Implicit Feature Extraction for Sentiment Analysis in Consumer Reviews

Kim Schouten and Flavius Frasincar

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands
{schouten,frasincar}@ese.eur.nl

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 F_1-measure of 38% and 64% is obtained on these data sets, respectively.

1 Introduction

Every day a vast amount of consumer reviews are written on the Web, where customers express their opinions about a product or service [1]. Not only do they describe their general sentiment or attitude towards the product or service, oftentimes specific aspects or features of that product or service are discussed in great detail [5]. This leaves researchers and companies alike with a valuable source of information about consumer sentiment.

However, in order to achieve the fine grained information that is needed for such analyses, the various aspects, or features, of a product or service must be recognized in the text first. Examples of such features include 'price', 'service', parts of a product like 'battery', or different meals and ingredients for restaurants. In most cases, these features are literally mentioned in the text. However, this is not always the case, as demonstrated in the example below, which is taken from the product review data set [4]:

"I like my phones to be small so I can fit it in my pockets."

Evidently, the feature referred to here is the size of the product, even though the word 'size' is never mentioned. However, words like 'small' and 'fit' give away that the feature implied here is the product's size. Unfortunately, detecting the implicit features is not always this straightforward.

"I love the fact I can carry it in my shirt or pants pocket and forget about it."

The above example is an actual sentence in the product review data set, and according to the available annotations, the implicit feature in this case is also its size, however, it is easy to see that weight would also have been a good candidate.

2 Related Work

Earlier works that focus on implicit feature extraction are [6], where implicit features are found using semantic association analysis based on Point-wise Mutual Information, and [3], which uses co-occurrence Association Rule Mining to link opinion words as antecedents to implicit features as consequents.

Instead of linking opinion words to implicit features, [7] constructs a co-occurrence matrix between notional words and explicit features, using these co-occurrences to imply a feature in a sentence when no explicit feature is present. For each feature f_i, a score is computed that essentially is the sum of the co-occurrence frequencies $c_{i,j}$ between that feature i and the words j in the sentence.

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

where v is the number of words in the sentence, f_i is the ith feature for which *score* is computed, w represents the jth word in the sentence, $c_{i,j}$ is the co-occurrence frequency of feature i and lemma j, and o_j is the frequency of lemma o in the data set.

3 Method

To deal with the two violated assumptions mentioned in Sect. 2, a small but significant change is made in the construction of the co-occurrence matrix: instead of explicit features, manually annotated implicit features are used. This results in direct co-occurrence data between words and the implicit features that are to be determined. This change renders the two violated assumptions irrelevant, but it also introduces a dependence on annotated data.

Furthermore, all words are considered as context for implicit features. However, some word categories might be more useful as context words than others. Therefore we investigate a Part-of-Speech filter in which all combinations of word categories (i.e., noun, verb, adjective, and adverb) are tested as context words.

4 Results Analysis

All evaluations are performed using 10-fold cross-validation, with all sentences without an implicit feature being removed from the test set. The evaluation

metric is the F_1-measure. Since precision and recall do not differ that much, only F_1-measure is reported. The two data sets that are used are a set of product reviews [4] and a set of restaurant reviews [2]. Note that for the latter, the "miscellaneous" category is removed as it is not really an implicit feature. Testing all mentioned Part-of-Speech filters pointed to the combination of nouns (NN), verb (VB), and adjectives (JJ) as the most effective context to find implicit features. This was the case for both data sets. The performance metrics using this context are given in Table 1.

Table 1: Evaluation results on both data sets, denoted as F_1-measure.

	original method		revised method	
	all words	only NN, VB, and JJ	all words	only NN, VB, and JJ
product data	0.14	0.14	0.32	0.38
restaurant data	0.21	0.21	0.57	0.64

In general, one can conclude that directly creating the co-occurrence matrix with implicit features instead of indirectly with explicit features is a good strategy. The performance gain is significant, which will offset the disadvantage of needing labeled data. In terms of overall performance, the revised algorithm works best with a Part-of-Speech filter that only allows nouns, verbs, and adjectives. Concerning data sets, the revised algorithm works best with the restaurant data, which is relatively large and has only five different implicit features to choose from. Using the product data results in the worst performance, due to its limited size and increased difficulty: it has more different implicit features than the restaurant data and less instances per unique implicit feature. This makes it hard to properly train the algorithm.

5 Conclusion

The detection of features from reviews is important when measuring consumer sentiment on a fine-grained level. Adding the detection of implicit features, while a difficult task because the features themselves do not appear in the sentence, can increase the overall coverage of an aspect-level sentiment analysis tool. Besides a base method [7], two main revisions were discussed and evaluated on two data sets [2,4].

There are two main conclusions that can be drawn from the performed evaluation. The first is that it is much better to count the co-occurrence frequency between annotated implicit feature and notional words than to count the co-occurrence frequency between explicit features and notional words. Since the number of implicit features is usually much smaller than the number of explicit features, this will greatly reduce the size of the co-occurrence matrix as well,

yielding better performance in terms of system load and processing time. The only drawback would be that this method is more domain dependent, as annotations of implicit features are required to train the system (i.e., do the counting). The second is that filtering which words are allowed as context from which the implicit feature is derived is indeed helpful, albeit only for the revised method. A combination of nouns, verbs, and adjectives turns out to be most informative to extract the right implicit feature.

Possible directions for future work might include an extension to deal with more than one implicit feature in a sentence. While this is arguably not useful for the product review data, roughly one sixth of the restaurant review sentences has more than one implicit feature, rendering this a good way of reducing the number of false negatives. Another option might be to introduce a weighting scheme for the co-occurrences where the co-occurrence with different words can be weighted differently, based on for example additional domain or world knowledge. This could, for example, be taken from structured data like ontologies.

Acknowledgments

The authors are partially supported by the Dutch national program COMMIT. We also would like to thank Sven van den Berg, Marnix Moerland, Marijn Waltman, and Onne van der Weijde, for their contributions to this research.

References

1. R. Feldman. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4):82–89, 2013.
2. 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.
3. Z. Hai, K. Chang, and J. Kim. Implicit Feature Identification via Co-occurrence Association Rule Mining. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text processing (CICLing 2011)*, volume 6608, pages 393–404. Springer, 2011.
4. 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.
5. B. Liu. *Sentiment Analysis and Opinion Mining*, volume 16 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool, 2012.
6. Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su. Hidden Sentiment Association in Chinese Web Opinion Mining. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 959–968. ACM, 2008.
7. 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.