Building a dictionary of affixal negations

Chantal van Son **Emiel van Miltenburg** **Roser Morante**
Vrije Universiteit Amsterdam
{c.m.van.son, emiel.van.miltenburg, r.morantevallejo}@vu.nl

Abstract

This paper discusses the need for a dictionary of affixal negations and regular antonyms to facilitate their automatic detection in text. Without such a dictionary, affixal negations are very difficult to detect. In addition, we show that the set of affixal negations is not homogeneous, and that different NLP tasks may require different subsets. A dictionary can store the subtypes of affixal negations, making it possible to select a certain subset or to make inferences on the basis of these subtypes. We take a first step towards creating an affixal negation dictionary by annotating all direct antonym pairs in WordNet using an existing typology of affixal negations. By highlighting some of the issues that were encountered in this annotation experiment, we hope to provide some insights into the necessary steps of building a negation dictionary.

1 Introduction

Affixal negations can be defined as words marked with a negative affix (in English, either the prefixes *un-, in-, dis-, a-, an-, non-, im-, il-, ir-*, or the suffix *-less*). As they typically flag the absence of particular features, detecting affixal negations is very useful for natural language processing tasks such as text mining, recognizing textual entailment, paraphrasing, or question answering. Despite their simple definition, affixal negations are very difficult to detect automatically without a substantial false positive rate. Blanco and Moldovan (2011) note:

> "No simple search could unequivocally distinguish between a negated word such as *ineffective* and the words that just happen to begin with the letters of a negative prefix, such as *invite*. The problem could be partially solved by checking if, after removal of the prefix, the word is still valid. This method mismarks *inform* as negation because *form* is a valid word. To complicate matters further, some words are valid both as negated base words and as words in their own right: The adjective *invalid* means *not valid*, while the noun *invalid* describes a disabled person." (Blanco and Moldovan, 2011, p. 232)

Blanco and Moldavan conclude that the field might be best served by a dictionary-based approach; once we have a list of affixal negations (ideally along with their antonyms), it becomes trivial to detect this kind of negation through a simple string-matching algorithm. Before we can produce such a list, however, we first need to agree on a set of annotation guidelines describing what constitutes an affixal negation, and what does not. This paper aims to highlight some of the main issues to be considered when building a negation dictionary, and reports on a first attempt to build one.

This paper is structured as follows. In Section 2, we explore the full range of lexical negation, explaining how regular antonyms and affixal negations are two sides of the same coin. We show that there are different semantic categories of lexical negation and argue that their relevance is determined by the task to be solved. Section 3 reports on an annotation experiment in which all antonym pairs in WordNet (Miller, 1995; Fellbaum, 1998b) were annotated with the subtypes of affixal negations defined by Joshi (2012).

Section 4 provides a follow-up discussion on the requirements of a negation dictionary (based on what we learned from the annotation experiment) and its limits for automatic detection. Finally, we conclude our paper in Section 5.

2 Defining lexical negation

This section aims to define affixal negation from a broad natural language processing perspective. We first discuss the Conan Doyle negation corpus (Morante and Daelemans, 2012), which has a narrow definition of 'affixal negation'. We argue that this definition is the result of the task that Morante and Daelemans (2012) envisioned for their corpus. Following this observation, we explore the range of lexical negations. First, in Section 2.2, we argue that there's hardly any *semantic* reason to not to study antonyms along with affixal negation, since both are marked and express an opposition to something else. Then, in Section 2.3, we review some literature on semantic categorization of lexical negation, revealing that there is a rich landscape of affixal negations beyond the commonly studied subclass of direct negations.

2.1 Affixal negation

Affixal negation can be defined as a type of negation that is marked by the presence of a negative affix. However, not every affixal negation is relevant for each task; its relevance is determined by the semantics of the affixal negation and the goal of the task at hand. For example, Morante and Daelemans (2012) included affixal negations as part of their annotations of negation information at sentence level in two Conan Doyle stories. In the guidelines that were provided for these annotations, Morante et al. (2011) describe their main goal as follows:

> In these guidelines we aim at describing how to annotate the words that express negation and the part of a sentence that is affected by the negation words. The words that express negation are called *negation cues* and the part of the sentence that is affected by a negation cue is called the *scope*. [...] The final goal of annotating negation cues and their scope is to determine which events in the sentence are affected by the negation. (Morante et al., 2011, p. 3-4)

Morante et al. (2011) use a narrow definition of affixal negation, in which not all negative affixes are negation cues. According to the guidelines, a word with a negative affix is only considered an affixal negation if the meaning of the affixed word is a direct antonym of its non-affixed counterpart. So *unclear* is an affixal negation, because its meaning is the opposite of *clear*. This can be contrasted with examples such as *unspoken* (which does not mean 'not spoken', but 'understood without the need for words') and *disappear* (which does not mean 'not appear', but 'to pass out of sight; vanish'). Despite these words having some negative meaning component, they are not considered affixal negations.

The choice of what type of affixal negation to include in a dictionary or annotation task depends on the goal of the task to be solved. The narrow definition used by Morante et al. (2011) is a direct consequence of their main goal: to annotate information relative to the negative polarity of an event. The resulting corpus is meant to support the development of a system that can distinguish between facts and counterfacts. Therefore, they focus exclusively on negations that turn an event into a negated event, disregarding any expression that does not meet this criterion. As a consequence, affixal negations are only annotated if the affix negates the event or property expressed by its base. For other tasks, however, it may be relevant to include other kinds of affixal negations. In the context of sentiment analysis it all depends on whether or not the affixal derivative or its base is opinionated; words like *flawless* or *disqualify* should be included in a polarity lexicon (Wiegand et al., 2010), whereas words like *untie* or *backless* would be irrelevant. In the context of question answering, however, knowing what the word *backless* entails is essential to know the answer to the question *does the dress have a closed back?*

2.2 Regular antonyms

In the previous subsection we have argued that, depending on its goal, the task to be solved may require a certain subset of affixal negations. On the other hand, the full set of affixal negations may still not

be sufficient if the task requires taking all sorts of opposites into account. That is, regular antonyms might have to be considered in addition to affixal negations. After all, the difference between the two categories is only morphological. The items in (1) illustrate our point; all entail the falsehood of their positive counterpart (*tasteful, delicious, great*):

(1) a. distasteful (a 'true' affixal negation)
b. disgusting (only etymologically an affixal negation)
c. dead (a regular antonym)

Moreover, we might consider these items as points on a *continuous scale* going from explicitly (1a) to implicitly (1c) marked lexical items.[1] Joshi (2012) uses the term *lexical negation* to denote both affixal negations and antonyms, leading to the taxonomy in Figure 1 (the difference between direct and indirect negation will be discussed in the next section). This taxonomy, we argue, shows the full picture that NLP researchers interested in negation ought to consider.

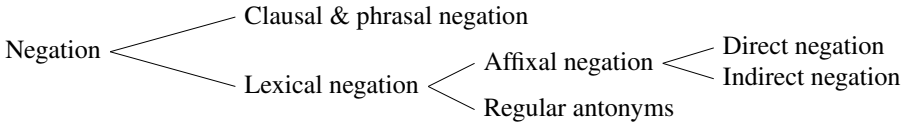

Figure 1: Taxonomy of negations, based on (Joshi, 2012).

To some extent, WordNet (Miller, 1995; Fellbaum, 1998b) and thesauri such as Roget's (Roget, 1911) already provide a collection of lexical negations. In WordNet, antonymy is defined as a lexical relation between individual lexemes that have clear opposite meanings (rather than between concepts, i.e. all the members of a synset). These 'direct antonym' pairs, such as *wet:dry* or *long:short*, are psychologically salient and have a strong associative bond between them resulting from their frequent co-occurrence (Fellbaum, 1998a). 'Indirect antonyms', then, result from similarity relations defined for the members of these direct antonym pairs. For example, *moist* and *humid* are classified as semantically similar to *wet*, and are therefore indirect antonyms of the lexeme *dry*. See Figure 2 for a schematic representation of these similarity and antonymy relations in WordNet. However, these resources do not further characterize the relations between the members of an antonymous pair. Mohammad et al. (2008) point out that WordNet does not encode the *degree* of antonymy between words; in this paper we aim to show that it is not so much the degree that should be encoded (we think that the distinction between direct and indirect antonyms already covers this for the most part), but *semantic categories* that enable distinguishing between, for example, *clear:unclear* and *appear:disappear*.

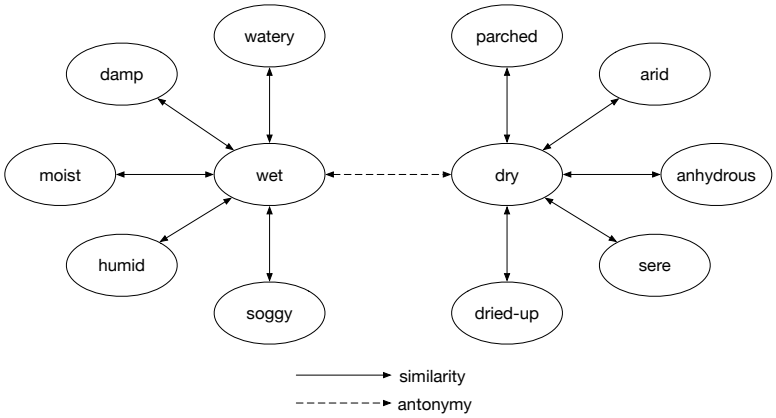

Figure 2: Similarity and antonymy relations in WordNet, from (Gross and Miller, 1990).

[1] See (Clark, 1976; Lehrer, 1985; Schriefers, 1985) and (Horn, 1989, Chapter 3) for more on the markedness of affixal negations and antonyms.

2.3 Semantic categories of lexical negation

As the examples in Section 2.1 illustrate, the set of affixal negations is not homogeneous. Joshi (2012) proposes grouping affixal negations into two main groups: direct and indirect. Direct negation expresses a direct opposition with its positive counterpart and "is characterized by the NOT-element in the derivative with respect to its base" (Joshi, 2012, p. 20). For example, *unhappy* can be paraphrased as *not happy*. Indirect negation, on the other hand, does not logically negate the existence of its base, yet still maintains a negative connotation (e.g. *dismount, debug*). Joshi (2012) further subcategorizes indirect negation into the types presented in Table 1. Knowledge of these subtypes is useful for making inferences about sentences containing indirect affixal negations. For example, the subtypes 'reversal of action' (e.g. in *she unlocked the door*) and 'removal' (e.g. in *dislodging a stone from the wall*) allow for inferences about previous states.

Category	Definition	Examples
Reversal of direction (ROD)	Indicating movement in an opposite direction (without negating the concept of movement indicated by the base).	*diverge*, *decrease*
Reversal of action (ROA)	Indicates an action performed to reverse another previous action.	*untie*, *disconnect*
Inferiority (INF)	Indicates a lower value or quality (without negating the existence of its base).	*hypoacid*, *hypotension*
Insufficiency (INS)	Gives a precision about the level, taken as negative in some contexts.	*subnormal*, *underestimate*
Badness/wrong (WRO)	Gives a precise description of someone's behaviour in a negative way.	*miscalculate*, *misjudge*
Over-abundance (OVA)	Indicates an excessive and undesired quantity of activity.	*hyperactive*, *overrate*
Pejorative (PEJ)	Pejorative indication of excessive behaviour.	*drunkard*, *braggart*
Opposition (OPP)	Indicates an opposition in notion, action, ideology, etc.	*anti-terrorist*, *antimatter*
Removal (REM)	Indicates the removal of something.	*debug*, *dislodge*

Table 1: Subtypes of indirect negation from (Joshi, 2012, p. 27). Definitions have been slightly reworded for clarity and some examples have been changed from Sanskrit or French to English for more uniformity.

Joshi's categorization system is organized in terms of the relation between the affix and the base. This can be contrasted with the taxonomy of Cruse (1986), which offers a characterization of the full domain of opposition relations between lexical items. Table 2 illustrates this, with a selection of opposition relations identified by Cruse. The overarching goal of (Cruse, 1986) is to describe the structural properties of the lexicon. Despite the differences between Joshi's and Cruse's approaches, we can also observe some similarities. For example, Cruse's category of 'reversives' strongly relates to Joshi's subtypes of 'reversal of action/direction' and 'removal'.

3 Building a negation dictionary

As noted by Blanco and Moldovan (2011), dealing with affixal negations seems to require a dictionary-based approach. We have shown that having a list of affixal negations may not be enough; we also need to specify the relation between the affix and the base in order to know what a word like *backless* or *miscalculate* entails. Furthermore, we have shown that affixal negations are part of a larger phenomenon that might either be called *lexical negation* (Joshi) or *lexical opposition* (Cruse). Ultimately, it seems to us that a dictionary-based approach should capture negation/opposition at this level, but creating such a dictionary goes beyond the scope of this paper. We will however take a step in this direction by testing the feasibility of creating a negation dictionary using Joshi's typology.

As a starting point for our negation dictionary, we have taken all pairs of direct antonyms from WordNet (Fellbaum, 1998b), which include both affixal negations and regular antonyms (WordNet does not make an explicit distinction between them). The full set comprises 3,557 antonym pairs and includes verbs, nouns, (satellite) adjectives and adverbs.[2]

[2]The dictionary is openly available at: `https://github.com/cltl/lexical-negation-dictionary`

Category	Definition	Examples
Directions	Pairs of terms which "denoting opposite directions indicate potential paths, which, if followed by two moving lines, would result in their moving in opposite directions."	*south:north, up:down*
Antipodal opposites	Pairs of terms for which "one term represents an extreme in one direction along some salient axis, while the other term denotes the corresponding extreme in the other direction."	*cellar:attic, head:toe, top:bottom, source:mouth, always:never, all:none*
Counterparts	Pairs of terms for which one term is the counterpart of the other, "in which essential defining directions are reversed."	*ridge:groove, hill:valley*
Reversives	"Pairs of verbs which denote motion or change in opposite directions."	*rise:fall, ascend:descend*
Sub: restitutives	"Pairs one of whose members necessarily denotes the restitution of a former state."	*damage:repair, kill:resurrect*
Sub: independent reversives	Pairs for which "there is no necessity for the final state of either verb to be a recurrence of a former state."	*narrow:widen, fill:empty*
Relational opposites: converses	"Those pairs which express a relationship between two entities by specifying the direction of one relative to the other along some axis."	*above:below, before:after, teacher:pupil*
Sub: direct converses	"Converse pairs in which the interchangeable noun phrases occupy central valency slots."	*follow:precede*
Sub: indirect converses	Converse pairs "where a central and peripheral noun phrase are interchanged."	*give:receive*

Table 2: Categories of directional oppositions from (Cruse, 1986).

3.1 Annotation tasks

We included the following information from WordNet about the antonym pairs in our dictionary: (1) the lemmas of both antonyms, (2) the lemma identifiers of both antonyms, (3) the definitions of both antonyms, and (4) the part of speech. Then, we performed the following three annotation steps to enrich the entries:

1. **Affixal or non-affixal:** For each antonym pair, we annotated whether the antonym pair contained an affixal negation or not. If applicable, the negative and the positive affixes were annotated as well.

2. **Direct or indirect:** For each affixal negation, we indicated whether it was a direct or an indirect negation according to the definitions provided by Joshi (2012).

3. **Subtype:** Each indirect affixal negation was classified into one of the nine subtypes defined by Joshi (2012): ROD, ROA, INF, INS, WRO, OVA, PEJ, OPP, or REM (see Table 1). In addition, we introduced a label LAC for affixal negations that indicate that some characteristic is lacking.

Table 3 shows a few simplified examples of the resulting entries in the dictionary. The tasks were performed by two annotators. A set of 500 randomly selected antonym pairs was annotated by both annotators in order to measure inter-annotator agreement.

Positive element	Negative element	POS	Positive affix	Negative affix	Direct/indirect	Subtype
structured	unstructured	a	NA	un-	direct	NA
inshore	offshore	a	in-	off-	indirect	ROD
colonize	decolonize	v	NA	de-	indirect	ROA
revolutionary	counter-revolutionary	a	NA	counter-	indirect	OPP
used	misused	a	NA	mis-	indirect	WRO
humerously	humerlessly	r	-ous	-less	indirect	LAC

Table 3: Simplified examples of entries of affixal negations in the dictionary (lemma identifiers and definitions are excluded for reasons of space).

3.2 Evaluation

Inter-annotator agreement was measured using Cohen's kappa for each of the three annotation tasks. For subtask (1), determining whether the antonym pair contained an affixal negation or not, we measured an IAA score of 0.80 (n=500). Most of the disagreements (58%) on this task were caused by mistakes of the annotators. The remaining 42% consisted of pairs where it was a bit more difficult to determine whether it should be considered an affixal negation or not. Examples are *onstage:offstage*, *intrusive:extrusive*, *concealing:revealing*. For subtask (2), indicating whether the affixal negation was direct or indirect, a rather low IAA score of 0.55 was obtained (n=268). Finally, we achieved an IAA of 0.76 (n=43) for subtask (3), the classification of indirect negations into their subtypes.

Table 4 represents the confusion matrix for the annotation of the subtypes; the 'direct' label is also included to show the disagreements between this label and each of the subtypes of indirect negation as well. What we can see from this confusion matrix is that one annotator annotated 35 antonym pairs as 'direct negation', whereas the other annotated these pairs as an indirect negation of the subtype 'opposition'. It appeared that it was not exactly clear what types of negation are covered by the 'opposition' type; although the definition provided by Joshi (2012) ("opposition in notion, action, ideology, etc.") can be understood in a very broad sense and seems similar to direct negation, the examples illustrating this subtype in (Joshi, 2012) are more specific (*anti-terrorist*, *antimatter*). Most of the disagreements (29/35) caused by this uncertainty regarding the definition of 'opposition' were on antonym pairs with an affixal negation starting with the prefix *non-*, such as *modern:non-modern*, *fictional:non-fictional*, *competitive:non-competitive*.

	LAC	**direct**	**OPP**	**OVA/INS**	**ROA**	**ROD**	**WRO**
INS	1	1	0	0	0	0	0
LAC	**18**	0	0	0	0	1	0
direct	0	**179**	35	0	3	1	0
OPP	0	0	**1**	0	0	0	0
OVA/INS	0	0	0	**1**	0	0	0
REM	0	0	0	0	1	0	0
ROA	0	6	0	0	**12**	1	0
ROD	0	0	1	0	2	**2**	0
WRO	0	0	0	0	0	0	**2**

Table 4: Confusion matrix for the annotation of subtypes

There was also some confusion between 'direct negation' and the subtype 'reversal of action', but most of them appeared to be mistakes (incorrectly annotated as 'direct'). Finally, the antonym pairs where both annotators recognized an indirect affixal negation but disagreed on the subtype were:

Antonym pair	**Annotator 1**	**Annotator 2**
arming:disarming	removal	reversal of action
content:discontent	reversal of direction	reversal of action
pressurise:depressurise	reversal of direction	reversal of action
conjunctive:disjunctive	reversal of direction	opposition
attachable:detachable	reversal of action	reversal of direction
merit:demerit	lack	reversal of direction
fluency:disfluency	insufficiency	lack

Table 5: Antonym pairs where both annotators recognized an indirect affixal negation but disagreed on the subtype.

4 Discussion

4.1 Annotating the relation between lexical items, or between affix and the base

Some words raised doubts for both annotators during the annotation process. One of these cases was the difference between the characterizations of verbal affixal negations and their inflected forms. For example, the antonym pair *fasten* ("become fixed or fastened") and *unfasten* ("become undone or untied")

is a clear example of reversal of action. However, *unfastened* ("not closed or secured") seems more of a direct negation with respect to its base *fastened* ("firmly closed or secured"). The difficulty with participles like this one, which are stored as adjectives in WordNet, is that they indicate a state that can be interpreted as a result of the action expressed by its verbal base (e.g. *unfasten*) - but not necessarily (it might never have been fastened at all). Similar doubts were raised regarding antonym pairs such as *spinous* ("having spines") and *spineless* ("lacking spiny processes"). Even though the affix *-less* clearly expresses the lack of something and both annotators annotated these cases as LAC, *spineless* is just a direct negation ("not having spines") in relation to its antonym *spinous*.

Both examples are related to the question: are we annotating the relation between the affix and its base (*spine:spineless*), or the oppositional relation between the two members of an antonym pair (*spinous:spineless*)? And if we are annotating the relation between the affix and its base, what exactly should be considered the base? The simple, uninflected form (*fasten*) or the lexeme with just the negative affix stripped off (*fastened*)? These are questions that were not explicitly answered for the annotation reported in this paper, but should in fact play a central role in any future effort to build a negation dictionary.

4.2 Coverage

As with any lexical resource, a negation dictionary is only as good as its coverage. And since affixal negation is a productive phenomenon, we can ask ourselves: what would be a good fallback strategy to detect and reason about affixal negations? As noted by Blanco and Moldovan (2011), cited in the introduction of this paper, simple string matching algorithms will produce many false positives. One way to reduce those false positives and increase coverage might be to train a classifier (using either word-level (Mikolov et al., 2013) or character-level (Kim et al., 2016) representations) to recognize (1) whether a word has a negative component, and (2) what kind of relation exists between the affix and the base. Training such a classifier still requires us to annotate negations, however, and to think about the relations that the classifier should learn.

5 Conclusion

We have argued that many NLP tasks could benefit from a negation dictionary, since this would solve some of the problems that are currently encountered when detecting negations in text. One of these problems is that it is difficult to distinguish between affixal negations and words that just happen to begin with the letters of a negative prefix. However, we have shown that a simple list of affixal negations would not suffice; there is a range of different kinds of affixal negations, and which of these are relevant to include depends on the NLP task that is to be supported by the list. In addition, we have noted that, from a semantic point of view, affixal negations are not that different from negative adjectives. A dictionary that is supposed to cover the complete spectrum of lexical negation should therefore include both affixal negations and antonyms. This paper does not offer the final solution to building the perfect negation dictionary. Nevertheless, we hope that it contributes its share to the discussion by highlighting some of the main issues to be considered when building one and by proposing some elements that we think such a dictionary should minimally include: the opposing pair of lexical items with their definitions, the type of relation between them, and what affix is used (if applicable).

6 Acknowledgements

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities, and by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019). We also thank four anonymous reviewers for their useful feedback.

References

Eduardo Blanco and Dan Moldovan. 2011. Some issues on detecting negation from text. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pages 228–233. AAAI.

Herbert H. Clark. 1976. *Semantics and Comprehension*. Mouton, The Hague.

D Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.

Christiane Fellbaum. 1998a. A semantic network of English: the mother of all WordNets. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 137–148. Springer.

Christiane Fellbaum. 1998b. *WordNet*. Wiley Online Library.

Derek Gross and Katherine J Miller. 1990. Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265–277.

Laurence R. Horn. 1989. *A natural history of negation*. CSLI Publications.

Shrikant Joshi. 2012. Affixal negation – direct, indirect and their subtypes. *Syntaxe et sémantique*, (1):49–63.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-aware neural language models.

Adrienne Lehrer. 1985. Markedness and antonymy. *Journal of linguistics*, 21(02):397–429.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 982–991. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. Conan Doyleneg: Annotation of negation in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1.0. Technical Report Series CTR-003, CLiPS, University of Antwerp, Antwerp.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company.

Heribert Johannes Schriefers. 1985. On semantic markedness in language production and verification.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68. Association for Computational Linguistics.