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

**Hyewon Jang**[⋆], **Qi Yu**[⋆], and **Diego Frassinelli**
Department of Linguistics, University of Konstanz, Germany
{hye-won.jang, qi.yu, diego.frassinelli}@uni-konstanz.de

## Abstract

Recent years have witnessed a growing interest in investigating what Transformer-based language models (TLMs) actually learn from the training data. This is especially relevant for complex tasks such as the understanding of non-literal meaning. In this work, we probe the performance of three black-box TLMs and two intrinsically transparent white-box models on figurative language classification of *sarcasm*, *similes*, *idioms*, and *metaphors*. We conduct two studies on the classification results to provide insights into the inner workings of such models. With our first analysis on feature importance, we identify crucial differences in model behavior. With our second analysis using an online experiment with human participants, we inspect different linguistic characteristics of the four figurative language types.

## 1 Introduction

In recent years, Transformer-based language models (TLMs) have achieved groundbreaking performance in various NLP tasks. Along with such progress, there has been an increasing demand for understanding the reasons for the decisions made by the TLMs, as this is often required for humans to trust the models (Gilpin et al., 2018).

As of now, researchers working on interpretability have mostly neglected the precise investigation of how TLMs models process non-literal language. Non-literal language, or *figurative language*, is a type of language where the intended meaning of an expression is incongruent with its literal meaning (Kalandadze et al., 2018; Gibbs and Colston, 2012). Typical cases of figurative language include *sarcasm*, e.g., saying 'lovely weather' on a stormy day, or *metaphors*, e.g., describing a person that always goes to bed late as a 'night owl' even though the person is not an actual owl. This discrepancy between the surface form and the intended message makes tasks involving figurative language complex both for humans and for models; therefore, it is harder for humans to trust the output of a model in such tasks without a precise understanding of the motivations behind specific models' decisions.

As the figurative meaning is not literally articulated in words, humans grasp it via pragmatic enrichment processes, i.e., inferring the speaker's communicative intention that is not uttered (Davis, 2019; Recanati, 2010; Grice, 1975). Although such processes often rely on non-deterministic factors such as social and cultural background (Colston and Katz, 2004), studies have shown that humans also utilize more explicit contextual cues to achieve pragmatic enrichment and identify the figurative meanings (see, e.g., Regel and Gunter, 2016; Kreuz and Caucci, 2007; Hanks, 2004; Kroll and Schepeler, 1987). These cues include contextual incongruity, semantic relations between words, or explicit syntactic forms. Given such multi-step nature of figurative language processing, we are interested in investigating the inner workings of TLMs in processing different types of figurative language. Specifically, we focus on two research questions (RQs): **RQ 1** - When the explicit cues that help the identification of the figurative meaning exist, do TLMs attend to them as humans do, or do TLMs adopt totally dissimilar strategies from humans? **RQ 2** - How do the performance and the feature attention behavior of TLMs compare to those of intrinsically interpretable white-box models such as regression models or decision-tree-based models? Would the attention mechanism enable them to grasp those cues better?

To explore these two questions, we probe three black-box TLMs along with two white-box models as baseline on the task of figurative language classification, using a dataset that provides a rich range of figurative language classes with different opacity degrees, i.e., some classes have obvious cue words, whereas others do not. Based on the

[⋆] Authors with equal contribution, listed alphabetically.

*Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832
July 9-14, 2023 ©2023 Association for Computational Linguistics

classification results, we conduct two analyses that compare 1) the behavior of different models and 2) the behavior of models vs. humans. Our main contributions are two-fold: First, we show that even though different TLMs achieve the same level of performance in the figurative language classification task, they show a striking discrepancy in the features they attend to, suggesting different levels of interpretability of different models. Second, we bridge existing work in psycholinguistics and theoretical linguistics with our data analysis results to gain a better understanding of figurative language processing in both machines and humans.[1]

## 2 Related Work

NLP researchers have attempted to build models that can comprehend figurative meaning. The public availability of large-scale annotated corpora, e.g., the Sarcasm Corpus V2 (Oraby et al., 2016), the VU Amsterdam Dataset of Metaphor (Steen, 2010), and the MAGPIE dataset for potentially idiomatic expressions (Haagsma et al., 2020), has encouraged the task of figurative language detection. Before the extensive use of neural networks, most studies have treated figurative language processing as a classification task and utilized theoretically-derived features. For example, incongruent sentiment expressions have often been used for sarcasm detection (Joshi et al., 2015; Riloff et al., 2013); while abstractness of words (Köper and Schulte im Walde, 2017; Turney et al., 2011) and topic transition information (Jang et al., 2016) have been used for metaphor detection. Recent studies have been using neural models (e.g., Gao et al., 2018; Wu et al., 2018; Do Dinh and Gurevych, 2016), and TLMs especially have shown good performance (see, e.g., Chakrabarty et al., 2022a, 2021; Avvaru et al., 2020; Dong et al., 2020; Liu et al., 2020). Some recent work using TLMs treats figurative language processing as a natural language inference (NLI) task instead of a classification task (He et al., 2022; Chakrabarty et al., 2021), which is a step closer to *comprehending* figurative language.

Despite the successful model performance of TLMs, little research has attempted to delve into their inner workings. Several studies have probed into whether knowledge of figurative meaning is encoded in TLMs (Chen et al., 2022; Dankers et al., 2022; Ehren et al., 2022; Tan and Jiang, 2021). Though this strand of work has confirmed that this knowledge is encoded in TLMs to some extent, it does not provide information about what motivates the output of the models in the task of figurative language processing. Our work attempts to fill this gap by inspecting the behavior of different models in processing different types of figurative language. We zoom into the most salient lexical properties used by different TLMs in distinguishing four types of figurative language and compare such properties with those used by humans.

## 3 Figurative Language Classification

As TLMs are not intrinsically explainable, one way of inspecting the reasons behind their decisions is by a post-hoc feature analysis. We treat figurative language processing as a classification task where the models classify 4 different types of figurative language. We then conduct analyses on the classification results to compare the behavior of 1) different models (Section 4) and 2) models vs. humans (Section 5). In this section, we report the results from our classification experiments using a variety of black-box and white-box models. All supplementary details of this experiment are provided in Appendix A.

### 3.1 Data

We use FLUTE (*Figurative Language Understanding through Textual Explanations*), an English-language dataset released for the Shared Task on Understanding Figurative Language 2022 (Chakrabarty et al., 2022b)[2]. We choose FLUTE as it is the most recent comprehensive dataset with a rich variety of figurative language types: *sarcasm*, *similes*, *idioms*, and *metaphors*. Even though the dataset is relatively small and imbalanced (see details below), we believe that it is beneficial for our research questions described in Section 1, as it brings together four different figurative language classes with varying lexical characteristics:

a) **Classes with apparent cues:** *Sarcasm* instances often contain words indicating positive sentiment and descriptions of a negative event or state (see Example (1)). *Simile* instances typically contain cues such as 'like' or 'as' (see Example (2)).

(1) Grad school was so comforting that I had no choice but to *drop out to keep my sanity*.

[1] All code, models, and experimental instructions are available at: https://github.com/CoPsyN/figurative-language-processing

[2] Publicly available at https://figlang2022shareddtask.github.io

(2) He was as graceful as a giraffe.

b) **Classes without apparent cues:** *Idioms* (see Example (3)) and *metaphors* (see Example (4)) do not come with obvious cues.

(3) Rule of thumb is escape while you're on the move.

(4) He felt a wave of excitement.

We assume that varying opacity degrees of different figurative language types provide a good test-bed for comparing the behavior of different models. Specifically, for the classes with obvious contextual cues, we investigate how well different models can capture these cues; for the classes without obvious contextual cues, we investigate how such models use contextual information to overcome the lack of clear cues.

As FLUTE is originally designed for an NLI task, each figurative sentence is the *hypothesis* paired up with its literal counterpart, the *premise*. For the purpose of our experiment, we reorganize the dataset by extracting all the *hypotheses* together with their original labels. In the original dataset, the same hypotheses are sometimes paired up with different premises and thus appear multiple times. We drop duplicates of such kind. As our focus is investigating the behavior of models and humans in processing *figurative language types with different characteristics* (with vs. without apparent cues), but not investigating *figurative language as a whole as opposed to literal language*, in our experiments we exclude the premises (i.e., the literal sentences). Table 1 summarizes the dataset after reorganization.

| | **With Apparent Cues** | | **No Apparent cues** | | |
|---|---|---|---|---|---|
| | Sarcasm | Simile | Idiom | Metaphor | **Total** |
| #Sentences | 2212 | 625 | 884 | 621 | 4342 |
| #Tokens | 45233 | 9062 | 14795 | 5692 | 74782 |

Table 1: Number of sentences and tokens for each figurative language class used in the analyses.

### 3.2 Models

We experiment with two types of models: *black-box* and *white-box*. Black-box models are the models whose predictions cannot be directly explained in ways that humans can understand whereas white-box models are the ones whose predictions can be interpreted at least by experts (Islam et al., 2021; Loyola-Gonzalez, 2019; Rudin, 2019). Given that the detection of figurative language is not the objective of this work, we fine-tune these models on our dataset and add a sequence classification head to identify the strongest lexical patterns characterizing various figurative language types (see Appendix A for details).

**Black-Box Models** We experiment with three TLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). All three models have frequently been used in former studies on figurative language processing and shown good performance (see studies mentioned in Section 2).

**White-Box Models** We experiment with four white-box models: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB). As input of the white-box models, each text is represented as a Tf-idf vector with the number of dimensions equaling to the vocabulary size of our dataset. We do not conduct any token selection (e.g., excluding infrequent tokens and/or stop words) to keep the features (i.e., the tokens) for the white-box models maximally comparable to those for the black-box models. We are aware though that it is impossible to keep the token sets for different models completely identical because they use different tokenizers and token representations. The use of Tf-idf as text representations guarantees the white-box models to be completely transparent for humans as each vector dimension corresponds to a word. This contrasts to representing a sentence as an average of pre-trained static word embeddings (e.g., Word2Vec or GloVe embeddings) as they are opaque by definition and averaging them adds another layer of opacity.

### 3.3 Results

Given the relatively small size of the dataset, we evaluate the performance of each model using a 10-fold cross-validation instead of a single hold-out test set. Table 2 shows the macro-F1 scores averaged from 10 folds: Among the white-box models, LR achieves the best performance, followed by RF. Among the black-box models, all three of them perform to a comparable degree.

### 3.4 Model Selection

In order to obtain informative features for the analysis of the model behavior, we only select the models with good performance in figurative language classification. As indicated by the F1-scores, TLMs

| Model | Macro-F1 |
|---|---|
| BLACK-BOX MODELS | |
| `bert-base-uncased` | **0.95** |
| `roberta-base` | **0.95** |
| `xlnet-base-cased` | 0.94 |
| WHITE-BOX MODELS | |
| `Logistic Regression (LR)` | **0.87** |
| `Decision Tree (DT)` | 0.77 |
| `Random Forest (RF)` | 0.85 |
| `Naive Bayes (NB)` | 0.68 |

Table 2: Macro-F1 from 10-fold cross-validation.

outperform white-box models by a large margin. But, given that white-box models constitute an inherently interpretable baseline, we include the two best-performing white-box models as reference points for our feature analysis. Figure 1 provides the per-class F1 scores of the selected models on the test set (20% of the full dataset) that we used for all of them for better comparability (see more details in Appendix B). Apart from the fact that the black-box models perform better than the white-box models for all classes, all models perform better in detecting the classes with obvious cues (*sarcasm* and *simile*) than the classes without obvious cues (*idiom* and *metaphor*).

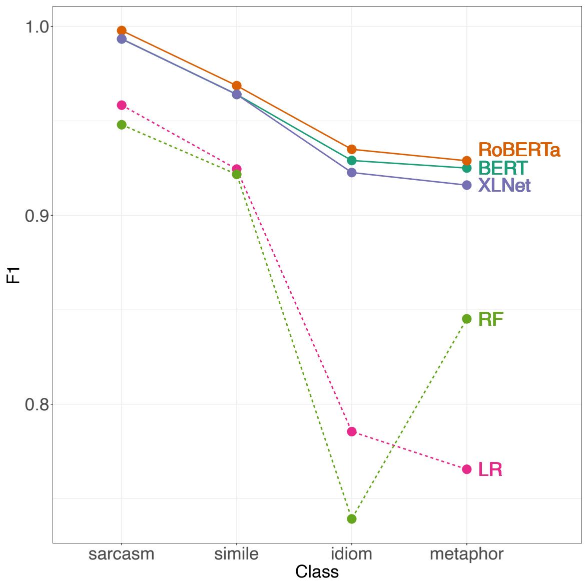


Figure 1: Per-class F1 scores of the selected models.

## 4 Feature Analysis 1: Models vs. Models

In our first analysis, we investigate the impact of each feature on the model predictions. We compare the features that different models deem important for each figurative language class (cross-model comparison). We also identify the common behavior of the five models for each figurative language class (cross-class comparison).

### 4.1 Methods

To maximize the comparability of our results, we use just one feature analysis method for all models: Shapley Additive Explanations (SHAP; Lundberg and Lee, 2017). SHAP returns the feature importance values by computing the Shapley Values of each feature, i.e., the feature's contribution towards a certain output of the model. Aside from its growing popularity in model explainability, we choose SHAP because it can be used for all the models selected in our analysis (both TLMs and white-box models). Also, SHAP is model-agnostic and provides *global* feature importance analysis methods based on the aggregations of Shapley Values, which allows us to compare and inspect the overall behavior of the models. Finally, SHAP allows us to conduct a per-class feature importance analysis, which is beneficial for our purpose of investigating the model behavior in processing different figurative language types.

For each model, we extract the top 20 features (i.e., tokens) with the highest mean Shapley Values for each figurative language class (*sarcasm*, *simile*, *idiom*, *metaphor*) as they are the most important features in the classification task.

To provide a linguistically informed and human-interpretable feature importance overview, we categorize the extracted tokens using the selected categories described below. For this mapping, we use LIWC (*Linguistic Inquiry and Word Count*; Pennebaker et al., 2015), a dictionary-based software that automatically maps individual words to linguistically motivated conceptual categories.

- **Function Words:** *articles*, *auxiliary verbs*, *conjunctions*, *interrogatives*, *negations*, *prepositions*, *pronouns*, *quantifiers*.
- **Content Words**[3]: *adjectives*, *adverbs*, *comparisons*, *verbs*.

[3] The original LIWC dictionaries for these four content words categories only cover *common* (i.e., high-frequency) words. To minimize this dictionary coverage bias, we manually map the words that belong to these categories but were not covered by LIWC to their corresponding categories.

- **Sentiment Words:** *negative emotion words*, *positive emotion words*.

The categories *function words* and *content words* are intentionally chosen to investigate the syntactic components each model attends to. We also include the category *comparisons* (subsumed under *content words*) and *sentiment words* because we assume they are typical cues for the classes *sarcasm* and *simile*, as mentioned in Section 3.1. We are interested in inspecting whether the models are able to actively use these cues for the classification.

## 4.2 Results

Figure 2 shows the results of the most important features given by the five models mapped to the selected linguistic categories. An elaborated list of all extracted tokens is given in Appendix C.

### 4.2.1 Cross-Class Comparison

**Classes With Apparent Cues** As shown in Figure 2, the class *sarcasm* displays the most obvious pattern: The category that most models attend to is *positive emotion words* (*posemo*). The category *adjective* (*adj*) also shows a high count, because most of the positive emotion words are also adjectives. As mentioned in Section 3.1, sarcasm instances in the dataset typically use a positive sentiment to describe a negative situation. Our feature analysis shows that models are generally able to capture these cues.

For *simile*, it can be observed from Figure 2 that four out of the five models (BERT, RoBERTa, XLNet, LR) attend to the category *comparisons* (*compare*), which contains the typical cues for similes including 'like' and 'as' (see examples in Section 3.1). The count values of the *comparison* words is not particularly high because not many variations exist for *comparison* words in the dataset (each variant adds 1 to the total count represented by the y-axis). Upon a closer inspection, we find that such words have the highest ranking among the top most important features in BERT and XLNet (BERT: 'resemble', 'resembled', 'like'; XLNet: 'like', 'resembled', 'resemble', 'similar', 'resembling'; see Appendix C for details), indicating that they can successfully capture these cues. These models also have a relatively high focus on *adj*, possibly because the words adjacent to the comparison words are often adjectives, e.g., 'Her words were like a sharp blade'.

**Classes Without Apparent cues** Despite our initial assumption that *metaphors* provide no apparent cues that models can rely on, all TLMs show relatively stronger attention to verbs for *metaphor* compared to the other classes (*sarcasm*, *simile*, and *idiom*). In fact, this observation is in line with what previous work has suggested, that verbs play a crucial role in the understanding of metaphors (Gibbs et al., 1997), as verbs are often the major component in creating metaphorical sentences (*predicative metaphor*; Glucksberg and McGlone, 2001). As such, some psycholinguistic work (Feng and Zhou, 2021; Chen et al., 2008; Wilson and Gibbs Jr, 2007) and computational work (Song et al., 2021) have been specifically dedicated to predicate metaphors.The words that the models attend to for *idiom* appear to show less transparent patterns: for all the models, these features show a sporadic pattern across the linguistic categories.

### 4.2.2 Cross-Model Comparison

**Black-Box vs. Black-Box** With a closer inspection of the top-ranking features of the five models for the two classes with obvious cues, i.e., *sarcasm* and *simile*, we find that RoBERTa shows a considerably different behavior compared to BERT and XLNet: Whereas BERT and XLNet focus on the expected features to classify these two classes, RoBERTa focuses on disparate features. This can be observed from the top 5 most important features of each model for *sarcasm* and *simile* in Table 3 (see Appendix C for details).

**Black-Box vs. White-Box** Interesting contrasts between these two model types emerge when inspecting the categories of the 20 most important features: As shown in Figure 2, white-box models show stronger attention to function words like *prepositions* and *pronouns* than the black-box models, whereas the black-box models attend to the content words more than the white-box models. This indicates that in the presence of function words, usually high-frequency tokens contributing little to the characterization of specific figurative language classes, TLMs are better than white-box models at tuning down their importance and capturing the more prominent cues. This difference could explain the overall higher performance of TLMs in all classes compared to white-box models, besides the fact that the Tf-idf vectors used as input for the white-box models are sparse and, usually, outperformed by dense vectors.

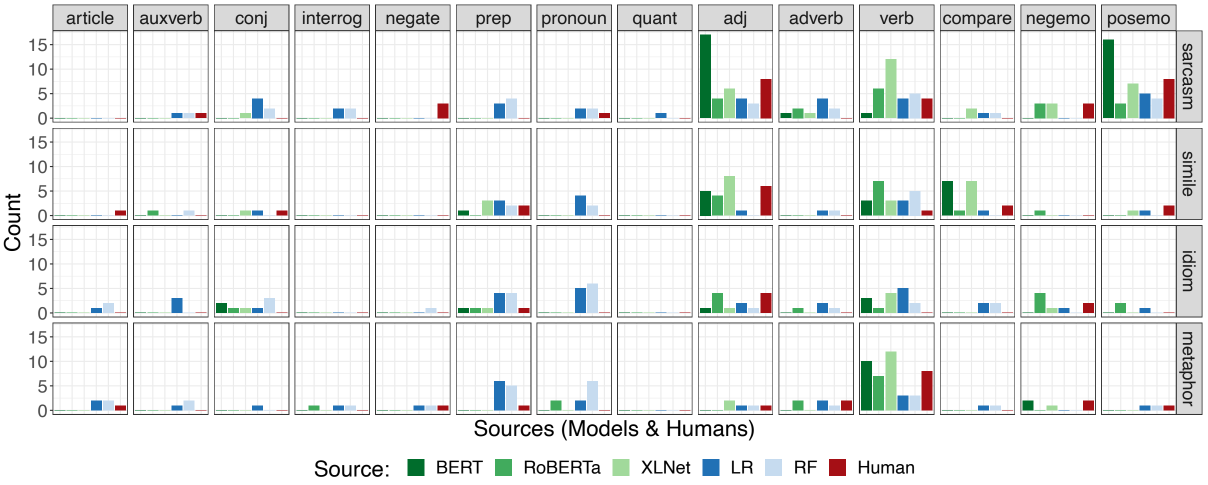


Figure 2: Mapping of the 20 most important features for the five models (in the order of BERT, RoBERTa, XLNet, LR, RF) to the corresponding linguistic categories. We also include the feature analysis results for the experiment with human participants (the rightmost bar in red) discussed in Section 5.

| | 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* |

Table 3: Top 5 most important features (tokens) of the black-box models for *sarcasm* and *simile*. The features are sorted in descending order by their SHAP values. **Purple**: positive emotion words; *Yellow*: comparison words.

### 4.3 Discussion

In summary, the behavior of BERT and XLNet with regard to *sarcasm* and *similes* are largely interpretable: As mentioned in Section 3.1, the results indicate that these models are good at capturing relevant cues for these classes (**RQ 1**). On the contrary, RoBERTa does not attend to these cues but to tokens that are difficult to linguistically motivate. White-box models tend to focus on high-frequency function words. This suggests that, even though white-box models are intrinsically interpretable, it is still difficult to identify the real motivations behind their output from the human's perspective (**RQ 2**).

For all five models, no clear patterns are observed among the most important features for the class *idiom*: Neither from the mapping results nor from a manual inspection of these tokens. *Idiom* is also one of the two classes, together with *metaphor*, where most models performed worst (see Figure 1). A precise classification of idiomatic sentences using lexical information is clearly very difficult for the models. This is not surprising considering that there are no obvious cues that the models can rely on because the vocabularies used in each expression of these two classes are highly idiosyncratic.

Verbs are the most important tokens for detecting *metaphors* for all TLMs, evidence supported by previous findings in theoretical research. However, even with the obvious cues, most models struggle with *metaphors*, which suggests that additional information than the identified cues is needed. One possible reason for the added difficulty could be the limited context provided in the dataset as the successful identification of *metaphors* in the text usually requires a larger amount of contextual information (Lemaire and Bianco, 2003; Inhoff et al., 1984; Ortony et al., 1978).

## 5 Feature Analysis 2: Models vs. Humans

The classification results in Section 3 suggest that models especially struggle with identifying *idioms* and *metaphors*. These results are in line with various studies in cognitive science and psycholinguistics showing that the processing of idioms and metaphors is also complex for humans. Idioms are defined as “constructions whose meanings cannot be derived from the meanings of its constituents" (Glucksberg and McGlone, 2001). Therefore, identifying idioms often relies on memory retrieval rather than syntactic and semantic analyses (Glucksberg and McGlone, 2001) and the speaker's familiarity to them (Cronk and Schweigert, 1992; Gibbs, 1980). Similarly, the difficulty of and the speaker's familiarity to metaphors are factors that influence metaphor processing (Schmidt and Seger, 2009). Drawing in on these intricacies, we build a classification task for human participants, aiming to investigate how human behavior differs from model behavior in figurative language classification, and whether humans also struggle more with identifying idioms and metaphors.

### 5.1 Methods

We extract the sentences that are misclassified by at least two models from the test set, as we assume that they are particularly tricky instances and thus interesting to inspect whether they are also difficult for humans. These include 7 *sarcasm*, 10 *simile*, 72 *idiom*, and 47 *metaphor* instances. To have a balanced number of sentences per class, for each class we randomly sample 7 misclassified sentences (henceforth, *difficult instances*). We also include 7 correctly classified sentences by all of our models as a control group (henceforth, *easy instances*), selecting 56 sentences in total. We ask 15 English native speakers based in the UK and the USA to classify these 56 sentences (presented in a randomized order) into one of the four classes (multiple-choice questions) and provide 1-3 words in each sentence that they consider as the most relevant for their classification decisions. We also add 3 attention-check questions where we ask participants to provide a keyword from the previous sentence. We conduct the experiment online using Google Forms[4] and Prolific[5]. The average duration was 26 minutes. Participants received a compensation of 9£/hour, a fair wage suggested by Prolific.

### 5.2 Results

#### 5.2.1 Human Classification Results

We collect the classification labels given by human participants for *easy* and *difficult* instances. For each instance, we extract the classification label that received the most votes by the participants (henceforth, *majority label*) and compare it with the ground-truth label.

Figure 3 depicts the proportions of the ground truth labels across majority labels for *difficult* and *easy* instances. Confusions are rarely found for *sarcasm* or *similes*. In contrast, more instances of *metaphors* and *idioms* are incorrectly classified. We observe that humans struggle in identifying *metaphors* more than *idioms*, which is in line with model behavior (see Figure 1). Whereas *metaphors* require more semantic processing, identifying *idioms* mainly requires the use of memory retrieval (Glucksberg and McGlone, 2001). Lastly, we find that when humans make 'wrong' judgments, they always classify instances of *metaphors* as *idioms*. Upon manual inspection, we find that most *metaphor* instances misclassified as *idioms* are highly conventionalized expressions (e.g., 'John <u>fell behind</u> his class mates').

Difficult instances for the models are also more difficult for humans. Among the *easy* instances, however, there is an exception to this general tendency, where 43% of these instances received the majority label *sarcasm*. The sentences in (5) - (6) are examples of this type of wrong classification.

(5) I wanted that gift as much as cancer.

(6) The formula was as well-known as the eleventh president of Zambia.

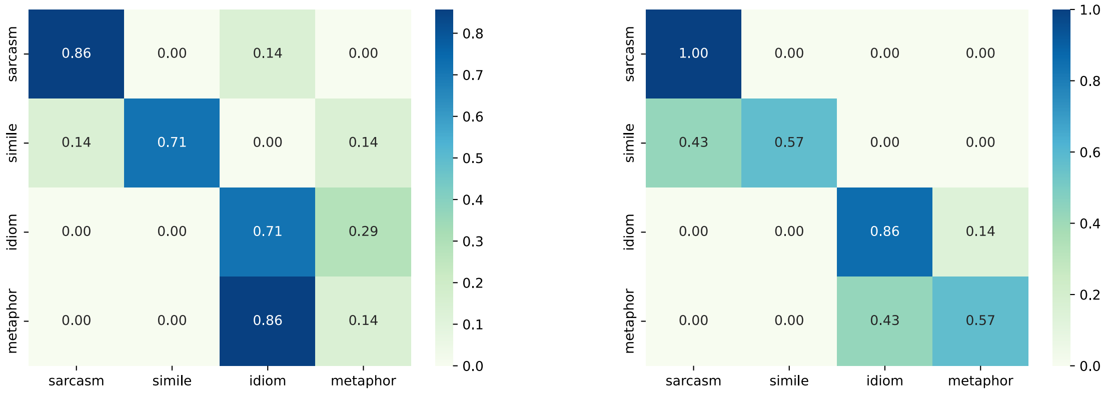
Whereas these examples are only labeled as *simile* in the dataset, it is evident that they could also be instances of *sarcasm*. Such instances have occurred possibly because the labeling scheme of our experiment was different from that of the original dataset: The 4 classes of figurative language in FLUTE stem from 4 different sources and the potential overlap between different labels was left unchecked. However, participants in our experiment had to select only one of the four classes for each statement, thus resulting in occasional confusion instances. Nevertheless, these examples reveal an intriguing pattern about human behavior in processing figurative language. When an instance could belong to more than one figurative language type, humans tend to make a choice based on the semantic information available. With these instances excluded, our assumption is confirmed: *Idioms* and *metaphors* are more opaque than *sarcasm* and *similes* and thus pose more difficulty for both humans and models.

#### 5.2.2 Important features

We aggregate all the words that the participants reported to have had the most influence on their classification decisions. For each class, we select

[4] https://forms.google.com

[5] https://www.prolific.co



(a) difficult

(b) easy

Figure 3: Normalized confusion matrix of the majority labels by human participants (x-axis) and the true labels provided in the dataset (y-axis) among 'difficult' (left) and 'easy' (right) instances.

the 20 most mentioned words (ties included) and map them to the linguistic categories mentioned in Section 4 (see Figure 2). Figure 2 shows that humans attend to *positive emotion words* and the related *adjectives* to identify sarcasm, cues that are also deemed important by BERT and XLNet. For *simile*, humans report *adjectives* as being the most indicative cues for their decision, followed by *comparison* words. BERT and XLNet also attend to *adjectives* and *comparison words*, but unlike the models, humans attend to adjectives more. This could be explained by findings in previous research that function words do not elicit more activation in the human brain than the content words (Diaz and McCarthy, 2009). Human participants also show a high degree of attention to verbs for *metaphor*, compared to other classes (*sarcasm*, *simile*, and *idiom*) and compared to all the linguistic categories. From the sentences that human participants correctly identified as metaphors (5 of all 14 sentences), we find that the most frequently mentioned words are always verbs (e.g., 'The tax cut will fertilize the economy.'). The result once again indicates the general importance of verbs in processing metaphors. No apparent patterns are found in the words that humans deemed most important for *idioms*.

### 5.3 Discussion

The results from the human annotation experiment show that the features that humans focus on to process different types of figurative language are largely in line with the features that BERT and XLNet attended to (**RQ 1**). The results also suggest that the degree of difficulty for humans in detecting different figurative language types generally matches the difficulty for machine learning models. A clear pattern is shown as to the opacity of the four figurative language classes: *Sarcasm* and *similes* are more transparent to detect, followed by *idioms* and then by *metaphors*. Our finding supports the assumption of Kreuz and Caucci (2007) that sarcasm can also be more formulaic than one might assume. However, future work should also investigate sarcastic sentences in a non-formulaic structure to have a full grasp of model performance in sarcasm processing.

## 6 Conclusion

With our two experiments, we provide insights into both the behavioral differences between different language models and the varying linguistic properties of several figurative language types. Our first analysis reveals contrasting behavior among the black-box models. This highlights different degrees of interpretability of the TLMs in the task of figurative language processing despite their similar performance. It also provides evidence-based indicators for choosing the best model that deals with rich linguistic information in an extended range of NLP applications. In the second analysis with human participants, we show that the general tendency found in the performance of all models is aligned with that of human participants; this manifests the varying complexity levels of different figurative language types.

## Acknowledgements

We thank Agnieszka Faleńska for reviewing our initial draft and providing insightful feedback that contributed to this final version.

## Limitations

All our experiments and the discussions thereof are based on a single dataset. Future work investigating similar topics should also involve datasets that provide more syntactic and semantic variations. We also acknowledge that the linguistic categories (syntactic categories including *function words*, *content words*, as well as *sentiment words*) we selected to conduct the feature analysis may not encompass all the properties relevant for figurative language processing. Lastly, a larger sample size of stimuli for the human experiment might be needed for a more robust support for the findings in this paper.

## Ethics Statement

We used and cited publicly available data and libraries for our experiments. According to the creators of the dataset FLUTE (see Section 3.1), the dataset does not contain any offensive context or information that uniquely identifies individuals.

Our experiment with human participants reported in Section 5 was carried out entirely anonymously and voluntarily. No personal information of the participants can be inferred from the collected data. The experiment is in line with the ethical regulations of the of the University of Konstanz (IRB 05/2021).

## References

Adithya Avvaru, Sanath Vobilisetty, and Radhika Mamidi. 2020. Detecting Sarcasm in Conversation Context Using Transformer-Based Models. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 98–103, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Evan Chen, Page Widick, and Anjan Chatterjee. 2008. Functional–anatomical organization of predicate metaphor processing. *Brain and language*, 107(3):194–202.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5875–5887, Dublin, Ireland. Association for Computational Linguistics.

Herbert L Colston and Albert N Katz. 2004. *Figurative language comprehension: Social and cultural influences*. Routledge.

Brian C Cronk and Wendy A Schweigert. 1992. The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics*, 13(2):131–146.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michele T Diaz and Gregory McCarthy. 2009. A comparison of brain activity evoked by single content and function words: an fmri investigation of implicit word processing. *Brain research*, 1282:38–49.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Xiangjue Dong, Changmao Li, and Jinho D. Choi. 2020. Transformer-based context-aware sarcasm detection in conversation threads from social media. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 276–280, Online. Association for Computational Linguistics.

Rafael Ehren, Laura Kallmeyer, and Timm Lichte. 2022. An analysis of attention in German verbal idiom disambiguation. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 16–25, Marseille, France. European Language Resources Association.

Yin Feng and Rong Zhou. 2021. Does embodiment of verbs influence predicate metaphors in a second language? Evidence from picture priming. *Frontiers in Psychology*, page 5036.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Raymond W Gibbs. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory & cognition*, 8(2):149–156.

Raymond W Gibbs, Josephine M Bogdanovich, Jeffrey R Sykes, and Dale J Barr. 1997. Metaphor in idiom comprehension. *Journal of memory and language*, 37(2):141–154.

Raymond W Gibbs and Herbert L Colston. 2012. *Interpreting figurative meaning*. Cambridge University Press.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Sam Glucksberg and Matthew S McGlone. 2001. *Understanding figurative language: From metaphor to idioms*. 36. Oxford University Press on Demand.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Patrick Hanks. 2004. The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7875–7887, Dublin, Ireland. Association for Computational Linguistics.

Albrecht Werner Inhoff, Susan D Lima, and Patrick J Carroll. 1984. Contextual effects on metaphor comprehension in reading. *Memory & Cognition*, 12(6):558–567.

Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. 2021. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rosé. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 216–225, Berlin, Germany. Association for Computational Linguistics.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Tamar Kalandadze, Courtenay Norbury, Terje Nærland, and Kari-Anne B Næss. 2018. Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain. Association for Computational Linguistics.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Neal EA Kroll and Eva M Schepeler. 1987. The comprehension and recall of similes as a function of context and cue. *Journal of psycholinguistic research*, 16(2):101–132.

Benoît Lemaire and Maryse Bianco. 2003. Contextual effects on metaphor comprehension: Experiment and simulation. In *Proceedings of the 5th international conference on cognitive modeling (ICCM2003)*, pages 153–158.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop* *on Figurative Language Processing*, pages 250–255, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Andrew Ortony, Diane L Schallert, Ralph E Reynolds, and Stephen J Antos. 1978. Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of verbal learning and verbal behavior*, 17(4):465–477.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin.

Francois Recanati. 2010. Pragmatic enrichment. In Delia Fara and Gillian Russell, editors, *Routledge Companion to Philosophy of Language*, pages 67–78. Routledge.

Stefanie Regel and Thomas C Gunter. 2016. What exactly do you mean? ERP evidence on the impact of explicit cueing on language comprehension. *Preproceedings of Trends in Experimental Pragmatics*, pages 115–120.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Gwenda L Schmidt and Carol A Seger. 2009. Neural correlates of metaphor processing: The roles of figurativeness, familiarity and difficulty. *Brain and cognition*, 71(3):375–386.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251, Online. Association for Computational Linguistics.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nicole L Wilson and Raymond W Gibbs Jr. 2007. Real and imagined body movement primes metaphor comprehension. *Cognitive science*, 31(4):721–731.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

## A Setup Details of Feature Analysis 1

For all experiments reported in Section 4, a random seed of 45 was used. Other hyperparameter settings are provided below.

**Black-Box Models** All black-box models were implemented using the Hugging Face's *Transformers* library.[6] All models were fine-tuned for 4 epochs with a learning rate of 2e-5 and a batch size of 16. The fine-tuning was conducted on a Quadro RTX 5000 GPU with a total memory of 16GB. As the dataset size is relatively small, for all models each training epoch was finished under 15 seconds.

**White-Box Models** All white-box models were implemented using the *scikit-learn* library (Pedregosa et al., 2011).[7] Table 4 summarizes the hyperparameters used. For the hyperparameters not specified in the table, the default values from *scikit-learn* were used.

| **Model** | **Hyperparameters** |
|---|---|
| LR | `solver` = 'sag',<br>`multi_class`='multinomial' |
| RF | `n_estimators` = 100 |

Table 4: Hyperparameters for the white-box models.

## B Model Performance

Table 5 shows the precision, recall and F1 of all models for each figurative language class. Figure 4 shows the confusion matrices of all models.

## C Most Important Features

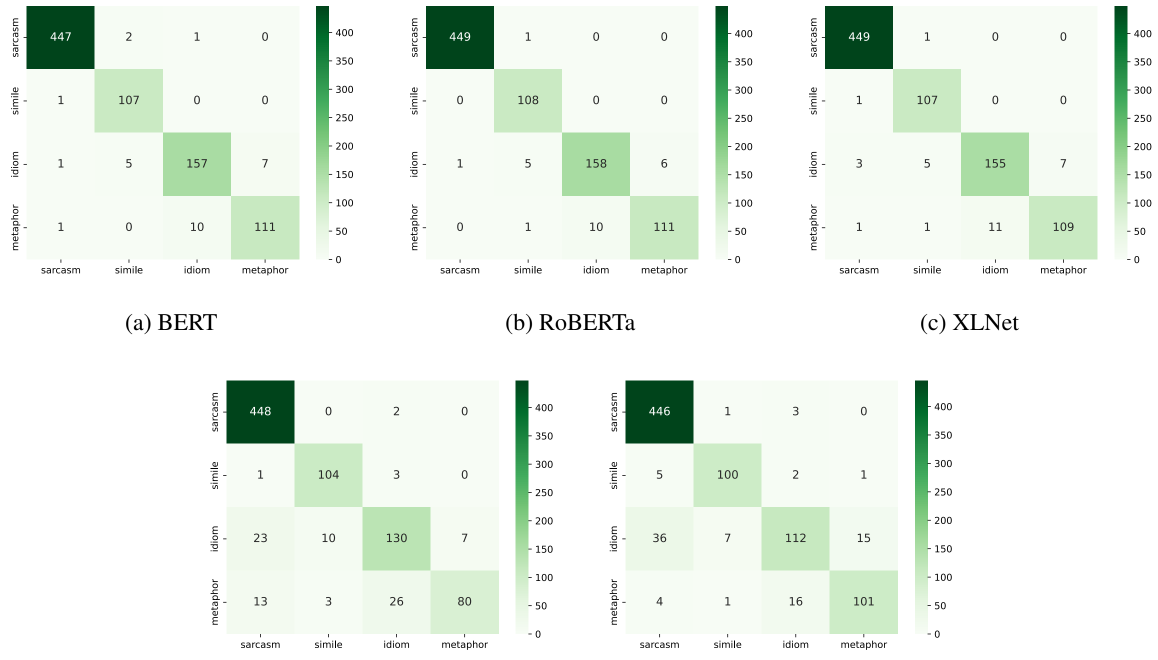
Tables 6-7 illustrate the per-class most important features extracted from the models and the human annotation experiment. For each class, we extracted the top 20 most important tokens (see Section 4).

[6] https://huggingface.co/docs/transformers/main/en/index

[7] https://scikit-learn.org/stable/

|  | BLACK-BOX MODELS |  |  |  |  |  |  |  |  | WHITE-BOX MODELS |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | BERT |  |  | RoBERTa |  |  | XLNet |  |  | LR |  |  | RF |  |  |
|  | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Sarcasm** | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.99 | 0.92 | 1.00 | 0.96 | 0.91 | 0.99 | 0.95 |
| **Simile** | 0.94 | 0.99 | 0.96 | 0.94 | 1.00 | 0.97 | 0.94 | 0.99 | 0.96 | 0.89 | 0.96 | 0.92 | 0.92 | 0.93 | 0.92 |
| **Idiom** | 0.93 | 0.92 | 0.93 | 0.94 | 0.93 | 0.93 | 0.93 | 0.91 | 0.92 | 0.81 | 0.76 | 0.79 | 0.84 | 0.66 | 0.74 |
| **Metaphor** | 0.94 | 0.91 | 0.93 | 0.95 | 0.91 | 0.93 | 0.94 | 0.89 | 0.92 | 0.92 | 0.66 | 0.77 | 0.86 | 0.83 | 0.85 |

Table 5: Precision (**P**), recall (**R**) and **F1** of all models for each figurative language class.



(a) BERT

(b) RoBERTa

(c) XLNet

(d) LR

(e) RF

Figure 4: Confusion matrices of the selected models (x-axis: predicted labels; y-axis: true labels).

| Class | Most Important Features | | |
|---|---|---|---|
| | BERT | RoBERTa | XLNet |
| **Sarcasm** | refreshing, thankful, proud, praised, thrilled, glad, awesome, delighted, excited, incredible, terrific, adorable, amazed, wonderful, fascinated, amazing, delightful, happily, planet, fantastic | increase, donated, videos, saving, boost, ankle, healing, elt, cried, attend, personally, satisfaction, celebrities, organization, somehow, shaken, frustrating, civic, cute, cooking | safest, refreshing, annoyed, scary, love, irritating, commend, cheered, hottest, ensured, celebrating, enjoying, pleasing, encourages, afterwards, approve, cheering, adore, makes, because |
| **Idiom** | differ, aye, halves, sevens, shove, eddie, nods, ways, guess, matthias, platt, hilt, plank, james, ava, plus, meantime, overboard, or, daylight | moist, damned, theory, arnold, eddie, until, toppled, caps, english, devils, els, nicely, production, hell, playing, words, fucking, jon, palace, bone | sticks, wire, messenger, beans, halves, lend, sides, hatch, platter, record, mark, splash, naked, broth, plus, hook, wolf, thieves, trades, source |
| **Simile** | resemble, resembled, like, arnold, predatory, titanium, resembling, slug, transparent, compared, twilight, charcoal, resemblance, similar, alligator, fragile, magazine, turtle, calculus, locomotive | mor, herd, slightest, movement, indicating, spicy, had, shield, disappears, understands, lining, messy, ays, colleague, resemble, balanced, towers, noble, descended, nationality | like, resembled, resemble, similar, resembling, richard, transparent, turtle, compared, as, salt, slot, liev, cchio, iva, religious, useful, smooth, unlike, juicy |
| **Metaphor** | eternity, disasters, prices, form, consumed, accusations, sings, gasoline, blossoms, summoned, drowned, trial, hunts, arguments, objections, tread, oath, clashed, communicated, fell | consumed, nos, time, rish, eled, which, imated, rah, given, jewel, crowned, theory, asionally, still, light, ving, charles, ift, these, looking | drizzle, tramp, rested, fect, scan, ravaged, eld, switched, shuddered, shiver, sighed, transported, rooms, spoke, reserve, proceeded, slight, dri, pivot, enne |

Table 6: Top most important features of each figurative language class extracted from the black-box models. The features in each cell are sorted in descending order by their SHAP values.

| Class | Most Important Features | | |
|---|---|---|---|
| | LR | RF | Humans |
| **Sarcasm** | that, when, am, like, how, and, love, out, her, for, great, proud, saw, got, thrilled, all, so, friend, just, beautiful | when, like, how, am, that, love, her, got, proud, great, for, thrilled, out, stone, fact, last, car, saw, feel, from | overjoyed, grateful, pleasant, great, praised, roses, crashed, excited, vomiting, myself, not, no, hero, terrible, proud, drops, happy, couldnt, deal, mistake |
| **Idiom** | like, my, really, me, you, your, and, go, an, after, let, been, excited, for, is, they, people, cut, under, back | my, like, me, and, go, really, as, you, your, it, people, not, an, for, with, made, the, ll, because, he | sink, leak, full, duck, cut, nth, swim, beans, broad, beam, lame, bag, baggage, dried, degree, hand, bear, flow, smell, of |
| **Simile** | me, my, really, as, time, makes, people, on, one, hills, good, eyes, husband, by, work, re, person, this, wanted, you | my, really, me, on, makes, person, to, time, eyes, world, skin, son, made, re, have, doesn, kids, husband, running, people | as, smooth, tough, dummies, crocodile, snowman, cancer, glass, affectionate, oak, an, angel, dream, gift, like, planted, deflated, day, fantastical, washboard |
| **Metaphor** | like, really, me, my, the, time, on, makes, an, clothes, into, made, but, good, where, is, by, without, of, people | like, really, my, me, the, was, makes, to, his, in, an, she, have, of, time, no, good, into, this, who | toppled, rose, fell, shoot, burning, gravely, never, cure, desire, fertilize, the, flicked, darkness, spirits, prescribes, behind, ravaged, dwell, leak, speed |

Table 7: Top most important features of each figurative language class extracted from white-box models and human annotations. The features in each cell are sorted in descending order by their SHAP values.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section "Limitations"*

☒ A2. Did you discuss any potential risks of your work?
*The case study (feature analyses on different models' decisions on figurative language processing) is not subject to potentially malicious or unintended harmful effects and uses, including but not limited to bias confirmation, harm of privacy, or adversarial attacks. The dataset used does not contain any sensitive or confidential information. The human annotation experiment was consented to by all participants, and no personal information of the participants can be inferred from the collected data. As the dataset size is relatively small, there is no significant environmental impact caused by model training.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract; Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*1*

☑ B1. Did you cite the creators of artifacts you used?
*3.1; 4.1; 5; Appendix A*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*4.1*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*3.1; 4.1; 5;*

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Section "Ethics Statement"*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*3.1; 4.1; 5.2*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*3.1; 4.1; 5.1*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

**C ☑ Did you run computational experiments?**

*4*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Appendix A*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*3; 4; Appendix A*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*3.4, 5.2; Appendix C*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Appendix A*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*5.1*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*1*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*5.1*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*5.2; Section "Ethics Statement"*

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Section "Ethics Statement"*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*5.1*