

Crowd-Sourcing A High-Quality Dataset for Metaphor Identification in Tweets

Omnia Zayed 

Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
omnia.zayed@insight-centre.org

John P. McCrae 

Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
john.mccrae@insight-centre.org

Paul Buitelaar 

Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
paul.buitelaar@insight-centre.org

Abstract

Metaphor is one of the most important elements of human communication, especially in informal settings such as social media. There have been a number of datasets created for metaphor identification, however, this task has proven difficult due to the nebulous nature of metaphoricity. In this paper, we present a crowd-sourcing approach for the creation of a dataset for metaphor identification, that is able to rapidly achieve large coverage over the different usages of metaphor in a given corpus while maintaining high accuracy. We validate this methodology by creating a set of 2,500 manually annotated tweets in English, for which we achieve inter-annotator agreement scores over 0.8, which is higher than other reported results that did not limit the task. This methodology is based on the use of an existing classifier for metaphor in order to assist in the identification and the selection of the examples for annotation, in a way that reduces the cognitive load for annotators and enables quick and accurate annotation. We selected a corpus of both general language tweets and political tweets relating to Brexit and we compare the resulting corpus on these two domains. As a result of this work, we have published the first dataset of tweets annotated for metaphors, which we believe will be invaluable for the development, training and evaluation of approaches for metaphor identification in tweets.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Language resources

Keywords and phrases metaphor, identification, tweets, dataset, annotation, crowd-sourcing

Digital Object Identifier 10.4230/OASICS.LDK.2019.10

Funding This work was supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

1 Introduction

Metaphor is an essential element of human cognition which is often used to express ideas and emotions. It is considered as an analogy between two concepts by exploiting common similarities. The sense of a concept such as “*war*” can be transferred to another concept’s sense such as “*illness*” by exploiting the properties of the first concept. This then can be expressed in our everyday language in terms of linguistic (conventional) metaphors such as “*attack cancer*” or “*beat the illness*” [11, 17]. Among the main challenges of the computational modelling of metaphors is their pervasiveness in language which means they do not only occur frequently in our everyday language but they are also often conventionalised to such



© Omnia Zayed, John P. McCrae, and Paul Buitelaar;
licensed under Creative Commons License CC-BY

2nd Conference on Language, Data and Knowledge (LDK 2019).

Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 10; pp. 10:1–10:17

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

10:2 Crowd-Sourcing A High-Quality Dataset for Metaphor Identification in Tweets

an extent that they exhibit no defined patterns. This has meant that achieving consistent annotations with higher inter-annotator agreement has been difficult and as such previous work has introduced restrictions, such as limiting the study to only a few chosen words of a certain syntactic type [1, 16, 32] or particular predefined metaphors [15, 31].

The widespread nature of Twitter communication has led to a growing interest in studying metaphors in such a context. People tend to use colloquial language in order to communicate on social media, and they may utilise figurative and metaphoric expressions more frequently. Twitter, which is the most popular microblogging application in the world, presents a new type of social media content, where users can express themselves through a tweet of limited characters. Processing metaphoric expressions in tweets can be very useful in many social media analysis applications such as political discourse analysis [3] and health communication analysis. Therefore, our goal is to create a dataset of tweets annotated for metaphors that offers comprehensive coverage of metaphoric usages as well as text genre. In order to achieve that, we need to design an annotation methodology that guarantees high annotator agreement at a large scale. Accordingly, the resulting dataset can be used for the development and evaluation of metaphor processing approaches in tweets.

There are different factors that affect the creation of datasets annotated for metaphor and their annotation scheme. Among these factors are the level of metaphor analysis and the type of metaphor, in addition to the task definition and the targeted application. Examples of metaphor types include conceptual, linguistic (conventional and novel) and extended metaphors. There exist different levels of metaphoric analysis of linguistic metaphors either on the word-level (token-level) or on the phrase-level. Processing metaphors on the word-level means looking at each word in a sentence and deciding whether it is used metaphorically or not given the context, while phrase-level processing looks at pairs of words such as verb-noun or adjective-noun pairs and check the metaphoricity of the verb or the adjective given its association with the noun. Various research has been done to address both levels of processing¹. The majority of previous approaches pertaining to metaphor identification have focused on formal well-structured text selected from a specific corpus to create datasets to model and evaluate their approaches. A common issue of all the available datasets is that they are specifically designed for a certain task definition focusing on a certain level of metaphor analysis which makes their annotation scheme difficult to generalise. Additionally, the majority of available datasets lack coverage of metaphors and text genres as they rely on predefined examples of metaphors from a specific domain during the creation process.

In this work, we introduce the first high-quality dataset annotated for phrase-level metaphor in English tweets. We propose a crowd-sourcing approach to create this dataset which is designed to ensure the dataset balance, coverage as well as high accuracy. Our approach employs an existing metaphor identification system to facilitate quick and accurate annotations as well as maintaining consistency among the annotators. We will outline the identification system used as well as the data sources in section 3 below. In this paper, we present our annotation methodology along with the results and analysis of the resulting dataset. We also provide a summary of the previous work done in past years to create annotated datasets for metaphor identification.

¹ We are not going to address it here as it is beyond the scope of this paper.

2 Related Work

In this section, we will discuss the most relevant research in terms of the dataset preparation and the annotation of linguistic metaphors. As discussed in the previous section, there are several factors that affect the dataset creation and the annotation scheme, including the task definition and the targeted application, which push the dataset creation towards a specific domain or text type. Past research in this area has focused on formal well-structured text such as news or has only targeted a selected examples of metaphors. The majority of researchers formulate their own annotation guidelines and definition of metaphor. One of the main challenges of this work is to introduce an annotation scheme that results in an expert annotated dataset for metaphor identification that have large coverage of metaphoric usages and text genres while maintaining high accuracy. Table 1 provides a detailed summary of the datasets annotated for linguistic metaphors.

TroFi Example Base [1] is one of the earliest metaphor datasets which consists of 3,737 manually annotated English sentences extracted from the 1987-1989 Wall Street Journal corpus (WSJ) covering the literal and metaphoric senses of 50 selected verbs. The dataset has been frequently used to evaluate approaches for metaphor analysis, however there is no information available regarding the inter-annotator agreement (IAA), so its value is questionable. Turney et al. [32] created a dataset of 100 sentences from the Corpus of Contemporary American English (COCA) [5] focusing on metaphoric adjectives. The dataset contains five selected adjectives forming twenty adjective-noun pairs which were manually annotated by five annotators.

Steen [30] employed the metaphor identification procedure (MIPVU) to annotate metaphors in a subset of the British National Corpus (BNC) [2], namely BNC Baby, in order to create the VU Amsterdam Metaphor Corpus (VUA) which has become one of the most popular existing metaphor datasets nowadays. The corpus consists of randomly selected texts from various text genres. Their collaborative annotation scheme annotates metaphors on the word-level, regardless of the word's syntactic type, considering a word as a metaphor as long as its most basic meaning, derived from corpus-based dictionaries, contradicts its current contextual meaning. The basic meaning is typically the most physical or concrete meaning which does not have to be the first sense listed under a word entry. The MIPVU employs two other dictionaries in addition to the corpus-based dictionary. The IAA was measured in terms of Fleiss' kappa [9] among four annotators which averaged 0.84. One of the issues with this procedure is that the sense of every word in the text is considered as a potential metaphor, even idioms or fixed collocations, which are considered inseparable lexical units. Moreover, the annotators have to go through a series of complex decisions starting from chunking the given text into lexical units, then discerning their basic meaning, and finally the metaphoric classification. The uniformity of the basic meaning interpretation may vary from one annotator to the other. Shutova and Teufel [27] adopted the MIPVU annotation scheme, with some modifications, to annotate linguistic metaphors on the word-level focusing on verbs in a subset of the BNC. The corpus comprises 761 sentences and 13,642 words. The IAA was evaluated by means of κ [29] which averaged 0.64 among three native annotators. The authors reported that the conventionality of some metaphors is a source of disagreement. A subset of the VUA corpus comprises around 5,000 verb-object pairs has been prepared in [34]. The adapted VUA subset is drawn from the training verbs dataset from the VUA corpus provided by the NAACL 2018 Metaphor Shared Task². The authors retrieved the

² <https://github.com/EducationalTestingService/metaphor/tree/master/NAACL-FLP-shared-task>

original sentences of around 17,240 annotated verbs, which yielded around 8,000 sentences. Then the verb-direct object relations were extracted using the Stanford parser [4]. The classification of each verb-noun pair was decided based on the metaphoric classification of the verb provided in the original corpus.

Hovy et al. [15] created a dataset by extracting sentences from the Brown corpus [10] to identify metaphors of any syntactic structure on the word-level. They used a list of 329 predefined metaphors as seed to extract sentences that contain the specified expressions. The dataset is manually annotated using crowd-sourcing through Amazon Mechanical Turk (MTurk) platform. The annotators were asked whether a highlighted word in a sentence was used metaphorically or not based on its original meaning. This approach is similar to ours but we annotated metaphoric expressions on the phrase-level focusing on verb-noun pairs. The IAA among seven annotators was 0.57. The annotated instances were filtered out yielding a final corpus consisting of 3,872 instances, out of which 1,749 contains metaphors. Mohler et al. [21] created a dataset focusing on linguistic metaphors in the governance domain. The dataset consists of 500 documents ($\sim 21,000$ sentences) manually annotated by three annotators which were extracted from political speeches, websites, and online newspapers. In 2016, the Language Computer Corporation (LCC) annotated metaphor datasets [22] was introduced. The English dataset was extracted from the ClueWeb09 corpus³. The freely available part of the dataset contains $\sim 7,500$ metaphoric pairs of any syntactic structure annotated by adopting the MIPVU scheme. There is no clear information regarding the number of annotators or the final IAA of this subset. Tsvetkov et al. [31] created a dataset of $\sim 2,000$ adjective-noun pairs which were selected manually from collections of metaphors on the Web. This dataset is commonly known as the TSV dataset and is divided into 1,768 pairs as a train set and 222 pairs as a test set. An IAA of 0.76 was obtained among five annotators on the test set. The annotators were asked to use their intuition to define the non-literal expressions.

Mohammad et al. [20] annotated different senses of verbs in WordNet [8] for metaphoricity. Verbs were selected if they have more than three senses and less than ten senses yielding a total of 440 verbs. Then the example sentences from WordNet for each verb were extracted and annotated by 10 annotators using crowd-sourcing through the CrowdFlower platform (currently known as Figure Eight). The verbs that were tagged by at least 70% of the annotators as metaphorical or literal were selected to create the final dataset. The dataset consists of 1,639 annotated sentences out of which 410 were metaphorical and 1,229 literal. This dataset, commonly known as the MOH dataset, had been used to model and evaluate systems identifying metaphoric verbs on the word-level. A subset of the MOH dataset has been adapted in [26] to extract the verb-subject and verb-direct object grammar relations, in order to model computational approaches that analyse phrase-level metaphors of verb-noun pairs. The final dataset consists of 647 verb-noun pairs out of which 316 instances are metaphorical and 331 instances are literal.

In an attempt to detect metaphors in social media, Jang et al. [16] acquired a dataset of 1,562,459 posts from an online breast cancer support group. A set of eight predefined words, that can appear either metaphorically or literally in the corpus, were employed to classify each post. An IAA of 0.81 was recorded in terms of Fleiss' kappa among five annotators on MTurk who were provided by a Wikipedia definition of metaphor. Twitter datasets of a figurative nature were prepared in the context of the SemEval 2015 Task 11 on Sentiment Analysis of Figurative Language in Twitter [12]. This dataset is referred to here as the SemEval 2015

³ <https://lemurproject.org/clueweb09/>

SAFL dataset. The dataset is originally designed to support the classification and sentiment analysis of tweets containing irony, sarcasm, and metaphors. The available training, and test sets were collected based on lexical patterns that indicate each phenomenon such as using the words “figuratively” and “literally” as lexical markers to collect the metaphoric tweets. Shutova et al. [28] studied the reliability of such technique and discussed that the dependence on lexical markers as a signal of metaphors is not sufficient. The training dataset contains 2,000 tweets which the organisers categorised as metaphoric tweets. We manually annotated a subset of arbitrary selected 200 tweets of the training dataset for use in our preliminary experiments.

Recently, Parde and Nielsen [23] exploited the VUA corpus to create a dataset of phrase-level metaphors annotated for novelty. In this work, 18,000 metaphoric word pairs of different syntactic structures have been extracted from the VUA corpus. Five annotators were then asked to score the highlighted metaphoric expression in a given context for novelty in a scale from 1 to 3. The annotation experiment was set up on MTurk and an IAA of 0.435 was achieved. Another work that addresses metaphor annotation for novelty is [6] focusing on word-level metaphors. Similar to [23], the authors exploited the VUA corpus to annotate 15,180 metaphors for novelty using crowd-sourcing. Different annotation experiments were set up on MTURK to decide: 1) the novelty and conventionality of a highlighted word, 2) the scale of novelty of a given metaphor, 3) scale of “unusualness” of a highlighted word given its context, and 4) the most novel and the most conventionalised metaphor from given samples. The annotators obtained an IAA of 0.39, 0.32 and 0.16 in terms of Krippendorff’s alpha for the first three tasks, respectively.

3 Data Preparation

Our aim is to prepare a high-quality annotated dataset focusing on balance, coverage, and representativeness. These factors [7] are central to building a corpus so we considered them besides the other factors discussed earlier. In this section, we discuss the data sources and the preparation steps for creating a dataset annotated for metaphor in tweets.

3.1 Sources

In order to avoid targeting specific topic genres or domains, we utilised two data sources to prepare our dataset which represents two categories of tweets. The first category is general domain tweets which is sampled from tweets pertaining to sentiment and emotions from the SemEval 2018 Task 1 on Affect in Tweets [19]. The second category of data is of a political nature which is sampled from tweets around Brexit [13].

Emotional Tweets. People tend to use figurative and metaphoric language while expressing their emotions. This part of our dataset is prepared using emotion related tweets covering a wide range of topics. The data used is a random sample of the Distant Supervision Corpus (DISC) of the English tweets used in the SemEval 2018 Task 1 on Affect in Tweets, hereafter SemEval 2018 AIT DISC dataset⁸. The original dataset is designed to support a range of emotion and affect analysis tasks and consists of about 100 million tweets⁹ collected using emotion-related hashtags such as “*angry, happy, surprised, etc*”. We

⁸ available online on: https://competitions.codalab.org/competitions/17751#learn_the_details-datasets

⁹ Only the *tweet-ids* were released and not the tweet text due to copyright and privacy issues.

Table 1 Summary of the datasets created for linguistic metaphor identification. *The dataset is not directly available online but can be obtained by contacting the authors.

	Level of Analysis	Syntactic Structure	text type	domain	crowd-source	IAA	annotators	size	available
Birke and Sarkar, 2006 (TroFi Example Base)	word-level	verb	selected examples (News)	open	no	-	-	3,737 sentences	yes ⁴
Steen, 2010 (VUA)	word-level	any	known-corpus (The BNC)	open	in-house	0.84	4	~200,000 word (~16,000 sentences)	yes ⁵
Shutova et al., 2010	word-level	verb	known-corpus (The BNC)	open	in-house	0.64	3	761 sentences	yes*
Turney et al., 2011	word-level	verb adjective	selected examples (News)	open	no	-	5	100 sentences	no
Hovy et al., 2013	word-level	any	known-corpus (The Brown Corpus)	open	yes*	0.57	7	3,872 instances	no
Mohler et al., 2013	word-level	any	selected examples	restricted (governance)	no	-	-	21,000 sentences	no
Tsvetkov et al., 2014 (TSV)	phrase-level	adj-noun	selected examples (News)	open	no	-	-	2,000 adj-noun pairs	yes ⁶
Jang et al. 2015	word-level	noun	selected examples (Social Media)	restricted (breast cancer)	yes	0.81	5	2,335 instances	no
Mohler et al., 2016 (LCC)	word-level	any	known-corpus (ClueWeb09)	open	no	-	-	7,500 metaphoric pairs	partially
Mohammad, 2016 (MOH)	word-level	verb	selected examples (WordNet)	open	yes	-	10	1,639 sentences	yes ⁷
Shutova, 2016 (adaptation of MOH)	phrase-level	verb-noun	selected examples (WordNet)	open	-	-	-	-	yes*
Our dataset	phrase-level	verb-direct obj	tweets	open	yes	0.70-0.80	5	2500 tweets	yes*

⁴ <http://natlang.cs.sfu.ca/software/trofi.html>

⁵ <http://ota.ahds.ac.uk/headers/2541.xml>

⁶ <https://github.com/ytsvetko/metaphor>

⁷ <http://saifmohammad.com/WebPages/metaphor.html>

retrieved the text of around 20,000 tweets given their published *tweet-ids* using the Twitter API¹⁰. We preprocessed the tweets to remove URLs, elongations (letter repetition, e.g. verrrry), and repeated punctuation as well as duplicated tweets then arbitrary selected around 10,000 tweets.

Political Tweets. Metaphor plays an important role in political discourse which motivated us to devote part of our dataset to political tweets. Our goal is to manually annotate tweets related to the Brexit referendum for metaphor. In order to prepare this subset of our dataset, we looked at the Brexit Stance Annotated Tweets Corpus¹¹ introduced by Grčar et al. [13]. The original dataset comprises 4.5 million tweets collected in the period from May 12, 2016 to June 24, 2016 about Brexit and manually annotated for stance. The text of around 400,000 tweets on the referendum day is retrieved from the published *tweet-ids*. These tweets contained a lot of duplicated tweets and re-tweets. We cleaned and preprocessed them similar to the emotional tweets as discussed above yielding around 170,000 tweets.

3.2 Initial Annotation Scheme

We suggested a preliminary annotation scheme and tested it through an in-house pilot annotation experiment before employing crowd-sourcing. In this initial scheme, the annotators are asked to highlight the words which are used metaphorically relying on their own intuition, and then mark the tweet depending on metaphor presence as “*Metaphor*” or “*NotMetaphor*”. In this experiment, 200 tweets were extracted from the SemEval 2015 SAFL dataset mentioned in Section 2. The tweets are sarcastic and ironic in nature due to the way they were initially collected by querying Twitter Search API for hashtags such as “*#sarcasm, #irony*”. The annotation is done by three native speakers of English from Australia, England, and Ireland. The annotators were given several examples to explain the annotation process. We developed a set of guidelines for this annotation experiment in which the annotators were instructed to, first, read the whole tweet to establish a general understanding of the meaning. Then, mark it as metaphoric or not if they suspect that it contains a metaphoric expression(s) based on their intuition taking into account the given definition of a metaphor. A tweet might contain one or more metaphors or might not contain any metaphors. Finally, the annotators were asked to highlight the word(s) that according to their intuition has a metaphorical sense.

The annotators achieved an inter-annotator agreement of 0.41 in terms of Fleiss’ kappa. Although the level of agreement was quite low, this was expected as the metaphor definition depends on the native speaker’s intuition. The number of annotated metaphors varies between individual annotators with maximum metaphors’ percentage of 22%. According to the annotators, the task seemed quite difficult and it was very hard to pick the boundary between metaphoric and literal expressions. A reason for this is perhaps the ironic nature of the tweets, with some authors deliberately being ambiguous. Sometimes the lack of background knowledge adds extra complexity to the task. Another important challenge is the use of highly conventionalised language. The question that poses itself here is how to draw a strict line about which expression should be considered as a metaphor and which is not.

We concluded from this initial experiment that the annotation task is not ready for crowd-sourcing due to the previously mentioned limitations. It would be still an expensive task in terms of the time and effort consumed by the annotators. We explored the usage of

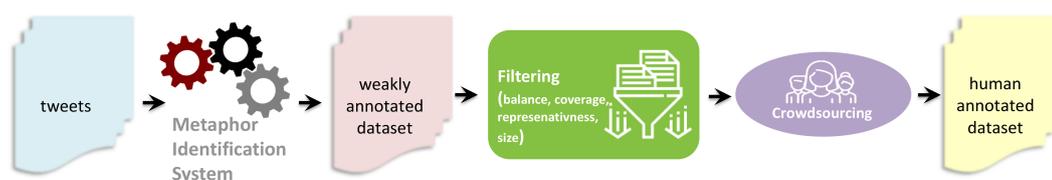
¹⁰ Twitter API: <https://developer.twitter.com/en/docs/api-reference-index>

¹¹ available online on: <https://www.clarin.si/repository/xmlui/handle/11356/1135>

WordNet as a reference for sense distinction on 100 tweets. An IAA agreement of 0.21 was achieved which is extremely low due to the annotators' disagreement on what they believed to be metaphors in their initial judgement, therefore they checked WordNet for different expressions. This initial pilot study also revealed that this dataset is not suitable for the annotation so we changed it as will be discussed in sub-section 3.1 to help improve the accuracy of the annotations.

3.3 Weakly Annotated dataset

In order to address the limitations of the initial annotation experiments, we prepared a weakly annotated dataset using a metaphor identification system, to reduce the cognitive load for annotators and maintain consistency. This system will be used to identify potential metaphoric expressions in tweets. Then, crowd-sourcing will be employed to ask a number of annotators to identify the expressions that are metaphorical in their judgement from these pre-identified ones. This way, the cognitive load on the annotators will be reduced while maintaining consistency. Figure 1 shows the process of creating our dataset.



■ **Figure 1** The proposed approach to create a dataset of tweets for metaphor identification.

Zayed et al. [34] introduced a weakly supervised system which makes use of distributed representations of word meaning to capture metaphoricity focusing on identifying verb-noun pairs where the verb is used metaphorically. The system extracts verb-noun pairs using the Stanford parser [4]. Then pre-trained word embeddings models are employed to measure the semantic similarity between the candidate pair and a predefined seed set of metaphors. The given candidate is classified using a previously optimised similarity threshold. We used this system to prepare a weakly annotated dataset using the data discussed in sub-section 3.1. The reason behind choosing this system is that it employs fewer lexical resources and does not require annotated datasets. Moreover, it is a weakly supervised system that employs a small seed set of predefined metaphors and is not trained on a specific dataset or text genre.

The arbitrarily selected tweets from both the emotional tweets and the political tweets subsets are used individually as input to the system which highlights the verb-direct object pairs using a parser as potential candidates for metaphor classification. A candidate is classified as a metaphor or not by measuring its semantic similarity to a predefined small seed set of metaphors which acts as an existing known metaphors sample. Metaphor classification is performed based on a previously calculated similarity threshold value. The system labelled around 42% and 48% as metaphorical expressions of the candidates from the emotional tweets subset and the political tweets subset respectively.

3.4 Dataset Compilation

Now that we have two weakly annotated subsets of emotional and political tweets, our approach for selecting the final subset of each category of tweets is driven by achieving the following factors:

1. **Balance:** the dataset should equally represent positive and negative examples.
2. **Verbs Representativeness (Verb Coverage):** the dataset should cover a wide range of verbs and a variety of associated nouns.
3. **Sense Coverage:** ideally each verb should appear at least once in its metaphoric sense and once literally. If the verb does not have one of these senses, more examples should be included. Moreover, unique object arguments of each verb should be represented.
4. **Size:** to ensure usability in a machine learning setting, the dataset should be sizeable.

To ensure verbs representativeness, we employed a set of 5,647 verb-object pairs from the adapted subsets of the MOH dataset (647 verb-direct object pairs) [26] and the VUA corpus (exactly 4,526 verb-direct object pairs) [34]. For each verb in the set¹², all the tweets that contain this verb are extracted without regard to the associated noun (object) argument or the initial metaphoric/literal classification of the weakly supervised system. This step yielded around 3,000 instances from the emotional tweets subset and 38,000 instances from the political tweets subset. For each verb, we randomly selected at least one metaphoric instance and one literal instance using the initial classification by the system to ensure balance, e.g. “*find love*” vs “*find car key*” and “*send help*” vs “*send email*”. Also we ensured the uniqueness of the noun argument associated with each target verb to ensure sense coverage within each subset and across both subsets meaning that the same verb appearing in both subsets has different nouns in order to cover a lot of arguments. Each instance should not exceed five words such as “*send some stupid memory*” or “*abandon a humanitarian approach*”. We observed that the parser more frequently made errors on these longer phrases and thus removing them eliminated many erroneous sentences. Moreover, from preliminary experiments, we realised that the annotators got confused when there are multiple adjectives between the verb and the direct object in a given expression and focused on them instead of the object. Although it might be argued that we could have selected a random set of the tweets but this will not achieve our goal of verb and sense coverage. Moreover, another approach to ensure verb representativeness would have been employing VerbNet [24] but we wanted to be sure that the majority of selected verbs have metaphoric usages.

Our final dataset comprises around 2,500 tweets of which around 1,100 tweets are emotional tweets of general topics and around 1,400 tweets are political tweets related to Brexit. Each tweet has a highlighted verb-object expression that need to be classified according to the metaphoricity of the verb given the accompanying noun (direct object) and the given context. Our next step is to employ crowd-sourcing to manually annotate these expressions. Table 2 shows examples of the different instances appeared in the emotional and political tweets subsets.

4 Annotation Process

4.1 Metaphor Definition

In this work, we adopt the most well-known definition of metaphor which is the conceptual metaphor theory (CMT) introduced by Lakoff and Johnson [17]. Therefore, we view a word

¹²The number of unique verbs (lemma) in this set is 1,134 covering various classes.

■ **Table 2** Examples of the instances appearing in the emotional and political tweets subsets and the corresponding classification of the employed weakly supervised system. *The human annotation disagrees with the system annotation on these examples.

Emotional Tweets	System Classification	Political Tweets	System Classification
accept the fact	metaphor	add financial chaos	not metaphor*
attract hate	metaphor	back #brexit cause	metaphor
break ego	not metaphor*	blame heavy rain	not metaphor
deserves a chance	metaphor*	claim back democracy	metaphor
have time	metaphor	claiming expenses	metaphor*
bring happiness	metaphor	have a say	metaphor
hold phone	not metaphor	hand over britain	not metaphor*
join team	not metaphor	make history	metaphor
win game	not metaphor	write your vote	not metaphor

or an expression as metaphoric if it has at least one basic/literal sense and a secondary metaphoric sense. The literal sense is more concrete and used to perceive a familiar experience while the metaphoric sense is abstract. Moreover, we consider Hank’s [14] view that the metaphoric sense should resonate semantically with the basic sense which means that the metaphorical sense corresponds closely with the literal sense sharing similar semantic features. For example, the metaphoric expression “*launch a campaign*” aligns with (resonates with) more literal, more concrete expressions such as “*launching a boat*”. In this work, we are interested in analysing verb-noun pairs where the verb could be used metaphorically and the noun is a direct object. Research has shown that the majority of metaphoric expressions clusters around verbs and adjectives [25]. We made some distinctions as follows:

Idioms and Similes. We make a distinction between metaphors and other figures of speech that they might be confused with, namely idioms and similes. Idioms such as “*blow the whistle, call the shots, pull the rug out, turn a blind eye, etc.*” were filtered manually.

Verbs with No Metaphorical Potential. We excluded auxiliary and modal verbs from our dataset assuming that they exhibit no potential of being used metaphorically.

Verbs with Weak Metaphorical Potential. In addition to verbs that exhibit strong potential of being metaphors, we are interested in investigating the metaphoricity of light verbs such as “*do, get, give, have, make, take*” and aspectual verbs such as “*begin, end, finish, start, stop*” as well as other verbs such as “*accept, choose, cause, remember, etc.*”. Section 5 presents an analysis of these verbs as they appeared in the proposed dataset. In order to ensure balance, our dataset contains verbs that exhibit both strong and weak metaphorical potential.

4.2 Annotation Task

The annotation task is concerned with the identification of linguistic metaphors in tweets. The main goal is to discern the metaphoricity of a target verb in a highlighted verb-object expression in a given tweet. We set up our annotation task on Amazon Mechanical Turk (MTurk). Five native English speakers were hired to annotate the dataset whose field of study is bachelor of arts with either English, journalism or creative writing.

Task Definition. The annotators were asked to review the tweets and classify the highlighted expression as being used metaphorically or not, based on the provided definition of metaphor and their intuition of the basic sense of the verb.

Guidelines. Each tweet has a highlighted expression of a verb-object (noun) expression. The annotators were instructed to follow a set of guidelines in order to classify the highlighted expression, which are:

1. Read the whole tweet to establish a general understanding of the meaning.
2. Determine the basic meaning of the verb in the highlighted expression. Then, examine the noun (object) accompanying the verb and check whether the basic sense of the verb can be applied to it or not. If it can not, then the verb is probably used metaphorically.
3. Select how certain they are about their answer.

These steps were represented in the task as three questions appearing to the annotators on MTurk as shown in Figure 2.

Reading the whole tweet is important as giving a decision based on reading the highlighted expression only is not enough and leads to inaccurate results. The annotators can skip the tweet if they do not understand it but we set a threshold for skipping tweets. If the annotator is confused about whether an expression is a metaphor or not they were asked to select the “don’t have a clue” option in question 3. The annotators were encouraged to add some notes regarding their confusion or any insights they would like to share. We provided the annotators with several examples to explain the annotation process and to demonstrate the definition of metaphor adopted by this work as well as showing how to discern the basic sense of a verb.

Task Design. We created three annotation tasks on MTurk. The first task is a demo task of 120 tweets from the emotional tweets subset. These tweets included 20 gold tweets with known answers which were obtained by searching the emotional tweets subset for metaphoric expressions (positive examples) from the MOH dataset as well as including some negative examples. This task acted as a training demo to familiarise the annotators with the platform and to measure the understanding of the task. Moreover, it acted as a test for selecting the best performing annotators among all applicants. The efficiency of each applicant is measured in terms of: 1) the time taken to finish the task, 2) the amount of skipped questions and 3) the quality of answers which is measured based on the gold tweets. We selected the top five candidates to proceed with the main tasks. The second task is the annotation of the emotional tweets subset and the third task was devoted to annotating the political tweets subset.

We designed our tasks as pages of 10 tweets each. Pages are referred to as a human intelligence tasks (HITs) by MTurk and annotators were paid per HIT (page). We

the #euref has **demolished my faith** in facts . when both sides have a haul of stats and figures that ' prove ' their side wins what 's the point ?

1. Do you understand the tweet?

Yes
 No

2. Is the highlighted expression used metaphorically?

Yes
 No

3. How certain are you of your answer?

certain
 mostly sure
 unsure
 don't have a clue

■ **Figure 2** A screenshot of the questions in the annotation task given to the annotators on MTurk.

10:12 Crowd-Sourcing A High-Quality Dataset for Metaphor Identification in Tweets

estimated the time taken to annotate around 200 tweets to be one hour; therefore, we paid 60 cents for each page. This comes down to \$12 per hour, which aligns with the minimum wage regulations of the country where the authors resided at the time of this publication.

4.3 Evaluation

Inter-annotator Agreement. The inter-annotator agreement (IAA) evaluation was carried out in terms of Fleiss’ kappa between the five annotators as shown in Table 3. As discussed earlier, we wanted to have a deeper look into light and aspectual verbs, as well as verbs with weak metaphoric potential, so we computed the IAA with and without these verbs for each subset of our dataset. As observed from the results, the annotators were able to achieve a substantial agreement (as for Landis and Koch [18] scale) on the demo task as well as the emotional tweets and the political tweets subsets. After the demo task, the annotators were instructed to pay extra attention to light verbs and to be consistent with similar abstract nouns as much as they can, meaning that “give hope” would often have the same annotation as “*give anxiety/faith*”. To ensure better performance and avoid distraction, we advised the annotators to annotate around 300 tweets per day and resume after reading the instructions again. Since we did not control this rule automatically, we verified that all annotators adhered to this rule by manually checking the time stamps of the annotated tweets.

■ **Table 3** Inter-Annotator Agreement between the five annotators using Fleiss’ kappa. The excluded verbs are light verbs, aspectual verbs, in addition to weak metaphoric potential verbs including “accept, choose, enjoy, imagine, know, love, need, remember, require, want”.

	partial exclusion (keep light verbs)	Fleiss’ kappa full exclusion	no exclusion
Demo Task (120 tweets)	0.627 (106 annotated instances)	0.715 (85 annotated instances)	0.623 (108 annotated instances)
Emotional Tweets (1,070 tweets)	0.742 (884 annotated instances)	0.732 (738 annotated instances)	0.701 (1,054 annotated instances)
Political Tweets (1,391 tweets)	0.806 (1,341 annotated instances)	0.805 (1,328 annotated instances)	0.802 (1,389 annotated instances)

Points of (Dis-)agreement. Tables 4 and 5 lists examples of the agreements and disagreements between the five annotators. The majority of disagreements centred around light verbs and verbs with weak metaphoric potential. The next section discusses the annotation results in detail and presents the statistics of the dataset.

5 Dataset Statistics and Linguistic Analysis

5.1 Statistics

The statistics of each subset of the dataset is presented in Table 6 focusing on different statistical aspects of our dataset. It is worth mentioning that the political tweets subset contains 431 more unique verbs that did not appear in the emotional tweets subset. The text of the political tweets was more understandable and structured. The emotional tweets subset contains some tweets about movies and games that sometimes the annotators found hard to understand.

■ **Table 4** Examples of agreements among the five annotators (100% majority vote).

		majority vote
Emotional Tweets	its great to be happy, but its even better to bring happiness to others.	metaphor
	make memories you will look back and smile at.	
	as long as the left stays so ugly, bitter, annoying & unlikeable, they will not win any elections...	not metaphor
Political Tweets	they forget this when they have money and start tweeting like they have all the answers	
	make or break moment today! together we are stronger! vote remain #strongerin #euref	metaphor
	...cameron can not win this #euref without your support. how many will lend their support to...	
	person's details taken by police for offering to lend a pen to voters, what a joke.	not metaphor
	in just a couple of days, no one will ever have to utter the word 'brexit' ever again	

■ **Table 5** Examples of disagreements among the five annotators (60% majority vote).

		majority vote
Emotional Tweets	someone should make a brand based off of triangle noodles that glow in the dark. call it illuminoodle...	metaphor
	smile for the camera, billy o. if you need a smile every day then #adoptadonkey @donkeysanctuary	
	cities are full of mundane spaces. imagine the potential to transform them into catalysts for positive emotions	not metaphor
Political Tweets	our captors are treating us well and we are very happy and well enjoying their kind hospitality	
	perhaps we can achieve a cohesive society when the referendum is over, but it does not feel like that is possible. #euref	metaphor
	#euref conspiracy theories predict people's voting intentions . will they sway today's vote?	
	democracy works there's still time. british people can not be bullied do not believe the fear #voteleave	not metaphor
	what's interesting here is not the figure but that it was from an online poll which has always favoured the leave .	

■ **Table 6** Statistics of the proposed dataset of tweets.

	Demo Task	Emotional Tweets	Political Tweets
# of tweets	120	1,070	1,390
# of unique verb-direct object (noun) pairs	119	1,069	1,390
Average tweet length	23.82	22.14	21.12
# of unique verbs (lemma) (in the annotated verb-noun pairs)	71	321	676
# of unique nouns (in the annotated verb-noun pairs)	102	725	706
% instances annotated as metaphors	63.15%	50.47%	58.16%
% instances annotated as not metaphors	36.84%	49.54%	41.84%
% instances annotated with agreement majority vote of 60%	20.17%	10.39%	12.29%
# of non-understandable tweets (skipped)	5.2%	1.68%	0.14%

5.2 Linguistic Analysis

As observed from the IAA values listed in Table 3, light and aspectual verbs as well as some other verbs represent a major source of confusion among the annotators. Although other researchers pointed out that they exhibit low potential of being metaphors and excluded them from their dataset, our dataset covers different examples of these verbs with different senses/nouns. The majority vote of the annotators on such cases could give us some insight on the cases where these verbs can exhibit metaphorical sense.

In the following paragraphs, we provide a linguistic analysis of the proposed dataset performed by manual inspection. The majority of annotators tend to agree that the verb “have” exhibits a metaphoric sense when it comes with abstract nouns such as “*anxiety, hope, support*” as well as other arguments including “*meeting, question, theory, time, skill, vote*”.

On the other hand, it is used literally with nouns such as “*clothes, friend, illness, license, money*”. The annotators find the light verbs “*get, give, and take*” to be more straightforward while discerning their metaphoric and literal usages. They agreed on their metaphorical usage with abstract nouns such as “*chance, happiness, smile, time, victory*” and their literal usage with tangible concepts including “*food, job, medication, money, notification, results*”. Regarding the verb “*make*” the annotators agreed that as long as the accompanied noun cannot be *crafted* then it is used metaphorically. Metaphors with this verb include “*difference, friends, money, progress, time*”, while literal ones include “*breakfast, mistake, movie, noise, plan*”.

The nouns occurring with the verb “*start*” in metaphoric expressions include “*bank, brand, friendship*”. Moreover, there are some rhetorical expressions such as “*start your leadership journey/living/new begining*”. The nouns appearing in the expressions classified as literal include “*argument, car, course, conversation, petition*”. The verb “*end*” occurred with “*horror, feud*” metaphorically and “*thread, contract*” literally according to the majority vote.

The annotators agreed that nouns such as “*food, hospitality, life, music*” occurring with the verb “*enjoy*” form literal expressions while the only metaphoric instance is “*enjoy immunity*”. In the case of the verb “*love*”, the majority of annotators agreed that it is not used metaphorically as you can love/hate anything with no metaphorical mapping between concepts. The disagreements revolve around the cases when the expression is an exaggeration or a hyperbole e.g. “*love this idea/fact/book*”. Expressions have stative verbs of thought such as “*remember and imagine*” are classified as non-metaphoric. The only debate was about the expression “*...remember that time when...*” as, according to the annotators, it is a well-known phrase (fixed expression). We looked at the verbs “*find and lose*” and they were easy to annotate following the mapping between abstract and concrete senses. They are classified as metaphors with abstract nouns such as “*love, opportunity, support*” as well as something virtual such as “*lose a seat (in the parliament)*”. However, it was not the case for the verb “*win*”. The majority agreed that expressions such as “*win award/election/game*” are literal expressions while the only disagreement was on the expression “*win a battle*” (3 annotators agreed that it is used metaphorically).

Annotating the verbs “*accept, reject*” was intriguing as well. The majority of annotators classified the instances “*accept the fact/prices*” as literal while in their view “*accept your past*” is a metaphor. An issue is raised regarding annotating expressions that contain the verbs “*cause, blame, need, want*”. Most agreed that “*need apology/job/life*” can be considered as metaphor while “*need date/service*” is not.

From this analysis, we conclude that following the adopted definition of metaphor helped the annotators to discern the sense of these verbs. Relying on the annotators’ intuition (guided by the given instructions) to decide the basic meaning of the verb led to some disagreements but it was more time and effort efficient than other options. Light verbs are often used metaphorically with abstract nouns. There are some verbs exhibiting weak metaphoric potential and classifying them is not as straightforward as other verbs. However, they might be used metaphorically on occasions, but larger data is required to discern these cases and find a solid pattern to define their metaphoricity. Hyperbole and exaggerations and other statements that is not meant to be taken literally need further analysis to discern its metaphoricity. Sharing and discussing the disagreements after each annotation task among the annotators helped to have a better understanding of the task.

6 Conclusion

This work proposes an annotation methodology to create a high-quality dataset of tweets annotated for metaphor using crowd-sourcing. Our approach is driven by achieving balance, sense coverage and verbs representativeness as well as high accuracy. We were able to introduce a better quality annotation of metaphors in spite of the conventionality of metaphors in our everyday language compounded by the challenging context of tweets. The employed approach resulted in a dataset of around 2,500 tweets annotated for metaphor achieving a substantial inter-annotator agreement despite the difficulty of defining metaphor. Although, we focused on annotating verb-direct object pairs of linguistic metaphors in tweets, this approach can be applied to any text type or level of metaphor analysis. The annotation methodology relies on an existing metaphor identification system to facilitate the recognition and selection of the annotated instances by initially creating a weakly annotated dataset. This system could be substituted by any other model to suit the type of targeted metaphors in order to reduce the cognitive load on the annotators and maintain consistency. Our dataset consists of various topic genres focusing on tweets of general topics and political tweets related to Brexit. The dataset will be publicly available to facilitate research on metaphor processing in tweets.

We are planning to use this dataset to create a larger dataset of tweets annotated for metaphors using semi-supervised methods. Additionally, an in-depth qualitative and quantitative analysis will be carried out to follow up on the conclusions that have been drawn in this work. Furthermore, we are interested in having a closer look at the metaphors related to Brexit on Twitter. We are also interested in investigating the metaphoric sense of verbal multi-word expressions (MWEs) by looking into the dataset released as part of the PARSEME shared-task [33].

References

- 1 Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 329–336, Trento, Italy, April 2006.
- 2 Lou Burnard. About the British National Corpus, 2009. URL: <http://www.natcorp.ox.ac.uk/corpus/index.xml>.
- 3 Jonathan Charteris-Black. Metaphor in Political Discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. Palgrave Macmillan UK, London, 2011.
- 4 Danqi Chen and Christopher Manning. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 740–750, Doha, Qatar, October 2014.
- 5 Mark Davies. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190, 2009.
- 6 Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1412–1424, Brussels, Belgium, November 2018.
- 7 David Evans. Compiling a corpus. *Corpus building and investigation for the Humanities*, 2007 (accessed December 23, 2018). URL: <https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit2.pdf>.
- 8 Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- 9 Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

- 10 W. Nelson Francis and Henry Kucera. The Brown Corpus: A Standard Corpus of Present-Day Edited American English. Technical report, Brown University Linguistics Department, 1979.
- 11 Dedre Gentner, Brian Bowdle, Phillip Wolff, and Consuelo Boronat. Metaphor Is Like Analogy. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors, *The analogical mind: Perspectives from cognitive science*, pages 199–253. The MIT Press, Cambridge, MA, USA, 2001.
- 12 Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 470–478, Denver, CO, USA, June 2015.
- 13 Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(6):1–25, July 2017.
- 14 Patrick Hanks. Three Kinds of Semantic Resonance. In *Proceedings of the 17th EURALEX International Congress*, pages 37–48, Tbilisi, Georgia, September 2016.
- 15 Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. Identifying Metaphorical Word Use with Tree Kernels. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 52–56, Atlanta, GA, USA, June 2013.
- 16 Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '15, pages 384–392, Prague, Czech Republic, September 2015.
- 17 George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago Press, Chicago, USA, 1980.
- 18 J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.
- 19 Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, SemEval '18, pages 1–17, New Orleans, LA, USA, June 2018.
- 20 Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, *Sem '16, pages 23–33, Berlin, Germany, 2016.
- 21 Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. Semantic Signatures for Example-Based Linguistic Metaphor Detection. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA, USA, June 2013.
- 22 Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. Introducing the LCC Metaphor Datasets. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC '16, pages 4221–4227, Portorož, Slovenia, May 2016.
- 23 Natalie Parde and Rodney Nielsen. A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC '18, pages 1535–1540, Miyazaki, Japan, May 2018.
- 24 Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 2006.
- 25 Ekaterina Shutova. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623, December 2015.
- 26 Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 160–170, San Diego, CA, USA, June 2016.
- 27 Ekaterina Shutova and Simone Teufel. Metaphor Corpus Annotated for Source-Target Domain Mappings. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10, pages 255–261, Malta, May 2010.

- 28 Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353, June 2013.
- 29 S. Siegel and N. Castellan. *Nonparametric statistics for the behavioral sciences*. Mc Graw-Hill, 1988.
- 30 Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company, 2010.
- 31 Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 248–258, Baltimore, MD, USA, June 2014.
- 32 Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Edinburgh, Scotland, UK, July 2011.
- 33 Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. Constructing an Annotated Corpus of Verbal MWEs for English. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, LAW-MWE-CxG-2018, pages 193–200, Santa Fe, NM, USA, August 2018.
- 34 Omnia Zayed, John Philip McCrae, and Paul Buitelaar. Phrase-Level Metaphor Identification using Distributed Representations of Word Meaning. In *Proceedings of the Workshop on Figurative Language Processing*, pages 81–90, New Orleans, LA, USA, June 2018.