Implicit Feature Detection for Sentiment Analysis

Kim Schouten & Flavius Frasincar

Abstract

Implicit feature detection is a promising research direction that has not seen much research yet. Based on previous work, where co-occurrences between notional words and explicit features are used to find implicit features, this research critically reviews its underlying assumptions and proposes a revised algorithm, that directly uses the co-occurrences between implicit features and notional words. The revision is shown to perform better than the original method, but both methods are shown to fail in a more realistic scenario.

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 Assumptions

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.

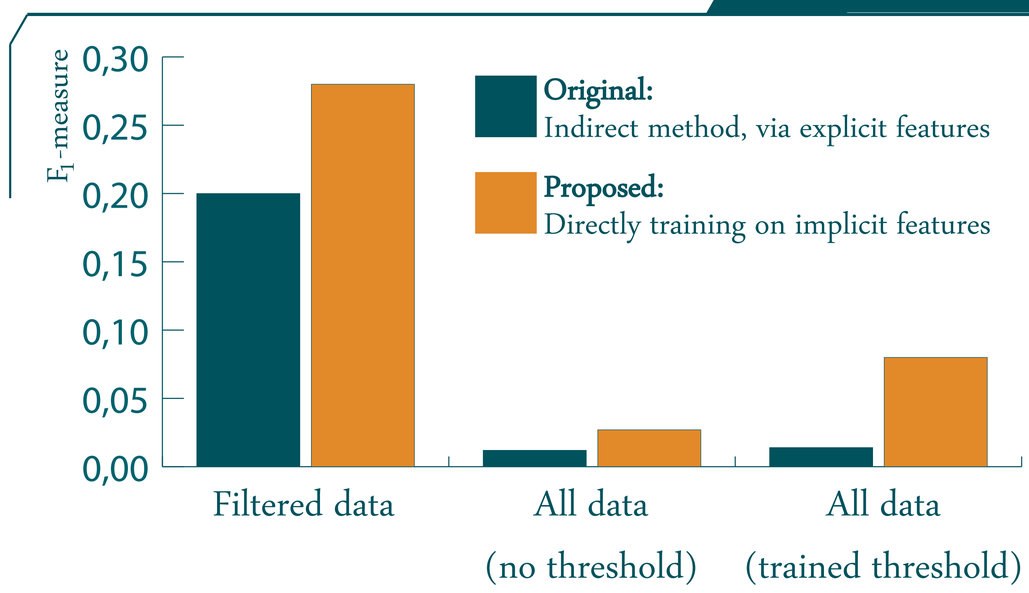
Drawback: this makes the method supervised instead of unsupervised.

Sentences are known to have an implicit feature

To test this assumption, we executed the algorithm on a data set that contained also sentences without implicit features and on a filtered version of that data set that contained only sentences with an implicit feature.

Drawback: a threshold on score needs to be trained to be able to not assign any implicit feature to a sentence. Otherwise, almost any sentence will be assigned an implicit feature, as there is bound to be one word in that sentence that co-occurs with some implicit feature.

Results



Data

The data set is a set of product reviews, for five different products, taken from Amazon [2].

# of implicit features per sentence:	0	1	2
# of sentences:	3797	140	8

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 all three settings that have been tested.

Removing the assumption that sentences are known to have one implicit feature turned out to be devastating for performance, yielding results that are unsuitable to work with. Even the addition of a trained threshold, while significantly boosting performance compared to not using such a threshold, cannot save the algorithm from failing miserably in this much more realistic scenario.

Future Work

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

1. Test on a different data set: we used a slightly updated version of a set of restaurant reviews [3], where the aspect categories are used as implicit features with the 'misc' category being removed. F-measure on this data set is ~0.6.

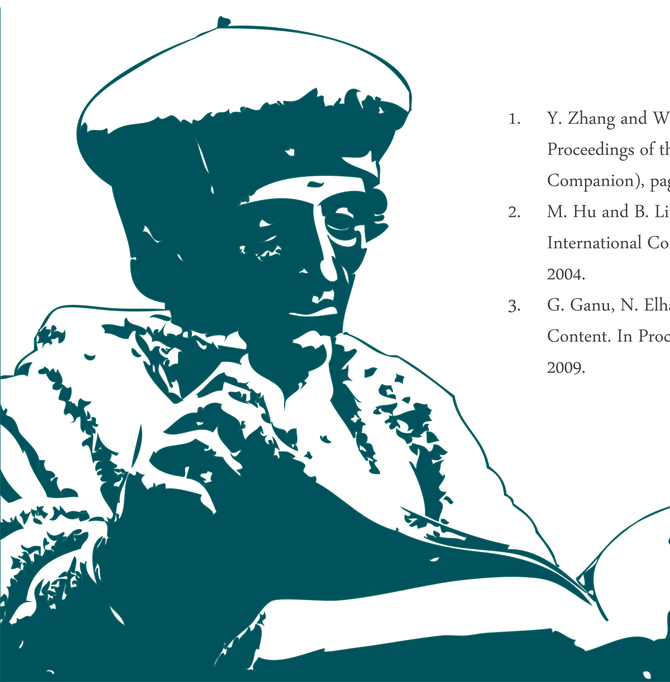
2. Test the effect of filtering words based on part-of-speech during the creation of the co-occurrence matrix: we found this to be helpful, yielding a minor improvement to performance.

3. 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.

4. Test the effect of having a separate classifier that determines whether a sentence has an implicit feature, and then letting this algorithm decide which one it is: this part of the proposed future work remains open.

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 Ratig Predictions using Review Content. In Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009), 2009.



ERASMUS UNIVERSITEIT ROTTERDAM

COMMIT/