Kim Schouten
PhD candidate
Erasmus University Rotterdam
schouten@ese.eur.nl
www.kimschouten.com

Implicit Feature Extraction for Sentiment Analysis in Consumer Reviews

Kim Schouten & Flavius Frasincar

Abstract

With the increasing popularity of aspect-level sentiment analysis, where sentiment is attributed to the actual aspects, or features, on which it is uttered, much attention is given to the problem of detecting these features. **While most aspects appear as literal words, some are instead implied by the choice of words.** With research in aspect detection advancing, we shift our focus to the less researched group of implicit features. By leveraging the co-occurrence between a set of known implicit features and notional words, we are able to predict the implicit feature based on the choice of words in a sentence. Using two different types of consumer reviews (product reviews and restaurant reviews), an F1-measure of 38% and 64% is obtained on these data sets, respectively.

Original Method

The original method of Zhang and Zhu [1] counts co-occurrences between explicit features and notional words in a sentence. In that way, a feature is found to implied by the words in the sentence if that feature co-occurs most with the words in the sentence throughout the corpus. The score is computed as

$$score_{f_i} = \frac{1}{v}\sum_{j=1}^{v}\frac{c_{i,j}}{o_j}$$

where

- f_i is the **i**th feature in the set of possible features;
- j is the **j**th lemma in the sentence;
- v is the number of lemmas in the sentence;
- $c_{i,j}$ is the co-occurrence between feature **i** and lemma **j**; and
- o_j is the occurrence frequency of lemma **j**.

Tested Hypotheses

Explicit features are a good proxy for implicit features

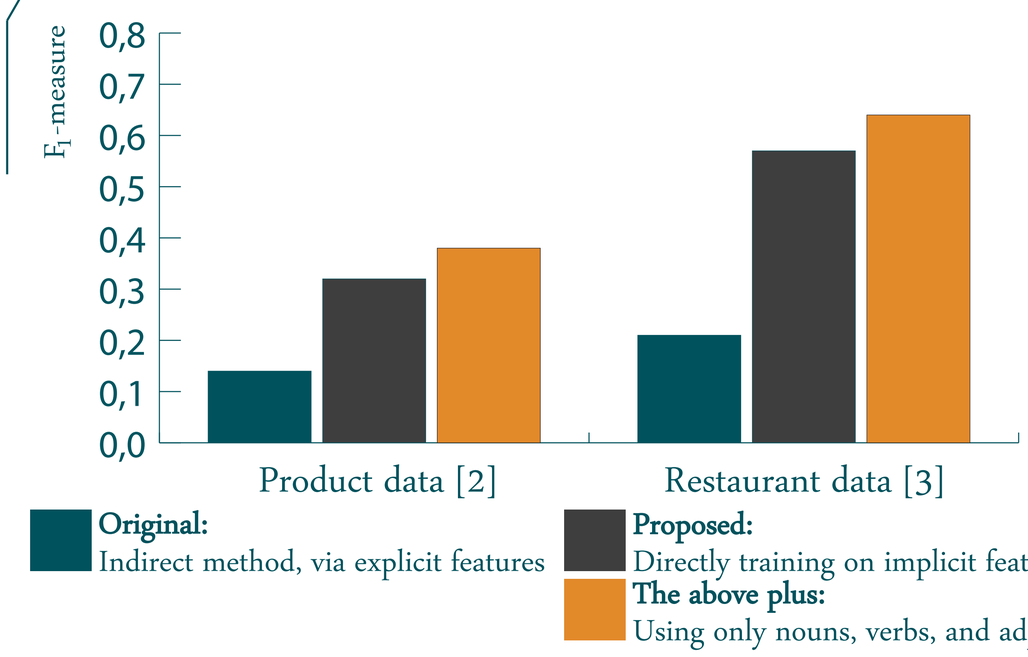
To test this assumption, we counted co-occurrences between annotated implicit features in a sentence and the words in a sentence. This directly links the used words in a sentence to the implicit feature.

Drawback: this makes the method supervised instead of unsupervised.

All word types contribute to performance

To test this assumption, the co-occurrence matrix is built from a limited set of words, instead of all words. This selection is done based on the Part-of-Speech tag of the words.

Results



Original:
Indirect method, via explicit features

Proposed:
Directly training on implicit features

The above plus:
Using only nouns, verbs, and adjectives

Data

The product data consists of a set of product reviews, for five different products, taken from Amazon [2]. The restaurant data is the same as used in SemEval-2014 Task 4 Aspect-Based Sentiment Analysis, except that the 'miscellaneous' category is removed, since it is not really an implicit feature.

Evaluation & Conclusion

Directly training on the implicit features clearly yields better results compared to using explicit features as an intermediary step. This is true for both data sets.

Only using certain word types for the co-occurrence matrix is helpful: when considering only nouns, verbs, and adjectives, the proposed method performs better on both data sets. The original method is unaffected by this change.

Future Work

Several possiblities for future work have been determined, and some of them have been tried in the mean time.

1. **Test the effect of employing word-sense disambiguation and then create the co-occurrence matrix based on synsets instead of lemmas:** in our experiments, this did not improve the performance.
2. **Test the effect of introducing separate thresholds for each implicit feature, enabling the algorithm to find more than one implicit feature in a sentence:** this seems to improve the performance, but introduces major overfitting issues that need to be dealt with.
3. **Introduce a weighting scheme into the co-occurrence matrix, given more weight to more relevant co-occurrences.** This would require additional domain knowledge, for example in the form of ontologies.

References

1. Y. Zhang and W. Zhu. Extracting Implicit Features in Online Customer Reviews for Opinion Mining. In Proceedings of the 22nd International Conference on World Wide Web Companion (WWW 2013 Companion), pages 103-104. International World Wide Web Conferences Steering Committee, 2013.
2. M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pages 168-177. ACM, 2004.
3. G. Ganu, N. Elhadad, and A. Marian. Beyond the Stars: Improving Rating Predictions using Review Content. In Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009), 2009.

ERASMUS UNIVERSITEIT ROTTERDAM

COMMIT/