Learning Interpretable Negation Rules via Weak Supervision at Document Level: A Reinforcement Learning Approach

Nicolas Pröllochs
Oxford-Man Institute
University of Oxford
nicolas.prollochs@eng.ox.ac.uk

Stefan Feuerriegel
ETH Zurich
sfeuerriegel@ethz.ch

Dirk Neumann
University of Freiburg
dirk.neumann@is.uni-freiburg.de

Abstract

Negation scope detection is widely performed as a supervised learning task which relies upon negation labels at word level. This suffers from two key drawbacks: (1) such granular annotations are costly and (2) highly subjective, since, due to the absence of explicit linguistic resolution rules, human annotators often disagree in the perceived negation scopes. To the best of our knowledge, our work presents the first approach that eliminates the need for word-level *negation* labels, replacing it instead with document-level *sentiment* annotations. For this, we present a novel strategy for learning fully interpretable negation rules via weak supervision: we apply reinforcement learning to find a policy that reconstructs negation rules from sentiment predictions at document level. Our experiments demonstrate that our approach for weak supervision can effectively learn negation rules. Furthermore, an out-of-sample evaluation via sentiment analysis reveals consistent improvements (of up to 4.66 %) over both a sentiment analysis with (i) no negation handling and (ii) the use of word-level annotations from humans. Moreover, the inferred negation rules are fully interpretable.

1 Introduction

Negations are a frequently utilized linguistic tool for expressing disapproval or framing negative content with positive words. Neglecting negations can lead to false attributions (Morante et al., 2008) and, moreover, impair accuracy when analyzing natural language; e. g., in information retrieval (Cruz Díaz et al., 2012; Rokach et al., 2008) and especially in sentiment analysis (Cruz et al., 2015; Wiegand et al., 2010). Hence, even simple heuristics for identifying negation scopes can yield substantial improvements in such cases (Jia et al., 2009).

Negation scope detection is sometimes implemented as unsupervised learning (e. g., Pröllochs et al., 2016), while a better performance is commonly achieved via supervised learning (see our supplements for a detailed overview): the resulting models thus learn to identify negation scopes from word-level annotations (e. g., Li and Lu, 2018; Reitan et al., 2015). We argue that this approach suffers from inherent drawbacks. (1) Such granular annotations are costly and, especially at word level, a considerable number of them is needed. (2) Negation scope detection is highly subjective (Councill et al., 2010). Due to the absence of explicit linguistic rules for resolutions, existing corpora often come with annotation guidelines (Morante and Blanco, 2012; Morante and Daelemans, 2012). Yet there are considerable differences: some corpora were labeled in a way that negation scopes consist of single text spans, while others allowed disjoint spans (Fancellu et al., 2017). More importantly, given the absence of universal rules, human annotators largely disagree in their perception of what words should be labeled as negated.

Motivational experiment. Since prevalent corpora were labeled only by a single rater, we now establish the severity of between-rater discrepancies. For this, we carried out an initial analysis of 500 sentences from movie reviews.[1] Each sentence contained at least one explicit negation phrase from the list of Jia et al. (2009), such as *"not"* or *"no."* Two human raters were then asked

[1] Details are reported in our supplementary materials.

Proceedings of NAACL-HLT 2019, pages 407–413
Minneapolis, Minnesota, June 2 - June 7, 2019.

to annotate negation scopes. They could choose an arbitrary selection of words and were not restricted to a single subspan, as recommended by Fancellu et al. (2017). The annotations resulted in large differences: only 50.20 % of the words were simultaneously labeled as "negated" by both raters. Based on this experimental evidence, we showcase there is no universal definition of negation scopes (rather, human annotations are likely to be noisy or even error-prone) and thus highlight the need for further research.

Contributions. To the best of our knowledge, our work presents the first approach that eliminates the need for *word-level* annotations of *negation labels*. Instead, we perform negation scope detection merely by utilizing shallow annotations at *document level* in the form of *sentiment labels* (e. g., from user reviews). Our novel strategy learns interpretable negation rules via weak supervision: we apply reinforcement learning to find a policy that reconstructs negation rules based on sentiment prediction at document level (as opposed to conventional word-level annotations).

In our approach, a single document d comes with a sentiment label y_d. The document consists of N_d words, $w_{d,1}, \ldots, w_{d,N_d}$, where the number of words can easily surpass several hundreds. Based on the sentiment value, we then need to make a decision (especially out-of-sample) for each of the N_d words, whether or not it should be negated. In this case, a *single* sentiment value is outnumbered by potentially *hundreds* of negation decisions, thus pinpointing to the difficulty of this task. Formally, the goal is to learn individual labels $a_{d,i} \in \{\text{Negated}, \neg\text{Negated}\}$ for each word $w_{d,i}$. Rewards are the errors in sentiment prediction at document level.

Strengths. Our approach exhibits several favorable features that overcome shortcomings found in prior works. Among them, it eliminates the need for manual word-level labels. It thus avoids the detrimental influence of subjectivity and misinterpretation. Instead, our model is solely trained on a document-level variable and can thus learn domain-specific particularities of the given prose. The inferred negation rules are fully interpretable while documents can contain multiple instances of negations with arbitrary complexity, sometimes nested or consisting out of disjoint text spans. Despite facing several times more negation decisions than sentiment labels, our experiments demonstrate that this problem can be effectively learned through reinforcement learning.

Evaluation. Given the considerable inconsistencies in human annotations of negation scopes and the lack of universal rules, we regard the "true" negation scopes as unobservable. Hence, we later compare the identified negation scopes with those from rater 1 and 2 only as a sensitivity check because of the fact that both raters have only 50.2 % overlap. Instead, we choose the following evaluation strategy. We concentrate on the performance of negation scope detection as a supporting tool in natural language processing where its primary role is to facilitate more complex learning tasks such as sentiment analysis. Therefore, we report the performance improvements in sentiment analysis resulting from our approach. For a fair comparison, we use baselines that only rely upon the same information as our weak supervision (and thus have no access to word-level negation labels). Our performance is even on par with a supervised classifier that can exploit richer labels during training.

2 Learning Negation Scope Detection via Weak Supervision

Intuition. The choice of reinforcement learning for weak supervision might not be obvious at first, but, in fact, it is informed by theory: it imitates the human reading process as stipulated by cognitive reading theory (Just and Carpenter, 1980), where readers iteratively process information word-by-word.

States and actions. In each learning iteration, the reinforcement learning agent observes the current state $s_i = (w_i, a_{i-1})$ that we engineer as the combination of the i-th word w_i in a document and the previous action a_{i-1}. This specification establishes a *recurrent* architecture whereby the previous negation can pass on to the next word. At the same time, this allows for nested negations, as a word can first introduce a negation scope and another subsequent negation can potentially revert it.

After observing the current state, the agent chooses an action a_t from of two possibilities: (1) it can set the current word to *negated* or (2) it can mark it as *not negated*. Hence, we obtain the following set of possible actions $A = \{\text{Negated}, \neg\text{Negated}\}$. Based on the selected action, the agent receives a reward, r_i which updates the knowledge in the state-action function

$Q(s_i, a_i)$. This state-action function is then used to infer the best possible action a_i in each state s_i, i. e., the optimal policy $\pi^*(s_i, a_i)$.

Reward function. The reward r_i depends upon the correlation between a given a document-level label (e. g., a rating in movie reviews) and the sentiment of a document. We predict the sentiment S_d of document d using a widely-used sentiment routine based on the occurrences of positively- and negatively-opinionated terms (see Taboada et al., 2011). If a term is negated by the policy, the polarity of the corresponding term is inverted, i. e., positively opinionated terms are counted as negative and vice versa. In the following, S_d^0 denotes the document sentiment without considering negations; S_d^π the sentiment when incorporating negations based on policy π.

When processing a document, we cannot actually compute the reward until we have processed all words. Therefore, we set the reward before the last word to $c \approx 0$, i. e., $r_i = c$ for $i = 1, \ldots, N_d - 1$. For the final word, the agent compares its performance in predicting the document label based on sentiment without considering negations S_d^0 to the sentiment when incorporating negations based on the current policy π^*. The former is defined by the absolute difference between the document label y_d and the predicted sentiment without negations S_d^0, whereas the latter is defined by the absolute difference between y_d and the adjusted sentiment using the current policy S_d^π. Then the difference between these values returns the terminal reward r_{N_d}. Thus the reward is

$$r_i = \begin{cases} 0, & \text{if } a_i = \text{Neg and } i < N_d, \\ c, & \text{if } a_i = \neg\text{Neg and } i < N_d, \\ \left|y_d - S_d^0\right| - \left|y_d - S_d^\pi\right|, & \text{if } i = N_d, \end{cases}$$

with a constant c (we use $c = 0.005$) that adds a small reward for default (i. e., non-negating) actions to avoid overfitting.

***Q*-learning.** During the learning process[2], the agent then successively observes a sequence of words in which it can select between exploring new actions or taking the current optimal one. This choice is made by ε-greedy selection according to which the agent explores the environment by selecting a random action with probability ε or, alternatively, exploits the current knowledge with probability $1 - \varepsilon$.

3 Experiments

Datasets. We use the following benchmark datasets with document-level annotations from the literature (cf. Hogenboom et al., 2011; Pröllochs et al., 2016; Wiegand et al., 2010):

IMDb: movie reviews from the Internet Movie Database archive, each annotated with an overall rating at document level (Pang and Lee, 2005).

Airport: user reviews of airports from Skytrax, each annotated with an overall rating at document level (Pérezgonzález and Gilbey, 2011).

Ad hoc: financial announcements with complex, domain-specific language (Pröllochs et al., 2016), labeled with the daily abnormal return of the corresponding stock.

Learning parameters. We perform 4000 learning iterations with a higher exploration rate as given by the following parameters[3]: exploration ε = 0.001, discount factor γ = 0 and learning rate α = 0.005. In a second phase, we run 1000 iterations for fine-tuning with exploration ε = 0.0001, discount factor γ = 0 and learning rate α = 0.001.

Policy learning. For each dataset, the reinforcement learning process converges to a stationary policy that shows reward fluctuations below 0.05 %. As part of a benchmark, we study the mean squared error (MSE) between y_d and the predicted sentiment S_d^0 when leaving negations untreated as our benchmark. For all datasets, the in-sample MSE decreases substantially (see Figure 1), demonstrating the effectiveness of our learning approach. The reductions number to 14.93 % (IMDb), 16.77 % (airport), and 0.91 % (ad hoc). The latter is a result of the considerably more complex language in financial statements.

Performance in Sentiment Analysis. We use 10-fold cross validation to compare the out-of-sample performance in sentiment analysis of reinforcement learning to benchmarks without word-level labels from previous works. The benchmarks consists of rules (Hogenboom et al., 2011; Taboada et al., 2011), which search for the occurrence of specific cues based on pre-defined lists and then invert the meaning of a fixed number of surrounding words. Table 1 compares the out-of-sample MSE between predicted sentiment and the

[2] We use Watkin's $Q(\lambda)$ with eligibility traces; see Sutton and Barto (1998) for details. At the beginning, we initialize the action-value function $Q(s, a)$ to zero for all states and actions. This also controls our default action when encountering unknown states or out-of-vocabulary (OOV) words. In such cases, the non-negated action is preferred.

[3] Further details regarding the learning parameters are provided in the supplementary materials.

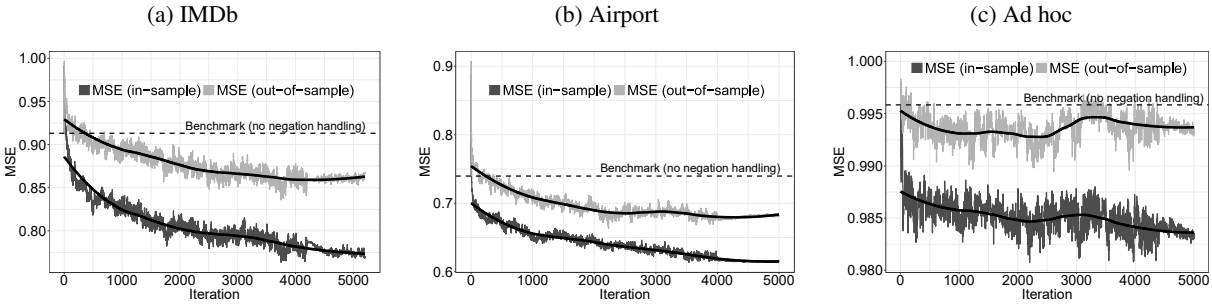

Figure 1: MSE between the document label and predicted sentiment across different learning iterations using 10-fold cross validation. Additional lines in black from smoothing.

document-level label:[4]

IMDb: Negating a fixed window of the next 4 words achieves the lowest error among all rules, similar to Dadvar et al. (2011). This rule reduces the MSE of the benchmark with no negation handling by 1.05 %. Our approach works even more accurately, and dominates all of the rules, reducing the out-of-sample MSE by at least 4.60 %.

Airport: Our method decreases the MSE by 4.66 % compared to the best-performing rule (negating a fixed window of the next 4 words).

Ad hoc: Even for complex financial language, reinforcement learning exceeds this benchmark method by 0.19 % in terms of out-of-sample MSE.

Altogether, our weak supervision improves sentiment analysis consistently across all datasets.[5]

Approach	IMDb	Airport	Ad hoc
Benchmark: no negation handling	0.9133	0.7415	0.9958
Negating all subsequent words	0.9160	0.7312	0.9949
Negating the whole sentence	0.9339	0.7811	0.9961
Fixed window of 1 word	0.9082	0.7212	0.9950
Fixed window of 2 words	0.9052	0.7130	0.9943
Fixed window of 3 words	0.9047	0.7146	0.9943
Fixed window of 4 words	0.9038	0.7134	0.9942
Fixed window of 5 words	0.9039	0.7136	0.9942
Proposed reinforcement learning for weak supervision	**0.8622**	**0.6798**	**0.9923**

Table 1: Out-of-sample MSE between sentiment S_d^π and the document label y_d. Lowest error in bold.

Comparison to human raters. For reasons of completeness, our supplements report the overlap with both human raters from our motivational experiment, which is in the range of 18.8 % to 25.2 %. However, these numbers should be treated with caution, as we remind that there is no universal definition of negation scopes and even the two human annotations reveal on 50.2 %. Moreover, our approach was not learned towards reconstructing these human annotations, since we focused on rules that achieve the greatest benefit in sentiment analysis.

Comparison to word-level classifiers. We also compared weak supervision against a supervised HMM classifier from Pröllochs et al. (2016) that draws upon granular word-level negation labels. Here we report the sentiment analysis on IMDb in order to be able to use the domain-specific negation labels from IMDb text snippets of our initial experiment. In comparison to our reinforcement learning, the supervised classifier results in a 5.79 % higher (and thus inferior) MSE. Yet our weak supervision circumvents costly word-level annotations.

Interpretability. Our method yields negation rules that are fully interpretable: one simply has to assess the state-action function $Q(s_i, a_i)$. Table 2 provides an example excerpt for the document *"this beautiful movie isn't good but fantastic."* The agent the starts by observing the first state given by the combination of the first word w_1 and the previous action a_0, i. e. $s_1 = (\textit{this}, \neg\text{Negated})$. According to the state-action table, the best action for the agent is to set this state to not negated ($a_1 = \neg$Negated). This pattern continues until it observes state $s_4 = (\textit{isn't}, \neg\text{Negated})$ in which the policy implies to initiate a negation scope ($a_4 =$ Negated). Subsequently, the negation scope is forwarded until the agent observes $s_6 = (\textit{but}, \text{Negated})$ in which it terminates the negation scope ($a_6 = \neg$Negated). Finally, the agent observes $s_7 = (\textit{fantastic}, \neg\text{Negated})$ in which it takes action $a_7 = \neg$Negated.

[4] We also experimented with performance comparisons in a classification task, yet our approach also yields consistent improvements in this evaluation.

[5] We also investigated the relationship between prediction performance and text length, finding only minor effects.

State=(w_i, a_{i-1})	Negated	¬Negated	$\pi^*(s_i, a_i)$
(*this*, ¬Negated)	0.0114	0.0502	¬Negated
(*beautiful*, ¬Negated)	0.0081	0.0779	¬Negated
(*movie*, ¬Negated)	0.0039	0.0506	¬Negated
(*isn't*, ¬Negated)	0.0700	0.0456	Negated
(*good*, Negated)	0.0578	0.0322	Negated
(*but*, Negated)	0.0120	0.0365	¬Negated
(*fantastic*, ¬Negated)	−0.0181	0.1708	¬Negated

Table 2: Excerpt of state-action function $Q(s_i, a_i)$ actions $A = \{\text{Negated}, \neg\text{Negated}\}$ and the learned policy π^* for IMDb reviews.

4 Related Work

State-of-the-art methods for detecting, handling and interpreting negations can be grouped into different categories (cf. Pröllochs et al., 2015, 2016; Rokach et al., 2008).

Rule-based approaches are among the most common due to their ease of implementation and solid out-of-the-box performance. These usually suppose a forward influence of negation cues based on which they invert the meaning of the whole sentence or a fixed number of subsequent words (Hogenboom et al., 2011). Furthermore, they can also incorporate syntactic information in order to imitate subject and object (Padmaja et al., 2014; Chowdhury and Lavelli, 2013). Negation rules have been found to work effectively across different domains and rarely need fine-tuning (Taboada et al., 2011). However, rule-based approaches entail several drawbacks, as the list of negations must be pre-defined and the selection criterion according to which rule a rule is chosen is usually random or determined via cross validation. In addition, rules cannot effectively cope with implicit expressions or particular, domain-specific characteristics.

Generative probabilistic models (e. g., hidden Markov models or conditional random fields) can partially overcome these shortcomings (Li and Lu, 2018; Reitan et al., 2015; Rokach et al., 2008), such as the difficulty of recognizing implicit negations. These process narrative language word-by-word and move between hidden states representing negated and non-negated parts. Such models can adapt to domain-specific language, but require more computational resources and rely upon ex ante transition probabilities. Although variants based on unsupervised learning avoid the need for any labels, practical applications reveal inferior performance compared to supervised approaches (Pröllochs et al., 2015). The latter usually depend on manual labels at a granular level, which are not only costly but suffer from subjective interpretations (Fancellu et al., 2017).

A third category of methods links the polarity shift effect of negations more closely to sentiment analysis tasks at sentence or document level. For example, text parts can be classified into a polarity-unshifted part and a polarity-shifted part according to certain rules (Li and Huang, 2009). Sentiment classification models are then trained using both parts (Li et al., 2010). Alternatively, rule-based algorithms can extract sentences with inconsistent sentiment and omit them from standard sentiment analysis procedures (Orimaye et al., 2012). Reversely, antonym dictionaries have been used to generate sentiment-inverted texts to classify polarity in pairs (Xia et al., 2016). Although such data expansion techniques usually enhance the performance of sentiment analysis, they require either complex linguistic knowledge or extra human annotations (Xia et al., 2015).

Research gap. In contrast to these methods, we propose a novel strategy for learning negation rules via weak supervision. Our model uses reinforcement learning to reconstruct negation rules based on an document-level variable and does not require any kind of manual word-level labeling or precoded linguistic patterns. It is able to recognize explicit as well as implicit negations, while avoiding the influence of subjective interpretations.

5 Conclusion

This paper proposes the first approach for negation scope detection based on weak supervision. Our proposed reinforcement learning strategy circumvents the need for word-level annotations with negation scopes, as it reconstructs negation rules based on a document-level sentiment labels. Our experiments show that our weak supervision is effective in negation scope detection; it yields consistent improvements (of up to 4.66 %) over a sentiment analysis without negation handling.

Our works suggests important implications. We are in line with growing literature (e. g., Fancellu et al., 2017) that reports challenges in resolving negation scopes through humans. Beyond prior works, our experiment reveals between-rater inconsistencies. While negation scope detection is widely studied as an isolated task, it could be beneficial when linking its evaluation more closely to context-specific uses such as sentiment analysis.

References

Md Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In *NAACL-HLT*, pages 765–771.

Isaac G. Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59. ACL.

Noa P. Cruz, Maite Taboada, and Ruslan Mitkov. 2015. A machine–learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

Noa P. Cruz Díaz, Maña López, Manuel J., Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology*, 63(7):1398–1410.

Maral Dadvar, Claudia Hauff, and Franciska de Jong. 2011. Scope of negation detection in sentiment analysis. In *Dutch-Belgian Information Retrieval Workshop*, pages 16–20.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Conference of the European Chapter of the ACL*, pages 58–63. ACL.

Alexander Hogenboom, Paul van Iterson, Bas Heerschop, Flavius Frasincar, and Uzay Kaymak. 2011. Determining negation scope and strength in sentiment analysis. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2589–2594.

Lifeng Jia, Clement Yu, and Weiyi Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. In *CIKM*, pages 1827–1830.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.

Hao Li and Wei Lu. 2018. Learning with structured representations for negation scope extraction. In *Proceedings of the ACL*, pages 533–539.

Shoushan Li and Chu-Ren Huang. 2009. Sentiment classification considering negation and contrast transition. In *Pacific Asia Conference on Language, Information and Computation*, pages 297–306. ACL.

Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. 2010. Sentiment classification and polarity shifting. In *International Conference on Computational Linguistics*, pages 635–643. ACL.

Roser Morante and Eduardo Blanco. 2012. Sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SemEval '12)*, pages 265–274.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul*.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *EMNLP*, pages 715–724.

Sylvester Olubolu Orimaye, Saadat M Alhashmi, and Eu-Gene Siew. 2012. Buy it-don't buy it: sentiment classification on amazon reviews using sentence polarity shift. In *Pacific Rim International Conference on Artificial Intelligence*, pages 386–399. Springer.

S. Padmaja, Sameen Fatima, and Sasidhar Bandu. 2014. Evaluating sentiment analysis methods and identifying scope of negation in newspaper articles. *International Journal of Advanced Research in Artificial Intelligence*, 3(11):1–6.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124.

Jose D. Pérezgonzález and Andrew Gilbey. 2011. Predicting skytrax airport rankings from customer reviews. *Journal of Airport Management*, 5(4):335–339.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. 2015. Enhancing sentiment analysis of financial news by detecting negation scopes. In *HICSS*, pages 959–968.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. 2016. Negation scope detection in sentiment analysis: Decision support for news-driven trading. *Decision Support Systems*, 88:67–75.

Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108.

Lior Rokach, Roni Romano, and Oded Maimon. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

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

Rui Xia, Feng Xu, Jianfei Yu, Yong Qi, and Erik Cambria. 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1):36–45.

Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li. 2015. Dual sentiment analysis: Considering two sides of one review. *Transactions on Knowledge and Data Engineering*, 27(8):2120–2133.