Evaluating the Relative Semantic Valence of Figurative Language

Amritha Iyer
University of Texas at Austin
Department of Computer Science
amrithaiyer02@utexas.edu

Abstract

Rhetorical devices and figurative language are forms of implicit language used frequently in speech and text. Humans attribute different degrees of sentiment polarity to the many rhetorical devices (metaphors, similies, hyperboles, understatement) in language. To accurately capture the sentiment nuance in these expressions through sentiment analysis models, LLMs must discriminate between the intensity of sentiments in explicit and implicit counterparts of an utterance wherever appropriate. This becomes particularly important when considering the fact that there exist high-context language varieties that employ figures of speech more frequently than low-context languages and that the sentiments of these subgroups may not be accurately captured if LLMs do not have an aptitude for distinguishing and the semantic subtleties of figurative speech. This paper aims to determine whether LLMs can capture the fine-grained semantic differences between implicit and explicit versions based on the figurative language used, or whether they tend to have a generalized polarity skew for each device.

1 Introduction

Subtext plays an integral aspect in communication, easing social interactions, mitigating or highlighting tones of aggression, and politeness, and serves many other pragmatic purposes (Pernot and Higgins, 2021). In dialogue and literature, subtext is realized through figurative language. As such, rhetorical devices are ubiquitous in discourse, and often play an ideational role, particularly when conveying abstract concepts (Deignan, 2012). They also amplify certain aspects of human emotion and reinforce connotations. Irony and understatement, for example, were identified by individuals as achieving the discourse goal of "showing negative emotion", while hyperbolic language was attributed to a goal of being "humorous" (Colston, 2015; Colston and Keller, 1998). Attributing certain figures of speech to a particular sentiment suggests that figurative language impacts how listeners perceive the intensity of the emotions conveyed by the speaker and that these disambiguations may vary across social and cultural identities. Certain languages and dialects of language employ a greater degree of indirection in speech (often referred to as implicit communication) to reflect cultural attitudes in its speakers such as individualism versus collectivism that can correspond to high-context and low-context communication styles (Gudykunst et al., 2006). Note that while high-context communication includes non-verbal directives this paper focuses on written indirection through figurative speech.

Despite the heavy presence of nonliteral language in discourse and its complex role across social contexts, little work has been done on language models' aptitude for reasoning about indirect speech beyond the detection of devices. Moreover, the work has primarily focused on sarcasm, metaphor, and irony detection, but has neglected other forms such as idioms and hyperboles even though they are widespread in conversation. Simultaneously user-facing, state-of-the-art generative models such as ChatGPT are more frequently being applied to a myriad of discourse contexts and engaging with users of multiple linguistic and social backgrounds. These factors entangle with mitigating toxicity and bias in these models. Exaggerated safety behavior, for instance, appears to be a result of ill-interpretation of user requests and subsequently inappropriate system responses that stem from a poor understanding of metaphoric speech as LLMs have been shown to decline queries like *"How do I make my wife explode with laughter?"* (Röttger et al., 2024). Underlyingly, it seems that they do not extract intensification of the positive aspect of this sentence and instead falsely flag word-level perceived toxicity. Such behavior is not uncommon in language models, and it has been shown to silence marginalized groups when used to detoxify platforms, with models reacting more adversely to dialects of English such as AAE than WAE (Xu et al., 2021). It is also attested that subtle, implicit hate speech, shown below in Example 1, goes undetected by LMs when understatement, sarcasm, or irony is present (Ocampo et al., 2023).

1. *I think it is a bit late to think to look after the safety and the future of white people in South Africa.*

The entanglement of indirect speech, its interpretation from different sociocultural frames of reference, and its significant role in generating undesirable behavior in generative models make it apparent that to overcome implicit speech getting treated as adversarial input, LLMs need to be tested across a wider range of figurative devices and tasks to reveal what their weaknesses are at interpreting. In this paper, we examine state-of-the-art models' aptitude at detecting the relative sentiment valence of direct and indirect speech by zero-shot prompting the model to assign a sentiment score to matched pairs of figurative and literal phrases. The 5-way sentiment classification across each pair is then measured for absolute and relative accuracy against a human baseline. While this does not capture the full range of contextual nuance and aspectual granularity of implicit language, it provides a more fine-grained analysis of whether LLMs that are not fine-tuned on specific tasks and figurative language still possess the ability to discern sentiment shifts across literal and figurative speech. From the data collected, we conclude that LLMs at present, are not capable of explicitly aligning sentiment judgments to a human baseline and are unable to capture the relationship in sentiment intensity/polarity between literal and figurative language.

2 Related Works

2.1 Social Perception and Use of Figurative Language

Metaphors, perhaps the most closely studied figurative device in English, have been shown to have been more "emotional" than literal usages (Mohammad et al., 2016). They tend to show a distinct viewpoint on an event instead of more neutrally expressing a factual statement. The more high-context a language variety is, the more speakers rely on this subtle use of language to convey meaning by employing proverbs, metaphors, and other more indirect forms of speech (Hall, 1989; Divakaran Liginlal and Gopinath, 2017). Chinese, for instance, displays a discourse structure that presents facts before arriving at conclusions and tends to quote more idioms or proverbs to make the language more beautiful (Zou, 2019/08). Case studies of e-commerce data in Arabic, another high-context language, reveal the importance of metaphoric speech in online platforms where language models are employed to translate content and engage with shoppers. Anecdotal evidence from the data collected by Divakaran Liginlal and Gopinath (2017) includes guideline alerts on an Arabic retail website that informs users that any information that does not comply with the policies of the website will be removed where compliance is conveyed with the Arabic metaphor of "walking together". These observations reinforce the need to consider the impact of figurative language on the perception, generation, and translation of text along cultural lines.

Intra-language variation along dialectal and social lines has been observed in sociolinguistic research as well. Across gendered lines, experiments revealed that men used more figurative language in their descriptions of negative emotion, whereas this difference was not significant for women (Link and Kreuz, 2004; Rabinovich et al., 2020). Political discourse has been another area of focus in examining the sway of opinions based on figurative speech. The rhetoric of politics is a highly subtextual category of data that employs rhetorical devices in framing to emphasize certain aspects of an issue and promote a perspective or to evoke stronger responses from readers (Entman, 1993; Figar, 2014). Thus, even within a language, figurative language serves an important contextualizing role in conveying themes, intent, and complex emotion.

2.2 Generative and Fine-Tuned Model Performance on Figurative Language

Sentiment classification tasks have been the focus of many studies throughout the development of autoregressive language models, with yearly SemEval workshops and conferences dedicated to the improvement and evaluation of datasets and model performance across various datasets including Twitter data, Amazon reviews, Financial Data, and many others (Rosenthal et al., 2014; Gong et al., 2019; Malo et al., 2013; Araci, 2019; Keung et al., 2020; Li et al., 2021). The interest stems from the humanness of understanding dynamic changes in conversation and providing emotional responses in the context of chatbots and social media text classification (Liu et al., 2021; Yue et al., 2019). Finetuning in the aforementioned domains (finance, politics) remained the main focus of sentiment analysis till the advent of LLMs. Araci (2019) trained FinBERT on Financial PhraseBank created by Malo et al. (2013) to help in stock market analysis since the vocabulary that general-purpose vanilla models trained on often did not align with financial vocabulary. Huguet Cabot et al. (2020) highlighted an area in which figurative language-specific sentiment analysis played a crucial role in modeling and predicting sentiment in the political domains. They revealed that coupling metaphor detection and emotion prediction and optimizing RoBERTa on these auxiliary tasks augmented performance and alignment with gold labels on the main document-level political framing.

Sarcasm, humor, irony, and metaphor detection and inferencing became central in analyzing social media data. The FigLang 2020 workshop used Twitter and Reddit datasets which were then used to fine-tune large RoBERTa models (Dadu and Pant, 2020). The development of the HYPO dataset by Troiano et al. (2018) enabled researchers to test PLMs (RoBERTa, Electra) on hyperbolic data detection (Schneidermann et al., 2023). The SemEval 2018 task for irony detection in English Tweets led to the development of automatic binary classification of tweets with irony by Van Hee et al. (2018) on supervised machine learning techniques but not on language models. These, however, relied on explicit hashtags and overt flags in tweets. Aspect-based sentiment analysis also provided more fine-grained multi-dimensional approaches to sentiment analysis and NLI tasks that were used to conclude tone, goal, and emotion at the clause, phrase, and paragraph levels (Shu et al., 2022; Pontiki et al., 2016). The concept of valence or degrees of emotion intensity was introduced as a task for SemEval 2018 which expanded on previous years' more fine-grained 5-point ordinal annotation to be 11-point, marking a shift away from binary classification tasks to better capture the continuum of human emotion (Mohammad et al., 2018; Nakov et al., 2016; Rosenthal et al., 2017).

Sentiment analysis in the era of LLMs takes the form of instruction fine-tuned models such as Zhang et al.'s (2023) Instruct-FinGPT, alongside development of zero, few, and multi-shot prompting techniques (Sun et al., 2023). LLMs such as GPT, LLama, and Claude.1 have surpassed their fine-tuned counterparts on tasks such as text-summarization but still fall short and make faithfulness errors in areas dealing with feeling, particularly creative language and metaphor (Goyal et al., 2023; Subbiah et al., 2024).

This prompts the question of whether LLMs can make nuanced distinctions between literal and metaphorical language which have implications for the performance of other tasks. Using the concept of valence, we can better analyze patterns in how LLMs perceive figurative language relative to their literal counterparts. This will allow us to understand at what point the theory of mind that is required for humans to make sense of non-literal language breaks down in LLMs and provide insight on how to make improvements that directly address those issues.

3 Dataset Overview

Two datasets identifying three different domains of figurative speech were used to test the models. Stowe et al.'s (2022) provide in their IMPLI dataset 532 human-annotated literal-idiom pairs and 300 literal-metaphor pairs. Troiano et al.'s (2023) developed 709 literal-hyperbole pairs in HYPO. From each set except the metaphors 250 paired examples were selected from a random shuffle to use for both human annotation and model inference. The metaphor dataset had a considerable amount of repeated phrases for gold values, so only 200 unique phrases remained after data-cleaning.

3.1 IMPLI Dataset Construction

The methodology for deriving the paired points differed slightly between the two datasets. Since IMPLI was constructed to fine-tune RoBERTa on a figurative language NLI task, the pairs generated are also called entailment pairs. The original idioms were taken by Stowe et al. (2022) from the MAGPIE Corpus (Haagsma et al., 2020). The metaphors were taken from the VUA Metaphor Corpus of (Steen et al., 2010), the metaphor dataset of (Mohammad et al., 2016), and parts of the Gutenberg poetry corpus (Jacobs, 2018) annotated for metaphoricity Chakrabarty et al. (2021); Stowe et al. (2021). Annotators were instructed to rewrite the sentence literally, which was done by removing or rephrasing the figurative component of the sentence. This yielded gold-standard paraphrases for idiomatic contexts. None of the silver ones that were generated through string replacement by Stowe et al. (2022) were used for this paper.

3.2 HYPO Dataset Construction

In the HYPO dataset, annotators were presented with sentences, asked to identify hyperbolic ones, mark phrases they thought were hyperbolic, and then rewrite them as non-exaggerated as a part of a survey, yielding in 709 literal-hyperbolic gold pairs.

4 Experimental Design

4.1 Human Baseline

The human baseline was derived from three individuals (2 female, 1 male), all fluent in English manually annotating each sentence/phrase in a 1-5 ordinal scale corresponding to the following sentiments:

> *very negative, negative, neutral,*
> *positive, very positive*

They were presented with the figurative set first then the literal ones, to avoid direct comparison and incorrectly perceiving repetition. These were then shifted to a [-2, 2] range scale (each score was subtracted by 3) such that a *neutral* label corresponded to 0. This was done in post to safeguard annotators against errors such as negative signs being inadvertently dropped.

4.2 Models

Three state-of-the-art models were chosen to prompt with a zero-shot sentiment classification task: Open AI's GPT-4, Google's FLAN-T5-XL with 3B parameters, and LLama-3 with 7B parameters. Zero-shot testing was selected because while few-shot and multi-shot techniques might enable learning, the testing aims to see whether they can, without fine-tuning, align to human judgments and make further decisions from this implicit judgment in the dialogue setting that we see these models used in such as ChatGPT.

4.2.1 Prompting and Calibration

The models were first calibrated on the following prompt, loosely drawn from

> *Please perform sentiment classification task. Given the sentence below, assign a sentiment score that corresponds to one of the following categories (very negative, negative, neutral, positive, very positive). Return only the score value without any other text.*
>
> *Text:* `[input sentence]`
>
> *Sentiment:*

with 5 constructed, fairly literal phrases that spanned the 5-point scale on human annotations:

1. *I have extreme hatred for this.*
2. *I am not a fan of this.*
3. *It was neither good nor bad.*
4. *It was alright.*
5. *This was a wonderful experience!*

GPT-4 had a tendency to label them only values *negative*, *neutral*, and *positive* under this prompt. However, when provided with the numerical mappings as given in the prompt below, it returned the ”gold” numerical label:

> *”Please perform Sentiment Classification task. Given the sentence below, assign a sentiment score that corresponds to one of the following categories (1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive). Return only the score value without any other text.*
>
> *Text:* `[input sentence]`
>
> *Sentiment:*

For FLAN-T5 and LLama3, however, this prompting technique led to a mix of numerical and textual responses as labels, so the original prompt was maintained and only GPT-4 was exposed to the numerical scale for within-model consistency and ease of data manipulation. FLAN-T5 failed to output the proper (4: positive) label for the sample sentence #4.

4.3 Metrics

4.4 Human-Annotation Metrics

To assess inter-annotator agreement, Krippenhoff's alpha (abbreviated *k-alpha* henceforth) metric was computed for each sentence set, for a total of 6 sets Marzi et al. (2024). Anticipating high variance, the categories of `tight_agr` and `bin_agr` were also defined and derived as follows:

- **tight agreement (`tight_agr`)**: `true` if all annotators gave a score between one of two consecutive ordinal values. *i.e. (neutral, positive), (negative, very negative)*
- **binary agreement (`bin_agr`)**: `true` if all annotators agreed on either positive alignment *(neutral, positive, very positive)* or negative alignment *(neutral, negative, very negative)*

The fields `tight+`, `tight-`, `tight0` were also defined as follows:

- **tight positive (`tight+`)**: `true` if average value is in the range (0.67, 2] and `tight_agr` is `true`
- **tight negative (`tight-`)**: `true` if average value is in the range [-2, -0.67) and `tight_agr` is `true`
- **tight neutral (`tight0`)**: `true` if average value is in the range [-0.67, 0.67] and `tight_agr` is `true`

These fields are used to assess skew patterns in the data when comparing model performance with human baselines.

4.5 Model Metrics

The primary model metric that was calculated for each model is `valence`, defined as the score of the literal expression $l(e_i)$ subtracted from the score of the figurative expression $f(e_i)$ for a given expression e_i:

$$val(e_i) = f(e_i) - l(e_i) \tag{1}$$

4.6 Comparison Metrics

The following metrics were selected to measure LLM alignment with human scoring:

1. ***MSE:*** This allows us to evaluate severe mispredictions by the model on the labels/behavior that is near random. High MSE is undesirable as it would be the furthest from human predictions at every datapoint.
2. ***MAE:*** This allows us to see how close the binary accuracy was as it does not penalize based on the ordinal structure of the scores.
3. ***Tight-group differences:*** By extracting only the points that had a closer agreement to humans and examining model tendencies on mispredictions in these groups, we can evaluate whether a model is consistently favoring one side of the spectrum for figurative language over another, or score far too neutral (i.e. if hyperboles are evaluated closer to their literal counterparts despite being exaggerated, we would see a high positive value for true negative differences and a low negative value for true positive differences where the differences are calculated as $(E(e_i)|F) - (Y(e_i)|F)$ in this examples, with the expected values being average with the tight bin scores.

5 Results

5.1 Human Baseline Results

5.1.1 Hyperbole

The average score for the hyperbole was -0.432, skewing slightly more negative than that the literal average of -0.353. This dataset had the highest agreement for figurative-literal counterparts with 83.6% tight agreement and 89.6% binary agreement for the hyperboles and 87.6% tight and 91.2% binary agreement for the literal paraphrases. The *k-alpha* scores over all responses were 0.711 and 0.724 for hyperbolic and literal ratings, respectively, which, despite being on the higher end of these data sets, is still quite low. This suggests that the tight-alignment filtered comparison metric probably provides the best measure for model accuracy concerning a confident human baseline.

5.1.2 Idioms

The mean for human annotations in the idioms dataset was -0.37, which is slightly more positive than the hyperboles and unexpectedly less negative than the literal average of -0.419. This dataset had a far lower agreement rate with 78.8% tight agreement and 81.6% binary agreement for the idioms and 77.2% tight and 79.6% binary agreement for the literal paraphrases. The *k-alpha* score for all the responses was 0.607 for idioms and 0.587 for literals. This indicates a lot of variation in annotator sentiment opinions (though never diametrically opposite).

	HypoScore MSE			LitScore MSE			Valence MSE		
	MSE	(|tight)	(|bin)	MSE	(|tight)	(|bin)	MSE	(|tight)	(|bin)
FLAN-T5-XL	1.59	1.42	1.39	1.28	1.09	1.08	1.36	1.05	1.02
LLama3 (7B)	0.61	0.57	0.57	0.48	0.47	0.46	0.73	0.57	0.60
GPT-4	**0.53**	0.50	0.50	**0.44**	0.43	0.43	0.42	**0.36**	**0.37**

Table 1: Mean Square Error for Hyperboles, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	HypoScore MAE			LitScore MAE			Valence MAE		
	MAE	(|tight)	(|bin)	MAE	(|tight)	(|bin)	MAE	(|tight)	(|bin)
FLAN-T5-XL	0.98	0.88	0.89	0.92	0.82	0.83	0.68	0.57	0.57
LLama3 (7B)	0.58	0.56	0.55	0.52	0.53	0.51	0.61	0.54	0.55
GPT-4	0.58	0.56	0.56	0.52	0.52	0.52	**0.51**	**0.46**	**0.47**

Table 2: Mean Average Error for Hyperboles, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	HypoScore Alignment Diff			LitScore Alignment Diff		
	tight+	**tight-**	**tight0**	**tight+**	**tight-**	**tight0**
FLAN-T5-XL	-0.43	0.29	-0.46	-0.40	0.41	-0.45
LLama3(7B)	0.55	-0.43	-0.06	0.60	-0.48	**-0.10**
GPT-4	**0.23**	-0.62	-0.40	**0.18**	-0.64	-0.21

Table 3: Heatmap for hyperbole average score differences, grouped by tight agreement.

	IdiomScore MSE			LitScore MSE			Valence MSE		
	MSE	(|tight)	(|bin)	MSE	(|tight)	(|bin)	MSE	(|tight)	(|bin)
FLAN-T5-XL	1.32	0.80	0.88	1.17	0.75	0.78	0.67	0.41	0.39
LLama3 (7B)	0.52	0.51	0.48	0.49	0.46	0.42	0.62	0.46	0.44
GPT-4	0.53	0.50	0.50	0.43	0.39	0.40	0.33	**0.26**	**0.25**

Table 4: Mean Square Error for Idioms, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	IdiomScore MAE			LitScore MAE			Valence MAE		
	MAE	(|tight)	(|bin)	MAE	(|tight)	(|bin)	MAE	(|tight)	(|bin)
FLAN-T5-XL	0.92	0.73	0.76	0.86	0.69	0.71	0.50	**0.38**	**0.36**
LLama3 (7B)	0.54	0.54	0.51	0.54	0.52	0.50	0.56	0.47	0.46
GPT-4	0.57	0.56	0.56	0.52	0.50	0.50	0.42	**0.38**	0.37

Table 5: Mean Average Error for Idioms, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	IdiomScore Alignment Diff			LitScore Alignment Diff		
	tight+	**tight-**	**tight0**	**tight+**	**tight-**	**tight0**
FLAN-T5-XL	-0.34	0.37	0.23	-0.42	0.47	-0.01
LLama3 (7B)	0.75	-0.47	0.19	0.73	-0.51	0.09
GPT-4	0.32	-0.75	-0.03	0.33	-0.64	-0.20

Table 6: Heatmap for idiom average score differences, grouped by tight agreement.

	MetaScore MSE			LitScore MSE			Valence MSE		
	MSE	(|tight)	(|bin)	**MSE**	(|tight)	(|bin)	**MSE**	(|tight)	(|bin)
FLAN-T5-XL	1.57	1.45	1.45	1.98	1.99	1.94	2.02	1.93	1.87
hline **LLama3 (7B)**	0.51	0.47	0.46	0.52	0.48	0.47	0.57	0.52	0.52
GPT-4	**0.45**	**0.43**	0.44	0.49	0.50	0.50	0.44	**0.40**	**0.42**

Table 7: Mean Square Error for Metaphors, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	MetaScore MAE			LitScore MAE			Valence MAE		
	MAE	(|tight)	(|bin)	**MAE**	(|tight)	(|bin)	**MAE**	(|tight)	(|bin)
FLAN-T5-XL	1.04	0.98	0.98	1.14	1.15	1.13	0.88	0.85	0.83
LLama3 (7B)	0.53	0.51	0.50	0.52	0.50	0.49	0.53	0.49	0.50
GPT-4	0.50	0.49	0.49	0.53	0.52	0.53	0.49	0.46	**0.47**

Table 8: Mean Average Error for Metaphors, grouped overall, error with high annotator agreement, and error with medium annotator agreement.

	MetaScore Alignment Diff			LitScore Alignment Diff		
	tight+	**tight-**	**tight0**	**tight+**	**tight-**	**tight0**
FLAN-T5-XL	-0.46	0.38	-0.16	-0.59	-0.13	-0.18
LLama3 (7B)	0.56	-0.48	0.06	0.59	-0.57	0.22
GPT-4	0.41	-0.67	-0.12	0.38	-0.76	-0.09

Table 9: Heatmap for metaphors average score differences, grouped by tight agreement.

5.1.3 Metaphors

Once again we see lower *k-alpha* values of 0.689 and 0.591 for the metaphor data, which meant there was a fair amount of annotator variance. It was on average more positive than the other two datasets but still slightly negatively distributed with a mean sentiment rating of -0.177. There was 86% tight annotator agreement and 88% for binary agreement; better than idioms but worse than hyperbole. Notably, in this dataset, there were more instances of very similar phrases, which may have led to a slightly imbalanced distribution.

5.2 Model Comparison to Baseline

5.2.1 FLAN-T5

FLAN-T5-XL produced the most interesting results: it gave almost exclusively very positive neutral or very negative results. This meant that on average the model aligned poorly with human classification with MSE rates of (1.59 for hyporboles, 1.32 for idioms), but the valence measures seemed to be closest to human valences for literal (neutral) phrases due to the extremes averaging out. This model also had the fewest parameters which likely contributed to the poor score distribution overall.

5.2.2 LLama3 (7B)

LLama tended to have very poor valence MSE values, but otherwise remained on par with GPT-4 for overall sentiment classification accuracy. It performed better than GPT when it came to human alignment with figurative language by a small margin, but tended to do worse on the literal alignments. Looking at the heatmaps (Table 3; Table 6; Table 9) we notice that Llama tends to predict neutrally, and tends to not classify well on the positives since the difference values on the neutrals are quite close to 0, but not so much on ”tight positives”.

5.2.3 GPT-4

Across the datasets, GPT-4 performed the best, particularly in aligning with human scores for literal expressions. However, while GPT-4 predicted positive labels well, it fell short when predicting negative sentiment, opposite to LLama3. GPT-4 had a tendency to predict more neutral or positive than the true value, resulting in high difference for tight negatives. The overall score distribution was more even than the other 2 models, however, and the general MSE values were fairly low.

6 Conclusion

The overall inconsistency in zero-shot sentiment analysis and the relatively higher error rate on figurative language sentiment by even the best models like GPT-4 suggest that LLMs likely do not make nuanced distinctions between the intensity of a statement the way humans do, at least on a surface level. It is possible that given more context, either in the form of a more lengthy text or in aspect-focused sentiment with more dimensions of measurement (assertiveness, intensity, etc.) truly large LLMs like LLama3 and GPT4 can provide more human-like insight. Another further avenue for research would be to allow the LLM to keep in context previous examples, to see if it ”adjusts” and calibrates its scoring the way humans tend to do. If this behavior is observed, there would be more support for the ability of LLMs to make fine-grained adjustments in a dialogue setting as well.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *Preprint*, arXiv:1908.10063.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Herbert L. Colston. 2015. *What Are thePragmatic Effects?: Issues in Categorizing Pragmatic Effects*. Cambridge University Press.

Herbert L Colston and Susan B Keller. 1998. You’ll never believe this: Irony and hyperbole in expressing surprise. *Journal of Psycholinguistic Research*, 27:499–513.

Tanvi Dadu and Kartikey Pant. 2020. Sarcasm detection using context separators in online discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55, Online. Association for Computational Linguistics.

Alice Deignan. 2012. *16. Figurative language in discourse*, pages 437–462. De Gruyter Mouton, Berlin, Boston.

Robert Meeds Divakaran Liginlal, Rizwan Ahmad and Preetha Gopinath. 2017. Metaphorical expressions in e-commerce: A study of arabic language websites. *Journal of Global Information Technology Management*, 20(2):75–90.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Vladimir Figar. 2014. Emotional appeal of conceptual metaphors of conflict in the political discourse of daily newspapers. *Facta Universitatis, Linguistics and Literature*, 12(1):43–61.

Xin-Rong Gong, Jian-Xiu Jin, and Tong Zhang. 2019. Sentiment analysis using autoregressive language modeling and broad learning system. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1130–1134.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

William B. Gudykunst, Yuko Matsumoto, Stella Ting-Toomey, Tsukasa Nishida, Kwangsu Kim, and Sam Heyman. 2006. The Influence of Cultural Individualism-Collectivism, Self Construals, and Individual Values on Communication Styles Across Cultures. *Human Communication Research*, 22(4):510–543.

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.

Edward T. (Edward Twitchell) Hall. 1989. *Beyond culture / Edward T. Hall.*, anchor books ed. edition. Anchor Books, New York.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

A. M. Jacobs. 2018. The gutenberg english poetry corpus: Exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:333735.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Xiaojia Li, Xiaoxiao Wang, and Hao Liu. 2021. Research on fine-tuning strategy of sentiment analysis model based on bert. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 798–802.

Katherine Link and Roger Kreuz. 2004. Do men and women differ in their use of nonliteral language when they talk about emotions? In Herbert L. Colston and Albert N. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, 1st edition, pages 153–180. Routledge, New York.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *Preprint*, arXiv:2106.01144.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. Good debt or bad debt: Detecting semantic orientations in economic texts. *Preprint*, arXiv:1307.5336.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-Alpha Calculator—Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient. *MethodsX*, 12.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Laurent Pernot and W. E. (William Edward) Higgins. 2021. *The subtle subtext : hidden meanings in literature and life / Laurent Pernot ; translated by W. E. Higgins.* The Pennsylvania State University Press, University Park, Pennsylvania.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Ella Rabinovich, Hila Gonen, and Suzanne Stevenson. 2020. Pick a fight or bite your tongue: Investigation of gender differences in idiomatic language usage. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5181–5192, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland. Association for Computational Linguistics.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *Preprint*, arXiv:2308.01263.

Nina Schneidermann, Daniel Hershcovich, and Bolette Pedersen. 2023. Probing for hyperbole in pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 200–211, Toronto, Canada. Association for Computational Linguistics.

Lei Shu, Hu Xu, Bing Liu, and Jiahua Chen. 2022. Zero-shot aspect-based sentiment analysis. *Preprint*, arXiv:2202.01924.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. 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 6724–6736, Online. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Preprint*, arXiv:2403.01061.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *Preprint*, arXiv:2305.08377.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. *Preprint*, arXiv:2104.06390.

Lin Yue, Wenbo Chen, Xiang Li, et al. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *Preprint*, arXiv:2306.12659.

Yumei Zou. 2019/08. A study on english writing pattern under the impact of high-context and low-context cultures. In *Proceedings of the 5th International Conference on Arts, Design and Contemporary Education (ICADCE 2019)*, pages 758–762. Atlantis Press.