

Joint Rumour Stance and Veracity

Anders Edelbo Lillie, Emil Refsgaard Middelboe, Leon Derczynski
ITU Copenhagen

This research is mostly based on Danish language data, and slightly on English and German. #benderrule

IT UNIVERSITY OF COPENHAGEN

Let's talk about rumours

- An Oregon mother was arrested after a dog attacked her and ate her.
- The “correct spelling” of the term “happy wedding” is “smiling family”.
- People with autism commonly have difficulties moving fingers, toes, palms and forefinger because of a deficiency of retinonic acid
- Nordstrom has discontinued its popular ‘Peanut Butter Snub Pie’.
- The United Nations said that God made humans immortal.
- A sign in Hawaii warns prospective bride-swappers that a baby bride will appear in a haunted house attraction.
- Kale mask could finally make your face attractive.

Let's talk about rumours

- An Oregon mother was arrested after a dog attacked her and ate her.
- The “correct spelling” of the term “happy wedding” is “smiling family”.
- People with autism commonly have difficulties moving fingers, toes, palms and forefinger because of a deficiency of retinonic acid
- Nordstrom has discontinued its popular ‘Peanut Butter Snub Pie’.
- The United Nations said that God made humans immortal.
- A sign in Hawaii warns prospective bride-swappers that a baby bride will appear in a haunted house attraction.
- Kale mask could finally make your face attractive.



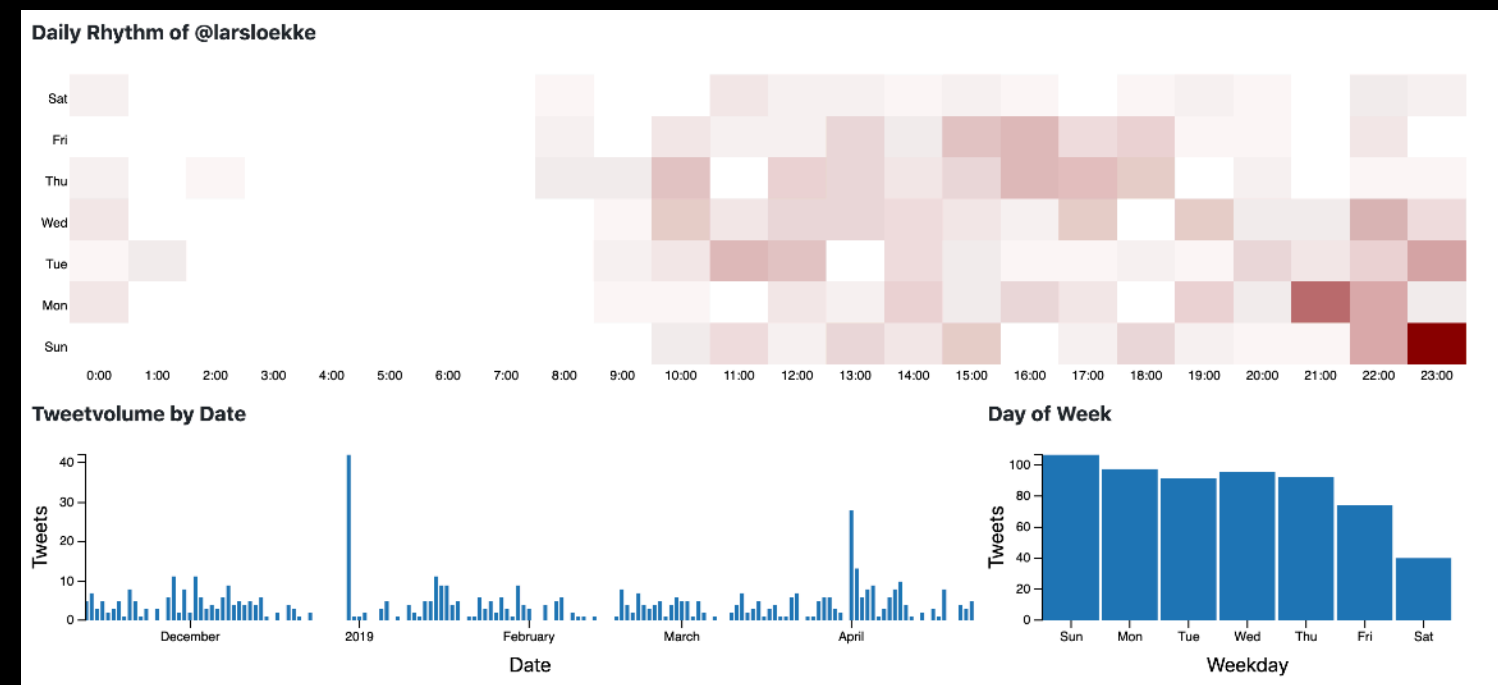
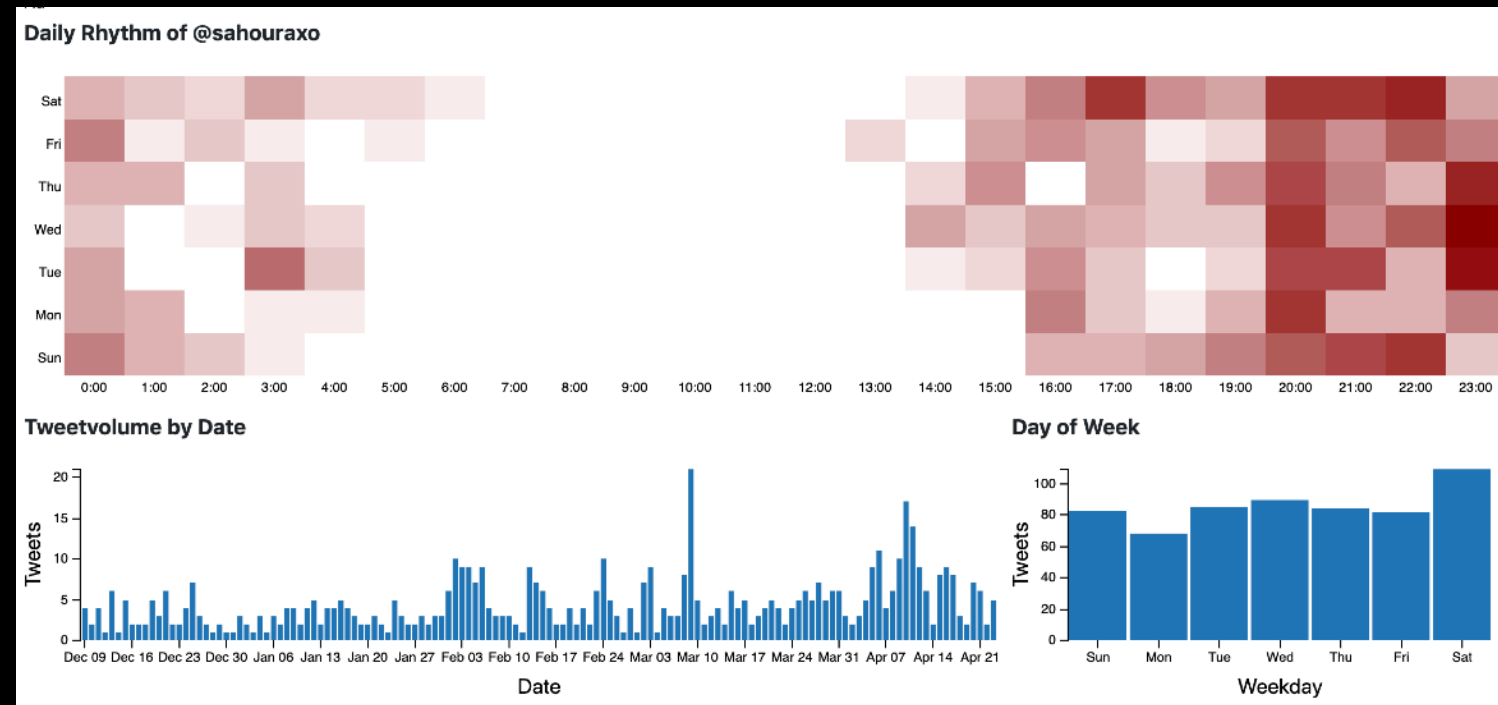
Generate automatically - using GPT2 model

Also trivial to generate article: workload imbalance for checkers

How can we detect misinformation?

- Account behaviour
- Network
- Verifying what it says
- Reactions to claims: stance detection

- Timeframes may be fixed
- The top account claims to be a Lebanese journalist in Israel
- The bottom account is a broad-appeal Danish politician (ex-?)
- The time they tweet, tells us who they are trying to reach



Amplified by the same route

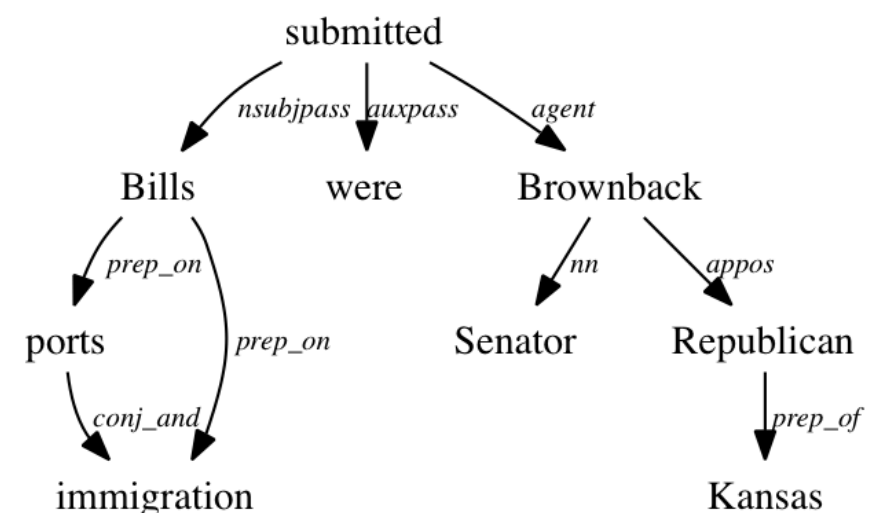
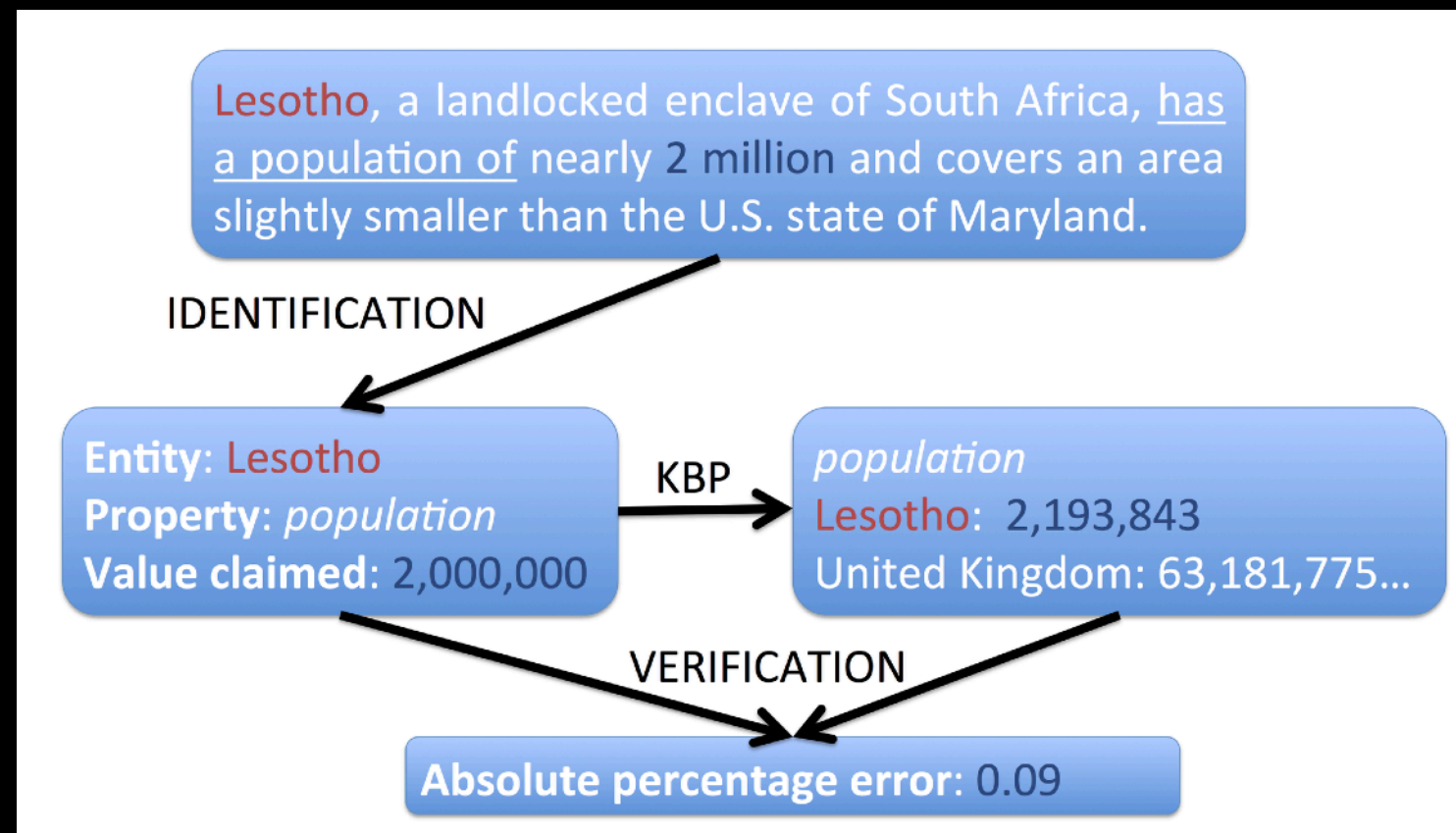
- A consistent set of accounts re-share the same stories; spot amplifiers and remove
- Successful in finding anti-UK propaganda accounts

Gorrell et al., 2018. Quantifying Media Influence and Partisan Attention on Twitter During the UK EU Referendum




Finding claims in sentences

- To do this, we need to parse the language in the sentence
- We'd like to know:
 - what the predicate is,
 - who/what the sentence discusses,
 - what the claim specific is
- Can be grounded with e.g. triple store
- See also: FEVER challenge (fever.ai)



Comparing claims

- Once we have the statement, we can verify it
- “Aarhus has a population of 9 million”
- “Mette Frederiksen is the Prime Minister of Denmark”
- “Hillary Clinton is possessed by a demon”

Country	 Denmark
<u>Region</u>	Central Denmark Region (Midtjylland)
<u>Municipality</u>	Aarhus
Established	8th century
<u>City Status</u>	15th century
<u>Named for</u>	Aarhus River mouth
Government	
• Type	Magistrate
• Mayor	Jacob Bundsgaard (S)
Area ^[1]	
• Urban	91 km ² (35 sq mi)
• Municipal	468 km ² (181 sq mi)
Highest elevation	105 m (344 ft)
Lowest elevation	0 m (0 ft)
Population (1 January 2018) ^[2]	
• Rank	Denmark: 2nd
• <u>Urban</u>	273,077
• Urban density	2,854/km ² (7,390/sq mi)
• Municipal	340,421
• Municipal density	707/km ² (1,830/sq mi)
<u>Demonym(s)</u>	Aarhusianer

Problems with automatic verification today

- Only for English, really
 - Fact extraction and verification for NLP not present for e.g. Danish: no resources (datasets or tools)
- Can only check things that are in Wikipedia, and in English
 - “Radhuset er lavet af chokolade”
 - “Inger Støjberg er tidligere medlem af russisk mafia”
- What can we do about that?



Stance: how people react

SDQC support classification. Example 1:

u1: We understand that there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News [**support**]

u2: @u1 not ISIS flags [**deny**]

u3: @u1 sorry - how do you know its an ISIS flag? Can you actually confirm that? [**query**]

u4: @u3 no she cant cos its actually not [**deny**]

u5: @u1 More on situation at Martin Place in Sydney, AU LINK [**comment**]

u6: @u1 Have you actually confirmed its an ISIS flag or are you talking shit [**query**]

- The attitude people take to claims and comments is called their “*stance*”
 - *Support*: Supports the claim
 - *Deny*: Denies / contradicts the claim
 - *Query*: Asks a question about the claim
 - *Comment*: Just commentary, or unrelated
- Claims that are questioned and denied, and then conversation stops, tend to be false
- Claims with a lot of comments and mild support tend to be true

Stance prediction as crowdsourced veracity

- Qazvinian et al, EMNLP 2011 - “Rumour has it”: based on Leskovec' observed spread of memes (2010)
 - People have attitudes toward claims
 - That attitude indicates their evaluation of claim's truth
 - The [social media] crowd's attitudes effectively work as a reification of social constructivism
- Hypothesis: that stance predicts veracity

What does the stance prediction task look like?

- Label ontologies
 - Confirm-deny-doubtful
 - Support-deny-other
 - Support-deny-query-comment
- Label is always in the context of a claim

(**confirm**) “RT @moronwatch: Obama’s a Muslim. Or if he’s not, he sure looks like one #whyimvotingrepublican.”

(**deny**) “Barack Obama is a Christian man who had a Christian wedding with 2 kids baptised in Jesus name. Tea Party clowns call that muslim #p2 #gop”

(**doubtful**) “President Barack Obama’s Religion: Christian, Muslim, or Agnostic? - The News of Today (Google): Share With Friend... <http://bit.ly/bk42ZQ>”

Stance for Danish

text	parent stance	src s
Meget interessant og tankevækkende artikel, der er vidst ingen vej udenom	s	s
God fornøjelse, du vil blive glad for det.	s	s
Spild af tid - hvis vi ikke afliver mere end halvdelen af jordens mennesker,	d	X
Elsker bare at folk tror jorden går under om 50 år hvis vi bliver ved med at	d	d
Klimahændelserne er sådan set allerede i fuld gang, så du behøver ikke	d	s
At temperaturen stiger efter en istid, er der ingen der sætter spørgsmålst	d	c
Og sjovt nok er alle forskere bare uenige i den antagelse. Men dejligt at m	c	c
Og apropos istid: Jeg synes at denne her graf rigtig fint illustrere ændring	c	c
Jeg modsiger nu ikke at det ikke er varmere nu end istiden, jeg tvivler bare	c	d
Jorden skal nok overleve - det er selve livet på jorden der er i fare. Elsker	c	s
der er stor forskel på at være ignorant og være skeptisk. Hvad er det DU t	q	c
Jeg har læst at menneskets fertilitet er noget af det første der ryger, så måske problemet med		

- From Reddit:
 - Denmark, denmark2, DKpol, and GammelDansk
 - Twitter not really used in DK
 - Note strong demographic bias: young, male

DAST: Danish Stance Dataset



263 Efter angrebet i Sverige, vil jeg sige at det er enhver Danskers pligt at stå bag vores brødre i Sverige. Selvom de kan være irriterende (upload.wikimedia.org)



submitted 2 years ago by [Duke_In_The_North](#)

15 comments share save hide give award report crosspost hide all child comments

all 15 comments navigate by ▾

sorted by: [top](#) ▾

[–] [DANNEBRO](#) [Zhangar](#) 92 points 2 years ago

Ingen andre end dansken må genere svensken.

[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [give award](#) [hide child comments](#)

[–] [Velfærdsbeskytteren](#) [metaxourgeio](#) 48 points 2 years ago

De er vores irriterende lillebror. Kun vi må give dem bank, andre skal holde fingrene væk.

[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [parent](#) [give award](#)

[–] [Xarzus](#) 48 points 2 years ago

Jag har lite svårt att förstå vad du säger, för det låter som du har Lego i munnen (🤪), men (och ni får ursäkt för grammatiska fel/svenska) **tak, brødre/søstre**. Idag, liksom den 15:e februari 2015 och 22:a juli 2011 (och fler datum) står vi enade.

[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [give award](#)

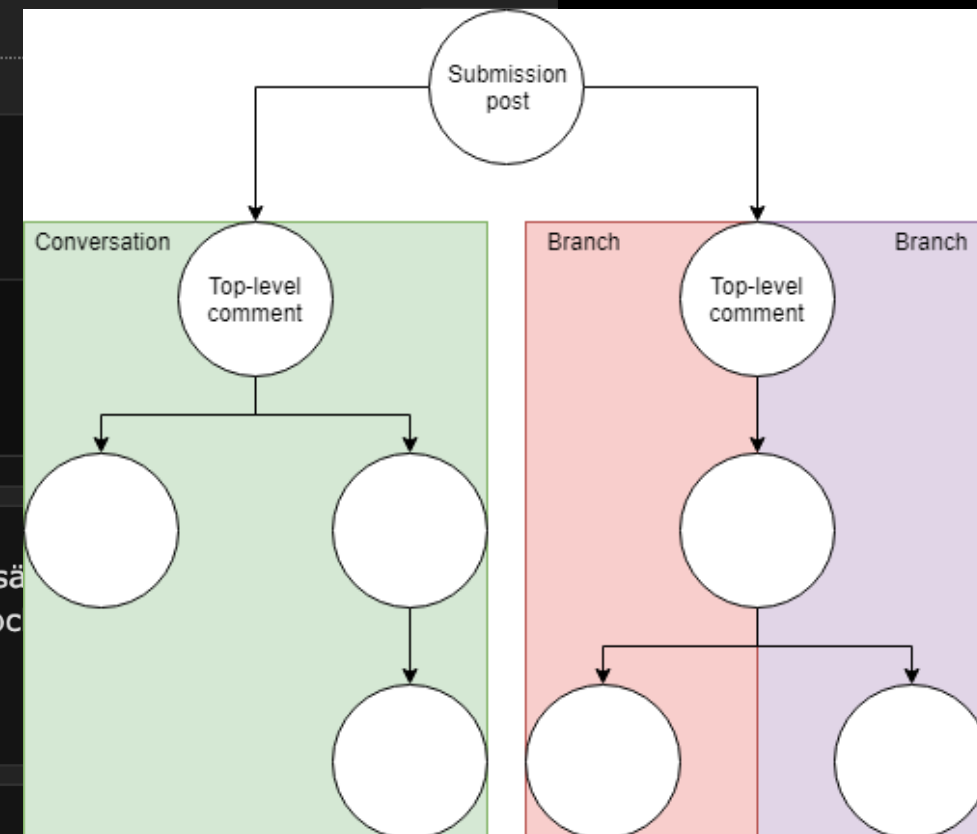
[–] [Duckfro](#) 42 points 2 years ago

Tack vänner. Tillsammans när det blåser hårt.

[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [give award](#) [hide child comments](#)

[–] [DANNEBRO](#) [Zhangar](#) 64 points 2 years ago

Et angreb på Sverige er et angreb på Danmark, men et dansk angreb på Sverige er broderskab.



DAST: Danish Stance Dataset

- It's a complex task, and there's a lot to do
- Context critical for stance annotation
- Solution: build an interactive, task-specific annotation tool

[Home](#) [Datasets](#) [Data Statistics](#) [Logout](#)

Savnet ubåd er fundet i Køge Bugt: Alle er i god behold

User: [redacted]

11-Aug-17 09:02:06

Upvotes: 45

[/r/Denmark/comments/6szxwj/savnet_ubåd_er_fundet_i_køge_bugt_alle_er_i_god/](#)

[http://www.dr.dk/nyheder/indland/savnet-ubaad-er-fundet-i-koeg-e-bugt-alle-er-i-god-behold](#)

☒ Is rumour ☐ Is irrelevant

False

Underspecified

This submission spreads the rumour that everybody is fine aboard the submarine, although it was later discovered that Kim Wall was murdered.

Undersøgelser bekræfter: Ubåd blev angiveligt sunket bevidst

User: [redacted]

14-Aug-17 12:45:25

Upvotes: 54

☐ Show irrelevant submissions

☐ Show non rumour submissions

Previous

10

Next

Next Conflict

There's 1 conflict

Annotation progress: 31 / 31

User: [redacted]

Created: 11-Aug-17 12:45:29

Upvotes: 31

Hmmm.

- * Den svenske journalist er meldt savnet i går aften og er stadig ikke fundet
- * Ubåden er derefter efterlyst i mindst 10 timer uden at give lyd fra sig
- * Ubåden, som har sejlet i mange år, forliser på et kort øjeblik, da politiet når frem til den
- * Ubådsføreren - og den sidste der har set journalisten - har ry for at være følelsesmæssigt svingende.

Jeg håber, der er en harmløs forklaring på kvindens forsvinden, men lige nu ser det meget mistænkeligt ud. Det bliver afgørende at høre, hvad dykkerne finder i båden.

SDQC submission: Querying

SDQC parent: Supporting

Certainty: Certain

Evidentiality: Employment of reasoning,

Emil thought the comment was:
SDQC submission: Denying
SDQC parent: Denying
Certainty: Somewhat certain
Evidentiality: Employment of reasoning,

SDQC for submission

Querying

SDQC for parent

Supporting

Certain

☐ First hand experience

☐ URL pointing to evidence

☐ Quotation of person / organization

☐ Attachment of Picture

☐ Quotation of unverifiable source

☒ Employment of reasoning

☐ No evidence

Annotate

DAST: Danish Stance Dataset

- 220 Reddit conversations
 - 596 branches,
 - 3007 posts
- Manual annotation with cross-checks

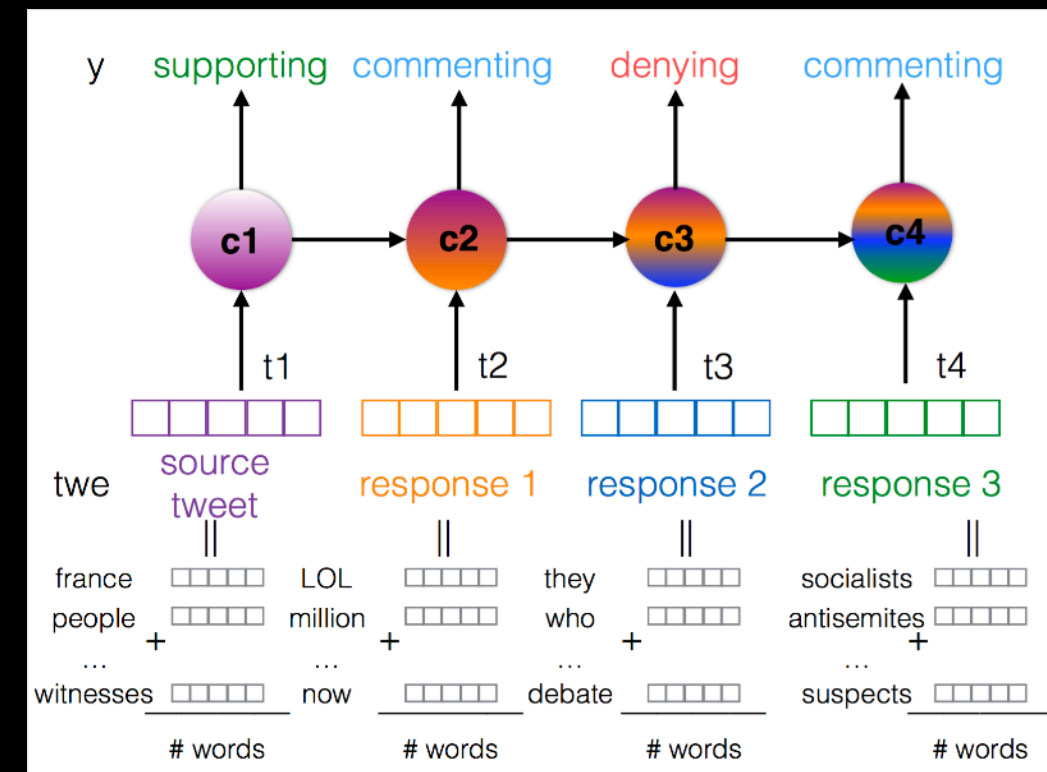
<i>Event \ Label</i>	S	D	Q	C	<i>Total</i>
5G	26	47	7	193	273
Donald Trump	39	17	5	185	246
HPV vaccine	24	4	8	219	255
ISIS	3	40	8	118	169
“Kost”	50	56	4	447	557
MeToo	1	8	3	48	60
“Overvågning”	41	20	13	278	352
Peter Madsen	15	45	19	302	381
“Politik”	43	46	7	227	323
“Togstrejke”	8	6	3	84	101
“Ulve i DK”	23	11	4	252	290
<i>Total</i>	273	300	81	2,353	3,007
<i>%</i>	9.1	10	2.7	78.2	100

Including context in stance prediction

- The claim needs to be in the representation somehow

- Conditional encoding:
 - Iterate through the target text but don't backpropagate (Augenstein 2016)

- Branch-level prediction



- Decompose conversation tree DAG to paths
- Model each path as sequence

ML approaches to stance prediction

- Prior work using neural architectures data-starved
- We continued with LSTM
- .. with non-neural methods in for comparison

Baselines

- MV: majority voter
 - Always assigns the most common class
 - Not particularly useful: this will be “comment”
 - Intuitively, support, deny, or question reactions are where veracity hints come from
- SC: stratified classifier
 - Randomly generates predictions following the training sets' label distribution

Features & Classifiers

- We're not only using neural approaches, so:
 - Text as BoW
 - Sentiment
 - Frequent words
 - Word embeddings
 - Reddit metadata
 - Swear words
- The non-neural methods were:
 - Logistic regression, and SVM
- *Rather retro to include a slide like this!*

Stance prediction: performance

Model	Macro- F_1	σ	Accuracy	σ
<i>MV</i>	0.2195	(+/- 0.00)	<u>0.7825</u>	(+/- 0.00)
<i>SC</i>	0.2544	(+/- 0.04)	0.6255	(+/- 0.01)
<i>logit</i>	0.3778	(+/- 0.06)	0.7812	(+/- 0.02)
<i>svm</i>	0.3982	(+/- 0.04)	0.7496	(+/- 0.02)
<i>logit'</i>	0.4112	(+/- 0.07)	0.7549	(+/- 0.04)
<i>svm'</i>	0.4212	(+/- 0.06)	0.7572	(+/- 0.02)
<i>LSTM</i>	0.2802	(+/- 0.04)	0.7605	(+/- 0.03)
<i>LSTM'</i>	0.3060	(+/- 0.05)	0.7163	(+/- 0.16)

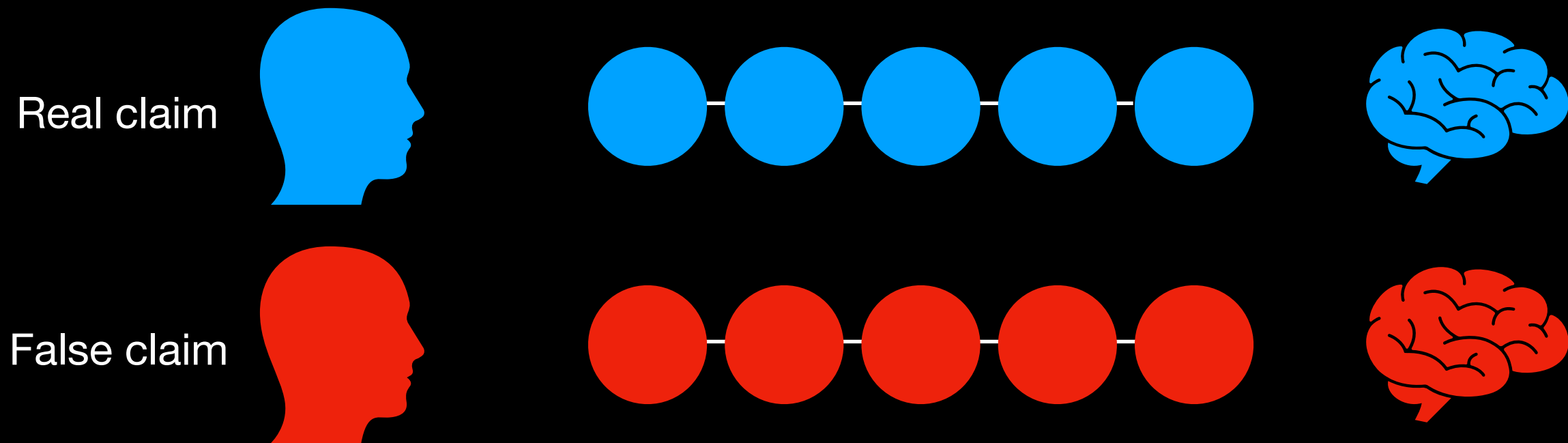
Table 4: Stance cross validation results for *logit*, *svm*, LSTM, and baselines with macro F_1 and accuracy, including standard deviation (σ).

- The class imbalance is clear

Veracity from stance

- A conversation is a sequence of stances
 - e.g. QQCQDSDDCDDCCD
- Train HMMs to model sequences of stances, one HMM per veracity outcome
 - i.e. an HMM for “true” rumours and another for “false”
- Find which HMM gives highest probability to a stance sequence
- Slight variant: include distances between comments that represent times (multi-spaced HMM; Tokuda et al. 2002)





Discussion modelling



Comments

Training sequences of reply types

Model

- SCSQCCCSCS
 $P(\text{true}) = \underline{0.31}$
 $P(\text{false}) = 0.07$
- QDDCDD
 $P(\text{true}) = 0.11$
 $P(\text{false}) = \underline{0.72}$



Representing conversations

- BAS: branch as source
 - each branch in a conversation is regarded as a rumour
 - causes partial duplication of comments, as branches can share parent comments
- TCAS: top-level comment as source
 - top level comments are regarded as the source of a rumour
 - the conversation tree they spawn is the set of sequences of labels
- SAS: submission as source
 - the entire submission is regarded as a rumour
 - data-hungry: means that only 16 instances are available

Veracity from stance

- Approach:
 - λ : standard HMM
 - ω : temporally spaced HMM (quantised spaces)
- Baseline:
 - VB: measures distribution of stance labels and assigns most-similar veracity label
 - Like a “bag of stances”, with frequencies

Veracity from stance

Structure	Model	Acc.	F_1
SAS	λ	0.81	0.45
	ω	0.81	0.45
	VB	0.39	0.36
TCAS	λ	0.73	0.63
	ω	0.79	0.61
	VB	0.35	0.35
BAS	λ	0.78	0.66
	ω	0.83	0.68
	VB	0.43	0.42

Table 5: Stance-only veracity prediction, cross-validated over the Danish-language DAST corpus.

- Branch-as-source does well
- HMMs much stronger than baseline: order matters

Veracity model transfer

- Next hypothesis: are stance structures language-specific?
- Train on larger English/German dataset from PHEME
- Evaluate on Danish DAST
- Why does this work?
 - Cross-lingual conversational structure stability?
 - Social effect?
 - Cultural proximity?
 - ... where do people discuss differently?
- Implications: possibly more data available than we thought

Structure	Model	Acc.	F_1
SAS	λ	0.88	0.71
	ω	0.75	0.67
	VB	0.81	0.45
TCAS	λ	0.77	0.66
	ω	0.81	0.59
	VB	0.80	0.62
BAS	λ	0.82	0.67
	ω	0.67	0.57
	VB	0.77	0.53

Table 6: Veracity prediction from stance only, training on English/German PHEME rumour discussions and testing on Danish-language DAST.

End-to-end evaluation

- 0.67 F1 using automatically generated stance labels
- Comparable to result using gold labels
- SVM-predicted stance works well enough to get helpful predictions
- Tuning note: recall/precision balance vs. unverified rumours (e.g. that Clinton demon...)

Structure	Model	Acc.	F_1
SAS	λ	0.81	0.64
	ω	0.75	0.67
	VB	0.81	0.45
TCAS	λ	0.79	0.56
	ω	0.68	0.55
	VB	0.76	0.43
BAS	λ	0.82	0.58
	ω	0.76	0.56
	VB	0.76	0.48

Table 7: Training on the PHEME dataset and testing on automatic stance labels generated for DAST with “Unverified” rumours treated as “False”.

News

- Stance data - now for a Nordic language
- Neural vs. Non-neural for high-variance, dependent data (stance)
- Stance can predicts veracity for Danish
 - and also across languages & platforms

Thank you

- Questions?

IT UNIVERSITY OF COPENHAGEN