Automated Identification of Social Media Bots using Deepfake Text Detection

Sina Mahdipour Saravani, Indrajit Ray, and Indrakshi Ray

Department of Computer Science

Colorado State University

Colorado State University

Motivation

- Social Media is the ubiquitous tool of real-time, large-scale communication
  - Huge user population

    → Broad impacts

- With such potentials

  → Malicious uses
    - Using bots for propagate misinformation and spam

      → influence economy, politics, healthcare, etc.

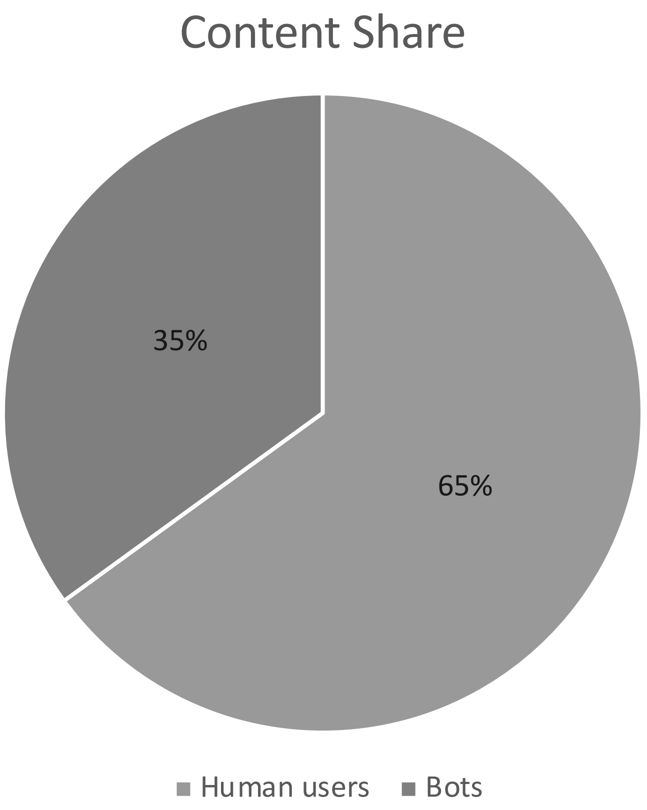
Motivation

- Examples of malicious use:
  - Syrian civil war
  - Boston marathon bombing
  - Cynk © 220-fold drop in market price

- Objectives:
  - Political gain
  - Financial gain

Motivation

- Among 9% to 15% of accounts are bots (over 48 million) [1]
- 35% of content is produced by bots [1]
- “Near half of Twitter accounts pushing to reopen America may be bots.” [2]



[1] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1.
[2] https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/

Motivation

- Social bots are different in their sophistication and capabilities
  - Some simply retweet or post content generated by human controller in large quantities
  - Some are way more complex and capable of generating content and interacting with people without human help
    - Progress in Natural Language Generation

    → Bots are now very deceptive and hard to detect

Account-level

- Account-level bot detection:
  - Network relationships (followers and friends)
  - Usage pattern
  - Account name and creation time
  - Content and sentiment of all/several posts

- This is expensive, since a large amount of data is required for each account under assessment

Content-level

- Account-level metadata may not be available
- If account is a cyborg, account-level mechanism tend to fail
  - Cyborg: human-assisted bot or bot-assisted human

- What is the solution?
  - Content-level bot detection
    - Decide based on a single content observation
    - Given a content from an online social network (OSN), determine whether it is produced by a bot or a human user

Content-level

- Low performance of humans in detecting the bot-generated text

- Text classification problem in Natural Language Processing (NLP)
  - Bots use Deep Learning for generating text content

  → Unsuitability of shallow syntactic and semantic NLP features for bot detection

  → Use Deep Learning as a natural candidate to detect them

Our Contributions

- Investigate the state-of-the-art NLP architectures and report their performance on detecting bot-generated text
  - Improve the state-of-the-art accuracy by 2 percent
  - Performance reported on a real-world, deceptive dataset
- Adapt a neural component (NeXtVLAD) from computer vision to NLP and assess its performance
- Real-time applicability of the approach by nature

Hankyol Lee, Youngjae Yu, and Gunhee Kim. 2020. Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context. In *Proceedings of the Workshop on Figurative Language Processing*, pages 12–17. Association for Computational Linguistics.

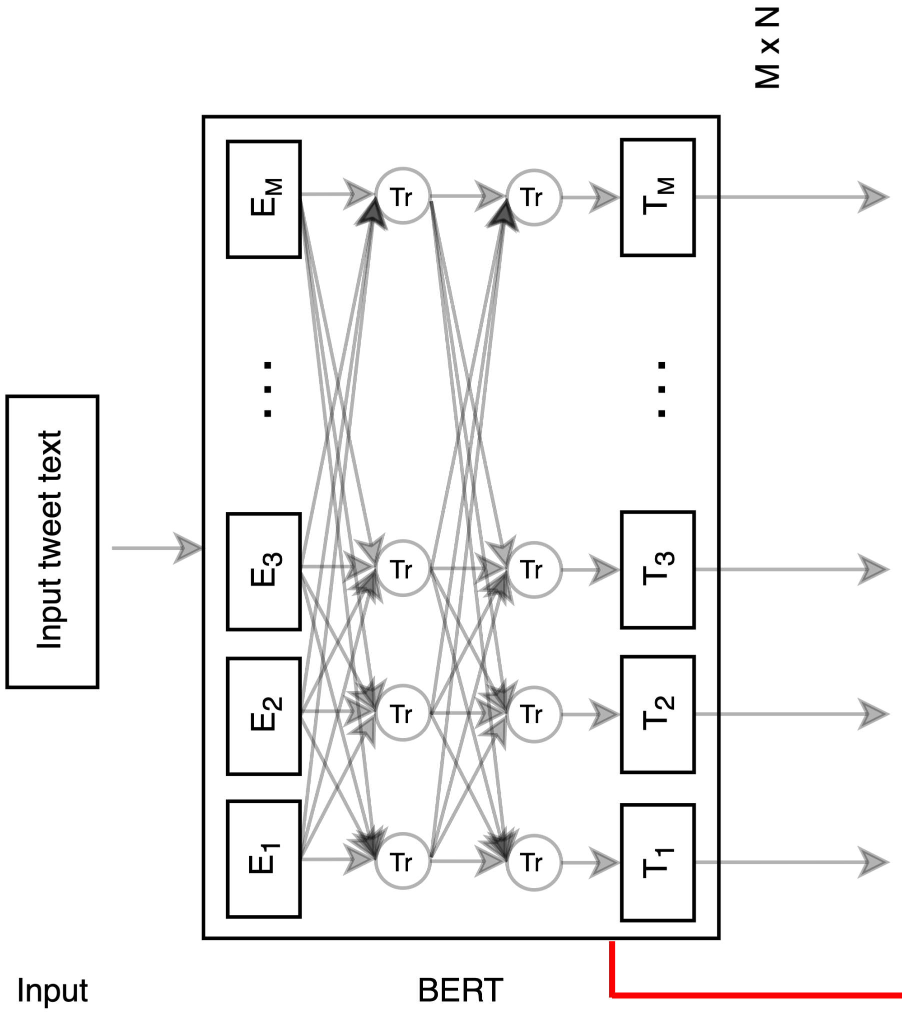
Dataset

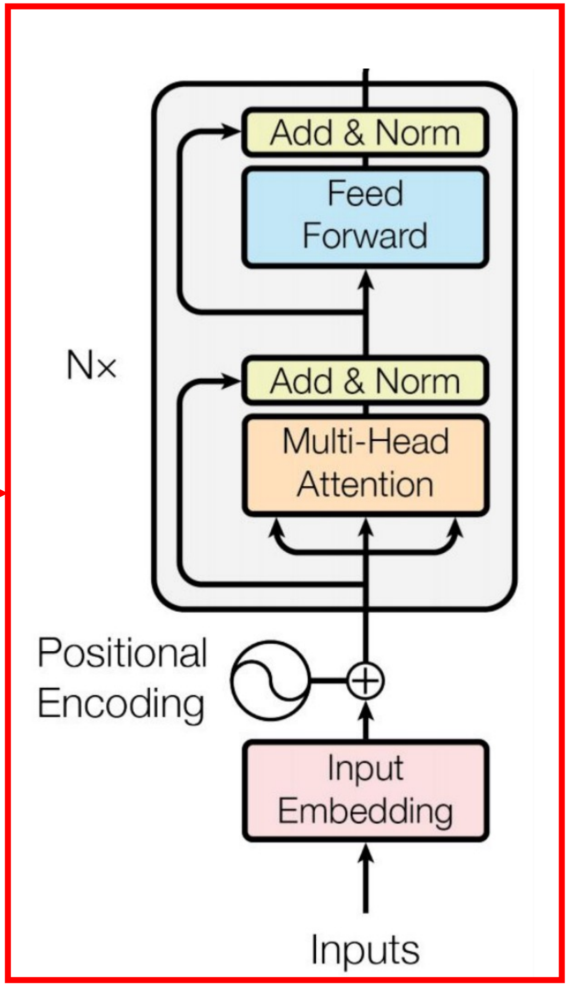
- Deepfake
- Real-world

Tweet text	**Label**
the world needs more whale stories. I would love to know what whalefacts are hiding in them.	GPT-2 Bot
I will make [FOLLOWERS OF A RELIGION] victims. They come into the United States but should have been crippled so I flourish. I can do it. @USERNAME #debate	RNN Bot
it literally what time of gucci shorts or not tolerate Libra slander on my face	Other Bot
I think if i put my mind to it, I could put a tree in my house like they do at the Cherry hill mall	Human

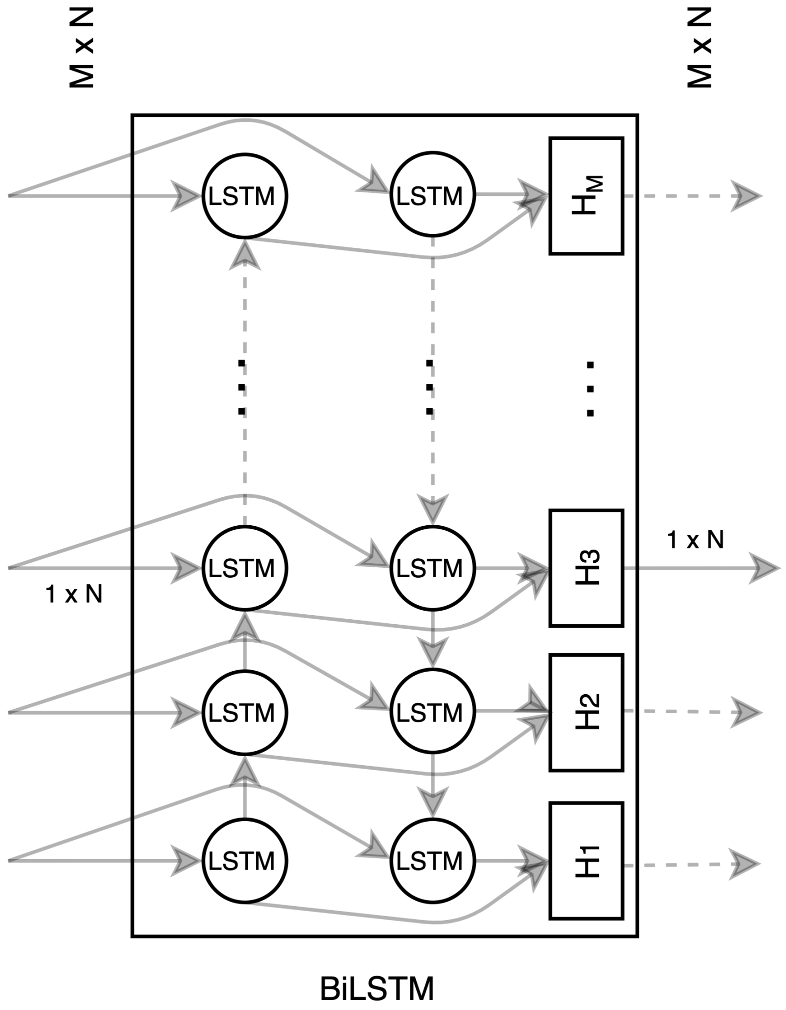
Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. "TweepFake: About detecting deepfake tweets." Plos one 16, no. 5 (2021): e0251415.

Transformer





BiLSTM

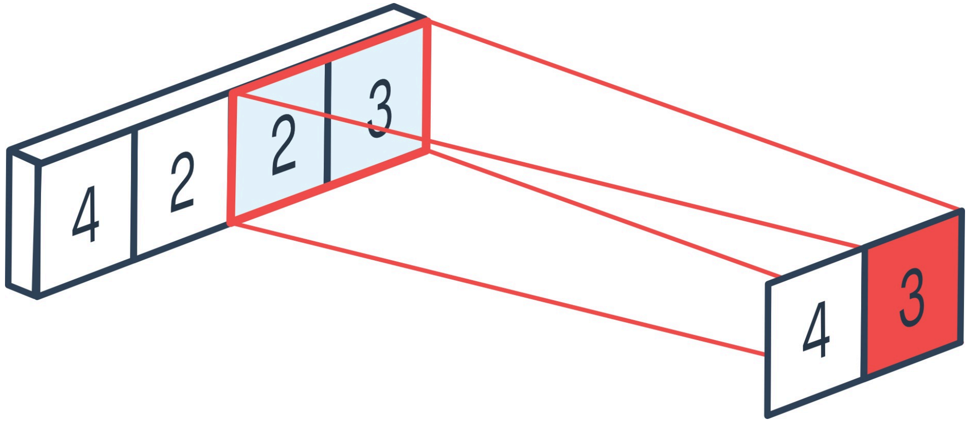


- Each LSTM cell includes:
  - Linear transformation
  - tanh and softmax activation functions

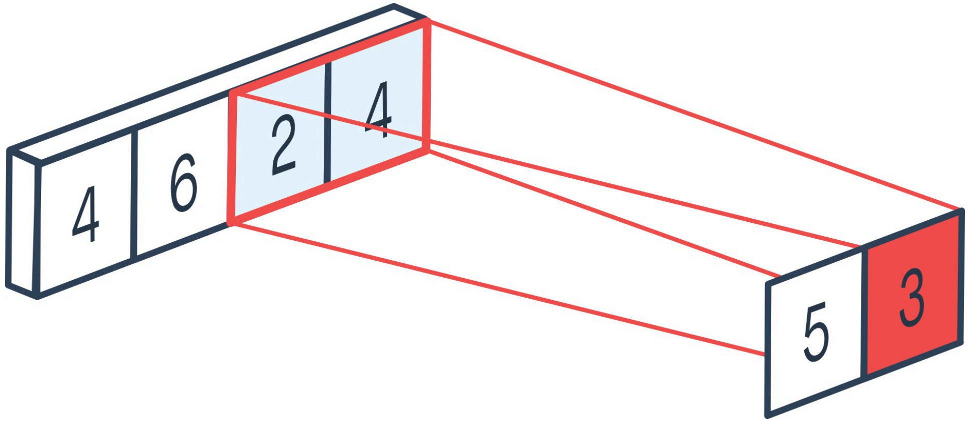
- LSTM for enhancing temporal information

Non-parametric Pooling

- Maximum Pooling:



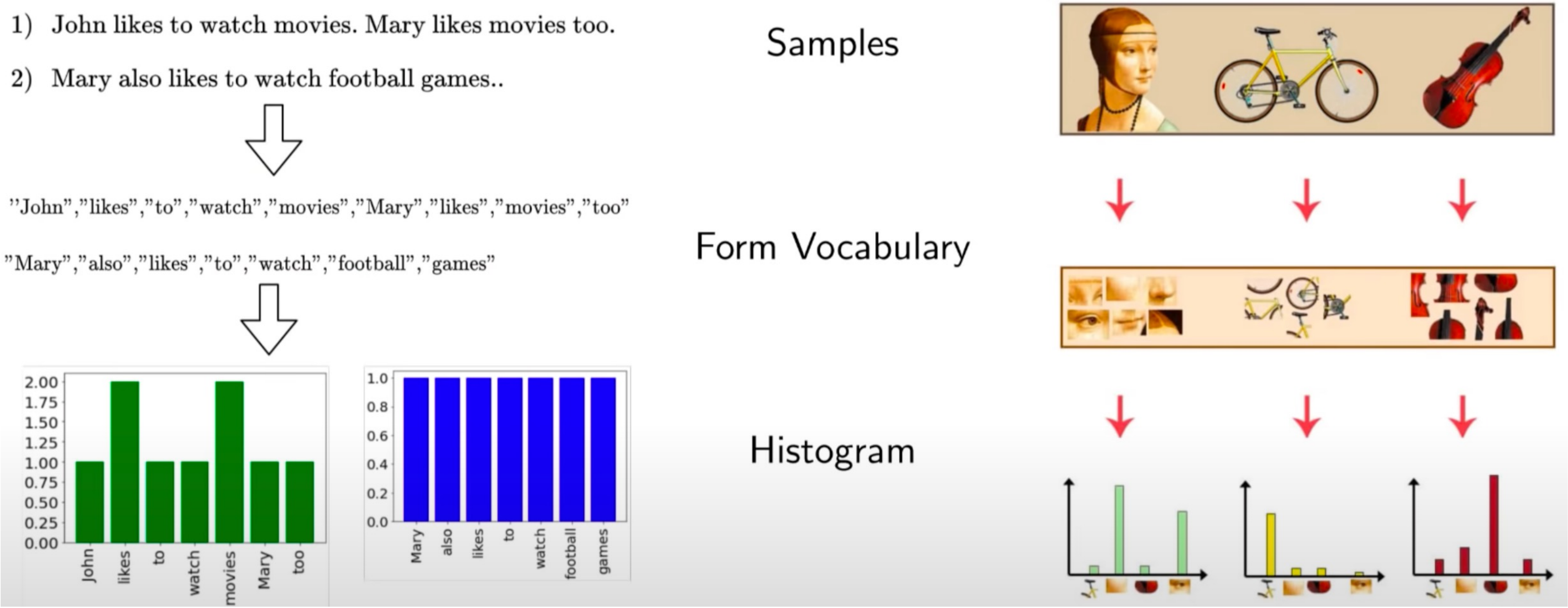
- Average Pooling:



https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/max-pooling-block-1d
https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/average-pooling-1d

Parametric Pooling: NeXtVLAD

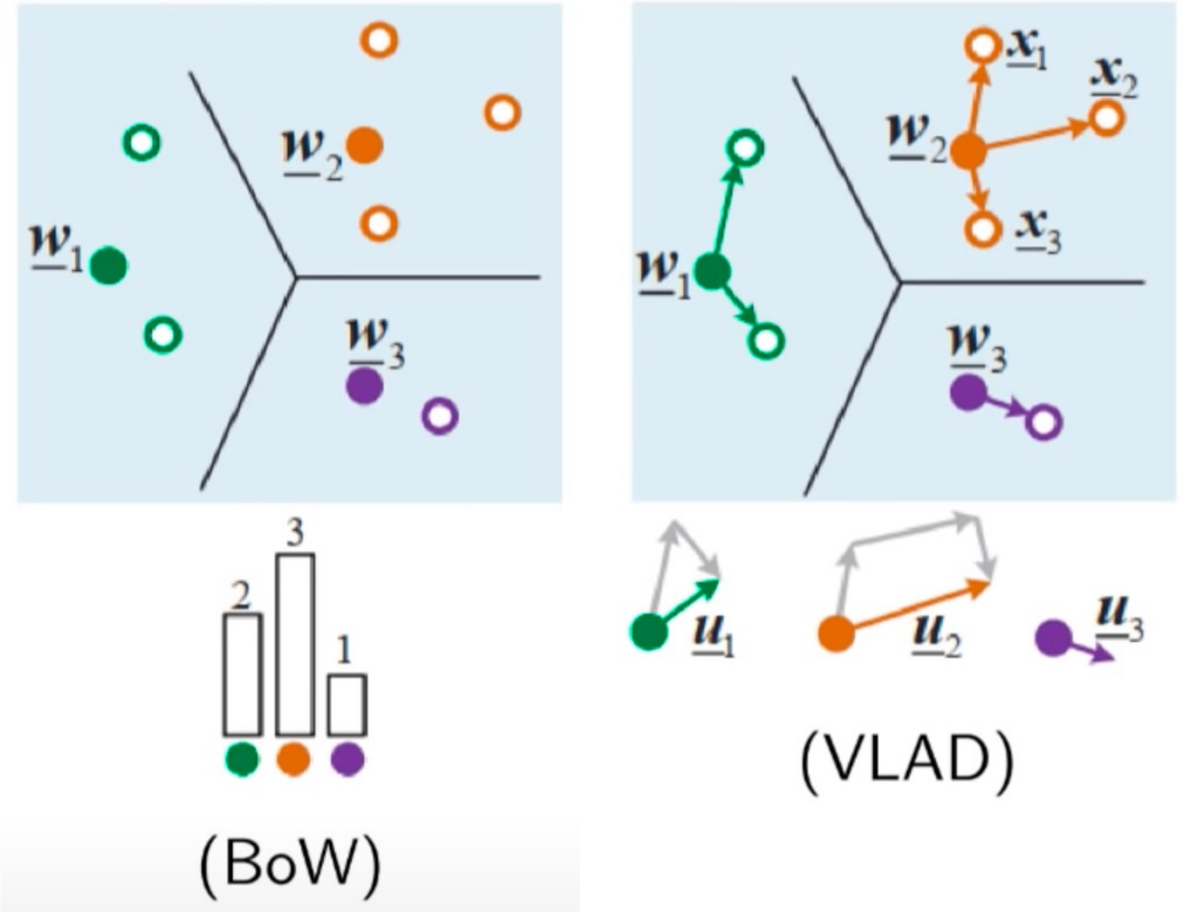
Bag of Visual Words



Vineeth N Balasubramanian, https://www.youtube.com/watch?v=CwJPEMcuAxY

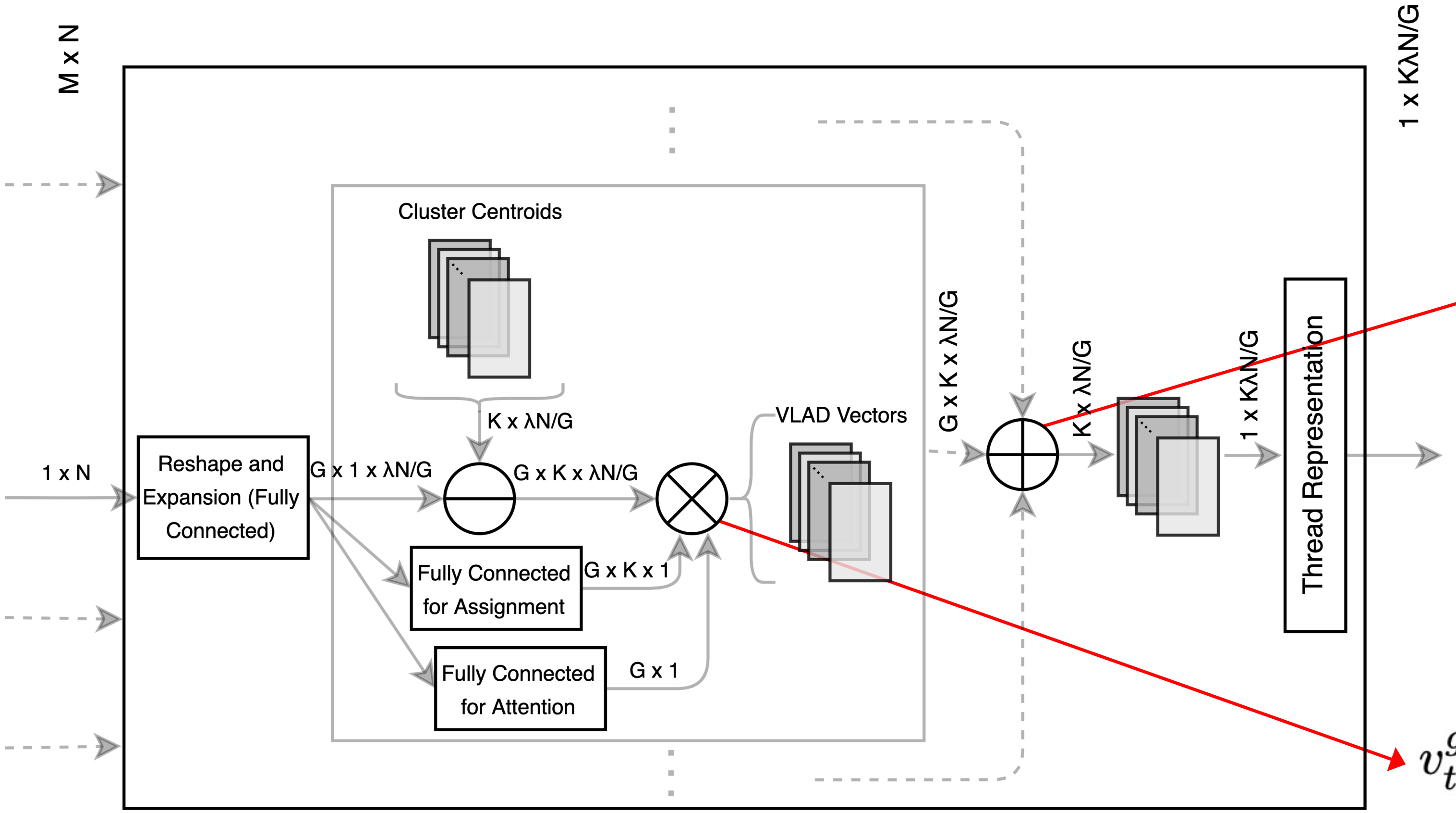
VLAD

- Vector of Locally Aggregated Descriptors
- Built on top of Bag of Visual Words
- Difference vector instead of presence frequency
  - Considering K clusters of all features



Vineeth N Balasubramanian, https://www.youtube.com/watch?v=CwJPEMcuAxY

Parametric Pooling: NeXtVLAD

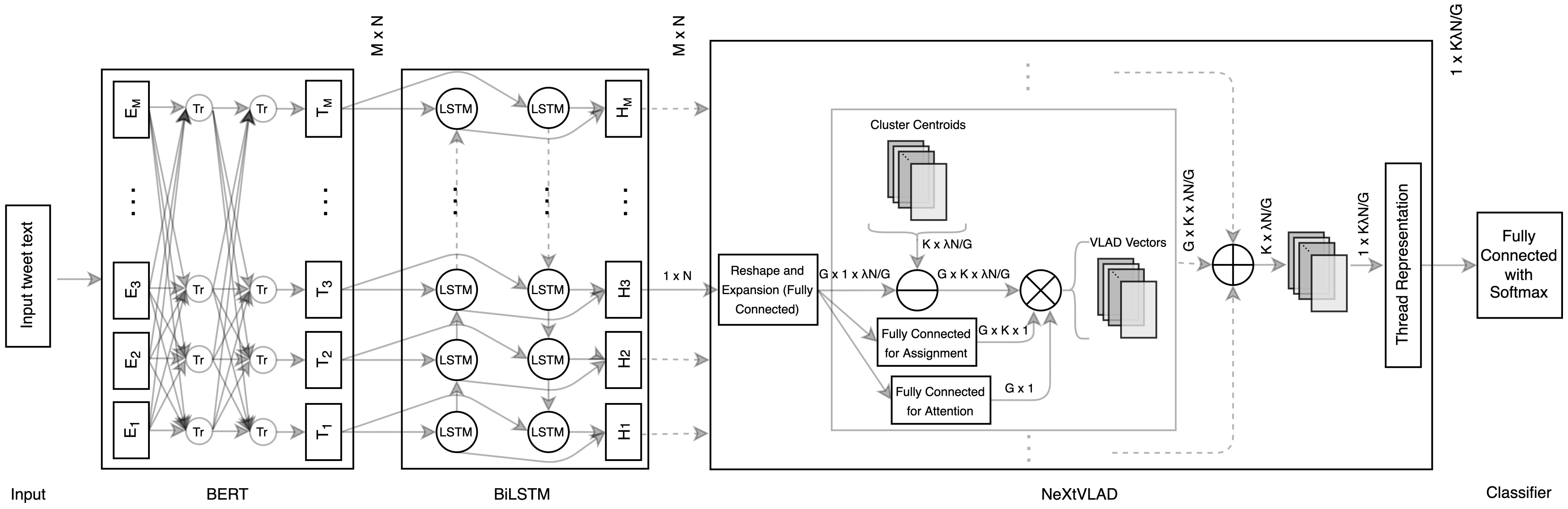


NeXtVLAD

$$r_{ki} = \sum_{t,g} v_{tki}^{g}$$

$$v_{tki}^{g} = \alpha_g(\dot{x}_t)\alpha_{gk}(\dot{x}_t)(\tilde{x}_{ti}^{g} - c_{ki})$$

The architecture



Experiments

- Architecture modifications:
  - $BERT_{Large}$
  - $XLNET_{Base}$
  - CTBERT
  - BERTweet
  - BiLSTM
  - NeXtVLAD
  - Average Pooling
  - Maximum Pooling

- Hyperparameter modifications:
  - Num. training epochs
  - Num. NeXtVLAD clusters
  - Learning rate

Experiments

For reported experiments in the paper:

Hyperparameter	Value
Num. of training epochs	8
Initial learning rate	10^{-6}
Batch size	1
Dropout rate	0.25
Num. of warmup steps	2000
Dropout rate	0.25
BERT's max length	512
NeXtVLAD's expansion	4
NeXtVLAD's num. of clusters	128

Experiments

Configuration (Accuracy)	**Model**	**Pre-Training**	**Pooling**	**num. of NeXtVLAD clusters**	**post-BiLSTM Operation**
Cfg 1 (0.92)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	128	Addition
Cfg 2 (0.91)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	2	Addition
Cfg 3 (0.92)	T+Cl	CTBERT-v2	—	—	—
Cfg 4 (0.88)	T+Bi+NV+Cl	$\text{BERT}_{\text{Large-Cased}}$	NeXtVLAD	2	Addition
Cfg 5 (0.91)	T+Bi+AP+Cl	CTBERT-v2	Avg Pooling	—	Addition
Cfg 6 (0.91)	T+Bi+MP+Cl	CTBERT-v2	Max Pooling	—	Addition
Cfg 7 (0.91)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	128	Concatenation
Cfg 8 (0.87)	T+Bi+NV+Cl	$\text{XLNET}_{\text{Base-Cased}}$	NeXtVLAD	128	Addition
Cfg 9 (0.91)	T+Cl	BERTweet	—	—	—
Cfg 10 (0.91)	T+Bi+NV+Cl	BERTweet	NeXtVLAD	128	Addition

Results

Model	Human			Bot			All
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [11]	0.91	0.88	0.89	0.89	0.97	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 1	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 2	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) Cfg 3	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) Cfg 4	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) Cfg 5	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) Cfg 6	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 7	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) Cfg 8	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) Cfg 9	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) Cfg 10	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier	0.83	0.81	0.82	0.82	0.83	0.82	0.82

Results

Model	Human			Bot			All
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [11]	0.91	0.88	0.89	0.89	0.97	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 1	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 2	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) Cfg 3	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) Cfg 4	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) Cfg 5	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) Cfg 6	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 7	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) Cfg 8	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) Cfg 9	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) Cfg 10	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier	0.83	0.81	0.82	0.82	0.83	0.82	0.82

Results

	Human			Bot			All
Model	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [11]	0.91	0.88	0.89	0.89	0.97	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 1	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 2	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) Cfg 3	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) Cfg 4	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) Cfg 5	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) Cfg 6	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 7	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) Cfg 8	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) Cfg 9	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) Cfg 10	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier	0.83	0.81	0.82	0.82	0.83	0.82	0.82

Results

Model	Human			Bot			All
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [11]	0.91	0.88	0.89	0.89	0.97	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 1	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 2	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) Cfg 3	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) Cfg 4	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) Cfg 5	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) Cfg 6	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 7	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) Cfg 8	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) Cfg 9	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) Cfg 10	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier	0.83	0.81	0.82	0.82	0.83	0.82	0.82

Results

Model	Human			Bot			All
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [11]	0.91	0.88	0.89	0.89	0.97	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 1	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 2	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) Cfg 3	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) Cfg 4	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) Cfg 5	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) Cfg 6	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) Cfg 7	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) Cfg 8	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) Cfg 9	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) Cfg 10	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier	0.83	0.81	0.82	0.82	0.83	0.82	0.82

Discussions and Conclusions

- Reinforce that domain-specific pretraining is important and can improve the performance
- NeXtVLAD achieves comparable performance to other pooling options
  - However, the performance jump is not enough to justify the computational cost of its incorporation
  - Needs further and deeper assessment for general conclusion

Discussions and Conclusions

- As the decoding strategies for text generation models are optimized to deceive humans, they introduce statistical abnormalities that help in automatic identification
- May not be the case if an attacker tunes the model in an adversarial setup

Discussions and Conclusions

- The only cost of deploying our trained model is a feed-forward pass through the network
  - → Can be used in real-time applications for bot-generated text detection

Future Directions

- Still room for improvement
- Defense against adversarial attacks
  - Robustness

Questions?

Thanks for your attention!

code/link to data @

https://github.com/sinamps/bot-detection

This work was supported in part by funds from NIST under award number 60NANB18D204, and from NSF under award number CNS 2027750, CNS 1822118 and from NIST, Statnett, Cyber Risk Research, AMI, ARL, and from DoE NEUP Program contract number DE-NE0008986.