Global Explainability for understanding opinions on social media

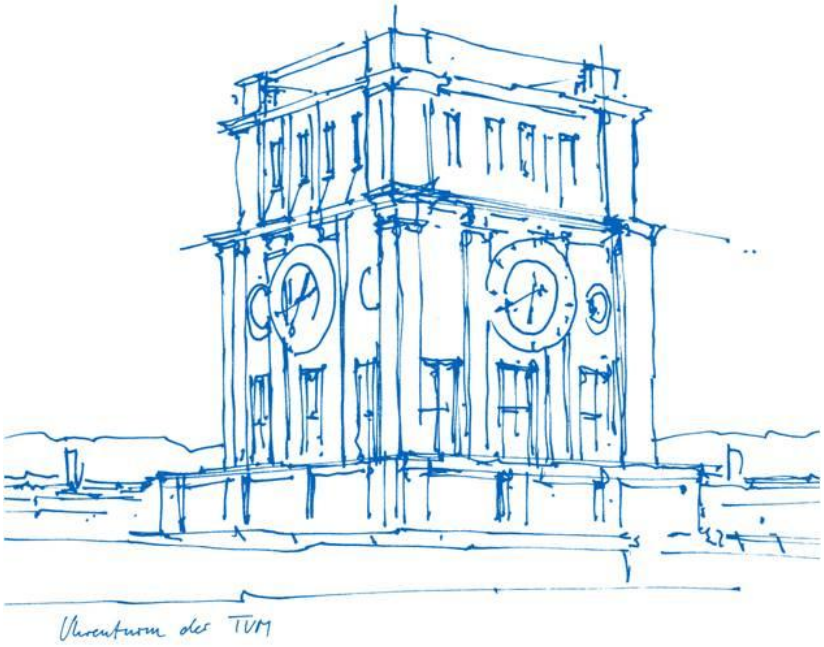
Technische Universität München

Fakultät für Informatik

NLP Lab Course, SS21

19.07.2021

Amine Bentellis, Arpit Karwasara and Moritz Schüler



Introduction

- **Predict** some characteristics of a post (**ideology, impact**)
- **Explain** the model on a **global** level

figure taken from [7]

Agenda

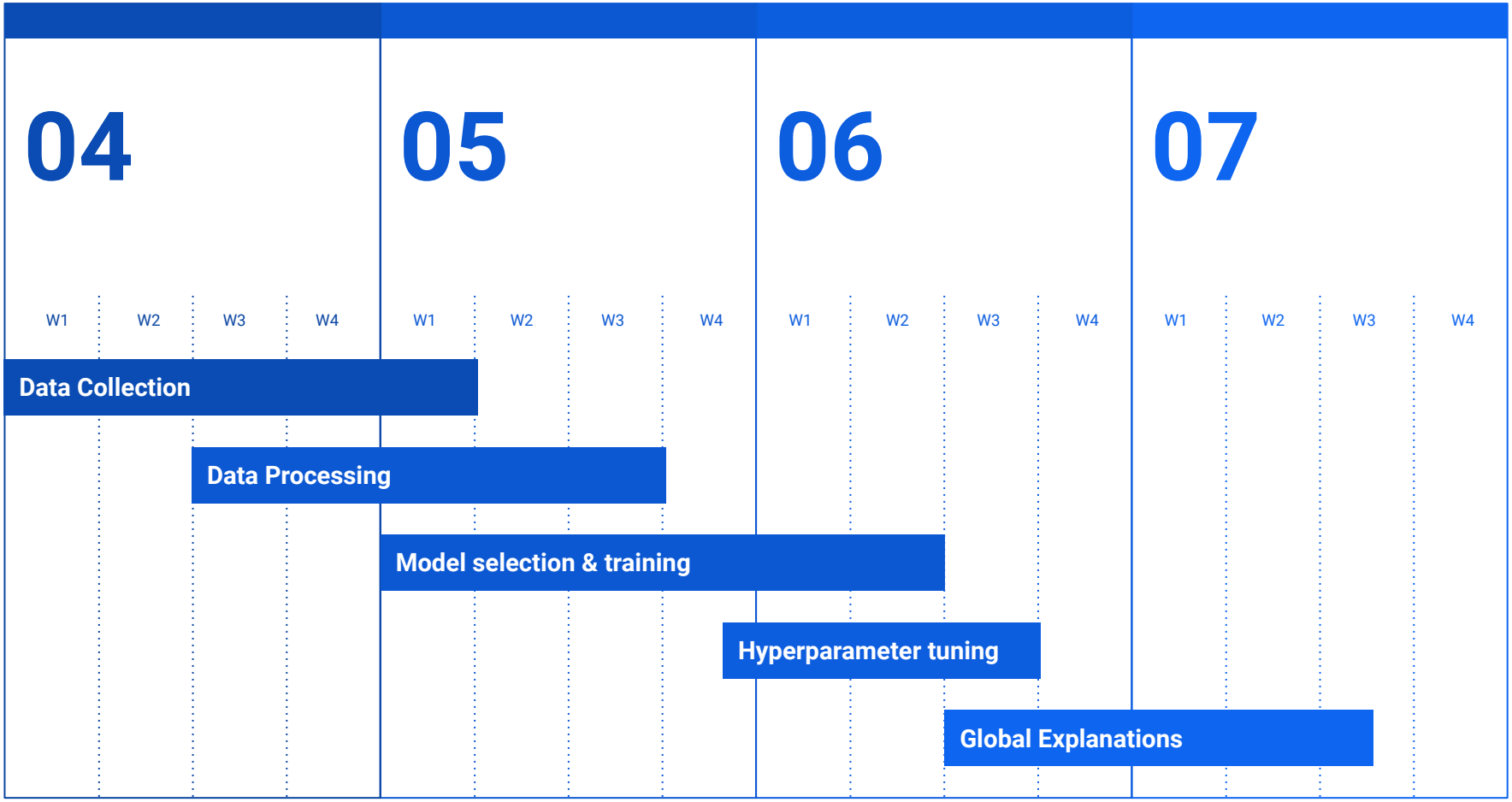
- Related work
- Timeline
- Data collection & processing
- Model
- Explanations
- Application on Stance

Related Work

- Related research
  - NLP for Social media [1]
  - Stance Prediction [2]
  - SAGE[3], SHAP[4], TCAV[5], Layer Attribution[6]
- Previous work
  - Local explanations
  - Focused on vision, not much for NLP

→ SAGE/Layer Attribution for NLP in upcoming slides

Timeline


04
05
06
07
W1
W2
W3
W4
Data Collection
Data Processing
Model selection & training
Hyperparameter tuning
Global Explanations

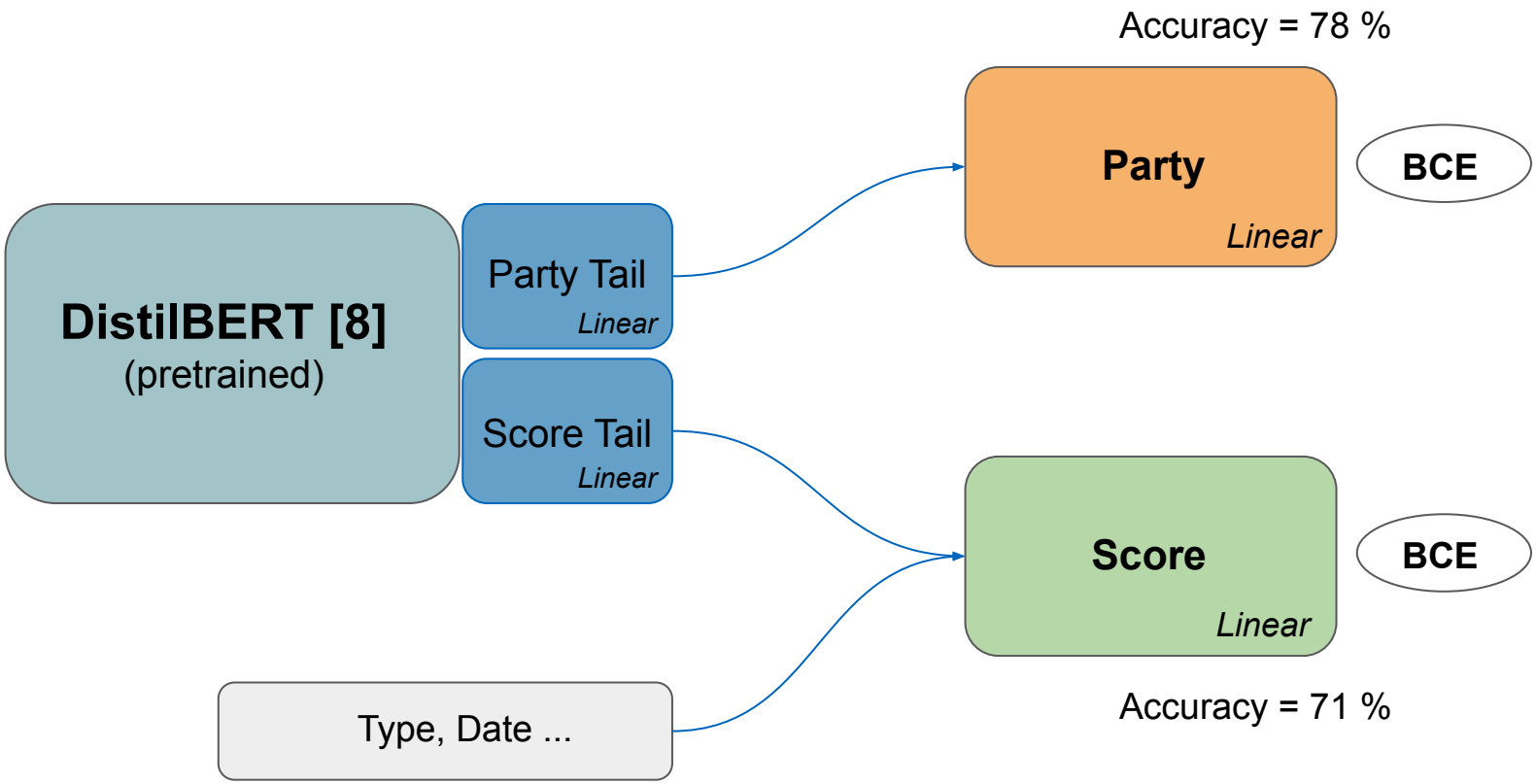
Crowdtangle Dataset [7]

Two lists:

- Republican Party (83 different pages)
- Democratic Party (105 different pages)

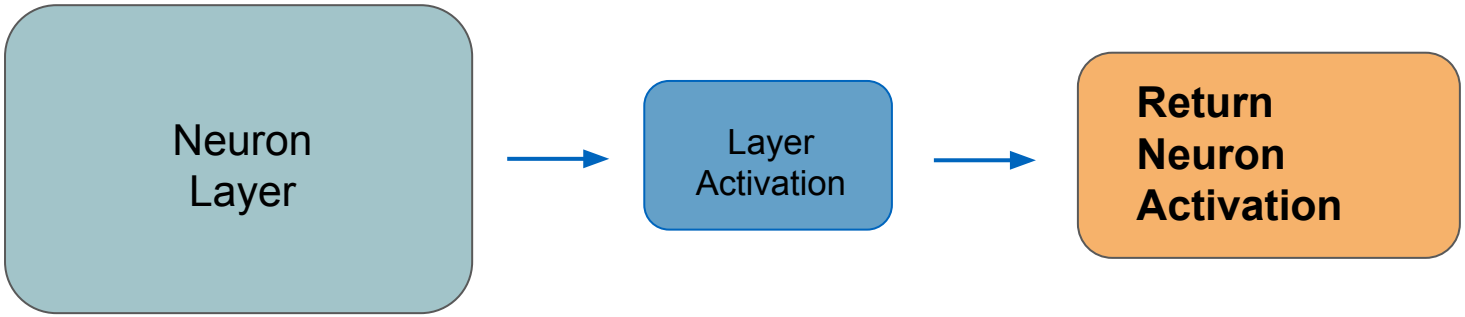
	Page Name	Total Interactions	Interaction Rate	Avg. Posts Per Day	Views on Owned Videos	Page Followers	Growth % and #
	Average Total	13,510.19	0.572%	1.88	6,346.86	246,238.76	+0.06%
1	 U.S. Senator Bernie Sanders	304,353	0.345%	1.71	144,069	7,356,273	-0.04% -2,699
2	 U.S. Senator Elizabeth Warren	262,595	0.379%	2.86	53,036	3,463,871	-0.04% -1,312
3	 Senator Chuck Schumer	76,431	0.335%	5.57	27,921	584,745	+0.05% +288
4	 Alexandria Ocasio-Cortez	72,796	0.432%	1.29	80,290	1,873,576	-0.01% -279

Network architecture

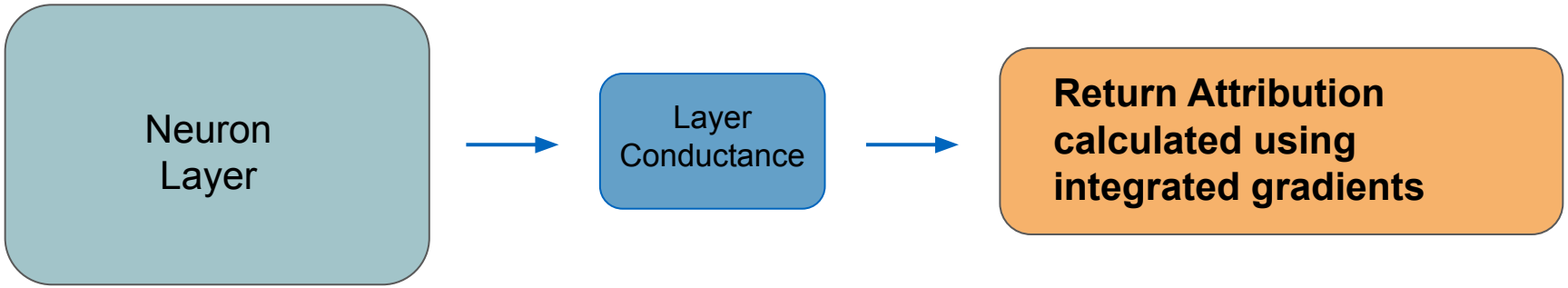


Explanations

Layer Attribution[6]

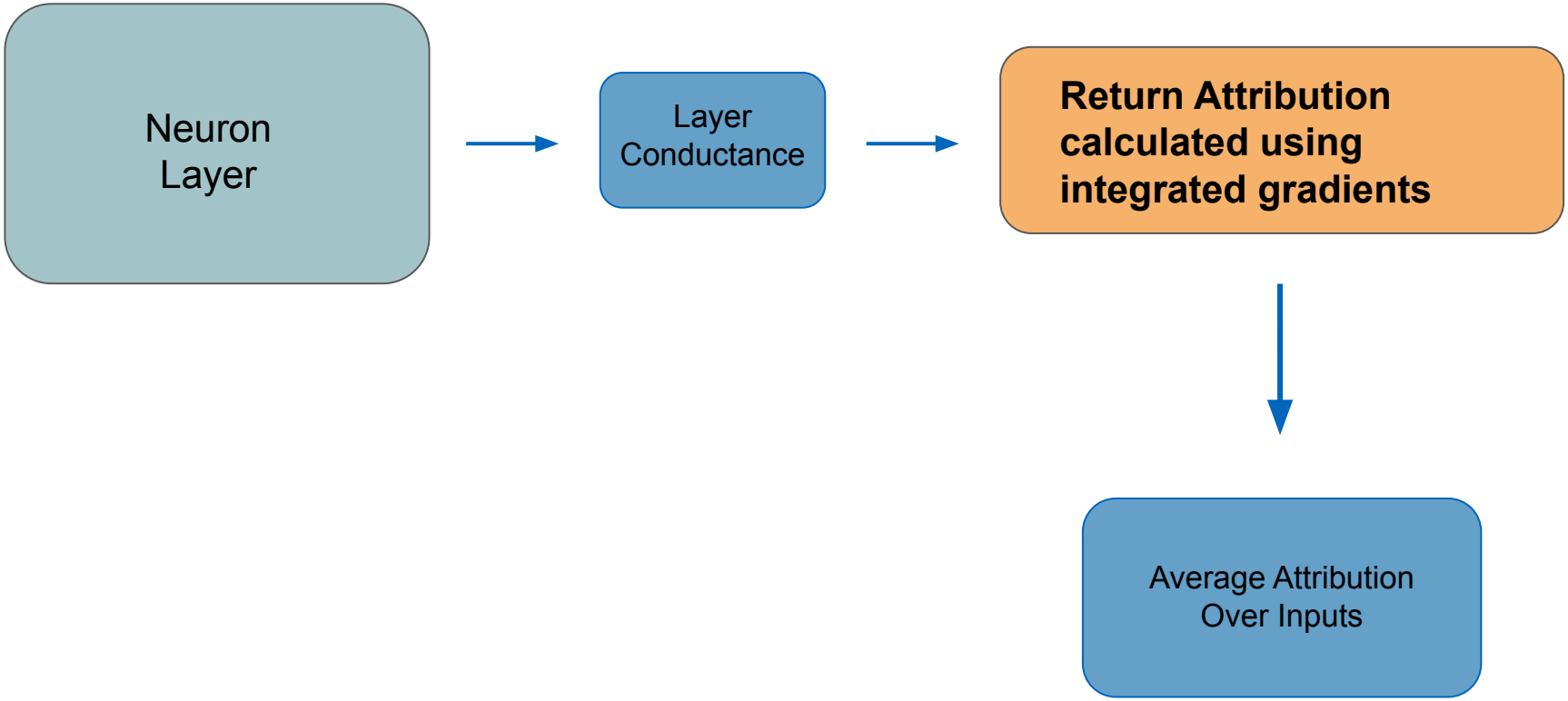


Layer Activation Approach



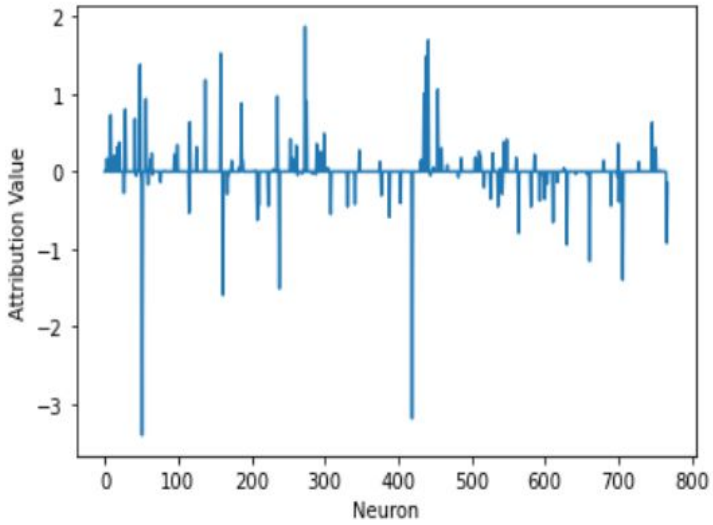
Layer Conductance Approach

Layer Attribution on our data

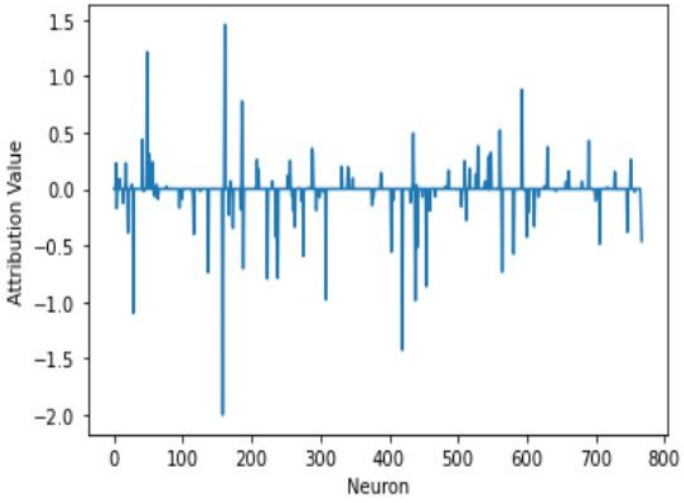


Layer Attribution [4]

AVERAGE HIDDEN LAYER NEURON ATTRIBUTION (FOR BERT TAIL)



DEMOCRATS



REPUBLICANS

PREDICTION CHANGES IF NEURONS WITH MORE ATTRIBUTION ARE DROPPED

SAGE[3]

SAGE answers the question

how much does the model depend on each feature overall?

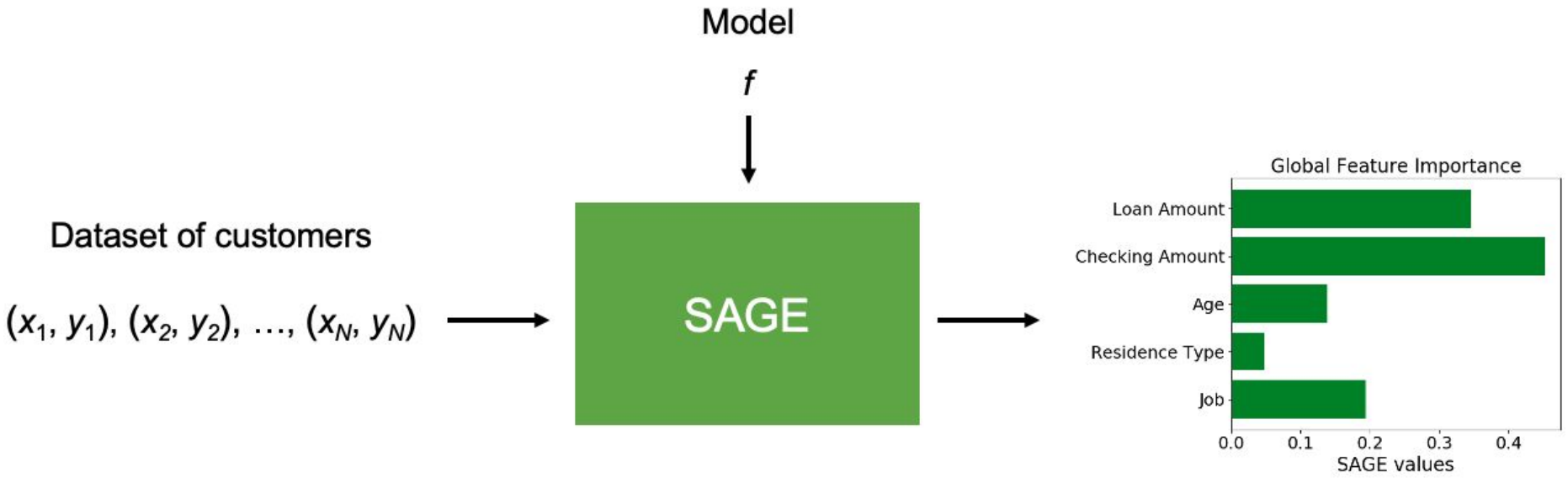
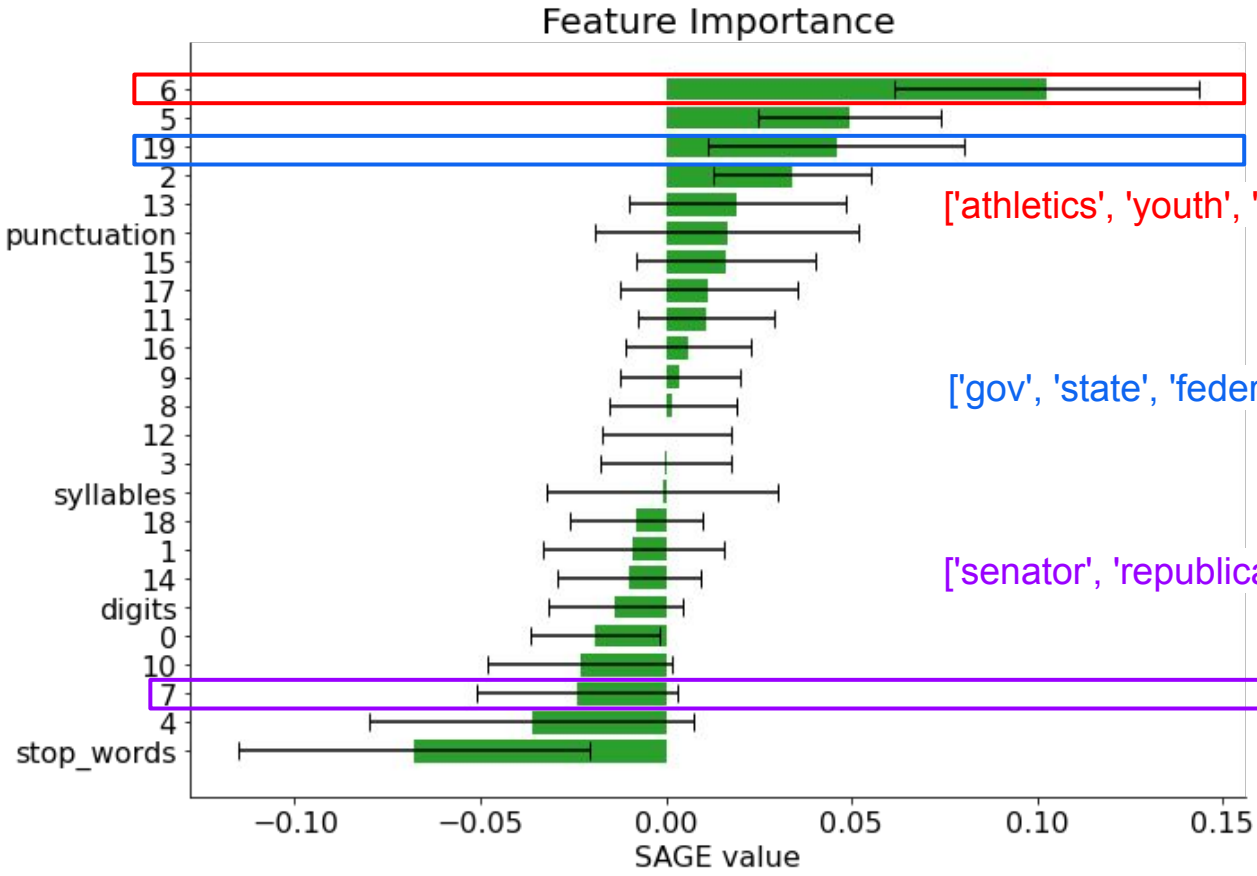


figure taken from [10]

SAGE on our data

Features → [‘kelly’, ‘joe’, ‘glenn’, ‘atkinson’, ‘jon’, ‘chuck’ …]



['athletics', 'youth', 'student', 'school', 'sports', ...]

['gov', 'state', 'federal', 'america', 'corporation', 'jobs', ...]

['senator', 'republican', 'vote', 'republicans', 'bill', ...]

Stance Prediction

Sem Eval Stance Dataset [9]

Tweets

Tweet	Target	Train/Te..	Stance	Opinion T..	Sentiment la..
If abortion is not wrong, then nothing is wrong. Powerful words from Blessed Mother..	Legalization o..	Train	AGAINST	Target	pos
Mary, Help of Christians persecuted everywhere, pray for us! #HolyLove #UnitedHear..	Legalization o..	Train	AGAINST	Other	pos
TY Michael @ASavageNation "I'd do anything to help him; he's right, he's telling the tr..	Donald Trump	Test	FAVOR	Target	pos
1 Cor 15:58 ...stand firm...Always give yourselves fully to the work of the #Lord...your l..	Atheism	Train	AGAINST	Other	pos

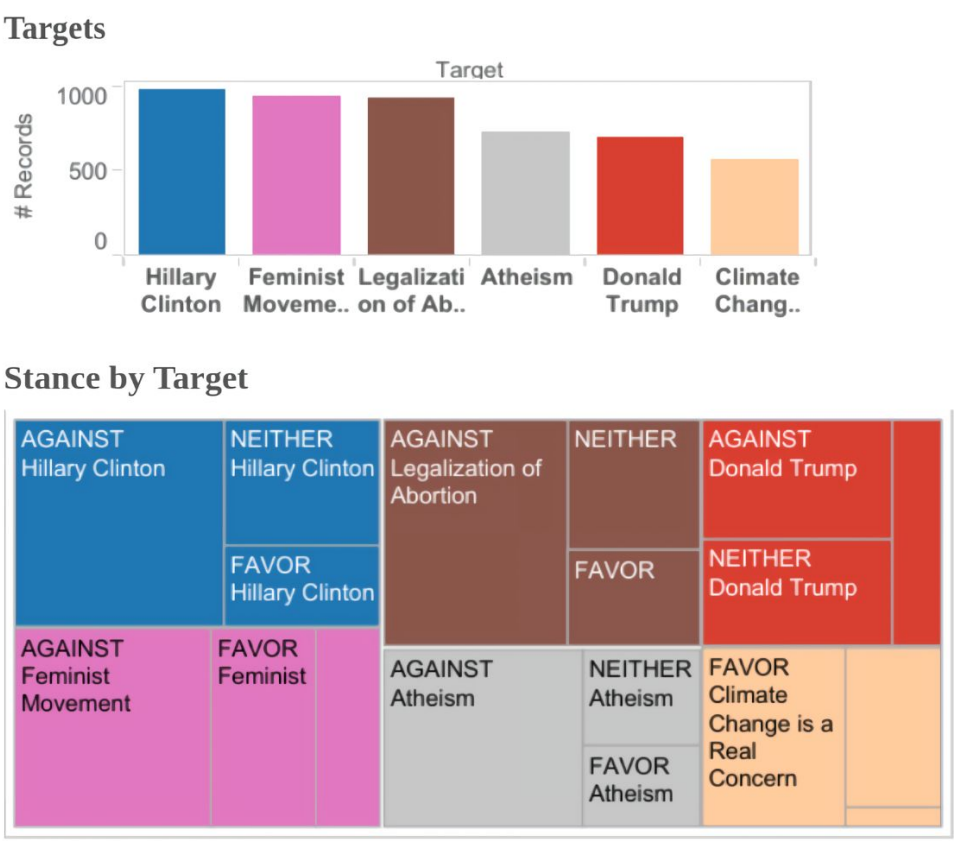
Stance: whether a sentence is in *favor* or *against* a target topic

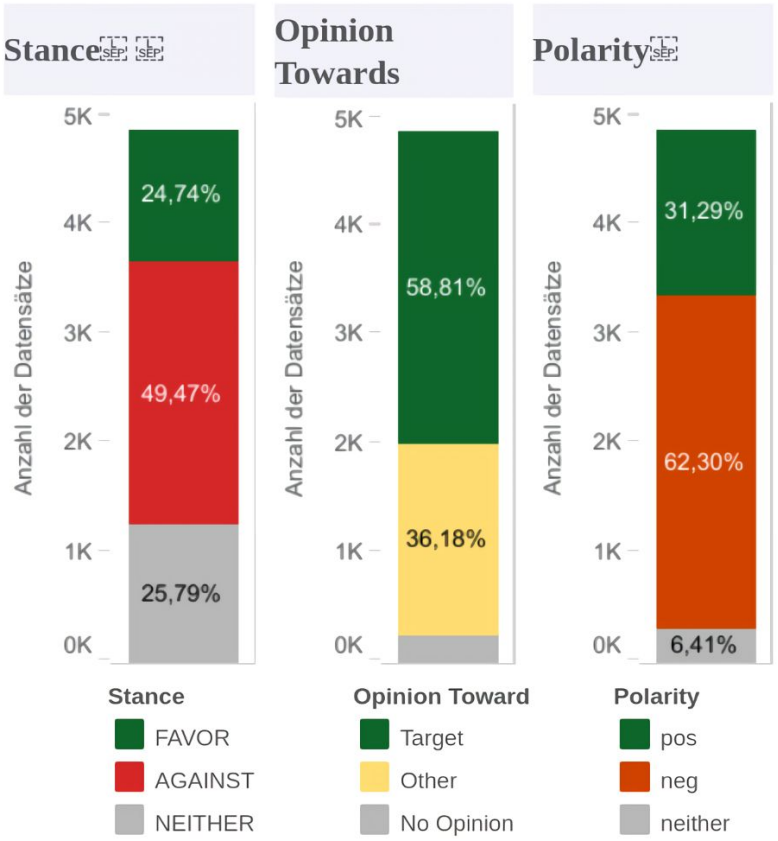
Goal: given a sentence predict the *stance* & the *target*

Source: annotated twitter dataset

figure taken from [11]

Sem Eval Stance Dataset [9]





X by Y Matrices

Opinion Toward

Stance	Target	Other	No Opinio..
FAVOR	94,69%	4,73%	0,58%
AGAINST	71,03%	28,31%	0,66%
NEITHER	0,96%	81,45%	17,60%

Sentiment labels

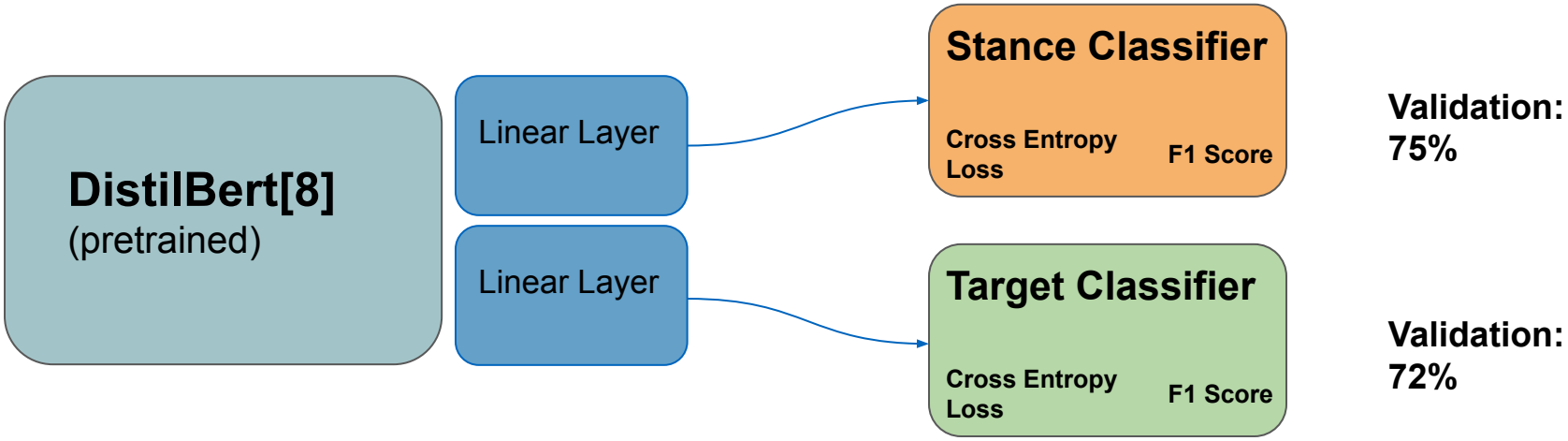
Stance	pos	neg	neither
FAVOR	40,25%	51,70%	8,05%
AGAINST	27,94%	69,12%	2,95%
NEITHER	29,14%	59,39%	11,46%

Sentiment labels

Opinion To..	pos	neg	neither
Target	29,92%	65,36%	4,71%
Other	32,58%	61,63%	5,79%
No Opinion	38,11%	31,15%	30,74%

figure taken from [11]

Network architecture

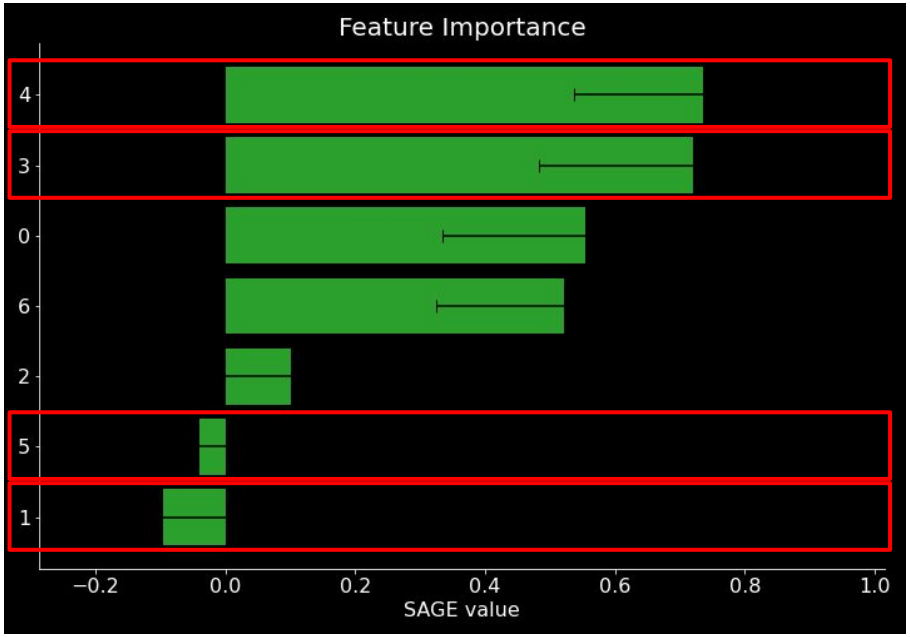


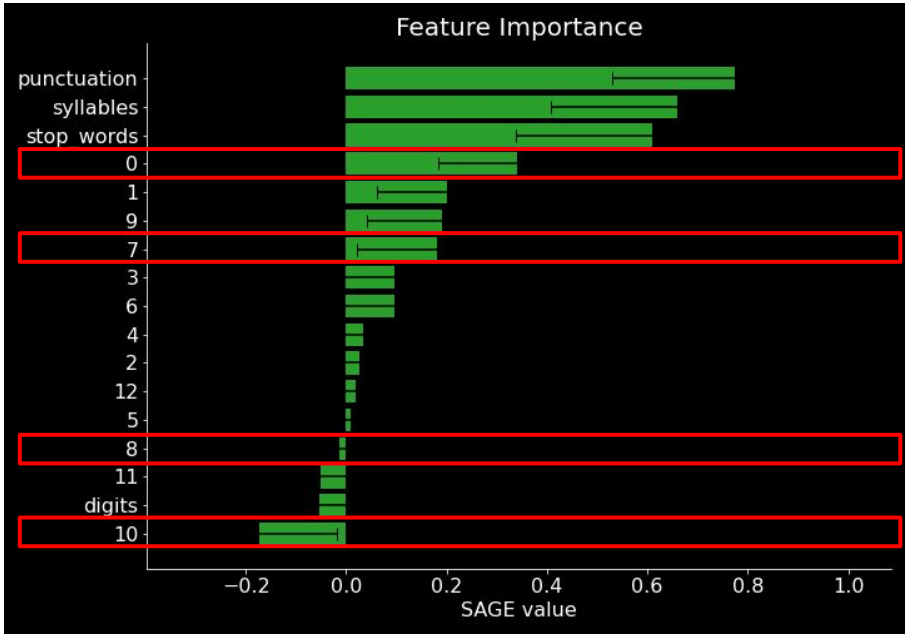
Results

```
============
Results
============
FAVOR     precision: 0.6468 recall: 0.5000 f-score: 0.5640
AGAINST   precision: 0.7426 recall: 0.7385 f-score: 0.7405
------------
Macro F: 0.6523
```

Perspective:
Winning model 2016:
67% Macro F1 score

Sage for Stance





1: [netflix, 'potter', 'volvo', 'auto', 'minute', 'second', ...]

3: ['in', 'or', 'into', 'were', 'during', 'against', ...]

4: ['!', '#', '(', ')', '@', '?', ...]

5: ['00', '27', '230', '730', '2014', '58', ...]

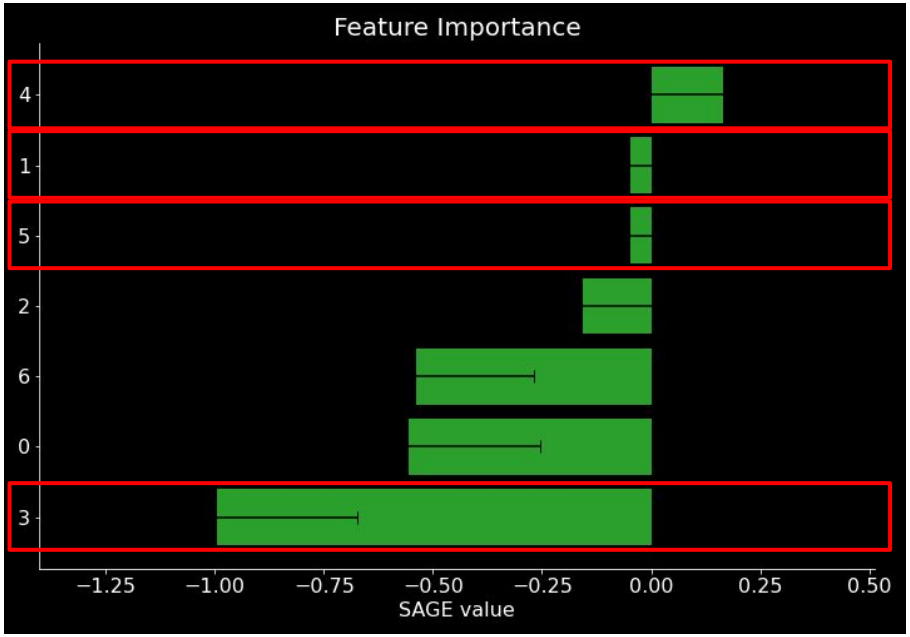
0: ['wonder', 'amazed', 'pro', 'folks', 'band', 'lineup', ...]

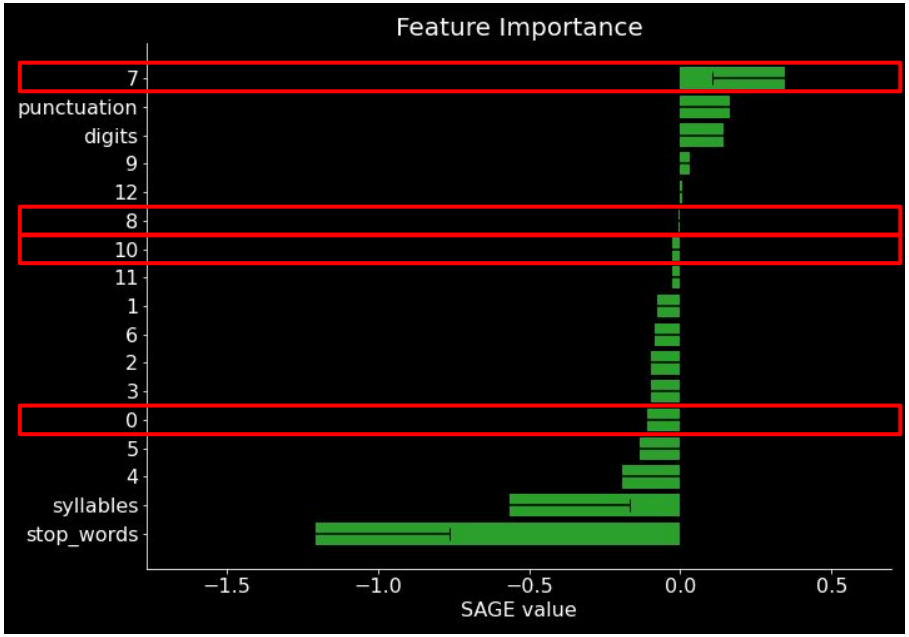
7: ['chill', 'passive', 'painful', 'historic', 'cure', 'base', ...]

8: ['netflix', 'potter', 'volvo', 'auto', 'minute', 'second', ...]

10: ['fish', 'soul', 'yoga', 'scientist', 'weapon', 'died', ...]

Sage for Target





1: [netflix, 'potter', 'volvo', 'auto', 'minute', 'second', ...]

3: ['in', 'or', 'into', 'were', 'during', 'against', ...]

4: ['!', '#', '(', ')', '@', '?', ...]

5: ['00', '27', '230', '730', '2014', '58', ...]

0: ['wonder', 'amazed', 'pro', 'folks', 'band', 'lineup', ...]

7: ['chill', 'passive', 'painful', 'historic', 'cure', 'base', ...]

8: ['netflix', 'potter', 'volvo', 'auto', 'minute', 'second', ...]

10: ['fish', 'soul', 'yoga', 'scientist', 'weapon', 'died', ...]

Challenges and Future Research

- Challenges
  - Global explanations on NLP
  - Data collection
- Achievements
  - Decent accuracy
  - Simple explanations
- Future Research
  - In depth analysis of explanations
  - Use different methods (XAI, Clustering)

References - Interesting libraries

- Sage: https://github.com/iancovert/sage
- Captum: https://captum.ai/
- Ray Tune: https://docs.ray.io/
- Pytorch Lighting: https://pytorch-lightning.readthedocs.io/
- Crowdtangle: https://www.crowdtangle.com/

Questions Time!!

References

- [1] Boberg et al. “Pandemic Populism: Facebook Pages of alternative news media and the corona crisis - a computational content analysis”. 2020. https://arxiv.org/pdf/2004.02566.pdf
- [2] Hardalov et al. “Cross-Domain Label-Adaptive Stance Detection”. 2021. https://arxiv.org/pdf/2104.07467.pdf
- [3] Covert et al. “Understanding Global Feature Contributions With Additive Importance Measures”. 2020. https://arxiv.org/abs/2004.00668

References

- [4] Lundberg et al. “A Unified Approach to Interpreting Model Predictions”. 2017. https://arxiv.org/abs/1705.07874
- [5] Kim et al. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. 2018. https://arxiv.org/pdf/1711.11279.pdf
- [6] Dhamdhere et al. “How important is a neuron?”. 2018. https://arxiv.org/abs/1805.12233
- [7] Facebook. “CrowdTangle”. 2021. https://www.crowdtangle.com/

References

- [8] Sanh et al. “DistilBert, a distilled version of BERT: smaller, faster, cheaper and lighter”. 2020. https://arxiv.org/abs/1910.01108
- [9] Mohammad et al. “SemEval-2016 Task 6: Detecting Stance in Tweets”. 2016. https://aclanthology.org/S16-1003.pdf
- [10] Covert. “Explaining machine learning models with SHAP and SAGE”. 2020. https://iancovert.com/blog/understanding-shap-sage/
- [11] Mohammad et al. “The SemEval-2016 Stance Dataset”. 2019. **http://www.saifmohammad.com/WebPages/StanceDataset.htm**