APPDIA: A DISCOURSE-AWARE TRANSFORMER-BASED STYLE TRANSFER MODEL FOR OFFENSIVE SOCIAL MEDIA CONVERSATIONS

Katherine Atwell*, Sabit Hassan*, Malihe Alikhani

*equal contribution

Department of Computer Science

University of Pittsburgh

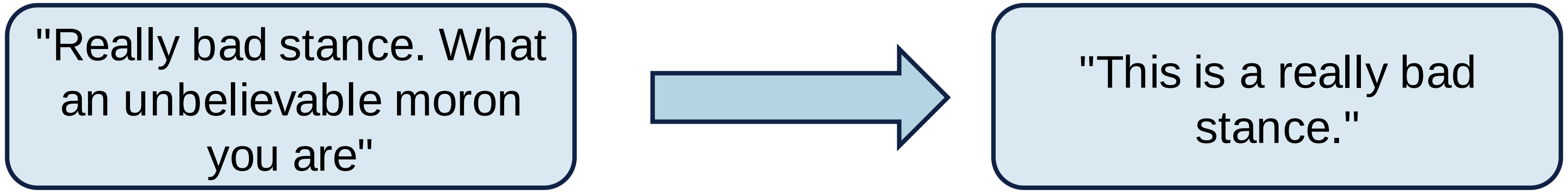
MOTIVATION

Problem

- Hateful content on social media can harm users' well-being (Waldron, 2012; Gülaçtı, 2010)
- The scale of large social-media platforms makes human moderation challenging (Dosono and Semaan, 2019)

Moderation isn't always the answer

- Sometimes comments have important contributions beyond their offensiveness
- If offensiveness can be limited, the comment can be **retained**



"Really bad stance. What an unbelievable moron you are"

"This is a really bad stance."

HOW CAN WE REDUCE OFFENSIVENESS?

Task

- **Style-transfer** task
- Offensive → inoffensive

Approach

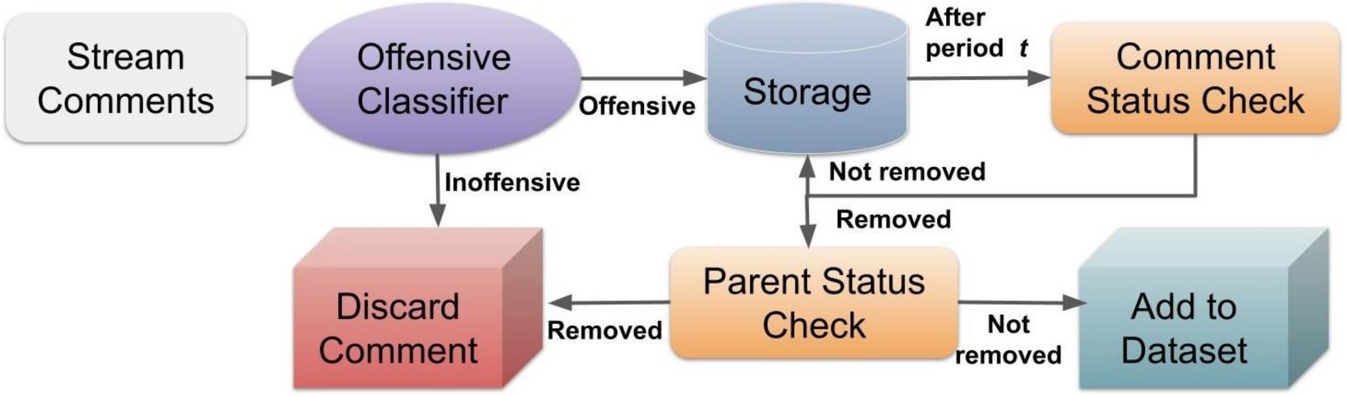
- Create a corpus with inoffensive paraphrases of offensive text
- Propose a discourse-aware model

DATA COLLECTION

Dataset

- 14 Reddit subreddits
- Topics including gender, politics, Q&A, and personal views
- 5.8k comments for annotation

Data Collection Pipeline



ANNOTATION

Protocol

- Rewrite to remove offensiveness while preserving the comment's meaning
- Discard comments that cannot be rewritten to be inoffensive
- Make **local** (removal of words) or **global** (rephrases of text) changes when necessary

Original Comment	Rewritten Comment	Global/Local	Reason
What backward b******k nowhere country do you live in?	What country do you live in?	Local	Xenophobia, cursing
To hell with peaceful protest. Protestors should drag DeathSantis out of his home and have a public trial	Peaceful protest won't work. Protestors should go for a public trial.	Global	Threats of violence

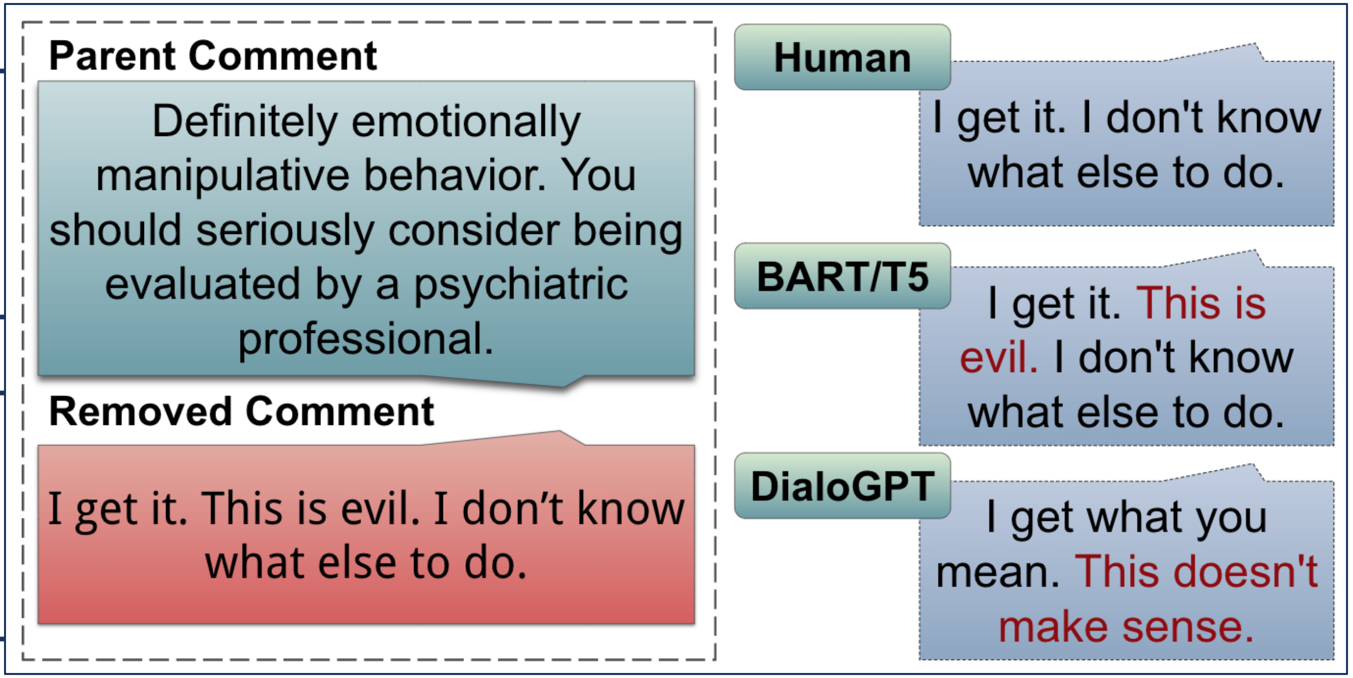
Are Pretrained Language Models Effective?

BART/T5

- Retains meaning
- Does not reduce offensiveness

DialoGPT

- Reduces offensiveness
- Substantially alters meaning



Parent Comment

Definitely emotionally manipulative behavior. You should seriously consider being evaluated by a psychiatric professional.

Removed Comment

I get it. This is evil. I don’t know what else to do.

Human

I get it. I don't know what else to do.

BART/T5

I get it. This is evil. I don't know what else to do.

DialoGPT

I get what you mean. This doesn't make sense.

Are Pretrained Language Models Effective?

BART/T5

- Retains meaning
- Does not reduce offensiveness

DialoGPT

- Reduces offensiveness
- Substantially alters meaning

Compared Against Annotated Text			
Model	**BLEU**	**BERTScore**	**SafeScore**
BART	65.1	68.1	44.7
T5	**65.3**	**69.2**	51.3
DialoGPT	42.5	47.2	**66.3**

Compared Against Original Text			
Model	**BLEU**	**BERTScore**	**SafeScore**
BART	**76.2**	**78.4**	44.7
T5	74.8	78.0	51.3
DialoGPT	45.3	49.4	**66.3**

DISCOURSE COHERENCE FRAMEWORKS

Purpose

- Capture rhetorical relations between units of text (or trees with multiple text units)
- Encode structural information about text organization

Our task

- Have had success with several generation tasks (even with pretrained language models)
- **Have not yet been employed for style-transfer**
- Can allow us to capture relations within and between comments

Hypothesis: Discourse frameworks can help style-transfer models preserve semantic meaning for offensiveness transfer

PENN DISCOURSE TREEBANK (PDTB)

Idea

- Captures rhetorical relations between two units of text (no hierarchy)
- Relations can be **implicitly** or **explicitly** signaled

Explicit **Contingency.Cause.result**

In addition, its machines are typically easier to operate, so **customers require less assistance from software.**

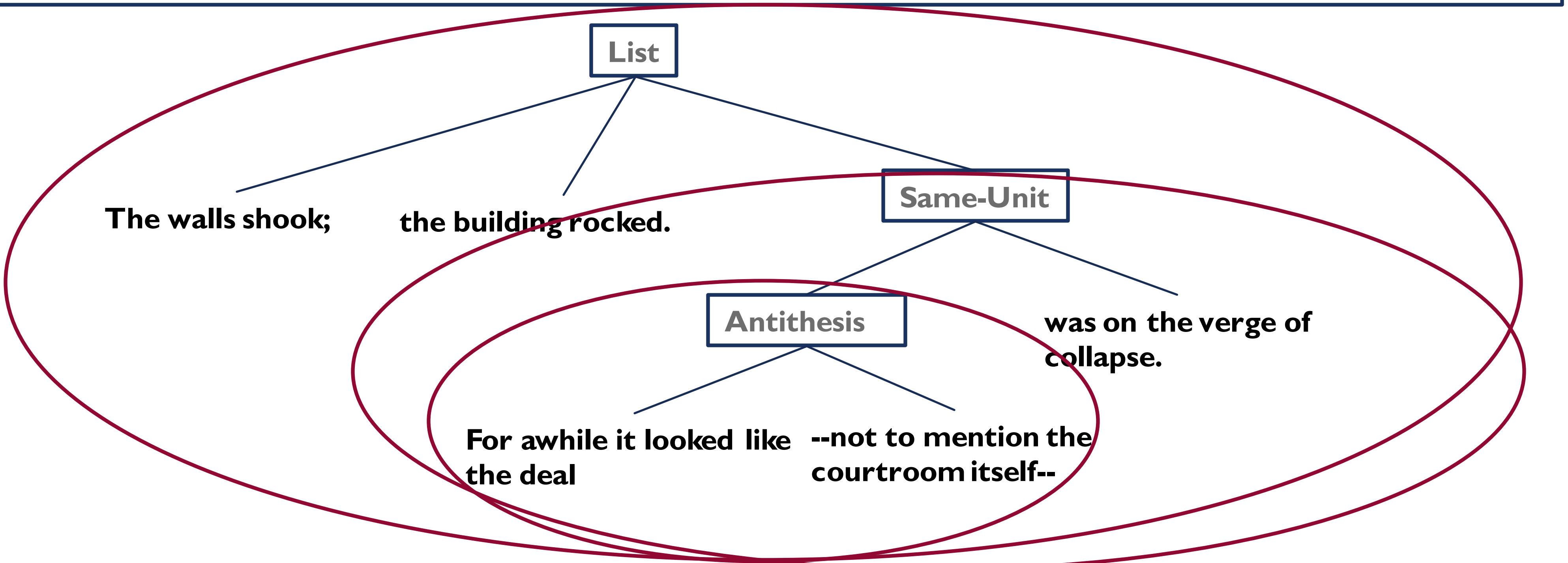
Implicit **Expansion.Restatement.specification**

I never gamble too far. [Implicit = in particular.] **I quit after one try.**

RST DISCOURSE TREEBANK (RST-DT)

Idea

- Based on **Rhetorical Structure Theory**
- Hierarchical discourse relations
- No implicit/explicit distinction


List
The walls shook;
the building rocked.
Same-Unit
Antithesis
was on the verge of collapse.
For awhile it looked like the deal
--not to mention the courtroom itself--

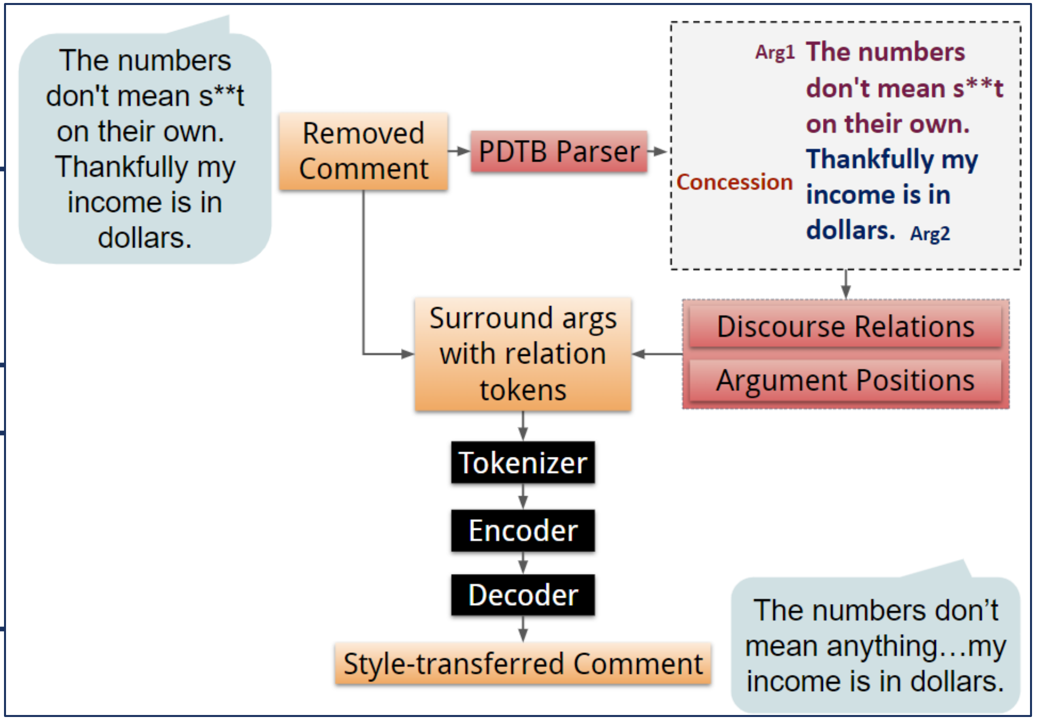
DISCOURSE-AWARE MODEL: PDTB

Parse comment text in isolation

- Explicit relations (Lin et al, 2014)
- Implicit relations (Kim et al, 2020)

Insert argument pair positions and relations

- Update tokenizer
- Resize embedding


The numbers don't mean s**t on their own. Thankfully my income is in dollars.
Removed Comment
PDTB Parser
Arg1
The numbers don't mean s**t on their own.
Concession
Thankfully my income is in dollars.
Arg2
Discourse Relations
Argument Positions
Surround args with relation tokens
Tokenizer
Encoder
Decoder
Style-transferred Comment
The numbers don't mean anything…my income is in dollars.

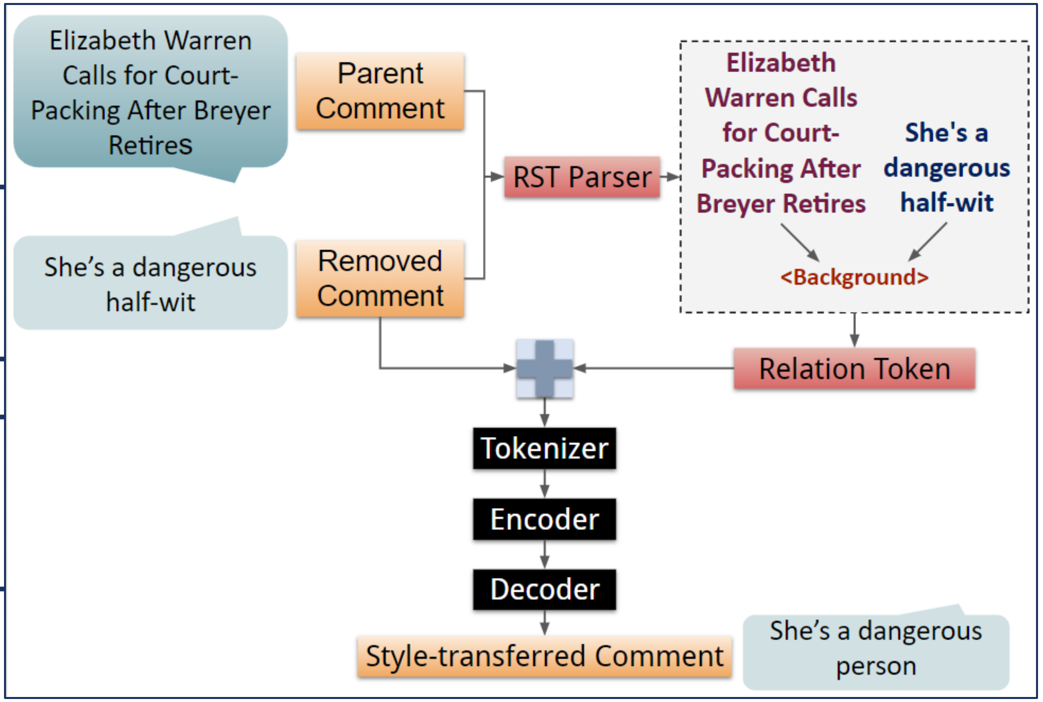
DISCOURSE-AWARE MODEL: RST-DT

Get root relation between comment and parent

- Concatenate and run EDU segmenter (Li et al., 2018)
- Predict relation on resulting EDUs (Wang et al., 2017)

Prepend relation to input

- Update tokenizer
- Resize embedding


Elizabeth Warren Calls for Court-Packing After Breyer Retires
Parent Comment
She's a dangerous half-wit
Removed Comment
RST Parser
Elizabeth Warren Calls for Court-Packing After Breyer Retires
She's a dangerous half-wit
<Background>
Relation Token
Tokenizer
Encoder
Decoder
Style-transferred Comment
She's a dangerous person

EXPERIMENTS: DISCOURSE-AWARE MODELS

Discourse-aware model improves on baseline

- **4.2** BLEU and **3.4** BERTScore improvement
- RST-DT has higher impact compared to PDTB
- Combining RST-DT and PDTB yields highest improvement

Compared Against Annotated Text

Model	BLEU	BERTScore	SafeScore
Baseline	42.5 (0.0)	47.2 (0.0)	66.3 (0.0)
RST-augmented	45.2 (2.6)	**50.6 (3.4)**	65.7 (-0.7)
PDTB-augmented	44.4 (1.9)	48.7 (1.5)	65.3 (-1.0)
RST+PDTB-augmented	**46.7 (4.2)**	50.3 (3.1)	**67.7 (1.3)**

Compared Against Original Text

Model	BLEU	BERTScore	SafeScore
Baseline	45.3 (0.0)	49.4 (0.0)	66.3 (0.0)
RST-augmented	47.9 (2.5)	**52.8 (3.4)**	65.7 (-0.7)
PDTB-augmented	47.2 (1.9)	50.8 (1.4)	65.3 (-1.0)
RST+PDTB-augmented	**49.6 (4.3)**	52.6 (3.2)	**67.7 (1.3)**

EXPERIMENTS: DISCOURSE-AWARE MODELS

Discourse-aware model improves on baseline

- **4.2** BLEU and **3.4** BERTScore improvement
- RST-DT has higher impact compared to PDTB
- Combining RST-DT and PDTB yields highest improvement

Compared Against Annotated Text

Model	BLEU	BERTScore	SafeScore
Baseline	42.5 (0.0)	47.2 (0.0)	66.3 (0.0)
RST-augmented	45.2 (2.6)	**50.6 (3.4)**	65.7 (-0.7)
PDTB-augmented	44.4 (1.9)	48.7 (1.5)	65.3 (-1.0)
RST+PDTB-augmented	**46.7 (4.2)**	50.3 (3.1)	**67.7 (1.3)**

Compared Against Original Text

Model	BLEU	BERTScore	SafeScore
Baseline	45.3 (0.0)	49.4 (0.0)	66.3 (0.0)
RST-augmented	47.9 (2.5)	**52.8 (3.4)**	65.7 (-0.7)
PDTB-augmented	47.2 (1.9)	50.8 (1.4)	65.3 (-1.0)
RST+PDTB-augmented	**49.6 (4.3)**	52.6 (3.2)	**67.7 (1.3)**

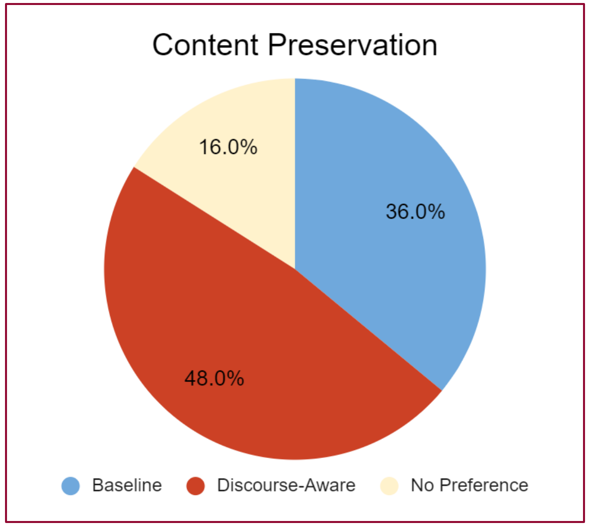
MODEL OUTPUTS

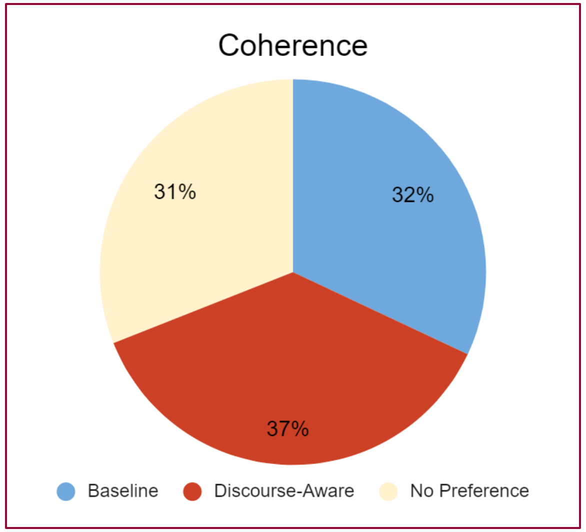
Original Comment	Really bad stance. What an unbelievable moron you are.
Human	This is a really bad stance.
Baseline	That is a bad stance, you are
Discourse-aware	That's a bad stance, and you are being unbelievable.
Original Comment	Just like Republicans creating themselves narratives to …..
Human	Just like Republicans creating themselves narratives as fuel to ….
Baseline	Just like Democrats creating themselves a narrative to justify …
Discourse-aware	Republicans create their own narratives to justify themselves.

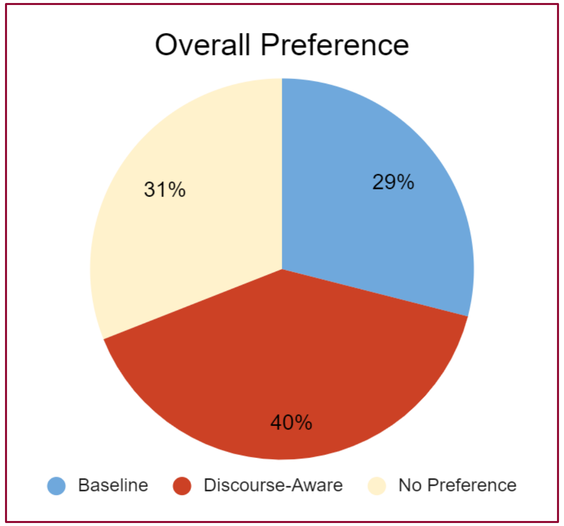
HUMAN EVALUATION

100 pairs of output from baseline and discourse-aware

- Coherence
- Content Preservation
- Overall Preference



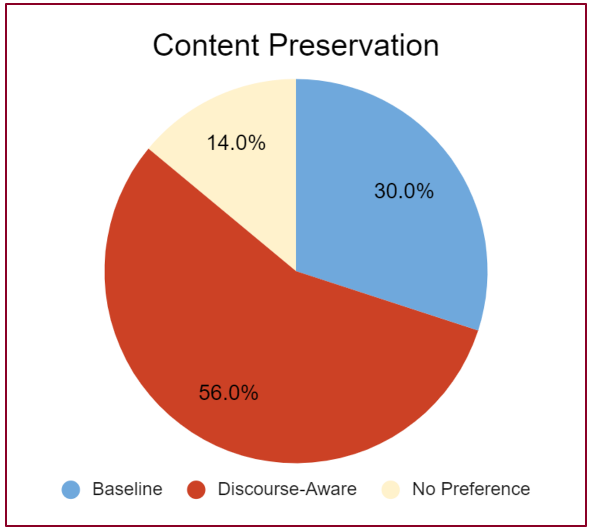


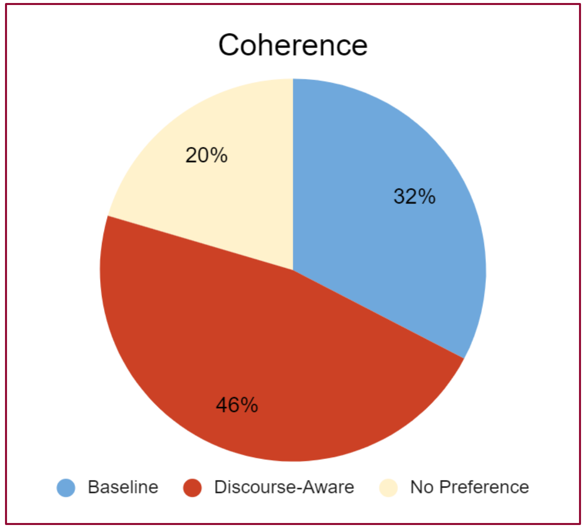


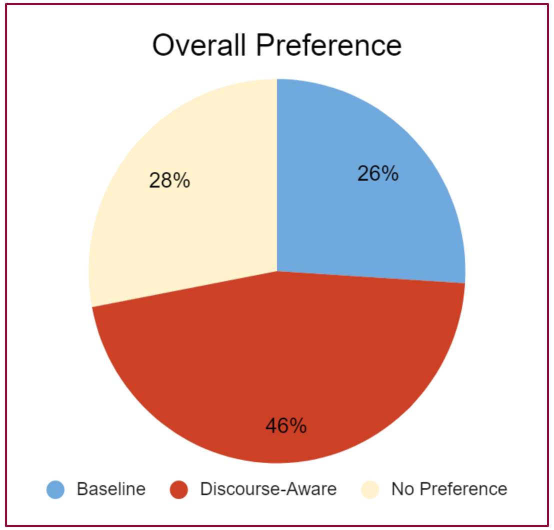
HUMAN EVALUATION (CONTD.)

Evaluate on instances with discourse relations

- Greater difference in preference
- Discourse-aware is more important for text with discourse structure







LIMITATIONS

Original Comment	You guys love oppression. Let me get some barbecue recipes. I know you Klansmen are always having cookouts.	Type of Error
Human	You guys seem to love oppression. Let me get some barbecue recipes. I know you people are always having cookouts.	N/A
Baseline	You guys love persecution. Let me go get some recipes. I know that you are always had cooking out.	Incoherent
Discourse-aware	You guys like oppression. I'll get some recipes.	Missing parts
Original Comment	May Mother Russia spare you from your certain demise.	N/A
Human	May Mother Russia spare you from your fate.	N/A
Baseline	May Mother Russia protect you from certain demise.	Change of meaning
Discourse-aware	May Mother Russia spare you.	Missing parts

CONCLUSION AND FUTURE WORK

Discourse frameworks help content preservation

- Existing parsers are not accurate
- Need for further discourse-annotated corpora

Style transferring offensive text can improve online environment

- Improve psychological well-being
- Motivate healthy engagement

THANK YOU! ANY QUESTIONS?

Thanks to DARPA for funding this project

QR for Github repository