# Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans

Hyewon Jang†, Qi Yu† and Diego Frassinelli, University of Konstanz

## Research Questions

**How do Transform-based Language Models (TLMs) process figurative language?**

- **RQ 1:** Do TLMs attend to explicit cues that help identify certain figurative meaning?
- **RQ 2:** How does the feature attention behavior of TLMs compare to that of white-box models?

## Dataset

### FLUTE

- 4 figurative language classes with varying *opacity.*

### Obvious cues

[**Sarcasm**] I love$_{[positive\ sentiment]}$ how my house got so trashed$_{[negative\ event]}$.

[**Simile**] The cancer made her like$_{[comparison]}$ a dried flower.

### No obvious cues

[**Idiom**] Rule of thumb is escape while you're on the move.

[**Metaphor**] He felt a wave of excitement.

## Experimental Setup

### 1 Figurative Language Classification (4-way classification)

- **TLMs:** BERT, RoBERTa, XLNet.
- **White-box:** Logistic Regression, Random Forest with tf-idf.

### 2 Feature Importance Analysis Using SHAP

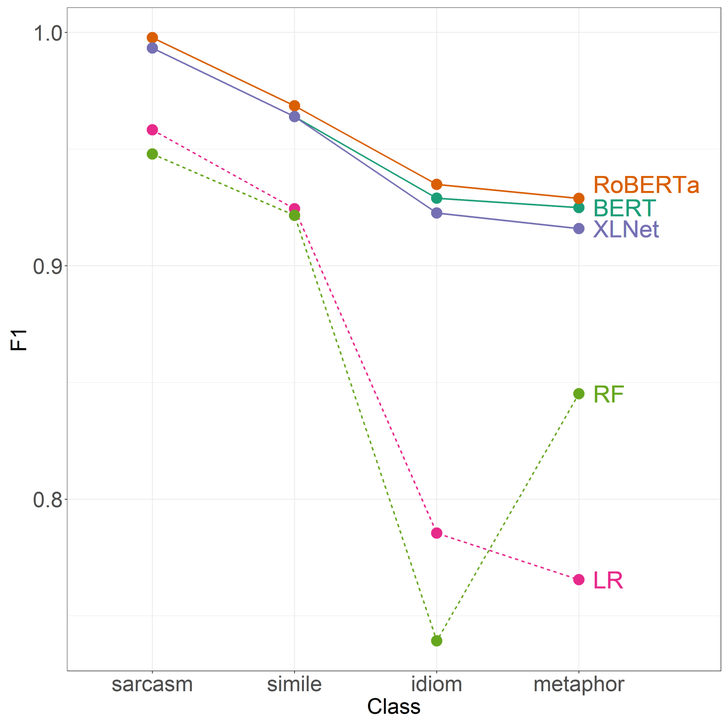
- Extract most important tokens per class with **SHAP**.

### 3 Feature Importance Analysis Using Online Experiments

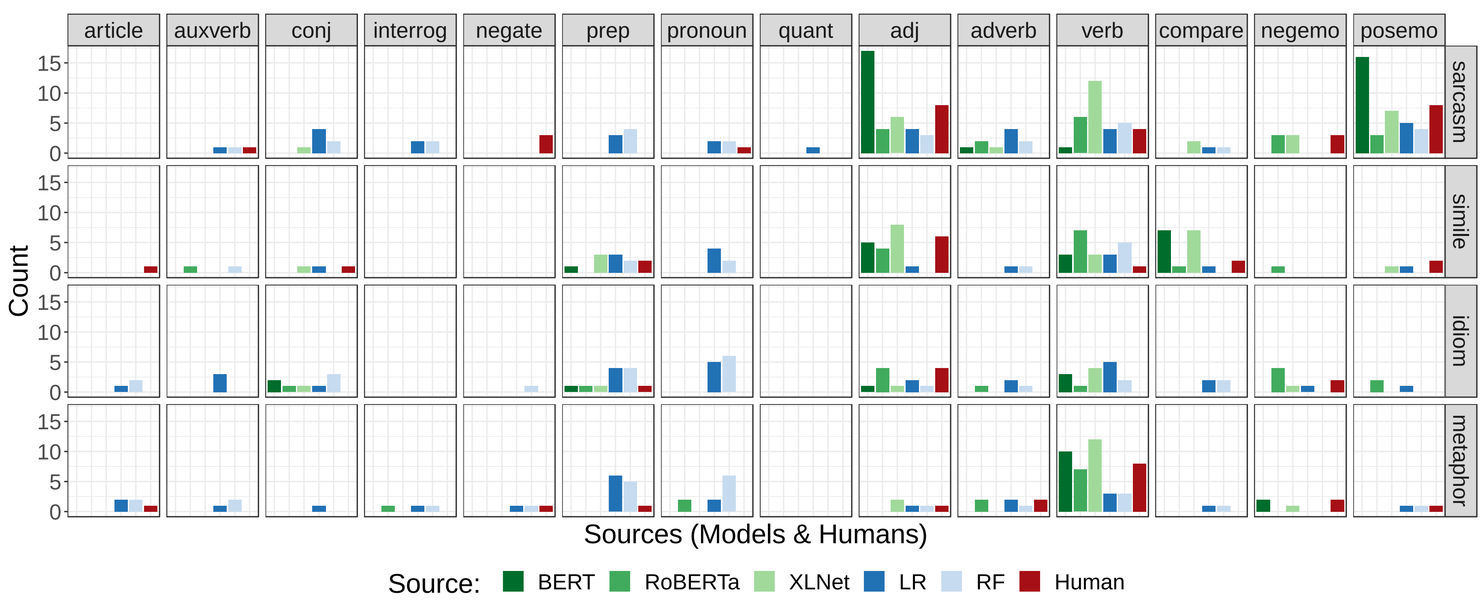
- Classification experiment on a subset of dataset.
- 56 sub-samples: 7 hard instances + 7 easy instances (per class).
- 15 annotators classified each instance and provided 1-3 most important words.

## Results

### Classification results by models (F1)




### 20 Top Features Mapped to Linguistically Motivated Categories (LIWC)




## Discussion

### Models vs. Models

5 most important features from SHAP

| | BERT | RoBERTa | XLNet |
|---|---|---|---|
| Sarcasm | refreshing, **thankful**, **proud**, **praised**, **thrilled** | increase, donated, videos, saving, boost | **safest**, refreshing, annoyed, scary, **love** |
| Simile | *resemble*, *resembled*, *like*, Arnold, predatory | mor, herd, slightest, movement, indicating | *like*, *resembled*, *resemble*, *similar*, *resembling* |

- BERT and XLNet show higher interpretability than RoBERTa for *sarcasm* and *simile*.
- White-box models focus on high-frequency function words (still opaque behavior).

### Cross-class comparison (Models)

- Sarcasm: positive emotion words.
- Similes: comparison words.
- Metaphors: verbs.
- Idioms: sporadic patterns.

### Models vs. Humans

- Difficult sentences for models are also difficult for humans.
- Confusions are rare for sarcasm or similes.
- Wrong judgments are always classified as idioms (highly conventionalized metaphors).
- Humans make choices based on the semantic information available.

### Cross-class comparison (Humans)

- Sarcasm: positive emotion words and adjectives.
- Similes: comparison words and adjectives.
- Metaphors: verbs.
- Idioms: sporadic patterns.