Detecting speculations, contrasts and conditionals in consumer reviews

Maria Skeppstedt[1,2] **Teri Schamp-Bjerede**[3] **Magnus Sahlgren**[1] **Carita Paradis**[3] **Andreas Kerren**[2]
[1]Gavagai AB, Stockholm, Sweden
{maria, mange}@gavagai.se
[2]Computer Science Department, Linnaeus University, Växjö, Sweden
andreas.kerren@lnu.se
[3]Centre for Languages and Literature, Lund University, Lund, Sweden
{teri.schamp-bjerede, carita.paradis}@englund.lu.se

Abstract

A support vector classifier was compared to a lexicon-based approach for the task of detecting the stance categories *speculation*, *contrast* and *conditional* in English consumer reviews. Around 3,000 training instances were required to achieve a stable performance of an F-score of 90 for *speculation*. This outperformed the lexicon-based approach, for which an F-score of just above 80 was achieved. The machine learning results for the other two categories showed a lower average (an approximate F-score of 60 for *contrast* and 70 for *conditional*), as well as a larger variance, and were only slightly better than lexicon matching. Therefore, while machine learning was successful for detecting *speculation*, a well-curated lexicon might be a more suitable approach for detecting *contrast* and *conditional*.

1 Introduction

Stance taking – including attitudes, evaluations and opinions – has received a great deal of attention in the literature (Hunston and Thompson, 2000; Biber, 2006; Hunston, 2011; Fuoli, 2015), and many studies of speakers' expression of feelings have been carried out in the fields of sentiment analysis and opinion mining with pre-defined or automatically detected categories related to sentiments and opinions. At its most basic level, such analyses use categories of positive, negative or (sometimes) neutral sentiment (Täckström and McDonald, 2011; Feldman, 2013), while other types of analyses use more fine-grained categories of sentiments or attitudes, such as happiness, anger and surprise (Schulz et al., 2013). There are, however, additional aspects or types of stance taking, e.g., contrasting of different opinions (Socher et al., 2013), indications of the degree of likelihood of a conveyed message (Biber, 2006) or expression of conditional statements (Narayanan et al., 2009). Detecting such aspects is an integral part of a high quality sentiment analysis system, as they modify the opinions expressed. In this study, the automatic detection of three such stance categories is investigated:

(1) *Speculation*: "the possible existence of a thing [that] is claimed – neither its existence nor its non-existence is known for sure" (Vincze, 2010, p. 28).

(2) *Contrast*: "Contrast(α,β) holds when α and β have similar semantic structures, but contrasting themes, i.e. sentence topics, or when one constituent negates a default consequence of the other" (Reese et al., 2007, p. 17).

(3) *Conditional*: "describe[s] implications or hypothetical situations and their consequences" (Narayanan et al., 2009, p. 1).

There are previous studies on automatic detection of *speculation* and related stance categories. Results are, however, reported for models trained on large annotated corpora, which are expensive to obtain (Uzuner et al., 2011; Cruz et al., 2015). Here, lexicon-based methods – as well as machine learning models trained on a smaller amount of training data – are instead evaluated for the task of detecting *speculation*, *contrast* and *conditional*. The categories are specifically compared with regards to the following research questions: (a) Are machine learning or lexicon-matching the more suitable method for detecting these three stance categories? (b) How does the amount of used training samples affect the performance of trained machine learning models?

2 Previous research

Speculation has been explored in, e.g., biomedical texts (Vincze et al., 2008; Velupillai, 2012; Aramaki et al., 2014), consumer reviews (Konstantinova et al., 2012), tweets (Wei et al., 2013) and

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 162–168,
Lisboa, Portugal, 17 September, 2015.

Wikipedia texts (Farkas et al., 2010). Biomedical text annotation has also included classification into different levels of uncertainty (Velupillai et al., 2011), as well as into the categories *present*, *absent*, *possible*, *conditional* and *hypothetical* (Uzuner et al., 2011). Some schemes annotate uncertainty markers/cues and their scope (Vincze et al., 2008), while others annotate speculation towards certain types of entities (Velupillai et al., 2011; Aramaki et al., 2014), or categorise text chunks, e.g., sentences or tweets, according to whether they contain speculation or not (Farkas et al., 2010; Wei et al., 2013).

Some systems for automatic detection of *speculation* are modelled as text classification problems, often using support vector classifiers (SVCs) trained on word n-grams (Uzuner et al., 2011; Wei et al., 2013). Others are modelled as named entity recognition systems and use structured prediction for detecting text chunks that function as cues for speculation (Tang et al., 2010; Clark et al., 2011).

The SFU Review corpus, which consists of English consumer generated reviews of books, movies, music, cars, computers, cookware and hotels (Taboada and Grieve, 2004; Taboada et al., 2006), is often used for sentiment analysis. This corpus has been annotated for speculation by Konstantinova et al. (2012), according to a modification of guidelines created by Vincze et al. (2008), in which cues for *speculation* and *negation*, and their scope, were annotated. Inter-annotator agreement was measured on 10% of the corpus, resulting in an F-score and a Kappa score of 89 for the agreement on speculation cues. The same corpus has also been annotated by Taboada and Hay (2008) for Rhetorical Structure Theory categories (Taboada and Mann, 2006, pp. 426–427). A total of 36 different categories were annotated, including *condition*, *contrast* and *concession*[1]. In contrast to the annotations by Konstantinova et al., these annotations were not checked for reliability.

Cruz et al. (2015) trained an SVC to detect the speculation cues annotated by Konstantinova et al., and achieved an F-score of 92. Their lexicon matching approach, which was built on a list of the four most frequent speculation cues, achieved a lower F-score of 70. The SVC was clearly successful, as results slightly better than the inter-annotator agreement were achieved. Since the results were achieved by 10-fold cross-validation on the entire set of annotated data, they were, however, also expensive in terms of annotation effort. The present study, therefore, explores if similar results can be achieved with fewer training samples. In addition, the lexicon matching is here further explored, as it was performed with a very limited lexicon by Cruz et al. (2015).

3 Methods

A lexicon-based and a machine learning-based approach for detecting the three stance categories were compared. The SFU Review corpus annotations by Konstantinova et al. (2012) and by Taboada and Hay (2008) were used for all experiments. These annotations were performed independently and at different times, with Konstantinova et al. segmenting the corpus into *sentences*, while Taboada and Hay used *segments*, which are often shorter. The two segmentation styles were reconciled, by using the sentence boundaries of the Konstantinova et al. corpus, except when the corresponding segment in the Taboada and Hay corpus was longer than this sentence boundary. In such cases, the segment annotated by Taboada and Hay was used as the sentence boundary.[2]

The *speculation* category in the Konstantinova et al. corpus was used for investigating *speculation*, and the *condition* category in the Taboada and Hay corpus for investigating the category *conditional*. Although these categories were somewhat overlapping, since *condition* was included in *speculation*, the categories were employed as defined and annotated in the previous studies. Since the two related categories *contrast* and *concession* are often conflated by annotators (Taboada and Mann, 2006), annotations of these categories in the Taboada and Hay corpus were combined, forming the merged category *contrast*. The speculation classification format previously used in the first of the CoNLL-2010 shared tasks (Farkas et al., 2010) and by Wei et al. (2013) was applied, that is an entire sentence was classified as either belonging to a stance category or not. The procedure used in CoNLL-2010 for transforming the data into this format was adopted, i.e., if either the scope of a *speculation* cue or a segment annotated for *concession/contrast* or *condition* was present

[1] Concession is defined by Mann and Thompson (1983) as "the relationship [that] arises when the speaker acknowledges, in one part of the text, the truth of a point which potentially detracts from a point in another part of the text."

[2] Ill-formed XML files from the Taboada-Hay corpus were discarded, making the corpus used a subset of the original.

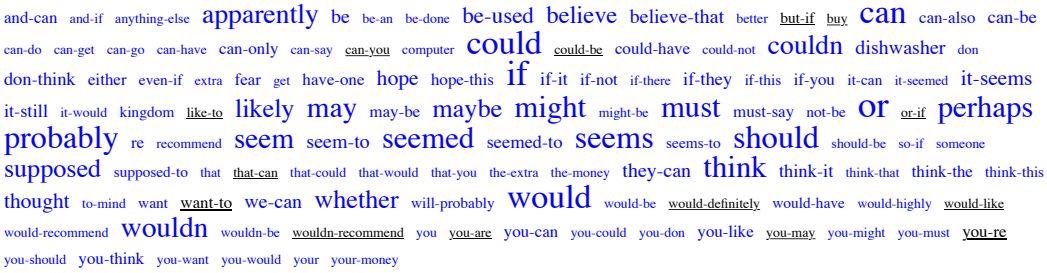

Figure 1: SVC-features selected for *speculation*, displayed in a font size corresponding to their feature weight. (Negative features underlined and displayed in black.)

# sentences	Spec.	Contr.	Cond.	Total
Training	1,184	432	220	5,027
Evaluation	1,217	459	230	5,028

Table 1: Frequencies of categories in data used.

in a sentence, the sentence was categorised as belonging to this category (or categories, when several applied). The sentence list was randomly split into two halves – as training and evaluation data (Table 1).

3.1 Machine learning-based approach (SVC)

A support vector classifier model, the LinearSVC included in Scikit learn (Pedregosa et al., 2011), was trained with bag-of-words and bag-of-bigrams as features. A χ^2-based feature selection was carried out to select the n best features. Suitable values of n and the support vector machine penalty parameter C were determined by 10-fold cross-validation on the training data.

The training and feature selection was carried out for different sizes of the training data; starting with 500 training samples and increasing sample size stepwise with additional 500, up to 5,000 samples. A separate classifier was always trained for each of the three categories, and the categories were evaluated separately.

3.2 Lexicon-based approach (Lexicon)

The lexicon-based approach used three lists of marker words/constructions, one list for each category of interest. Sentences containing constructions signalling any of the three categories were classified as belonging to that category. The lists were created by first gathering seed markers; for *speculation* from constructions listed by Konstantinova et al. (2012) and from a previous resource collected with the aim of detecting speculations in clinical texts (Velupillai et al., 2014), and for *contrast* from constructions listed by Reese et al. (2007). These seeds were then expanded with neighbours in a distributional semantics space (Gavagai, 2015) and from a traditional synonym lexicon (Oxford University Press, 2013). Finally, the expanded lists of candidates for *speculation* and *contrast* markers were manually filtered according to the suitability of included constructions as stance markers. From the list created for *speculation*, a subset of markers signalling *conditional* was selected to create the list for this category.

The final lists contained 191 markers for *speculation*, 39 for *contrast* and 26 for *conditional*.

		Prec.	Recall	F-score
Spec.	SVC	88.59%	95.07%	91.72
	Lexicon	83.41%	78.47%	80.86
Contr.	SVC	54.31%	69.93%	61.14
	Lexicon	43.07%	83.22%	56.76
Cond.	SVC	62.80%	80.00%	70.36
	Lexicon	57.18%	84.78%	68.30

Table 2: Precision, recall and F-score for the two approaches, when using all available training data.

4 Results

Results on the evaluation set for the two approaches (lexicon-matching and the SVC when using all training data) are shown in Table 2. Features selected when obtaining these SVC results are shown in a font size corresponding to their model weight in Figures 1 and 2, and markers found in the evaluation data when using the lexicon-based approach are shown in Figure 3.

Different training data sizes were evaluated with

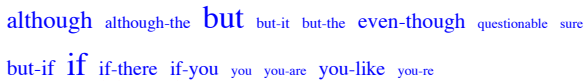

Figure 2: SVC-features selected for *contrast* (first row) and for *conditional* (second row).

bootstrap resampling (Kaplan, 1999). For each data size, 50 different models were trained, each time with a new random sample from the pool of training data. Figure 4 displays all results.

5 Discussion

Both approaches were clearly more successful for detecting *speculation* than for detecting *contrast* and *conditional*. When using the entire training data set, the SVC results for *speculation* were slightly higher than the human ceiling (an SVC F-score of 92, compared to an inter-annotator agreement of 89). The F-scores for *contrast* and *conditional* were, however, considerably lower (approximately 30 points lower and 20 points lower than *speculation*, respectively). The SVC results for the two latter categories also remain unstable for larger training data samples, but stabilise for *speculation* (Figure 4).

The higher F-score for *speculation* than for *contrast* and *conditional*, as well as its higher stability, might be explained by this category being more frequent than the other two. However, there seems to be a much greater variety in the way in which *speculation* is expressed, as shown by the number of SVC-features selected for this category and the number of markers that lead to true positives in the lexical approach, compared to what was the case for the other two categories. Lower recall was also achieved for the lexical approach for detecting *speculation*, despite the many stance markers used for this category. Therefore, it would seem reasonable to hypothesise that, while many training samples would be required for *speculation*, a smaller number of samples should be enough for the other categories. Language is, however, highly contextually adaptable, allowing the same construction to express different phenomena (Paradis, 2005; Paradis, 2015), and frequent English markers for *contrast* and *conditional* seem to be polysemous to a larger extent than *speculation* markers. E.g., 'while' sometimes expresses *contrast*, although it more often has a temporal meaning (Reese et al., 2007), which results in 30 true positives and 70 false positives when it is used as a marker for *contrast* in the lexicon-matching approach. Similarly, 'if' is, by far, the most frequently used marker for expressing *conditional*, as previously observed by Narayanan et al. (2009), and as shown here in the lexical approach, in which 98% of the true positives contained this marker. Despite that, 'if' is also used to indicate indirect questions and as a more informal version of 'whether' (Oxford University Press, 2013), which has a potential to give rise to false positives. In the scheme used by Konstantinova et al., on the other hand, most readings of 'if' were covered by their broad definition of *speculation*.

In addition, it cannot be disregarded that annotations from two different sources were used for the experiment, and that part of the differences in performance, therefore, might be attributed to differences in annotation quality. For the Konstantinova et al. corpus, there is a reliability estimate, which does not exist for the Taboada and Hay corpus. The Taboada and Hay annotation scheme might also be more difficult – as it included 36 annotation categories – and thus more error prone.

Comparing the SVC approach and the lexicon matching, it can be concluded that the only case in which machine learning clearly outperforms lexicon matching is when the SVC for detecting *speculation* is trained on at least 1,500–2,000 training samples. For the categories *contrast* and *conditional*, on the other hand, it can be observed that (1) the machine learning results are unstable, and (2) only very few features – and only positive ones – are used by the models. One point of applying machine learning for text classification is to be able to create models that are complex enough to overcome weaknesses of a lexicon-matching approach, e.g., weaknesses arising from the use of polysemous expressions. Despite being trained on more than 5,000 training samples, only a few features were, however, selected as relevant for *contrast* and *conditional*. Therefore, for automatic detection, it might be more resource efficient to focus the effort on further curation of the lexicons used, rather than on annotation of training data. The complexity of the model for *speculation* seems, however, to exceed what could easily be captured with lexicon-matching, since more features, including negative ones, were used. This further motivates the suitability of machine learning for the task of detecting *speculation*.

It should also be noted that SVC results for

I-think:40 *I-think:4* allegedly:1 almost-certainly:1 *and/or:3* apparently:7 appear:1 *appear:1* as-long-as:2 *as-long-as:3* *assume:2* assuming:2 *assuming:1* *assuming-that:1* believe:20 *believe:12* can-be:25 *chances-are:1* *condition:1* considered:5 *considered:3* could-be:14 doubt:1 *doubt:2* either:19 *either:16* estimate:2 expect:12 *expect:11* feels-like:1 *feels-like:2* *gives-the-impression:1* guess:8 *guess:8* guessing:1 have-a-feeling:1 if:288 *if:19* *implausible:2* *indicate:1* indicated:1 *indicated:1* indicating:1 *indicating:1* it-appears:1 it-can:7 *it-can:1* it-could:7 likely:4 *likely:1* may:43 *may:7* may-be:10 maybe:13 *maybe:3* might:33 *might:1* might-be:5 no-obvious:1 not-sure:6 *not-sure:4* or:220 *or:22* perhaps:15 *plausible:1* *points-to:1* possible:3 *possible:10* possibly:7 potential:4 *potential:3* probably:38 *probably:1* *question:3* seem:17 seemed:28 seeming:1 seems-like:7 should:63 *should:1* shouldn't:4 skeptical:1 *skeptical:2* suggest:9 *suggest:2* suggested:2 suggests:2 *suggests:1* suppose:9 supposedly:1 *supposedly:3* suspect:4 *suspect:1* *suspicion:1* think:66 *think:10* thought:29 *thought:8* *unconvinced:1* *under-the-impression:1* unless:5 *unless:12* unsure:2 *unsure:1* *versus:2* *vs:2* wether:1 whether:9 with-the-understanding-that:1 wonder-if:3 *wonder-why:2* wondering:1 *wondering:4* wondering-if:1 would:175 *would:5*

albeit:1 although:25 *although:9* anyway:1 *anyway:12* *at-the-same-time:3* but:287 *but:249* despite:4 *despite:5* even-if:1 *even-if:13* *even-so:2* however:9 *however:52* *in-contrast:2* in-spite-of:1 *in-spite-of:2* *on-the-contrary:1* on-the-other-hand:4 *on-the-other-hand:8* *regardless:2* still:17 *still:62* *that-said:4* then-again:1 *then-again:1* though:22 *though:28* whereas:2 while:33 *while:70* yet:13 *yet:20*

as-long-as:3 *as-long-as:2* *assuming-that:1* *condition:1* if:192 *if:115* *unless:17* *wether:1* whether:1 *whether:8* *with-the-understanding-that:1*

Figure 3: Constructions leading to true positives (in green) and false positives (in black/italic) for the lexicon-based approach (and number of occurrences as true or false positive). The first group shows constructions for *speculation*, the second group for *contrast* and the third for *conditional*.

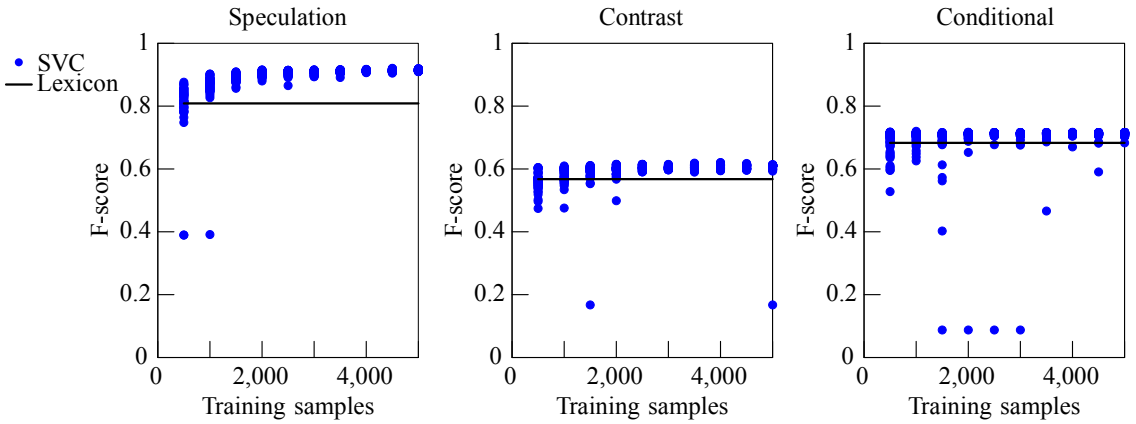

Figure 4: All 50 evaluation results for each random resampling, for each evaluated training data size.

speculation stabilise on high levels already with 3,000 training instances. This shows that results comparable to those of previous studies can be achieved with a smaller amount of training data. For instance, the most closely comparable study by Cruz et al. (2015) achieved the F-score of 92 for detecting speculation cues using 10-fold cross-validation on the entire SFU review corpus. For *contrast* and *conditional* on the other hand, it is difficult to make comparisons to previous studies, as such studies are scarce, but e.g., Clark et al. (2011) achieved an F-score of 89 and 42, respectively, for detecting the related categories *hypothetical* and *conditional*.

In future work, inclusion of additional features for training models for stance detection will be attempted (e.g., syntactic features or distributional features), and the usefulness of applying the detection on extrinsic tasks, such as sentiment analysis (Narayanan et al., 2009), will be further evaluated.

6 Conclusion

For detecting sentences with *speculation*, an SVC trained on bag-of-words/bigrams performed around 10 points better than a lexicon matching approach. When using between 3,000-5,000 training instances, the model performance was stable at an approximate F-score of 90, which is just above the inter-annotator agreement F-score. For detecting *conditional* sentences and sentences including *contrast*, however, the results were lower (an F-score of around 60 for *contrast* and around 70 for *conditional*). On average, the F-score for the machine learning models for these two categories was a few points better than for the lexicon-based methods, but these better results were achieved by models that only used eight features (which were all positive). This, together with the fact that the machine learning models showed a large variance, indicates that a lexicon-based approach, with a well-curated lexicon, is more suitable for detecting *contrast* and *conditional*.

Acknowledgements

This work was funded by the StaViCTA project, framework grant "the Digitized Society – Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council (Vetenskapsrådet).

References

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the ntcir-11 mednlp-2 task. In *Proceedings of NTCIR-11*.

Douglas Biber. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2):97–116.

Cheryl Clark, John Aberdeen, Matt Coarr, David Tresner-Kirsch, Ben Wellner, Alexander Yeh, and Lynette Hirschman. 2011. Mitre system for clinical assertion status classification. *J Am Med Inform Assoc*, 18(5):563–7.

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*, pages 526–558.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Stroudsburg, PA. Association for Computational Linguistics.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, April.

Matteo Fuoli. 2015. A step-wise method for annotating appraisal. (Under review).

Gavagai. 2015. The Gavagai living lexicon. lexicon.gavagai.se.

Susan Hunston and Geoffrey Thompson. 2000. *Evaluation in Text 'Authorial Stance and the Construction of Discourse'*. Oxford University Press, Oxford.

Susan Hunston. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language.* Routledge, New York and London.

Daniel Kaplan. 1999. Resampling stats in matlab. http://www.macalester.edu/˜kaplan/Resampling/ (accessed August 2015).

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğanur, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

William C. Mann and Sandra A. Thompson. 1983. Relational propositions in discourse. Technical report, No. ISI/RR-83-115, Marina del Rey, CA: Information Sciences Institute.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oxford University Press. 2013. Oxford thesaurus of English. Digital Version 2.2.1 (156) on Mac OS X.

Carita Paradis. 2005. Ontologies and construals in lexical semantics. *Axiomathes*, 15(4):541–573.

Carita Paradis. 2015. Meanings of words: Theory and application. In Ulrike Hass and Petra Storjohann, editors, *Handbuch Wort und Wortschatz (Handbücher Sprachwissen-HSW Band 3)*. Mouton de Gruyter, Berlin.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis, and Jason Baldridge. 2007. Reference manual for the analysis and annotation of rhetorical structure. timeml.org/jamesp/annotation_manual.pdf (accessed May 2015).

Axel Schulz, Tung Dang Thanh, Heiko Paulheim, and Immanuel Schweizer. 2013. A fine-grained sentiment analysis approach for detecting crisis related microposts. In *Proceedings of the 10th International ISCRAM Conference*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161.

Maite Taboada and Montana Hay. 2008. The SFU review corpus. www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html (accessed May 2015).

Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8:423–459.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genoa, Italy. European Language Resources Association (ELRA).

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In Paul Clough, Colum Foley, Cathal Gurrin, GarethJ.F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 368–374. Springer Berlin Heidelberg.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Natural Language Learning*, Stroudsburg, PA. Association for Computational Linguistics.

Özlem. Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality Levels of Diagnoses in Swedish Clinical Text. In A. Moen, S. K. Andersen, J. Aarts, and P. Hurlen, editors, *Proc. XXIII International Conference of the European Federation for Medical Informatics – User Centred Networked Health Care*, pages 559–563, Oslo, August. IOS Press.

Sumithra Velupillai, Maria Skeppstedt, Maria Kvist, Danielle Mowery, Brian E Chapman, Hercules Dalianis, and Wendy W Chapman. 2014. Cue-based assertion classification for swedish clinical text–developing a lexicon for pyConTextSwe. *Artif Intell Med*, 61(3):137–44, Jul.

Sumithra Velupillai. 2012. *Shades of Certainty – Annotation and Classification of Swedish Medical Records*. Doctoral thesis, Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden, April.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra1, and János Csirik. 2008. The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Veronika Vincze. 2010. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 28–31, Stroudsburg, PA. Association for Computational Linguistics.

Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2013. An empirical study on uncertainty identification in social media context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 58–62, Stroudsburg, PA. Association for Computational Linguistics.