Georgia Institute of Technology

Jacob Eisenstein
20 Aloha St.
Seattle, WA 98109
USA

June 21, 2019

This document contains my review of the dissertation of Uladzimir Sidarenka.

Overall comments

This thesis is entitled "Sentiment Analysis of German Twitter", but its ambitions are much greater than this title implies: it features not only a new and detailed corpus, but also new methods for sentiment lexicon induction, aspect-based opinion mining, lexicon-driven neural attention, and discourse-based sentiment analysis. Many of these methods have applications well beyond German Twitter, and in many cases, beyond sentiment analysis as well. These ambitions are a strength of thesis, but it occasionally struggles to find a balance between its linguistic and machine learning goals.

The thesis demonstrates an aptitude and passion for scientific communication that far exceeds even the high expectations that accompany a PhD dissertation. In each chapter, the work is introduced with an overview of the history of the sub-field that is illuminating, engaging, and (usually) concise. The only hesitation I have in complimenting this writing style is that the author's own methodological contribution sometimes gets a little lost. Rather than using the history to set the stage for the novel contribution offered by the dissertation, the survey of related work sets up a "bake off" application of prior work to the specific case of German. As a result, the novel methodological aspects of chapters 3-5 come as a bit of a surprise, and are less grounded in prior work than would be ideal.

Another strength of the dissertation is a set of very thorough evaluations, including comprehensive ablations and error analyses that lend linguistic heft to the technical contributions.

While the novel methods were not motivated as well as I would have liked in the document, I think that they target important problems for sentiment analysis and document analysis more generally: in particular, how to combine the linguistic knowledge inherent in hand-crafted lexicons with the distributional information built into word embeddings, and how to leverage discourse structure in document-level analysis.

A weakness of the dissertation is that it was less solidly grounded in machine learning than in linguistics and natural language processing. This would not necessarily be a problem in an NLP dissertation, but for the aforementioned ambitions toward methodological innovation in sophisticated areas of machine learning. Some of the proposed machine learning techniques seem to be closely adjacent to existing methods such as linear discriminant analysis (relevant to the embedding projection method in chapter 3) and hidden-variable CRFs (relevant to the latent CRF methods in chapter 6). These existing methods generally have better theoretical properties (e.g., convergence), and in any case are the logical starting points for research

of this kind. Similarly, there are technical issues that raise some questions, such as the combination of softmax and hinge loss in chapter 5, and the clarity of the RDP method in chapter 6. Pushing this research through the peer review process might have identified these issues and connections at an earlier stage.

As a final comment, I'll note that in many ways, this dissertation represents much more work than expected for a PhD thesis – more experiments, more dataset contributions, more new methods, more replications of existing methods, and more detailed discussion of prior research. I believe that it has much to offer future researchers, even if the specific methods proposed here are not directly adopted.

Detailed notes

I will now move to more specific comments and recommendations.

Chapter 1 is a very well researched and well written introduction to sentiment analysis, and **chapter 2** provides a very well organized and comprehensive approach to both data collection and annotation. For replicability, it would be good to include the complete keyword lists alongside the annotator instructions in an appendix.

I want to poke a little at the definition of targets as "entities or events evaluated by opinions." How generally should I think of entities? For example:

- Attributes, e.g. "Her tact is lacking but her honesty is commendable."
- Propositions, e.g. "I'm glad I'm not allergic to ice cream."
- Sets, e.g. "I'm tired of Californians who think they know about pizza."
- Hypothetical entities, e.g. "Dragon meat would probably not taste very good."

My impression is that all of these would count, but perhaps then the term "entities" is a little misleading, at least as it is typically used in computational linguistics.

The notion of annotating comparative relations ("I'd rather have a bottle in front of me than a frontal lobotomy") is new to me, and I think it fills a significant gap in existing annotations of targeted sentiment.

I wonder whether the initial low levels of agreement stemmed from a lack of clarity in the original instructions. The approach of initial annotations, followed by adjudication and then further annotations, seems reasonable for generating high-quality annotations once. But it might be difficult to extend such a corpus with new annotators; those annotators would have to be trained in the same way, and adjudication of their disagreements might result in a different equilibrium than the first group.

The qualitative analysis in section 2.6.4 is quite thorough, and a nice complement to the quantitative results. Example 2.6.2 raises an interesting distinction, and I'm not quite satisfied with the position that both interpretations are correct, since it's not clear to me what is being interpreted: if it is the text that's fine, but I worry that it's really the task.

It was very interesting to see agreement across topics and categories; again, this section is unusually thorough.

I didn't understand the correlation analysis in table 2.6.

Chapter 3 offers an extremely comprehensive exploration of techniques for identifying lists of sentiment words, including both manual and automated techniques, and re-implementing a huge amount of prior work. These results are bolstered by their link to well-validated instance-level annotations, which are another contribution of the dissertation. Overall, the chapter has convinced me that the problem is much more challenging than I would have guessed, although I would emphasize that the "precision" numbers are all lower bounds, since more annotations would presumably yield more true positives for any lexicon. The exploration of seed word sets was particularly useful.

One place that I would like to see the work go further is in cross-lingual comparisons. In particular, I wonder whether the relatively low F1 results are a property of German, and if so, why? Alternatively, they could be a property of validating in this particular way, and English word lists might fare similarly if compared against a similar set of annotations.

If I understand correctly, the goal of the novel projection technique is find some projection such that each pair of (positive, negative) words are far apart in the projected space. A classical technique for projection with a very similar objective is Linear Discriminant Analysis. Unfortunately, understanding of this technique has been confused by the textbook by Hastie et al, who treat it as a baseline classification algorithm, when in fact, it can be used more generally for dimensionality reduction. This is explained in the older textbook by Duda and Hart, and their treatment is summarized in this blogpost:
`https://sebastianraschka.com/Articles/2014_python_lda.html`

The relevant point for this thesis is that a very similar objective can be optimized by solving a generalized eigenvalues problem, with global optimality guaranteed. This seems advantageous in comparison with the proposed method, which finds only a local optimum by gradient ascent.

Chapter 4 concerns the problem of fine-grained aspect-based sentiment analysis. While I am not an expert in this domain, the evaluations seem reasonable. Based on this description of preprocessing, it would not be possible to exactly replicate these preprocessing steps: it would be necessary to know exactly which emoticons, which misspellings, and which slang terms were normalized. I appreciate that this information is already published, but for completeness it might have been nice to include it here. I was also glad to read that the whole pipeline is available online. It was also good to see the ablation experiments that demonstrate the merit of normalization for this problem.

While a detailed recapitulation of the issue of label bias is interesting, it seems a bit outside the scope of this thesis. Nonetheless, I appreciate the passion for communicating technical material, and I hope that the author has the opportunity to do this in his future career.

The approach described in section 4.3 seems very reasonable, but I am troubled by the extreme amount of overfitting. It might help to remind readers of this size of the training and test sets, and to indicate how many features from the training set are unseen in the test set, and vice versa. I would also like to see how F1 evolved across the space of regularization parameters, and to know how the final regularization parameter was selected.

Page 69 notes that the weaker performance of RNN-based models should be viewed as evidence that hand-crafted features contain additional information. Perhaps. Another difference is the role of structured prediction, which is employed in the CRF, mitigating the label bias problem. It is possible to combine structured prediction with RNNs, as described in chapter 7.6 of my textbook (an early citation is the LSTM-CRF of Huang et al 2015, although I am not sure this is the first). I suspect such an approach might close the gap between RNNs and CRFs, without adding any features.

The use of alternative graph structures in CRFs is quite interesting and creative. It seems that the semi-markov model yields significant improvements, which is also interesting and worth further exploration. I would have liked to know more about how inference and learning was implemented in these structures, since the "off-the-shelf" Viterbi and forward-backward algorithms are not immediately applicable to Semi-Markov and Tree-structured models. I was unable to understand the specific factor graphs from Figure 4.2, and would have liked more mathematical detail on this.

Chapter 5 moves to approaches to message-level classification. It includes a large number of datasets. I am skeptical of the use of emoticons to label tweets, despite the fact that this is done in prior work: there's good evidence that the "smiley" emoticon is used for many pragmatic purposes aside from indicating sentiment, such as softening face-threatening speech acts (e.g., Skovholt et al 2014).

There is a very nice overview of lexicon-based classification, giving equal focus to the key historical events as well as the most relevant techniques for today.

I was surprised by the choice of TreeTagger, since it is over 20 years old. Would a CRF and LSTM-based tagger, trained on Universal Dependencies data, not do as well or better? Similarly, I wonder why MateParser rather than something more contemporary.

The evaluation of polarity-changing factors in 5.3.1 is interesting. I would like to know more about why these factors generally do not help. Is it because the theory about them is wrong (i.e. negation and intensification don't work the way that we think they do), or because these systems are not capturing them accurately? Or are they captured accurately but not incorporated effectively?

5.4 is a nice summary of prior work on this topic – great history. This history is then translated into an empirical evaluation. As noted above, it would interesting to understand the contrasts, if any, between the results in the two German corpora and the results for the same methods on English SemEval data. The feature analysis in Table 5.6 was particularly helpful in understanding the differences between the three main systems, although the character features for MHM are hard to interpret. I'm less interested in the contrast between classifiers, since these differences are unlikely to be specific to the problem of sentiment analysis in German. There are well-known relationships between these three classifiers in particular: Naive Bayes is unlikely to do well in feature-rich settings, and (linear) SVM and logistic regression generally give similar performance, since they are optimizing similar objectives.

5.5 again begins with a nice history. I particularly appreciate the reference to Yessanlina and Cardie 2011, an idea that represents an intriguing alternative evolutionary path for the

neural revolution in NLP.

The neural architecture proposed in this section is interesting, although I'm not ready to accept the motivating arguments on page 106. The BiLSTM is trained to produce a representation at each token that simultaneously captures polarity and salience, with the latter factor controlling the attention weights. You can think of these two aspects as living on different parts of the vector h_i, although of course in practice they will be superposed. The point is that there is no doubt that the BiLSTM has sufficient expressive power to represent both of these aspects of each token. I would motivate the modeling approach differently: lexicons of polar terms and shifters provide an alternative way to incorporate linguistic knowledge into sentiment classification, and an impractical number of labeled examples would be required in order to get the same information through supervised training alone.

The failure of nearly all systems to predict negative labels raised questions in my mind about the validity of this evaluation, particularly since this problem was not observed with the simpler feature-based classifiers in previous sections. Logistic regression is equivalent to a softmax output layer and a cross-entropy objective, so I would expect it to suffer the same problem if the issue really is skew in the label distribution. Page 108 states that the models were trained on a "categorical hinge loss", but this may not be a good idea in combination with a softmax output layer, which would make it impossible to satisfy the margin constraint. Typically softmax is used in combination with a cross-entropy loss; otherwise, this monotonic transformation is unnecessary. It was good to see that initialization of the word embeddings seemed to resolve this issue, but there may be some remaining problem which, if addressed, could unlock even better performance.

I would relabel "distant supervision" as "semi-supervised learning" or "weak supervision", as "distant supervision" typically refers to supervision from type-level resources such as knowledge bases.

The normalization results are quite interesting, since to my knowledge similar benefits have not been obtained for English. I wonder whether this is a specific feature of German, or whether the normalization used here is just more effective.

Chapter 6 moves to discourse-driven sentiment analysis, and begins with a comprehensive history of both discourse parsing and its application to sentiment analysis.

A few technical points arise in this chapter. The objective arrived at in the Latent CRF (pg 137, eq 6.3), is very nearly the same as the latent variable perceptron. The key difference is that the latent variable perceptron allows y' to include the correct label; when $y = y'$, there is no gradient from the example. Some derivations are found in my textbook, and the idea goes back to Collins. A true latent-marginalized CRF is called the "hidden CRF" by Quattoni et al (2007), where they do achieve summation over all y_h. While this is not possible in general, I think it is possible in tree-structured models like the one considered here. The use of the ratio of marginal probabilities is an interesting solution, but again it requires identifying a single incorrect prediction y'.

The recursive Dirichlet process is an intriguing Bayesian approach to the problem, although I was not able to fully understand what was done here. My main confusion is in the training

and inference of the model, and in the relationship between the two. In Bayesian models of this sort, the "training" phase usually involves estimating the parameters of the priors, which in this case would seem to be ζ, μ_r, and Σ_r. These parameters would then be re-used across all documents. Alternatively, we can treat the priors as true latent variables, and either marginalize over them or sample them. Similarly, we would sample or marginalize over the latent variables M_r, β, α, and z. Marginalization might involve computing some variational distribution, e.g., $Q(M_r, \beta, \alpha, z)$. Pages 144-145 seem to suggest that some sort of Q distribution was computed, but I would need more details to understand how this was done. For completeness, I would also like to have more details about the use of the sparsemax function here.

As a minor nomenclature quibble, I don't think it's correct to call this method a Recursive Dirichlet *Process*. The Dirichlet Process is a nonparametric model which induces Dirichlet distributions over all partitions of a sample space. In the RDP, there is only a single Dirichlet distribution over the space of vectors of size 3. In other respects, the model seems well designed, and is an interesting step beyond Bhatia et al.

As usual, this chapter includes a comprehensive evaluation of many different techniques. The discourse driven techniques perform similarly to the "no discourse" baseline, with the gap probably not meeting the standard of statistical significance. It seems likely that these techniques might be better suited to longer texts, where the no-discourse baseline is likely to be weaker.

I was particularly interested to see the evaluation with gold RST trees, as shown in table 6.2. But I couldn't understand why the No-Discourse method also improved in this setting.

Regarding the **conclusions**, I like the idea of providing "rules of thumb" for future work in this domain and elsewhere (pg 160). The advice not to use randomly initialized word embeddings is the only point with which I would quibble, and only because I am anecdotally aware of a lot of other research that seems to have come to the opposite conclusion. I was also happy to see so much of the code and pipeline made available on github.

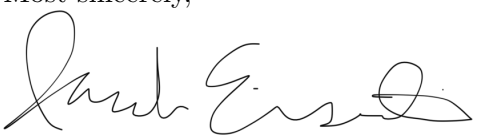
As a minor point, if you run bibtex with `--min-crossrefs=100`, then bibliographical entries like (2014) will be removed.

Grade

The dissertation was a pleasure to read, and represents an achievement that the student

should be proud of. I recommend a grade of **magma cum laude**.

Most sincerely,



Jacob Eisenstein
Assistant Professor
School of Interactive Computing
College of Computing
Georgia Institute of Technology
1 617 913 2859