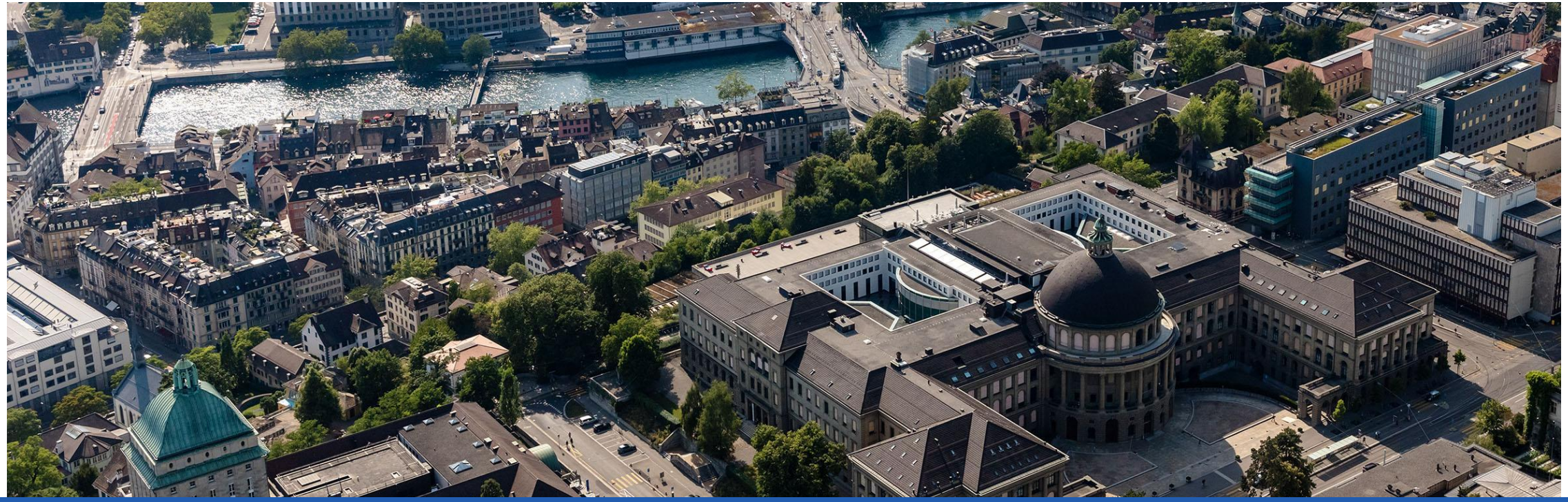
ETH zürich

13.6.2022



Controllable & Conditional Clickbait Title Generation and Detection [NLP & LSS Course project]

Student: Dusan Svilarkovic
Advisors: Dominik Stammbach, Afra Amini, Elliott Ash

Motivation

- **Recent developments of LLMs brought to ethical debates on open usage of models.**
- Inherent bias and toxicity of data & models has been exploited by fine-tuning for adversarial LLMs like **Grover** (2019) and recently **GPT-4chan** (2022)**.
- **Automated clickbait title generation will eventually be a reality**

* Defending Against Neural Fake News (2019), Zellers et al. https://arxiv.org/abs/1905.12616
** GPT-4chan: This is the worst AI ever, https://www.youtube.com/watch?v=efPrtcLdcdM, https://gpt-4chan.com/

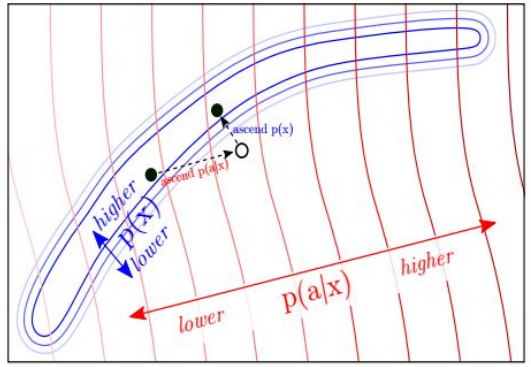
Motivation

- Recent developments of LLMs brought to ethical debates on open usage of models.
- Inherent bias and toxicity of data & models has been exploited by fine-tuning for adversarial LLMs like **Grover** (2019) and recently **GPT-4chan** (2022)**.
- Automated clickbait title generation will eventually be a reality
- **Can we create a classifier able to detect them?**
- **Can we create a tunable clickbait title generation?**
- **Can we make adversarial learning approach to further improve them**

* Defending Against Neural Fake News (2019), Zellers et al. https://arxiv.org/abs/1905.12616
** GPT-4chan: This is the worst AI ever, https://www.youtube.com/watch?v=efPrtcLdcdM, https://gpt-4chan.com/

Previous work

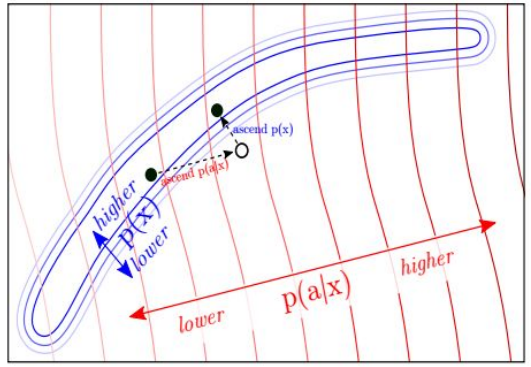
- **PPLM (2019) -** using gradient ascent for controllable text generation with frozen text generation model GPT-2 (P(x)) and BERT discriminator (P(a | X)).



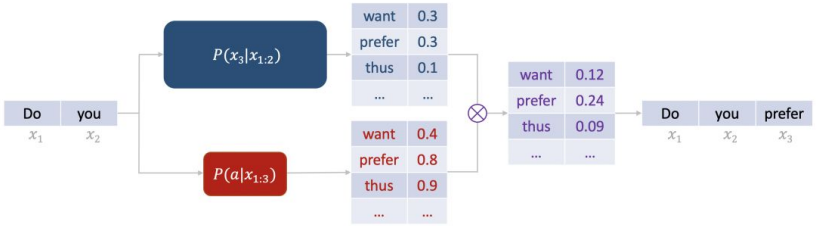
PPLM (2019)

Previous work

- **PPLM (2019) -** using gradient ascent for controllable text generation with frozen text generation model GPT-2 (P(x)) and BERT discriminator (P(a | X)).
- **Fudge (2021) -** using probability multiplication for decoding steps on controllable text generation with frozen text generation GPT-2 (P(x)) and ML text classifier.



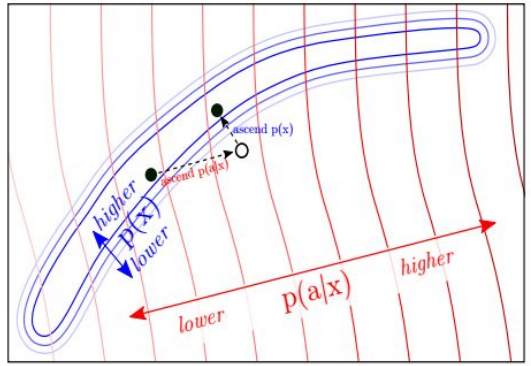
PPLM (2019)



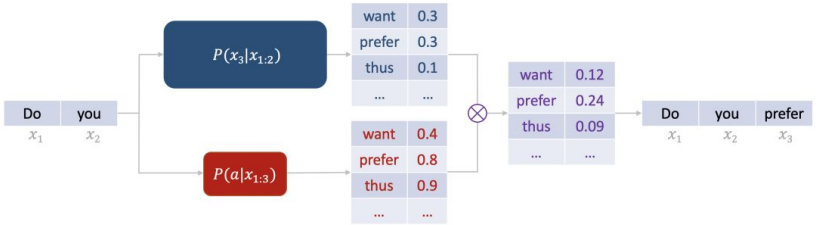
Fudge (2021)

Previous work

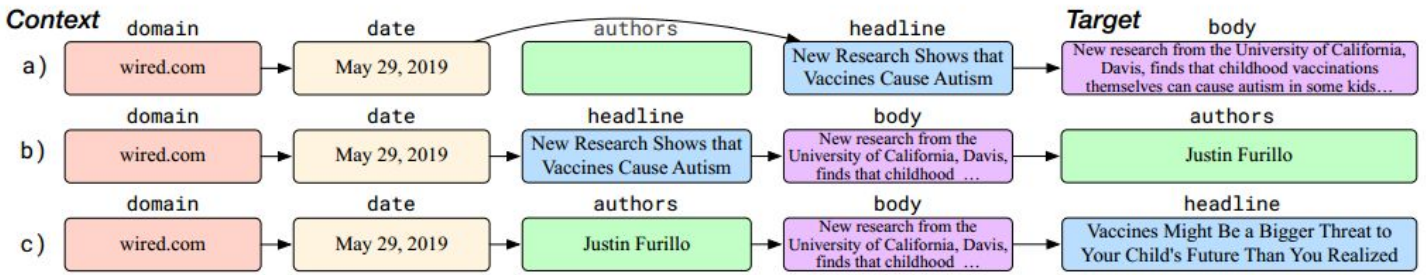
- **PPLM (2019) -** using gradient ascent for controllable text generation with frozen text generation model GPT-2 (P(x)) and BERT discriminator (P(a | X)).
- **Fudge (2021) -** using probability multiplication for decoding steps on controllable text generation with frozen text generation GPT-2 (P(x)) and ML text classifier.
- **Grover (2019) -** fake news generation model based on GPT-2 for generating articles with meta-attributes.



PPLM (2019)



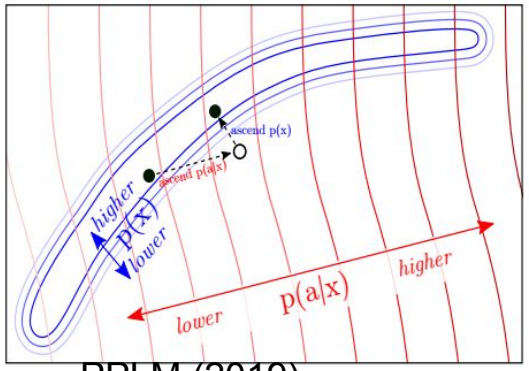
Fudge (2021)



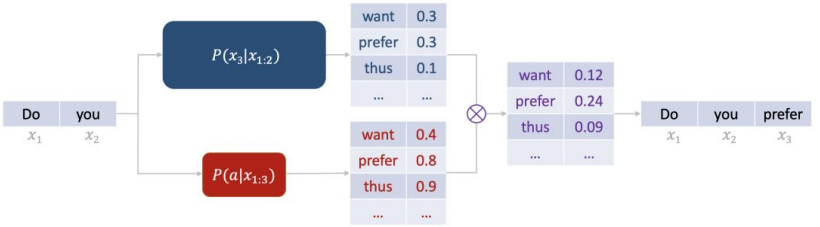
Grover (2019)

Previous work - Trade-off

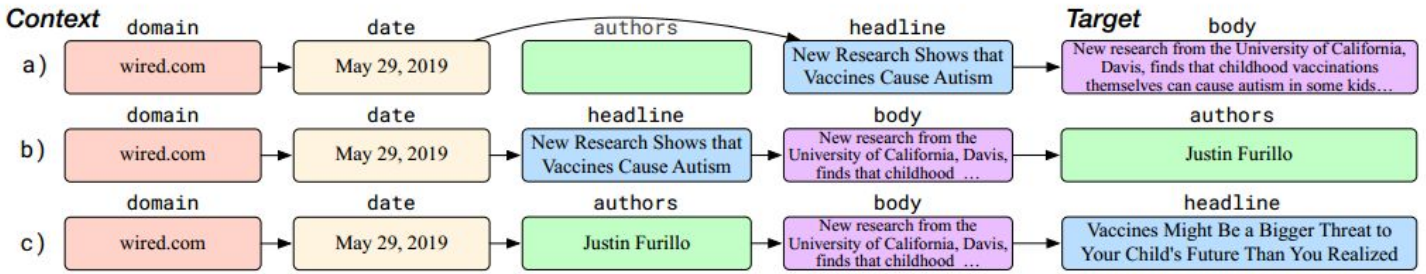
- **PPLM (2019) [Tunable attribute classifier, no conditioning]**- using gradient ascent for controllable text generation with frozen text generation model GPT-2 (P(x)) and BERT discriminator (P(a | X)).
- **Fudge (2021)[Tunable attribute classifier, no conditioning]** - using probability multiplication for decoding steps on controllable text generation with frozen text generation GPT-2 (P(x)) and ML text classifier.
- **Grover (2019) [Untunable, but conditioned]** - fake news generation model based on GPT-2 for generating articles with meta-attributes.



PPLM (2019)



Fudge (2021)



Grover (2019)

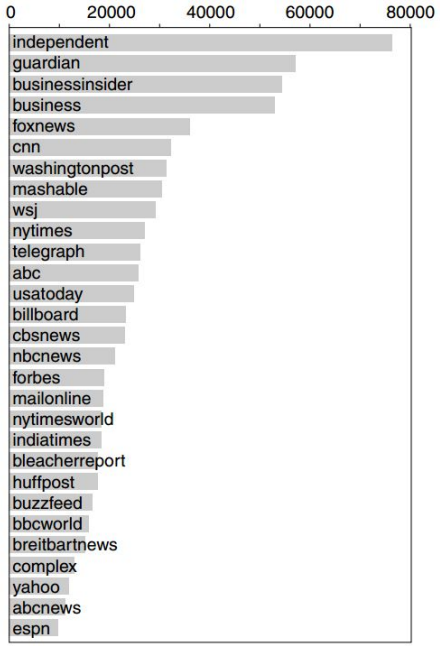
Previous work - Trade-off

- **PPLM (2019) [Tunable attribute classifier, no conditioning]-** using gradient ascent for controllable text generation with frozen text generation model GPT-2 (P(x)) and BERT discriminator (P(a | X)).
- **Fudge (2021)[Tunable attribute classifier, no conditioning] -** using probability multiplication for decoding steps on controllable text generation with frozen text generation GPT-2 (P(x)) and ML text classifier.
- **Grover (2019) [Untunable, but conditioned] -** fake news generation model based on GPT-2 for generating articles with meta-attributes.
- We need tunable and conditioned text generation!

Data Sources

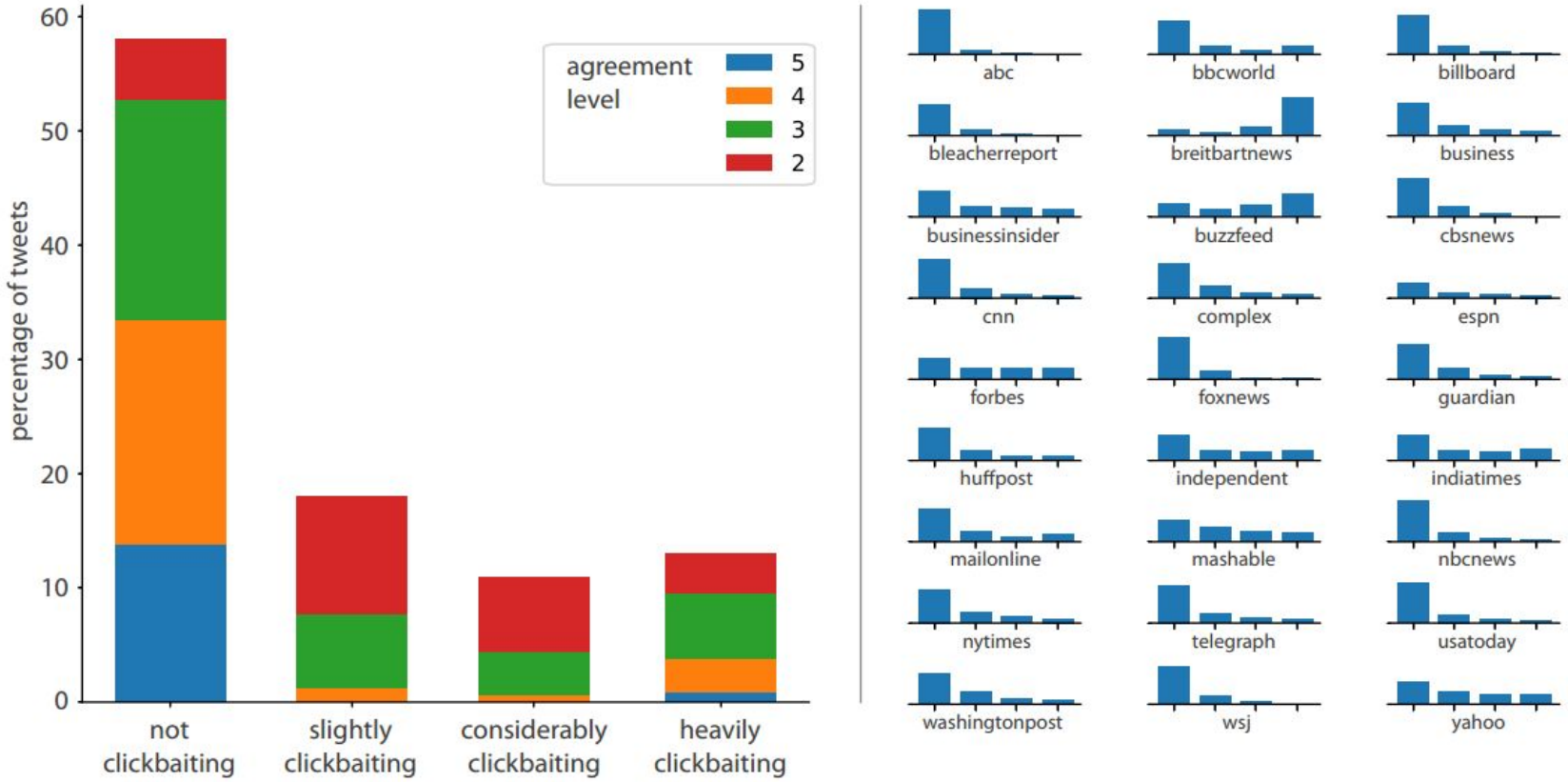
- Using Webis Clickbait 2017 Tweets Dataset (period Dec 1 2016 – Apr 30 2017)
- Graded labelling using **MTurk** workers and clickbait scales : 0.0, 0.33, 0.66, 1.0.
- Relevant attributes: **title**, **article content** and **tweet headline** and **clickbait score**.
- 27 publishers (abc, bbcworld, billboard, bleacherreport, breitbartnews, business,businessinsider, buzzfeed, cbsnews, cnn,complex, espn, forbes, foxnews, guardian,huffpost, independent, indiatimes, mailonline, mashable, nbcnews, nytimes,telegraph, usatoday, washingtonpost, wsj,yahoo)
- The most data
- Public leaderboard

Publication	Genre	Teaser	Acquisition	Annotation (Scale, People, κ)			Articles	Size
Agrawal (2016)	News	Headline	Gatekeeper-based	Binary	3	0.84	No	2,388
Potthast et al. (2016)	News	Tweet	Importance-based	Binary	3	0.35	Yes	2,992
Biyani et al. (2016)	News	Headline	Reputation-based?	Binary	?	?	?	4,073
Chakraborty et al. (2016)	News	Headline	Reputation-based	Binary	3	0.79	No	15,000
Rony et al. (2017)	News	Headline	Reputation-based	Binary	3	0.79?	No	32,000
Webis-Clickbait-17	News	Tweet	Importance-based	Graded	5	0.36	Yes	38,517



Crowdsourcing a Large Corpus of Clickbait on Twitter (2017), Potthast et al.

Data Sources - Ground truth overview



Crowdsourcing a Large Corpus of Clickbait on Twitter (2017), Potthast et al.

Methods for our approach

- Previous works used GPT-2 (decoder based) for non-conditional text generation.
- We generate **text-to-title summarizations**
- Text summarization : pre-trained **Pegasus**
- Clickbait classifier : fine-tuned **MPNet**
- **Approaches:**

Methods for our approach

- Previous works used GPT-2 (decoder based) for non-conditional text generation.
- We generate **text-to-title summarizations**
- Text summarization : pre-trained **Pegasus**
- Clickbait classifier : fine-tuned **MPNet**
- **Approaches:**
    1. Binary text-to-title clickbait generation:
        a. **Grover**-like **GPT-2**, **I/O** : (**article content**, **clickbait/non-clickbait title**)
        b. **Pegasus** text-to-title generation,
            - **I** : article content, **O** : clickbait title

Methods for our approach

- Previous works used GPT-2 (decoder based) for non-conditional text generation.
- We generate **text-to-title summarizations**
- Text summarization : pre-trained **Pegasus**
- Clickbait classifier : fine-tuned **MPNet**
- **Approaches:**
    1. Binary text-to-title clickbait generation:
        a. **Grover**-like **GPT-2**, **I/O** : (**article content**, **clickbait/non-clickbait title**)
        b. **Pegasus** text-to-title generation,
            - **I** : article content, **O** : clickbait title

    2. Tunable text-to-title clickbait generation:
        a. Modified **PPLM** with **GPT-2**->**Pegasus** and BERT/BoW classifier->MPNet
            - **I:** article content, clickbait grade in [0,1] range.
            - **O:** clickbait title
        b. Modified **Fudge** with **GPT-2**->**Pegasus** and classifier->MPNet
            - **I:** article content, **condition lambda** in [1,100] range.
            - **O:** clickbait title

Preliminary results

Clickbait classifier (MPNet):

F1 : 0.678 (2nd), **Precision:** 0.707 (3rd), **Recall** : 0.65(2nd), **Accuracy** : **0.853** (close 1st)

Leaderboard

VIRTUAL MACHINE	RUN	MEAN SQUARED ERROR	MEDIAN ABSOLUTE ERROR	F1 SCORE	PRECISION	RECALL	ACCURACY	NORMALIZED MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR	EXPLAINED VARIANCE
whitebait	2017-08-17-09-48-47	0.043	0.139	0.565	0.699	0.474	0.826	0.583	0.165	0.422
zingel	2017-08-29-23-13-43	0.033	0.131	0.683	0.719	0.65	0.856	0.453	0.149	0.566
arowana	2017-08-31-14-51-47	0.039	0.141	0.656	0.659	0.654	0.837	0.532	0.161	0.526
houndshark	2017-08-31-21-22-14	0.108	0.173	0.451	0.478	0.427	0.753	1.464	0.239	-0.432
pike	2017-09-01-04-00-19	0.045	0.109	0.604	0.711	0.524	0.836	0.607	0.156	0.413
houndshark	2017-09-01-08-35-31	0.099	0.321	0.023	0.779	0.012	0.764	1.353	0.282	0.003
torpedo	2017-09-01-08-47-33	0.079	0.236	0.65	0.53	0.841	0.785	1.077	0.241	0.346

Classification report

	precision	recall	f1-score	support
0	0.89	0.92	0.90	14464
1	0.71	0.65	0.68	4515
accuracy			0.85	18979
macro avg	0.80	0.78	0.79	18979
weighted avg	0.85	0.85	0.85	18979

Preliminary results

Clickbait classifier (MPNet):

F1 : 0.678 (2nd), **Precision:** 0.707 (3rd), **Recall** : 0.65(2nd), **Accuracy** : **0.853** (close 1st)

Leaderboard

VIRTUAL MACHINE	RUN	MEAN SQUARED ERROR	MEDIAN ABSOLUTE ERROR	F1 SCORE	PRECISION	RECALL	ACCURACY	NORMALIZED MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR	EXPLAINED VARIANCE
whitebait	2017-08-17-09-48-47	0.043	0.139	0.565	0.699	0.474	0.826	0.583	0.165	0.422
zingel	2017-08-29-23-13-43	0.033	0.131	0.683	0.719	0.65	0.856	0.453	0.149	0.566
arowana	2017-08-31-14-51-47	0.039	0.141	0.656	0.659	0.654	0.837	0.532	0.161	0.526
houndshark	2017-08-31-21-22-14	0.108	0.173	0.451	0.478	0.427	0.753	1.464	0.239	-0.432
pike	2017-09-01-04-00-19	0.045	0.109	0.604	0.711	0.524	0.836	0.607	0.156	0.413
houndshark	2017-09-01-08-35-31	0.099	0.321	0.023	0.779	0.012	0.764	1.353	0.282	0.003
torpedo	2017-09-01-08-47-33	0.079	0.236	0.65	0.53	0.841	0.785	1.077	0.241	0.346

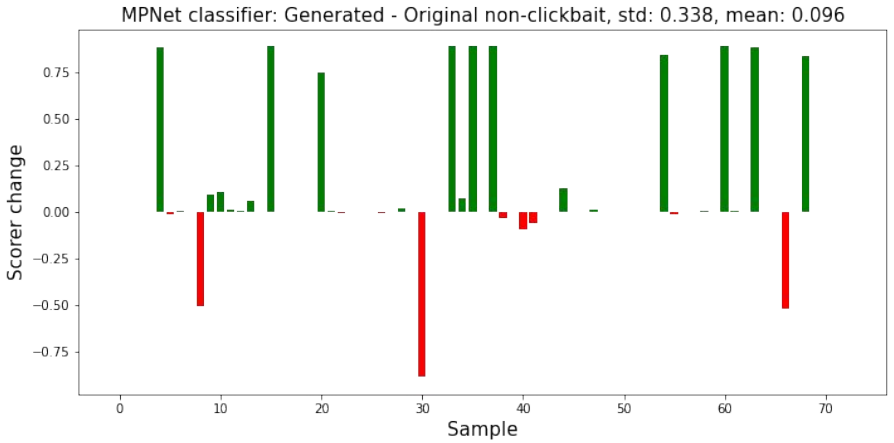
Classification report

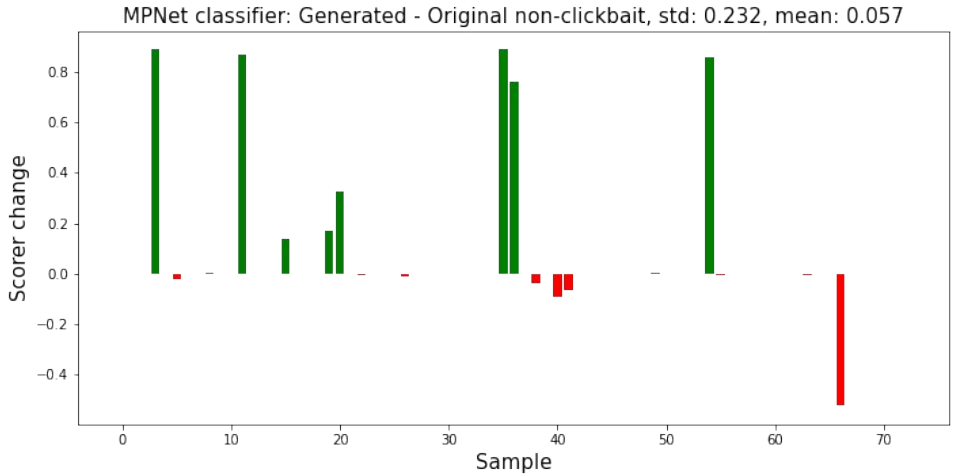
	precision	recall	f1-score	support
0	0.89	0.92	0.90	14464
1	0.71	0.65	0.68	4515
accuracy			0.85	18979
macro avg	0.80	0.78	0.79	18979
weighted avg	0.85	0.85	0.85	18979

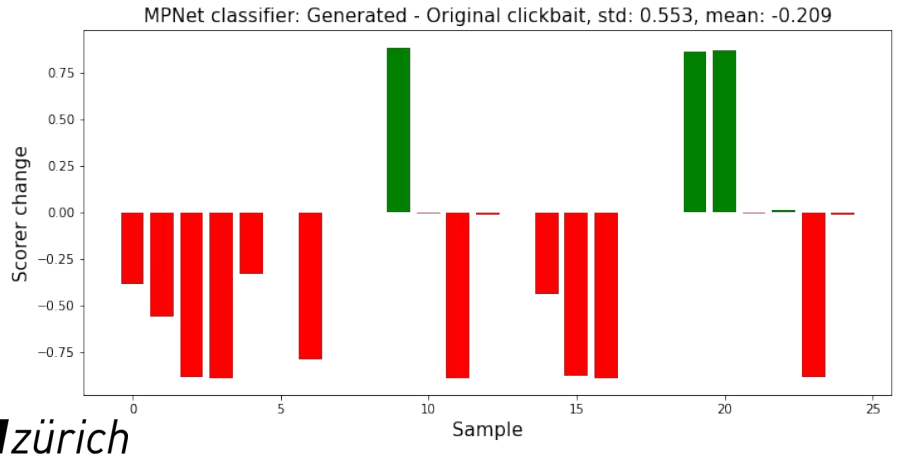
Preliminary results

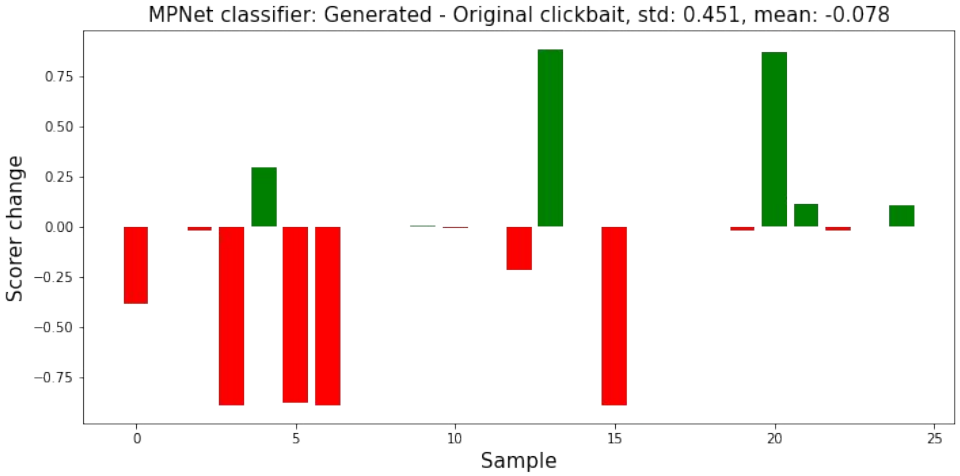
Pegasus (2020): Binary summarization

- better titles when conditioning on clickbait title only
- could be because Pegasus starts capturing more news-like titles





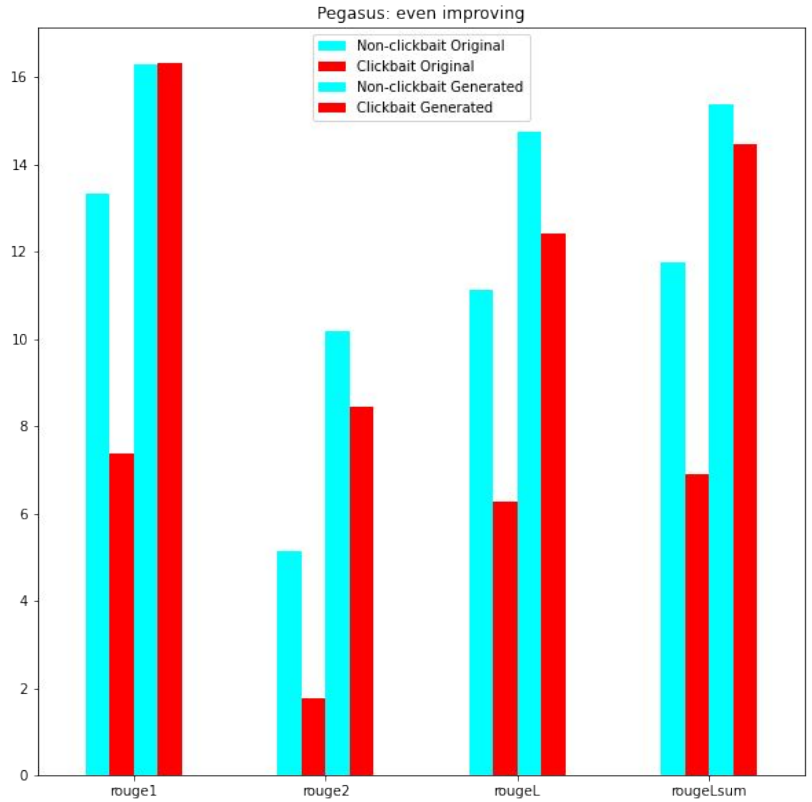




Preliminary results

Pegasus (2020): Binary summarization

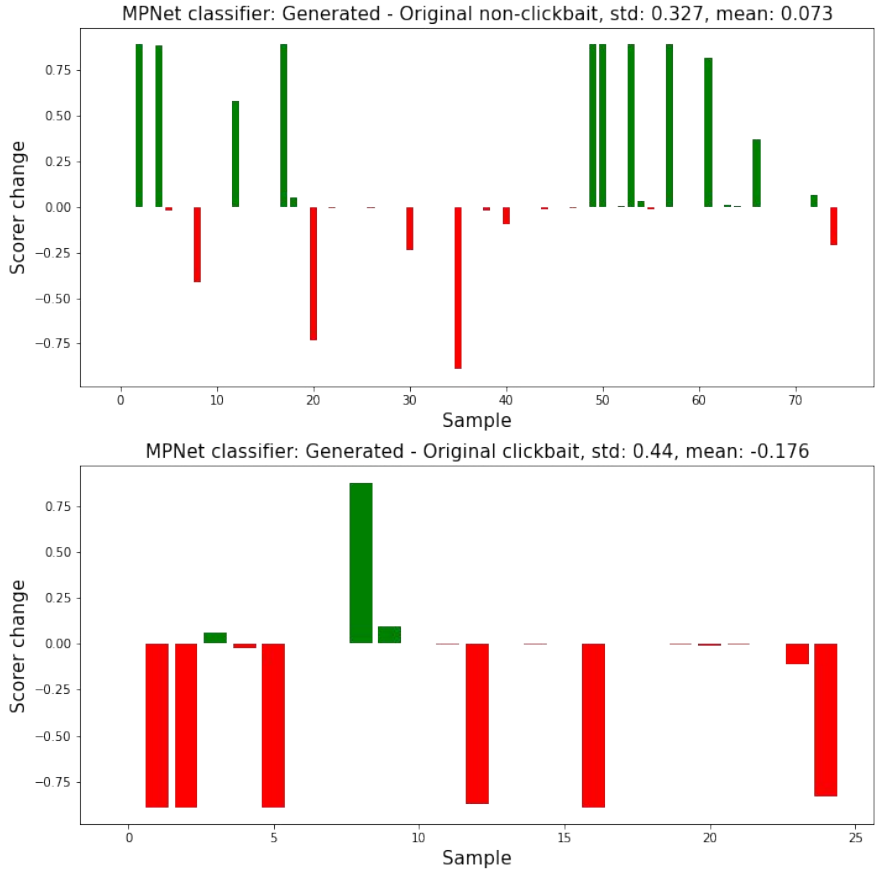
- better titles when conditioning on clickbait title only
- could be because Pegasus starts capturing more news-like titles



Preliminary results

Fudge (2021)

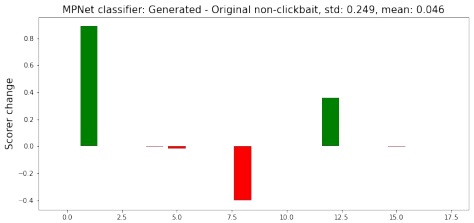
- for conditional lambda = 1.0, solid push to clickbaity titles

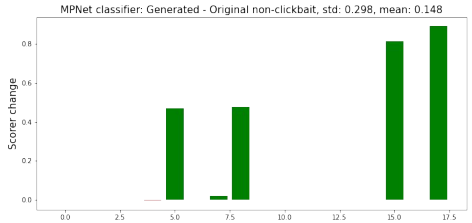


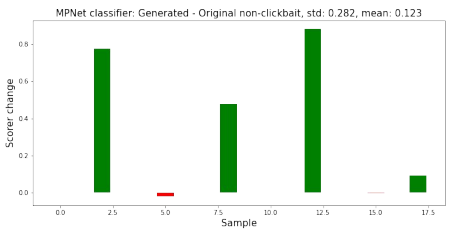
Preliminary results

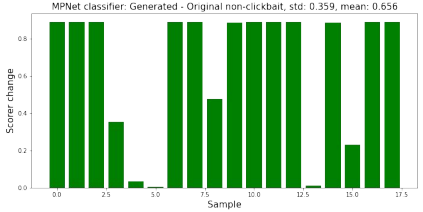
Fudge (2021) - Controllability

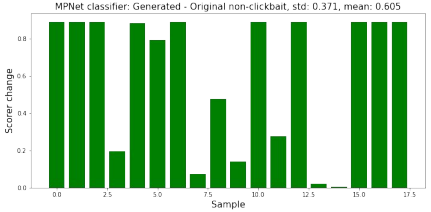
- increasing condition lambda increases clickbaitiness of titles
- need hyperparameter search for condition_lambda/clickbaitiness correlation

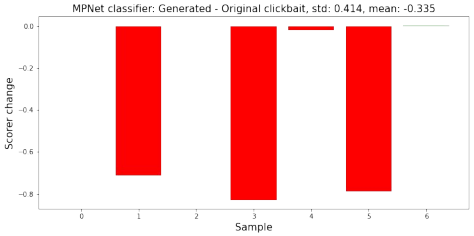


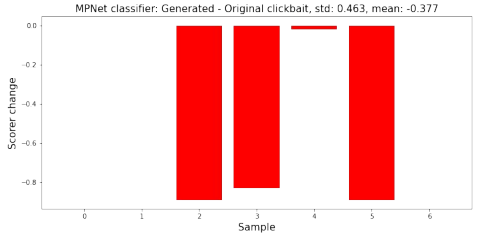


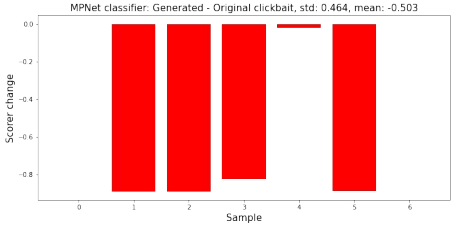


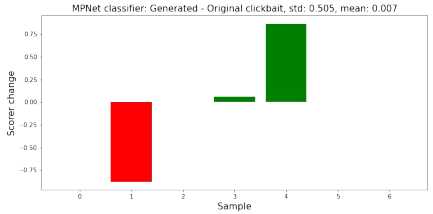


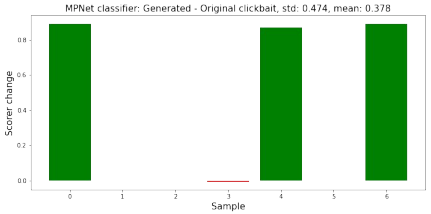










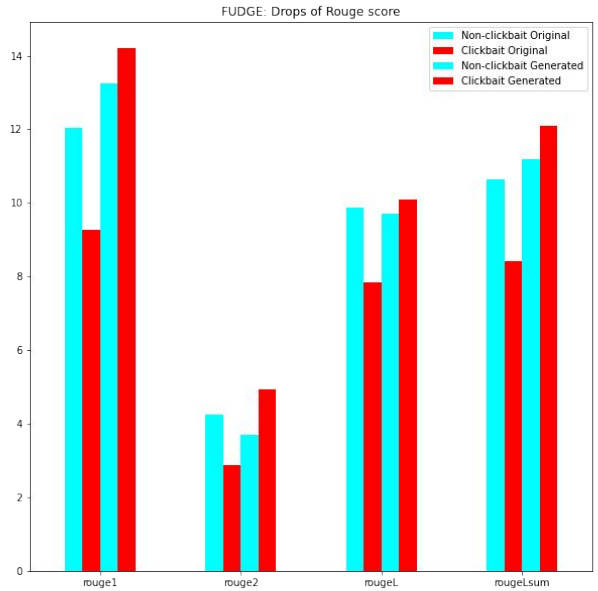


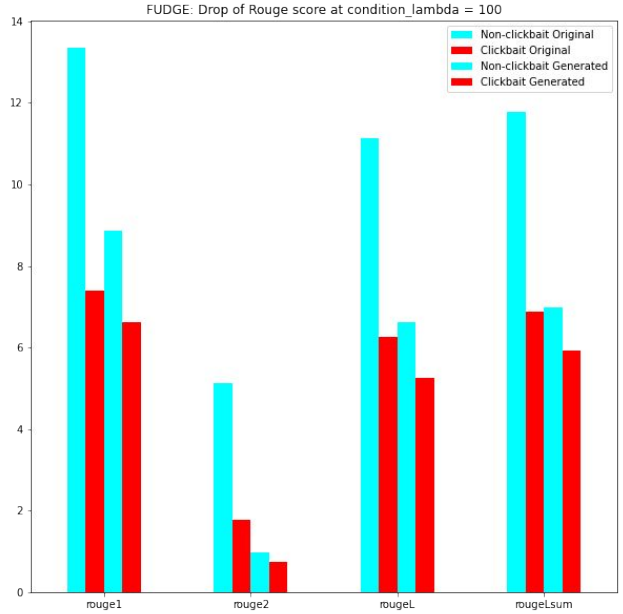
MPNet's difference in Generated - Original title score.
Top is originally non-clickbaity, bottom is already clickbaity.
From left to right : condition_lambda set to [0.0, 1.0, 5.0, 50.0, 100.0]

Preliminary results

Fudge (2021) - Controllability

- increasing condition lambda increases clickbaitiness of titles
- need hyperparameter search for condition_lambda/clickbaitiness correlation



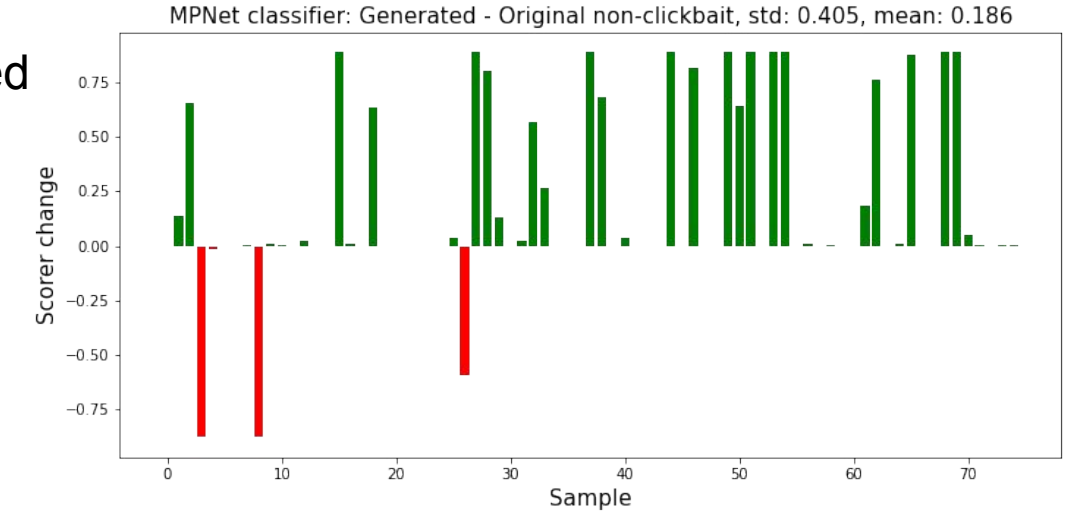


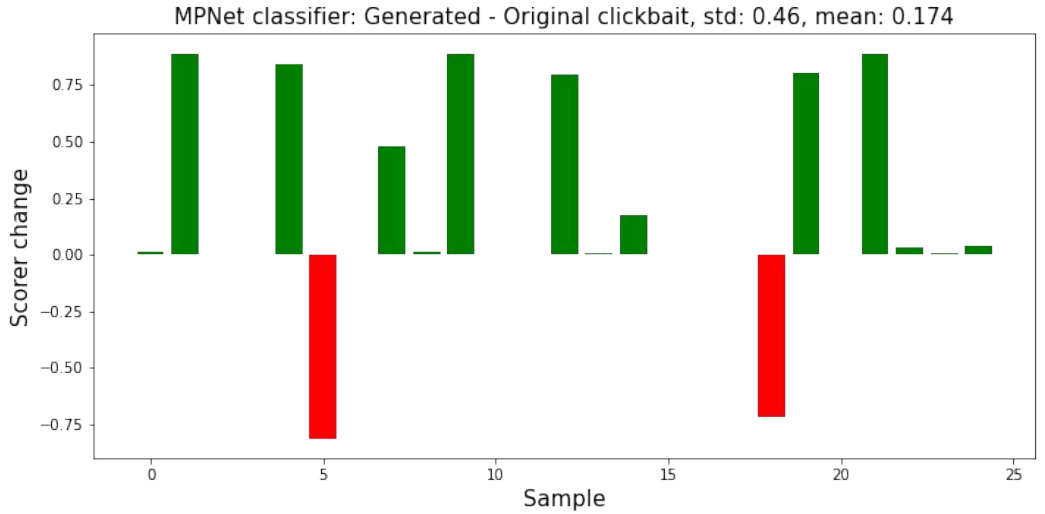
Rouge scores when condition_lambda = 1.0 vs when condition_lambda = 100.0

Changes over time

PPLM (2019):

- with class_label pushed to “clickbait”, works as intended
- attribute control **P(a | X)** not reacting properly
- needs hyperparameter search for better control

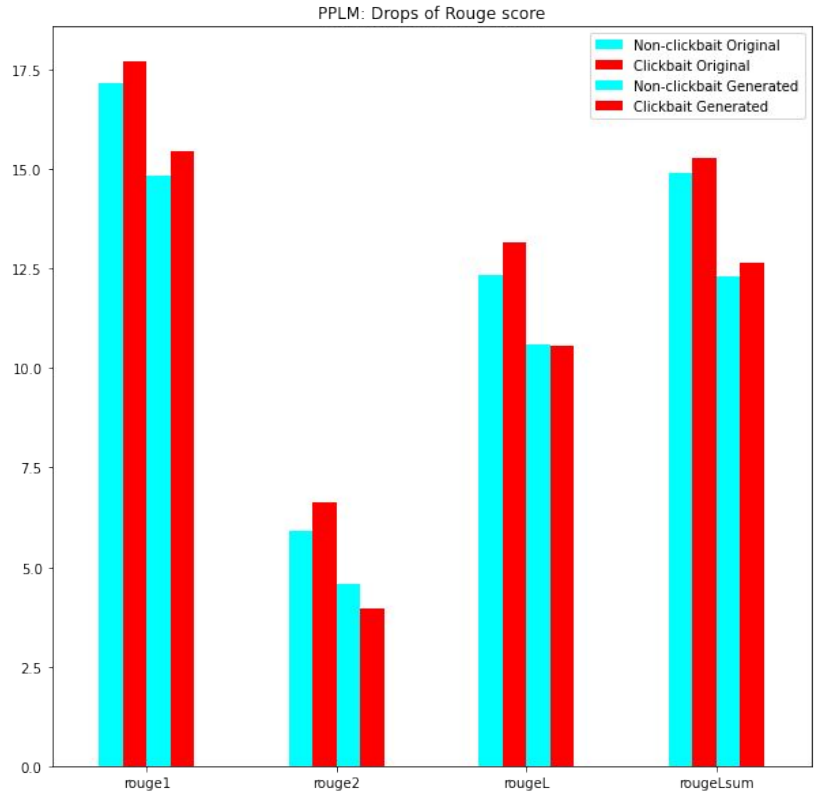


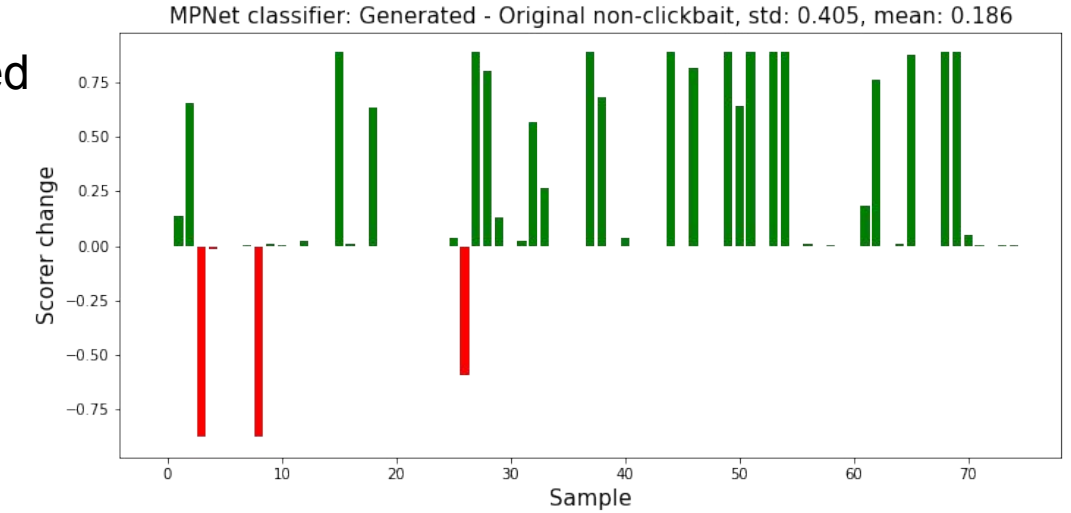


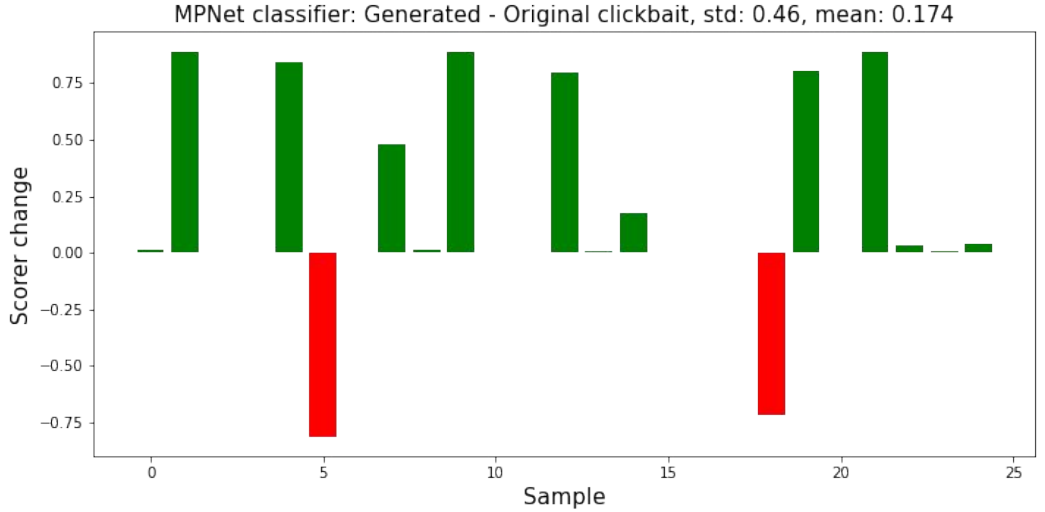
Changes over time

PPLM (2019):

- with class_label pushed to “clickbait”, works as intended
- attribute control **P(a | X)** not reacting properly
- needs hyperparameter search for better control







Timeline

- Conditional text generation needs further fine-tuning for better control
- Hyperparameter search focus in the upcoming period
- Generation of titles is slow (100 samples in PPLM in 3h)!
- **Political alignment/Topic control?**
- Estimate end of experiments and raw draft start: end of June

ETH zürich

FIN !

Questions?

Additional dataset information

Corpus Acquisition	
Platform:	Twitter
Crawling period:	Dec 1 2016 – Apr 30 2017
Crawled tweets:	459,541
Publishers:	27. Names: abc, bbcworld, billboard, bleacherreport, breitbartnews, business, businessinsider, buzzfeed, cbsnews, cnn, complex, espn, forbes, foxnews, guardian, huffpost, independent, indiatimes, mailonline, mashable, nbcnews, nytimes, telegraph, usatoday, washingtonpost, wsj, yahoo
Filters:	- No videos in tweets. - Exactly one hyperlink in tweet. - Article archiving succeeded. - Main content extraction succeeded.
Recorded fields:	12. Names: postId, postTimestamp, postText, postMedia, postPublisher, targetUrl, targetTitle, targetDescription, targetKeywords, targetParagraphs, targetCaptions, targetWarcArchive
Sampling strategy:	Maximally 10 tweets per day and publisher
Sampled tweets:	38,517

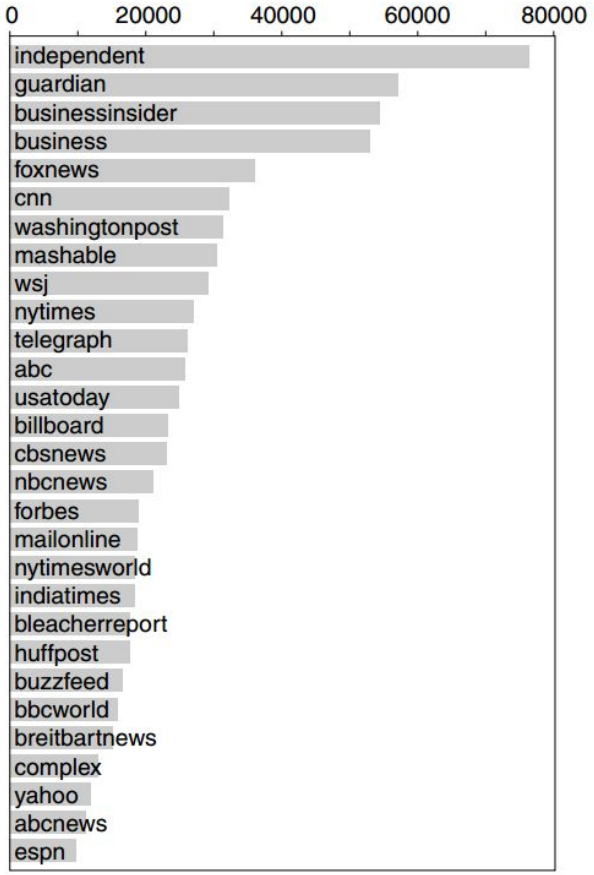


Table 2: Corpus acquisition overview (left), and number of tweets crawled from every publisher (right).