P5 Final Presentation
UID & Authorship Obfuscation

Nick Abegg

Research Question

Explore the applicability of UID metric in the task of obfuscation. Specifically investigate if UID can be used as a guiding metric to result in successful obfuscation in which an automated authorship attributor misattributes an obfuscated article.

Intro

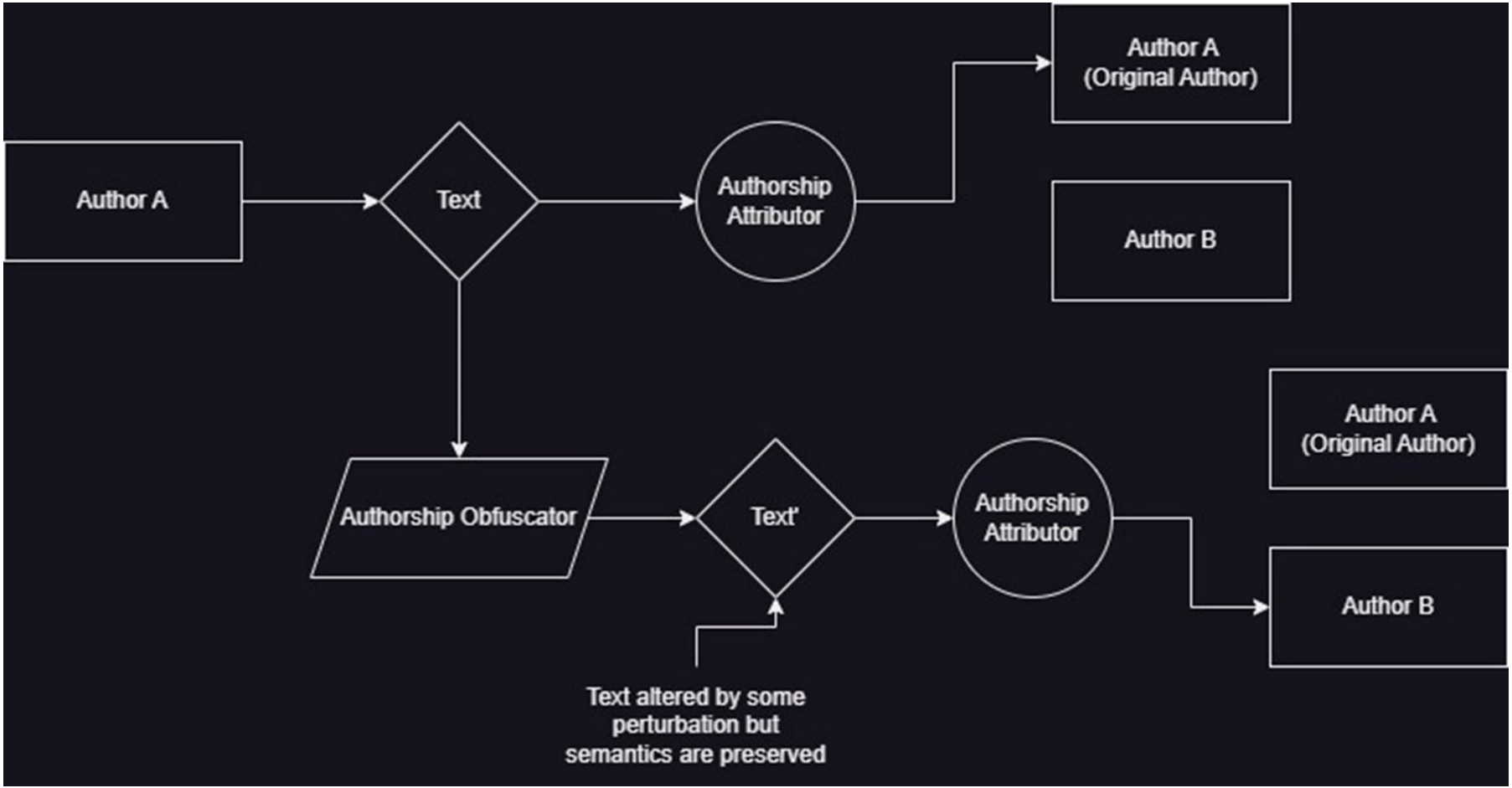
- **Authorship Obfuscation** vs. Authorship Attribution
- Psycholinguistic theory/UID
- Synonym Swap Review

Authorship Attribution and Obfuscation

Authorship Attribution - Process of taking a text of unknown authorship and attributing the authorship amongst a number of known authors

Authorship Obfuscation - Take some text and obscure the original author through some perturbation.

Authorship Obfuscation is successful if it is capable of deceiving an AA into attributing the incorrect author of a text


Author A
Text
Authorship Attributor
Author A (Original Author)
Author B
Authorship Obfuscator
Text'
Authorship Attributor
Author A (Original Author)
Author B
Text altered by some perturbation but semantics are preserved

Psycholinguistic/UID

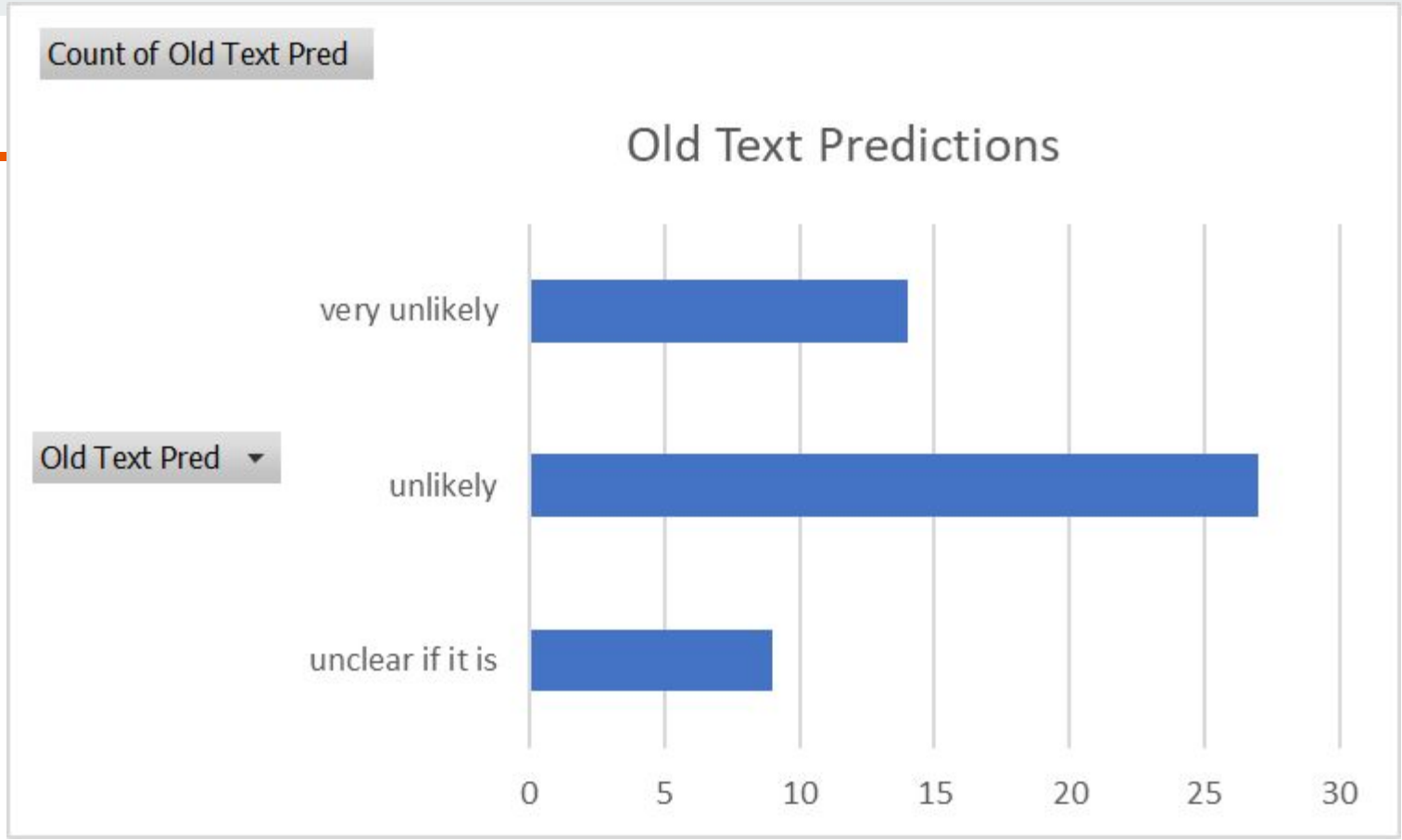
Uniform Information Density (UID) Theory - suggest that humans optimize their speech and text so as to uniformly distribute information over an entire message (Frank and Jaeger [2008])

- Do so in effort to maximize efficiency in communication
- Both humans and language models follow UID patterns which can be quantified in a variety of ways (Venkatraman et al. [2023]), (Meister et al. [2021]).
- Possibly altering the UID score of an article could lead to successful obfuscation

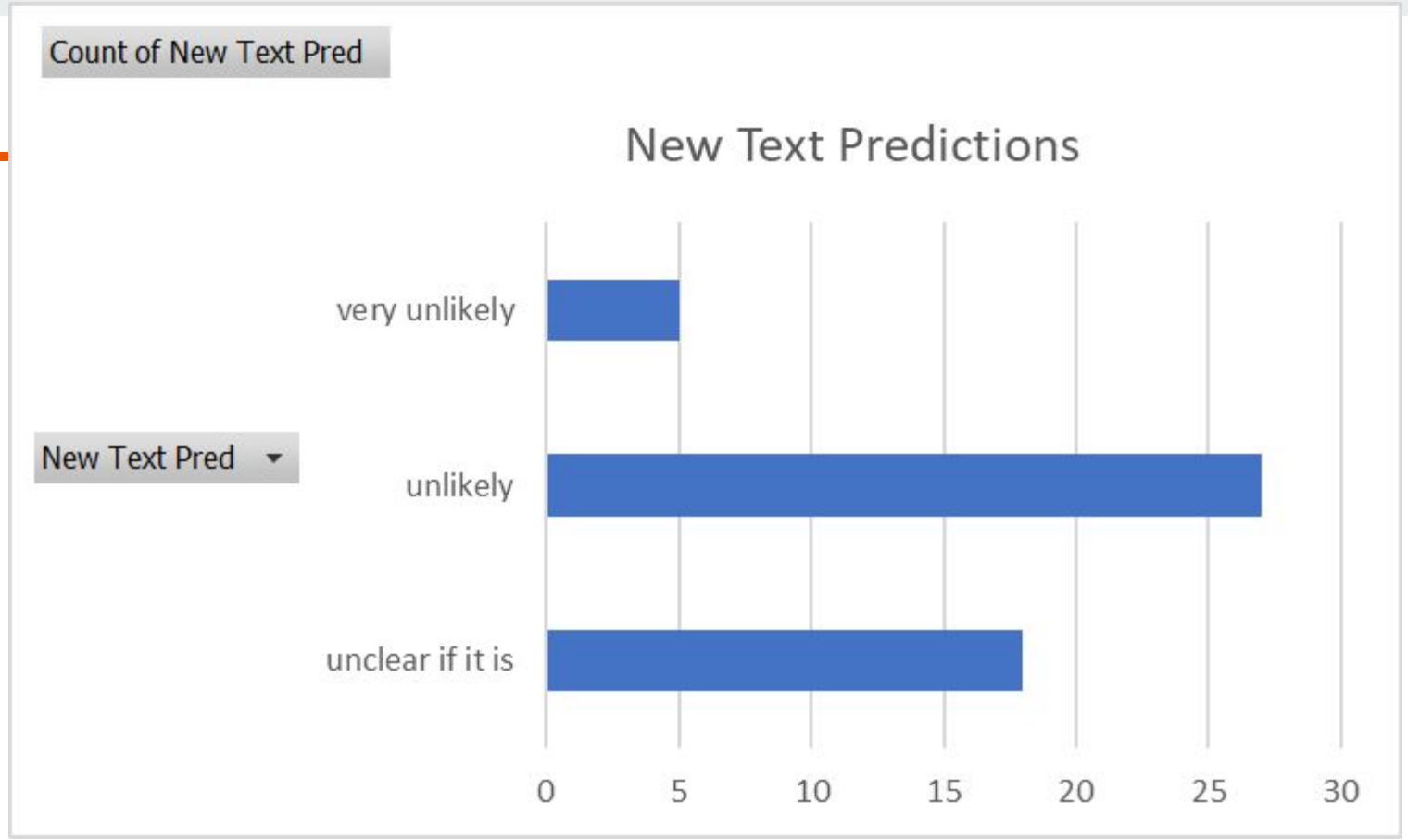
Synonym Swap

- First algorithm developed, very simple
- Find most likely synonyms for a particular word in a sentence, swap most probable according to GPT-2 Language Model
- Demonstrated that even a small change to an article could throw off an attributor's labels
- Quality of swaps were poor as a result of GPT-2 and Wordnet

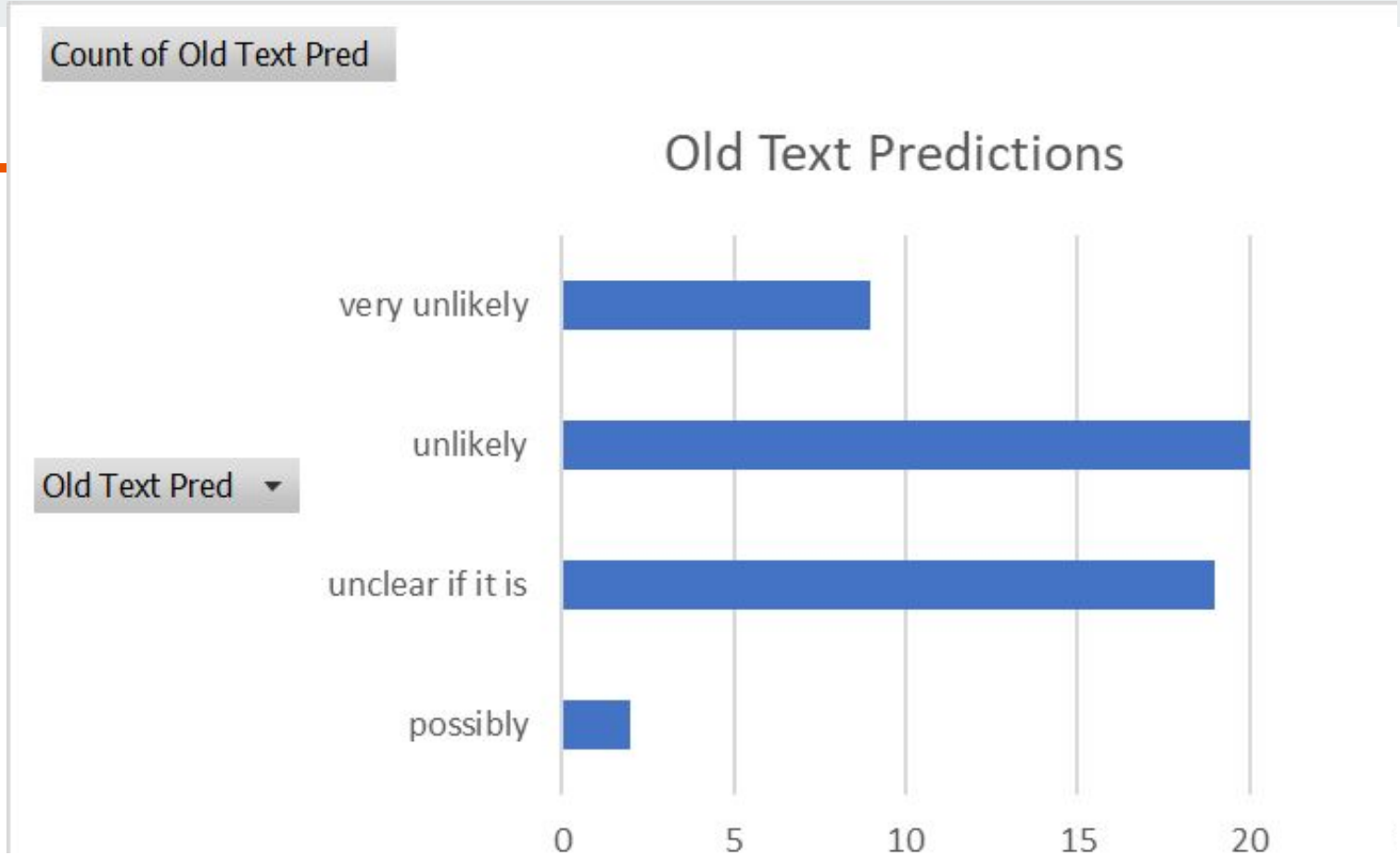
Unaltered Human Generated Text



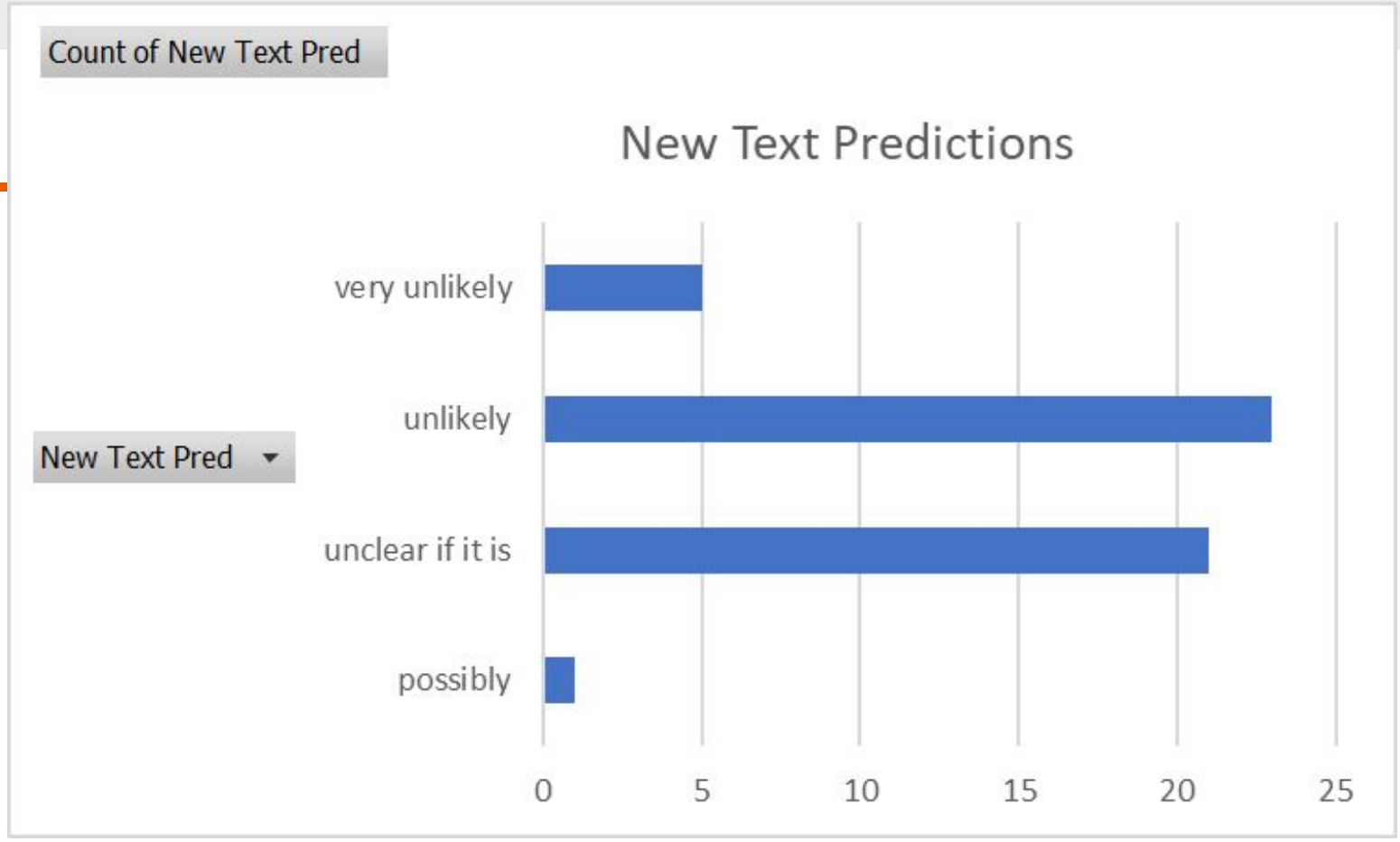
Altered Human Generated Text



Unaltered Machine Generated Text

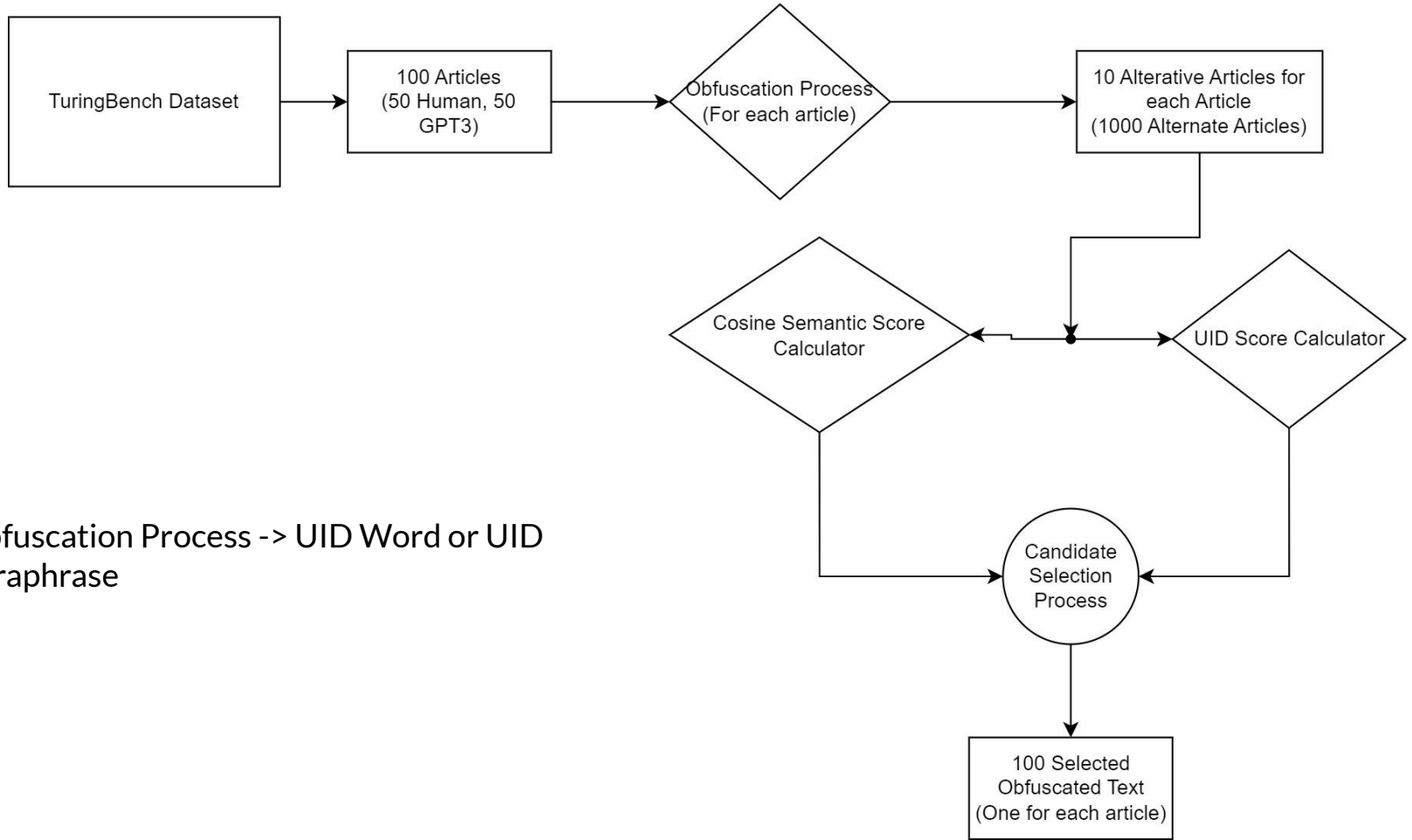


Altered Machine Generated Text



Problems with Synonym Swap

- WordNet does not account for context of the sentence
- GPT-2 cannot examine the tokens after the target word and thus, misses much needed context
- Only generated 1 new sentence based on the most probable token
- Does not account for UID or semantic similarity



Obfuscation Process -> UID Word or UID Paraphrase

TuringBench

- Dataset containing 200,000 articles human and machine-generated text
- The authors consisted of 19 language models and human authored text
- Language models: GPT-1, GPT-2_small, GPT-2_medium, GPT-2_large, GPT-2_xl, GPT-2_PyTorch, GPT-3, GROVER_base, GROVER_large, GROVER_mega, CTRL, XLM, XLNET_base, XLNET_large, FAIR_wmt19, FAIR_wmt20, TRANSFORMER_XL, PPLM_distil, PPLM_gpt2
- Human text from The Washington Post, CNN, Breitbart
- Pooled a small subset of articles from dataset, 50 human, 50 GPT-3

(Uchendu et al. [2021]).

Cosine Semantic Similarity

- Using scikit-learn, we calculate the cosine semantic similarity
- Calculated for each alternate article compared to the original article
- The cosine similarity between two vectors A and B is given by the formula:

cosine_similarity = dot_product(A, B) / (norm(A) * norm(B))

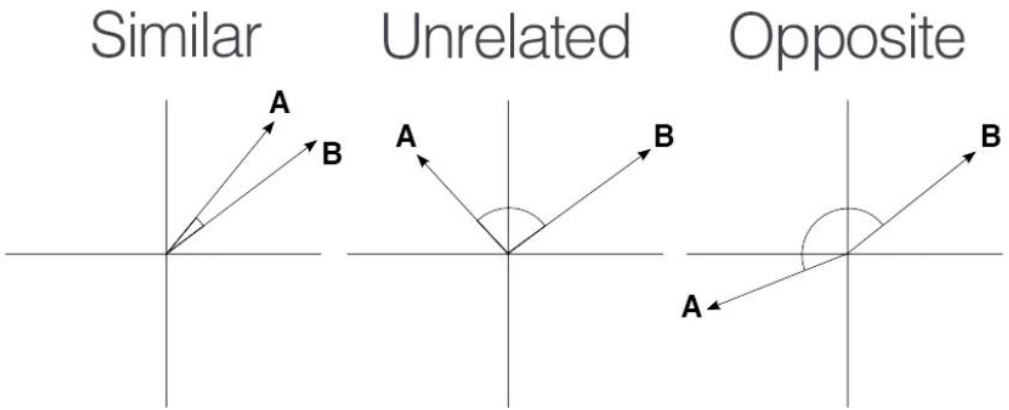


Image: https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded

UID Score

For each article and its alternates we also calculated the UID Scores

- UID Score 1 -> UID Score(Variance)
- UID Score 3 -> UID Score(Difference^2)

For both, need surprisal -> the negative log probability of a token given the previous tokens

Variance - obtained by calculating the variance of the surprisals for all tokens within the article.

Difference^2 - calculate the average of the squared differences in surprisals for every two consecutive tokens within the article.

UID Word Swap (UWS)

- UWS was created first, hoped to iterate and improve on Synonym Swap
- UWS utilized DistilBERT in a Masked Language Model approach to conduct the word swap
- Should address both the inability of GPT-2 to have context of tokens later in the sentence and Wordnet's poor synonym selection

The man cried out in pain

Synonym Swap:
The man [MASKED] ~~out in pain~~

The man wept…
The man sobbed…
The man grieved…

UID Word Swap:
The man [MASKED] out in pain

The man yelled…
The man screamed…
The man howled…
The man shrieked…

UID Word Swap

```python
def bert_swap(masked):

    # get tokens for masked sentence
    inputs = tokenizer(masked, return_tensors="pt")

    # calculates the probability of the sentence
    with torch.no_grad():
        logits = model(**inputs).logits

    # retrieve index of [MASK]
    mask_token_index = int((inputs.input_ids == tokenizer.mask_token_id)[0].nonzero(as_tuple=True)[0])

    # get 10 most probable tokens
    most_prob_tokens = torch.argsort(logits[0,mask_token_index])[-10:]

    # decode those tokens
    most_prob_words = tokenizer.decode(most_prob_tokens)

    most_prob_words = most_prob_words.split()
    prob_words_list.append(most_prob_words)

    most_prob_sentences = []

    # insert the 10 most probable tokens into 10 seperate sentences
    for i in range(len(most_prob_words)):
        new_sentence = masked.replace('[MASK]', most_prob_words[i])
        most_prob_sentences.append(new_sentence)

    return most_prob_sentences
```

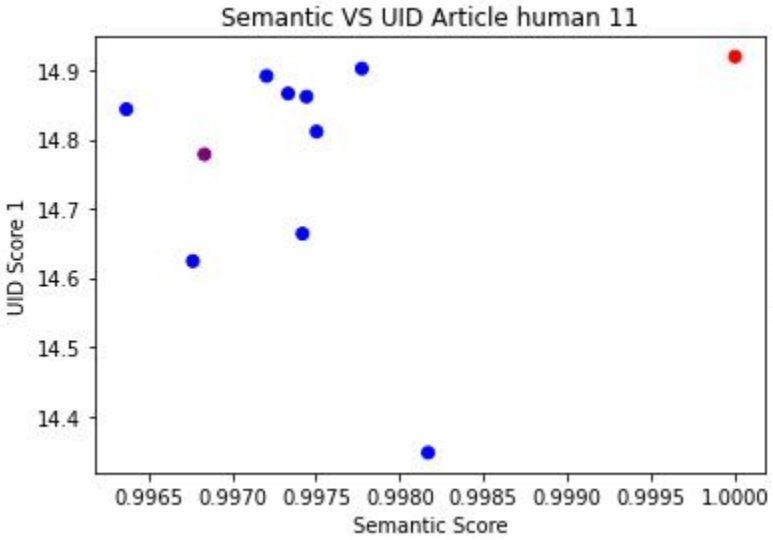
UID Word Swap Example

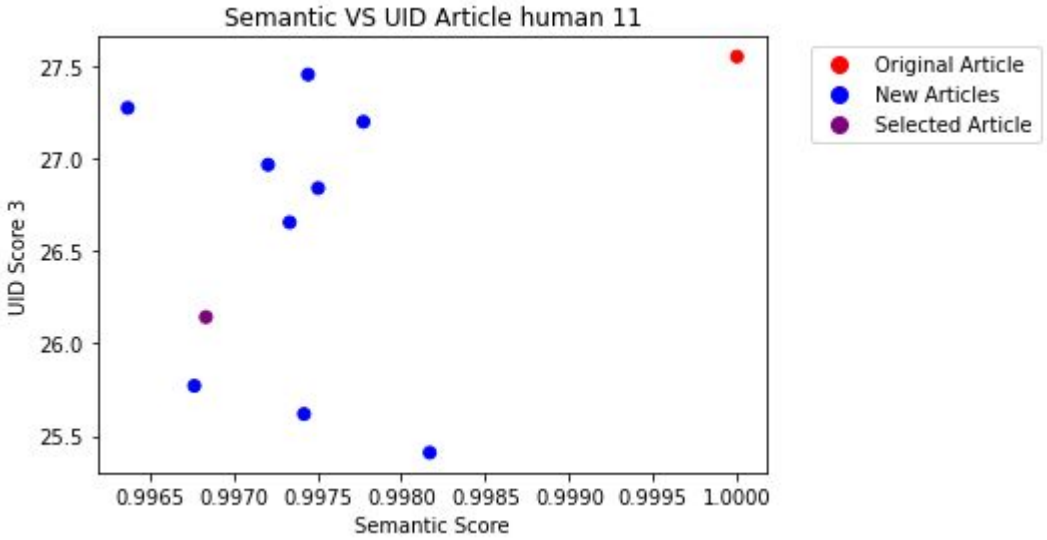
washington ( ) at a time when president donald trump seems to permeate nearly every aspect of american discourse, it might come as surprise that the first movie from barack and michelle obama's production company, higher ground, never mentions him by name.but subtlety is part power factory, new netflix documentary charts r ening factory in dayton, ohio. over course two hours, movie, directed seasoned documentarians steven bognar julia reichert, serves quiet historical political corrective, offering their portrait state america's industrial heartland prodding viewers rethink who, exactly, project.american starts on december 23, 2008, crowd gathers learn general motors plant dayton has shuttered. then fast-forwards 2015, ens enterprise, fuyao glass america, arm shanghai-based company manufactures automotive glass. one man makes fuyao's expanded mission crystal clear: what we're doing melding cultures together: chinese culture s culture. so we are truly global organization.as some critics have pointed out, is, important ways, commentary unpredictability globalization; york times review frames underscoring haves have-nots.but it's also much more than that. arrives moment white house continues make vociferous, bold claims about economy, particularly manufacturing. that's despite increasing concerns economists warnings history recession could be horizon. there's sobering contrast between trump's rhetoric how job growth ballooned during his presidency reality broader slowdown slamming states -- including ohio helped win 2016 presidential election.read

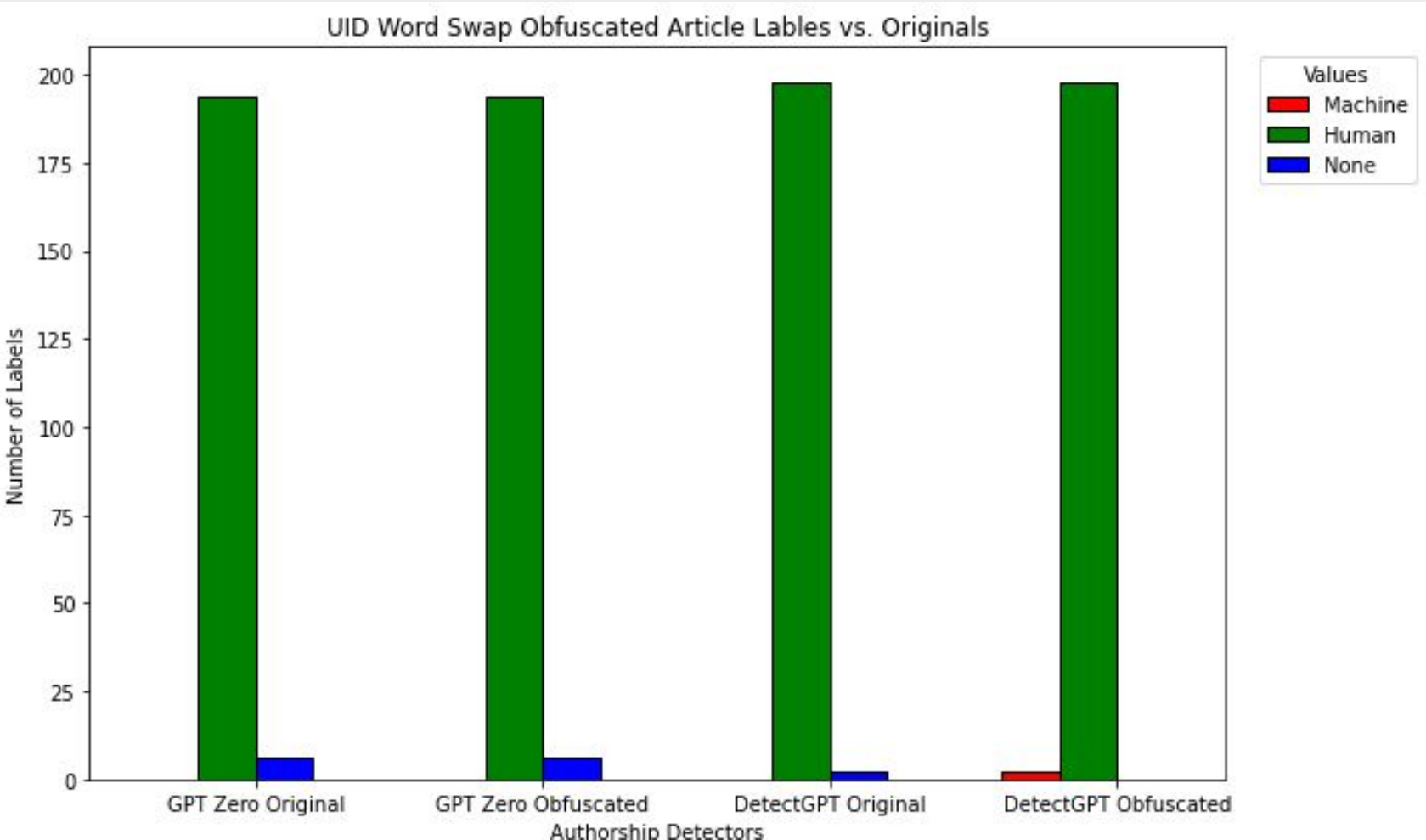
washington ( ) at a time when president donald trump seems to permeate nearly every aspect of american discourse , it might come as surprise that the first movie from barack and michelle obama 's publishing company , higher ground , never mentions him by name.but subtlety is part power factory , new netflix documentary charts r ening factory in dayton , ohio . over course two hours , movie , directed seasoned documentarians steven bognar julia reichert , serves quiet historical political corrective , offering their portrait state america 's industrial heartland helps viewers rethink who , exactly , project.american starts on december 23 , 2008 , crowd gathers learn general motors plant dayton has shuttered . then fast-forwards 2015 , ens enterprise , fuyao glass corporation , arm shanghai-based company manufactures automotive glass . one man makes fuyao 's expanded mission crystal clear : what we 're doing bringing cultures together : chinese culture s culture . so we are truly global organization.as some critics have pointed out , is , important ways , commentary upon globalization ; york times review frames underscoring haves have-nots.but it 's also much more than that . arrives moment white house continues make vociferous , conflicting claims about economy , particularly manufacturing . that 's despite increasing concerns economists thought history recession could be horizon . there 's sobering contrast between trump 's rhetoric how job growth ballooned during his political reality broader slowdown slamming states -- including ohio helped win 2016 presidential election.read

Human (Labeled Human before and after)

UWS


Semantic VS UID Article human 11
UID Score 1
Semantic Score
14.9
14.8
14.7
14.6
14.5
14.4
0.9965
0.9970
0.9975
0.9980
0.9985
0.9990
0.9995
1.0000


Semantic VS UID Article human 11
UID Score 3
Semantic Score
27.5
27.0
26.5
26.0
25.5
0.9965
0.9970
0.9975
0.9980
0.9985
0.9990
0.9995
1.0000
Original Article
New Articles
Selected Article


UID Word Swap Obfuscated Article Lables vs. Originals
Number of Labels
0
25
50
75
100
125
150
175
200
GPT Zero Original
GPT Zero Obfuscated
DetectGPT Original
DetectGPT Obfuscated
Authorship Detectors
Values
Machine
Human
None

UWS Challenges

Main problem, lack of diversity amongst UID scores

Just swapping one word each sentence didn’t seem enough to push the UID score in any major way

Way to drastically alter the articles to get more significant UID variation

UID Paraphrase (UP)

- Developed after UID Word Swap, hoped to result in greater variation to UID without losing too much semantic similarity
- Utilize diverse beam search to paraphrase each sentence in each article
- Attempted the process on mostly every sentence in the article

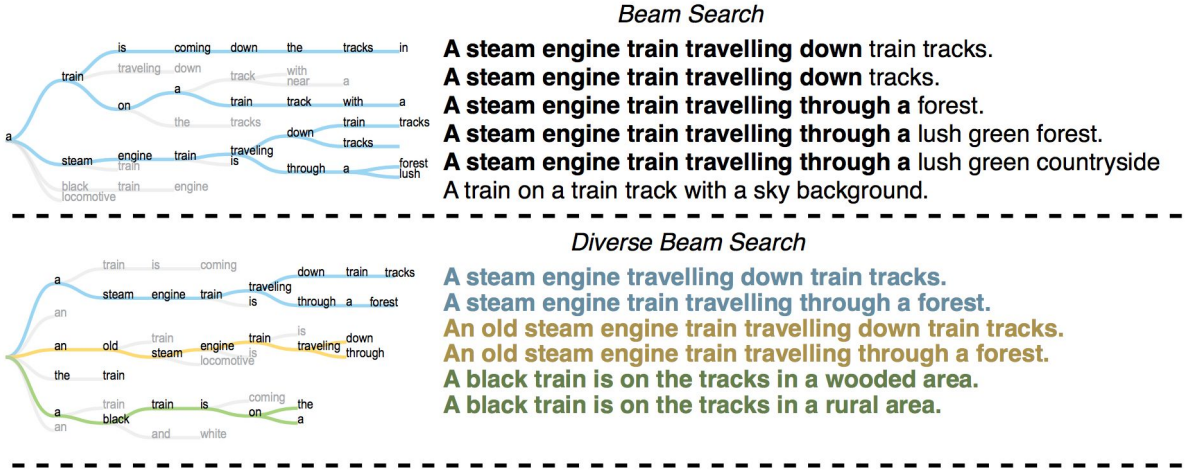
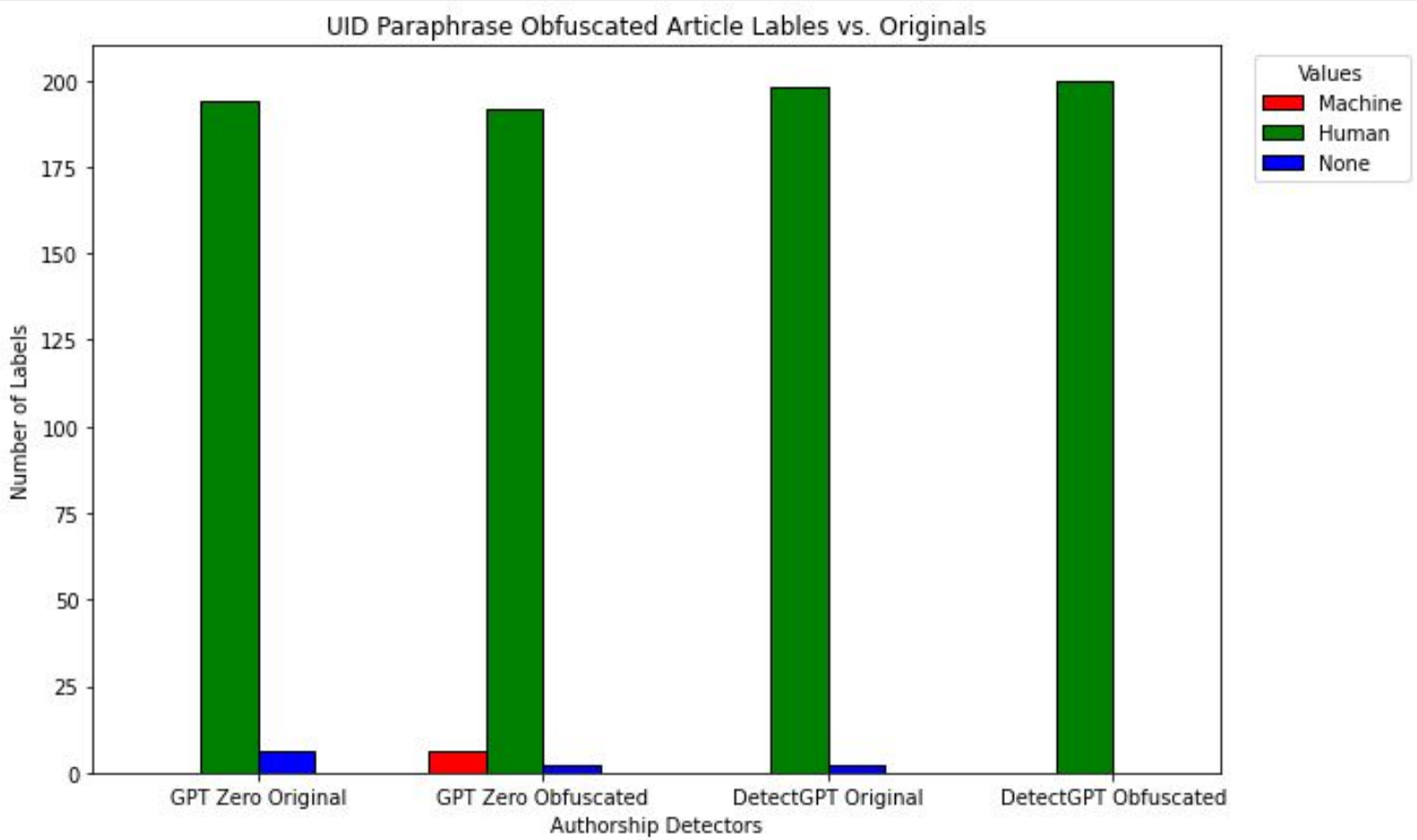


Image: https://github.com/kweonwooj/papers/issues/32

```python
def ParaSwap(sentence):
    new_articles = []
    # Diverse Beam search
    context =  sentence
    text = context + " </s>"
    encoding = phrase_tokenizer.encode_plus(text, padding=True, return_tensors="pt")
    input_ids,attention_mask  = encoding["input_ids"], encoding["attention_mask"]
    phrase_model.eval()
    diverse_beam_outputs = phrase_model.generate(
        input_ids=input_ids,
        attention_mask=attention_mask,
        num_beams=10,
        num_beam_groups=10,
        num_return_sequences=10,
        diversity_penalty=0.70,
        max_length=500,  # Increase the max_length value as needed
        )
    print ("\n\n")
    print ("Original: ",context)
    for beam_output in diverse_beam_outputs:
        new_articles.append(phrase_tokenizer.decode(beam_output, skip_special_tokens=True,clean_up_tokenization_spaces=True
    return new_articles
```


UID Paraphrase Obfuscated Article Lables vs. Originals
Number of Labels
0
25
50
75
100
125
150
175
200
GPT Zero Original
GPT Zero Obfuscated
DetectGPT Original
DetectGPT Obfuscated
Authorship Detectors
Values
Machine
Human
None

UID Paraphrase Problems

Original Articles	Selected Articles	True Labes	GPTZero_Orig	GPTZero_S	DetectGPT	DetectGPT_Selected
related. as the 15th lok sabl	related. The 15th session of l	Human	1	1	1	1
related. as the 15th lok sabl	related. The 15th session of l	Human	1	1	1	1
markets ended the last mor	The last month saw a rise of a	Human	1	1	1	1
markets ended the last mor	The last month saw a rise of a	Human	1	1	1	1
it was never a question that	It was never a question that c	Human	1	1	1	1
it was never a question that	It was never a question that c	Human	1	1	1	1
related. india said that the v	related. The global bank ban	Human	1	1	1	1
. exit g rgaon rallyaug 25, 2(	. What's your must do list for	GPT-3	1	1	1	1
. exit g rgaon rallyaug 25, 2(	. What's your must do list for	GPT-3	1	1	1	1
researchers in denmark wer	As the country's wolves have	GPT-3	1	1	1	1
researchers in denmark wer	As the country's wolves have	GPT-3	1	1	1	1
. new delhi: days before the	. On twitter, the delhi's body	GPT-3			1	1
. new delhi: days before the	. On twitter, the delhi's body	GPT-3			1	1
'google s'est associe avec la	The best s'est associe with la	GPT-3	1	1	1	1
'google s'est associe avec la	The best s'est associe with la	GPT-3	1	1	1	1

UID Paraphrase Example

. sheena gupta, 32, who is based out of mumbai, always tries to lead the conversation her friend kotak's facebook wall when they meet. i can't tell you number times hear: hey, what did post on my wall? has a group friends in delhi and mumbai created through social networking platforms, said.'

'. When they meet, sheena gupta, 32, who is based out of mumbai, continues to lead the discussion with her friend kotak's facebook wall... Hey, what did you see on my wall? i can't tell you how many times i can't tell you how many times i can't tell you how many times i can't tell you how many times i hear. According to delhi and mumbai, a group of friends in delhi and mumbai formed through social media platforms.'

GPT-3 UID Score Variance (Labeled: Human switched to machine)

UID Paraphrase Example

political novice, who polled between 30 50 percent private polls, forced confront state lawmaker mike kennedy, started raising money polling around 20 his race, public policy polling.tue, 09:33:12 -0700brazil will not roll back current mix, finance minister: di luzio(sharecast brazil's government existing policies that aimed reassuring investors undo incoherently assembled patchwork, minister henrique meirelles told journalists on tuesday, what'

finance minister mike kennedy, who polled between 30 percent and 31 percent private polls, prompting state lawmaker mike kennedy, who polled around 30 percent private pollsoutput: tue tue tue tue tue tue tue tue tue l's l's tue tue reassuredy tuedy tue tuedy tuedy tuedy tuedy's tuedy's tuedy'sy'stuedy'stuedy'stuedy'stuedy'stuedy'stuedy'stuedy'stue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tue tuedy'stuedy'stuedy'stuedy'stue tuedy'stuedy'stue tue tue tue tue tuedy'stue tue tue tue tue tue tue tue tue tue tue

GPT-3 UID Score Variance (Labeled: Human switched to machine)

Candidate Selection Process (for both UWS and Paraphrase)

- Process attempts to cause the biggest change for obfuscation with regards to UID score whilst still being semantically similar
- Calculates the differences for both UID scores for each article
- Sorts the differences
- Finds the highest difference article that passes the similarity score threshold

```
def candidate_select(num_article, article):
    index = num_article * 11
    max_index = index + 11

    UID1_Difference_list = []
    UID2_Difference_list = []

    original_article = article_list[index]
    alternate_list = article_list[index+1:max_index]

    original_UID1 = column_data1[index]
    alternate_UID1 = column_data1[index+1:max_index]

    original_UID2 = column_data2[index]
    alternate_UID2 = column_data2[index+1:max_index]

    for i in range(len(alternate_UID1)):
        UID1_Difference_list.append(abs(original_UID1-alternate_UID1[i]))

    for i in range(len(alternate_UID2)):
        UID2_Difference_list.append(abs(original_UID2-alternate_UID2[i]))

    sorted_UID1 = sorted(UID1_Difference_list)
    sorted_UID2 = sorted(UID2_Difference_list)

    for i in reversed(sorted_UID1):
        score_index = UID1_Difference_list.index(i) + 1
        if score_list[score_index] >= .85:
            selected_article_list.append(alternate_list[score_index])
            break


    for i in reversed(sorted_UID2):
        score_index = UID2_Difference_list.index(i) + 1
        if score_list[score_index] >= .85:
            selected_article_list.append(alternate_list[score_index])
            break
```

UID Paraphrase Challenges

Degeneration/Repetition - fixed or improved via tuning of parameters for the diverse beam search

Detectors Labeling - fixed by expanding the number of articles and/or utilizing a different detector

UID Candidate Selection Process - Make the process more sophisticated so as to pick higher quality candidates

Incorporate UID into the perturbation/obfuscation process rather than after the perturbations has been made

Conclusion

Explore the applicability of UID metric in the task of obfuscation amongst a number of obfuscation methods. Specifically investigate if UID can be used as a guiding metric to result in successful obfuscation in which an automated authorship attributor misattributes an obfuscated article.

Nothing to indicate that UID can be used as a guiding metric…

But study was very small scaled and restricted due to time

Know obfuscation is a feature we can use to distinguish between human and machine authors

Just a matter of developing the correct obfuscation method to exploit this

Future Improvements

- Increase scale of research
  - More articles, different LMs
  - Tune parameters for UP (diverse beam search)
  - Tune parameters for detectors
  - Increase sophistication/complexity of UWS
  - Improve candidate selection
    - denoising/post-processing on alternates
- Try different detectors
- Further implement UID as a more sophisticated guiding metric

References

A. F. Frank and T. Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. Proceedings of the Annual Meeting of the Cognitive Science Society, 30, 2008.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURING-BENCH: A benchmark environment for the turing test in the age of neural text generation. September 2021.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. Revisiting the uniform information density hypothesis. September 2021.

Saranya Venkatraman, He He, and David Reitter. How do decoding algorithms distribute information in dialogue responses? In Findings of the Association for Computational Linguistics: EACL 2023, pages 953–962, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.