NewsMet: A 'Do It All' dataset of *contemporary* Metaphors in News headlines

Rohan Joseph[1], Timothy Liu[2], Aik Beng Ng[2], Simon See[1,2], and Sunny Rai[1,3]

[1]École Centrale School of Engineering and Sciences, Mahindra University.
[2]NVIDIA AI Technology Center.
[3]Department of Computer and Information Science, University of Pennsylvania.
rohan18545@mechyd.ac.in, {aikbengn, timothyl, ssee}@nvidia.com,
sunny.rai@seas.upenn.edu

Abstract

Metaphors are highly creative constructs of human language that grow old and eventually die. Popular datasets used for metaphor processing tasks were constructed from dated source texts. In this paper, we propose NewsMet, a large high-quality contemporary dataset of news headlines hand-annotated with metaphorical verbs. The dataset comprises headlines from various sources including political, satirical, reliable and fake. Our dataset serves the purpose of evaluation for the tasks of metaphor interpretation and generation. The experiments reveal several insights and limitations of using LLMs to automate metaphor processing tasks as frequently seen in the recent literature. The dataset is publicly available for research purposes[1].

1 Introduction

Metaphors are creative cognitive constructs designed to communicate an idea in an evocative fashion (Khaliq et al., 2021). Research on computational metaphor processing has explored a variety of questions related to the detection of metaphorical speech in text (Choi et al., 2021; Zhang and Liu, 2022) and its interpretation by readers (Rai et al., 2019; Aghazadeh et al., 2022). The task of metaphor generation (Ottolina and Pavlopoulos, 2022; Li et al., 2022; Stowe et al., 2021b) has recently gained traction due to the growing ability of LLMs to forge common sensical connections.

Metaphors are highly creative constructs of human language that grow old and eventually die (Rai et al., 2017). The vast majority of studies on metaphors in the English language still rely on datasets such as TroFi (Birke and Sarkar, 2006), VUA Metaphor Corpus (Steen et al., 2010), and LCC (Mohler et al., 2016) that contain archaic source texts (See Table A1). For instance, the *news* genre in VUA Metaphor corpus (Steen et al., 2010) which is derived from BNC Baby[2] has the latest headline from the year 1994.

Dead metaphors in these archaic source texts are essentially ineffective training samples. Moreover, the object of interest that is, metaphors from *contemporary* world texts are lacking. For instance, consider the metaphor *heal* in "love *heals* soul" that does not require much thought to comprehend versus a phrase like "Amtrak dining car *heals* nation" taken from a headline in our dataset.

In this paper, we propose NewsMet, a large high-quality contemporary dataset of news headlines hand-annotated with metaphorical verbs. Metaphors are a commonly used figurative construct in news headlines to better explain a complex event or scenario. For instance, consider the phrases <companies, *pushing*, boundaries/reforms> vs <companies, *pushing*, microchip implants>. The metaphorical verb *push* with *pushing implants* is a relatively new use. News headlines thus provide an evolving linguistic backdrop with new entities to learn contemporary metaphor use. Metaphors are also routinely used to make a political or social argument. Metaphorical language imprints emotions that one may not have anticipated otherwise and therefore, metaphorical news could be an interesting data source to evaluate the detection of hyperpartisan content.

In this paper, we make the following contributions:

- We present a large dataset of high-quality contemporary metaphors from *natural* settings published during 2017-2018. The dataset comprises headlines from sources identified as *reliable, fake, bias* etc. More information is provided in Table 3.
- We investigate the quality of predictions gen-

[1]https://github.com/AxleBlaze3/NewsMet_Metaphor_Dataset

[2]http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf

Findings of the Association for Computational Linguistics: ACL 2023, pages 10090–10104
July 9-14, 2023 ©2023 Association for Computational Linguistics

erated by Large Language Models (LLMs) for the task of metaphor detection, interpretation and generation concerning (a) correctness, (b) likelihood and (c) *goodness* of the generated predictions.

We believe that the proposed dataset will be an invaluable resource for researchers studying metaphor processing, providing a rich and diverse set of samples. We further believe that metaphors in news headlines will help understand and tackle the growing bias and hyperpartisan in digital media. Additionally, the dataset could be utilised to evaluate natural language understanding in LLMs.

2 Background

Early approaches for metaphor processing focused on analyzing restricted forms of linguistic context such as the subject-verb-object (SVO) type grammatical relation, using hand-crafted features (Bollegala and Shutova, 2013; Rai et al., 2018). Later, the approaches evolved to capture implicit relationships in long text through word embeddings and large language models. Rai and Chakraverty (2020) and Tong et al. (2021) provide a detailed discussion on these approaches as well as existing datasets for the tasks of linguistic metaphor detection and interpretation. Metaphor generation research in particular has recently gained traction with quite a few approaches exploiting neural language models as their underlying knowledge base (Stowe et al., 2021a; Chakrabarty et al., 2021).

To evaluate machine-generated metaphorical content, BigBench (BIG-bench collaboration, 2021) proposed four metaphor-related classification tasks (figure_of_speech_detection, metaphor_boolean, metaphor_understanding and identify_odd_metaphor). These tasks use newly curated datasets from sources such as online literature and existing datasets. However, the average size of the 4 datasets mentioned is 255, limiting their usefulness for the task of evaluation. Moreover, the aforementioned BigBench tasks and some other newly proposed datasets like Do Dinh et al. (2018) and IMPLI (Stowe et al., 2022) use source texts from old datasets, such as the BNC corpus (dated 1975-1995) and SemEval 2013 Task 5 (data collected in 2009). This limits their novelty as the source text remains similar to old datasets. Alnajjar et al. (2023) curated multi-modal *Ring That Bell* corpus of permissively licensed YouTube videos that have human-authored closed captions in English with metaphors annotated by human experts. This dataset has visual and audio clues that provides additional context for text interpretation.

Existing datasets for metaphor interpretation (See Table A2) and generation (See Table A3) are included in the appendix A. Bizzoni and Lappin (2018a) posed the metaphor interpretation task as an entailment problem and provided a collection of 200 metaphorical sentences with four paraphrases. Liu et al. (2022a) introduced a corpus of over 10k creative sentences based on the Winograd Schema to test common sense reasoning of models for figurative text. An example is "The dinner has the flavor of a rubber duck." with two paraphrases. However, the text itself reveals the inherent property (*flavor*) which makes it inappropriate for the task of metaphor interpretation. Zayed et al. (2020a) built a corpus of 1350 verb-object metaphoric pairs with "dictionary definitions". Recently, Chakrabarty et al. (2022) released FLUTE having 750 metaphors with two paraphrases. This dataset is however built using old sources.

To the best of our knowledge, there is no gold-label dataset to evaluate metaphor generation tasks. Chakrabarty et al. (2021) provide a method to generate silver labels using a BERT model finetuned on VUA and LCC (See Table A1). The authors ensure quality by considering sentences with probability > 0.95 (Pg 4252, Sec 2.1). However, it is worth noting that 83.38% of the predicted samples had probabilities > 0.95.

Through this work, we aim to bridge the gap for a high-quality contemporary dataset of English metaphors. The proposed dataset has fresh, expert-authored, thought-provoking metaphors that were used in contemporary world contexts. Additionally, we help address the gold data scarcity challenging metaphor interpretation and generation tasks by providing samples for all three use cases i.e. detection, interpretation and generation.

3 Proposed Approach

Figurative text annotation is a non-trivial task that demands significant cognitive effort and time. Below are the key concerns that we considered while designing the annotation pipeline:

- Sparsity: Headlines with *metaphors* are likely rare. A random sampling of headlines for manual annotation tasks thus may lead to a skewed distribution favoring literal headlines.

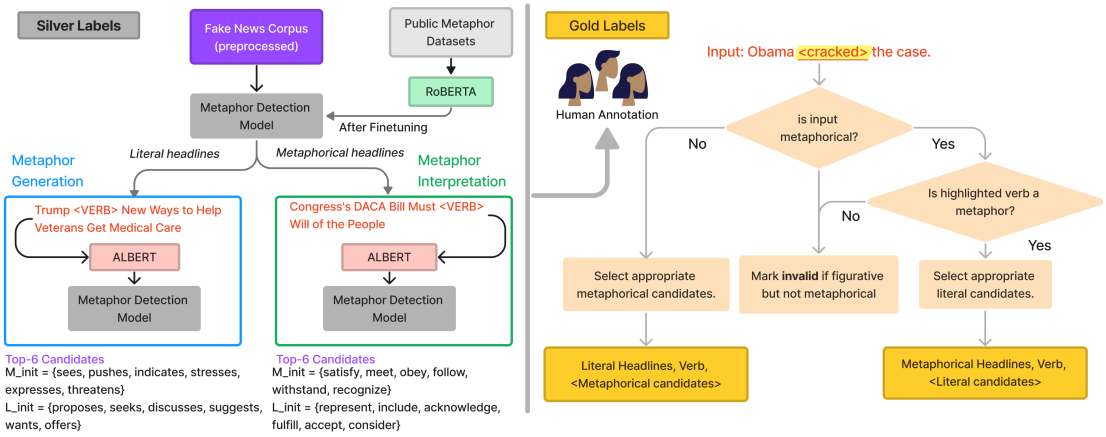

Figure 1: Here, we present the pipeline for news headlines annotation. In **Silver Labels** phase, we automatically predict plausible candidates for tasks (a) Metaphor Detection, (b) Metaphor Generation and (c) Metaphor Interpretation. The silver labels are manually verified and corrected as required in **Gold Labels** phase.

- Subjectivity: Metaphors are highly subjective cognitive constructs. It is thus important to capture diverse perspectives from multiple annotators for the tasks namely metaphor interpretation and generation.
- Skilled Annotators: Humans are adept at identifying *good* vs *bad* metaphors. However, they struggle when asked to create metaphors. Without skilled experts, it may be difficult to think of *good* metaphors.

To alleviate these concerns, we use state-of-the-art large language models (LLMs) for generation of *silver candidates*. With the help of a LLM, we identify *possibly* metaphorical and literal headlines that are then randomly sampled for manual annotation. The intent is to improve the likelihood of seeing metaphorical headlines during manual annotation. We also generate a diverse set of metaphorical and literal candidates using an LLM for human evaluation. We believe that human annotators will be able to appreciate novel interpretations and metaphorical mappings when presented, and will be quick to discard incorrect or absurd connections. This will ensure diversity as well as help us overcome the need for highly skilled annotators. The proposed annotation pipeline is illustrated in Figure 1.

3.1 News Headlines Corpus

For dataset creation, we use the open source Fake News Corpus (Szpakowski, 2020) that contains news articles from various sources known for *clickbait, reliable, fake* articles. At the time of data collection, the latest news article in this dataset was scraped on Feb 28, 2018. The publication date was identified using the URL associated with the news article. We considered headlines published from 2017-2018.

We started with a subsample of $100k$ headlines having > 7 words. The threshold was decided empirically to ensure sufficient context. For this study, we focus on verb metaphors, hence we identified headlines that contained a single SVO triplet[3]. We also ensured the verb in the SVO triplet is the ROOT when parsed using spaCy (Honnibal et al., 2020). For example, consider the headline:

"Jeff Sessions due to face Democrats' Russia questions next week."[4]

The detected SVO triplet is {Sessions, *face*, questions} and *face* is the ROOT as seen in Figure 2. Hence this headline will be retained.

An example of a headline that would be dropped is:

"France Threatens Brexit Deal Unless UK Takes More Calais Migrants." [5]

Here, the detected SVO is {UK, Takes, Migrants} but the ROOT obtained after dependency parsing is *Deal*. We thus remove this headline from the set. The intent behind this filtering is to identify the subject and object associated with a verb which is often critical information when determining metaphoricity.

[3] https://github.com/NSchrading/intro-spacy-nlp/blob/master/subject_object_extraction.py
[4] Rawstory
[5] Breitbart

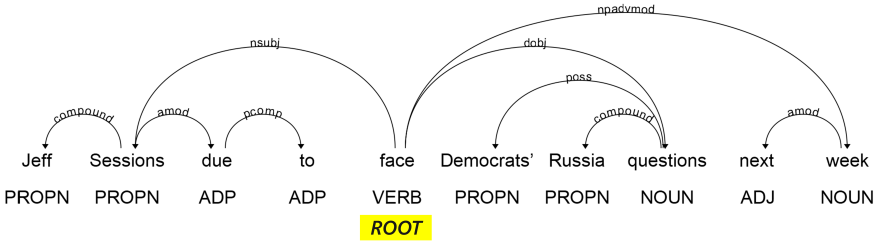

Figure 2: Dependency parse tree as generated by spaCy. Here, *face* is the ROOT.

Dataset	Class	$F1_{test_1}$	$F1_{test_2}$
D_{imbal}	L	0.78	0.76
	M	0.78	0.71
D_{bal}	L	0.71	0.58
	M	0.69	0.23

Table 1: Performance Evaluation of Models finetuned on D_{imbal} and D_{bal}. The train validation split is 90 : 10. L stands for *Literal* and M for *Metaphorical*.

We refer to the ROOT as the *focus verb* in the rest of the paper. At this stage, we have 28742 unique headlines. We pick a random sample of $15k$ headlines for the next stage keeping in mind the manual effort for annotation.

3.1.1 Metaphor Detection

To tackle *sparsity*, we finetuned a state-of-the-art large language model, RoBERTa (Liu et al., 2019) to automate the task of identifying possible metaphorical headlines. The model is openly available and well-suited for the task of text classification. We utilized the publicly available datasets listed in Table A1 for finetuning. This collection has a total of $62k$ samples of which $52k$ were metaphorical. We call this set D_{imbal}. To reduce the imbalance between the literal and metaphorical classes, we added $29k$ unique literal sentences from WikiQA (Yang et al., 2015) that is known to be highly objective. We also empirically verified the literalness of sentences by manually inspecting a random sample of 200. Adding these sentences increased the literal samples to $39k$. We then randomly sampled an equal number of metaphorical samples from D_{imbal}. We call this new set D_{bal} having an equal number of metaphorical and literal samples i.e. $39k$ samples each.

We manually curated two balanced *test sets* (a) $test_1$ consisting of a random sample from D_{imbal}. Please note, these samples were not used for training purposes. (b) $test_2$ having headlines from fake news corpus. Both test sets had 50 samples each. The Fleiss kappa (Fleiss et al., 1981) obtained for $test_1$ was 0.7 and Cohen kappa (Cohen, 1960) for $test_2$ was 0.76. Both test sets are provided at the link[6]. The performance of both models on these test sets is summarized in Table 1. The model trained on D_{imbal} substantially outperformed the other. Hence, we use the RoBERTa finetuned on D_{imbal}. We denote this model as M_{met_det}.

Out of the $15k$ headlines curated earlier, 10061 were predicted as metaphorical by M_{met_det} and 4939 as literal.

3.1.2 Candidate Sets for Metaphor Interpretation and Generation

For this task, we mask the *focus verb* in each headline and extract the top 200 candidate replacements generated by a LLM. As our goal is masked word replacement, we did not consider Causal Language Models such as GPT-n. After manually inspecting the quality of candidates generated by Masked Language Models including ALBERT (Lan et al., 2019), DeBERTa (He et al., 2020) and RoBERTa, we picked ALBERT for our task.

We denote the unfiltered initial set of top 200 candidates as C_{init}. As a postprocessing step, we filtered candidates that were purely non-alphabetic. Duplicate lemmas within the candidate sets were removed. We also removed non-verb candidates. The non-verb candidates were detected by substituting them in place of the focus verb and using spaCy's POS tagger (Honnibal et al., 2020). We will denote this filtered set of candidates as C_{filter}.

To segregate metaphorical candidates from literal candidates, we replace the *focus verb* in a given headline with the candidate verbs $c \in C_{filter}$ and predict the metaphoricity of the new headline containing the respective candidate using M_{met_det}.

[6] https://github.com/AxleBlaze3/NewsMet_Metaphor_Dataset/tree/main/data/custom_test_sets

We only consider the top 6 candidates to limit the effort and ensure the quality of the manual annotation task discussed in Section 3.2. Headlines having < 6 candidates were dropped. We refer to the top 6 candidates of the literal partition as L_{init}, metaphorical partition as M_{init} and the combined data as D_{init} respectively.

Consider the headline,

"Handling of Police Killing **Spurs** Grand Jury Inquiry Into Prosecutor." [7]

In this case, **spurs** is the focus verb. Here, C_{init} = { prompting, prompted, threatens, ..., headline, catalyze, 1944 }. On passing the candidates through M_{met_det}, we have M_{init} = { threatens, requires, triggered, sends, raises, starts } and L_{init} = { prompting, enters, begins, announces, asks, causing }

We considered a total of 2592 original headlines with candidate sets for gold label annotations. Of these 2592 headlines, 1430 were metaphorical and 1162 were literal as per predictions by M_{met_det}.

3.2 Gold Labels

At this stage, we have a collection of news headlines with silver labels that is <Headline, Focus verb, Class (M/L), M_{init}, L_{init} >. We design the gold label annotation process as illustrated in Figure 1 and described below.

- **Task-1**: Identify if the given text is metaphorical or literal. We use the guidelines[8] as provided in the Metaphor Identification Procedure VU University Amsterdam (MIP-VU) (Pragglejaz_Group, 2007) for annotation. Annotators were encouraged to use the Merriam-Webster Dictionary[9] to help identify basic and contextual meanings of lexical units. A headline is marked *invalid* if it contains figurative text that is not metaphorical. This includes metonymy, idioms, sarcasm and so on. The annotators are duly explained meanings of these linguistic constructs with examples contrasting them with metaphorical use.

- **Task-2** Identify if the highlighted verb is metaphorical. This is an important step as the generation of metaphorical or literal candidates are performed by masking the focus verb.

- **Task-3** Verify the semantic appropriateness and metaphoricity of the metaphorical or literal predictions provided by LLMs (RoBERTa and ALBERT) as is the case.
  - If the sentence and the focus verb are metaphorical, then annotators were asked to identify the candidates from the given literal set which make the sentence more literal while preserving the original meaning of the sentence when substituted in place of the focus verb. We denote this set as L_{final}.
  - Analogously, if the sentence and the focus verb are literal, then the annotators are prompted to identify candidates from the given metaphorical candidate set that makes the sentence more metaphorical while preserving the original meaning of the sentence when substituted in place of the focus verb. We denote this set as M_{final}.

3.2.1 Annotation Interface

We used Streamlit[10] an open-source app framework to build our annotation interface (see Figure A1). In a particular week, each annotator was given a maximum of 100 headlines to annotate. There were four questions to be answered.

Q1 Is the sentence metaphorical?

Q2 Is the focus verb metaphorical?

Q3 Which of the candidates makes the sentence more metaphorical / literal?

Q4 Which of the candidates satisfies the above condition while preserving the meaning of the sentence?

We implemented a quality check (QC) mechanism to evaluate the trustworthiness of annotators. A QC question is triggered randomly that has a definitive answer for Q1 and Q2. Annotations by annotators who performed poorly on the QC metric are discarded.

3.2.2 Human Annotators

A total of 15 annotators volunteered for our task. Each annotator was 18-22 years old and a native of India. Twelve annotators identified themselves as male and the remaining as female. Each annotator

[7] New York Times

[8] http://www.vismet.org/metcor/documentation/MIPVU.html

[9] https://www.merriam-webster.com/

[10] https://streamlit.io/

Group	D_{gold}	D_{gold+}	
Invalid	314	-	-
H_m & V_m	389	1009	
H_m & V_l	205	-	
H_m & (V_m ‖ V_l)	594	1009	=1603
H_l & V_l	611	455	=1066
	=1205	=1464	= 2669

Table 2: Distribution of NewsMet Dataset. Here, H_m indicates metaphorical headline, V_m indicates metaphorical verb and H_l, V_l are the literal counterparts. *Invalid* indicates headlines that were figurative but not metaphorical such as idioms and metonymies.

Type	$\#H_m$	$\#H_{(m+l)}$	$\%_m$
Political	213	428	49.7%
Satire	85	178	47.7%
Reliable	80	157	50.9%
Fake	71	161	44%
Bias	35	73	47.9%
Clickbait	30	50	60%
Conspiracy	27	52	51.9%
Unknown	40	76	52.6%
Others	13	30	43.3%
Total	594	1205	

Table 3: Type Distribution in NewsMet. Here, $\#H_m$ indicates the number of metaphorical headlines. $\#H_{m+l}$ indicates the total number of headlines including metaphorical and literal. $\%_m$ indicates the proportion of metaphorical headlines. The type label *Others* includes types with less than 10 articles such as Hate, JunkSci, Unreliable.

was a fluent speaker of English. To ensure that the annotators had a uniform understanding of the task, each annotator completed a brief training before undertaking the quality check task. After quality check evaluation, we selected 8 annotators who demonstrated a thorough understanding of the task to carry out the final set of annotations on streamlit.

4 Results & Discussion

4.1 NewsMet Dataset

A total of 1519 headlines containing 795 unique focus verbs were manually annotated. The final distribution of labels is provided in Table 2. D_{gold} indicates the set of hand-annotated original headlines. Over 44% of the headlines irrespective of type including *reliable* were marked as metaphorical (see Table 3).

The set of literal headlines can be further transformed into metaphorical headlines by replacing the focus verb with verbs from the verified candidate list M_{final}. Likewise, the metaphorical headlines can be converted to literal headlines by picking candidates from L_{final}. For instance, consider the literal headline below,

'Trump **blames** Obama again for Russian hacking — but still refuses to do anything about it'[11].

Here, **blames** is the focus verb and M_{final}= { slammed, kicks }. If we substitute *slammed* in place of **blame**, the transformed headline with the metaphorical verb is as follows:

Trump *slammed* Obama again for Russian hacking — but still refuses to do anything about it.

We use D_{gold+} to denote this expanded set of headlines. We thus have 1009 new metaphorical headlines and 455 new literal headlines in addition to the original 1519 headlines in our corpus. This combined set of 2669 headlines can be used for training figurative text detection models. To the best of our knowledge, the news genre in VUA Metaphor corpus[12] has 1451 metaphorical sentences out of 1704.

A total of 266 headlines (H_m & V_m) have at least one literal interpretation whereas, 445 headlines (H_l & V_l) have at least one metaphorical interpretation.

4.2 Quality of Silver Labels

4.2.1 Correctness

Out of 1430 headlines predicted as metaphorical by M_{met_det}, 41% were annotated as metaphorical by human annotators. That is, a false positive rate of 59%. Likewise, 27% were false negatives that is, metaphorical headlines were predicted as literal by M_{met_det}. It is thus important that finetuned Metaphor Detection models are properly validated and tested on out-of-domain corpora.

Out of 389 valid headlines considered for metaphor interpretation (that is H_m & V_m), 32% had no correct candidate and another 32% had only one correct candidate. Whereas for the task of metaphor generation (that is H_l & V_l), 27% were found to have no correct candidate and almost 24% were assigned only one correct candidate by human annotators. Interestingly, metaphor interpretation

[11] Rawstory

[12] https://github.com/jayelm/broader-metaphor

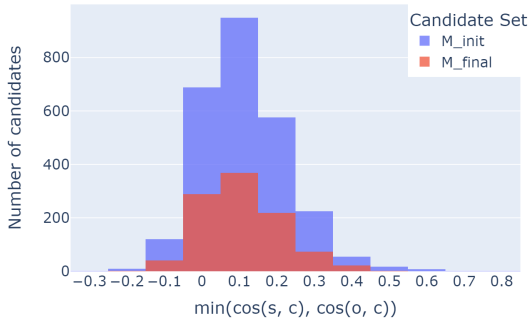

Figure 3: Metaphoricity i.e. $\min(\cos(s,c),\cos(o,c))$ distribution for generated candidates from M_{init} and M_{final}.

has a higher error rate compared to the relatively difficult and *creative* task of metaphor generation.

4.2.2 Likelihood of finding the *first* candidate

To determine the position of the first discovered candidate, we use the candidate's index position as in C_{init} before any postprocessing (as described in Section 3.1.2).

Metaphor Generation: 89% of the candidates in M_{init} were found within the interval $[0, 10]$ and that increased to 96.5% on stretching the interval to $[0, 20]$. In contrast, only 54% of the candidates in M_{final} were in the interval $[0, 10]$ and it went up to 66.2% for the interval $[0, 20]$. There were no candidates found in 27% of the cases. The high percentage of headlines without any correct candidates indeed indicates that M_{met_det} was overconfident in labeling metaphors.

Metaphor Interpretation: On repeating the same experiment on L_{init}, 94% of the candidates were within the interval $[0, 10]$. This went up to 98.7% when we considered the interval as $[0, 20]$. However, only 57% of the candidates in L_{final} were located within the interval of $[0, 10]$. This increased to 65% for the interval $[0, 20]$. There were no candidates found in 32% of the cases. This could be attributed to the lack of an explicit mechanism to filter out candidates semantically farther from the *focus verb*.

4.2.3 Goodness of Prediction

The *goodness* of prediction is computed only for the task of metaphor generation. The aim is to probe the metaphoricity and diversity of the generated metaphorical candidate set.

Algorithm 1 Measuring $Diversity$

Input: $H \leftarrow headline, fv \leftarrow focus\_verb, M_{init}$
Output: $clusters, diversity$

$clusters \leftarrow \{\phi\}$ ▷ set of clusters of similar words
$index \leftarrow 0$
for i in range(len(M_{init})) **do**
  $c_i \leftarrow M_{init}[i]$
  $H \leftarrow H.swap(fv, c_i)$
  $synset_1 \leftarrow lesk(H, c_i, verb)$
  for j in range(i+1,len(M_{init}) **do**
    $c_j \leftarrow M_{init}[j]$
    $H \leftarrow H.swap(fv, c_j)$
    $synset_2 \leftarrow lesk(H, c_j, verb)$
    $sim\_score \leftarrow synset_1.lch\_sim(synset_2)$
    ▷ Threshold is decided empirically. Here, it is 1.7
    if $sim\_score > threshold$ **then**
      if $c_i \in clusters$ **then**
        $clusters[find(c_i)].union(c_j)$
      else if $c_j \in clusters$ **then**
        $clusters[find(c_j)].union(c_i)$
      else
        $clusters[index++].union(c_i, c_j)$
      end if
    end if
  end for
end for
$diversity = |clusters|$
return $diversity, clusters$

Metaphoricity: We leverage the notion of *incongruity* (Wilks, 1975) to determine the metaphoricity of the generated candidates with respect to their subject(s) and object(o) in a given headline. We make a simple assumption that a metaphorical candidate $c \in M_{init}$ is incongruous either with its subject or object in the headline. A lower similarity indicates a higher distance between the candidate word and its surrounding context and therefore, higher metaphoricity (Yu and Wan, 2019). We thus consider the function $\min(\cos(s,c),\cos(o,c))$ when determining metaphoricity.

We use the cosine similarity function found in the Gensim library (Řehůřek and Sojka, 2010) and GloVe-300d embeddings (Pennington et al., 2014) to estimate the dissimilarity for our experiments. We plot the distribution of candidates in M_{init} in Figure 3.

The plot does reflect a pattern for lower cosine similarity and therefore higher metaphoricity, which is in tune with the hand-annotated candidates in M_{final}. We also note that candidates having low $\min(\cos(s,c),\cos(o,c))$ are more likely to be marked metaphorical by human annotators compared to other machine-generated candidates.

Diversity: A metaphor is a mapping between a TARGET domain and SOURCE domain. Consider

Headline (**focus verb**)	**Clusters**$_i$	*diversity*
Far Cry 5': Coop Mission ***has*** a massive problem with mission progress	{confronts, faces}$_1$,{creates, causes}$_2$, {finds}$_3$, {tackles}$_4$	4
Wait... Peter Strzok ***Discussed*** 'Insurance Policy' Against Trump Presidency With Andrew McCabe?	{sells, buying, traded}$_1$, {wants}$_2$, {gets}$_3$, {considers}$_4$	4
Could One of These Four Screenplays ***Win*** the Oscar?	{steal, snatch, take}$_1$, {grab}$_2$, { snare}$_3$, { scoop}$_4$	4
Sens. Cory Booker, Al Franken and Elizabeth Warren ***propose*** that the U.S. 'prevent genocide'	{demanded ask}$_1$, {insist, require, recommend}$_2$, { suggest}$_3$	3

Table 4: A subset of Headlines with $clusters$ and $diversity$ as generated by Algorithm 1. Here, $_i$ indicates the cluster number, *diversity* is the number of clusters.

the metaphorical phrase, 'My car *drinks* gasoline.'. Here, the mapping is CAR IS ANIMATE and *drink* is the linguistic manifestation from the domain ANIMATE.

In an attempt to quantify the variety within the candidate set, we measure the *diversity*. That is, the number of clusters formed after grouping conceptually similar words. This is an approximation to count the unique SOURCE DOMAINS in M_{init}. Our algorithm for clustering conceptually similar words is provided in Algorithm 1.

The input is { headline, focus verb, M_{init} }. The objective is to group candidates having *strong is-a* relationship. Using Lesk Algorithm (Lesk, 1986), we first disambiguate the word sense to identify the right WordNet (Fellbaum, 2010; Loper and Bird, 2002) synset for a candidate word $c \in M_{init}$ as in lines 3-6. Using Leacock-Chodorow similarity (Leacock et al., 1998), we then determine the similarity between the synsets of any two candidates and accordingly cluster as in lines 7-18. We empirically decided the similarity threshold as 1.7 to be clustered together. We provide a few examples in Table 4. For instance, words such as *selling, buying* and *trading* are essentially representing an overlapping idea. Likewise, *create, cause* also have a shared meaning. On manual analysis of the clustered sets, we found them to be coherent and valid.

Out of 611 headlines, we found that 238 (38.9%) had a score of 6 for diversity that is, every candidate word was in its own cluster (#clusters = 6 as there are only six candidates). 161 (26.3%) had 5, 115 (18.8%) had 4, 65 (10.6%) had 3, 26 (4.2%) had 2 and only 6 (0.9%) had one. This is an encouraging result that indeed supports the use of LLMs to generate diverse sets for the task of metaphor generation.

Model	A	P	R	F1
M_{met_det}	0.54	**0.70**	0.35	0.46
$M_{D_{imbal}+D_{gold+}}$	**0.61**	0.68	**0.57**	**0.62**

Table 5: Baseline Performance Evaluation for the task of Metaphor Detection. RoBERTa model finetuned on D_{gold+} in addition to D_{imbal} performs significantly better than M_{met_det}.

4.3 Baseline Performance

We compared the performance of RoBERTa trained on $D_{imbal} + D_{gold+}$ with M_{met_det} (RoBERTa trained on D_{imbal} (See Sec 3.1.1)). We split D_{gold+} into 80% train, 10% validation and 10% test. While doing so, we ensured the test set T_{gold+} of 546 samples had no overlap with the train and validation sets in terms of headlines and their respective transformations.
The performance on T_{gold+} is summarized in Table 5. $M_{D_{imbal}+D_{gold+}}$ showcased a performance improvement of 7% in accuracy, 16% in F1 score and 22% in recall. We also performed McNemar's statistical significance test and obtained a statistic of 57 with $p < 0.01$ indicating the gain in performance is statistically significant.

5 Conclusion

In this paper, we proposed NewsMet formed from contemporary news headlines, hand-annotated with metaphorical verbs. The samples are provided with metaphorical and literal interpretations. This dataset is useful for building and evaluating automated systems to detect, interpret as well as generate metaphors. Our analysis cautions against the

blind use of fine-tuned metaphor detection models to annotate new corpora. However, LLMs could be of great help in curating diverse metaphorical candidate sets. The proposed dataset has a variety of news sources such as *reliable* and *bias* that can be useful in understanding the role of metaphors in news framing and hyperpartisan content. Machine translation of figurative text is an under explored research area. The literal candidates associated with metaphorical samples can be used to automatically evaluate the quality of translations. It would also be interesting to evaluate LLMs for the task of generating culturally coherent metaphors in translated news headlines.

Limitations

The proposed dataset is annotated for verb metaphors in particular. However, other lexical units including adjectives and adverbs should also be studied in order to truly understand the role of metaphors in news. It is important to examine the diversity of the generated ideas when performing metaphor generation. In this study, we proposed a simple approach to cluster words using WordNet. However, the metric is far from perfect and can be improved. For the task of candidate generation, we performed word masking to generate metaphorical and literal substitutes as we were curious about the ability of LLMs to generate relevant metaphorical mappings while preserving the underlying semantic idea. Direct substitution of metaphorical candidates resulted in syntactically incoherent sentences in a few cases. It may be better to paraphrase the sentence after selecting the metaphorical mapping (Ottolina and Pavlopoulos, 2022).

Ethical Concerns and Broader Impact

We created the dataset from a publicly available news headlines dataset. This ensures that data is free from (a) anonymity concerns, (b) obscenities and (c) any stereotyping or bias. As the task is cognitively intensive, we only assigned at most 150 headlines to each annotator. All annotators were duly acknowledged and appreciated by Nvidia AI Technology Center for their contribution. The original dataset of news headlines is the under Apache License 2.0. We are thus permitted to modify and redistribute it.

Generating metaphors carries concerns due to the implicit potential to craft misleading text. The usage of metaphors has been shown to resonate emotionally with readers (Citron and Goldberg, 2014). This should not be a concern with our data as we only release generated candidates that preserve the underlying semantic meaning of the source headline.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang. 2023. Ring that bell: A corpus and method for multimodal metaphor detection in videos.

BIG-bench collaboration. 2021. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018a. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55.

Yuri Bizzoni and Shalom Lappin. 2018b. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PLOS ONE*, 8(9):1–10.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

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.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Oana David, Ellen Dodge, J Hong, Elise Stickles, and E Sweetser. 2014. Building the metanet metaphor repository: The natural symbiosis of metaphor analysis and construction grammar. In *The 8th International Construction Grammar Conference (ICCG 8)*, pages 3–6.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Mohammed Abdul Khaliq, Rohan Joseph, and Sunny Rai. 2021. #covid is war and #vaccine is weapon? covid-19 metaphors in india. In *Proceedings of the 18th International Conference on Natural Language Processing (Long Papers)*, pages 431–438.

George Lakoff and Mark Johnson. 1980. Metaphors we live by.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Claudia Leacock, Martin Chodorow, and George A Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. Nominal metaphor generation with multitask learning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 225–235, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022a. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. 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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Saif M. Mohammad, Ekaterina Shutova, and Peter D Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*, Berlin, Germany.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Giorgio Ottolina and John Pavlopoulos. 2022. Metaphorical paraphrase generation: Feeding metaphorical language models with literal texts.

Brad Pasanek. 2015. The mind is a metaphor: Browse the database.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pragglejaz_Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, and Yash Kukreti. 2017. Soft metaphor detection using fuzzy c-means. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 402–411. Springer.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, and Yash Kukreti. 2018. A study on impact of context on metaphor detection. *The Computer Journal*, 61(11):1667–1682.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, Divyanshu Sharma, and Ayush Garg. 2019. Understanding metaphors using emotions. *New Generation Computing*, 37(1):5–27.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

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

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021a. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021b. 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.

Maciej Szpakowski. 2020. Fake news corpus. Available at https://github.com/several27/FakeNewsCorpus.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. 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 861–871.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020a. Figure me out: a gold standard dataset for metaphor interpretation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5810–5819.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020b. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5810–5819, Marseille, France. European Language Resources Association.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Appendix

A.1 Existing datasets

Dataset	Source	Time Period
Tsvetkov et al. (2014)	Web	2014
LCC (Mohler et al., 2016)	ClueWeb09 and debate politics forum	2009-unknown
TroFi (Birke and Sarkar, 2006)	WSJ	1987-1989
Master Metaphor List (Lakoff and Johnson, 1980)	published books, papers and research seminars	1991 edition
MetaNet (David et al., 2014)	Original	2013-2016
Mohammad et al. (2016)	Wordnet	Not Specified.
VUAMC (Steen et al., 2010)	BNC-Baby	1975-1995
The Mind is a Metaphor (Pasanek, 2015)	Various Sources	aim : 1660-1819
Grothe [13]	Various Sources	BC-21st century
WikiQA (Yang et al., 2015)	Wikipedia	Not Specified

Table A1: Metaphor Detection: Datasets used for finetuning RoBERTa

Dataset	Source	Time Period	**#M**	**#NM**	Label Types
Shutova (2010)	Mixed (BNC corpus)	2010	761	-	Gold
Bizzoni and Lappin (2018b)	Original (Crowd sourced)	2018	200	-	Gold
Zayed et al. (2020b)	Twitter	2020	1300	-	Gold
Liu et al. (2022b)	Hand written *metaphors* + *similes* (Crowdsourced)	2022	10256	-	Gold

Table A2: Existing Datasets for the Metaphor Interpretation Task

Dataset	Source	Time Period	**#M**	**#NM**	**Label Types**
MERMAID (Chakrabarty et al., 2021)	Gutenberg Poetry	1991-2016	90000	-	Silver

Table A3: Existing Datasets used for the Metaphor Generation task

A.2 Model Parameters

A.2.1 RoBERTa

No. of Parameters: We use the RoBERTa base checkpoint (125M parameters)[14].
No of Epochs: We finetuned the model for 7 epochs and save the best model based on validation accuracy.
Training Time: 2 hours
Training hyper parameters: We use popular parameters that is, learning rate: $2e-5$, dropout as 0.3 and AdamW as the optimizer.

A.2.2 ALBERT

We make use of the pre-trained albert-xxlarge-v2[15] checkpoint without finetuning.
No. of Parameters: 223M parameters

[14] https://huggingface.co/roberta-base
[15] https://huggingface.co/albert-xxlarge-v2

A.2.3 Hardware Configuration

We made use of Google colab[16] to fine-tune RoBERTa and make predictions. The service provides a variety of single GPU instances (commonly Nvidia T4 or P100) and assigns one based on availability. Total GPU hours equated to approximately 8.

A.3 Streamlit Annotation Interface

The interface is provided in Figure A1.

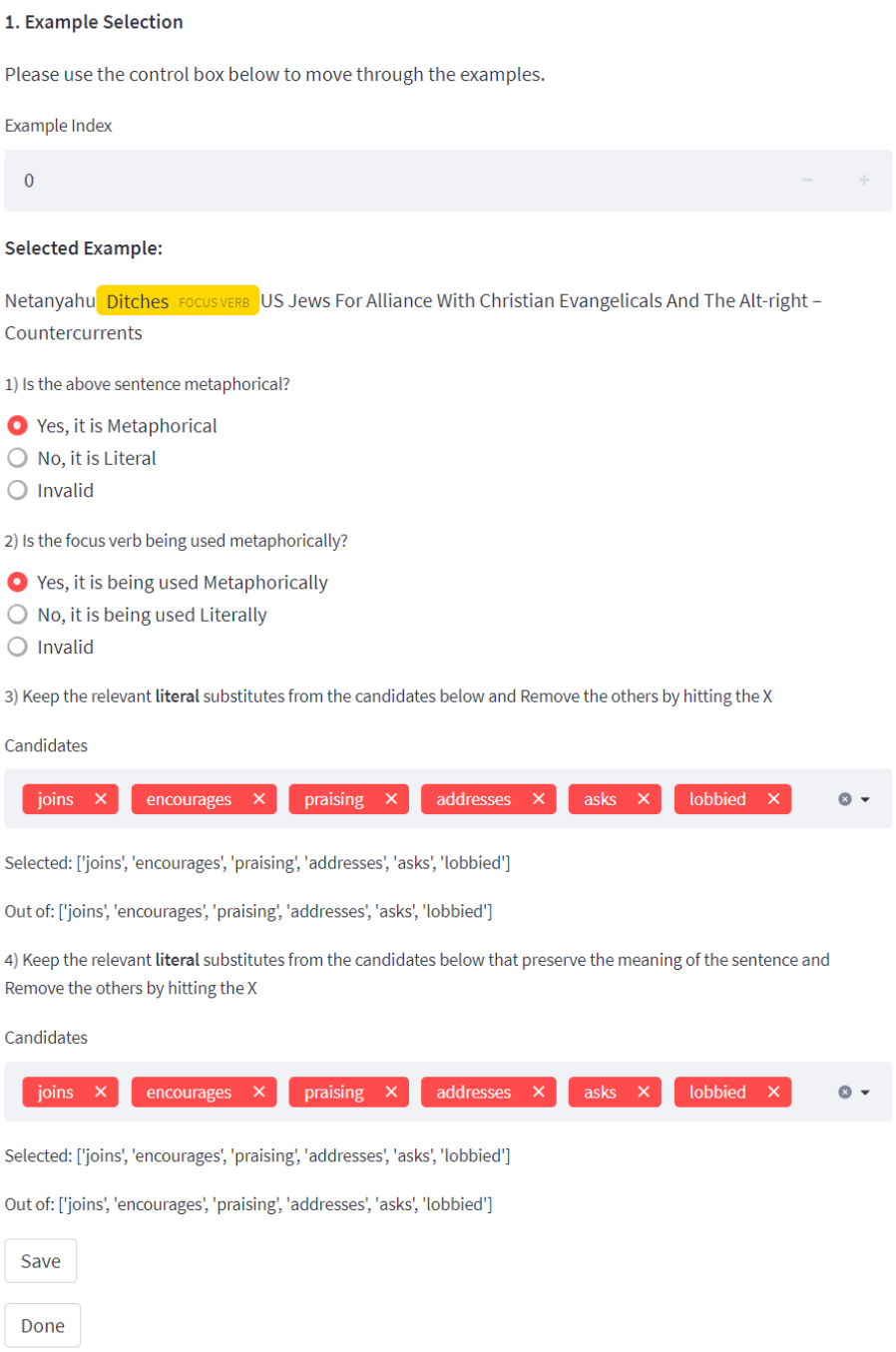

Figure A1: Interface for Annotation. Human annotators used this interface to (a) verify the metaphoricity of predicted verb metaphor and (b) identify semantically appropriated literal or metaphorical candidates as applicable.

[16] https://colab.research.google.com/

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 5

☑ A2. Did you discuss any potential risks of your work?
Section 6

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

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

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
On first mention of each resource across the paper (various sections)

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3 and Section 6

☑ 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)?
Section 6 we discuss redistribution of the dataset

☒ 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?
We use existing publicly available news headlines

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 3, Section 6, A

☑ 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.
In Section 3

C ☑ Did you run computational experiments?

Section 3

☑ 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 A1

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 and A1

☑ 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?
Section 3, Section 4

☑ 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.)?
Section 3, Section 4

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

Section 3

☑ 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.?
Section 3

☑ 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)?
Section 3

☑ 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?
Section 6, We did not collect data from people. Our annotators annotated existing available data (no personal data involved) and understood the task.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We are labelling existing public datasets with no human or social experimentation involved.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3, full details will be available after the double-blind review period