Natural Language Engineering (2024), **30**, pp. 256–293
doi:10.1017/S1351324923000153

ARTICLE

Determining sentiment views of verbal multiword expressions using linguistic features

Michael Wiegand[1,2], Marc Schulder[2,3] and Josef Ruppenhofer[4]

[1]Digital Age Research Center, Alpen-Adria-Universität Klagenfurt, Klagenfurt am Wörthersee, Austria, [2]Sprach- & Signalverarbeitung, Universität des Saarlandes, Saarbrücken, Germany, [3]Institut für Deutsche Gebärdensprache, Hamburg, Germany, and [4]Leibniz-Institut für Deutsche Sprache, Mannheim, Germany
Corresponding author: M. Wiegand; Email: michael.wiegand@aau.at

(Received 5 September 2021; revised 11 March 2023; accepted 17 April 2023; first published online 15 May 2023)

Abstract

We examine the binary classification of sentiment views for verbal multiword expressions (MWEs). Sentiment views denote the perspective of the holder of some opinion. We distinguish between MWEs conveying the view of the speaker of the utterance (e.g., in "*The company* ***reinvented the wheel***" the holder is the implicit speaker who criticizes *the company* for creating something already existing) and MWEs conveying the view of explicit entities participating in an opinion event (e.g., in "*Peter* ***threw in the towel***" the holder is *Peter* having given up something). The task has so far been examined on unigram opinion words. Since many features found effective for unigrams are not usable for MWEs, we propose novel ones taking into account the internal structure of MWEs, a unigram sentiment-view lexicon and various information from Wiktionary. We also examine distributional methods and show that the corpus on which a representation is induced has a notable impact on the classification. We perform an extrinsic evaluation in the task of opinion holder extraction and show that the learnt knowledge also improves a state-of-the-art classifier trained on BERT. Sentiment-view classification is typically framed as a task in which only little labeled training data are available. As in the case of unigrams, we show that for MWEs a feature-based approach beats state-of-the-art generic methods.

Keywords: Sentiment analysis; Opinion mining; Lexical semantics; Opinion holder extraction; Multiword expressions

1. Introduction

While there has been much research in sentiment analysis on the tasks of subjectivity detection and polarity classification, there has been less work on other types of categorizations that can be imposed upon subjective expressions. In this article, we focus on the views that an opinion expression evokes. We refer to them as *sentiment views*. We study this phenomenon on English language data.

1.1. The concept of sentiment views

Sentiment views are tuples of entities where the first represents the holder of an opinion and the second the target of the opinion. An opinion expression evokes at least one view. We distinguish two types of views, according to where the holder of the opinion is. If the holder is a participant in the event denoted by the opinion expression, we say the expression bears an **actor view**. Otherwise, the holder is a sentient entity to whom the opinion expression is attributed as their speech or thought. We refer to this as **speaker view**.

We illustrate the distinction we have in mind considering the following four examples:

(1) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject.

(2) The government **wasted**$_{\textit{speaker-view}}$ a lot of money.

(3) Party members were **disappointed**$_{\textit{actor-view}}$ by the election outcome.

(4) All representatives **praised**$_{\textit{actor-view}}$ the final agreement.

The verb *excelled* in (1) expresses the view of the speaker producing the sentence. The speaker thereby evaluates *Sarah*'s educational performance positively. While the speaker's positive assessment of *Sarah* is lexically expressed by the opinion expression and not defeasible (1a), the sentence does not reliably convey what *Sarah*'s sentiment is towards *virtually every subject* or her own performance (1b)–(1d). Thus, there is no actor view attached to *excelled*.

(1a) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject. *She did badly.

(1b) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject even though she disliked all of them.

(1c) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject because she is really interested in all of them.

(1d) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject even though she didn't think so.

(1e) Sarah **excelled**$_{\textit{speaker-view}}$ in virtually every subject and she thought so herself.

In the following, we explain for each opinion expression in (2)–(4) the respective sentiment view that it conveys.

In (2), the implicit speaker of the utterance evaluates the government's spending policy. Therefore, *wasted* conveys a speaker view. While the speaker has a negative sentiment towards the government, the reader cannot tell what the government's sentiment is towards *a lot of money*. It may be positive but it could also be negative or neutral. For instance, the government could regard money as a means to achieve something. However, they may not specifically have a particularly positive sentiment towards the money itself. Since in (2), *the government* is the only entity participating in the event evoked by *wasted* that is also eligible to be an opinion holder, and given that we cannot infer it having a specific sentiment towards *a lot of money*, *wasted* does not convey an actor view.

In (3), the situation is different. *Party members* is some entity participating in the event evoked by *disappointed,* and it has a negative sentiment towards *the election outcome* since if something disappoints someone, then one has typically a negative sentiment towards it. Therefore, *disappointed* conveys an actor view. On the other hand, we cannot infer any sentiment of the implicit speaker towards any of the entities participating in the event evoked by *disappointed*, that is *party members* and *the election outcome*. Of course, the implicit speaker may have some specific sentiment to either of these entities or both, but this is not conveyed by the sentence and would have to be established by the wider context in which the sentence is embedded. Therefore, we can conclude that *disappointed* does not convey a speaker view.

The situation is similar in (4) where *all representatives* is some entity participating in the event evoked by *praised*. If someone praises something, then we can infer that they also have a positive sentiment towards it. So, in (4), *all representatives* have some positive sentiment towards *the final agreement*. Therefore, the given opinion expression *praised* conveys an actor view. At the same time, similar to (3), we cannot infer any obvious sentiment of the implicit speaker towards any of the entities participating in the event evoked by *praised*, that is *all representatives* and *the final agreement*. Therefore, *praised* does not convey a speaker view.

As the previous examples showed, sentiment-view classification is considered a ***binary* classification**. That is, either an opinion expression conveys an actor view or it conveys a speaker view.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Actor views are largely defined as opinion words whose opinion holder is a (syntactic) dependent of the opinion expression, typically its agent or patient. As a consequence, we would still refer to *praised* from (4) as an actor-view word if we replaced *all representatives* by a first-person pronoun (5). That explicit holder may refer to the speaker but the definition of actor views, namely that there is an explicit opinion holder which is also a dependent of the opinion expression,[a] still holds. On the other hand, in (6) and (7), *wasted* is still categorized as a speaker-view word despite the presence of the opinion holder *the opposition*. However, that noun phrase is not a syntactic dependent of *wasted*. Such types of holders are also referred to as *nested sources* (Wiebe, Wilson, and Cardie 2005).

(5) [I]$^{\mathit{agent}}_{\mathit{Holder}}$ **praised**$_{\mathit{actor\text{-}view}}$ the final agreement.

(6) [The opposition]$_{\mathit{Holder}}$ criticized that the government **wasted**$_{\mathit{speaker\text{-}view}}$ a lot of money.

(7) According to [the opposition]$_{\mathit{Holder}}$, the government **wasted**$_{\mathit{speaker\text{-}view}}$ a lot of money.

Strictly speaking, most actor views are biased by the author's beliefs and opinions on the entities whose (actor) view is described. This is so since the sentiment of the actor is often not directly observable but a matter of interpretation. This means that the author needs to derive someone's sentiment from their actions unless that person verbally expresses their sentiment towards something or someone else explicitly. Therefore, the term *actor view* is actually not fully informative. Still, we do not introduce a new term (e.g., *assumed actor view*) in this article, in order to be consistent with previous work, particularly Wiegand *et al.* (2016) upon whose results our work is largely built.

In this work, we use a fairly wide notion of the concept *opinion expression*. Following the annotation scheme of the MPQA corpus (Wiebe *et al.* 2005), that is the English reference corpus for fine-grained sentiment analysis, we subsume all types of *privates states* by that term, that is, not only sentiments or evaluations but also all possible forms of mental and emotional states (Quirk *et al.* 1985). Therefore, emotion-evoking words (8)-(14) also fall under our definition of opinion expression.

(8) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **happy**$_{\mathit{actor\text{-}view}}$.

(9) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **sad**$_{\mathit{actor\text{-}view}}$.

(10) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **relaxed**$_{\mathit{actor\text{-}view}}$.

(11) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **exhausted**$_{\mathit{actor\text{-}view}}$.

(12) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **angry**$_{\mathit{actor\text{-}view}}$.

(13) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **devastated**$_{\mathit{actor\text{-}view}}$.

(14) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **surprised**$_{\mathit{actor\text{-}view}}$.

Such opinion expressions typically evoke actor views. For instance, in (8)–(14), the emotional state originates from the respective agent, that is *Mary*. In a strict sense, that agent has no sentiment towards another entity. However, we could extend all sentences by some patient so that the agent has some sentiment towards it (15)–(21). On the other hand, in none of these sentences can we infer any sentiment of the speaker towards *Mary*.

(15) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **happy**$_{\mathit{actor\text{-}view}}$ about her family's encouragement.

(16) [Mary]$^{\mathit{agent}}_{\mathit{Holder}}$ is **sad**$_{\mathit{actor\text{-}view}}$ about the passing of her cat.

[a] In (5), the first-person pronoun *I* is the agent of *praised*.

(17) [Mary]$^{agent}_{Holder}$ is **relaxed**$_{actor\text{-}view}$ about the additional tasks she has been assigned to.

(18) [Mary]$^{agent}_{Holder}$ is **exhausted**$_{actor\text{-}view}$ by the long journey.

(19) [Mary]$^{agent}_{Holder}$ is **angry**$_{actor\text{-}view}$ about Peter's constant opposition.

(20) [Mary]$^{agent}_{Holder}$ is **devastated**$_{actor\text{-}view}$ by the court's strict sentence.

(21) [Mary]$^{agent}_{Holder}$ is **surprised**$_{actor\text{-}view}$ by that sudden job offer.

1.2. Sentiment views of verbal multiword expressions

So far, sentiment views have only been examined for opinion words which are unigrams. However, **multiword expressions (MWEs)**, particularly verbal MWEs, can similarly either convey speaker views (22)–(23) or actor views (24)–(25).

(22) Minecraft is a game that always **keeps up with the times**$_{speaker\text{-}view}$.

(23) His latest remarks only **added fuel to the fire**$_{speaker\text{-}view}$.

(24) Trump **draws the line**$_{actor\text{-}view}$ at gay marriage and abortion.

(25) Russia **sees eye to eye**$_{actor\text{-}view}$ with the coalition on Syria airstrike targets.

In (22) and (23), the implicit speaker evaluates the agent, that is *Minecraft* in (22) and *his latest remarks* in (23). Therefore, in both sentences the given MWEs evoke a speaker view. Neither *Minecraft* nor *his latest remarks* are entities that are opinion holders. Therefore, since no other entities are evoked by the respective MWEs, no actor view is conveyed. In (24) and (25), the respective agents of the given MWEs have some sentiment towards the patient of the given MWEs, that is *gay marriage and abortion* in (24) and *the coalition* in (25). [(25) also conveys that *the coalition* has the same sentiment as *Russia*.] Therefore, actor views are evoked in these sentences by the MWEs. In neither of the sentences can we infer any sentiment of the implicit speaker towards any of the entities participating in the event evoked by the respective MWEs, that is *Trump* and *gay marriage and abortion* in (23), on the one hand, and *Russia* and *the coalition* in (25), on the other hand. Both MWEs, that is *draws the line* and *sees eye to eye*, can be considered metaphors since Trump is not literally drawing any lines nor do members of the Russian government and the coalition physically stand opposite each other and look members of the other group in the eyes. It is the authors of those two respective sentences who chose to make use of this form of figurative language in order to achieve a certain effect on the readers. For instance, the sentences may become more vivid or expressive than non- or less figurative language. Still, such stylistic devices should not be mistaken to be reliable clues for a speaker view. The authors may want to reach a certain effect on the readers, and, as already discussed in Section 1.1, in many situations they need to derive the actors' views from some other observable actions. However, neither of these indicates that the authors themselves have any sentiment towards those entities. Thus, no speaker view is conveyed by the respective MWEs in (24) and (25).

Sentiment-view classification has been shown to be effective for other tasks in sentiment analysis, particularly opinion role extraction (Wiegand and Ruppenhofer 2015; Deng and Wiebe 2016). In opinion role extraction, the task is to extract the strings from a sentence which represent the opinion holder and the opinion target of a particular opinion expression. (26) and (27) represent sentences with identical structures but the agent noun phrase *the committee* conveys different opinion roles.

(26) [The committee]$^{agent}_{Holder}$ **pulled the plug**$_{actor\text{-}view}$.

(27) [The committee]$^{agent}_{Target}$ **delivered the goods**$_{speaker\text{-}view}$. *(Holder: implicit speaker)*

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

In (26), *the committee* made an assessment about something, for example some funding, and concluded it had to prevent it from continuing. In other words, the committee had a negative sentiment towards this implicit patient of *pulled the plug*.[b] It pulled the plug. Therefore, in this sentence *the committee* is a holder. In (27), on the other hand, the implicit speaker of that utterance made some clearly positive evaluation about the committee's actions. It did something it was expected or obliged to do. It delivered the goods. Therefore, in this sentence *the committee* is a target. Only the knowledge of sentiment views helps us to assign opinion roles correctly in (26) and (27). In (27), the opinion expression *delivered the goods* conveys a speaker view (i.e., it denotes a positive evaluation made by the speaker). Such expressions possess an implicit opinion holder, typically the speaker of the utterance. As a consequence, unlike (26), where the opinion expression *pulled the plug* conveys an actor view,[c] *the committee* in (27) can only represent an opinion target.

In this article, we explore features for the automatic classification of sentiment views of MWEs. The task is to determine the sentiment view of a given set of MWEs out of context (*binary classification*). This lexeme-level classification of MWEs is even more pressing than the classification of unigram opinion words. The latter can also be largely learnt from a text corpus labeled with sentiment-view information, since a great proportion of opinion words occurs in such corpora. For instance, Johansson and Moschitti (2013) demonstrate this on the MPQA corpus (Wiebe *et al.* 2005). However, MWEs occur less frequently. In the MPQA corpus, only 5% of a large list of MWEs[d] can be found, while for the opinion words of the Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005), which are exclusively unigrams, 51% are included. Even though they are less frequent than unigrams, MWEs occur *regularly*. Jackendoff (1997) even argues that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. Since this is an estimate for all types of MWEs, we inspected a random sample of 1000 sentences drawn from the North American News Text Corpus (LDC95T21) with respect to the type of MWEs we consider in this work (i.e., lexicalized verbal MWEs). Thus, we hope to offer a more precise quantitative estimate. On that sample, we identified 166 verbal MWEs. Therefore, on average more than 15% of the sentences of that corpus contain a verbal MWE. However, of the 157 unique MWEs of our sample, only 8 MWEs occur more than once. Consequently, in a typical text corpus we have to expect the vast majority of MWEs to be singletons. This makes it very difficult for traditional context-based classifiers to learn information for those individual MWEs from labeled text corpora. Still, MWEs, particularly verbal MWEs, are very relevant to sentiment analysis, as a large proportion conveys subjective information. In our above sample, 154 of the 166 MWE mentions were considered to convey a subjective context. This amounts to more than 90%. We also examined a random sample of 1000 sentences of the two other corpora we consider in this work (Section 3.4), that is the corpus from Jindal and Liu (2008) and UKWAC (Baroni *et al.* 2009). Here, too, subjective MWEs constitute a clear majority of all MWEs (namely 84% in the sample of the corpus from Jindal and Liu (2008) and 79% in the sample of UKWAC).

Many features found effective for the detection of sentiment views on unigrams (Wiegand *et al.* 2016) are less effective for the detection on MWEs. The reason for this is that in common lexical resources, such as WordNet, FrameNet and subcategorization lexicons, MWEs are only sparsely represented. This justifies tackling MWEs as a separate research question in this article.

The **contributions** of our article are the following: Apart from introducing a new gold standard for this novel task (Section 3.2) and adjusting the features proposed for unigram words to MWEs (Sections 5.1.1–5.1.5), we address several new directions for the analysis of MWEs

[b] (26) has two entities participating in the event evoked by *pulled the plug*. However, only the agent, that is, *the committee* is explicitly realized. The patient is implicit. The sentence could be modified to include an explicit patient: $[\textit{The committee}]^{\textit{agent}}_{\textit{Holder}}$ $\textbf{\textit{pulled the plug}}_{\textit{actor-view}}$ $[\textit{on the funding}]^{\textit{patient}}_{\textit{Target}}$.

[c] That MWE does not convey a speaker view. Though it may be possible that the implicit speaker of the utterance also has some opinion on the committee's decision, the sentence does not specify the sentiment of that opinion.

[d] This statistic is based on the union of MWEs found in all lexical resources taken into consideration in this work.

(Sections 5.1.6–5.1.8). Firstly, we analyze the internal structure of an MWE itself and show that light-verb constructions, on the one hand, and idiomatic MWEs, on the other hand, have clear tendencies towards different sentiment views (Section 5.1.6). Secondly, the semantic similarity of MWEs to unigram parts of speech is found to be predictive (Section 5.1.7). Thirdly, we investigate to what extent a unigram sentiment-view lexicon helps to determine the sentiment views of MWEs (Section 5.1.8). Fourthly, for many types of features that we consider for lexical acquisition related to MWEs, we also explore variants that use information from a lexical resource that has a much wider coverage of MWEs than WordNet or FrameNet. We investigate what information of that resource can be effectively used. Finally, we extrinsically evaluate whether information about sentiment views can improve the task of opinion role extraction (Section 6). In this context, we will demonstrate that extraction systems trained on corpora labeled with fine-grained sentiment information (i.e., MPQA corpus) miss opinion holders that a classifier based on sentiment views is able to detect.

Our work deals with a semantic categorization problem which so far has only been studied for unigrams but is now extended to MWEs. Our insights may be relevant in light of recent interest in MWE analysis as reflected by the SemEval Shared Task on *Detecting Minimal Semantic Units and their Meanings (DiMSUM)* (Schneider *et al.* 2016). Exploring different kinds of features—several of which can be seen as standard features for term categorization—we provide detailed analysis of which ones work and which ones do not, proposing alternatives where possible.

1.3. Outline

The remainder of this article is structured as follows. Section 2 discusses related work, while Section 3 describes the data and annotation we use in our experiments. In the following two sections, we explore two different strategies for the automatic classification of sentiment views of MWEs: In Section 4, we investigate in how far a classifier without access to labeled MWEs can be designed with the help of a unigram sentiment-view lexicon. Through graph-based label propagation, we explore how sentiment views of MWEs can be best inferred from unigram opinion words. In Section 5, we present the second strategy. We examine features for supervised sentiment-view classification of MWEs using a small fraction of labeled MWEs. In Section 6, we evaluate the usefulness of sentiment views in the task of opinion role extraction. Section 7 concludes this article.

2. Related work

In this section, we situate our research in the context of prior related work. On the one hand, we discuss previous research on sentiment views (Section 2.1), that is the subtask in sentiment analysis we extend for MWEs in this article. On the other hand, we also review previous research on sentiment analysis for MWEs (Section 2.2). While MWEs have not yet been examined with respect to sentiment views, there exists research on MWEs related to other aspects of sentiment analysis such as polarity. Finally, we also discuss the relevance of metaphor (Section 2.3), a linguistic phenomenon which can be often observed with MWEs, depending on whether an MWE represents a metaphor or not may have an impact on its sentiment view.

2.1. Previous work on sentiment views

The annotation scheme of the MPQA corpus (Wiebe *et al.* 2005) was the first work to include the distinction between different sentiment views. The two sentiment views are referred to as *direct subjectivity* (=actor view) and *expressive subjectivity* (=speaker view). In subsequent research, some approaches were proposed to distinguish these two categories in the MPQA corpus. The most extensive works are Breck *et al.* (2007) and Johansson and Moschitti (2013). Since MPQA

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

provides annotation regarding sentiment in context, sentiment views are exclusively considered in contextual classification. The fact that it is the opinion words that convey those views, as we do in this article, is insufficiently addressed. Johansson and Moschitti (2013) focus on optimizing a machine-learning classifier, in particular to model the interaction between different subjective phrases within the same sentence. Breck *et al.* (2007) address feature engineering and partly acknowledge that sentiment views are a lexical property by deriving features from lexical resources, such as WordNet (Miller *et al.* 1990), the verb categories from Levin (1993), and FrameNet (Baker, Fillmore and Lowe 1998).

Maks and Vossen (2012b) link sentiment views to opinion words as part of a lexicon model for sentiment analysis. Maks and Vossen (2012a) also examine a corpus-driven method to induce opinion words for the different sentiment views. The authors, however, conclude that their approach, which sees news articles as a source for actor views and news comments as a source for speaker views, is not sufficiently effective.

The works most closely related to ours are Wiegand *et al.* (2016), Deng and Wiebe (2016), and Wiegand and Ruppenhofer (2015) who all successfully distinguish sentiment views on the lexeme level out of context:

Wiegand *et al.* (2016) take into account the opinion adjectives, nouns, and verbs from the Subjectivity Lexicon (Wilson *et al.* 2005). As a gold standard, these words are manually annotated with sentiment-view information. This set of opinion words *exclusively* comprises unigrams. Various types of features, both syntactic and semantic, are examined for automatic classification. These features present a baseline for our work. We will discuss them and the resources that they are derived from in Section 5.1.

Deng and Wiebe (2016) do not employ a manual gold standard of sentiment views but heuristically derive the sentiment view of opinion words from the context-level annotation of the MPQA corpus. For all those opinion expressions, embeddings are induced. The classifier to categorize expressions as actor and speaker views is solely trained on these embeddings. No further features are considered. The knowledge of sentiment views is then incorporated into a classifier as a feature to extract opinion holders from the MPQA corpus. Deng and Wiebe (2016) employ a different terminology. Actor views are referred to as *participant opinions* whereas speaker views are referred to as *non-participant opinions*. Since that study was conducted on the MPQA corpus, this approach allows no conclusions to be made about verbal MWEs, since only very few verbal MWEs are contained in the MPQA corpus (see Section 1).

Wiegand and Ruppenhofer (2015) examine sentiment views exclusively on opinion verbs using graph-based label propagation. They distinguish between two types of actor views, namely agent views and patient views. The former take their opinion holder as an agent and their target as a patient (28)–(29), while the latter align their roles inversely (30)–(31).

(28) $[\text{Peter}]^{agent}_{Holder}$ $\textbf{loves}_{agent\text{-}view}$ $[\text{Mary}]^{patient}_{Target}$.

(29) $[\text{Peter}]^{agent}_{Holder}$ $\textbf{criticizes}_{agent\text{-}view}$ $[\text{Mary}]^{patient}_{Target}$.

(30) $[\text{Mary}]^{agent}_{Target}$ $\textbf{pleases}_{patient\text{-}view}$ $[\text{Peter}]^{patient}_{Holder}$.

(31) $[\text{Mary}]^{agent}_{Target}$ $\textbf{disappoints}_{patient\text{-}view}$ $[\text{Peter}]^{patient}_{Holder}$.

This distinction between different subtypes of actor views does not exist among nouns or adjectives as illustrated by (32)–(37). The opinion holders and targets of opinion nouns (33) & (36) and opinion adjectives (34) & (37) typically align to the same argument positions. Consequently, opinion nouns and opinion adjectives are only categorized into actor views and speaker views.

(32) $[\text{Peter}]^{agent}_{Holder}$ $\textbf{criticizes}^{verb}_{agent\text{-}view}$ $[\text{Mary}]^{patient}_{Target}$.

(33) $[\text{Peter's}]^{agent}_{Holder}$ $\textbf{criticism}^{noun}_{actor\text{-}view}$ $[\text{of Mary}]^{patient}_{Target}$ was immense.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

(34) [Peter]$^{agent}_{Holder}$ is **critical**$^{adj}_{actor\text{-}view}$ [of Mary]$^{patient}_{Target}$.

(35) [Mary]$^{agent}_{Target}$ **surprises**$^{verb}_{patient\text{-}view}$ [Peter]$^{patient}_{Holder}$.

(36) [Peter's]$^{agent}_{Holder}$ **surprise**$^{noun}_{actor\text{-}view}$ [over Mary]$^{patient}_{Target}$ was immense.

(37) [Peter]$^{agent}_{Holder}$ is **surprised**$^{adj}_{actor\text{-}view}$.

Despite all those previous research efforts on sentiment views, so far MWEs have not been explicitly addressed for this classification task.

2.2. Previous work on sentiment analysis for multiword expressions

Previous work in the area of MWEs, in general, has focused on methods for the automatic *detection* of MWEs (Hashimoto and Kawahara 2008; Tsvetkov and Wintner 2011; Constant, Sigogne and Watrin 2012; Green, de Marneffe, and Manning 2013; Schneider *et al.* 2014a; Constant *et al.* 2017). Our work is rather different in that we consider a set of given MWEs and try to categorize them. Categorization tasks are also the predominant tasks in sentiment analysis which is the subject whose related work we discuss in this subsection.

There has been significant work on computing the sentiment of phrases (Moilanen and Pulman 2007; Liu and Seneff 2009; Socher *et al.* 2013). However, only arbitrary sequences of tokens in sentences are considered as phrases rather than specific *lexicalized* phrases such as MWEs.

Some methods used in sentiment analysis work equally well for unigrams and MWEs. Graph-based label propagation, such as the one proposed by Velikovich *et al.* (2010) for polarity classification, is a prime example. We will take such a type of classifier into account with our graph-based baseline (Section 4) which bears a great resemblance to the approach of Velikovich *et al.* (2010).

The only works in sentiment analysis that specifically address MWEs are Moreno-Ortiz *et al.* (2013); Beigman Klebanov *et al.* (2013); Williams *et al.* (2015) and Jochim *et al.* (2018). Moreno-Ortiz *et al.* (2013) report on the manual annotation of a Spanish polarity lexicon exclusively comprising MWEs. Beigman Klebanov *et al.* (2013) present an elicitation study on the polarity of noun-noun compounds. They find that polarity information is highly compositional. They also represent the polarity of noun-noun compounds in sentiment profiles and show that this representation helps to improve sentence-level polarity classification. Williams *et al.* (2015) similarly report improvements on that task by incorporating the polarity of idioms that have been manually compiled. Jochim *et al.* (2018) present a polarity lexicon for idioms extracted from Wiktionary.

In summary, despite all those previous research efforts on MWEs, research with regard to sentiment has primarily been restricted to polarity classification. So far, sentiment views have not yet been considered.

2.3. Metaphors and multiword expressions

As mentioned in Section 1.2, MWEs are often used in a figurative sense. Indeed, several of the MWEs of our gold standard are lexicalized with a figurative meaning. More specifically, these expressions can be seen as metaphors. That is, in these expressions, a meaning transfer occurs via similarity of conceptual domains (Lakoff and Johnson 1980). For instance, in the MWE *play the second fiddle*, there is a transfer from the domain of an ORCHESTRA to the domain of general (business) HIERARCHY. Ideally, we would like to investigate whether the status of an MWE as metaphorical or not can help us in establishing the sentiment view of the expression. In this work, however, we refrain from considering the property of being a metaphor as an explicit feature. The reason for this is that although there has been a considerable body of work dealing with the detection of metaphors (Turney *et al.* 2011; Tsvetkov *et al.* 2014; Shutova 2015; Veale, Beigman

Klebanov, and Shutova 2016), we consider this task as an unsolved task in NLP that is at least as difficult as the task of determining the sentiment view of MWEs. For example, we are not aware of any publicly available system to detect metaphors. Therefore, metaphor detection will not be considered a plausible auxiliary task for the classification of sentiment views.

However, although we do not explicitly consider metaphor detection in this work, some of our features may approximate the distinction between metaphors and non-metaphors. For example, one feature we will consider distinguishes between the type of verbal MWEs in our dataset (Section 5.1.6), namely light-verb constructions (e.g., *have a laugh* or *take care*) and idioms (e.g., *hit the nail on the head* or *dip one's toe in the water*). Our observation is that while light-verb constructions usually have a literal meaning for the noun involved, idioms are more likely to incorporate figurative language, or more precisely metaphors.

Moreover, some of the features that we employ may also reflect properties that are considered for metaphor detection. One predictive feature is the determination of the degree of *concreteness* of an expression (Beigman Klebanov, Leong, and Flor 2015; Maudslay *et al.* 2020). Expressions that exhibit a high degree of concreteness are more likely to be used in a metaphorical way. For example, in *sea of sadness*, the noun *sea* denotes something concrete but it is used to specify the degree of an abstract concept, that is *sadness*. While we do not explicitly measure concreteness in our work, we employ features that take into consideration the semantic classes of component words of an MWE (Section 5.1.4). We will consider the semantic classes that are represented by the so-called *lexicographer files* from WordNet (Miller *et al.* 1990). We believe that there are certain semantic classes in this set that imply concreteness, for example *noun.animal*, *noun.food*, *noun.location*, *noun.plant* etc.

There has also been research in sentiment analysis looking into metaphors. However, that work focuses on tasks other than the detection of sentiment views, namely mainly the categorization of affect and polarity of metaphors (Kozareva 2013; Strzalkowski *et al.* 2014), so we cannot apply these methods to our task.

3. Data and annotation

In this section, we introduce the data and annotation we employ for our experiments. Next to an existing resource for sentiment-view classification based on unigrams (Section 3.1), we will introduce a new gold standard with MWEs labeled with sentiment-view information (Section 3.2). Moreover, we briefly discuss *Wiktionary* (Section 3.3), a web-based dictionary that is collaboratively produced. This resource plays a significant role in our experiments since it contains considerably more MWEs than the lexical resources previously employed for the categorization of sentiment views. Finally, we also present the different corpora we consider in this article (Section 3.4). Text corpora are vital for the methods based on distributional similarity.

3.1. Unigram sentiment-view lexicon

In this article, we heavily use the publicly available sentiment-view lexicon from Wiegand *et al.* (2016). In that lexicon all opinion adjectives, nouns, and verbs from the Subjectivity Lexicon (Wilson *et al.* 2005) are categorized either as conveying an actor view or a speaker view (see also Table 1). Table 2 illustrates entries from that lexicon. The crucial difference between this lexicon and our MWE gold standard lexicon (Section 3.2) is that the former lexicon exclusively contains *unigram* opinion words.

So far, this lexicon has only been employed as a gold standard for *evaluating* unigram sentiment-view classifiers. In this article, however, we will use this resource as a means of *building* sentiment-view classifiers for MWEs. In order to harness this lexicon for MWEs we can establish similarities between entries from this unigram lexicon and MWEs.

Table 1. Unigram lexicon with sentiment-view information from Wiegand *et al.* (2016)

	Actor View		Speaker View	
Part of Speech	Freq	Proportion	Freq	Proportion
adjective	223	8.9	2279	91.1
noun	487	29.1	1189	70.9
verb	618	52.6	557	47.4

Table 2. Illustration of entries from the unigram sentiment-view lexicon from Wiegand *et al.* (2016)

Actor View			Speaker View		
Adjective	Noun	Verb	Adjective	Noun	Verb
annoyed	assessment	approve	arrogant	abuse	blaspheme
bewildered	bereavement	despair	boring	beauty	cohere
content	enjoyment	fear	deceptive	corruption	deserve
depressed	intention	hope	enjoyable	dishonesty	excel
glad	loathing	imagine	harmful	harmony	fool
interested	objection	like	informative	ingenuity	plagiarize
lonely	promise	mourn	maladjusted	purity	qualify
pleased	regret	oppose	pathetic	ruthlessness	shine
tired	remark	salute	reasonable	success	tarnish
worried	wish	worship	tolerable	waste	worsen

3.2. MWE gold standard lexicon

The MWE gold standard lexicon represents the dataset on which we will carry out our experiments. In this work, we exclusively consider **verbal MWEs**. We define a verbal MWE as a sequence of tokens which includes at least one full verb and one noun where the verb is the syntactic head of the phrase (e.g., *pull the plug* or *beat around the bush*). These MWEs are also referred to as verb-noun MWEs (Liebeskind and HaCohen-Kerner 2016; Taslimipoor *et al.* 2017). We do not consider phrasal verbs (e.g., *take off* or *go out*) as part of our set of verbal MWEs since such expressions are widely covered by lexical resources, such as WordNet (Miller *et al.* 1990).[e] We, therefore, believe that the methods found effective for unigrams in Wiegand *et al.* (2016), which heavily rely on those lexical resources, should similarly work for phrasal verbs. In our work, we want to specifically look at the more difficult subtypes of MWEs, that is verb-noun MWEs.

Other types of MWEs, such as nominal MWEs (e.g., *golf club* or *nut tree*) or prepositional MWEs (e.g., *by car* or *on summer vacation*), are not considered in this work either. They have a much lower proportion of subjective expressions.

For our MWE gold standard lexicon, we consider a union of samples from two different resources: MWEs from Wiktionary[f] [thus following Jochim *et al.* (2018) and Kato *et al.* (2018)]

[e] We found more than 1900 phrasal verbs in this resource.

[f] https://en.wiktionary.org

Table 3. (Verbal) MWEs in different lexical resources

Resource	Type of Resource	Verbal MWEs
Wiktionary	online dictionary	4040
WordNet	lexical ontology	656
FrameNet	lexical ontology	120
COMLEX/NOMLEX	subcategorization lexicon	0
Subjectivity Lexicon	sentiment lexicon	0

and MWEs from SAID—the Syntactically Annotated Idiom Database (Kuiper *et al.* 2003). SAID itself is a compilation of several other dictionaries (Long, 1979; Cowie et al. 1983). We chose MWEs from Wiktionary, since, of the set of lexical resources commonly used for NLP, it contains by far the most (verbal) MWEs, as shown in Table 3. We also consider SAID because we want to have a varied dataset for MWEs. Only using MWEs from Wiktionary may have features drawn from that resource look unreasonably good since our dataset would exclusively contain entries for which Wiktionary would always also provide information.[g] We did not consider using corpora annotated for MWEs, since the number of *unique* verbal MWEs contained is usually far too small. For example, on the training set of the PARSEME corpus (Ramisch *et al.* 2018), we found only 96 unique verbal MWEs.

For our final gold standard, we sampled 800 MWEs per resource, that is Wiktionary and SAID.[h] We annotated those MWEs regarding their sentiment view. Only 6% of our data were considered as non-subjective and hence as not conveying any sentiment view. The high degree of subjectivity among (verbal) MWEs can be explained by their nature. Many MWEs represent some form of idiom. Nunberg *et al.* (1994) present *figuration* (i.e., the property of having a figurative meaning), *proverbiality*, *informality*, and *affect* as prototypical characteristics of such expressions. These characteristics also strongly imply subjective language.

Since the set of non-subjective MWEs is very small, we exclude it from our final gold standard. Previous work on unigram words (Wiegand *et al.* 2016) similarly decoupled the classification of sentiment views from subjectivity detection. The final dataset therefore only comprises two categories: *actor-view* and *speaker-view*. The sampling from the two resources was done independently of each other which resulted in a small overlap of MWEs. The final dataset contains 1355 *unique* MWEs. Table 4 illustrates entries from our new dataset.

The MWEs comprising our gold standard do not represent anywhere near the full set of English verbal MWEs. Otherwise, an automatic categorization would not be necessary in the presence of our gold standard. The classification approach that we propose in this paper, which works well with few labeled training data, would also be helpful for categorizing sentiment views on much larger sets of MWEs.

Despite the syntactic and semantic similarities between unigram verbs and verbal MWEs, we refrained from distinguishing between the two subtypes of actor views for opinion verbs, that is *agent views* and *patient views* as proposed by Wiegand and Ruppenhofer (2015) (see also Section 2.1). The reason for this is that among the verbal MWEs of our gold standard,

[g] SAID does not provide any word definitions or other information that could be harnessed as a feature for classification which is why we only use it for the compilation of our gold standard.

[h] We did not apply random sampling as we had to avoid very rare MWEs. Such MWEs, mostly those from Wiktionary, posed difficulties in manual annotation that were too great. The annotators could not reliably specify the sentiment view or even decide whether the entries are actually correct.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 4. Illustration of entries from the MWE gold standard lexicon

Actor View		Speaker View	
arrive at the conclusion	bear in mind	bend the truth	add fuel to the fire
break the news	come to grips	bring to the table	do justice
draw the line	express an interest	come full circle	face the music
hammer home	fall in love	fit the bill	get away with murder
have second thoughts	get wind of	grease the wheels	have a future
leap for joy	keep an eye on	jump the shark	look for trouble
read the riot act	make a statement	mean business	make a mistake
shed tears	put out feelers	rise to the surface	pay the price
take pleasure	sit on the fence	stand a chance	strike gold
throw in the towel	wage war against	weather the storm	work wonders

Table 5. The MWE gold standard lexicon

Property	Frequency
no. of MWEs	1355
actor-view MWEs	562 (41.5%)
speaker-view MWEs	793 (58.5%)
average token length	3.33
MWEs with at least one unigram opinion word	448 (33.1%)

the proportion of patient-view expressions is less than 2%. We consider such a low number of instances to be insufficient for carrying out classification experiments.[i]

Table 5 shows some further statistics of our gold standard, including class distribution. Both sentiment views have a significant share. For the annotation of the MWEs, the same annotator as for the unigram lexeme-level annotation from Wiegand *et al.* (2016), a trained linguist and one coauthor of this article, was employed. We also adhere to the annotation process proposed in that work. That is, the basis of the annotation were various dictionaries (e.g., *Macmillan Dictionary*) which provide both a word definition and example sentences for the MWEs. The example sentences represent prototypical contexts in which the relevant opinion expression, in our case an MWE, may occur. From such contexts, it is fairly straightforward to derive the respective sentiment view. For the annotation guidelines, however, some additions for MWEs were necessary:

- We emphasized that the annotators were to annotate the meaning of the MWE (and not of individual constituents).
- In order to be in line with previous work (Deng and Wiebe 2016; Wiegand *et al.* 2016), we consider the two sentiment views to be mutually exclusive categories. However, we

[i] For comparison, on the set of unigram verbs that is used in Wiegand and Ruppenhofer (2015), 16% of the verbs had been classified as patient-view words.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

observed that there are a few MWEs that actually simultaneously convey actor and speaker view. For example, the MWE *try one's best* does not only convey that the agent has a positive sentiment towards the goal which it wants to achieve (i.e., *actor view*), but it also conveys a positive evaluation of the speaker towards the agent, who is trying hard (i.e., *speaker view*). We deliberately did not introduce a new category, that is *conveying both actor and speaker view*, since it would have further increased the complexity of our annotation scheme. Our impression was that there are actually not that many MWEs which equally convey both sentiment views. So, for our classification experiments we would not have ended up with sufficient labeled instances for all three categories. Moreover, by maintaining the concept of two mutually exclusive categories, we preserve the compatibility of our new dataset with the data of previous work, particularly the unigram sentiment-view lexicon (Section 3.1). In practice, in case our annotators faced an MWE they thought to convey both actor and speaker view, they were to prefer the sentiment view that they think is more prominent. For example, in (38), we would recommend to label the MWE as conveying a speaker view since it is more prominent than the actor view.

(38) Peter **tried his best**$_{\textit{speaker-view/actor-view}}$.

Typically, the goal that the agent of *try one's best*, that is *Peter* in (38), has (in other words, the target of this explicit opinion holder), is not realized as a dependent of the MWE and remains implicit. This suggests that the resulting actor view is not that prominent.

- Another issue that we addressed in the guidelines is the treatment of MWEs with multiple senses. As we carry out an out-of-context annotation, it becomes problematic if we face an MWE with two meanings that also convey different sentiment views. For example, the MWE *take a back seat* has two meanings (according to *Macmillan Dictionary*):

(38) *to deliberately become less active, and give up trying to control things:* I'll be happy to **take a back seat**$_{\textit{actor-view}}$ when Robin takes over.

(39) *to become less important:* Other issues must **take a back seat**$_{\textit{speaker-view}}$ to this crisis.

Due to the lack of robust word-sense disambiguation, we are pursuing a lexeme-level annotation rather than a sense-level annotation. Therefore, we can only assign one sentiment view to each MWE. In principle, the annotators were to consider the most common reading which typically coincides with the first sense listed in the lexicon. However, the annotators were to focus on a figurative sense of an MWE if the first sense had a literal reading, since in most cases, the figurative sense of MWEs (e.g., *throw in the towel* in the sense of *stop trying to do something* or *close one's eyes* in the sense of *ignore something bad*) represents the subjective reading that we are interested in. In general, the share of MWEs having multiple senses is fairly moderate. For example, for the set of all verbal MWEs from Wiktionary, we computed an average of only 1.2 word senses while for the set of unigram verbs, we computed an average of 3.1 senses.

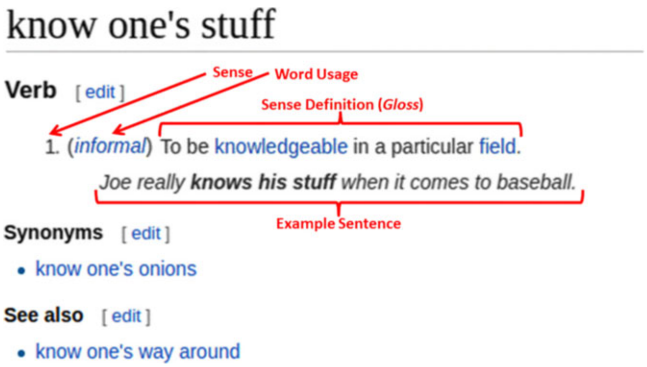
All novel data created as part of this research including annotation guidelines is publicly available.[j]

On a sample of 400 MWEs, we computed an interannotation agreement between the main annotator and the first author of this article. We obtained an agreement of Cohen's $\kappa = 0.62$. This score can be considered substantial (Landis and Koch 1977).

Although we achieved a good agreement for this task, it is not perfect. The few systematic disagreements that we found were MWEs that actually convey both actor and speaker view and

[j] https://doi.org/10.5281/zenodo.7423947

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press



know one's stuff

Verb [edit]

1. (*informal*) To be knowledgeable in a particular field.

*Joe really **knows his stuff** when it comes to baseball.*

Synonyms [edit]

- know one's onions

See also [edit]

- know one's way around

Figure 1. Illustration of a Wiktionary-entry.

the annotators chose a different view as the prominent one. Moreover, there were also occasionally MWEs having more than one (figurative) sense and the annotators annotated the sentiment view of different senses.

3.3. Wiktionary

We now turn to Wiktionary as we also use some information from this resource for feature engineering. Wiktionary is a freely available web-based dictionary. One major difference between this resource and the more commonly used WordNet is that it is written collaboratively by volunteers rather than linguistic experts. Despite possible concerns about its linguistic accuracy, this dictionary seems more suitable for our work than WordNet since it has a much wider coverage of MWEs, as shown in Table 3. There are more than six times as many MWEs according to our definition (Section 3.2).

Figure 1 illustrates a typical Wiktionary-entry for MWEs. Similar to WordNet, we find for each sense information on its usage, a definition (typically referred to as *gloss*) and one example sentence. Optionally, there are links to both synonyms and near-synonyms (*See also*). In our work, we focus on glosses rather than example sentences as the former are known to be predictive for lexicon categorization tasks (Esuli and Sebastiani 2005; Andreevskaia and Bergler 2006; Choi and Wiebe 2014; Kang *et al.* 2014). In order to process Wiktionary automatically, we use JWKTL (Zesch, Müller, and Gurevych 2008). Due to the lack of robust word-sense disambiguation, we will always consider the union of all sense descriptions of a given MWE. We think that working on the lexeme level instead of the sense level will only marginally affect our results since MWEs tend to be less ambiguous than unigrams. (In the previous subsection, we already provided figures on the degree of ambiguity of verbal MWEs compared to unigram verbs.)

3.4. Corpora

Some of the classification approaches we employ require corpora. For example, they are required for inducing word embeddings. (We cannot use pre-trained embeddings, such as *CommonCrawl*, since they only encode unigrams and no MWEs.) We consider three different corpora as displayed in Table 6. In principle, every corpus-based method can be implemented with the help of either of those corpora. Following Wiegand *et al.* (2016), the first corpus we use is NEWS—the North American News Text Corpus (LDC95T21). Although it is the smallest corpus, it has the advantage of comprising well-written text. Our second corpus LIU is the set of reviews from Jindal and Liu (2008). It contains more sentiment-related text and is considerably larger. Our final corpus UKWAC (Baroni *et al.* 2009) is a corpus crawled from the web. It is twice as large as LIU but does

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 6. The different corpora used

		Coverage [percent]		
Corpus	Tokens	Unigram Lex.	Verbal MWEs*	Property
NEWS	∼0.17B	99.76	57.69	*clean text*
LIU	∼1.17B	99.38	71.63	*much sentiment*
UKWAC	∼2.25B	98.69	64.58	*large corpus*

*The union of **all** (verbal) 6996 MWEs in Wiktionary and SAID (i.e., not only those that comprise the MWE gold standard) are considered.

not focus on sentiment-related text. Table 6 shows that while the different corpora vary in terms of coverage of MWEs, they all cover the lexical units from our unigram sentiment-view lexicon (Section 3.1) equally well.

4. Graph-based label propagation using a unigram sentiment-view lexicon

Our first set of experiments tries to uncover the sentiment views of MWEs without access to any MWEs manually labeled with sentiment-view annotation. Instead, we infer the sentiment views of MWEs with the help of opinion unigrams as encoded in our unigram sentiment-view lexicon (Section 3.1). We describe this approach in Section 4.1 and the results of our experiments in Section 4.2. This method presents a baseline that will be compared against supervised classifiers using minimal amounts of labeled training instances of MWEs in Section 5.

4.1. The classification approach

We consider a graph-based classifier. The nodes in the graph represent opinion expressions. They comprise all unigram opinion words from the sentiment-view lexicon (Section 3.1) and all MWEs from our gold standard (Section 3.2). In total, there are 6708 nodes, that is, 5353 nodes corresponding to each unigram opinion expression and 1355 corresponding to each MWE from our gold standard. All unigram opinion words are labeled while all MWEs are unlabeled. The nodes are connected by edges. Edge weights are computed by distributional similarity. We employ cosines of *Word2Vec*-vector representations (Mikolov *et al.* 2013) for these weights. We can compute the similarity between each possible pair of nodes for which we obtain a vector representation. As shown in Table 6, while almost all unigram opinion words are represented in all corpora we experiment with, the coverage of MWEs varies considerably. This also has an impact on the connectedness of the resulting graph. If we only consider the nodes representing words or MWEs for which there is a vector representation, we actually can produce a fully connected graph.[k]

We follow Mikolov *et al.* (2018) in that we represent MWEs as one artificial word, that is we concatenate the tokens of the MWE in the corpus on which *Word2Vec* is run (e.g., *kick_the_bucket*). We induce vectors with 500 dimensions leaving any other parameter of

[k] Since that graph structure, that is, a graph in which each opinion expression for which there is a vector representation can be connected to any other opinion expression with a vector representation, is too resource-intensive for the graph-based classifier we want to use, we only consider for each node the 10 edges with the highest distributional similarity to another node. This cutoff value was not tuned on our dataset but was taken from previous work (Wiegand and Ruppenhofer 2015; Wiegand *et al.* 2016).

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

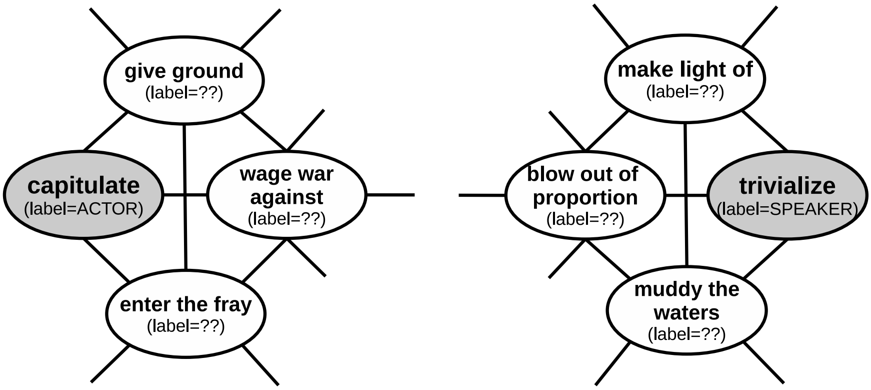


Figure 2. Illustration of graph-based approach (*unigram opinion words are labeled seeds*).

Word2Vec at its default configuration.[1] (Thus we follow the parameter settings from Wiegand and Ruppenhofer (2015) who perform the same classification approach on a set of unigram opinion verbs. We assume that verbal MWEs and unigram opinion verbs largely share both syntactic and semantic properties, so that the settings for unigram verbs should be equally applicable for verbal MWEs.) All unigram opinion words are used as labeled seeds. Subsequently, we run label propagation in order to determine the labels of the MWEs. Figure 2 illustrates this graph structure. For label propagation, we consider the Adsorption label propagation algorithm as implemented in *junto* (Talukdar *et al.* 2008). Adsorption is a general framework for label propagation, consisting of a few nodes annotated with labels and a graph structure containing the set of all labeled and unlabeled nodes. This method labels all nodes based on the graph structure, ultimately producing a probability distribution over labels for each node in the graph.

This type of classifier exclusively draws its knowledge from *distributional similarity*. Table 7 illustrates the 10 most similar unigram opinion words for two different MWEs (*sit on the fence* and *bend the truth*). It shows that most of the similar words share the sentiment view of the MWE. This is an important pre-requisite in order to make graph-based label propagation work on our particular setting.

Related to this classification set-up there are two different aspects we want to examine:

- What type of corpus should be used as a basis to induce the vector representation of MWEs?
- What type of unigram opinion words should be included in the graph?

4.1.1. Corpus choice for MWEs

Since sparsity is an issue for (verbal) MWEs, the type of corpus from which we induce our vector representation is also likely to have an impact. We consider the different corpora from Section 3.4, that is NEWS, LIU, and UKWAC (Table 6). We want to find out whether larger corpora really produce better results for such a classification approach (i.e., UKWAC and LIU vs. NEWS) and whether a high concentration of sentiment information has a further impact on the results (i.e., LIU vs. UKWAC).

[1] The parameter settings are as follows: objective: continuous bag of words (cbow); number of negative examples: 5; minimal count of words: 5; number of training iterations: 5; threshold for occurrence of words: 1e-3; starting learning rate: 0.05.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 7. The 10 most similar unigrams for two different MWEs (embeddings were induced on the corpus LIU); unigrams conveying a sentiment view other than that of the MWE are in **bold** type

MWE: sit on the fence *(actor)*	MWE: bend the truth *(speaker)*
undecided *(actor)*	dissemble *(speaker)*
adamant *(actor)*	**impugn *(actor)***
unsure *(actor)*	fabricate *(speaker)*
debate *(actor)*	propagandize *(speaker)*
cautious *(speaker)*	disingenuous *(speaker)*
unconcerned *(speaker)*	defame *(speaker)*
apprehensive *(actor)*	moralize *(speaker)*
skeptical *(actor)*	misinterpret *(speaker)*
leery *(actor)*	exaggerate *(speaker)*
enthuse *(speaker)*	distort *(speaker)*

Table 8. Label propagation on different corpora and vector representations

Corpus	Acc	Prec	Rec	F1
NEWS	54.12	62.81	52.55	57.22
UKWAC	57.93	68.47	54.14	60.47
LIU	**63.14**	**68.48**	**60.42**	**64.20**
LIU (only verbs)	52.83	63.22	56.52	59.69

4.1.2. Subsets of unigram opinion words relevant for MWEs

Regarding the issue of what types of unigram opinion words should be represented, we want to examine how far the part of speech of those different opinion words plays a role. Syntactically speaking verbal MWEs have the greatest resemblance to opinion verbs. Yet also including opinion nouns and opinion adjectives may add some extra information to the graph that could help in label propagation.

4.2. Results for graph-based label propagation

Table 8 shows the results of our experiments. We report accuracy and macro-average precision, recall and F-score on our gold standard (Section 3.2). The largest corpus (i.e., UKWAC) does not produce best performance. The best-performing configuration is on LIU. That is, the best classification can be obtained by a corpus which yields the highest sentiment concentration (see Table 6). For this particular corpus, Table 8 also shows the performance of a graph which only comprises verbal MWEs and unigram verbs. That graph produces much worse classification performance than the original graph (i.e., a graph in which unigram nouns and adjectives are also included). From this, we conclude that verbal MWEs also draw significant information from unigram opinion nouns and adjectives. This is further supported by Table 9 which shows the average

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 9. Average proportion of the different parts of speech among the 10 most similar unigram opinion words (embeddings were induced on LIU)

Part of Speech	Proportion
verbs	73.06
adjectives	19.47
nouns	12.70

Note that the sum of those proportions exceeds 100%. This is due to the fact that since our *Word2Vec* representation does not incorporate parts of speech information, some opinion words are ambiguous (for instance, *love* can function both as a verb and a noun).

distribution of the three parts of speech among the 10 most similar opinion unigrams for each of our verbal MWEs. More than 30% of these unigrams are either unigram nouns or adjectives.

In our subsequent experiments, we use the best corpus, that is LIU, for all features using corpus-based information and distributional similarity.

5. Feature-based approach using supervised learning

In this section, we present features for a supervised learning approach to the classification of sentiment views of MWEs. Unlike the method presented in the previous section, this method requires MWEs as labeled training data. After discussing the specific features we devised for MWEs (Section 5.1), we briefly describe Markov Logic Networks, the supervised classifier in which we integrate our features (Section 5.2). We also present the global constraints that we incorporate into this classifier. Then, we present the baseline supervised classifiers against which we compare Markov Logic Networks (Section 5.3). This is followed by the presentation of our experimental results (Section 5.4). We conclude this section with an error analysis (Section 5.5).

5.1. Feature design

Our features for this task can be divided into three different units, which we call **representation foci** (Table 10). Each individual feature is defined as part of one of these foci. The most straightforward representation focus is the MWE itself (`MWE`). Features that operate on this focus are features that are applied to the entire MWE or to individual component tokens. Another set of features considers corpus-based mentions of the MWE (`CORP`). These features typically exploit the context words of MWE mentions. Our final representation focus considers the information provided by the Wiktionary-entry of an MWE (`WIKT`). Table 11 provides a summary of our features. It also assigns each individual feature its representation focus. We now present all these features which we further group into subsets sharing the same resource. We first discuss the features inspired by previous work (Wiegand *et al.* 2016) and show how they need to be adjusted for MWEs (Sections 5.1.1–5.1.5). Then, we present the completely novel features (Sections 5.1.6–5.1.9). Wiegand *et al.* (2016) is the only previous work on lexeme-level sentiment-view classification that explores diverse features.[m] Therefore, we can only consider features from this work as a reference feature set.

[m] Deng and Wiebe (2016) only consider word embeddings as a feature representation of opinion words, and Wiegand and Ruppenhofer (2015) is not a feature-based but a graph-based approach to sentiment-view classification.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 10. The different representation foci

Focus	Description
MWE	Considers the MWE as a whole or individual tokens of the MWE.
CORP	Considers corpus-mentions of the MWE.
WIKT	Considers information of Wiktionary-entry corresponding to the MWE.

Table 11. Summary of all features used

Feature	Focus	Description
*PATT**	CORP	Co-occurrence of MWEs with *prototypical opinion holders* for the extraction of actor-view MWEs; co-occurrence with reproach patterns for the extraction of speaker-view MWEs.
SUBC	CORP	Arguments subcategorized by MWE automatically extracted from text corpus.
*FN-direct**	MWE	FrameNet-frame(s) containing MWE.
FN-uni	MWE	FrameNet-frame(s) containing unigrams of MWEs.
FN-gloss	WIKT	FrameNet-frame(s) containing unigrams of Wiktionary-gloss(es) of MWE.
*WN-direct**	MWE	WordNet-lexicographer file(s) containing MWE.
WN-uni	MWE	WordNet-lexicographer file(s) containing unigrams occurring in MWE.
WN-gloss	WIKT	WordNet-lexicographer file(s) containing unigrams occurring in Wiktionary-gloss(es) of MWE.
POLAR-uni	MWE	Most frequent polarity of unigrams contained in MWE according to Subjectivity Lexicon.
POLAR-gloss	WIKT	Most frequent polarity of unigrams contained in Wiktionary-gloss(es) of MWE.
STRUC-light	MWE	Does MWE represent a light-verb construction?
STRUC-length	MWE	Number of tokens comprising MWE.
POS	WIKT	Most frequent parts of speech tag in Wiktionary-gloss(es) of MWE.
ULEX-uni	MWE	Most frequent sentiment view of the unigram opinion words contained in MWE according to unigram sentiment-view lexicon (Section 3.1).
ULEX-gloss	WIKT	Most frequent sentiment view of the unigrams occurring in Wiktionary-gloss(es) of MWE.
ULEX-distr	CORP	Most frequent sentiment view of the 10 most distributionally similar unigrams of MWE.
USAGE	WIKT	Usage information of the word sense (Fig. 1) provided by Wiktionary-entries of MWE (bag-of-words feature).

*Features from Wiegand *et al.* (2016) that can be immediately applied to MWEs (*the other features from that work either employed resources which only hold unigram entries or they turned out to be too sparse in our initial exploratory experiments*).

5.1.1. Pattern-based approaches (PATT)

Wiegand *et al.* (2016) proposed pattern-based approaches for this task. Actor-view words are identified by extracting opinion words from a corpus that frequently occur with *prototypical opinion holders*, that is common nouns that act as explicit opinion holders, such as *opponents* or *critics* as

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

in (38) and (39). By definition, explicit opinion holders are indicative of actor views since speaker views have the speaker of the utterance as an implicit holder.

(38) Opponents$_{\textit{prototyp. opinion holder}}$ **claim**$_{\textit{actor-view}}$ these arguments miss the point.

(39) Critics$_{\textit{prototyp. opinion holder}}$ **argued**$_{\textit{actor-view}}$ that the proposed limits were unconstitutional.

Prototypical opinion holders may similarly co-occur with actor-view opinion expressions being multiword expressions as in (40) and (41).

(40) Opponents$_{\textit{prototyp. opinion holder}}$ **come to the conclusion**$_{\textit{actor-view}}$ that the effects of social media on youths are not all positive.

(41) Critics$_{\textit{prototyp. opinion holder}}$ **find fault**$_{\textit{actor-view}}$ with the government for not providing sufficient supervision of the banks.

Therefore, in order to extract actor-view MWEs with the help of prototypical opinion holders, we simply extract verbal MWEs instead of unigram opinion words occurring with them.

Speaker-view words can be extracted with the help of *reproach patterns*, for example *blamed for X* as in (42). Such patterns are motivated by the fact that reproaches are usually speaker-view words. Similar to the pattern-based method to extract actor-view opinion expressions, the pattern-based method to extract speaker-view opinion expressions can be applied to MWEs in the same way in which it has been applied to unigrams. The only difference is that we extract verbal MWEs (43) instead of unigram opinion words (41) occurring with these patterns.

(42) The US was blamed for **misinterpreting**$_{\textit{speaker-view}}$ climate data.

(43) The US was blamed for **closing their eyes**$_{\textit{speaker-view}}$.

5.1.2. Subcategorization (SUBC)

Actor-view MWEs usually require two explicit arguments, that is opinion holder and target. Consequently, those MWEs should have two obligatory arguments x_i (e.g., x_1 *draws the line at* x_2 or x_1 *expresses interest in* x_2). For speaker views, on the other hand, only one argument is required since the holder is the implicit speaker of the utterance (e.g., x_1 *makes an error* or x_1 *adds fuel to the fire*). We want to use subcategorization information to capture this tendency. Unfortunately, the publicly available subcategorization lexicons, that is COMLEX (Grishman, McKeown and Meyers 1994) and NOMLEX (Macleod *et al.* 1998), do not contain any MWE entries. So we cannot follow Wiegand *et al.* (2016) in using these resources for lookup. Instead, we extract the information from a corpus.

If we were to extract from a corpus subcategorization information for a unigram, for example a verb, we would simply extract the labels of all (immediate) dependency relations connecting the verb with its dependents. However, since we are now dealing with an MWE, we need to extract the labels of those dependency labels for all tokens representing content words of the MWE. Since there may be dependency relations between the tokens of the same MWE, we can omit these labels from the final list of labels of dependency relations, as we are only interested in the relation of the MWEs to other tokens that are not part of the MWE itself. We will illustrate this with the two sentences (44) and (45):

(44) [Russia]$_{\textit{subj}}$ **agrees** [with the coalition]$_{\textit{pobj\_with}}$ [on Syria airstrike targets]$_{\textit{pobj\_on}}$.

(45) [Russia]$_{\textit{subj}}$ **sees** [**eye**]$_{\textit{dobj}}$ [**to eye**]$_{\textit{pobj\_to}}$ [with the coalition]$_{\textit{pobj\_with}}$ [on Syria airstrike targets]$_{\textit{pobj\_on}}$.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Since these two sentences are synonymous, we should extract the same subcategorization information for the respective predicates, that is *agrees* in (44) and *sees eye to eye* in (45). We can compute for (44) that *agrees* has as dependents a subject (*subj*), and two prepositional objects (*pobj_with* and *pobj_on*). For (45), we consider all dependents of the single tokens of the MWE *sees eye to eye* that represent content words.[n] The dependents of *sees* are a subject (*subj*), and one direct object (*dobj*) and three prepositional objects (*pobj_to*, *pobj_with* and *pobj_on*). Two of these dependents are actually part of the MWE itself, that is *dobj* and *pobj_to*. These two dependents are omitted, which results in the remaining dependents being exactly those of the single-word predicate in (44). (The two remaining content words of the MWE, *eye* and *eye* do not have any dependents themselves, so we have nothing further to consider for subcategorization.) Our corpus is parsed by the Stanford parser (Klein and Manning 2003).

5.1.3. FrameNet (FN)

FrameNet (Baker *et al.* 1998) is a semantic resource that collects words with similar semantic behavior in semantic frames. Wiegand *et al.* (2016) found that information from FrameNet is useful for sentiment-view classification. Different frames are associated with different sentiment views. For example, the frame PREVARICATION contains speaker-view opinion expressions, such as *deceive*, *lie,* or *mislead*, while the frame TAKING_SIDES contains actor-view opinion expressions, such as *endorse*, *oppose,* or *support*. Replicating the FrameNet-feature from Wiegand *et al.* (2016) (*FN-direct*), that is, looking up the frames of our MWEs in FrameNet, results in poor coverage. We could only identify less than 3% of MWEs in that resource. For example, a speaker-view MWE *bend the truth* has a semantics similar to *lie* but due to the coverage limitations of FrameNet, this MWE is not included in the frame which *lie* includes, that is PREVARICATION.

We introduce two methods to exploit information from that resource more effectively for our task despite the fact that most of our MWEs are not included as such. For the first method, we look up every unigram in the MWE to be categorized in FrameNet and consider the union of frames found for the individual unigrams (*FN-uni*). For the second method, we look up every unigram occurring in the Wiktionary-gloss(es) of the MWE and consider the union of frames found for those words (*FN-gloss*). Of course, we are aware that these two features make the simplifying assumption that the meaning of an MWE can be reduced to the meaning of its composite tokens. While this is not true for several MWEs, we hope that our two features can at least partly compensate the sparsity of MWEs in FrameNet.

5.1.4. WordNet (WN)

WordNet (Miller *et al.* 1990) is the largest available ontology for English and a popular resource for sentiment analysis, in general. Wiegand *et al.* (2016) established that there are correspondences between sentiment views and the WordNet-*lexicographer files*, also referred to as *supersenses* (Flekova and Gurevych 2016), that is a set of 45 coarse-grained classes into which each synset is categorized. There are particular lexicographer files that predominantly include opinion expressions conveying a particular sentiment view. For instance, the lexicographer file change contains many speaker-view opinion expressions, such as *barbarize*, *damage* or *facilitate*.

Similar to FrameNet (Section 5.1.3), looking up the lexicographer files of our MWEs directly (*WN-direct*) only results in a coverage of 13.4%. We, therefore, introduce two methods to exploit this information more effectively. For the first method, we look up every unigram occurring in an MWE and consider the union of lexicographer files found (*WN-uni*). For the second method, we look up every unigram occurring in the Wiktionary-gloss(es) of an MWE and consider the union of lexicographer files found for those words (*WN-gloss*).

[n] We do not consider the function words of the MWE, that is, in our example *to*, since these words do not function as predicates having a subcategorization frame.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

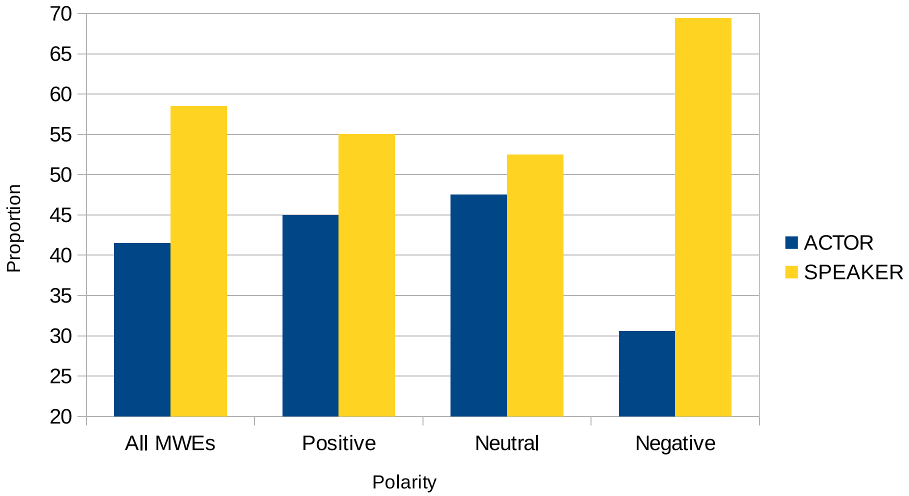


Figure 3. Polarity distribution among the different sentiment views.

5.1.5. Polarity (POLAR)

Figure 3 shows the distribution of polarity labels manually assigned to our MWEs.[o] Among the MWEs with a negative polarity there is a notably higher proportion of speaker views. Therefore, polarity information may be helpful for our task.

In order to determine the polarity of MWEs *automatically*, we look up the polarity of the opinion words occurring in an MWE and calculate for each MWE the most frequently observed polarity (*POLAR-uni*). We obtain polarity information of opinion words from the *Subjectivity Lexicon* (Wilson *et al.* 2005). Our second feature (*POLAR-gloss*) considers the most frequently observed polarity in the Wiktionary-gloss(es) of an MWE.

5.1.6. Internal structure of MWEs (STRUC)

We can divide the set of our verbal MWEs into two subcategories: idiomatic MWEs (e.g., *hit the nail on the head* or *dip one's toe in the water*) and light-verb constructions (e.g., *have a laugh*, *take care* or *give voice*) (Baldwin and Kim 2010). While the former MWEs can assume varying shapes, the latter MWEs typically comprise a light verb (e.g., *have*, *take*, *give*) followed by a noun. Our first feature checks whether an MWE represents a light-verb construction (*STRUC-light*). In order to avoid overgeneration we restrict ourselves to constructions in which the noun is a deverbal noun (e.g., *decision*, *kiss*, *sigh*).[p] We detect such nouns with the help of NOMLEX (Macleod *et al.* 1998). Figure 4 displays the distribution of sentiment views among light-verb constructions. There is a notably larger proportion of actor views.

Our second feature counts the number of tokens comprising the MWE (*STRUC-length*). Figure 5 shows the distribution. Short MWEs are more likely to represent actor views while longer MWEs tend to represent speaker views.

Obviously, *STRUC-length* is also related to *STRUC-light*. The more tokens an MWE comprises, the less likely it is to represent a light-verb construction. Since our detection of light-verb constructions is fairly coarse, *STRUC-length* could be regarded as a back-off feature for *STRUC-light*.

[o] In addition to manually annotating our gold standard with respect to sentiment views (Section 3.2), we also annotated those MWEs with respect to polarity.

[p] Due to the lack of any publicly available system, a more advanced detection as in the fashion of Vincze *et al.* (2013) or Chen *et al.* (2015) is beyond the scope of this work.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

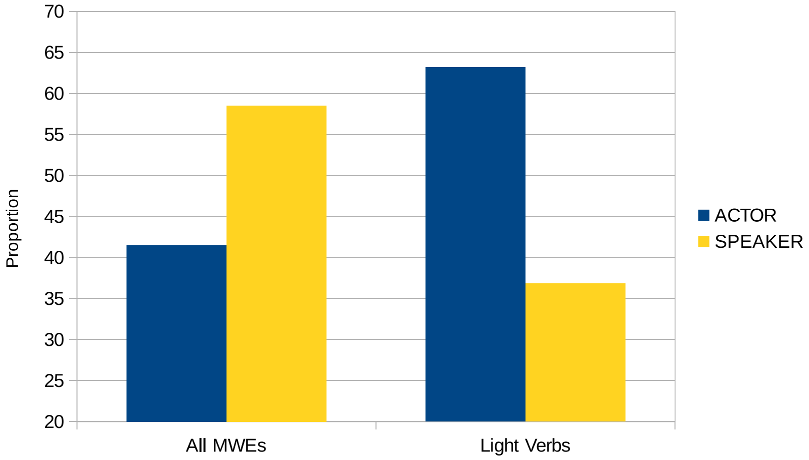


Figure 4. Sentiment-view distribution on light-verb constructions.

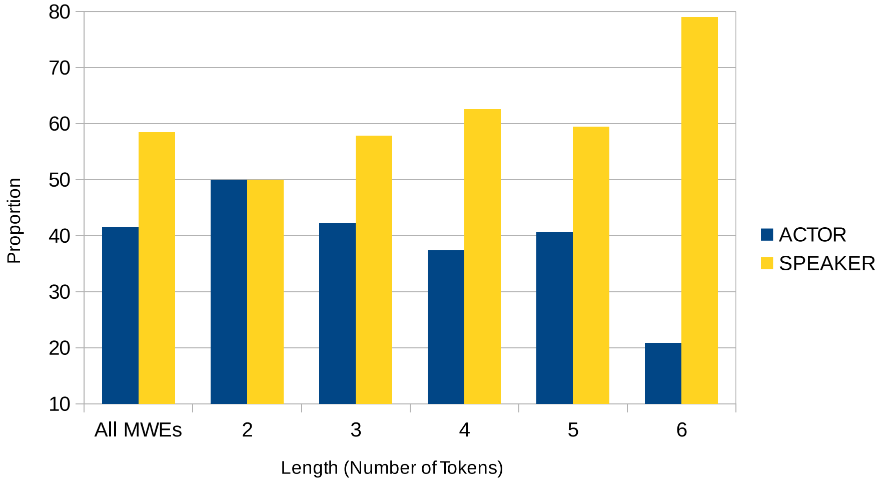


Figure 5. View distribution and token length.

5.1.7. Part of speech (POS)

The distribution of sentiment views among unigram adjectives (Table 1) is heavily skewed towards speaker views. We assume that some of our MWEs are *adjective-like* in nature. We also hypothesize that those MWEs that denote properties as adjectives do are also much more likely to convey speaker views. (46) and (47) are examples of such MWEs. The adjective-like nature becomes obvious if one considers the Wiktionary-glosses of these MWEs. Adjectives dominate these glosses.

(46) MWE: hit the spot$_{speaker\text{-}view}$; GLOSS: *To be particularly* ***pleasing***$_{adj}$ *or* ***appropriate***$_{adj}$; *to be just* ***right***$_{adj}$.

(47) MWE: go by the wayside$_{speaker\text{-}view}$; GLOSS: *To become* ***obsolete***$_{adj}$ *or* ***outmoded***$_{adj}$.

In order to detect adjective-like MWEs, we compute the most frequent part of speech (we only count adjectives, adverbs, nouns, and verbs) in the gloss(es) of an MWE. In (46) and (47) the most frequent part of speech are adjectives.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

5.1.8. Unigram sentiment-view lexicon (ULEX)

We also use the unigram sentiment-view lexicon (Section 3.1) that we already harnessed for building a graph-based classifier (Section 4) for feature engineering in supervised classification. *ULEX-uni* computes the most frequent view of the unigram opinion words contained in the MWE itself. Table 5 already stated that only 30% of our MWEs contain an opinion word. So this feature can only be of limited help. *ULEX-gloss* computes the most frequent view of the unigrams occurring in the Wiktionary-gloss(es) of the MWE. *ULEX-distr* establishes the connection between MWEs and unigram opinion words by distributional similarity. For each MWE, we extract the 10 most similar words from the unigram view lexicon and use the most frequent sentiment view associated with these unigrams as a feature. Similarity is computed on the basis of the cosine of vector representations between our MWEs and the words of the unigram sentiment-view lexicon. The vectors are induced with *Word2Vec* using the best induction configuration established in the context of graph-based label propagation in Section 4.

5.1.9. Usage information (USAGE)

Many sense descriptions of a Wiktionary-entry contain in parentheses some information on the usage of the sense (Figure 1). This is typically information on the speech register in which an expression is commonly used (e.g., *informal*, *vulgar* etc.). We consider all this information and encode it as a bag-of-words feature.

5.2. Markov logic networks (MLN) and global constraints

Markov Logic Networks (MLN) are a supervised classifier combining first-order logic with probabilities. MLN are a set of pairs (F_i, w_i) where F_i is a first-order logic formula and w_i a real valued weight associated with F_i. The probability distribution that is estimated is a log-linear model $P(X = x) = \frac{1}{Z} exp\left(\sum_{i=1}^{k} w_i n_i(x)\right)$ where $n_i(x)$ is the number of groundings[q] of F_i in x and Z is a normalization constant. As an implementation, we use *thebeast* (Riedel 2008).

From a practical perspective, MLN can be used in the same way as other traditional supervised learning algorithms, such as SVM or logistic regression. Rather than encoding features, in MLN, we encode so-called *local constraints*. These constraints produce similar classification performance as the equivalent features in traditional supervised learning algorithms. However, in addition, MLN allow us to formulate *global constraints*. While local constraints, similar to features in traditional supervised learning, describe observed properties on individual instances,[r] global constraints describe relations between different instances. This enables a classifier to make predictions for some instance not only on the basis of the *features* with which it has been (individually) observed. We can also exploit the similarity (or dissimilarity) between two instances. This may be advantageous if two instances *a* and *b* share a considerable degree of similarity but we have only observed sufficiently predictive features (i.e., local constraints) for instance *a*. While traditional supervised learning algorithms would struggle to make an appropriate prediction for instance *b* (since no predictive features have been observed with it), in MLN, we can enforce by a similarity constraint that instances *a* and *b* should also be assigned the same class labels. This enables the classifier to project the features observed for instance *a* also to our sparse instance *b*.

It is precisely because of this additional expressiveness of MLN due to the global constraints that we have chosen this classifier for our given task. We employ MLN as they allow us to formulate constraints holding between individual MWEs in addition to the ordinary features (Sections 5.1.1–5.1.9). Wiegand *et al.* (2016) report performance increases by incorporating similarity constraints. The only similarity constraint that can be translated to the setting of MWEs

[q] Grounding means that all variables in a formula are replaced by some constants.

[r] In our given task, the set of instances represents the set of MWEs for which we want to determine the sentiment view.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 12. Global constraints enforcing sentiment-view consistency as incorporated in MLN

Abbreviation	Constraint as Logic Formula
CONSTR-distr	$\forall x[\forall y[\forall z[\forall u[[MWE(x) \wedge MWE(y) \wedge Word2Vec\text{-}Similar(x,y) \wedge ViewOf(z,x) \wedge ViewOf(u,y)] \rightarrow (z == u)]]]]$ (Distributionally similar MWEs share the same sentiment view.)
CONSTR-wikt	$\forall x[\forall y[\forall z[\forall u[[MWE(x) \wedge MWE(y) \wedge Synonym(x,y) \wedge ViewOf(z,x) \wedge ViewOf(u,y)] \rightarrow (z == u)]]]]$ (MWEs linked with Wiktionary *Synonymy*-links share the same sentiment view.)
	$\forall x[\forall y[\forall z[\forall u[[MWE(x) \wedge MWE(y) \wedge See\text{-}also(x,y) \wedge ViewOf(z,x) \wedge ViewOf(u,y)] \rightarrow (z == u)]]]]$ (MWEs linked with Wiktionary *See-also*-links share the same sentiment view.)

is a *distributional similarity constraint* using *Word2Vec*-embeddings. Wiegand *et al.* (2016) also employ another distributional similarity constraint based on the similarity metric from Lin (1998). This metric cannot be applied to MWEs since it is defined on dependency relation triples which are based on individual tokens. Such a representation is incompatible with MWEs, which are multi-token expressions. Neither could we employ the *morphological similarity constraint* from Wiegand *et al.* (2016) in which morphological relatedness between different opinion unigrams was established with the help of the WordNet link *derivational related form*. Since WordNet only contains a small fraction of MWEs (Table 3), such a constraint would have applied to too few MWEs to be effective.

For our distributional similarity constraint, we compute for each MWE the five most similar other MWEs.[s] Our constraint *CONSTR-distr* (Table 12) requires all those distributionally similar MWEs to possess the same sentiment view. Distributional similarity is established in the same manner as we computed it in the feature *ULEX-distr* (Section 5.1.8). The major difference is that this constraint computes the similarity *between* MWEs rather than between MWEs and unigram opinion words.

We add a further **novel** constraint that is derived from Wiktionary. Several Wiktionary-entries may be connected via *Synonymy*-links and *See-also*-links (i.e., near-synonyms). As Figure 1 shows, these links occurring on Wiktionary-entries also tend to be MWEs. Moreover, all these words tend to convey the same sentiment view. For example, the MWEs from Figure 1 (i.e., *know one's stuff*, *know one's onions*, and *know one's way around*) all convey a speaker view. Our constraint *CONSTR-wikt* (Table 12) demands that MWEs that are linked via *Synonymy*- or *See-also*-links should possess the same sentiment view.

5.3. *Supervised baseline classifiers (SVM and BERT)*

As supervised baseline classifiers, we examine Support Vector Machines (SVM) (Joachims 1999) and BERT (Devlin *et al.* 2019). Both of these classifiers differ from MLN in that they cannot incorporate global constraints.

SVM are an efficient and robust traditional learning method. They are trained on the features we devised for this task (Section 5.1). As a tool, we used *SVM*^{light}.[t] The tuning of hyperparameters is only critical if the data on which the classifier is trained has a fairly skewed class distribution. Since this is not the case with our dataset (41%:59%, Table 5), we use this tool with its standard configuration.

Transformers, such as BERT, are currently the most advanced supervised classifiers. They produce a word representation that takes context into account. This property may also be useful for

[s] In order to avoid overfitting we took this value from Wiegand *et al.* (2016) who used it for their verb constraint.
[t] http://svmlight.joachims.org

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

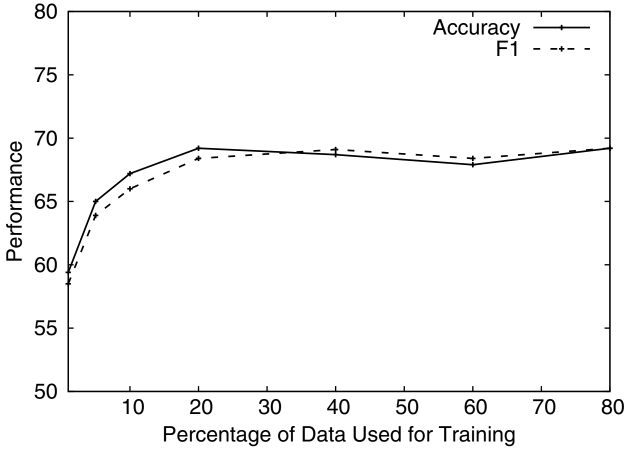


Figure 6. Learning curve using MLN, the best supervised classifier.

the classification of MWEs since a representation of words forming an MWE should not consider the words in isolation (as traditional word embeddings do) but with respect to the other words that are part of the MWE. We train BERT on the sequence of tokens that an MWE comprises. As a model, we take the BERT-Large model (Cased: 24-layer, 1024-hidden, 16-heads, 340 M parameters). Since we only have little labeled training data we fine-tune the model by adding a layer on top of the pre-trained BERT-Large model. Since we do not want to overfit the classifier, we run the model in its standard configuration (batch size: 32, learning rate: 5e-5, number of epochs: 3).

5.4. Experiments using supervised classification

In our experimental set-up, we largely follow Wiegand *et al.* (2016). The type of features that we use are, like the features from Wiegand *et al.* (2016), frequently occurring features. Such features typically only require a small amount of labeled training data. This property is essential for the task that we examine in this article. If we established that such a task was only feasible by using the greatest part of the lexicon as training data (and therefore have it manually annotated), there would be little benefit in *learning* this task. Instead, one could stick to the one-time effort of manually annotating the *entire* lexicon.

Following Wiegand *et al.* (2016) we just use 20% of our dataset (Section 3.2) for training a supervised classifier. The remaining 80% are used for testing. We do this experiment 10 times using different random partitions of training data and report the average performance over those experiments. As in our experiments from Section 4, we report accuracy and macro-average precision, recall, and F-score.

In order to show that the small proportion of training data proposed by Wiegand *et al.* (2016) is also sufficient for our new dataset, Figure 6 shows a learning curve in which we vary the proportion of data used as training data. (80% is the largest possible training set since we require 20% of the dataset to be used as test data.) We used the best supervised classifier (i.e., MLN) from our experiments which we will detail in the remainder of this section. The figure clearly indicates that beyond using 20% of the dataset for the training, the classifier no longer systematically improves.[u] Thus, we have shown that the experimental set-up from Wiegand *et al.* (2016) (i.e., using 20% of the dataset for training) is also a suitable setting for our dataset comprising MWEs.

[u] We also tested various amounts of training data for SVM, that is, the learning algorithm for which we obtained the second-best performance in our evaluation. We obtained similar results to those we obtained for MLN, so we assume that the learning curve in Figure 6 is fairly representative.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 13. Comparison of features groups

	MLN			
Configuration	Acc	Prec	Rec	F1
majority	58.77	29.39	50.00	37.02
USAGE	44.09	45.91	46.33	46.12*
POLAR	47.68	49.85	49.83	49.84*
POS	55.34	50.90	51.94	51.41*
PATT	58.83	50.29	52.72	51.47*
SUBC	56.29	56.25	55.62	55.93*
STRUC	55.01	56.79	56.36	56.58*
WN	60.23*	58.83	58.60	58.71*
FN	62.27*	61.29	61.61	61.45*
ULEX	**65.87***	**65.74**	**65.97**	**65.85***

*Significantly better than **majority** using a paired *t*-test at $p < 0.05$.

Table 14. Comparison of representation foci

	MLN			
Focus	Acc	Prec	Rec	F1
WIKT	56.5	58.4	58.1	58.2
MWE	64.6	63.4	62.9	63.2
CORP	64.4	64.8	65.0	64.9
WIKT + MWE	65.1	64.1	63.9	64.0
WIKT + CORP	66.4*	66.2	66.7	66.4*
CORP + MWE	67.6*	66.8	66.6	66.7*
WIKT + CORP + MWE	**68.5***†	**67.7**	**67.5**	**67.6***†

Statistical significance testing (paired *t*-test at $p < 0.05$): *better than CORP (i.e., best individual focus); †better than CORP + MWE (i.e., best pair).

Table 13 shows the performance of individual feature groups (Table 11) using MLN (Section 5.2). As a baseline, we use a majority-class classifier. With respect to F-score, all features outperform that baseline. *WN*, *FN*, and *ULEX* also outperform it with regard to accuracy. The most effective feature group is *ULEX*. This means that for sentiment-view classification of MWEs, the knowledge of sentiment views of unigrams is most helpful.

Table 14 displays the performance of the different representation foci (Table 10). The weakest focus is WIKT, the strongest is CORP. The table also examines all possible combinations of those foci. It shows that the information of the different representation foci is, to some extent, complementary. Even though WIKT is the weakest focus, it always helps to increase performance when added to another focus.

It is worth noting that MWE represents a fairly strong representation focus. Given that many of these features (Table 11) consider individual tokens as a proxy of their respective MWE,

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 15. Comparison of different classifiers

Configuration	Acc	Prec	Rec	F1
majority	58.8	29.4	50.0	37.0
MLN with Wiegand 2016-features	58.3	55.2	53.9	54.5
BERT (trained on sequence of tokens)	62.9	59.0	62.0	60.5
label propagation (LIU)	63.1	68.5	60.4	64.2
MLN with best indiv. feature group (i.e., *ULEX*)	65.9	65.7	66.0	65.9
SVM with all features (i.e., `WIKT` + `CORP` + `MWE`)	68.2	67.9	65.2	66.5
MLN with all features (i.e., `WIKT` + `CORP` + `MWE`)	68.5	67.7	67.5	67.6^{*}
MLN with all features + *CONSTR-distr*	68.7	67.9	67.8	67.8^{*}
MLN with all features + all constraints	$69.2^{*\dagger}$	68.5	68.3	$68.4^{*\dagger}$
MLN with all feat. + all constr. + label prop. (LIU)	$69.5^{\dagger}$	69.0	69.2	$69.1^{*\dagger}$
MLN with all feat. + all constr. + label prop. (all corpora)	**69.8***†	**69.4**	**69.8**	**69.6***‡†

Statistical significance testing (paired *t*-test at $p < 0.05$): *better than **SVM with all features**; †better than **MLN with all features**; ‡better than **MLN with all features + all constraints.**

we conclude that sentiment-view information of MWEs is compositional to a large degree. We even observed compositionality of sentiment-view information on several idiomatic MWEs, whose *meaning* is non-compositional. For example, the speaker-view idioms *jump the shark* or *raise the devil* both contain a speaker-view noun, that is *shark* and *devil.*[v] Given this observation, we conclude the sentiment view of an MWE can be computed in a compositional manner more effectively than its meaning. However, we want to emphasize that the classification of sentiment views cannot be completely solved by such a compositional approach. For example, the MWE *have a good time* is an actor-view word (which can be regarded as a synonym of *relax*[w]). However, if we derived the sentiment view from the adjective *good* which the MWEs contained, we would produce a wrong classification since *good* is a speaker-view adjective.

Table 15 compares further classifiers. We also consider the subset of our features (Table 13) that could be literally taken from Wiegand *et al.* (2016) (i.e., *PATT*, *FN-direct*, *WN-direct*) and applied to MWEs. Those features may largely outperform the majority-class classifier, but already the most generic classifier BERT outperforms that classifier by a large degree. The best label propagation (using LIU) also outperforms BERT which means that that distributional similarity of opinion words to MWEs is very predictive.

Table 15 also shows that the strongest individual feature group, that is *ULEX*, can be improved by adding the remaining features. On top of that, a further performance increase is obtained by applying our global constraints (Table 12). The *combination* of all global constraints (i.e., *CONSTR-distr* and *CONSTR-wikt*) is most effective and yields better results than only the distributional constraint (*CONSTR-distr*), which is the only constraint we originally took from Wiegand *et al.* (2016). SVM are roughly on a par with MLN using no global constraints. MLN using all global constraints also outperform SVM.

[v] Both *shark* and *devil* do not convey their literal meaning in these idioms. However, they both preserve their sentiment view.

[w] We consider *relax* a typical emotion verb. In Section 1.1, we explained why such opinion expressions are considered to convey an actor view.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

We can further increase F-score by adding the prediction of label propagation (Section 4). However, only if we include the prediction of all three corpora do we obtain a further significant improvement. This means that the information in the different corpora is complementary.

Overall, the fairly low performance of BERT may come as a surprise. However, we offer two major reasons that explain this result: First, in this article, we deal with a task in which the instances to be classified, that is our MWEs, only comprise very few tokens, that is about 3 on average (Table 5). Therefore, BERT has only very few observations on the basis of which it has to make a prediction. Typically, BERT is applied to much longer sequences of words, that is full sentences or even a list of sentences (Devlin *et al.* 2019). Second, the strong performance of our feature-based approach (i.e., MLN) suggests that predictive information for this task is contained in various resources (e.g., unigram sentiment-view lexicon, Wiktionary, subcategorization lexicons, WordNet, FrameNet, etc.). BERT, on the other hand, by design, is typically trained on generic features, in our case: the sequence of word tokens that comprise the labeled MWEs. Obviously, the information contained in these features is limited and insufficient for producing reasonable performance for the task at hand. It is not competitive with the rich information contained in the various resources that we included in our feature-based approach.

5.5. Error analysis of sentiment-view classification

In this subsection, we report on the error analysis we conducted based on the output of our best classifier for the classification of sentiment views (i.e., *MLN with all feat. + all constr. + label prop. (all corpora)* in Table 15). Two major error types are possible: either an MWE actually conveying an actor view is predicted to convey a speaker view or an MWE actually conveying a speaker view is predicted to convey an actor view. Based on the output of our best classifier, both types of errors occur at a similar rate. The former accounts for 44.7% of the errors whereas the latter accounts for 55.3%. Therefore, we consider both types of errors in our error analysis.

5.5.1. Actor views miscategorized as speaker views

Our analysis of the internal structure of MWEs in Section 5.1.6 suggests that while MWEs conveying an actor view tend to be light-verb constructions, MWEs conveying a speaker view are rarely such constructions and, therefore, tend to be idiomatic MWEs. We observed that concrete nouns are often a part of the latter MWEs (e.g., *grease the wheels*, *play the second fiddle*, or *twist the knife*), possibly since idiomatic constructions often include figurative language. As a consequence, our best classifier considers the mention of such concrete nouns to be a strong predictor for speaker views. However, there are also actor-view MWEs that contain concrete nouns. Several of them tend to be misclassified as speaker views, such as *run for the hills*, *test the waters*, *throw in the sponge*, or *smell a rat*.

As pointed out in Section 5.1.5, among MWEs with a negative polarity there are only comparatively few MWEs conveying an actor view (Figure 3). As a result of this polarity bias, many negative actor-view MWEs, such as *go on strike*, *see red* or *wage war against*, are mistakenly categorized as conveying a speaker view.

5.5.2. Speaker views miscategorized as actor views

We observed a bias for MWEs containing light verbs, for example *have*, *make*, *take*, to be categorized as actor-view words. This is because many MWEs being light-verb constructions convey an actor view (Figure 4). For instance, among the MWEs containing the light verb *take* in our dataset, almost 70% convey an actor view. This number is particularly significant since, in general, the MWEs conveying an actor view represent 41.5% of the minority class in our dataset (Table 5). While many MWEs being light-verb constructions convey an actor view, there are also light-verb

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

constructions that convey a speaker view. However, many of them, such as *make trouble* or *take the liberty*, get misclassified.

6. Extrinsic evaluation: opinion role extraction

In this section, we examine the impact of sentiment-view classification of MWEs on the task of opinion role extraction. With the examples (26) and (27) from Section 1, we already illustrated how the knowledge of sentiment views may help on this task. We are not aware of any publicly available tool conducting opinion target extraction that we could use as a baseline. Therefore we leave aside the task of opinion target extraction and only focus on opinion holder extraction. Since neither the MPQA corpus (Section 1) nor any other publicly available dataset annotated with opinion role information contains a sufficient number of verbal MWEs, we had to create a new dataset for this particular task. Section 6.1 introduces this new dataset. This is followed by a description of the different classifiers we compare (Section 6.2). In Section 6.3, we present the results of our experiments and conclude with an error analysis in Section 6.4.

6.1. Data and annotation for opinion holder extraction

We sampled from the NEWS corpus (Section 3.4) sentences with mentions of the verbal MWEs of our gold standard (Section 3.2). We chose the news domain since it typically displays a high number of different opinion holders.[x] We sampled in such a way that each sentence contains a different MWE. In total, our new dataset comprises 1167 sentences.[y]

For each sentence, the verbal MWE contained in it was presented to our annotator in addition to the sentence itself. Only the opinion holders of verbal MWEs were to be annotated. Opinion holders of any further opinion expressions other than MWEs were ignored. The reason for this is that we want to exclusively evaluate the performance on verbal MWEs. Due to the ambiguity of opinion expressions (Akkaya, Wiebe and Mihalcea 2009), some MWE mentions do not convey any subjectivity. This typically concerns MWEs that can be used both literally (in which case they usually do not convey subjectivity) and figuratively (in which case they usually do convey subjectivity). Examples are *catch dust*, *go to town* or *pull the plug*. As a consequence, no opinion holders are evoked in sentences containing MWEs with non-subjective usage. In order to maintain a realistic evaluation of classification performance, we kept those sentences in our dataset and if an opinion holder was extracted from these sentences by a classifier, then this was counted as an error. 6.9% of our sentences contain verbal MWEs in a context in which they do not display any subjectivity. In other words, MWEs, such as *catch dust*, are predominantly used in a figurative sense. The strong bias towards figurative (i.e., subjective) senses is also in line with the results of our corpus-based study from Section 1. We ascribe it to the fact that MWEs tend to be less ambiguous than unigrams (Section 3.2).

In order to verify the reliability of our annotation of opinion holders, we computed an interannotation agreement between the main annotator and the first author of this article on a sample of 200 sentences. We measured a substantial interannotation agreement of Cohen's $\kappa = 0.76$. It is also much higher than the agreement we achieved for annotating our MWE gold standard lexicon (Section 3.2). This can be explained by the fact that this in-context annotation task, in general, is easier than the previous out-of-context task. This is also supported by the fact that we did not identify any notable issues responsible for the few remaining instances in which the above two annotators disagreed.

[x] Despite its high concentration of sentiment, LIU does not qualify as source for this task since on product reviews, opinion holder extraction is not relevant—the overwhelming majority of opinion holders represent the writers of the corresponding reviews.

[y] Some MWEs of our gold standard are not contained in NEWS.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Table 16. Derivation of opinion holders from sentiment views

Senti. View	Holder	Example
actor view	agent	[Trump]$^{agent}_{Holder}$ **draws the line**$_{actor\text{-}view}$ at gay marriage and abortion.
speaker view	*implicit*	[A lot of people]^{agent} **reinvent the wheel**$_{speaker\text{-}view}$ with e-commerce.

6.2. Classifiers for opinion holder extraction

As a baseline classifier, we consider ***MultiRel*** from Johansson and Moschitti (2013). It is currently the most sophisticated opinion holder extractor that is publicly available. *MultiRel* incorporates relational features taking into account interactions between multiple opinion cues. It has been trained on the MPQA corpus. *MultiRel* has only very little knowledge of verbal MWEs since they are rare on MPQA (Section 1). However, it is also a context-based classifier. As a consequence, it may potentially also detect unknown opinion expressions (i.e., words or phrases that have not been observed in the training data, i.e., the MPQA corpus) and identify their respective holders, provided the context of these expressions is sufficiently indicative.

The second baseline, ***BERT*** is a transformer fine-tuning the pre-trained BERT-Large model (Devlin *et al.* 2019) on MPQA. (We also made exploratory experiments with LSTM-CRFs (Marasović and Frank, 2018) which had been previously proposed for opinion holder extraction but BERT outperformed that sequential classifier on our data by a large degree.) We also mark the opinion expressions in the input, which in the case of our new corpus correspond to MWEs. We encode this information as *positional markers* (Zhang and Wang 2015) (e.g., *His latest remarks only <**oe**>added fuel to the fire</**oe**>*.). Thus, unlike *MultiRel*, *BERT* knows about the presence of the MWE in each input sentence but, like *MultiRel*, it does not know their sentiment view.

Our third baseline, ***SpeakerHolder*** is also aware of all the verbal MWEs from our MWE gold standard lexicon (Section 3.2). Like *MultiRel* and *BERT*, this classifier does not include any specific knowledge of the sentiment views the MWEs convey. *SpeakerHolder* simply always assigns the opinion holder as the implicit speaker of the utterance. We consider the implicit speaker since on our set of verbal MWEs (Table 5) the MWEs conveying a speaker view were more frequent than those MWEs conveying an actor view. Opinion expressions conveying a speaker view typically have the implicit speaker of the utterance as their opinion holder (27). So, this is the best *a priori* guess we can make.

View, like *BERT* and *SpeakerHolder*, possesses knowledge of all the verbal MWEs from our gold standard but it also incorporates knowledge about the sentiment views that those MWEs convey. However, it does not take that knowledge from the gold standard lexicon (Section 3.2) but the output of the best learning method from Section 5.4 (Table 15). We do so since we want to have a realistic estimate of the impact of having knowledge of sentiment views on opinion holder extraction. The derivation of opinion holders from sentiment views is illustrated in Table 16. In the case of actor views the opinion holder is the agent of the verbal MWE. (Agents represent by far the most frequent argument position of explicit opinion holders (Bethard *et al.* 2006).) In case of speaker views, there is no (explicit) opinion holder.

Table 17 summarizes the different classifiers we consider. All classifiers except for *MultiRel* and *BERT* are rule-based classifiers. *View* is also a rule-based classifier but it uses the output of a statistical sentiment-view classifier which has been trained on out-of-context data.

All sentences from our dataset may serve as test instances. In order to make a fair evaluation of *View*, we have to exclude those sentences from the test set in which those MWEs occur that have been used for training the sentiment-view classifier it incorporates. (In other words, *View* should incorporate a sentiment-view classifier which makes predictions for *unseen* MWEs.) Since we actually evaluated 10 different sentiment-view classifiers in Section 5.4 with different training instances (and reported their average performance), we also evaluate 10 different extractors

Table 17. Summary of different classifiers used for opinion holder extraction

	Classifier			
Property	MultiRel	BERT	SpeakerHolder	View
Has been trained on MPQA?	√	√		
Has *explicit* knowledge of MWEs?		√	√	√
Has knowledge about sentiment views of MWEs?				√

Table 18. Comparison of different opinion holder extraction systems

Classifier	Prec	Rec	F1
MultiRel	60.49	11.72	19.64
SpeakerHolder	42.57	40.07	41.28
View	61.55	48.40	54.19*
BERT	66.00	50.74	57.37*
BERT + View	65.41	64.36	64.88*†

Statistical significance testing (paired *t*-test at $p < 0.05$): *better than *SpeakerHolder*; †better than *BERT*.

for *View* where each time another of those sentiment-view classifier instances is incorporated. In order to produce a meaningful comparison, the other classifiers (i.e., *MultiRel*, *BERT*, and *SpeakerHolder*) will be evaluated on exactly the same 10 test sets. Similar to Section 5.4, we will report the average performance.

In this experimental set-up, we do not consider learning from our novel dataset in which verbal MWEs are annotated in context with their respective opinion holders. This design choice is deliberate and it also reflects a realistic scenario. As we have pointed out in Section 1, even though they occur regularly, MWEs are much less frequent than unigrams. Due to that sparsity, even if we annotated much larger portions of *contiguous* text than are contained in MPQA and thus obtained a significant number of MWE mentions, in general, each *individual* MWE would still most likely be only observed once. (Such distributional behavior of MWEs, in general, was also reported in Schneider *et al.* (2014b).) Unfortunately, with a dataset comprising mostly singleton MWEs, we cannot train a supervised classifier that learns characteristic syntactic relations between each individual MWE and their respective opinion holders.

6.3. Results for opinion holder extraction

Table 18 shows the performance of the different classifiers on opinion holder extraction. *MultiRel* is the worst performing classifier. It performs so poorly since it can hardly detect any MWEs and is therefore unable to predict their opinion holders. *SpeakerHolder* has a notably higher F-Score than *MultiRel* despite the fact that it always assigns the opinion holder to the implicit speaker of the utterance. This can be explained by the fact that this type of opinion holder is the most frequent type in our dataset. However, its lacking discriminative power between different types of opinion holders is reflected by its low precision, which is the worst of all classifiers examined in this evaluation.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

BERT is a strong baseline. Unlike for the task of sentiment-view classification in Section 5.4, *BERT* benefits from the fairly large amount of labeled training data (about 10,000 labeled sentences) available for opinion holder extraction in the MPQA corpus. Apparently, the classifier can transfer knowledge from MPQA, in which hardly any MWEs are contained, to the MWEs in our new corpus. Our observation is that it learns likely opinion holders irrespective of their specific opinion expression. For example, personal pronouns *he* or *we* at the beginning of a sentence have a high likelihood to be opinion holders, in general. In the absence of those likely opinion holders in a sentence, *BERT* predicts an implicit opinion holder or none at all. *View*, however, reasons on the basis of the sentiment view of the MWE in a sentence. It detects 40% more explicit opinion holders than *BERT* is able to detect. Still, *BERT* detects explicit opinion holders that *View* does not detect, namely holders that are not an agent of the MWE. Given these complementary capabilities, we combine *BERT* and *View* by adding the explicit opinion holders detected by *View* to the *BERT* output. This combination significantly outperforms *BERT* and thus underscores the importance of sentiment views.

6.4. Error analysis of opinion holder extraction

We manually inspected the output of our best opinion holder extraction system (i.e., the combination of *BERT* and *View*) in order to identify systematic errors made by that classifier. Though that model has the best balance between precision and recall, with a score of about 64–65% it still evidently lacks in both correctness and coverage. Apart from false positives caused by the few sentences in our dataset that convey non-subjective use of the given MWEs (as already discussed in Section 6.1), we identified the following patterns in the misclassification:

As pointed out in the previous section, *BERT* often recognizes explicit opinion holders based on noun phrases that were often observed as opinion holders in the MPQA corpus, for example *he* or *we* at the beginning of a sentence. Therefore, unlike *View*, *BERT* may also detect explicit opinion holders that are no agents of the MWE if that explicit opinion holder coincides with an opinion holder frequently observed in the MPQA corpus. While we observed only few cases in which this behavior produces false positives, we found that this also causes certain words or phrases not to be considered as opinion holders. This may produce some false negatives. For example, the personal pronoun *it* is often not recognized as an opinion holder, as in (48), presumably since it too often represents inanimate entities who are not eligible to represent opinion holders.

(48) But [it]$_{\textit{actual Holder}}$ hasn't yet **made up its mind**$_{\textit{actor-view}}$ that it wants to join.

With regard to *BERT*, there may also be context words which are learned to trigger opinion holders. For example, in the MPQA corpus, the verbal predicate *said* precedes an opinion holder in most cases. We noticed that this causes several false positives in our test data, such as (49).

(49) [Arlacchi]$_{\textit{predicted Holder}}$ said it was the first time [Riina]$^{\textit{agent}}_{\textit{actual Holder}}$ had **named names**$_{\textit{actor-view}}$.

This is notable insofar as we actually indicate the opinion expression in a sentence for which an opinion holder is to be extracted in both training and testing by a positional marker (Section 6.2). In (49), this opinion expression is the MWE *named names*. However, despite this form of prompting which should suggest the noun phrase *Riina* to be a more plausible opinion holder since it is the agent of that MWE, *BERT* mostly predicts noun phrases preceding the verbal predicate *said* as opinion holders.

The second component of our combined classifier, that is *View*, is more precise when it comes to the extraction of explicit opinion holders. However, since it is based on a lexicon whose distinction between actor and speaker views is far from perfect (according to our evaluation in

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Section 5.4, it achieves an F-score 69.6%), it also adds some noise to our opinion holder extraction system. In most cases, when the lexicon predicts an incorrect sentiment view, this also means that an incorrect opinion holder is extracted from a given sentence. However, even a perfect lexicon would not guarantee a perfect opinion holder extraction system. As outlined in Table 16, for MWEs with an actor view we consider as opinion holder the agent of the respective MWE. Our manual inspection of the system output revealed that there is a handful of MWEs which realize the opinion holder in a different argument position, such as (50) and (51).

(50) But he indicated that it would be acceptable to Congress if [it]^{agent} **passes muster**$_{actor\text{-}view}$ [with parents' groups]$_{Holder}$.

(51) Though Sartori has his critics, [he]^{agent} is generally **held in high esteem**$_{actor\text{-}view}$ [in Italy]$_{Holder}$.

For our combined classifier, these instances of opinion holders are very unlikely to be extracted since on the MPQA corpus (i.e., the dataset on which *BERT* has been trained), explicit opinion holders that appear in an argument position other than the agent are very rare (Wiegand and Ruppenhofer 2015; Zhang, Liang and Fu 2019). Therefore, in most cases, the BERT-based component will not detect these opinion holders either.

7. Conclusion

Verbal MWEs are important for sentiment analysis because a very large proportion of those MWEs convey sentiment. We assessed different methods for the novel task of classifying the sentiment view conveyed by (verbal) MWEs. MWEs occur regularly in written texts but since they are considerably less frequent than unigrams, labeled corpora for sentiment analysis only include very few distinct MWEs. As a consequence, knowledge regarding MWEs cannot be directly learnt from those corpora. We presented an approach which considered MWEs out of context.

By just applying simple label propagation involving an existing unigram sentiment lexicon annotated with sentiment-view information and no labeled MWEs, a reasonable classification of MWEs can be obtained already outperforming BERT trained on the sequence of word tokens which an MWE comprises. We found, however, that in order to obtain the best possible performance for that approach, the corpus from which the underlying similarity graph is generated is vital. That corpus needs to be large and contain a high concentration of sentiment information.

Beyond the label propagation approach, we found that even better classifiers can be obtained by a supervised learning approach that exploits a limited number of labeled MWEs and features specifically tailored to the task at hand. Similar to the sentiment-view classification on unigram opinion expressions, features based on surface patterns, subcategorization, polarity, and the two lexical resources FrameNet and WordNet are beneficial. For most of the features that rely on some lexicon lookup, unlike with unigram opinion expressions, we cannot look up MWEs in these respective lexicons in the same fashion since they are too sparsely represented. Instead, we looked up individual tokens comprising the MWEs in those resources or drew information from Wiktionary. We established that Wiktionary contains a significantly larger number of verbal MWEs than the previously examined lexical resources. Therefore, in Wiktionary, many MWEs can be looked up directly. In the case of subcategorization features, instead of consulting a lexical resource holding that information we directly extracted subcategorization information from a large corpus.

We also proposed entirely novel features for MWE classification. For instance, we addressed the internal structure of the MWEs. Both the awareness of light-verb constructions and the token length of MWEs are predictive for this task. We also inspected global constraints as yet another type of information and obtained further increases in classification performance by incorporating

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

distributional similarity and semantic similarities as encoded in Wiktionary. Best classification performance was achieved by including various types of features, global constraints, and the output of label propagation.

Finally, we also carried out an extrinsic evaluation by incorporating the knowledge of sentiment views in a rule-based opinion holder extraction system. This approach can improve state-of-the-art opinion holder extraction based on BERT. We also explained how our rule-based classifier based on sentiment views and a statistical classifier based on BERT complement each other.

In this article, we exclusively examined verbal MWEs. However, we also showed that for that analysis not only knowledge about verbs is relevant. In our graph-based experiments, we achieved better results by employing a similarity graph that includes not only unigram opinion verbs but also opinion nouns and adjectives. Opinion adjectives generally play an important role for sentiment-view classification since on this part of speech the distribution is highly skewed towards speaker views. As a consequence, the information whether an MWE is adjective-like also proved to be beneficial for this classification task. This subset of MWEs, like adjectives, is much more likely to convey a speaker view.

We believe that our research also produced some insights that can be of value in other semantic classification tasks on (verbal) MWEs. For instance, we showed that the most effective individual method for MWE classification is inferring categories of MWEs from categories of similar unigram expressions. For other classification tasks, we would therefore recommend harnessing existing unigram resources that have been labeled for the same task. Moreover, we assume that the problem of sparsity of MWEs in common lexical resources will be similarly present in other classification tasks. We also believe that the alternatives we proposed (e.g., using Wiktionary as a resource with a much larger coverage of MWEs or decomposing MWEs into tokens and looking up that information in the established lexical resources) should also be equally applicable. Finally, the issues of corpus coverage and vector representation of MWEs could be similarly relevant for other tasks.

As part of future work, we may consider how our approach can be applied to other languages. Simply relying on machine translation is unlikely to be a satisfactory solution since the translation of MWEs, in general, is a hard problem in NLP (Constant *et al.* 2017). Therefore, it may be more worthwhile to replicate our proposed method in the target language. However, given that all other languages tend to be less well resourced than English, we envisage that the replication of our proposed method will require certain modifications to compensate for the scarcity of resources.

Acknowledgements. The authors would like to thank Stephanie Köser for contributing to the annotation of the dataset presented in this article. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References

Akkaya C., Wiebe J. and Mihalcea R. (2009). *Subjectivity word sense disambiguation.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, pp. 190–199.

Andreevskaia A. and Bergler S. (2006). *Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses.* **In** *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, pp. 209–216.

Baker C.F., Fillmore C.J. and Lowe J.B. (1998). *The Berkeley FrameNet project.* **In** *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Montréal, Quebec, Canada, pp. 86–90.

Baldwin T. and Kim S.N. (2010). Multiword expressions. In Indurkjya N. and Damerau F. J. (eds), *Handbook of Natural Language Processing*. Boca Raton, FL: CRC Press, Taylor and Francis Group, pp. 267–292.

Baroni M., Bernardini S., Ferraresi A. and Zanchetti E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* **43**(3), 209–226.

Beigman Klebanov B., Burstein J. and Madnani N. (2013). Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds. *ACM Transactions on Speech and Language Processing* **10**(3), 12.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Beigman Klebanov B., Leong C.W. and Flor M. (2015). *Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples.* **In** *Proceedings of the Workshop on Metaphor in NLP*, Denver, CO, USA, pp. 11–20.

Bethard S., Yu H., Thornton A., Hatzivassiloglou V. and Jurafsky D. (2006). Extracting opinion propositions and opinion holders using syntactic and lexical cues. **In** *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht: Springer-Verlag, pp. 125–141.

Breck E., Choi Y. and Cardie C. (2007). *Identifying expressions of opinion in context.* **In** *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, pp. 2683–2688.

Chen W.-T., Bonial C. and Palmer M. (2015). *English light verb construction identification using lexical knowledge.* **In** *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Austin, TX, USA, pp. 2375–2381.

Choi Y. and Wiebe J. (2014). *+/–EffectWordNet: Sense-level lexicon acquisition for opinion inference.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1181–1191.

Constant M., Eryiğit G., Monti J., van der Plas L., Ramisch C., Rosner M. and Todirascu A. (2017). Multiword expression processing: A survey. *Computational Linguistics* **43**(4), 837–892.

Constant M., Sigogne A. and Watrin P. (2012). *Discriminative strategies to integrate multiword expression recognition and parsing.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju Island, Korea, pp. 204–212.

Cowie A.P., Mackin R. and McCaig I. (eds) (1983). *Oxford Dictionary of Current Idiomatic English: Phrase, Clause and Sentence Idioms*. Oxford: Oxford University Press.

Deng L. and Wiebe J. (2016). *Recognizing opinion sources based on a new categorization of opinion types.* **In** *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, New York City, NY, USA, pp. 2775–2781.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* **In** *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, Minneapolis, MN, USA, pp. 4171–4186.

Esuli A. and Sebastiani F. (2005). *Determining the semantic orientation of terms through gloss classification.* **In** *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, Bremen, Germany, pp. 617–624.

Flekova L. and Gurevych I. (2016). *Supersense embeddings: A unified model for supersense interpretation, prediction, utilization.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, pp. 2029–2041.

Green S., de Marneffe M.-C. and Manning C.D. (2013). Parsing models for identifying multiword expressions. *Computational Linguistics* **39**(1), 195–227.

Grishman R., McKeown C. and Meyers A. (1994). *COMLEX syntax: Building a computational lexicon.* **In** *Proceedings of the International Conference on Computational Linguistics (COLING)*, Kyoto, Japan, pp. 268–272.

Hashimoto C. and Kawahara D. (2008). *Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii, USA, pp. 992–1001.

Jackendoff R.S. (1997). *The Architecture of the Language Faculty. Linguistic Inquiry Monographs*. Cambridge, CA, MIT Press.

Jindal N. and Liu B. (2008). *Opinion spam and analysis.* **In** *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, Palo Alto, CA, USA, pp. 219–230.

Joachims T. (1999). Making large-scale SVM learning practical. In Schölkopf B., Burges C. and Smola A. (eds), *Advances in Kernel Methods - Support Vector Learning*. Cambridge, CA, MIT Press, pp. 169–184.

Jochim C., Bonin F., Bar-Haim R. and Slonim N. (2018). *SLIDE – A sentiment lexicon of common idioms.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, pp. 2387–2392.

Johansson R. and Moschitti A. (2013). Relational features in fine-grained opinion analysis. *Computational Linguistics* **39**(3), 473–509.

Kang J. S., Feng S., Akoglu L. and Choi Y. (2014). *ConnotationWordNet: Learning connotation over the word+sense network.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, USA, pp. 1544–1554.

Kato A., Shindo H. and Matsumoto Y. (2018). *Construction of large-scale English verbal multiword expression annotated corpus.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, pp. 2495–2499.

Klein D. and Manning C.D. (2003). *Accurate unlexicalized parsing.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, pp. 423–430.

Kozareva Z. (2013). *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 682–691.

Kuiper K., McCann H., Quinn H., Aitchison T. and van der Veer K. (2003). *A Syntactically Annotated Idiom Database (SAID) v.1*. Documentation to a LDC Resource.

Lakoff G. and Johnson M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy* **77**(8), 453–486.

Landis J.R. and Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174.

Levin B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, University of Chicago Press.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Liebeskind C. and HaCohen-Kerner Y. (2016). *A lexical resource of Hebrew verb-noun multi-word expressions.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, pp. 522–527.

Lin D. (1998). *Automatic retrieval and clustering of similar words.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL/COLING)*, Montreal, Quebec, Canada, pp. 768–774.

Liu J. and Seneff S. (2009). *Review sentiment scoring via a parse-and-paraphrase paradigm.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, pp. 161–169.

Long T.H. (ed) (1979). *The Longman Dictionary of English Idioms*. Harlow: Longman.

Macleod C., Grishman R., Meyers A., Barrett L. and Reeves R. (1998). *NOMLEX: A lexicon of nominalizations.* **In** *Proceedings of EURALEX*, Liège, Belgium, pp. 187–193.

Maks I. and Vossen P. (2012a). *Building a fine-grained subjectivity lexicon from a web corpus.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, pp. 3070–3076.

Maks I. and Vossen P. (2012b). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems* **53**(4), 680–688.

Marasović A. and Frank A. (2018). *SRL4ORL: Improving opinion role labeling using multi-task learning with semantic role labeling.* **In** *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, New Orleans, LA, USA, pp. 583–594.

Maudslay R.H., Pimentel T., Cotterell R. and Teufel S. (2020). *Metaphor detection using context and concreteness.* **In** *Proceedings of the Workshop on Figurative Language Processing*.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). *Efficient estimation of word representations in vector space.* **In** *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA.

Mikolov T., Grave E., Bojanowski P., Puhrsch C. and Joulin A. (2018). *Advanced in pre-training distributed word representations.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, pp. 52–55.

Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K. (1990). Introduction to wordNet: An on-line lexical database. *International Journal of Lexicography* **3**(4), 235–244.

Moilanen K. and Pulman S. (2007). *Sentiment construction.* **In** *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Moreno-Ortiz A., Pérez-Hernández C. and Del-Olmo M.Á. (2013). *Managing multiword expressions in a lexicon-based sentiment anaylsis system for spanish.* **In** *Proceedings of the Workshop on Multiword Expressions (MWE)*, Altanta, GA, USA, pp. 1–10.

Nunberg G., Sag I.A. and Wasow T. (1994). Idioms. *Language* **70**(3), 491–538.

Quirk R., Greenbaum S., Leech G. and Svartvik J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ramisch C., Cordeiro S.R., Savary A., Vincze V., Mititelu V.B., Bhatia A., Buljan M., Candito M., Gantar P., Giouli V., Güngör T., Hawwari A., Inurrieta U., Kovalevskaitė J., Krek S., Lichte T., Liebeskind C., Monti J., Escartín C.P., QasemiZadeh B., Ramisch R., Schneider N., Stoyanova I., Vaidya A. and Walsh A. (2018). *Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions.* **In** *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG)*, Santa Fe, NM, USA, pp. 222–240.

Riedel S. (2008). *Improving the accuracy and efficiency of MAP inference for Markov Logic.* **In** *Proceedings of the Annual Conference on Uncertainty in AI (UAI)*, Helsinki, Finland, pp. 468–475.

Schneider N., Danchik E., Dyer C. and Smith N.A. (2014a). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* **2**, 193–206.

Schneider N., Hovy D., Johannsen A. and Carpuat M. (2016). *SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM).* **In** *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, San Diego, CA, USA, pp. 558–571.

Schneider N., Onuffer S., Kazour N., Danchik E., Mordowanec M.T., Conrad H. and Smith N.A. (2014b). *Comprehensive annotation of multiword expressions in a social web corpus.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 455–461.

Shutova E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics* **41**(4), 579–623.

Socher R., Perelygin A., Wu J.Y., Chuang J., Manning C.D., Ng A.Y. and Potts C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, WA, USA, pp. 1631–1642.

Strzalkowski T., Shaikh S., Cho K., Broadwell G.A., Feldman L., Taylor S., Yamrom B., Liu T., Cases I., Peshkova Y. and Elliot K. (2014). *Computing affect in metaphors.* **In** *Proceedings of the Workshop on Metaphor in NLP*, Baltimore, MD, USA, pp. 42–51.

Talukdar P.P., Reisinger J., Pasca M., Ravichandran D., Bhagat R. and Pereira F. (2008). *Weakly-supervised acquisition of labeled class instances using graph random walks.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, HI, USA, pp. 582–590.

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press

Taslimipoor S., Rohanian O., Mitkov R. and Fazly A. (2017). *Investigating the opacity of verb-noun multiword expression usages in context.* **In** *Proceedings of the Workshop on Multiword Expressions (MWE)*, Valencia, Spain, pp. 133–138.

Tsvetkov Y., Boytsov L., Gershman A., Nyberg E. and Dyer C. (2014). *Metaphor detection with cross-lingual model transfer.* **In** *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, USA, pp. 248–258.

Tsvetkov Y. and Wintner S. (2011). *Identification of multi-word expressions by combining multiple linguistic information sources.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland, UK, pp. 836–845.

Turney P. D., Neumann Y., Assaf D. and Cohen Y. (2011). *Literal and metaphorical sense identification through concrete and abstract context.* **In** *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland, UK, pp. 680–690.

Veale T., Beigman Klebanov B. and Shutova E. (2016). *Metaphor: A Computational Perspective.* Synthesis Lectures on Human Language Technologies. Williston, VT: Morgan & Claypool Publishers.

Velikovich L., Blair-Goldensohn S., Hannan K. and McDonald R. (2010). *The viability of web-derived polarity lexicons.* **In** *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, Los Angeles, CA, USA, pp. 777–785.

Vincze V., I.N. T. and Zsibrita J. (2013). Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing* **10**(2), 25–25.

Wiebe J., Wilson T. and Cardie C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* **2**(3), 164–210.

Wiegand M. and Ruppenhofer J. (2015). *Opinion holder and target extraction based on the induction of verbal categories.* **In** *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, Beijing, China, pp. 215–225.

Wiegand M., Schulder M. and Ruppenhofer J. (2016). *Separating actor-view from speaker-view opinion expressions using linguistic features.* **In** *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, San Diego, CA, USA, pp. 778–788.

Williams L., Bannister C., Arribas-Ayllon M., Preece A. and Spasić I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications* **42**(21), 7375–7385.

Wilson T., Wiebe J. and Hoffmann P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* **In** *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada, pp. 347–354.

Zesch T., Müller C. and Gurevych I. (2008). *Extracting lexical semantic knowledge from Wikipedia and Wiktionary.* **In** *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, pp. 1646–1652.

Zhang D. and Wang D. (2015). Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006.*

Zhang M., Liang P. and Fu G. (2019). *Enhancing opinion role labeling with semantic-aware word representations from semantic role labeling.* **In** *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, Minneapolis, MN, USA, pp. 641–646.

Cite this article: Wiegand M, Schulder M and Ruppenhofer J (2024). Determining sentiment views of verbal multiword expressions using linguistic features. *Natural Language Engineering* **30**, 256–293. https://doi.org/10.1017/S1351324923000153

https://doi.org/10.1017/S1351324923000153 Published online by Cambridge University Press