Automated Analysis of Online Behaviour on Social Media

Erik Tjong Kim Sang

Brown bag talk, 3 February 2020

netherlands
eScience center

by SURF & NWO

Netherlands eScience Center

Mission: enable digitally enhanced research

Methods:

- **Cooperate with researchers from all disciplines**
- **Develop research software**
- **Share ideas and tools**

Automated Analysis of Online Behaviour on Social Media (2017)

Goal: study the behavior of politicians on social media

Means: tens of thousands of manually annotated tweets

Problem: new tweets need annotation, which requires a lot of work

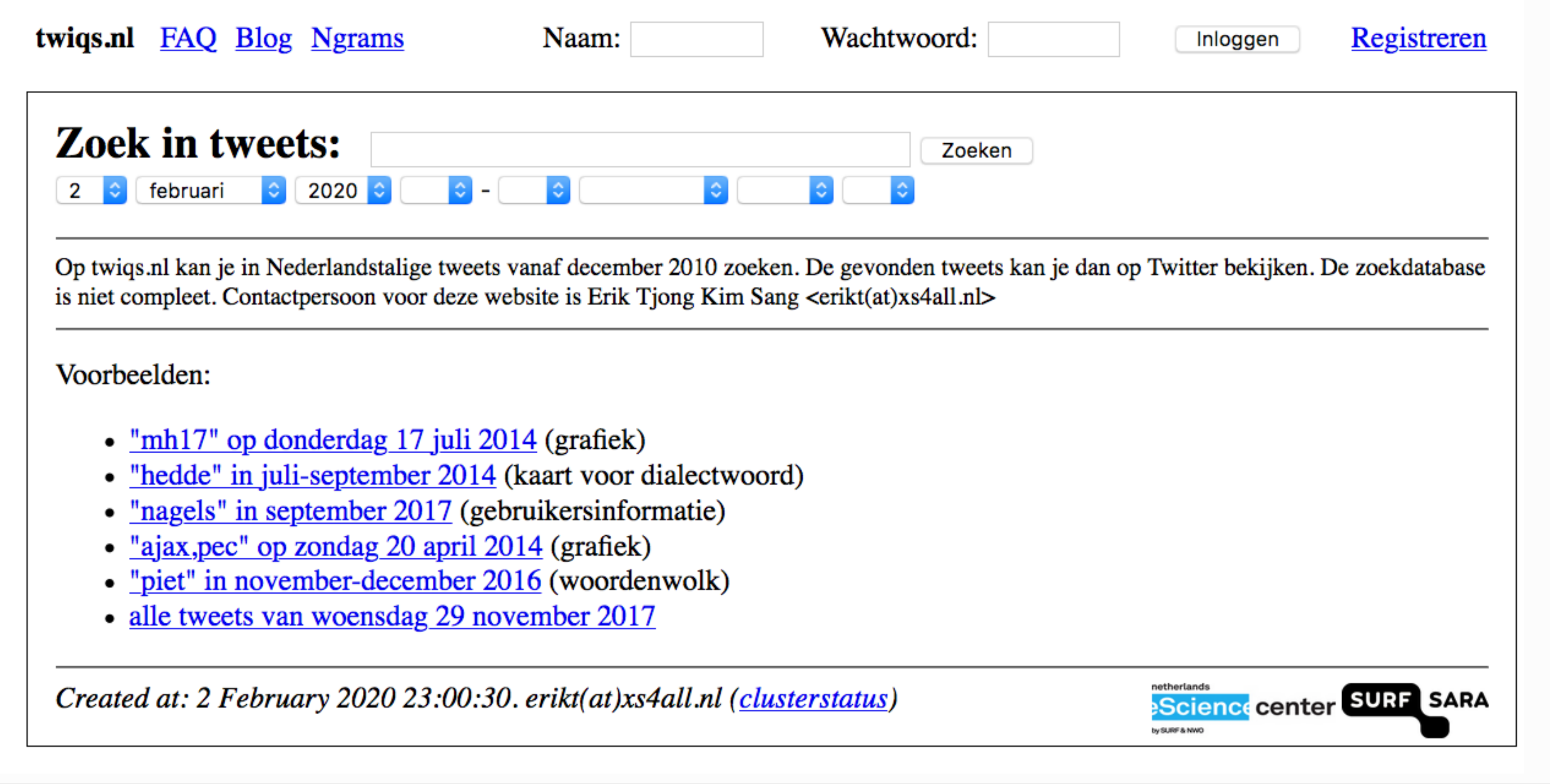
Social Media Analysis

We started collecting Dutch tweets in 2010

Our collection contains almost 4 billion tweets

We performed several tasks with the data

Project TwiNL (2012-2013)



Zoek in tweets:

2 februari 2020 -

Op twiqs.nl kan je in Nederlandstalige tweets vanaf december 2010 zoeken. De gevonden tweets kan je dan op Twitter bekijken. De zoekdatabase is niet compleet. Contactpersoon voor deze website is Erik Tjong Kim Sang <erikt(at)xs4all.nl>

Voorbeelden:

- "mh17" op donderdag 17 juli 2014 (grafiek)
- "hedde" in juli-september 2014 (kaart voor dialectwoord)
- "nagels" in september 2017 (gebruikersinformatie)
- "ajax,pec" op zondag 20 april 2014 (grafiek)
- "piet" in november-december 2016 (woordenwolk)
- alle tweets van woensdag 29 november 2017

Created at: 2 February 2020 23:00:30. erikt(at)xs4all.nl (clusterstatus)

Election prediction (2011)

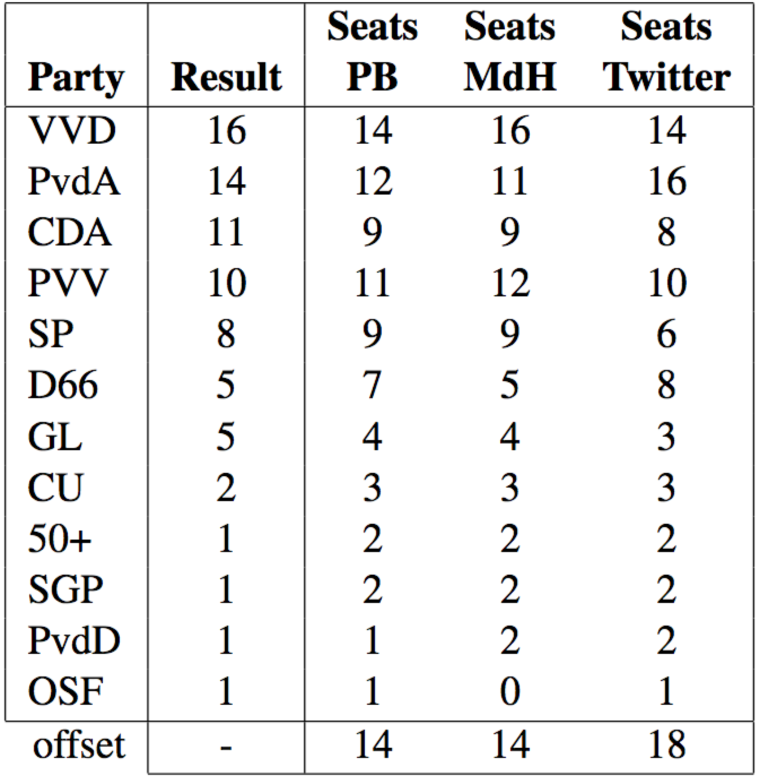
Predict Dutch Senate elections based on tweets

Challenges:

- **Small data sets (after filtering)**
- **Data contain positive and negative party mentions**

Paper: Tjong Kim Sang & Bos (2012)

Results



Party	Result	Seats PB	Seats MdH	Seats Twitter
VVD	16	14	16	14
PvdA	14	12	11	16
CDA	11	9	9	8
PVV	10	11	12	10
SP	8	9	9	6
D66	5	7	5	8
GL	5	4	4	3
CU	2	3	3	3
50+	1	2	2	2
SGP	1	2	2	2
PvdD	1	1	2	2
OSF	1	1	0	1
offset	-	14	14	18

PB: Politieke Barometer (poll)
MdH: Maurice de Hond (poll)
Twitter: our predictions

Take-away messages

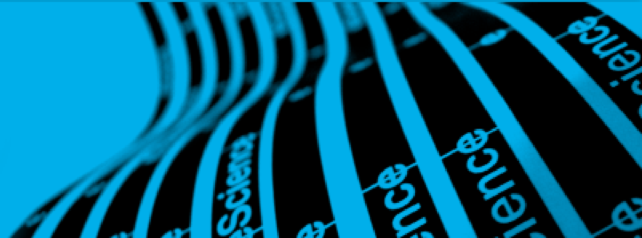
Focusing on certain topics and certain topics leaves little data to analyze

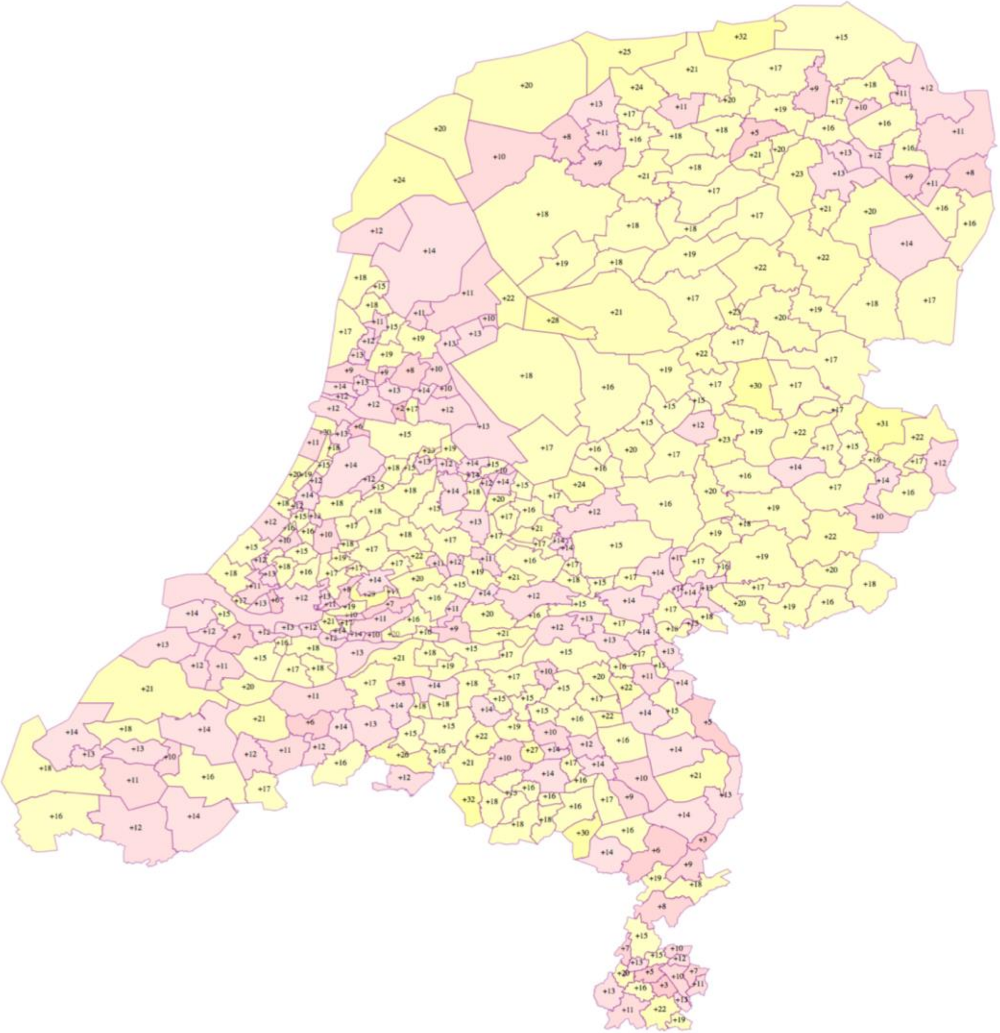
Twitter users are not a good representation for Dutch (or any other) society

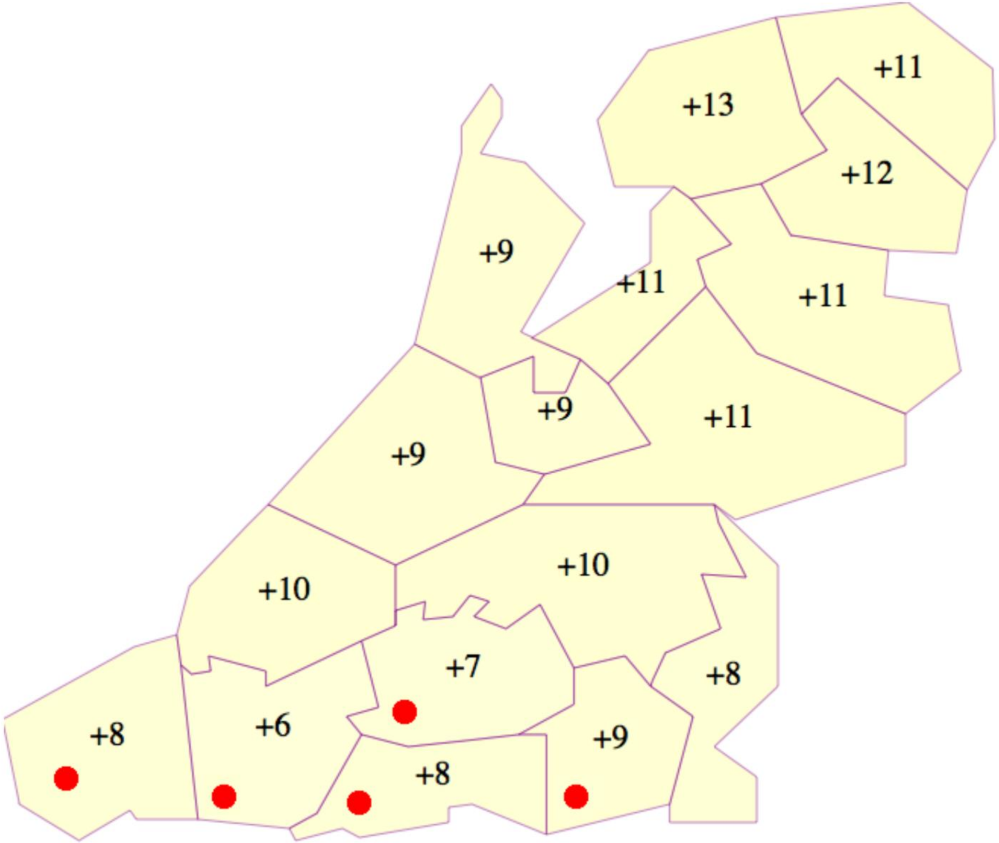
Happiness estimation

Can we estimate the happiness of people in geographical regions?

Yes, by applying automatic sentiment analysis to tweets attached to locations

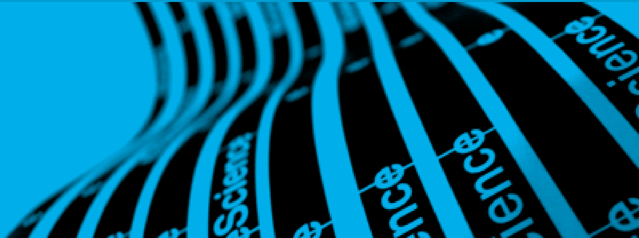





+13
+11
+12
+9
+11
+11
+9
+11
+9
+10
+10
+7
+8
+8
+6
+9
+8

Take-away message

Know your methods very well: they might affect the difference that you will measure



Automated Analysis of Online Behaviour on Social Media (2017)

Goal: study the behavior of politicians on social media

Means: tens of thousands of manually annotated tweets

Problem: new tweets need annotation, which requires a lot of work

SARA
netherlands
eScience center
egi

1. Research Questions

RQ1: Which discursive practices do politicians and journalists use on Twitter and how do these change?
RQ2: To what extent do institutional differences between agents still matter, or event exist, now they all have the power to publish on social media?

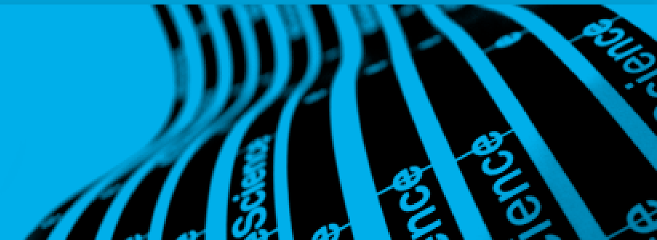
Hypothesis: Networked communication is blurring the distinctive but interdependent role of journalists and politicians. Now that they can both broadcast information, the online behavior of politicians and journalists converges.

Online behaviour: project planning

- We have 55,000 labeled political tweets
- We have 250,000 unlabeled tweets
- Train a machine learner to accurately classify the unlabeled tweets
- Target performance: 67% accuracy (= human performance)

Labels: intention of political tweets

- **CampaignTrail: Tomorrow we'll make a difference: at 16:00 @mariannethieme @PartijvdDieren is a guest in #zeelandkiest http://t.co/D0SoxDkZ @omroepzeeland**
- **Critique: You don't need to be an economist to understand that we need to get rid of mortgage interest tax deduction. #EenVandaag**
- **News: Roemer will not give the right a majority – VK Dossier Elections of 2012 – VK http://t.co/Zg7YZgWn #SP**

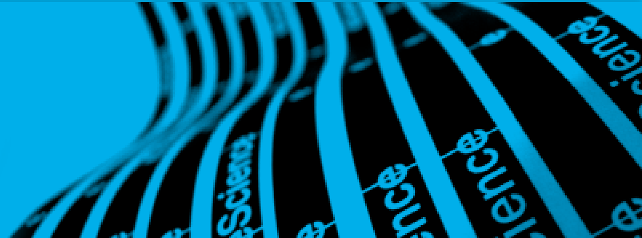


Approach

- Use a standard method for text classification: fastText

- Build language models from the unlabeled data

- Look for methods for selecting additional data for labeling (active learning)



Active learning

For machine learning we often have a few labeled data next to a large amount of unlabeled data

More labeled data leads to better models

Active learning involves selecting the best unlabeled data for future labeling

Data selection in active learning

Two approaches for finding the best unlabeled data:

1. Build a data model which can estimate the difficulty of assigning a label to data (uncertainty sampling)
2. Build different models and check how often they disagree on the new labels (query-by-committee)

The next step is to select the most difficult data items

Selecting difficult data items

There are different methods for selecting difficult data items, depending on the method used:

Uncertainty sampling:

- data items with labels with the lowest scores
- data items with labels with close competing scores

Query-by-committee:

- data items with high entropy of assigned labels

Banko & Brill (2001): Task

Confusable disambiguation involves finding out which variant of a set of easily confusable words fits in a sentence

Examples with { to , too , two } :

- This is version TO of "TO sad TO tell you"
- TO schools set TO close for being TO small
- Congrats TO John for his TO awards TO

Banko & Brill: Active learning

This study uses ten models for determining initial labels for the unlabeled data

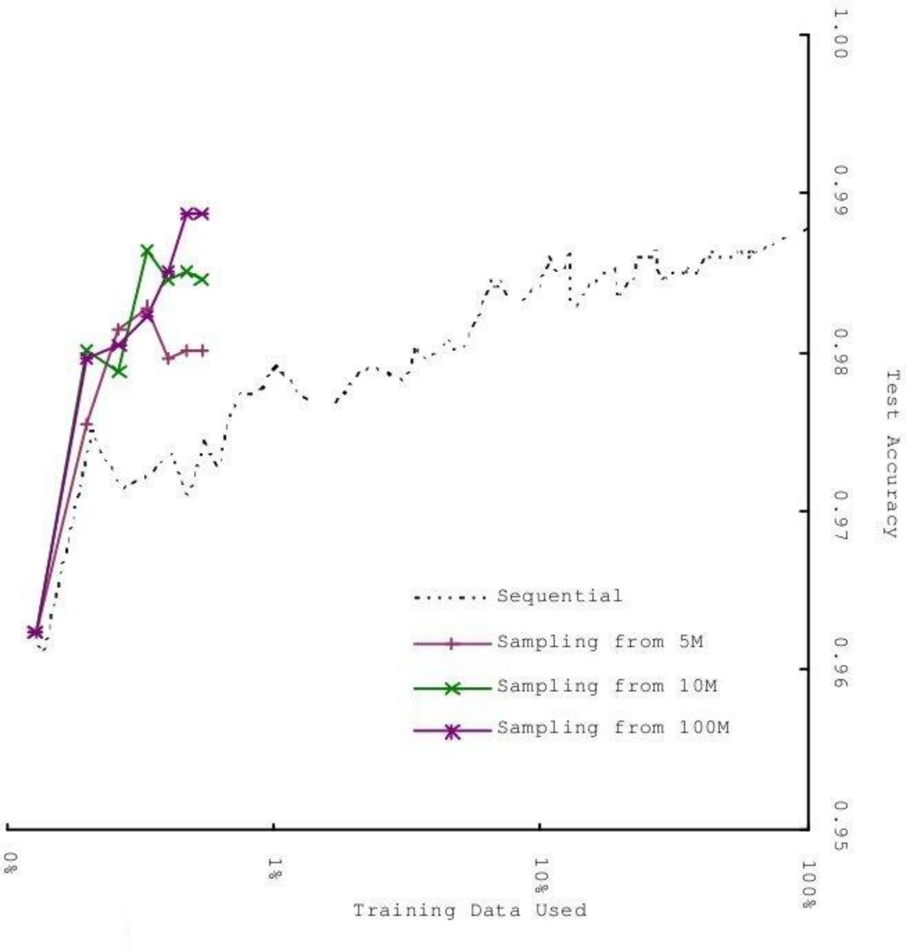
The difficulty of a data items was estimated by the entropy (H) of its labels:

H(A A A A A A A A A A) = 0

H(A A A A A A A A A B) = 0.47

H(A A A A A B B B B B) = 1

Banko & Brill: Results



Our approach

We compared four different active learning methods:

1. Lowest confidence score
2. Highest competing confidence score
3. Entropy of label distribution
4. Longest text

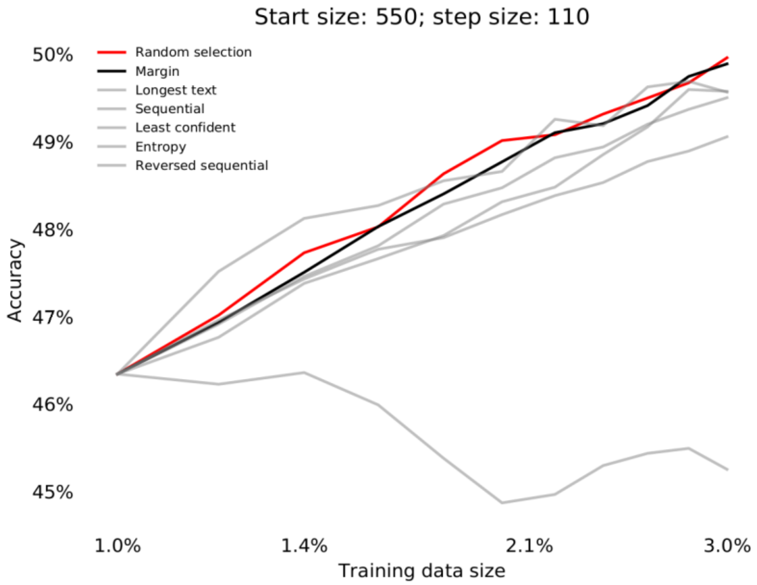
And three baseline methods, e.g. random selection

Machine learning method

We used fastText in all experiments

FastText is developed by Facebook for text classification

TKS: Results



Train size	Accuracy	Method
80.0%	55.6±0.3%	All training data
3.0%	50.0±0.9%	Random selection
3.0%	49.9±0.9%	Margin
3.0%	49.6±0.7%	Longest text
3.0%	49.6±0.9%	Sequential
3.0%	49.5±1.0%	Least confident
3.0%	49.1±0.8%	Entropy
3.0%	45.3±1.3%	Revrsd sequential
1.0%	46.3±0.8%	Baseline

Result

No smart method outperforms the baseline methods

Recent developments

New data has been annotated with two methods

Data with two different labels has been annotated

Next: determine effects to additional data on performance