Explanations in NLP

- Neural models in NLP have complex behavior. How can we understand them?
- QA: why did the model prefer *Stewart* over *Devin Funchess*?

QID: 1f4b668a0343453b9d4bf3edc86daf63
Question: who caught a 16-yard pass on this drive ?
Answer: devin funchess

Start Distribution

there would be no more scoring in the third quarter , but early in the fourth , the broncos drove to the panthers 41-yard line . on the next play , ealy knocked the ball out of manning 's hand as he was winding up for a pass , and then recovered it for carolina on the 50-yard line . a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal , cutting the panthers deficit to one score at 16â€“10 . the next three drives of the game would end in punts .

Lipton (2016)

Explanations in NLP

- Neural models in NLP have complex behavior. How can we understand them?
- Sentiment:

	DAN	Ground Truth
this movie was not good	negative	negative
this movie was good	positive	positive
this movie was bad	negative	negative
the movie was not bad	negative	positive

- Left side: predictions model makes on individual words
- Tells us how these words combine

Iyyer et al. (2015)

Why explanations?

- **Trust:** if we see that models are behaving in human-like ways and making human-like mistakes, we might be more likely to trust them and deploy them

- **Causality:** if our classifier predicts class *y* because of input feature *x*, does that tell us that *x* causes *y*? Not necessarily, but it might be helpful to know

- **Informativeness:** more information may be useful (e.g., predicting a disease diagnosis isn't that useful without knowing more about the patient's situation)

- **Fairness:** ensure that predictions are non-discriminatory

Lipton (2016)

What are explanations?

- Some models are naturally **transparent**: we can understand why they do what they do (e.g., a decision tree with <10 nodes)

- Explanations of more complex models

  - **Local explanations:** highlight what led to this classification decision. (Counterfactual: if they were different, the model would've predicted a different class)

  - **Text explanations:** describe the model's behavior in language

  - **Model probing:** auxiliary tasks, challenge sets, adversarial examples to understand more about how our model works

Lipton (2016); Belinkov and Glass (2018)