*Natural Language Engineering* (2021), **27**, pp. 153–179
doi:10.1017/S135132492000039X

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# Automatic generation of lexica for sentiment polarity shifters

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

[1]Universität des Saarlandes, Sprach- & Signalverarbeitung, C7 1, 66123 Saarbrücken, Germany, [2]Institut für Deutsche Gebärdensprache, Gorch-Fock-Wall 7, 20354 Hamburg, Germany and [3]Institut für Deutsche Sprache, R 5, 6-13, 68161 Mannheim, Germany
*Corresponding author. E-mail: marc.schulder@uni-hamburg.de

(Received 17 May 2019; revised 27 March 2020; accepted 15 May 2020; first published online 9 July 2020)

**Abstract**

Alleviating pain is good and abandoning hope is bad. We instinctively understand how words like *alleviate* and *abandon* affect the polarity of a phrase, inverting or weakening it. When these words are content words, such as verbs, nouns, and adjectives, we refer to them as *polarity shifters*. Shifters are a frequent occurrence in human language and an important part of successfully modeling negation in sentiment analysis; yet research on negation modeling has focused almost exclusively on a small handful of closed-class negation words, such as *not*, *no*, and *without*. A major reason for this is that shifters are far more lexically diverse than negation words, but no resources exist to help identify them. We seek to remedy this lack of shifter resources by introducing a large lexicon of polarity shifters that covers English verbs, nouns, and adjectives. Creating the lexicon entirely by hand would be prohibitively expensive. Instead, we develop a bootstrapping approach that combines automatic classification with human verification to ensure the high quality of our lexicon while reducing annotation costs by over 70%. Our approach leverages a number of linguistic insights; while some features are based on textual patterns, others use semantic resources or syntactic relatedness. The created lexicon is evaluated both on a polarity shifter gold standard and on a polarity classification task.

**Keywords:** Sentiment analysis; Sentiment polarity; Lexical semantics; Lexicon generation; Negation content words

## 1. Introduction

In natural language processing, the field of sentiment analysis is concerned with the detection and analysis of opinions and evaluative statements in language. While this involves several tasks, such as determining the opinion holder, target, and intensity, the vast majority of research focuses on determining the *polarity* (also referred to as *valence*) of a text, that is, whether it is positive, negative, or neutral.

The basis for determining the polarity of a text is knowing the polarity of individual terms within it. Knowing that *to pass* in (1) is a positive term allows us to infer that *"pass the exam"* is a positive phrase and that the entire sentence is positive. The polarity of an expression can also be influenced by a number of phenomena, for example, by *negation*. The best-established cause of negation is *negation words*, such as *no*, *not*, *neither*, or *without*. In (2), the negation *not* affects the positive polarity of *"pass the exam"*, resulting in a negative polarity for the sentence.[a]

[a]In example sentences, phrase scopes are indicated by square brackets. The polarity of a word or phrase is indicated by a superscript + (positive), −; (negative), or ∼ (neutral). The key phenomenon of the example is highlighted in bold and identified in a subscript.

(1) Peter [passed$^{+}$ the exam]$^{+}$.

(2) Peter [did **not**$_{\textit{negation}}$ [pass the exam]$^{+}$]$^{-}$.

Negation words are not, however, the only words that can affect the polarity of a phrase. Many content words, so-called *polarity shifters*, can have a very similar effect. The negated statement in (2), for example, can also be expressed using the verb *fail*, as seen in (3). Polarity shifters are not limited to verbs. The nominal (4) and adjectival forms (5) of *fail* exhibit the same kind of polarity shifting.

(3) Peter [**failed**$_{\textit{shifter}}$ to [pass the exam]$^{+}$]$^{-}$.

(4) Peter's [**failure**$_{\textit{shifter}}$ to [pass the exam]$^{+}$]$^{-}$.

(5) Peter's [**failed**$_{\textit{shifter}}$ attempt to [pass the exam]$^{+}$]$^{-}$.

Handling these nuances of compositional polarity is essential, especially for phrase- and sentence-level polarity classification. While significant research has been performed on the topic of compositional polarity, it has mostly focused on negation words (Wiegand *et al.* 2010). One reason for this is the availability of lexical resources for negation words and lack thereof for polarity shifters. Negation words are usually function words, of which there are few. Polarity shifters, on the other hand, are content words (e.g., verbs, adjectives, and nouns), which are far more numerous. *WordNet* (Miller *et al.* 1990), for example, contains over 10,000 verbs, 20,000 adjectives, and 110,000 nouns. At the same time, most individual content words occur far less frequently than individual function words. Overall, however, polarity shifters can be expected to occur more frequently than negation words (Schulder *et al.* 2018b). The challenge is therefore how to create a large lexicon of polarity shifters while keeping the required annotation effort manageable.

While previous work in compositional sentiment analysis included research on specific linguistic issues, such as the truth or falsity of complement clauses (Nairn, Condoravdi, and Karttunen 2006) or the inference of implicit opinions, that is, opinion implicatures (Deng and Wiebe 2014), the lexical resources created as part of that research do not sufficiently cover polarity shifters. We demonstrate this for the *effect lexicon* (Choi and Wiebe 2014) that was created for computing opinion implicatures. Previous work also exclusively focused on verbs while we also consider nouns and adjectives.

Prior to our own efforts, even the most complex negation lexicon for English (Wilson, Wiebe, and Hoffmann 2005) contained only 30 polarity shifters. While corpora for the training of negation handling exist (Szarvas *et al.* 2008; Socher *et al.* 2013), they are inadequate for learning how to handle most polarity shifters (as we show in Section 7.2). Creating a comprehensive lexicon through manual annotation alone would be prohibitively expensive, as it would require the annotation of many tens of thousands of words. Instead, we introduce a bootstrapping approach that allows us to filter out the majority of words that do not cause polarity shifting (non-shifters), reducing the manual annotation effort by over 72%, saving hundreds of work hours.

The structure of our bootstrapping approach is detailed in Figure 1. We begin by having a human annotator label a small number of randomly sampled verbs, which are used to evaluate a variety of linguistic features and to train a supervised classifier. This classifier is used to classify the remaining unlabeled verbs. Verbs that the classifier considers shifters are manually verified by our annotator, while those classified as non-shifters are discarded. This ensures the high quality of the lexicon while significantly reducing the annotation load. Once the verb lexicon is complete, the process is repeated for nouns and adjectives. As we already have the verb lexicon at this point, we use it as a resource in the feature design for nouns and adjectives.

This article presents and significantly extends our work in Schulder *et al.* (2017). Our goal is to create a large lexicon of English polarity shifters through the use of bootstrapping and to show its use for improving polarity classification in sentiment analysis. While Schulder *et al.* (2017)

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

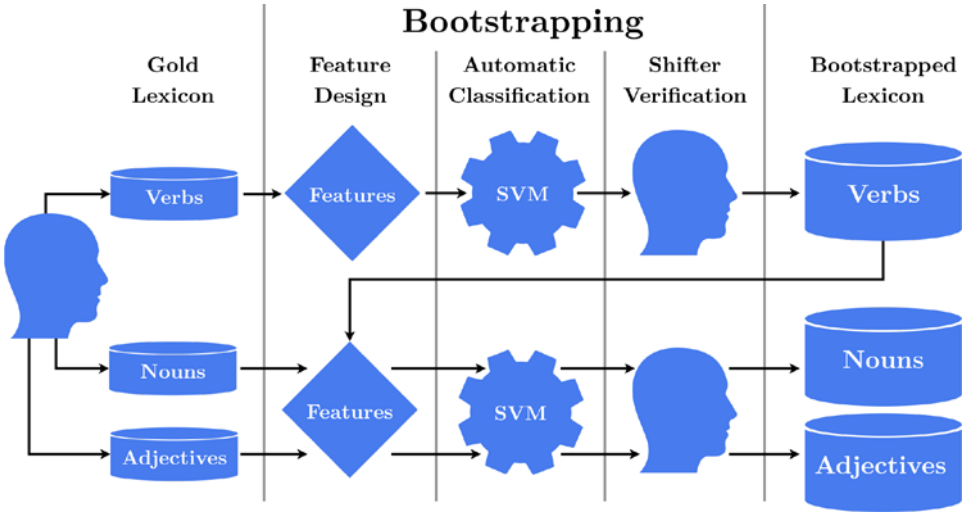



**Fig. 1.** Workflow for creating the polarity shifter lexicon.

focused on verbs, we now extend the lexicon to also include nouns and adjectives, thus creating a *general lexicon of polarity shifters*. This requires both adapting the verb features to other parts of speech and the introduction of entirely new features. Schulder *et al.* (2017) also included an extrinsic evaluation to show that explicit knowledge of polarity shifters is essential to correctly model phrase polarities. We expand this evaluation to also investigate the impact of word sense disambiguation of shifters on polarity classification.

The remainder of this article is structured as follows: in Section 2, we provide a formal definition of polarity shifters, and information on related works. Section 3 introduces our gold standard and other resources. In Section 4, we describe the features used in our bootstrap classifier, and in Section 5, we evaluate both features and classifier. The actual bootstrapping of the lexicon is performed in Section 6. In Section 7, we compare the information our lexicon provides against compositional polarity classifiers without explicit knowledge of shifters and against a verbal shifter lexicon that differentiates by word sense. Section 8 concludes this article.

We make all data annotated as part of our research, that is, the entire polarity shifter lexicon as well as the gold standard of the extrinsic evaluation, *publicly available*.[b]

## 2. Background

### *2.1 Polarity shifters*

*Polarity shifting* occurs when the sentiment polarity (or valence) of a word or phrase is moved toward the opposite of its previous polarity (i.e., from positive toward negative or vice versa). The phenomenon was first brought to the attention of the research community by Polanyi and Zaenen (2006), who observed that the prior polarity of individual lexical items could be shifted by (a) specific lexical items, (b) the discourse structure and genre type of a text, and (c) sociocultural factors. In subsequent research, the meaning of the term *shifter* was narrowed to refer to lexical items that affect phrasal polarity. For the purposes of this work, we further require that shifters must be *open class* words (e.g., verbs, adjectives, or nouns) to differentiate them from closed-class negation words.

*Polarity shifters are defined by their ability to negate or diminish facts or events that were either previously true or presupposed to occur.* In (6), the speaker expects that their daughter would

[b] https://doi.org/10.5281/zenodo.3365601

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

receive a scholarship, as she applied for one. This did not happen, as being denied a scholarship implies not receiving it. In (7), the speaker expects that their amount of pain will continue at the same level, but due to the medication the amount of pain is reduced. These examples also show that shifting can occur in either direction, as in (6) a positive polarity is shifted to negative and in (7) a negative polarity is shifted to positive.

(6) My daughter was [**denied**$_{\textbf{\textit{shifter}}}$ the [scholarship]$^{+}$]$^{-}$.

(7) The new treatment has [**alleviated**$_{\textbf{\textit{shifter}}}$ my [pain]$^{-}$]$^{+}$.

In our work, we use the term *polarity shifter* in a descriptive spirit. We do so since, on the one hand, we take a data-driven approach to identify the class of items that we deal with and, on the other hand, we do not believe that the lexical items in question can be fully subsumed under existing categorizations (see Section 2.1.1).

### *2.1.1 Polarity shifting and related concepts*

**Polar expressions** are words that inherently express an evaluation (or sentiment, appraisal, *etc.*). Clear examples are the adjective *good*, verb *like*, and noun *hate*. When polar expressions are combined with other polar expressions and both have the same opinion holder, then the one that takes the other as a syntactic argument dominates in the overall sentiment, but the argument's sentiment is co-present:

(8) I [[like]$^{+}$ [annoying]$^{-}$ people]$^{+}$.

(9) I [[hate]$^{-}$ having [fun]$^{+}$]$^{-}$.

(10) She just [hates]$^{-}$ that [lovely]$^{+}$ man.

In cases like (8) and (9), the effect can seem contradictory. This is not the case when opinion holders differ, as in (10), where we understand *lovely* to be the evaluation of the speaker of the sentence, whereas *hate* is that of the referent of *she*. Sentiments by different opinion holders do not interact to cause any form of shifting.

While some polarity shifters are also polar expressions, their polarity does not dictate the shifting direction. For example, the shifter *destroy* is of negative polarity, but shifts both positive and negative words, as seen in (11) and (12).

(11) Smoking [[**destroys**]$^{-}_{\textbf{\textit{shifter}}}$ your [health]$^{+}$]$^{-}$.

(12) The medication [[**destroys**]$^{-}_{\textbf{\textit{shifter}}}$ [cancer cells]$^{-}$]$^{+}$.

**Negation** can be performed syntactically by function words such as *not*, *never*, *nowhere*, or *no*. Syntactic negation words create syntactic negation scopes that license the use of so-called negative (syntactic) polarity items such as *ever* and which block positive (syntactic) polarity items such as *hardly*:

(13) He [**hasn't**$_{\textbf{\textit{negation}}}$ ever [helped]$^{+}$me]$^{-}$.

(14) *He [**hasn't**$_{\textbf{\textit{negation}}}$ hardly [helped]$^{+}$ me]$^{?}$.

(15) *He [has ever [helped]$^{+}$ me]$^{?}$.

(16) He [has hardly [helped]$^{+}$ me]$^{-}$.

Lexical negation words include negation as part of their internal semantic representation. Sometimes, this is reflected by their morphological structure (in the verb *unmake* and the adjective *homeless*) but not necessarily so (see the verbs *stop* and *abstain from*). Lexical negation does not yield syntactic negation scopes:

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

(17) *He has ever [abstained]$^{-}$ from smoking.

(18) *I have ever been [homeless]$^{-}$.

Negation does not introduce sentiment when there is none to begin with:

(19) John is not my brother, he is my neighbor.

+/−**Effect verbs** are ones that imply a positive or negative effect on an entity participating in some state or event. Predications with effect verbs are not polar unless relevant entities are valued positively or negatively:

(20) Penguins [**lack**$_{-effect}$ the ability to fly]$^{\sim}$.

(21) Peter [**lacks**$_{-effect}$[ambition]$^{+}$]$^{-}$.

(22) Max [has [energy]$^{+}$ to **burn**$_{+effect}$]$^{+}$.

The sentiment polarity associated with predications of effect verbs comes about via compositional inference, unlike with polar expressions where sentiment polarity is directly coded. As shown by the work of Anand and Reschke (2010), Deng, Choi, and Wiebe (2013), and Ruppenhofer and Brandes (2016), the calculation of the sentiment polarity usually depends on the property of several arguments of a predicate and its semantic class. (The examples in (20)–(22) illustrate verbs to do with possession.)

Note that the effect of effect verbs is not necessarily *binary*: a special subclass of what are treated as effect verbs is ones that entail a scalar change such as *increase*, *decrease*, and *reduce*. For sentiment analysis purposes, scalar effect verbs such as *cut* in (23) are usually treated like non-scalar effect verbs such as *eliminate* in (24).

(23) The company [**cut**$_{-effect}$ my [bonus]$^{+}$]$^{-}$.

(24) The company [**eliminated**$_{-effect}$ my [bonus]$^{+}$]$^{-}$.

Effect verbs may also be multilayered and come with a built-in polarity on one of the arguments. For instance, *spare* as a verb of negated giving enforces a construal of the potential theme as having negative sentiment polarity:

(25) Thankfully, they [**spared**$_{-effect}$ me the [trauma]$^{-}$]$^{+}$ of choosing dessert by offering the sampler platter.

(26) Thankfully, they [**spared**$_{-effect}$ me the [joy]$^{+}$]$^{?}$ of choosing dessert by offering the sampler platter.

Using *joy* in (26) rather than *trauma* in (25) clashes with the built-in construal of the theme as negative. The verbs *lack* and *spare* in (20), (21), and (25) show that conceptually half the effect verbs are lexical negation words. But note that they are complemented by words without lexical negation such as *have* in (22).

Polarity shifting covers intensionally syntactic negation as well as positive and negative effect predicates. However, and importantly, it sets aside those effect predicates whose overall polarity is lexically prespecified and which impose a sentiment polarity on one or more of the arguments that are relevant for the calculation of the effect's polarity with other items in its class. For instance, while the verb *abuse* has a negative effect on its patient like *rough up* does, *rough up* is not a shifter because it still results in a negative sentence even if its object is valued neutrally.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

#### *2.1.2 What counts as shifting?*

Most commonly, the term *shifting* is used to refer to a change between discrete polarity classes, for example, from positive to negative or vice versa. There is no consensus on whether this includes shifting toward neutral polarity or not. In (27), it is unclear whether the polarity of *wasn't excellent* should be considered negative or neutral.

(27) Let's say, the movie [**wasn't** [excellent]$^{+}$]$^{-/\sim}$.

Choi and Cardie (2008) state that the positive polarity of *excellent* is flipped to negative. Taboada *et al.* (2011) disagree arguing the negation of *excellent* is not synonymous with its antonym *atrocious* and should be considered neutral.

Another question is whether intensification (e.g., *extremely dangerous*) should also be considered shifting. Polanyi and Zaenen (2006) include it as it affects the polar intensity of a phrase. However, intensifiers serve to strengthen a given polarity and prevent it from being replaced with a different polarity. A *good movie* cannot be bad at the same time, but can be even more positive than *good* already implies (e.g., *"The movie was good. In fact, it was excellent."*) This is incompatible with our definition of shifting. Therefore, we do not consider intensifiers to be shifters.

#### *2.1.3 Compositionality of phrasal polarity*

To determine the polarity of a phrase, we observe (a) the polarity of its lexical items and (b) how their polarity is influenced by contextual elements (Moilanen and Pulman 2007; Anand and Reschke 2010). Following the principles of semantic compositionality, the scope of most contextual elements is limited to specific syntactic constituents (Moilanen and Pulman 2007; Choi and Cardie 2008). In (28), the verbal shifter *defeat* affects the polarity of its direct object, while in (29) the verb *falter* shifts the polarity of its subject. However, it is not just shifting and negation that can influence phrasal polarity. Connectives such as *however* or *but* influence which parts of a sentence affect the overall polarity of the phrase (Polanyi and Zaenen 2006). In (29), the positive polarity of *enthusiasm* is counteracted by the connective *despite* and in (30) the connective *but* indicates that the positive polarity of the second half of the sentence takes precedence over the negative polarity of the first half.

(28) [The hero]$^{+}_{subj}$ [**defeated**$_{\textbf{\textit{shifter}}}$[the villain]$^{-}_{dobj}$]$^{+}$.

(29) [[[My enthusiasm]$^{+}_{subj}$ **faltered**$_{\textbf{\textit{shifter}}}$]$^{-}$ **despite**$_{\textbf{\textit{connective}}}$ their [encouragement]$^{+}$]$^{-}$.

(30) [[The battle was gruesome,]$^{-}$ **but**$_{\textbf{\textit{connective}}}$ [we prevailed]$^{+}$]$^{+}$.

Modal operators like *if* and *could* introduce hypotheticals that do not directly impact the polarity of events (e.g., *"If Mary were a bad person, she would be mean to her dogs"* conveys no negative opinion about Mary) (Polanyi and Zaenen 2006) or may even shift polarities (*"this phone would be perfect if it had a bigger screen"* implies the phone is *not* perfect) (Liu *et al.* 2014). Deriving the polarity of a phrase is therefore not just a matter of enumerating all polarities and shifters therein.

### *2.2 Related work*

The majority of works on the topic of the computational processing of negation concern themselves chiefly with the handling of negation words and with determining their scope. For more information on these topics, we refer the reader to the survey on negation modeling in sentiment analysis by Wiegand *et al.* (2010). We shall instead focus our discussion on works that address polarity shifters specifically.

There are few resources providing information about polarity shifters. Even fewer offer any serious coverage. The most complex general negation lexicon was published by Wilson *et al.*

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

(2005). It contains 30 polarity shifters. The *BioScope* corpus (Szarvas *et al.* 2008), a text collection from the medical domain, has been annotated explicitly for negation cues. Among these negation cues, Morante (2010) identifies 15 polarity shifters. *EffectWordNet* (Choi and Wiebe 2014) lists almost a thousand verbs with *harmful effects*, a phenomenon similar to shifting. However, this similarity is not close enough to provide reliable classifications by itself (see Section 5.1).

Alternatively, one can learn negation implicitly from corpora. The *Stanford Sentiment Treebank* (SST) (Socher *et al.* 2013) contains compositional polarity information for 11,855 sentences. Each sentence is syntactically parsed and each tree node is annotated with its polarity. Negation can be inferred by changes in polarity between nodes. Socher *et al.* (2013) show that a neural network polarity classifier trained on this treebank can successfully identify negation words. However, as individual shifters are far less frequent than negation words, the size of the treebank is not sufficient for handling shifters, as we show in Section 7.2.

The work that is most closely related to our own effort of bootstrapping lexicon creation is that of Danescu-Niculescu-Mizil, Lee, and Ducott (2009) who create a lexicon of downward-entailing operators (DE-Ops), which are closely related to polarity shifters. Leveraging the co-occurrence of DE-Ops with negative polarity items, they use unsupervised machine learning to generate a ranked list of DE-Ops. Of the 150 highest ranked items, a human annotator confirmed 60% as DE-Ops.

Our own first contribution to the topic of polarity shifters was Schulder *et al.* (2017), in which we bootstrap a lexicon of 980 English verbal shifters and evaluate their use for polarity classification. This work is covered as part of this article (for details see Section 1). In Schulder, Wiegand, and Ruppenhofer (2018a), we adapt our bootstrapping approach and its features to German and introduce cross-lingual features that leverage the lexicon of Schulder *et al.* (2017). Schulder *et al.* (2018b) relies entirely on manual annotation to create a lexicon of 1220 English verbal shifters. Like Schulder *et al.* (2017), it covers English verbs, but it provides additional information by assigning shifter labels for individual word senses and by annotating the syntactic scope of the shifting effect. Wiegand, Loda, and Ruppenhofer (2018) examine such sense-level information for shifting-specific word sense disambiguation. They conclude that while generally possible, this task would require large amounts of labeled training data.

## 3. Resources

To bootstrap a shifter lexicon, we require several resources. In Section 3.1, we define the vocabulary of our lexicon and use a subset to create a gold standard. In Section 3.2, we describe additional resources necessary for our feature extraction.

### 3.1 Gold standard

For our lexicon, we need to define the underlying vocabulary. To this end, we extract all verbs, nouns, and adjectives from *WordNet 3.1*.[c] This amounts to 84,174 words: 10,581 verbs, 55,311 nouns, and 18,282 adjectives. From here on, all references to *"all words"* refer to this selection. We use some of them to create a shifter gold standard in this section. For the remaining words, we bootstrap shifter labels in Section 6.

To train and test our classifiers, we create a polarity shifter gold standard for verbs, nouns, and adjectives. We extract a random sample of 2000 words per part of speech from the vocabulary, to

[c]We exclude words that by definition cannot be shifters. These include proper names, abbreviations, words containing digits and proverbial and compositional expressions.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 1.** Distribution of polarity shifters in gold standard. For each part of speech, a random sample of 2000 words was taken from *WordNet*.

| | Verbs | | Nouns | | Adjectives | |
|---|---|---|---|---|---|---|
| | Frequency | Percentage | Frequency | Percentage | Frequency | Percentage |
| Shifter | 304 | 15.20 | 107 | 5.35 | 129 | 6.45 |
| Non-shifter | 1696 | 84.80 | 1893 | 94.65 | 1871 | 93.55 |
| Total | 2000 | | 2000 | | 2000 | |

be labeled by an expert annotator with experience in linguistics and annotation work. To ensure that all possible senses of a word are considered, the annotator refers to word definitions in a number of different dictionaries.

Annotation is handled as a *binary classification* task. Each word is either a *"shifter"* or a *"Non-shifter"*. Following our definition from Section 2.1, to qualify as a shifter, a word must allow polar expressions as its dependent and the polarity of the shifter phrase (i.e., the proposition that embeds both the shifter and the polar expression) must move toward a polarity opposite to that of the polar expression.

Our gold standard is annotated at the lemma level. When a word has multiple word senses, we consider it a shifter if at least one of its senses qualifies as a shifter (cf. Schulder *et al.* (2018b)). Word sense shifter labels would only be of use if the texts they were applied to were also word sense disambiguated. We do not believe that automatic word sense disambiguation is sufficiently robust for our purposes.

Annotating the gold standard took 170 work hours. To measure inter-annotator agreement, 10% of the gold standard was also annotated by one of the authors. This resulted in a Cohen's kappa (Cohen 1960) of 0.66 for verbs, 0.77 for nouns, and 0.71 for adjectives. All scores indicate substantial agreement (Landis and Koch 1977).

The raw percentage agreement for the 600 lemmas labeled by the annotators is 85.7%. The 86 disagreements break down to 34 among the verbs, 29 among the adjectives, and 23 among the nouns. One major type of divergence among the annotators are cases where a neutral literal and a polar metaphorical meaning coexist (e.g., in the noun *cushioning* or the adjective *geriatric*). Given that the annotations were performed out of context, annotators might have had different typical uses in mind when deciding on their lemma-level annotations.

Table 1 shows the distribution of shifters in our gold standard. Unsurprisingly, the majority of words are non-shifters. However, extrapolating from the shifter frequencies, we can still expect to find several thousand shifters in our vocabulary.

### *3.2 Additional resources*

**Sentiment polarity:** As polarity shifters interact with word polarities, some of our features require knowledge of *sentiment polarity*. We use the *Subjectivity Lexicon* (Wilson *et al.* 2005) to determine the polarity of individual words. Table 2 shows that of the shifters for which the polarity is known, the vast majority are negative.

**Text corpus:** Many of our features require a *text corpus*, for example, for word frequencies or pattern recognition. We use *Amazon Product Review Data* (Jindal and Liu 2008), a corpus of 5.8 million product reviews. The corpus was chosen both for its large size and its domain. Product reviews are a typical domain for sentiment analysis, as they are rich in opinions and polar statements and very focused on communicating the opinion of the author (Liu 2012). We expect that

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 2.** Shifter distribution in the *Subjectivity Lexicon* (Wilson *et al*. 2005).

| | | Verbs | | Nouns | | Adjectives | |
|---|---|---|---|---|---|---|---|
| | | Frequency | Percentage | Frequency | Percentage | Frequency | Percentage |
| Positive words | Shifter | 4 | 5.5 | 1 | 1.9 | 5 | 2.3 |
| | Non-shifter | 69 | 94.5 | 52 | 98.1 | 216 | 97.7 |
| Negative words | Shifter | 49 | 25.9 | 30 | 24.4 | 68 | 20.4 |
| | Non-shifter | 140 | 74.1 | 93 | 75.6 | 266 | 79.6 |

using a sentiment-rich text corpus will help avoid issues of sparsity that might arise in other corpora consisting more of neutral factual statements that cannot be affected by polarity shifters.

**Word embedding:** Some features rely on the distributional hypothesis that words in similar contexts have similar meanings (Firth 1957). To determine this distributional similarity, we use `Word2Vec` (Mikolov *et al*. 2013) to compute a word embedding vector space from our text corpus. Following the work of Wiegand and Ruppenhofer (2015) on the related task of verb category induction for sentiment roles, we use the continuous bag of words algorithm and generate a vector space of 500 dimensions. All other settings are kept at their default. The resulting word embedding is made publicly available.[d] To determine the similarity between two specific words, we compute their cosine similarity.

**Syntactic structure:** *Syntactic dependency relations* are often used to extract information from corpora through the application of text patterns (Jiang and Riloff 2018) and to collect complex distributional information (Shwartz, Goldberg, and Dagan 2016). We use the *Stanford Parser* (Chen and Manning 2014) to obtain syntactic information.

## 4. Feature design

We provide a variety of features for our bootstrap classifier. Section 4.1 describes how we define (for computational purposes) the scope that a polarity shifter can affect. In Section 4.2, we introduce features specifically designed for determining polarity shifters. Section 4.3 presents more generic features already established in a number of sentiment analysis tasks. Finally, in Section 4.4, we discuss means of applying shifter information across different parts of speech.

### *4.1 Shifting scope*

In preparation for our feature definitions, we define how we handle shifting scope, that is, which part of a sentence is affected by the shifter. A shifter can only affect expressions that it syntactically governs, but not every argument is within its shifting scope. In (31) and (32), the polarity of the direct object is shifted by *defeated*, but the polarity of the subject, which is outside the shifting scope, makes no difference.

(31) $[\text{The villain}]^{-}_{subj}$ $[\textbf{defeated}_{\textbf{\textit{shifter}}}$ $[\text{the hero}]^{+}_{dobj}]^{-}$.

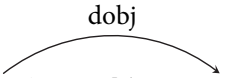
(32) $\text{Chance}_{subj}$ $[\textbf{defeated}_{\textbf{\textit{shifter}}}$ $[\text{the hero}]^{+}_{dobj}]^{-}$.

[d] https://doi.org/10.5281/zenodo.3370051

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

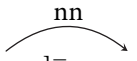
To determine the scope of a shifter, we rely on its dependency relations, which differ according to part of speech.[e] We consider the following dependency relations as identified by the *Stanford Typed Dependencies* tag set (de Marneffe and Manning 2008):

`dobj:` If the shifter is a verb, then its direct object is the scope.



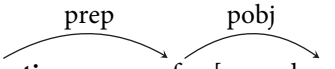

*Example:* The storm [**ruined**$_{shifter}$ [their party]$^{+}$]$^{-}$.

`nn:` If the shifter is the head of a noun compound, then the compound modifier of the compound is the scope.



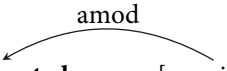

*Example:* It is a [[cancer]$^{-}$ **cure**$_{shifter}$]$^{+}$.

`prep_of:` If the shifter is a noun that is the head of the preposition *of*, then the object of that preposition is the scope.



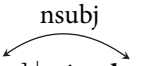

*Example:* It was the [**destruction**$_{shifter}$ of [my dreams]$^{+}$]$^{-}$.

`amod:` If the shifter is an attributive adjective, then the modified noun is the scope.




*Example:* The [**exonerated**$_{shifter}$ [convict]$^{-}$]$^{+}$ walked free.

`nsubj:` If the shifter is a predicative adjective, then its subject is the scope.




*Example:* The [[hero]$^{+}$ is **dead**$_{shifter}$]$^{-}$.

### *4.2 Task-specific features*

We begin with features specifically designed to identify polarity shifters. Each feature creates a word list ranked by how likely each word is to be a shifter.

#### *4.2.1 Features applicable to all parts of speech*

**Distributional similarity (SIM):** Shifters and negation words are closely related and words that occur in similar contexts as negations might be more likely to be shifters. Using our word embedding, we rank all words by their similarity to negation words. We use the intersection of negation words from Morante and Daelemans (2009) and the *valence shifter lexicon* (Wilson *et al.* 2005). We compute the *centroid* of these to create an embedding representation of the general concept of negation words. Potential shifters are ranked by their similarity to the centroid.

**Polarity clash (CLASH):** Shifters with a polarity of their own tend to shift expressions of the opposite polarity. For example, the negative verb *ruin* is a shifter that often has positive polar expressions like *career* or *enjoyment* in its scope:

[e]This definition of shifting scopes is a simplified representation designed to fit the needs of our data-driven features. For more detailed discussions that also address less frequent kinds of scopes, we refer the reader to Wiegand *et al.* (2018) and Schulder *et al.* (2018b).

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

(33) It [**ruined**$_{\textbf{\textit{shifter}}}$ her [career]$^{+}$]$^{-}$.

(34) The constant coughing [**ruined**$_{\textbf{\textit{shifter}}}$ my [enjoyment]$^{+}$ of the play]$^{-}$.

The more often the polarity of a word clashes with that of its scope (see Section 4.1), the more likely it is to be a shifter. Due to the rarity of shifters with positive polarity in the *Subjectivity Lexicon* (see Table 2) and the associated higher risk of including non-shifters, we limit our search to negative shifter candidates that occur with positive polar expressions. Each candidate is ranked by its relative frequency of occurring with positive polarity scopes.

**Heuristic using *"any"* (ANY):** For this feature, we leverage the similarity between shifters and DE-Ops, which are expressions that invert the logic of entailment assumptions (Ladusaw 1980). Under normal circumstances, such as in (35), a statement implies its relaxed form (35a), but not restricted forms like (35b). However, in (36), the DE-Op *doubt* inverts entailment, so that the relaxed form (36a) is not entailed, but the restricted form (36b) is.

(35) The epidemic spread quickly.

(35a) The epidemic spread.

(35b) *The epidemic spread quickly via fleas.

(36) We **doubt** the epidemic spread quickly.

(36a) *We **doubt** the epidemic spread.

(36b) We **doubt** the epidemic spread quickly via fleas.

The inversion of inference assumptions modeled by downward entailment is closely related to polarity shifting, as both relate to the non-existence or limitation of entities (van der Wouden 1997) (see also Section 2.1.1). This overlap means that DE-Ops often also qualify as polarity shifters or negation words.

Negative polarity items (NPIs) are words that are excluded from being used with positive assertions, a phenomenon strongly associated with both DE-Ops (Ladusaw 1980) and negation (Baker 1970; Linebarger 1980). For example, the NPI *any* may be used in negated contexts, such as with *not* in (37) or *deny* in (38), but not in positive assertions like (39). NPIs are strongly connected to DE-Ops, usually occurring in their scope (Ladusaw 1980), although their exact nature is still disputed (Giannakidou 2011). We hypothesize that a similar connection can be found between NPIs and polarity shifters, as exemplified in (38).

(37) They did [**not** give us <u>any</u> help$^{+}_{dobj}$]$^{-}$.

(38) They [**denied**$_{\textbf{\textit{shifter}}}$ us <u>any</u> help$^{+}_{dobj}$]$^{-}$.

(39) *They gave us <u>any</u> help.

Danescu-Niculescu-Mizil *et al.* (2009) use the co-occurrence of NPIs and DE-Ops to automatically list DE-Ops. We adapt a similar approach to identify shifters, collecting occurrences in which the NPI *any* is a determiner within the scope of the potential shifter, such as in (38). Potential shifters are sorted by their relative frequency of co-occurring with this pattern (ANY). As an additional constraint, we require that the head word of the scope must be a polar expression (ANY$_{\text{polar}}$). In (38) this requirement is met, as *help* is of positive polarity. A variety of NPIs was considered for this feature, but only *any* provided the required pattern frequencies to make an efficient feature.

#### *4.2.2 Features applicable to verbs*

The following features are only available for verbs, either due to their nature or due to the availability of resources.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**EffectWordNet (–EFFECT):** +/−Effect is a semantic phenomenon similar to polarity shifting. It posits that events may have beneficial (+*effect*) or harmful effects (−*effect*) on the objects they affect (Deng *et al.* 2013; Choi, Deng, and Wiebe 2014; Choi and Wiebe 2014). It was originally introduced in the context of *opinion inference*. In (40), people are happy about the event *"Chavez has fallen"* and *fall* has a harmful –effect on *Chavez*. It can be inferred that people have a negative opinion of Chavez due to their positive reaction to a harmful effect on him.

(40) I think people are happy because [[Chavez]$^{-}$ has **fallen**$_{-effect}$]$^{+}$.

(41) We don't want the public getting the idea that we [**abuse**$_{-effect}$ our [prisoners]$^{-}$]$^{-}$.

As a semantic concept, –effects bear some similarity to shifters. Often, the harmful effect that they describe is one of removal or weakening, that is, of shifting, such as in (40). However, despite their similarity, the two phenomena are not identical. In (41), *abuse* has a harmful –effect on the prisoners, but it does not shift the polarity.

While –effects and polarity shifting are not equivalent, their relatedness may be a useful source of information. We use *EffectWordNet* (Choi and Wiebe 2014), a lexical resource for verbs which provides effect labels for *WordNet* synsets. We label verbs as shifters when at least one of their word senses has a –effect and none have a +effect. As no inherent ranking is available, we use word frequency as a fallback.

**Particle verbs (PRT):** Particle verbs are phrasal constructs that combine verbs and adverbial particles, such as *"tear down"* or *"lay aside"*. Often the particle indicates a particular aspectual property, such as the complete transition to an end state (Brinton 1985). In *"dry (something) out"*, *out* indicates that we *"dry (something) completely"*. Shifting often involves the creation of a new (negative) end state of an entity, for example, through its removal or diminishment (see Section 2.1). We expect a significant number of particle verbs to be shifters, such as in (42) and (43).

(42) This [**tore down**$_{shifter}$ our great [dream]$^{+}$]$^{-}$.

(43) Please [**lay aside**$_{shifter}$ all your [worries]$^{-}$]$^{+}$.

We only consider particles which typically indicate a complete transition to a negative end state: *aside*, *away*, *back*, *down*, *off*, and *out*. The list of verbs is ranked via the frequency of the particle verb relative to the frequency of its particle.

### *4.3 Generic features*

The following features use general purpose semantic resources. They do not produce ranked lists and are only evaluated in the context of supervised classification.

**WordNet (WN):** *WordNet* (Miller *et al.* 1990) is the largest available ontology for English and a popular resource for sentiment analysis. *Glosses*, brief sense definitions, are a common feature for lexicon induction tasks in sentiment analysis (Esuli and Sebastiani 2005; Choi and Wiebe 2014; Kang *et al.* 2014). We expect that the glosses of shifters will share similar word choices. *Supersenses* (coarse semantic categories) and *hypernyms* (more general related concepts) have also been found to be effective features for sentiment analysis, as shown by Flekova and Gurevych (2016). We treat each shifter candidate as the union of its *WordNet* senses. Glosses are represented as a joint bag of words, while supersenses and hypernyms are represented as sets.

**FrameNet (FN):** *FrameNet* (Baker, Fillmore, and Lowe 1998) is a frame semantics resource (Fillmore 1967). It has been used for sentiment tasks such as opinion spam analysis (Kim *et al.* 2015), opinion holder and target extraction (Kim and Hovy 2006), and stance classification (Hasan and Ng 2013). *FrameNet* collects words with similar semantic behavior in semantic frames. We

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

assume that shifters cluster together in specific frames, such as AVOIDING, which consists exclusively of shifters like *desist*, *dodge*, *evade*, *shirk*, *etc.* The frame memberships of a word are used as its feature.

### *4.4 Cross-POS feature*

Following the workflow outlined in Section 1, the verb component of our shifter lexicon is created first. This means the verb lexicon is available to us as a resource when we bootstrap nouns and adjectives. We hypothesize that nominal (N) and adjectival (A) forms of a verbal shifter (V) will equally be shifters, as can be seen in (44)–(46).

(44) Smoking [**damages**$^{V}_{\textit{shifter}}$ his [health]$^{+}$]$^{-}$.

(45) Beware the [[health]$^{+}$ **damage**$^{N}_{\textit{shifter}}$]$^{-}$ caused by smoking.

(46) Constant chain smoking is the reason for his [**damaged**$^{A}_{\textit{shifter}}$ [health]$^{+}$]$^{-}$.

Using our bootstrapped lexicon of verbal shifters, we assign shifter labels to related nouns and adjectives. To determine related words, we use the following approach:

**Relatedness to verb (VerbLex):** To connect nouns with related verbs, we use the *WordNet* derivational-relatedness relation or, if unavailable, the *NOMLEX* nominalization lexicon (Macleod *et al*. 1998). For adjectives, these resources are too sparse. Instead, we match word stems (Porter 1980), a word's root after removal of its inflectional suffix, to approximate relatedness. Stems are not specific to a part of speech, for example, the verb *damage* and the adjective *damaged* share the stem *damag*.

## 5. Evaluation on gold standard

We now evaluate the features from Section 4. Section 5.1 presents a precision-based evaluation of task-specific features to determine their use in unsupervised contexts. In Section 5.2, we combine those features to increase recall. Section 5.3 investigates the amount of training data required for high-quality classifications. These evaluations are in preparation for bootstrapping a complete shifter lexicon in Section 6.

### *5.1 Analysis of task-specific features for verb classification*

The task-specific features (Section 4.2) were specifically designed for shifter classification. To determine their quality, we perform a precision-based evaluation. Each feature is run over the 2000 verbs in our gold standard (Section 3.1) and generates a ranked list of potential shifters. Features are evaluated on the precision of high-ranking elements of their list. We limit this evaluation to verbs as this classification represents the first step in our workflow (Figure 1), and decisions regarding the classification of verbs were made before gold standards for the other parts of speech were available. Therefore, cross-POS features are not evaluated at this point. Generic features are not evaluated in this phase as they do not generate ranked lists.

Table 3 shows the number of verbs retrieved by each feature, as well as the precision of the 20, 50, 100, and 250 highest ranked verbs. We compare our features against two baseline features. The first is a list of all gold standard verbs ranked by their frequency in our text corpus (FREQ). This is motivated by the observation that shifters are often polysemous words (Schulder *et al.* 2018b) and polysemy is mainly found in frequently used words. The second restricts the frequency-ranked list to negative polar expressions (NEGATIVE), as the ratio of shifters to non-shifters was greatest among these expressions (see Table 2).

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 3.** Analysis of task-specific features (Section 4.2) for the classification of verbs. Features generate a ranked list of potential shifters. Best results are depicted in bold.

| Feature | Retrieved | Prec@20 | Prec@50 | Prec@100 | Prec@250 |
|---|---|---|---|---|---|
| FREQ | 2000 | 10.0 | 18.0 | 22.0 | 22.0 |
| NEGATIVE | 189 | 30.0 | 30.0 | 29.0 | *n/a* |
| SIM | 1901 | 45.0 | 30.0 | 29.0 | 27.6 |
| CLASH | 107 | 45.0 | 46.0 | 37.0 | *n/a* |
| –EFFECT | 175 | 45.0 | 44.0 | 46.0 | *n/a* |
| PRT | 165 | 60.0 | 64.0 | 58.0 | *n/a* |
| ANY | 539 | 65.0 | 60.0 | 53.0 | 38.8 |
| $ANY_{polar}$ | 272 | 75.0 | 66.0 | 62.0 | 41.2 |
| $ANY_{polar+pageR}$ | 1901 | **80.0** | **70.0** | **63.0** | **45.2** |

Our similarity to negation words feature (SIM) retrieves most of the verbs, but barely outperforms the NEGATIVE baseline.[f] This may be due to general issues that word embeddings face when encoding function words, as they occur far more frequently and in more varied contexts than individual content words do. Other features show more promising results. The polarity clash (CLASH), *EffectWordNet* (–EFFECT), and verb particle (PRT) features all clearly outperform the baselines. These features create small lists of candidates, but as these lists barely overlap with each other (12% overlap), grouping them into a set of features results in better coverage.

Nevertheless, the features, especially CLASH and –EFFECT, still contain many false-positive classifications. In the case of CLASH, this can be due to errors in the polarity labels that are compared. These labels were determined using the lexical polarities of the *Subjectivity Lexicon*, without taking word sense differences or contextual phenomena into account, so clashes may have been missed or erroneously detected. Mistakes in the –EFFECT feature are caused by its assumption that all –effect words are also shifters, a simplification that does not always hold up, as we discussed in Section 4.2.2.

The heuristic using the NPI *any* (ANY) is our strongest feature, especially when limited to polar scopes ($ANY_{polar}$). To improve it even further, we apply *personalized PageRank* (Agirre and Soroa 2009). *PageRank* ranks all nodes in a graph by how highly connected they are and *Personalized PageRank* allows prior information to be taken into account. As graph we provide a network of distributional similarities between verbs, based on our word embedding, and the output of $ANY_{polar}$ is used as prior information. The reranked list ($ANY_{polar+pageR}$), which includes all verbs found in the word embedding, does indeed improve performance. Accordingly, we use this form of the ANY feature in all future experiments.

In conclusion, we find that to fulfill the high coverage requirements of our lexicon, rather than only being used individually, the presented features will have to be combined using machine learning techniques. Even apparently weak features may contribute to performance when used in concert with others. We will discuss these multi-feature machine learning approaches in the next section.

[f]To ensure that the weak performance of SIM was not due to using the centroid, we also investigated using individual negation words, but observed no consistent improvements.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

### *5.2 Classification of complete gold standard*

In this section, we evaluate the classification of the entire gold standard. Section 5.2.1 introduces the classifiers and Section 5.2.2 evaluates their performance. As some noun and adjective classifiers rely directly on the output of the verb classifier (see description of workflow in Section 1), we evaluate each part of speech separately. Based on these results, we choose the classifier configurations for boot-strapping our shifter lexicon in Section 6.

#### *5.2.1 Classifiers*

We consider classifiers that work *with* and *without labeled training data*. They are compared against a majority class baseline that labels all words as non-shifters.

**Label Propagation (LP):** Given a set of seed words and a word-similarity graph, the classifier propagates the seed labels across the graph, thereby labeling the remaining words. It requires no labeled training, as our seeds are automatically determined by heuristics. We use the *Adsorption* LP algorithm (Baluja *et al.* 2008) as implemented in `Junto` (Talukdar *et al.* 2008). All hyperparameters are kept at their default. As word-similarity graph, we use the same graph as for the *PageRank* computation in Section 5.1. As shifter seeds, we use the 250 highest-ranked words from $\text{ANY}_{\text{polar+pageR}}$, our best task-specific shifter feature. As non-shifter seeds, we use 500 words[g] that can be considered anti-shifters, that is, words that create a strong polar stability which prevents shifting. Similar to ANY, we determine anti-shifters through the use of co-occurrence patterns. We seek words that co-occur with the adverbials *exclusively*, *first*, *newly*, and *specially*, as they show attraction to verbs of creation (non-shifters) while at the same time being repelled by verbs of destruction (shifters).

A major advantage of LP is that it works without explicitly labeled training data, thus avoiding the cost of expert annotation. While our automatically generated seeds are far from flawless ($\text{ANY}_{\text{polar+pageR}}$ provided only 45.2% precision), they are the best seeds available to us without resorting to fully supervised methods. Indeed, we will find in Section 5.2.2 that, given the quality of the seeds, performance for verb classification is surprisingly good. We also experimented with using fewer seeds, as the precision of $\text{ANY}_{\text{polar+pageR}}$ is better at earlier cutoffs, but this provided no improvements.

For nouns and adjectives, we lack strong enough features for LP. ANY suffers from sparsity issues and our anti-shifter patterns are only applicable to verbs. We therefore provide LP exclusively for verbs.

**Mapping from verb lexicon (VerbLex):** To make up for the lack of a LP classifier for nouns and adjectives, we introduce the cross-POS feature as a stand-alone classifier. While it relies heavily on information about verbal shifters, no additional labeled training data for nouns or adjectives is required.

**Support vector machine (SVM):** As supervised classifier, we choose an SVM as implemented in `SVM`$^{\texttt{light}}$ (Joachims 1999). SVMs require labeled training data but allow arbitrary feature combinations (unlike LP, which must encode all information in the choice of seeds and graph weights). We evaluate a number of feature combinations. We train one classifier using the task-specific features ($\text{SVM}_{\text{T}}$) from Section 4.2 and one using generic features ($\text{SVM}_{\text{G}}$) from Section 4.3.[h] A third classifier combines both feature sets ($\text{SVM}_{\text{T+G}}$). For nouns and adjectives, we also add the output of the best-performing cross-POS feature from Section 4.4 ($\text{SVM}_{\text{T+G+V}}$).

[g] We use twice as many non-shifters as we use shifters to account for the higher frequency of non-shifters while avoiding overfitting to the statistics from Table 1.

[h] We also performed an ablation study to see whether any individual features were detrimental or irrelevant to the performance of our classifier. However, all features contributed positively, even those with weak individual performance, so we chose to include all of them in their respective SVM feature sets.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 4.** Classification of polarity shifters for individual parts of speech. SVM features are grouped as task-specific (T), generic (G), and VerbLex (V). The evaluation is run as a 10-fold cross validation. All metrics are macro-averages.

| | Verbs | | | Nouns | | | Adjectives | | |
|---|---|---|---|---|---|---|---|---|---|
| Classifier | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| $Baseline_{maj}$ | 42.4 | 50.0 | 45.9 | 47.3 | 50.0 | 48.6 | 46.8 | 50.0 | 48.3 |
| LP | 68.6 | 56.7 | 62.0* | *n/a* | *n/a* | *n/a* | *n/a* | *n/a* | *n/a* |
| VerbLex | *n/a* | *n/a* | *n/a* | 82.6 | **74.1** | 78.1* | 66.0 | 56.9 | 61.0* |
| $SVM_T$ | 65.5 | 69.7 | 67.5* | 62.1 | 61.9 | 61.9 | 59.4 | **66.9** | 62.9* |
| $SVM_G$ | 79.6 | 74.4 | 76.9* | 70.1 | 56.6 | 62.4 | **74.4** | 60.5 | 66.3† |
| $SVM_{T+G}$ | **80.7** | **77.6** | **79.1*** | 70.4 | 57.6 | 63.1 | 72.8 | 62.1 | **66.6†** |
| $SVM_{T+G+V}$ | *n/a* | *n/a* | *n/a* | **84.6** | 73.8 | **78.7*** | 70.0 | 62.9 | **66.1†** |

*: F1 is better than previous classifier (paired *t*-test with $p < 0.05$).
†: F1 is better than VerbLex (paired *t*-test with $p < 0.05$).

$SVM^{light}$ is configured to use a cost factor for training errors on positive examples of $j = 5$ (default is $j = 1$). In our case, positive examples, that is, shifters, are the minority, so the heightened cost factor prevents the classifier from always favoring the Non-shifter majority. All other hyperparameters are kept at their default.

Training and testing are performed using 10-fold cross-validation on the 2000 gold standard words of the respective part of speech. We report the averaged performance across the 10 runs.

#### *5.2.2 Evaluation of classifiers*

Table 4 shows the performance of the classifiers on our polarity shifter gold standard (Section 3.1), presenting macro-averaged results for each part of speech.

**Classification of verbs:** Both LP and SVM clearly outperform our baseline. Furthermore, all versions of SVM outperform LP, indicating that labeled training data is beneficial and that a combination of features is better than only the strongest feature. While the generic features ($SVM_G$) outperform the task-specific ones ($SVM_T$), combining both provides a significant performance boost ($SVM_{T+G}$).

**Classification of nouns:** Comparing SVM performance of verbs and nouns, we see that the noun classifier is less successful, mainly due to the considerably lower performance of the generic features ($SVM_G$), likely caused by the lower frequency of shifters among nouns (5% instead of 15%). Transferring verb labels to nouns via VerbLex provides performance similar to that of the best verb classifier.

**Classification of adjectives:** For adjectives, the performance of task-specific and generic features is similar to that for nouns. VerbLex performs worse than it did for nouns, as it has to use the fallback stem-matching approach (see Section 4.4). Nevertheless, the general idea of relatedness across parts of speech still holds and the performance of VerbLex is significantly above the baseline.

### *5.3 How much training data is required?*

We use automatic classification to reduce the amount of human annotation required for a lexicon of adequate size. At the same time, our SVM classifiers require annotated training data to create

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

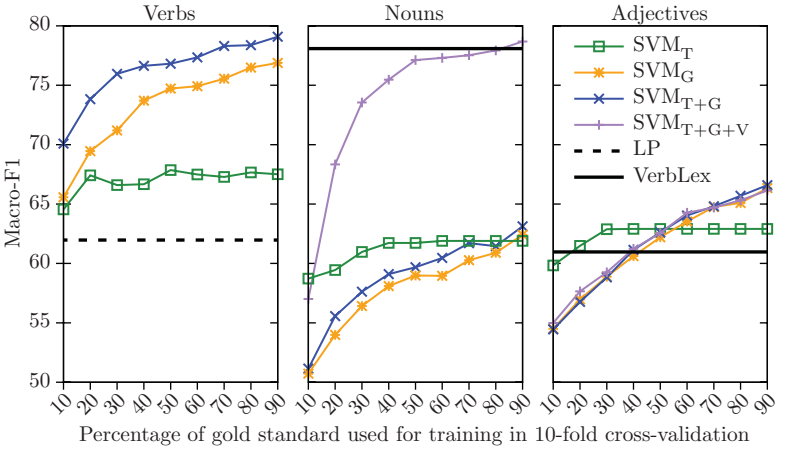



**Fig. 2.** Learning curves for supervised training. This repeats the evaluation of Section 5.2 but reduces the amount of training data. At 90% training data, this task is identical to the one reported in Table 4.

high-quality candidate lists for the verification step. We are faced with a trade-off between pre- and post-processing annotation efforts.

Figure 2 shows a learning curve of how the classifiers from Table 4 perform with varying amounts of training data. LP and mapping from the verb lexicon (VerbLex) require no labeled training data so their performance is constant. For all parts of speech, task-oriented features (SVM$_T$) reach their full potential early on, but plateau in performance. Generic features (SVM$_G$) require larger amounts of training data to compensate for the low frequency of shifters, especially for nouns and adjectives. Regarding verbs, SVM is able to outperform LP even with small amounts of training data. In Table 4, it seemed that SVM$_{T+G+V}$ offered no improvement over VerbLex for noun classification, but now we can see that this was a side effect of the training size and may change with more training data. For adjectives, on the other hand, combining feature groups provides no improvement. This suggests that adjectival shifters operate under different conditions than verbal and nominal shifters.

## 6. Bootstrapping the lexicon

In this section, we bootstrap the remaining unlabeled vocabulary of 8581 verbs, 53,311 nouns, and 16,282 adjectives. For this, we train classifiers for each part of speech on their full gold standard of 2000 words. All words that the classifiers *predict* to be shifters are then verified to remove false-positive classifications (Section 6.1). This is done by the same expert annotator who worked on the gold standard. Words predicted to be non-shifters are not considered further. This classifier-based pre-filtering approach (Choi and Wiebe 2014) allows us to ensure the high quality of the lexicon while keeping the annotation workload manageable. The verified bootstrapped shifters are then combined with those from the gold standard to create a single large polarity shifter lexicon (Section 6.2).

### *6.1 Evaluation of bootstrapping*

For bootstrapping verbs, we use SVM$_{T+G}$ as it is clearly the best available classifier (see Table 4). For nouns and adjectives, the choice is not as clear as the VerbLex classifier introduces a strong new resource: our own verbal shifter lexicon. For nouns, the unsupervised VerbLex outperformed our supervised classifier SVM$_{T+G}$, suggesting that training data for nominal shifters may

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

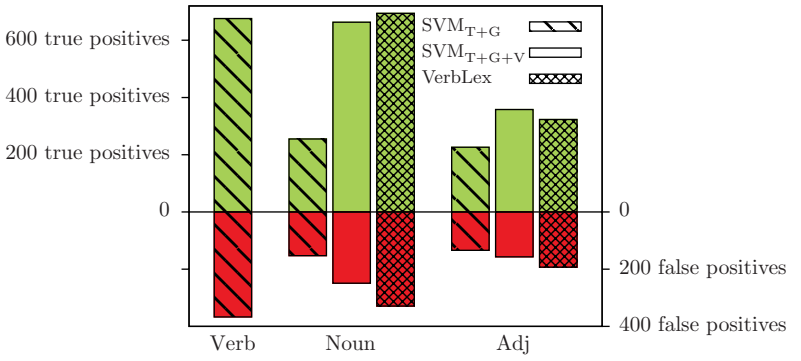



**Fig. 3.** Bootstrapping of shifters that were not part of the gold standard (compare Table 1). Each bar represents the number of words that a classifier predicted to be shifters, separated by how many of them are actually shifters (true positives) and how many are misclassified non-shifters (false positives).

be unnecessary. For adjectives, on the other hand, the question is whether VerbLex can contribute to the supervised classifier at all, as SVM$_{T+G}$ outperforms SVM$_{T+G+V}$. To investigate both these questions further, we run three separate bootstrapping classifiers for nouns and adjectives: SVM$_{T+G}$, SVM$_{T+G+V}$, and VerbLex.

#### *6.1.1 Coverage-oriented evaluation*

Figure 3 illustrates the number of *predicted* shifters returned by the bootstrap classifiers. The overall size of a bar indicates the number of words predicted to be shifters. The bar is divided into true positives (actual shifters, as confirmed by a human annotator) and false positives (non-shifters mislabeled by the classifier).

We see that the SVM$_{T+G}$ classifier is considerably more conservative for nouns and adjectives than it is for verbs, labeling only 408 nouns and 360 adjectives as shifters, compared to the 1043 verbs. This is likely due to the lower frequency of shifters among those parts of speech in our training data (see Table 1). Adding VerbLex information to the classifier helps with this issue, as the output of SVM$_{T+G+V}$ shows. Compared to VerbLex, SVM$_{T+G+V}$ filters out more false-positive nouns without dropping many true positives. Therefore, its output has a higher precision, reducing the verification effort without incurring large losses in recall. For adjectives, this is even more obviously true, as SVM$_{T+G+V}$ has the same number of predicted shifters as VerbLex but contains 35 more true positives.

Comparing the output overlap of VerbLex and SVM$_{T+G}$, we find that just 27% of nouns and 59% of adjectives are returned by both classifiers. Unsurprisingly, SVM$_{T+G+V}$ overlaps almost entirely with the other classifiers, so its strength lies in improving precision instead. We conclude that for noun and adjective classification, VerbLex is the best starting point, but to increase coverage, supervised classification via SVM$_{T+G+V}$ offers the best balance of precision and recall. This may of course differ for languages with different typological properties. For a closer look at how the availability of mono- and cross-lingual resources affects the classification of shifters, we refer the reader to Schulder *et al.* (2018a).

#### *6.1.2 Precision-oriented evaluation*

Our SVM classifiers provide a confidence value for each label they assign. In Figure 4, we inspect whether higher confidences also translate into higher precision. For this, we rank the bootstrapped shifters of each classifier by their confidence value and then split them into four groups, from highest to lowest confidence. As VerbLex provides no confidence values, we limit this evaluation to the SVM classifiers.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

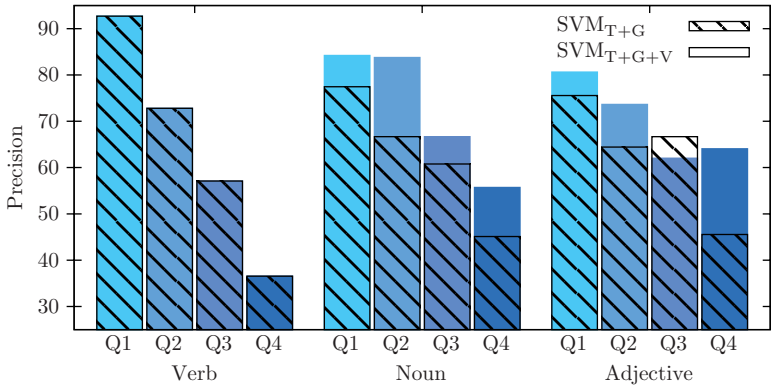



**Fig. 4.** Evaluation of the bootstrapping of shifters that were not part of the gold standard (see Section 3.1). SVM classifiers provide a confidence value for each label they assign. We split the set of potential shifters into quarters, sorting them from highest (Q1) to lowest (Q4) confidence. We report precision for each quarter.

We see a clear trend that high confidence means high precision. When employing a human annotator for a manual verification step is not feasible, a high precision lexicon can be ensured by limiting it to only items of high confidence. SVM$_{T+G+V}$ profits from the addition of VerbLex information, improving precision over SVM$_{T+G}$ in all but onequarter (while significantly increasing recall, as we saw in Figure 3).

### *6.2 Creating the complete lexicon*

With the bootstrapping process complete, we now consolidate our data into a single shifter lexicon containing all words verified by a human, that is, all words from the gold standard (Section 3.1) and all bootstrapped words (Section 6). In the case of nouns and adjectives, we combine the output of the three evaluated bootstrap classifiers.

Table 5 shows the annotation effort for each dataset and its balance of shifters versus Non-shifters. The benefit of the bootstrapping process is clearly visible. The percentage of shifters among bootstrap data is far higher than that among the randomly sampled gold standard. While the amount of bootstrap data that had to be annotated was roughly half of what was annotated for the gold standard, it contains more than twice as many verbal shifters, over seven times as many nominal shifters and four times as many adjectival shifters.

Our bootstrapping produced 1981 shifters among 3145 words. Based on the gold standard shifter frequencies, we assume that to find as many shifters by blindly annotating a random set of words, we would have had to annotate 24,000 additional words. Taking into account the 6000 words annotated for the gold standard, our approach reduces the annotation effort by 72%, a saving of 680 work hours.

## 7. Impact on sentiment analysis

Our main motivation for creating a large lexicon of shifters is to improve sentiment analysis applications. In this section, we investigate whether knowledge of shifters offers such improvements for *phrase-level polarity classification*. Apart from being an intermediate step in compositional sentence-level classification (Socher *et al.* 2013), polarity classification for individual phrases has been used for knowledge base population (Mitchell 2013), summarization (Stoyanov and Cardie 2011), and question answering (Dang 2008).

In this experiment, we deliberately decided against using established datasets annotated for sentiment at the sentence level. Sentence-level sentiment classification requires several linguistic

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 5.** Result of the lexicon generation workflow outlined in Figure 1. The complete lexicon contains both the gold and bootstrap lexica.

| | | Verbs | | Nouns | | Adjectives | |
|---|---|---|---|---|---|---|---|
| | | Frequency | Percentage | Frequency | Percentage | Frequency | Percentage |
| Gold | Annotated | 2000 | | 2000 | | 2000 | |
| | Shifter | 304 | 15.20 | 107 | 5.35 | 129 | 6.45 |
| | Non-shifter | 1696 | 84.80 | 1893 | 94.65 | 1871 | 93.55 |
| Bootstrap | Annotated | 1043 | | 1270 | | 832 | |
| | Shifter | 676 | 64.81 | 793 | 62.44 | 512 | 61.54 |
| | Non-shifter | 367 | 35.19 | 477 | 37.56 | 320 | 38.46 |
| Complete | Annotated | 3043 | | 3270 | | 2832 | |
| | Shifter | 980 | 32.21 | 900 | 27.52 | 641 | 22.63 |
| | Non-shifter | 2063 | 67.79 | 2370 | 72.48 | 2191 | 77.37 |

phenomena apart from shifting, making it too coarse for a focused evaluation. One might argue that the impact of shifter knowledge should be observable by providing it to a sentence-level classifier and judging its change in performance. However, classification of sentence-level sentiment is often facilitated by grasping the polarity of the more salient polar expressions of a sentence. For instance, in (47), a classifier does not have to correctly identify the polarity shifting caused by *dropped*. Instead, it can already read off the sentence-level polarity from the most salient polar expression, *luckily*. Typically, such heuristics are implicitly learned by sentence-level classifiers.

(47) $[[\text{Luckily}]^{+}, [\text{all } [\text{charges}]^{-} \text{ were } \mathbf{dropped}_{\textit{shifter}}]^{+}]^{+}$.

While cases like (47) might be frequent enough at the sentence level to dilute the apparent relevance of shifters, phrase-level classification offers far fewer such shortcuts, instead requiring a thorough handling of all involved phenomena.

In Section 7.1, we describe the experimental design of our classification task. Section 7.2 compares the performance of state-of-the-art compositional polarity classifiers (without knowledge of shifters) with an approach that uses our shifter lexicon. In Section 7.3, we compare our lemma-based shifter lexicon to a lexicon that was annotated for individual word senses (Schulder *et al.* 2018b).

### *7.1 Experimental setup*

The question we seek to answer in this experiment is whether a given classifier can correctly decide if the polarity of a word has changed in the context of a phrase. For example, the noun *passion* is of positive polarity, but in (48) it is shifted by the verb *lack*, resulting in the negative polarity phrase *"lack her usual passion"*.

(48) The book seemed to $[\mathbf{lack}_{\text{V}} \text{ [her usual } \mathbf{passion}^{+}_{\text{N}}]_{\text{NP}}]^{-}_{\text{VP}}$.

We limit this evaluation to cases involving verbal shifters. This ensures a fair comparison to the sense-level lexicon in Section 7.3, which only covers verbs. As previously argued by Schulder *et al.* (2018b), verbs are the most important part of speech for handling shifting, as they are the main syntactic predicates of clauses and sentences, which gives them the largest scope. Together with

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 6.** Annotation example for the sentiment analysis impact evaluation. The annotator determines the polarities of polar noun and verb phrase given the sentence context. If the polarities differ, the shifting label is *"shifted"* else it is *"not shifted"*.

| Field | Polarity | Content |
|---|---|---|
| Verb | | *"soothe"* |
| Sentence | | *"Norah Jones' voice could soothe any savage beast."* |
| Polar noun | negative | *"beast"* |
| Verb phrase | positive | *"soothe any savage beast"* |
| Shifting label | | shifted |

nouns, they are the most important minimal semantic units in text (Schneider *et al.* 2016). Verbs, however, occur more often with the syntactic arguments required for shifting than nouns do, that is, verbal shifters usually have subjects and objects whose polarity they can shift, but nominal shifters often occur without compound modifiers or prepositional objects, so there will be nothing to shift.

We select sentences in which the scope of the (potential) shifter is a noun. As expressions with neutral polarity are not expected to be affected by shifters, we require the noun to be of positive or negative polarity (as determined through the *Subjectivity Lexicon*). We do not consider sentences that also contain a negation word to avoid the complication of multiple polarity shifts canceling each other out. This gives us a verb phrase (VP) that contains a verb (the potential shifter) and a polar noun. The question that must be answered in our evaluation is whether the polarity of the polar noun and the VP are the same or different. In other words, whether the polarity has *"shifted"* or *"not shifted"*.

To create a dataset for the evaluation, we extract sentences from the *Amazon Product Review Data* text corpus (see Section 3.2) that contain a VP headed by a verb that has a polar noun as a dependent. The polarity of the noun is determined using the *Subjectivity Lexicon*. We begin by annotating 400 sentences in which the verb is a polarity shifter according to our shifter lexicon. Next, we annotate 2231 sentences where the verb is a non-shifter. This makes the distribution of verbal shifters and non-shifters in the sentences match that in the bootstrapping gold standard (Table 1). Since shifters are content words, they follow a power-law distribution (Zipf 1935). One advantage of our approach is that it is designed to take the long tail of such distributions into account. To cover such a variety of different shifters, rather than the most frequent ones, each shifter therefore only occurred once in our data. The annotation was performed by one of the authors. To determine inter-annotator agreement, another author also annotated 10% of the data (263 instances), resulting in a substantial agreement of Cohen's kappa of 0.72 (observed agreement of 91.3%). Among the 23 instances of disagreement, by far the most concern the polarity of the VP. There are just two instances in which the annotators only disagreed on the polarity of the noun.

Table 6 shows the fields of information provided to the annotator during creation of the dataset. For each sentence, the annotator chooses the polarities of the given noun and VP, labeling each as either *"positive"*, *"negative"*, or *"neutral"*. If a polarity would evaluate differently for the speaker, than for an event participant the sentiment of the speaker is annotated (e.g., *"She adores her idiot husband"* would be annotated as negative, as only the wife (the event participant) adores the husband, while the speaker considers the husband an idiot). The full sentence is provided to clarify the context in which the phrase is set. The verb that is the head of the VP (and therefore the potential shifter) is also explicitly defined to avoid confusion in cases where more than one verb occurs in the phrase. The field *shifting label* shows the label that classifiers will have to determine

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

in our evaluation. In the example in Table 6, the annotator labels the noun *beast* as negative and the VP *soothe any savage beast* as positive. Based on these polarities, the shifting label of the sentence is determined to be *"shifted"*. Classifiers may either provide the shifting label directly or provide noun and VP polarities. If the polarities are identical, the label is *"not shifted"*, otherwise it is *"shifted"*.

There is no consensus on how to model how far the polarity of a shifted word moves (see Section 2.1.2). Expressions like *"it wasn't excellent"* have been argued to convey either positive (Choi and Cardie 2008) or neutral polarity (Taboada *et al.* 2011; Kiritchenko and Mohammad 2016). Our evaluation is concerned with whether shifting occurs, rather than with the exact polarities or polar intensities involved. To accommodate both legitimate interpretations, we count both behaviors as shifting. As long as the polarity of the polar noun and that of the VP are not identical, we consider it shifted. Our own approach does not profit from this, as its decisions are solely driven by its knowledge of shifters and not by the polarities involved.

### *7.2 Comparison to existing methods*

In this section, we investigate whether knowledge of polarity shifters can be used to improve polarity classification. As baselines, we use a majority class classifier that labels all sentences as *"not shifted"* and two neural network classifiers optimized for handling negation at the phrase level: The *Recursive Neural Tensor Network tagger* (`RNTN`) by Socher *et al.* (2013) and the *ELMo bi-attentative classification model* (`ELMo`) by Peters *et al.* (2018).

`RNTN` is a compositional sentence-level polarity classifier that achieves strong performance on polarity classification datasets. Given the constituency parse of a sentence, it determines the polarity of each tree node. This allows us to extract the polarities it assigns to the relevant nouns and VPs in our data. `ELMo`, a recent classifier that uses bidirectional LSTMs, has been shown to further improve performance when applied to the same task.

One of the major strengths of `RNTN` (and by extension `ELMo`) is that it can learn polarity shifting effects caused by negation words implicitly from labeled training data, rather than requiring explicit knowledge of shifters or shifting rules. It does, however, require training sentences in which each node of a constituency parse tree has been labeled with polarity information. The creation of such data is expensive. To date, the only manually annotated dataset that provides such fine-grained polarity information is the *SST* (Socher *et al.* 2013), a set of 11,855 sentences from movie reviews. Unfortunately, resources like *SST* do not contain most shifters with sufficient frequency to train or test the ability of a classifier to handle polarity shifters. For example, *SST* only contains instances of 30% of the polarity shifters from our lexicon. Over a third of these occur only a single time. `RNTN` and `ELMo` are both trained on *SST*.

We use precomputed models and default hyperparameters as provided by the authors. The `RNTN` model of Socher *et al.* (2013) is available as part of *CoreNLP* (Manning *et al.* 2014)[i] and the `ELMo` model of Peters *et al.* (2018) is available as part of *AllenNLP* (Gardner *et al.* 2018).[j]

Our own approach (LEX) first determines the polarity of a given noun using the *Subjectivity Lexicon* and infers the polarity of the VP through our knowledge of polarity shifters. If the head verb of the VP is a shifter according to our lexicon, then the polarity of the VP is set to be the opposite of the polarity of the noun. If the verb is a non-shifter, then the VP receives the same polarity as the noun.

We evaluate our approach with two versions of our bootstrapped shifter lexicon from Section 6. $\text{LEX}_{\text{SVM}}$ uses the output of our best SVM classifier before its shifters were verified by a human annotator. $\text{LEX}_{\text{gold}}$ uses the final version of the shifter lexicon after the verification step. It should be considered as an upper bound to the expected performance of the LEX approach.

[i] http://nlp.stanford.edu/software/stanford-corenlp-full-2014-08-27.zip
[j] https://allennlp.s3.amazonaws.com/models/sst-5-elmo-biattentive-classification-network-2018.09.04.tar.gz

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 7.** Classifier performance for sentiment analysis task of determining whether shifting occurs between a polar noun and the VP that contains it (see Section 7.1).

| | Classifier | Precision | Recall | F1 |
|---|---|---|---|---|
| Baseline | Majority | 39.95 | 50.00 | 44.41 |
| | RNTN | 50.81 | 51.16 | 50.98* |
| | ELMo | 54.65 | 56.72 | 55.67* |
| Lemma Lexicon | $\text{LEX}_{\text{SVM}}$ | 81.63 | 80.95 | 81.29* |
| | $\text{LEX}_{\text{gold}}$ | **88.85** | **81.18** | **84.84*** |

*: F1 is better than previous classifier (paired permutation test with $p < 0.05$).

Results in Table 7 show that our approach clearly outperforms the baselines. RNTN and ELMo perform above the majority class baseline, but fail to detect most instances of shifting. $\text{LEX}_{\text{SVM}}$ provides a significant improvement over the other classifiers, identifying most instances of shifting correctly and coming fairly close to the upper bound of $\text{LEX}_{\text{gold}}$. Even $\text{LEX}_{\text{gold}}$ still contains a number of misclassifications, however. One potential reason for false positives is that some verbs are shifters in only some of their word senses. This issue is addressed in the upcoming section.

While RNTN performs slightly better than the majority class classifier, it still fails to detect most instances of shifting. One contributing factor is a lack of knowledge about shifters. The instances where RNTN makes mistakes are more likely to involve a shifter candidate that is not encountered in its *SST* training data (43.4% of false positives and 53.1% of false negatives). By contrast, the true positives that RNTN produces involve an unknown shifter candidate in only 11% of the cases. In the case of the true negatives, unexpectedly, the shifter candidate is unknown for 45.1% of the instances. Closer inspection shows that RNTN succeeds by sheer luck: since our test set presents it with many unseen polar nouns, it has many combinations of neutral-by-default nouns combined with neutral-by-default governing verbs, yielding a prediction of non-shifting, which happens to be correct. Notably, RNTN has such neutral–neutral combinations for 68.9% of the instances, whereas neutral–neutral combinations occur only in 11.4% of the cases in the gold standard. A second source of error is, unsurprisingly, the polar nouns. In the *SST*, on which RNTN is trained, many nouns that have a clear prior polarity behave idiosyncratically. For instance, *arrogance* is slightly positive in *SST*. Accordingly, RNTN erroneously sees shifting in our test instance *"Their marketing driven approach smacks of arrogance"*, where it recognizes *smack* and the larger VP *"smacks of arrogance"* as negative but the noun as positive.

Compared to $\text{LEX}_{\text{SVM}}$, RNTN produces many more false positives (6.3 : 1) by shifting when it should not and false negatives (1.4 : 1) by not shifting when it should shift. Much of this difference can be attributed to the fact that RNTN gets the correct polarity for only 1243 of the 2631 noun instances (47.2%). It classifies a large number of nouns as neutral. By contrast, the lexical lookup employed with $\text{LEX}_{\text{SVM}}$ gets the correct polarity for 2239 instances (85.1%). It is notable that $\text{LEX}_{\text{SVM}}$ never predicts any neutral VPs, which does not hurt performance much because there are few neutral VPs in the gold standard. RNTN, by contrast, massively overpredicts neutral polarity for VPs.

### *7.3 Comparison to sense-level lexicon*

Some words only cause shifting in some word senses. For example, *mark down* shifts in its sense of *"reduce in value"* in (49), but not in its sense of *"write down"* in (50).

(49) The agency $[[\textbf{marked down}]_{\text{V}}$ their $\textbf{assets}^{+}_{\text{N}}]^{-}_{\text{VP}}$.

(50) She $[[\textbf{marked down}]_{\text{V}}$ her $\textbf{highscore}^{+}_{\text{N}}]^{+}_{\text{VP}}$ in the rankings.

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

**Table 8.** Comparison between lemma- and sense-level shifter lexica on the sentiment analysis task (see Section 7.1). $\text{SENSE}_{\text{first}}$ assigns the first *WordNet* sense to each verb. $\text{SENSE}_{\text{oracle}}$ always chooses an appropriate word sense where possible.

| | Classifier | Precision | Recall | F1 |
|---|---|---|---|---|
| Sense Lexicon | $\text{SENSE}_{\text{first}}$ | 85.71 | 67.10 | 75.27 |
| | $\text{SENSE}_{\text{oracle}}$ | **92.45** | 74.34 | 82.41* |
| Lemma Lexicon | $\text{LEX}_{\text{gold}}$ | 88.85 | **81.18** | **84.84*** |

*: F1 is better than previous classifier (paired permutation test with $p < 0.05$).

When designing our approach for bootstrapping a shifter lexicon, we chose to assign a single label per lemma to avoid reliance on word sense disambiguation (see Section 3.1). Independently from the semi-automatic approach presented in this article, we also created a sense-level lexicon of verbal shifters (Schulder *et al.* 2018b) which was created entirely by hand and only contains verbs. Its sense inventory is based on the synset definitions of *WordNet* and provides an individual shifter label for each sense of a verb. We define a new classifier SENSE, which works like LEX except that it uses this sense-level lexicon and requires a word sense to be chosen for each potential shifter. $\text{SENSE}_{\text{first}}$ chooses the first sense of each word. As *WordNet* senses are ordered to list common uses first, this makes for a stronger baseline than randomly choosing a sense. To establish an upper bound for the performance of the sense-level lexicon, we also provide $\text{SENSE}_{\text{oracle}}$, which always chooses an appropriate word sense. Note that this does not guarantee perfect performance, as the oracle can only make a choice if a lemma has both shifter and non-shifter senses.

Table 8 compares the sense-level lexicon classifiers (SENSE) with $\text{LEX}_{\text{gold}}$. Looking at $\text{SENSE}_{\text{oracle}}$, we see that differentiating by word sense can improve precision by avoiding false-positive hits, but harms recall. This is due to systematic flaws in the coverage of the sense-level lexicon of Schulder *et al.* (2018b) caused by its reliance on the sense inventory of *WordNet*. In some cases, specific meanings of a word are simply not defined in *WordNet* and therefore missing from the sense-level lexicon. *WordNet* for example only lists the literal sense of *to derail* that pertains to trains, but not the metaphoric sense found in *"derail your chance of success"*. In other cases, the definition of a sense is so broad or vague that it is unclear which uses of the lemma should be considered. The use of sense-level labels without strong word sense disambiguation is also clearly detrimental, as shown by $\text{SENSE}_{\text{first}}$.

## 8. Conclusion

We presented a bootstrapping approach for creating a lexicon of English polarity shifters, using task-specific features, general semantic resources, and label mapping across different parts of speech. Words predicted to be shifters are verified by a human annotator to ensure a high-quality lexicon. Using this approach, we create a large lexicon of 2521 shifters while reducing the required manual annotation effort by over 72%. It is the first shifter lexicon of a larger size that covers not only verbs but also nouns and adjectives. The entire lexicon is made publicly available.

Our bootstrapping approach employs a variety of linguistic phenomena and resources. These include clashing polarities and the semantic properties of verb particles. Our strongest unsupervised feature, co-occurrence with the NPI *any*, is inspired by research on DE-Ops. For supervised classification, we also use semantic information from *WordNet* and *FrameNet*.

In expanding our bootstrapping efforts from verbs to nouns and adjectives, we observe both similarities and differences between the parts of speech. Nouns and adjectives offer a greater lexical variety than verbs, making it even more important to aid the annotation process through bootstrap classification. The number of shifters relative to the size of the vocabulary, however, is considerably lower in these parts of speech, making classification more challenging. To overcome

https://doi.org/10.1017/S135132492000039X Published online by Cambridge University Press

this challenge, we introduce new features that use the lexicon of verbal shifters which we already bootstrapped, transferring its shifter labels from verbs to nouns and adjectives. This proves an excellent feature for labeling nouns, due to the availability of derivational relatedness mappings between verbs and nouns. Results for adjectives are more mixed, as there are no equivalent mapping resources available for them, so using an ensemble of features for supervised classification remains the best approach.

In a polarity classification task, we show that our lexicon helps to avoid classification errors involving shifters which a state-of-the-art classifier is unable to handle. We also examine whether fine-grained labels for individual word senses further enhance performance but do not find evidence for this.

The obvious next step is to make use of the shifter lexicon that we have created, for example, by integrating it into sentiment analysis pipelines. It may also be of help in inference, as polarity shifters indicate negated conditions, which usually entail semantic phenomena such as downward entailment. Another future task is to use our bootstrapping approach to create shifter lexica for additional languages. In Schulder *et al.* (2018a), we showed the feasibility of mono- and cross-lingual bootstrapping for German, with a focus on the impact of resource availability. It will be especially interesting to see how well features adapt to non-Indo-European languages.

**Acknowledgments.** We would like to thank Amy Isard for her invaluable feedback during the writing and revision of this article. We would also like to thank the anonymous reviewers for their constructive comments.

**Financial support.** The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

## References

**Agirre E. and Soroa A.** (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of EACL*. Athens, Greece: ACL, pp. 33–41.

**Anand P. and Reschke K.** (2010). Verb classes as evaluativity functor classes. In *Proceedings of VERB*. Pisa, Italy: Scuola Normale Superiore, pp. 98–103.

**Baker C.F., Fillmore C.J. and Lowe J.B.** (1998). The Berkeley FrameNet project. In *Proceedings of COLING*. Vancouver, Canada: ACL, pp. 86–90.

**Baker C.L.** (1970). Double negatives. *Linguistic Inquiry* **1**(2), 169–186.

**Baluja S., Seth R., Sivakumar D., Jing Y., Yagnik J., Kumar S., Ravichandran D. and Aly M.** (2008). Video suggestion and discovery for Youtube: Taking random walks through the view graph. In *Proceedings of WWW*. Beijing, China: ACM, pp. 895–904.

**Brinton L.J.** (1985). Verb particles in English: Aspect or Aktionsart. *Studia Linguistica* **39**(2), 157–168.

**Chen D. and Manning C.D.** (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*. Doha, Qatar: ACL, pp. 740–750.

**Choi Y. and Cardie C.** (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of EMNLP*. Honolulu, Hawaii: ACL, pp. 793–801.

**Choi Y. and Wiebe J.** (2014). +/–EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of EMNLP*. Doha, Qatar: ACL, pp. 1181–1191.

**Choi Y., Deng L. and Wiebe J.** (2014). Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events. In *Proceedings of WASSA*. Baltimore, MD, USA: ACL, pp. 107–112.

**Cohen J.** (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46.

**Danescu-Niculescu-Mizil C., Lee L. and Ducott R.** (2009). Without a 'doubt'? Unsupervised discovery of downward-entailing operators. In *Proceedings of NAACL-HLT*. Boulder, CO, USA: ACL, pp. 137–145.

**Dang H.T.** (2008). Overview of the TAC 2008 opinion question answering and summarization tasks. In *Proceedings of TAC*. Gaithersburg, MD, USA: NIST.

**de Marneffe M.-C. and Manning C.D.** (2008). *Stanford Typed Dependencies Manual*. Technical report. Stanford University.

**Deng L. and Wiebe J.** (2014). Sentiment propagation via implicature constraints. In *Proceedings of EACL*. Gothenburg, Sweden: ACL, pp. 377–385.

**Deng L., Choi Y. and Wiebe J.** (2013). Benefactive/Malefactive event and writer attitude annotation. In *Proceedings of ACL*. Sofia, Bulgaria: ACL, pp. 120–125.

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

**Fillmore C.J.** (1967). The case for case. In Bach E. and Harms R. (eds), *Proceedings of the Texas Symposium on Language Universals*. New York, NY, USA: Holt, Rinehart and Winston, pp. 1–90.

**Firth J.R.** (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32.

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

**Gardner M.**, **Grus J.**, **Neumann M.**, **Tafjord O.**, **Dasigi P.**, **Liu N.F.**, **Peters M.**, **Schmitz M.**, and **Zettlemoyer L.** (2018). AllenNLP: A deep semantic natural language processing platform. In *Proceedings of NLP-OSS*. Melbourne, Australia: ACL, pp. 1–6.

**Giannakidou A.** (2011). Negative and positive polarity items: Variation, licensing, and compositionality. In Maienborn C., von Heusinger K. and Portner P. (eds), *Semantics: An International Handbook of Natural Language Meaning*, vol. **2**. Berlin, Germany: Mouton de Gruyter, pp. 1660–1712.

**Hasan K.S. and Ng V.** (2013). Frame semantics for stance classification. In *Proceedings of CoNLL*. Sofia, Bulgaria: ACL, pp. 124–132.

**Jiang T. and Riloff E.** (2018). Learning prototypical goal activities for locations. In *Proceedings of ACL*. Melbourne, Australia: ACL, pp. 1297–1307.

**Jindal N. and Liu B.** (2008). Opinion spam and analysis. In *Proceedings of WSDM*. Palo Alto, CA, USA: ACM, 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*. MIT Press, pp. 169–184.

**Kang J.S.**, **Feng S.**, **Akoglu L. and Choi Y.** (2014). ConnotationWordNet: Learning connotation over the word+sense network. In *Proceedings of ACL*. Baltimore, Maryland: ACL, pp. 1544–1554.

**Kim S.-M. and Hovy E.** (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. Sydney, Australia: ACL, pp. 1–8.

**Kim S.**, **Chang H.**, **Lee S.**, **Yu M. and Kang J.** (2015). Deep semantic frame-based deceptive opinion spam analysis. In *Proceedings of CIKM*. Melbourne, Australia: ACM, pp. 1131–1140.

**Kiritchenko S. and Mohammad S.M.** (2016). The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of WASSA*. San Diego, CA, USA: ACL, pp. 43–52.

**Ladusaw W.** (1980). *Polarity Sensitivity as Inherent Scope Relations. Outstanding Dissertations in Linguistics*. New York: Garland Publishing.

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

**Linebarger M.C.** (1980). The Grammar of Negative Polarity.

**Liu Y.**, **Yu X.**, **Liu B. and Chen Z.** (2014). Sentence-level sentiment analysis in the presence of modalities. In *Proceedings of CICLing*. Kathmandu, Nepal: Springer, pp. 1–16.

**Liu B.** (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* **5**(1), 1–167.

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

**Manning C.**, **Surdeanu M.**, **Bauer J.**, **Finkel J.**, **Bethard S. and McClosky D.** (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL*. Baltimore, Maryland, USA: ACL, pp. 55–60.

**Mikolov T.**, **Chen K.**, **Corrado G. and Dean J.** (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*. Scottsdale, AZ, USA.

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

**Mitchell M.** (2013). Overview of the TAC2013 knowledge base population evaluation: English sentiment slot filling. In *Proceedings of TAC*. NIST.

**Moilanen K. and Pulman S.** (2007). Sentiment construction. In *Proceedings of RANLP*. Borovets, Bulgaria: ACL.

**Morante R. and Daelemans W.** (2009). A metalearning approach to processing the scope of negation. In *Proceedings of CoNLL*. Boulder, CO, USA: ACL, pp. 21–29.

**Morante R.** (2010). Descriptive analysis of negation cues in biomedical texts. In *Proceedings of LREC*. Valletta, Malta: ELRA, pp. 1429–1436.

**Nairn R.**, **Condoravdi C. and Karttunen L.** (2006). Computing relative polarity for textual inference. In *Proceedings of the International Workshop on Inference in Computational Semantics*. Buxton, UK, pp. 67–76.

**Peters M.**, **Neumann M.**, **Iyyer M.**, **Gardner M.**, **Clark C.**, **Lee K. and Zettlemoyer L.** (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*. New Orleans, LA, USA: ACL, pp. 2227–2237.

**Polanyi**, **L. and Zaenen**, **A.** (2006). Contextual Valence Shifters. In Shanahan J.G., Qu Y. and Wiebe J. (eds), *Computing Attitude and Affect in Text: Theory and Applications*, vol. **20**. Dordrecht, Netherlands: Springer, pp. 1–10.

**Porter M.F.** (1980). An algorithm for suffix stripping. *Program* **14**(3), 130–137.

**Ruppenhofer J. and Brandes J.** (2016). Verifying the robustness of opinion inference. In *Proceedings of KONVENS*. Bochum, Germany, pp. 226–235.

**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 SemEval*. San Diego, CA, USA, pp. 558–571.

**Schulder M.**, **Wiegand M.**, **Ruppenhofer J. and Roth B.** (2017). Towards bootstrapping a polarity shifter lexicon using linguistic features. In *Proceedings of IJCNLP*. Taipei, Taiwan: Asian Federation of NLP, pp. 624–633.

**Schulder M.**, **Wiegand M. and Ruppenhofer J.** (2018a). Automatically creating a lexicon of verbal polarity shifters: Mono- and cross-lingual methods for German. In *Proceedings of COLING*. Santa Fe, NM, USA: ICCL, pp. 2516–2528.

**Schulder M.**, **Wiegand M.**, **Ruppenhofer J. and Köser S.** (2018b). Introducing a lexicon of verbal polarity shifters for English. In *Proceedings of LREC*. Miyazaki, Japan: ELRA, pp. 1393–1397.

**Shwartz V.**, **Goldberg Y. and Dagan I.** (2016). Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*. Berlin, Germany: ACL, pp. 2389–2398.

**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 EMNLP*. Seattle, WA, USA: ACL, pp. 1631–1642.

**Stoyanov V. and Cardie C.** (2011). Automatically creating general-purpose opinion summaries from text. In *Proceedings of RANLP*. Hissar, Bulgaria: ACL.

**Szarvas G.**, **Vincze V.**, **Farkas R. and Csirik J.** (2008). The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of BioNLP*. Columbus, OH, USA: ACL, pp. 38–45.

**Taboada M.**, **Brooke J.**, **Tofiloski M.**, **Voll K. and Stede M.** (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics* **37**(2), 267–307.

**Talukdar P.P.**, **Reisinger J.**, Paşca M., Ravichandran D., Bhagat R. and Pereira F. (2008). Weakly-supervised acquisition of labeled class instances using graph random walks. In *Proceedings of EMNLP*. Honolulu, HI, USA: ACL, pp. 582–590.

**van der Wouden T.** (1997). *Negative Contexts: Collocation, Polarity and Multiple Negation*. Routledge.

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

**Wiegand M.**, **Balahur A.**, **Roth B.**, **Klakow D. and Montoyo A.** (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in NLP*. Uppsala, Sweden, pp. 60–68.

**Wiegand M.**, **Loda S. and Ruppenhofer J.** (2018). Disambiguation of verbal shifters. In *Proceedings of LREC*. Miyazaki, Japan: ELRA, pp. 608–612.

**Wilson T.**, **Wiebe J. and Hoffmann P.** (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*. Vancouver, Canada: ACL, pp. 347–354.

**Zipf G.K.** (1935). *The Psycho-Biology of Language: An Introdution to Dynamic Philology*. Oxford, England: Houghton, Mifflin.

**Cite this article:** Schulder M, Wiegand M and Ruppenhofer J (2021). Automatic generation of lexica for sentiment polarity shifters. *Natural Language Engineering* **27**, 153–179. https://doi.org/10.1017/S135132492000039X