66.56
8	Pattern.en	7.57	8	Umigon	8.00	71.67
9	SO-CAL	9.00	9	AFINN	8.67	73.37
10	Emolex	12.29	10	SO-CAL	8.78	67.81

Validity of using text analysis at scale?

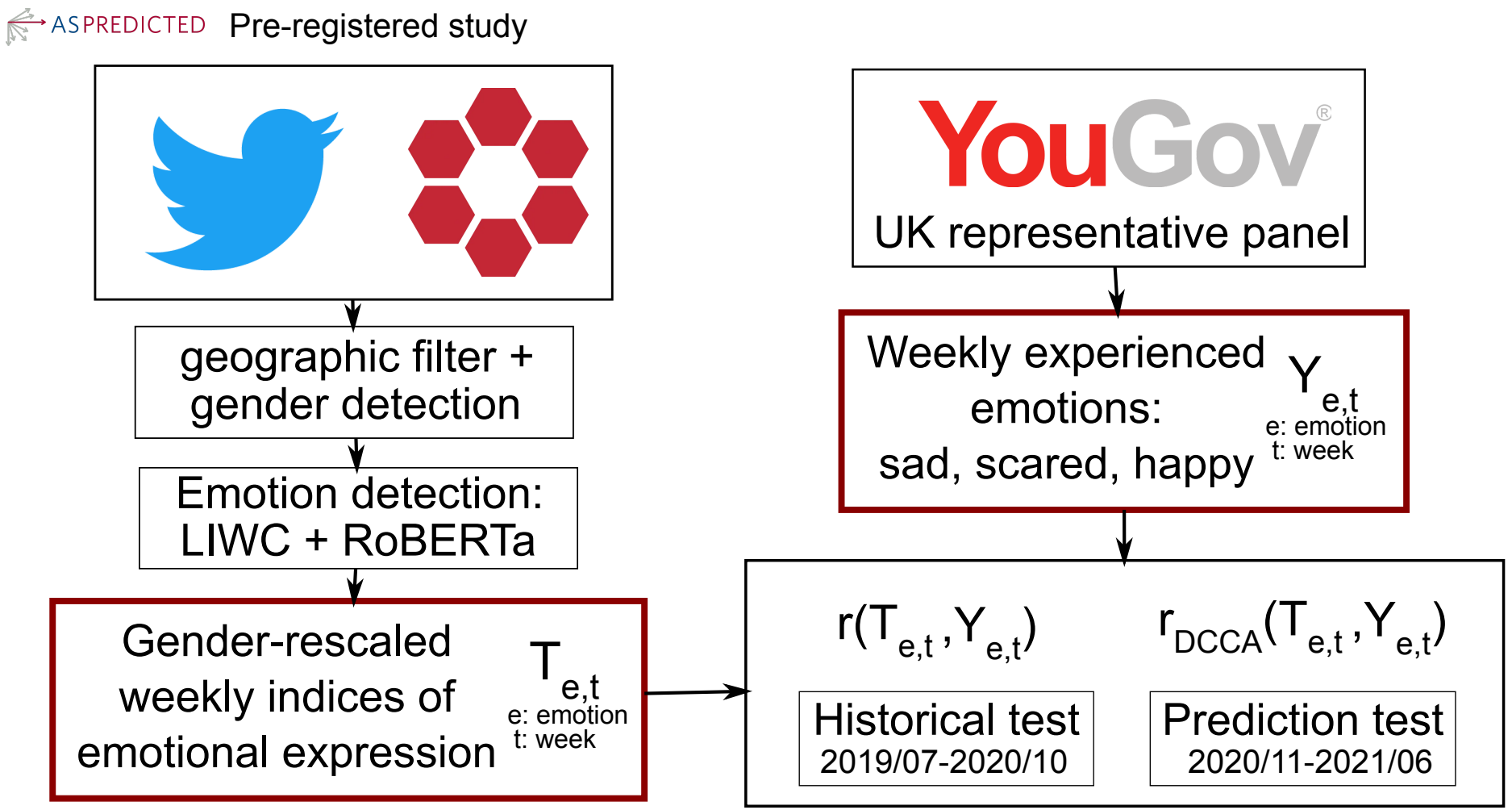


Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Golder & Macy, Science (2011)



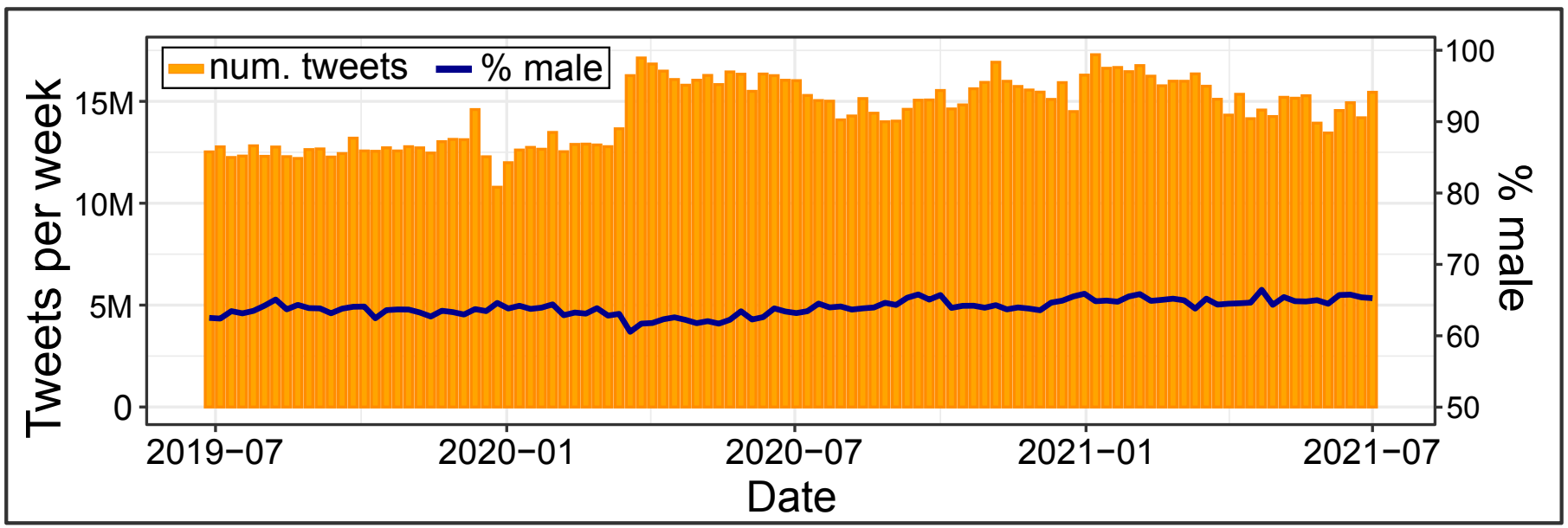
Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. Dodds et al. PLoS One (2011)

Validating a UK emotion macroscope

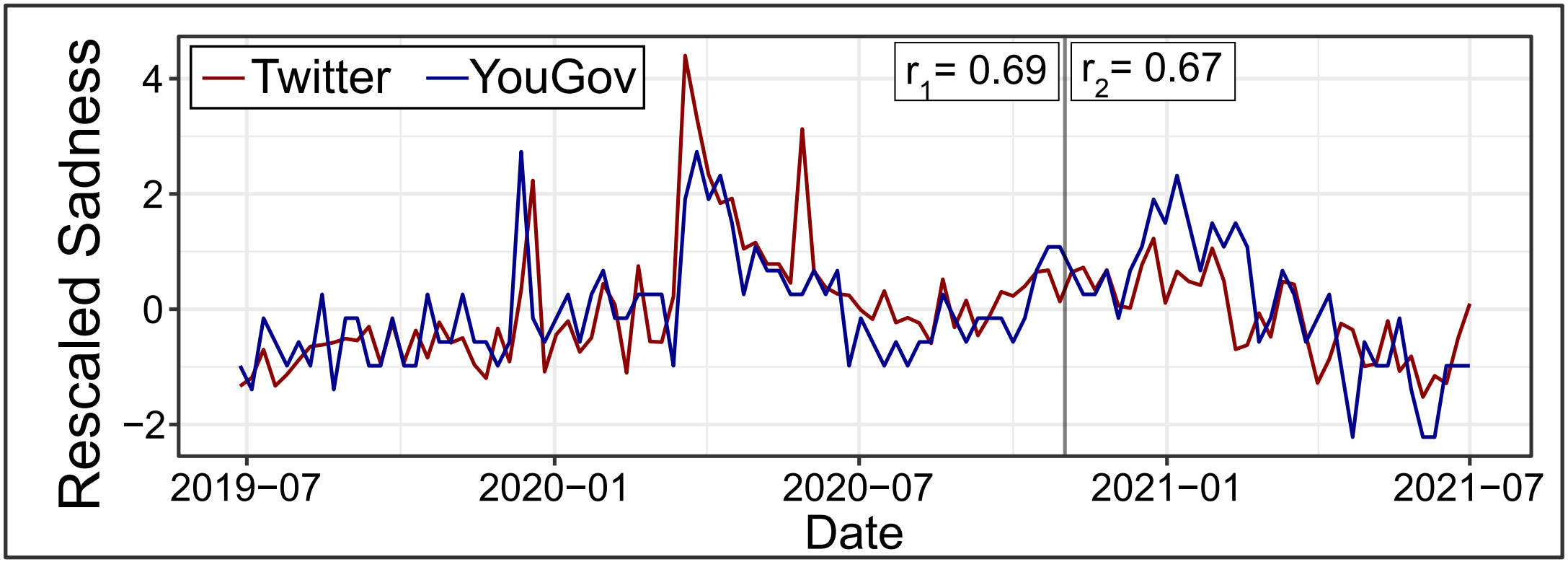


Validating a UK emotion macroscope

1. Two years of weekly representative UK emotion survey by YouGov
2. UK Twitter data for the same period: 1.5 Billion tweets (without RT)
3. Text analysis: dictionary-based (LIWC) and supervised (RoBERTa)
4. Gender detection of twitter users based on profile
5. Gender-rescaled time series of emotional expression

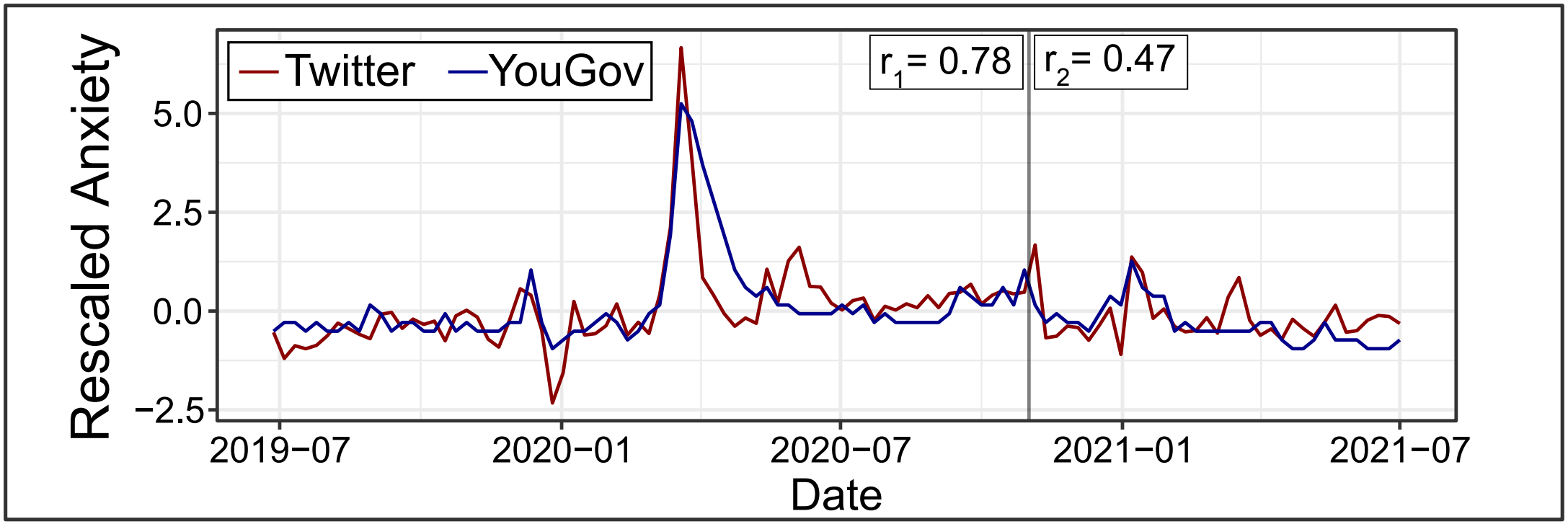


Sadness in Twitter and YouGov



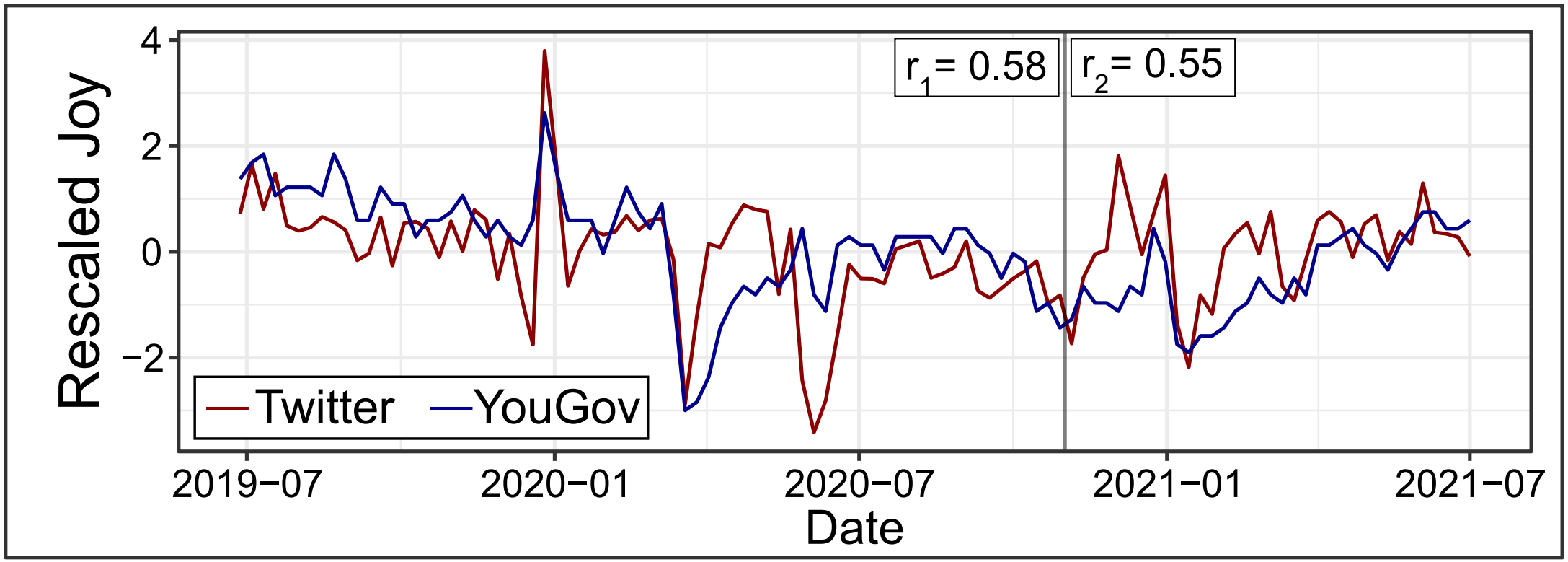
- Similar results with dictionary-based and supervised (r~0.65)

Anxiety in Twitter and YouGov



- Better results with dictionary-based method and with gender rescaling
- Results robust to autocorrelation and heteroskedasticity

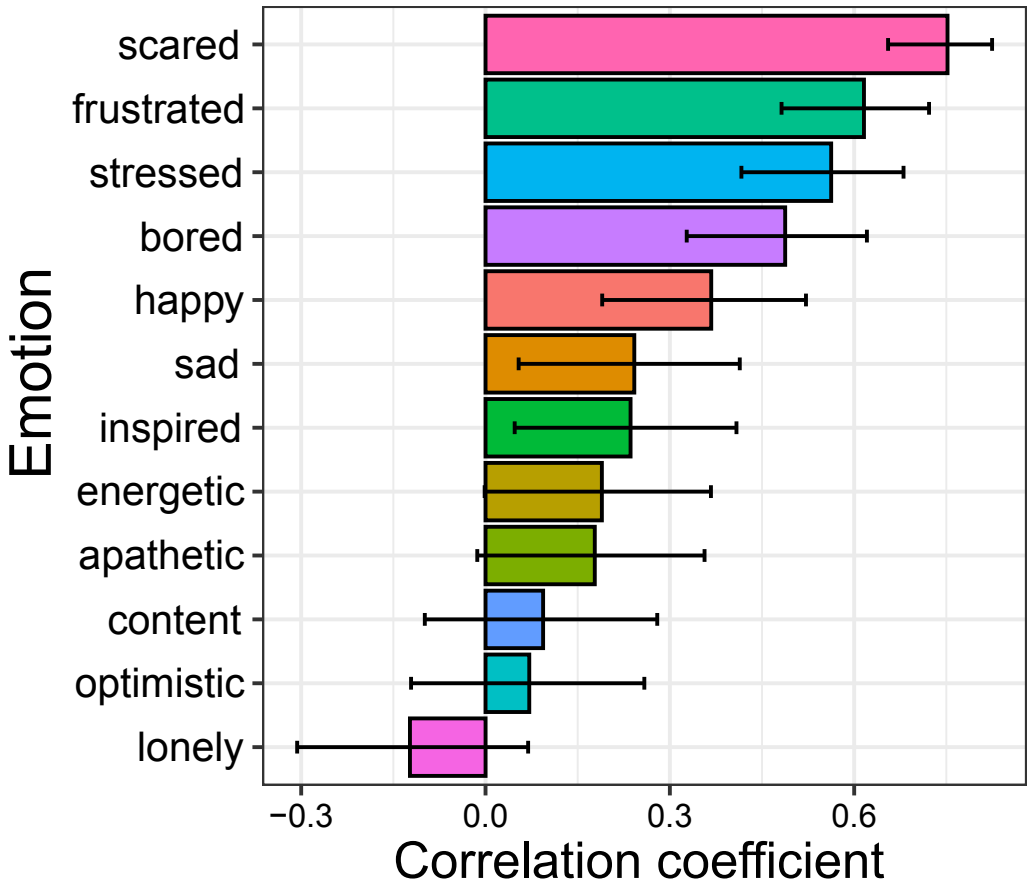
Joy in Twitter and YouGov



- Substantially better results with supervised method than dictionary-based

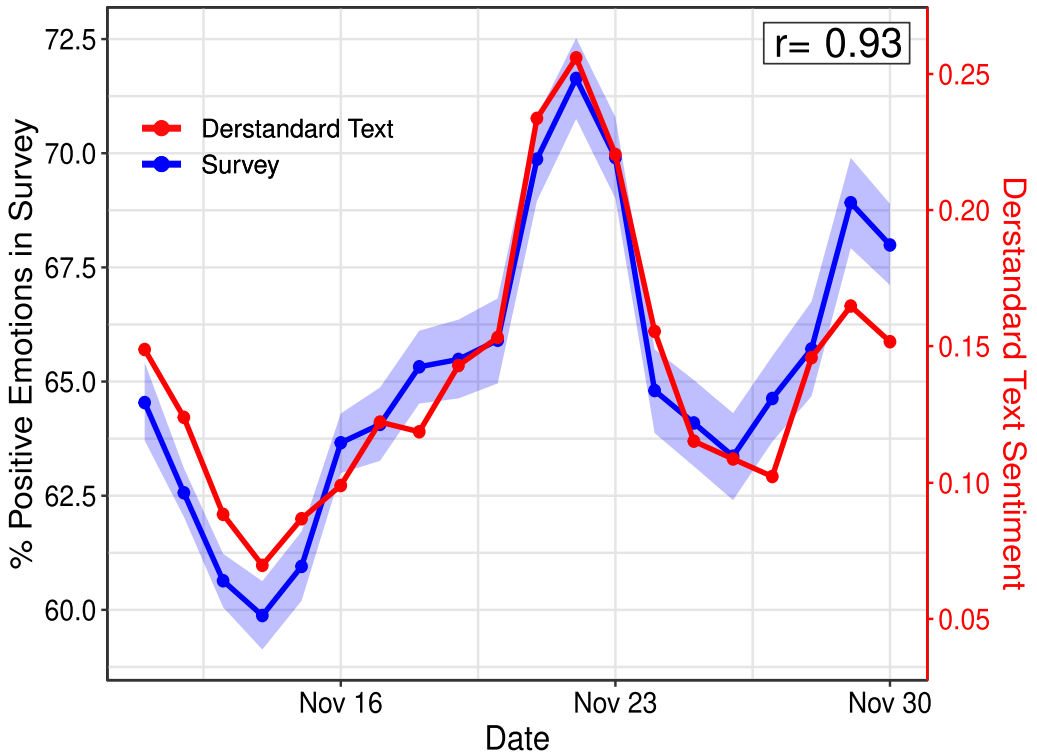
Exploring 12 emotional states

- Time series of number sentences like "I am [emotion]" on Twitter
- Weak correlations happen for infrequent emotions in text
- Comparison: US weekly pre-election polls correlate with 0.66
- Arxiv preprint at https://arxiv.org/abs/2107.13236

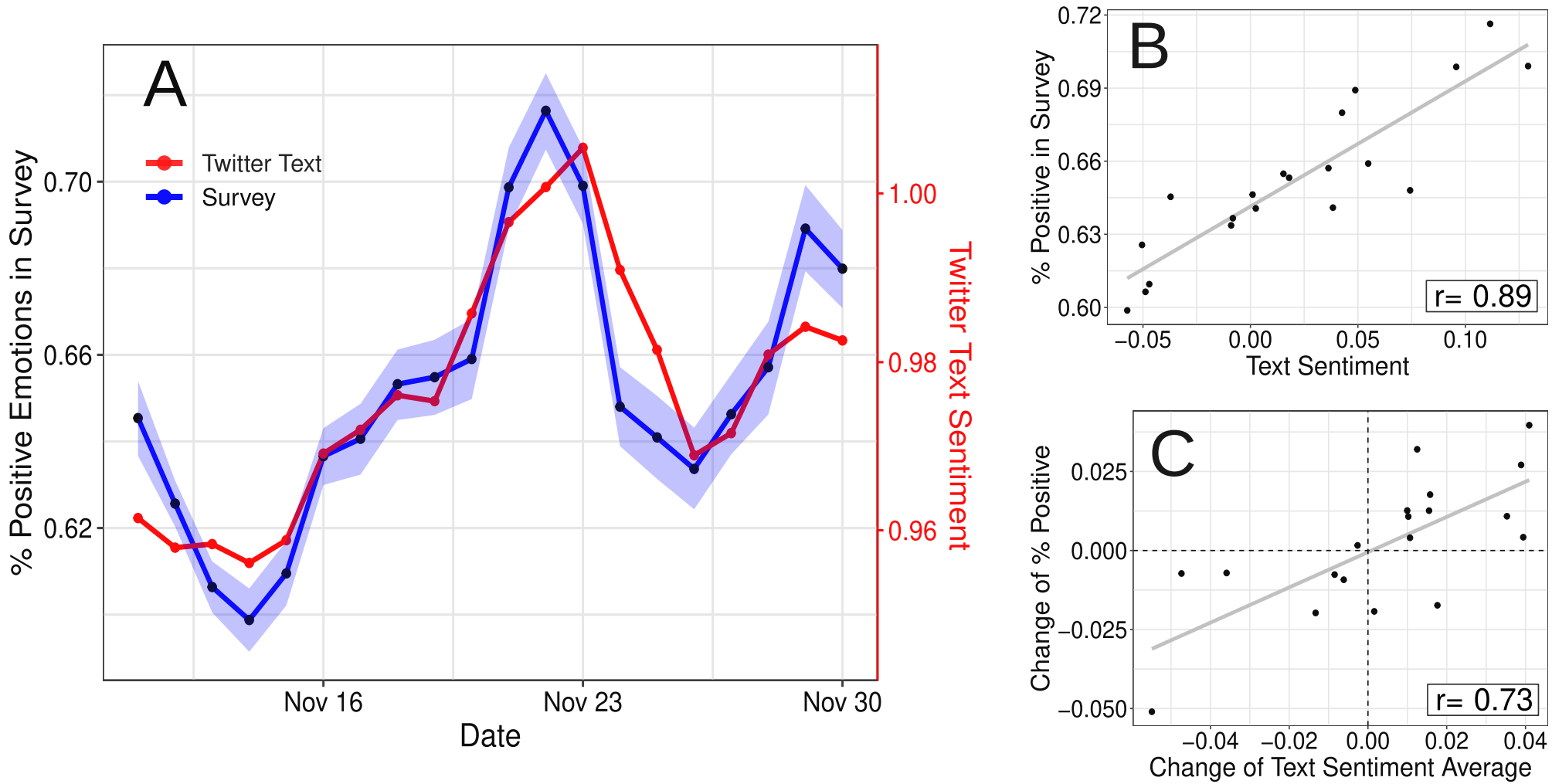


Study 2: Validating an Austrian macroscope

- 20-day emotion survey in derstandard.at (N=268,128)
- Daily frequency, 3-day windows
- Text from Der Standard forum (N=452,013)
- Austrian tweets (N=515,187) filtered as UK macroscope
- Compared dictionary-based (LIWC) and supervised model (GS)

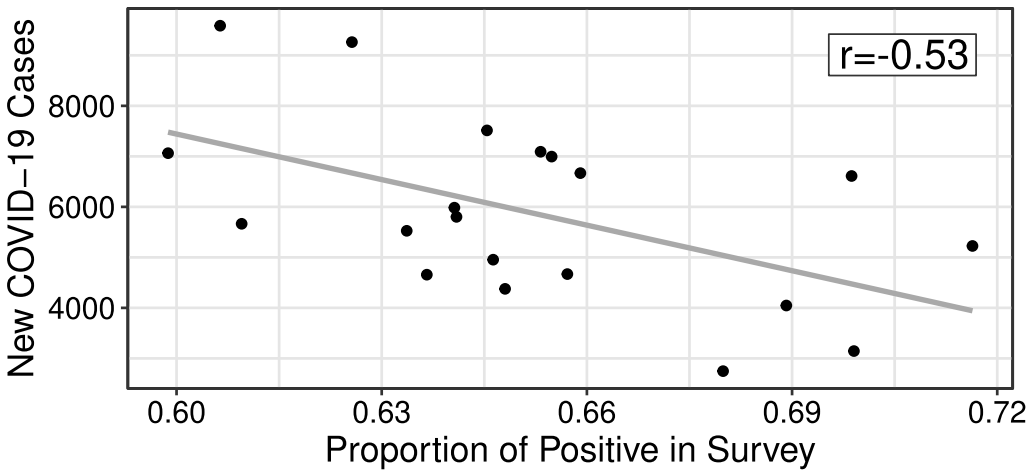


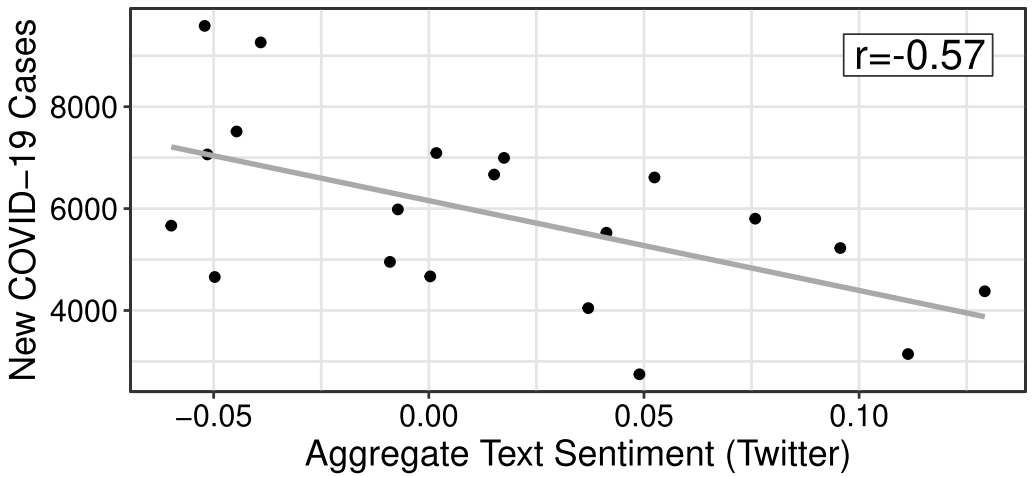
Twitter sentiment and Der Standard survey



Validating daily social media macroscopes of emotions. Max Pellert, Hannah Metzler, Michael Matzenberger, David Garcia. Scientific Reports (2022)

Correlations with new COVID-19 cases





- Do correlations attenuate due to additional social media measurement error?
- Survey emotion correlation with new cases as strong as Twitter sentiment
- Errors sources might be different: Need for conceptual validations

Summary

- When dictionary methods go wrong
  - The case of LIWC on pagers after 9/11
  - Reading the data matters! Visualizations, random samples, etc

- Supervised methods
  - Gold standards as annotated texts
  - The importance of the separation of training and test
  - Evaluation: Precision, Recall, and F_1

- Comparing and evaluation methods
  - SentiBench: a 2016 benchmark to compare methods
  - Validations beyond documents: time series in the UK and Austria