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., Regal 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
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.\footnote{All code, models, and experimental instructions are available at: https://github.com/CoPsyN/figurative-language-processing}

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 \cite{Oraby2016}, the VU Amsterdam Dataset of Metaphor \cite{Steen2010}, and the MAGPIE dataset for potentially idiomatic expressions \cite{Haagsma2020}, 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 expression have often been used for sarcasm detection \cite{Joshi2015, Riloff2013}; while abstractness of words \cite{Kopper2017, Turney2011} and topic transition information \cite{Jang2016} have been used for metaphor detection. Recent studies have been using neural models \cite{Gao2018, Wu2018, DoDinh2016}, and TLMs especially have shown good performance \cite{Chakrabarty2022a, Chakrabarty2022b, Jang2016, Avvaru2020, Dong2020, Liu2020}. Some recent work using TLMs treats figurative language processing as a natural language inference (NLI) task instead of a classification task \cite{He2022, Chakrabarty2021}, 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 \cite{Chen2022, Dankers2022, Ehren2022, Tan2021}. 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 \cite{Chakrabarty2022}, an English-language dataset released for the Shared Task on Understanding Figurative Language 2022 \cite{Chakrabarty2022b}. 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 \cite{Chakrabarty2022b}, 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)).

\begin{itemize}
  \item \textbf{(1)} Grad school was so comforting that I had no choice but to drop out to keep my sanity.
  \item \textbf{(2)}
\end{itemize}
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.

Rule of thumb is escape while you’re on the move.

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.

<table>
<thead>
<tr>
<th></th>
<th>With Apparent Cues</th>
<th>No Apparent cues</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Sentences</td>
<td>2212</td>
<td>625</td>
</tr>
<tr>
<td>#Tokens</td>
<td>45233</td>
<td>9062</td>
</tr>
</tbody>
</table>

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
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).

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLACK-BOX MODELS</strong></td>
<td></td>
</tr>
<tr>
<td>bert-base-uncased</td>
<td>0.95</td>
</tr>
<tr>
<td>roberta-base</td>
<td>0.95</td>
</tr>
<tr>
<td>xlnet-base-cased</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>WHITE-BOX MODELS</strong></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (LR)</td>
<td>0.87</td>
</tr>
<tr>
<td>Decision Tree (DT)</td>
<td>0.77</td>
</tr>
<tr>
<td>Random Forest (RF)</td>
<td>0.85</td>
</tr>
<tr>
<td>Naive Bayes (NB)</td>
<td>0.68</td>
</tr>
</tbody>
</table>

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

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**: adjectives, adverbs, comparisons, verbs.

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 a relatively higher importance and 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.

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.
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 con-
stituents” (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 Forms4 and Prolific5. 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 fell behind 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

4https://forms.google.com
5https://www.prolific.co
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


Xiangjue Dong, Changmao Li, and Jinho D. Choi. 2020. Transformer-based context-aware sarcasm detection


on Figurative Language Processing, pages 250–255, Online. Association for Computational Linguistics.


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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td><code>solver = 'sag', multi_class='multinomial'</code></td>
</tr>
<tr>
<td>RF</td>
<td><code>n_estimators = 100</code></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Figurative Language</th>
<th>BERT</th>
<th>RoBERTa</th>
<th>XLNet</th>
<th>LR</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarcasm</td>
<td>P 0.99</td>
<td>R 0.99</td>
<td>F1 0.99</td>
<td>P 0.99</td>
<td>R 0.99</td>
</tr>
<tr>
<td>Simile</td>
<td>P 0.94</td>
<td>R 0.96</td>
<td>F1 0.94</td>
<td>P 1.00</td>
<td>R 0.97</td>
</tr>
<tr>
<td>Idiom</td>
<td>P 0.93</td>
<td>R 0.93</td>
<td>F1 0.93</td>
<td>P 0.94</td>
<td>R 0.93</td>
</tr>
<tr>
<td>Metaphor</td>
<td>P 0.94</td>
<td>R 0.93</td>
<td>F1 0.91</td>
<td>P 0.94</td>
<td>R 0.93</td>
</tr>
</tbody>
</table>

Figure 4: Confusion matrices of the selected models (x-axis: predicted labels; y-axis: true labels).
<table>
<thead>
<tr>
<th>Class</th>
<th>BERT</th>
<th>RoBERTa</th>
<th>XLNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarcasm</td>
<td>refreshing, thankful, proud, praised, thrilled, glad, awesome, delighted, excited, incredible, terrific, adorable, amazed, wonderful, fascinated, amazing, delightful, happily, planet, fantastic</td>
<td>increase, donated, videos, saving, boost, ankle, healing, elt, cried, attend, personally, satisfaction, celebrities, organization, somehow, shaken, frustrating, civic, cute, cooking</td>
<td>safest, refreshing, annoyed, scary, love, irritating, commend, cheered, hottest, ensured, celebrating, enjoying, pleasing, encourages, afterwards, approve, cheering, adore, makes, because</td>
</tr>
<tr>
<td>Idiom</td>
<td>differ, aye, halves, sevens, shove, eddie, nods, ways, guess, matthias, platt, hilt, plank, james, ava, plus, meantime, overboard, or, daylight</td>
<td>moist, damned, theory, arnold, eddie, until, toppled, caps, english, devils, els, nicely, production, hell, playing, words, fucking, jon, palace, bone</td>
<td>sticks, wire, messenger, beans, halves, lend, sides, hatch, platter, record, mark, splash, naked, broth, plus, hook, wolf, thieves, trades, source</td>
</tr>
<tr>
<td>Simile</td>
<td>resemble, resembled, like, arnold, predatory, titanium, resembling, slug, transparent, compared, twilight, charcoal, resemblance, similar, alligator, fragile, magazine, turtle, calculus, locomotive</td>
<td>mor, herd, slightest, movement, indicating, spicy, had, shield, disappears, understands, lining, messy, ays, colleague, resemble, balanced, towers, noble, descended, nationality</td>
<td>like, resembled, resemble, similar, resembling, richard, transparent, turtle, compared, as, salt, slot, liev, chchio, iva, religious, useful, smooth, unlike, juicy</td>
</tr>
<tr>
<td>Metaphor</td>
<td>eternity, disasters, prices, form, consumed, accusations, sings, gasoline, blossoms, summoned, drowned, trial, hunts, arguments, objections, tread, oath, clashed, communicated, fell</td>
<td>consumed, nos, time, rish, eled, which, imated, rah, given, jewel, crowned, theory, asionally, still, light, ving, charles, ift, these, looking</td>
<td>drizzle, tramp, rested, fect, scan, ravaged, eld, switched, shuddered, shiver, sighed, transported, rooms, spoke, reserve, proceeded, slight, dri, pivot, enne</td>
</tr>
</tbody>
</table>

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

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.
A For every submission:

- ✔ A1. Did you describe the limitations of your work?
  
  "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?
  
  "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?

ulado text with the document. It appears that the text is related to evaluating the implementation and reporting of computational experiments and human annotator research in a research paper. The text includes sections on reporting parameters, computational budget, hyperparameter values, descriptive statistics, and the use of existing packages. Additionally, it covers the use of human annotators and the ethical considerations involved in such research.

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

ulado text with the document. It appears that the text is related to evaluating the implementation and reporting of computational experiments and human annotator research in a research paper. The text includes sections on reporting parameters, computational budget, hyperparameter values, descriptive statistics, and the use of existing packages. Additionally, it covers the use of human annotators and the ethical considerations involved in such research.