Research Group Social Computing
TUM Department of Informatics
Technical University of Munich

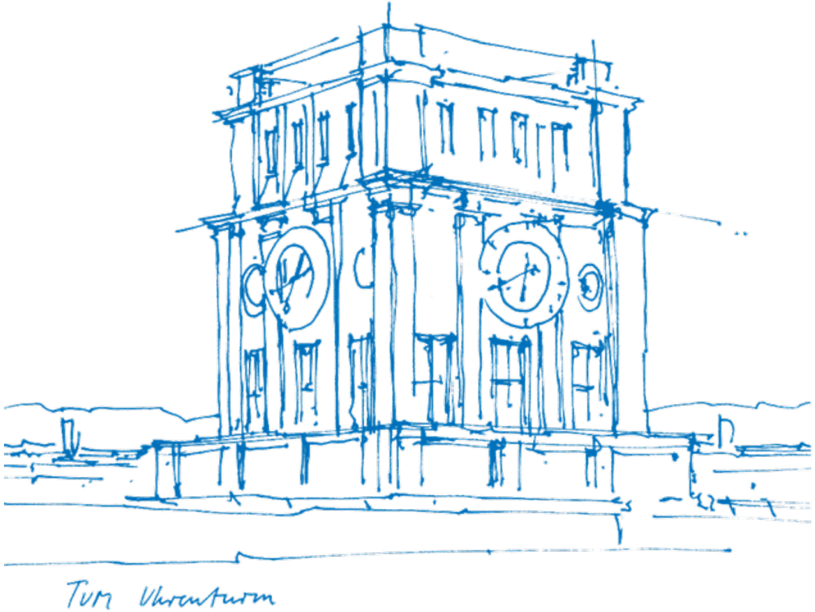
Global Explainability for Natural Language Processing Models

Guided Research - Final Presentation

B. Sc. Simon Klimek

Supervisor: M. Sc. Edoardo Mosca

April 16, 2021



Background - eXplainable Artificial Intelligence (XAI)

- How does a machine learning model come to a conclusion?
- Simple (linear) vs. complex (non-linear) models
- Humans can not comprehend a deep neural network!

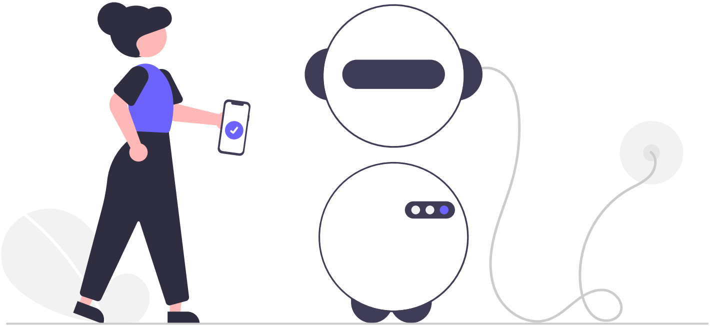


illustration made by undraw.co

Motivation

- First thought: What do we want to explain?
- Given: A tool allowing us to search and download posts from Facebook

Motivation

- First thought: What do we want to explain?
- Given: A tool allowing us to search and download posts from Facebook

- “The idea that the current crisis may be especially fertile ground for conspiracy beliefs may well be correct.”
  - 50% little evidence of conspiracy thinking
  - 25% a degree of endorsement
  - 15% consistent pattern of endorsement
  - 10% had very high levels of endorsement

[Freeman et al., 2020]

Motivation

- First thought: What do we want to explain?
- Given: A tool allowing us to search and download posts from Facebook

- “The idea that the current crisis may be especially fertile ground for conspiracy beliefs may well be correct.”
  - 50% little evidence of conspiracy thinking
  - 25% a degree of endorsement
  - 15% consistent pattern of endorsement
  - 10% had very high levels of endorsement
- “Higher levels of corona virus conspiracy thinking were associated with *less adherence to all government guidelines* and *less willingness to take diagnostic or antibody tests or to be vaccinated.*”

[Freeman et al., 2020]

Motivation

Figure: Capitol Riots - Jason Andrew for The New York Times [Williamson, 2021]

Related Work

Guidotti et al. [2018] - Survey giving us an overview of the state of the art XAI methods:

- **Model Explanation Problem (global)**
- Outcome Explanation Problem (local)
- Model Inspection Problem (global/local)
- Transparent Box Design Problem (global/local)

Related Work

Guidotti et al. [2018] - Survey giving us an overview of the state of the art XAI methods:

- **Model Explanation Problem (global)**
- Outcome Explanation Problem (local)
- Model Inspection Problem (global/local)
- Transparent Box Design Problem (global/local)

- All global explainability methods are surrogates.
- Listed no global XAI methods which can be used for NLP.

Related Work

Covert et al. [2020] - Understanding Global Feature Contributions With Additive Importance Measures

- Based on Shapley values
- Not method proposed for text based inputs
- Uses a group of pixels for images
- Use a closed unit (e.g. named entity)

Related Work

Covert et al. [2020] - Understanding Global Feature Contributions With Additive Importance Measures

- Based on Shapley values
- Not method proposed for text based inputs
- Uses a group of pixels for images
- Use a closed unit (e.g. named entity)

Here’s tomorrow’s Ancient Aliens lineup. Better wake up early. Starts at 7am eastern time in the USA.

The system is broken. It amazes me that a young child in kindergarten can be kicked out of school because a document isn’t checked and dated for a vaccine. We are all wearing masks!

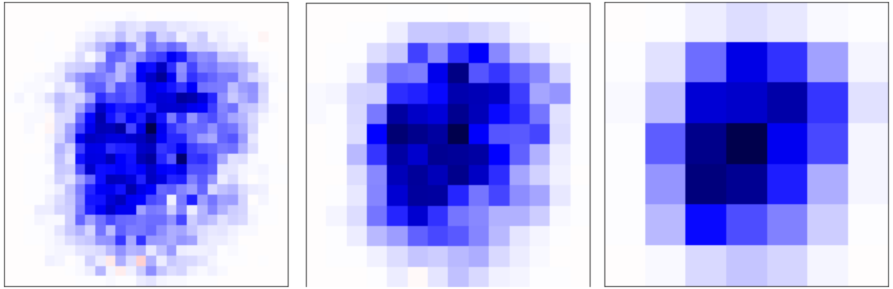
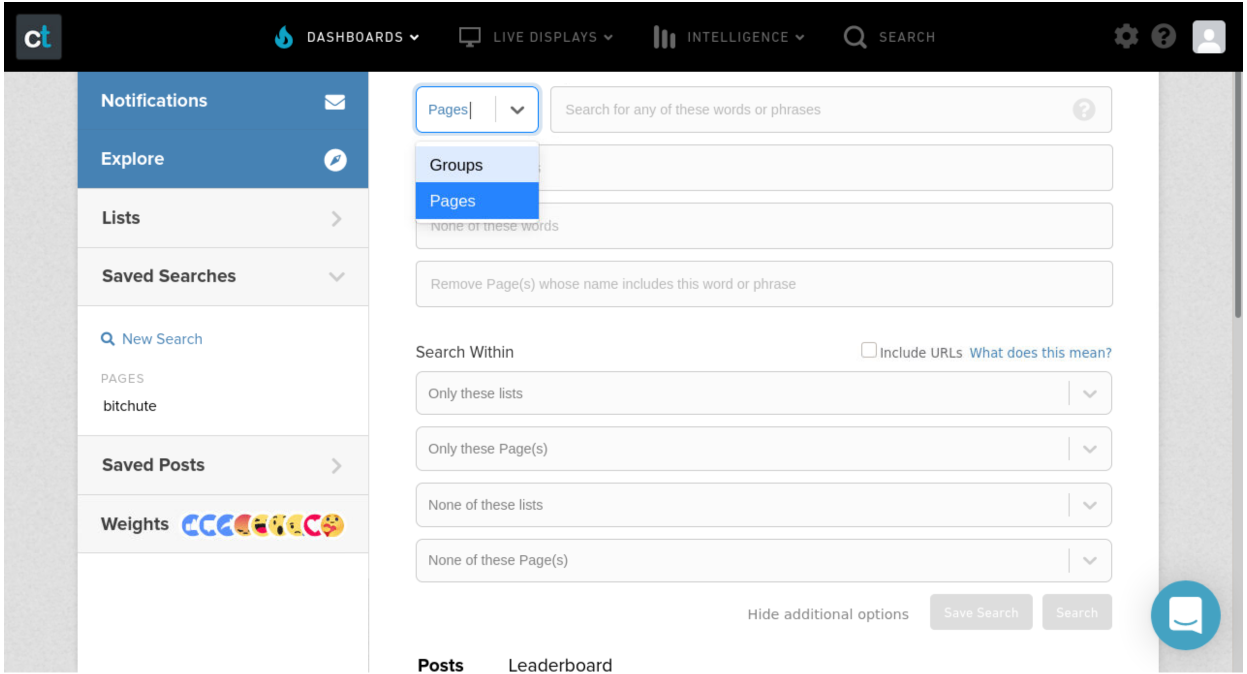


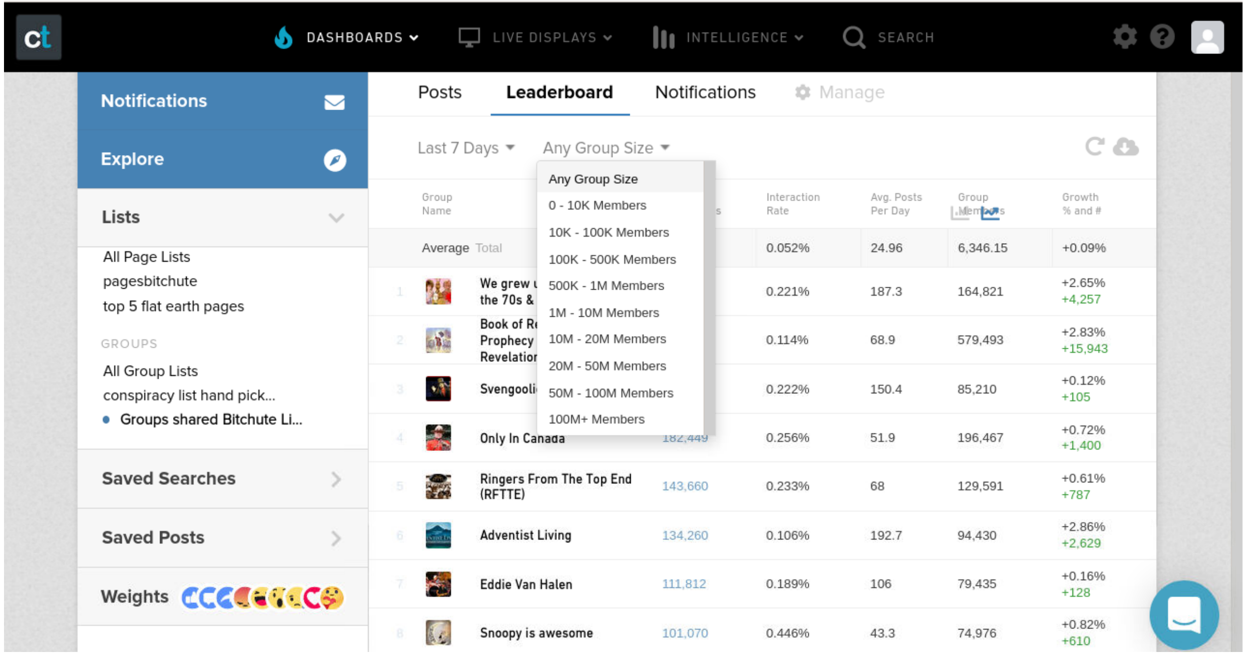
Figure: from left to right: individual pixels, 2x2 super pixels, 4x4 super pixels

Dataset Retrieval

CrowdTangle, “a public insights tool owned and operated by Facebook” [Shiffman, 2021]



Dataset Retrieval



All Page Lists
pagesbitchute
top 5 flat earth pages

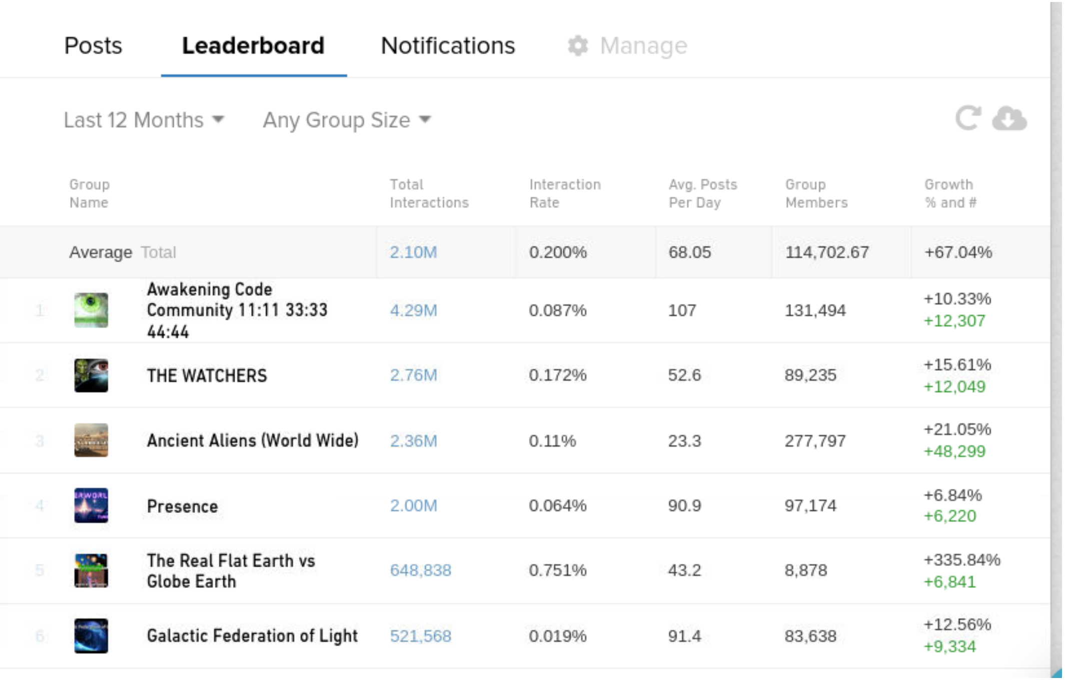
GROUPS

All Group Lists
conspiracy list hand pick...
● Groups shared Bitchute Li...

Any Group Size
0 - 10K Members
10K - 100K Members
100K - 500K Members
500K - 1M Members
1M - 10M Members
10M - 20M Members
20M - 50M Members
50M - 100M Members
100M+ Members

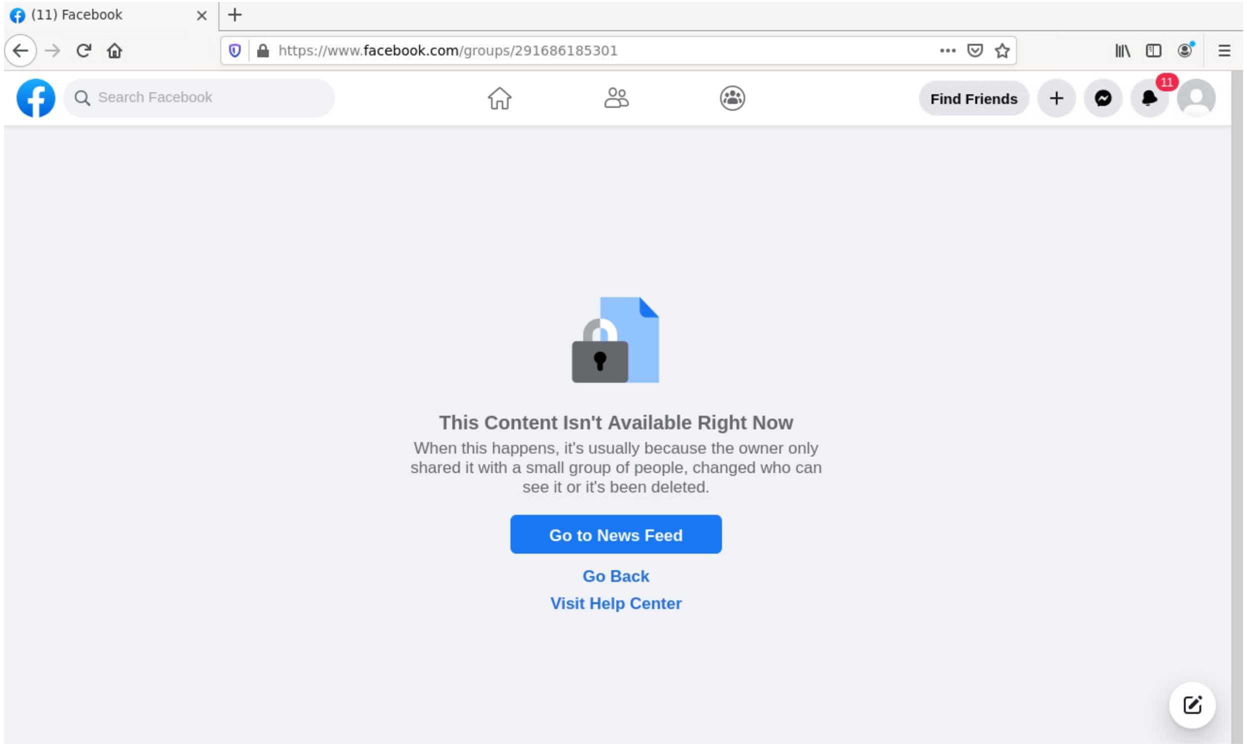
	Group Name	s	Interaction Rate	Avg. Posts Per Day	Group Members	Growth % and #
	Average Total		0.052%	24.96	6,346.15	+0.09%
1	We grew u the 70s &		0.221%	187.3	164,821	+2.65% +4,257
2	Book of Re Prophecy Revelatior		0.114%	68.9	579,493	+2.83% +15,943
3	Svengooli		0.222%	150.4	85,210	+0.12% +105
4	Only In Canada	182,449	0.256%	51.9	196,467	+0.72% +1,400
5	Ringers From The Top End (RFTTE)	143,660	0.233%	68	129,591	+0.61% +787
6	Adventist Living	134,260	0.106%	192.7	94,430	+2.86% +2,629
7	Eddie Van Halen	111,812	0.189%	106	79,435	+0.16% +128
8	Snoopy is awesome	101,070	0.446%	43.3	74,976	+0.82% +610

Dataset Retrieval



	Group Name	Total Interactions	Interaction Rate	Avg. Posts Per Day	Group Members	Growth % and #
	Average Total	2.10M	0.200%	68.05	114,702.67	+67.04%
1	Awakening Code Community 11:11 33:33 44:44	4.29M	0.087%	107	131,494	+10.33% +12,307
2	THE WATCHERS	2.76M	0.172%	52.6	89,235	+15.61% +12,049
3	Ancient Aliens (World Wide)	2.36M	0.11%	23.3	277,797	+21.05% +48,299
4	Presence	2.00M	0.064%	90.9	97,174	+6.84% +6,220
5	The Real Flat Earth vs Globe Earth	648,838	0.751%	43.2	8,878	+335.84% +6,841
6	Galactic Federation of Light	521,568	0.019%	91.4	83,638	+12.56% +9,334

Dataset Retrieval



This Content Isn't Available Right Now

When this happens, it's usually because the owner only shared it with a small group of people, changed who can see it or it's been deleted.

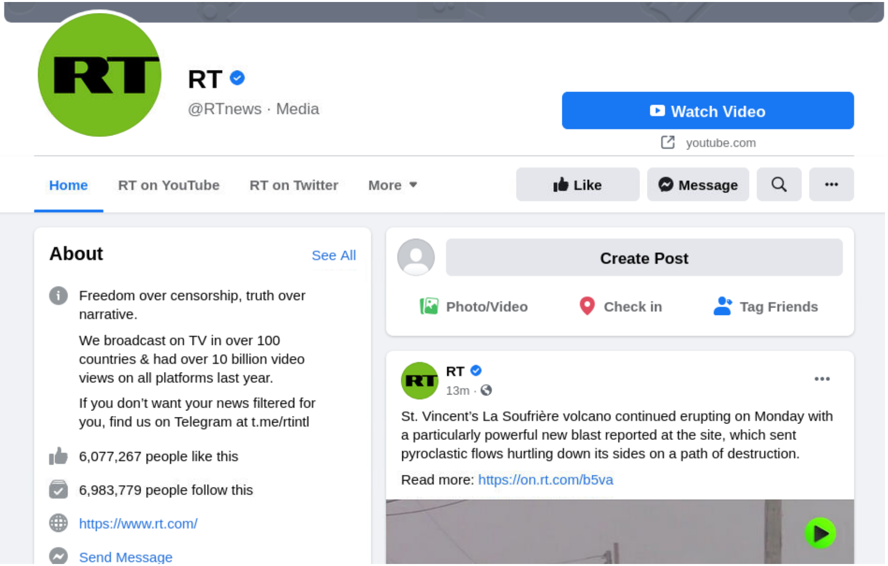
Dataset Retrieval

- Goal: at least 30,000 posts
- Posts had to be at least two month old

Dataset Retrieval

- Goal: at least 30,000 posts
- Posts had to be at least two month old

- Result: two datasets: conspiracy theory groups and RTnews (Russia Today)[1]



Freedom over censorship, truth over narrative.

We broadcast on TV in over 100 countries & had over 10 billion video views on all platforms last year.

If you don't want your news filtered for you, find us on Telegram at t.me/rtintl

6,077,267 people like this

6,983,779 people follow this

https://www.rt.com/

Send Message

St. Vincent's La Soufrière volcano continued erupting on Monday with a particularly powerful new blast reported at the site, which sent pyroclastic flows hurtling down its sides on a path of destruction.

Read more: https://on.rt.com/b5va

Figure: RT Facebook page

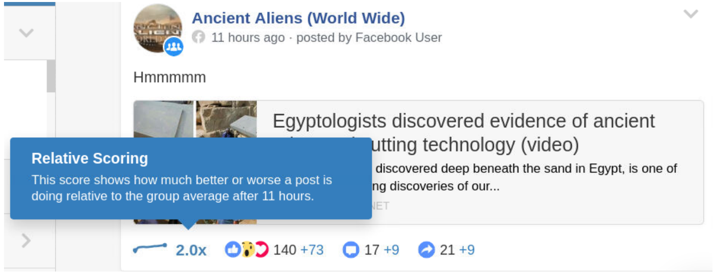
[1] A paper analyzing how RT shares conspiracy theories and false information was written by Yablokov [2015]

Dataset Preprocessing

- Duplicate posts have been removed

Dataset Preprocessing

- Duplicate posts have been removed



Ancient Aliens (World Wide)

11 hours ago · posted by Facebook User

Hmmmmm

Egyptologists discovered evidence of ancient
utting technology (video)

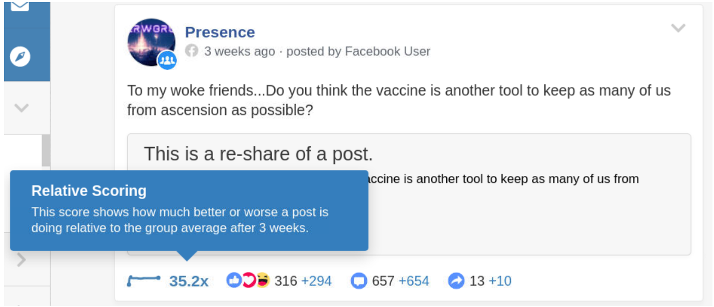
discovered deep beneath the sand in Egypt, is one of
ng discoveries of our...

ET

Relative Scoring

This score shows how much better or worse a post is doing relative to the group average after 11 hours.

2.0x 140 +73 17 +9 21 +9



Presence

3 weeks ago · posted by Facebook User

To my woke friends...Do you think the vaccine is another tool to keep as many of us from ascension as possible?

This is a re-share of a post.

ccine is another tool to keep as many of us from

Relative Scoring

This score shows how much better or worse a post is doing relative to the group average after 3 weeks.

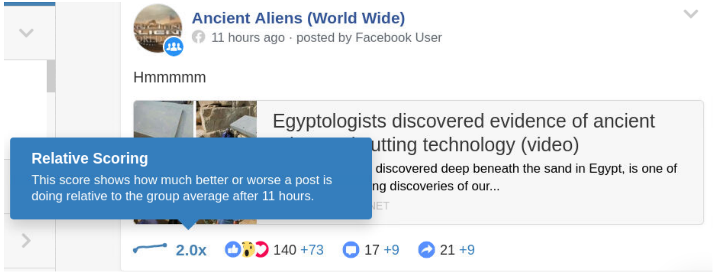
35.2x 316 +294 657 +654 13 +10

Dataset Preprocessing

- Duplicate posts have been removed
- Our score formula

$$\text{score} = \frac{\#\text{reactions}}{\#\text{group avg number of reactions}}$$

illustration made by undraw.co



Ancient Aliens (World Wide)

11 hours ago · posted by Facebook User

Hmmmmm

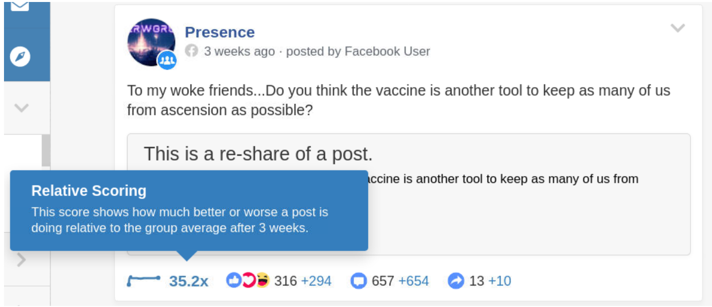
Egyptologists discovered evidence of ancient
utting technology (video)

discovered deep beneath the sand in Egypt, is one of
ng discoveries of our...

Relative Scoring

This score shows how much better or worse a post is doing relative to the group average after 11 hours.

2.0x 140 +73 17 +9 21 +9



Presence

3 weeks ago · posted by Facebook User

To my woke friends...Do you think the vaccine is another tool to keep as many of us from ascension as possible?

This is a re-share of a post.

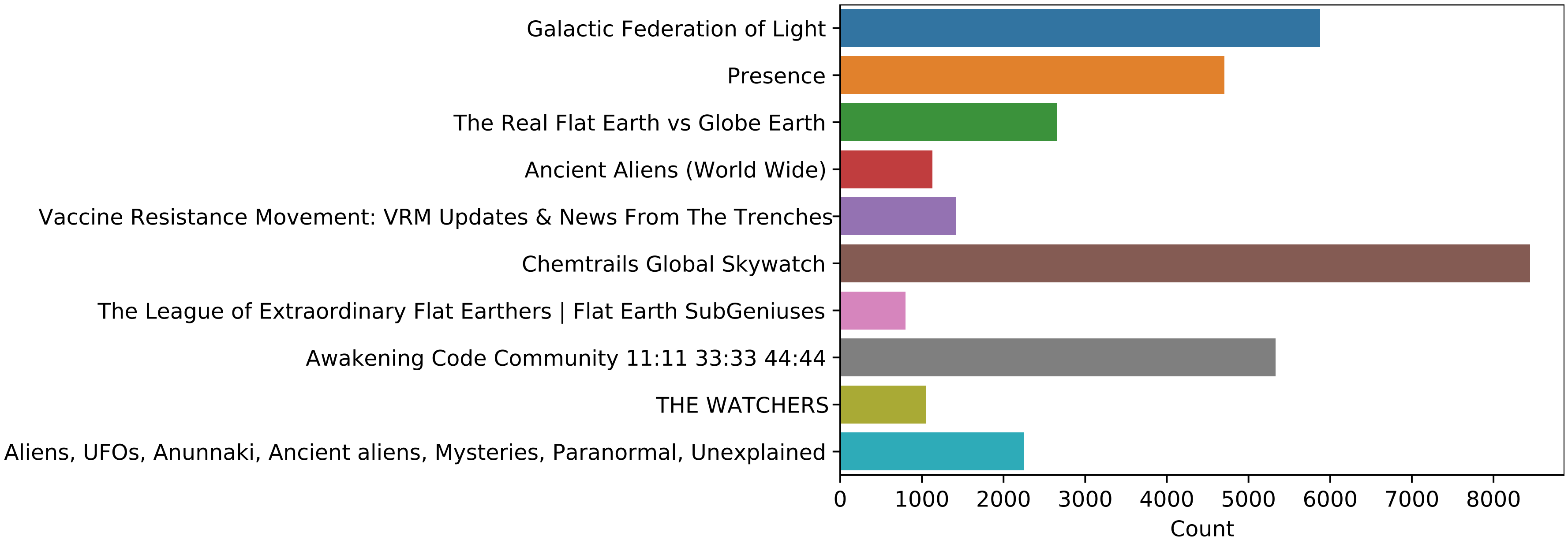
ccine is another tool to keep as many of us from

Relative Scoring

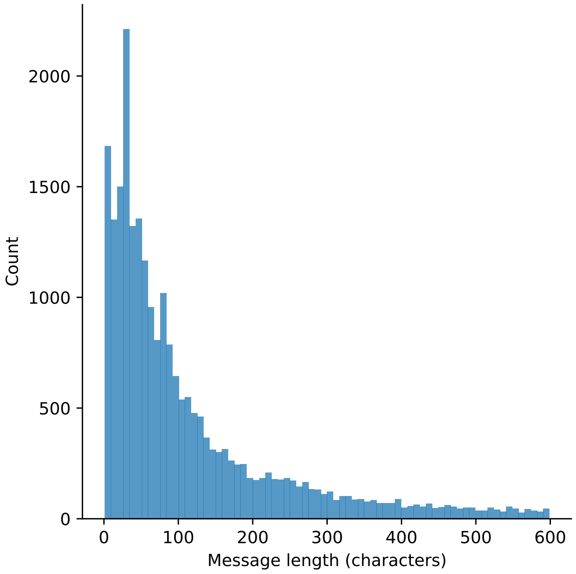
This score shows how much better or worse a post is doing relative to the group average after 3 weeks.

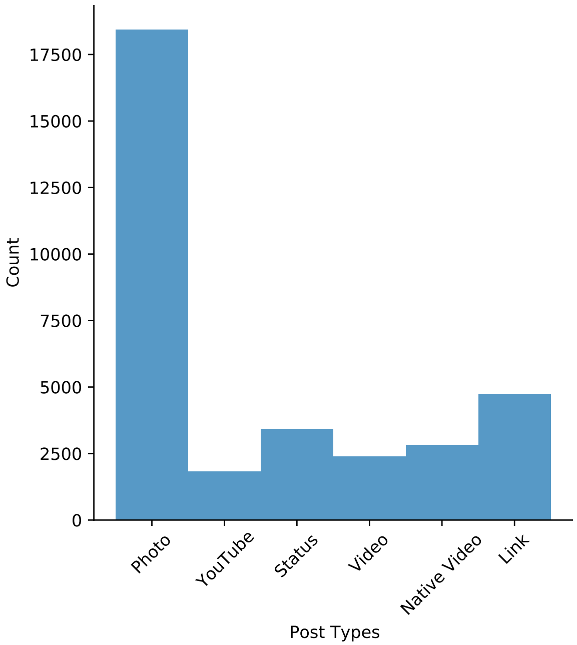
35.2x 316 +294 657 +654 13 +10

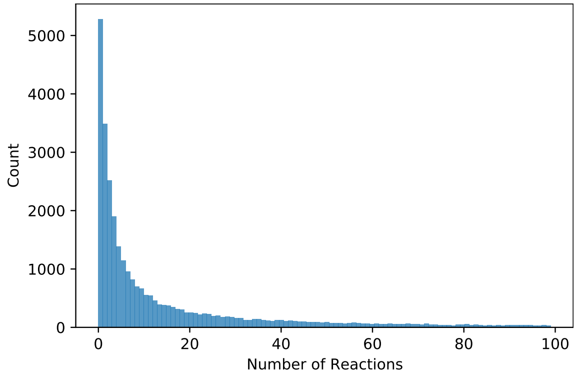
Dataset 1: Conspiracy Theory Groups



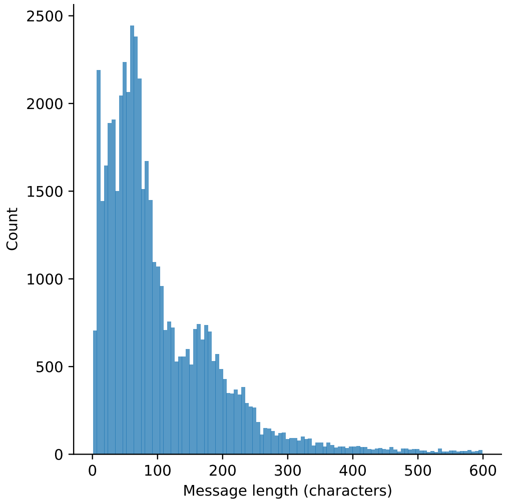
Dataset 1: Conspiracy Theory Groups

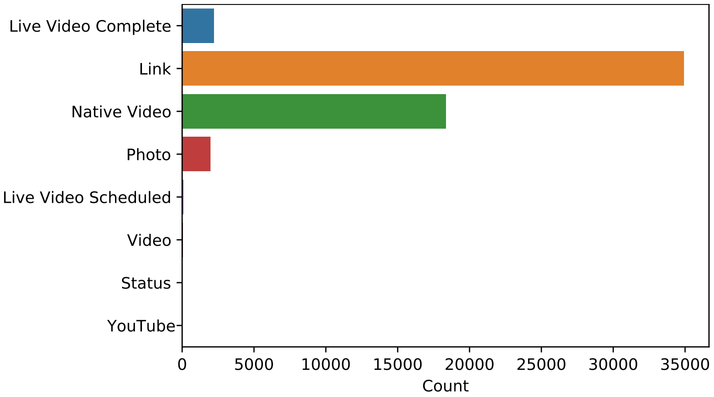


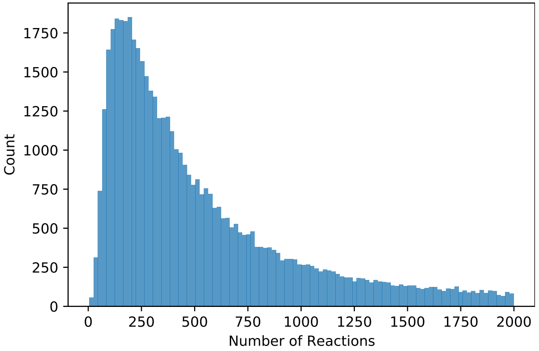




Dataset 2: rtnews







Experiments

- Baseline: linear regression
  - features: group, post type and top level domain
- Neural model based on DistilBert [Sanh et al., 2019b]
- BiLSTM model
  - categorical features + post text
  - optimizer: Adam
- 60-20-20 split (train / validation / test)

Results

Table: mean squared error for all models and datasets

	conspiracy theory groups	**rtnews**
linear model	6.68	177.51
DistilBert	6.46	177.00
BiLSTM	6.54	177.44

Results

Table: mean squared error for each group

Group	Linear Model	DistilBert based	BiLSTM
Galactic Federation of Light	12.71	12.49	12.58
Presence	3.51	3.39	3.46
The Real Flat Earth vs Globe Earth	2.18	2.16	2.07
Ancient Aliens (World Wide)	2.08	1.90	1.94
Vaccine Resistance Movement VRM Updates . . .	5.43	5.24	5.47
Chemtrails Global Skywatch	7.10	7.00	6.99
The League of Extraordinary Flat Earthers . . .	1.22	0.84	0.95
Awakening Code Community 1111 3333 4444	7.75	7.42	7.61
THE WATCHERS	2.21	2.01	2.12
Aliens, UFOs, Anunnaki, Ancient aliens, . . .	6.72	5.97	6.21

Results

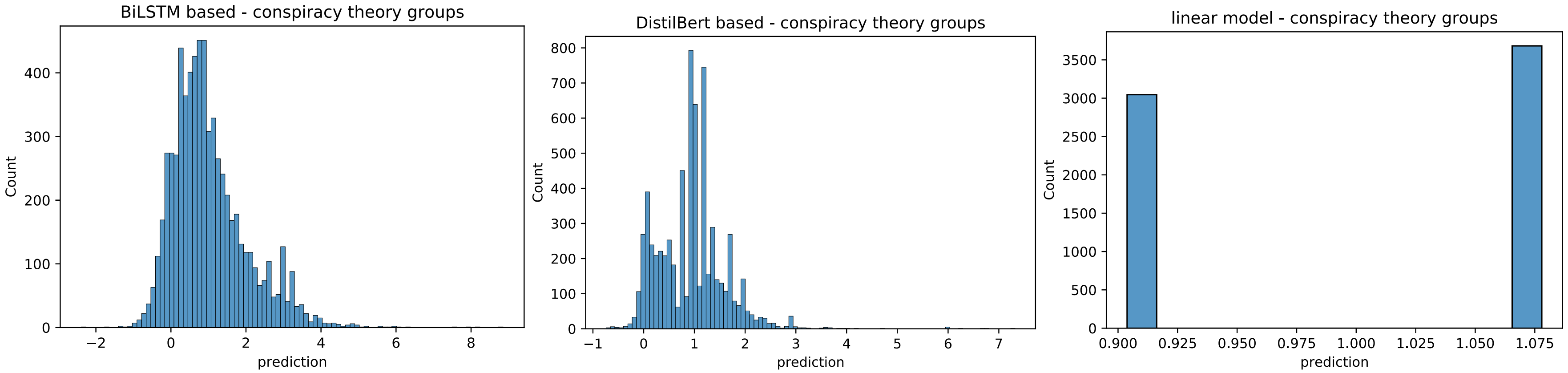


Figure: distribution of the model's score variable prediction for the conspiracy theory group dataset

Results

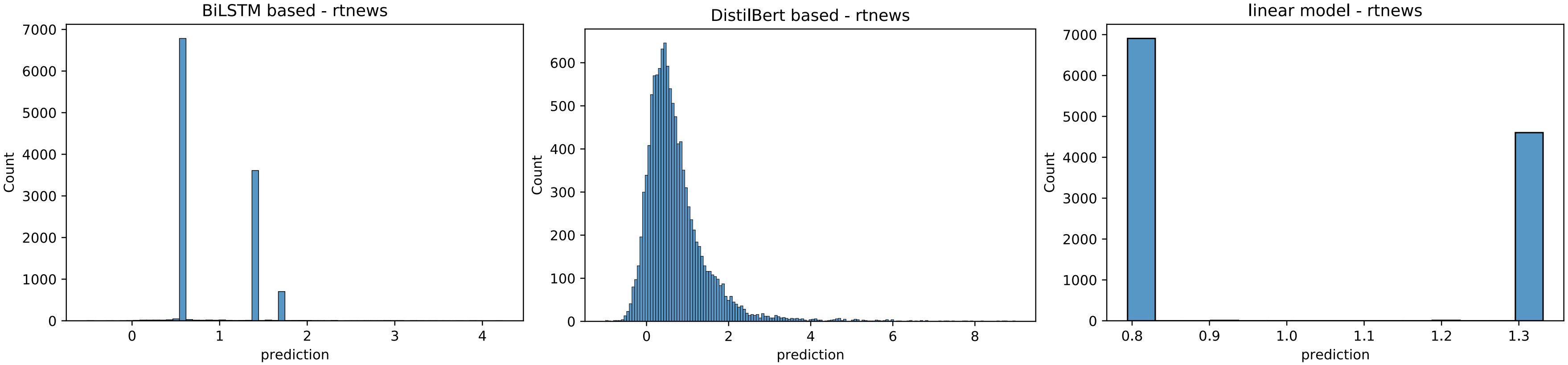


Figure: distribution of the model's score variable prediction for the rtnews dataset

Discussion

- We could not outperform the baseline model
- The baseline model is already comprehensible for humans

- Uniform distribution of reactions
- Use of other topics / groups
- No relation at all?

Thank you for your attention

Feel free to ask questions!

References I

I. Covert, S. M. Lundberg, and S.-I. Lee. Understanding global feature contributions with additive importance measures. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223. Curran Associates, Inc., 2020.

D. Freeman, F. Waite, L. Rosebrock, A. Petit, C. Causier, A. East, L. Jenner, A.-L. Teale, L. Carr, S. Mulhall, and et al. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in england. *Psychological Medicine*, page 1–13, 2020. doi: 10.1017/S0033291720001890.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

S. Krishnan and E. Wu. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, pages 1–6, 2017.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

References II

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019a. URL `http://arxiv.org/abs/1910.01108`.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019b.

N. Shiffman. Citing crowdtangle data, 2021. URL `https://help.crowdtangle.com/en/articles/3192685-citing-crowdtangle-data`.

E. Williamson. Rioters followed a long conspiratorial road to the capitol. *The New York Times*, Jan 2021. URL `https://www.nytimes.com/2021/01/27/us/politics/capitol-riot-conspiracies.html`.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

I. Yablokov. Conspiracy theories as a russian public diplomacy tool: The case of Russia Today (RT). *Politics*, 35(3-4):301–315, 2015.