Linguistic Sentiment Features for Newspaper Opinion Mining

Thomas Scholz and Stefan Conrad

Heinrich-Heine-University, Institute of Computer Science, Düsseldorf, Germany
{scholz,conrad}@cs.uni-duesseldorf.de

Abstract. The sentiment in news articles is not created only through single words, also linguistic factors, which are invoked by different contexts, influence the opinion-bearing words. In this paper, we apply various commonly used approaches for sentiment analysis and expand research by analysing semantic features and their influence to the sentiment. We use a machine learning approach to learn from these features/influences and to classify the resulting sentiment. The evaluation is performed on two datasets containing over 4,000 German news articles and illustrates that this technique can increase the performance.

Keywords: Opinion Mining, Sentiment Analysis, Media Response Analysis.

1 Introduction

Every day, many news texts are published and distributed over the internet (uploaded newspaper articles, news from online portals). They contain potentially valuable opinions. Many organisations analyse the polarity of sentiment in news items which talk about them. How is the media image about company XY? Is the sentiment changing after the last advertising campaign? For instance, a Media Response Analysis (MRA) answers these questions [12]. In a MRA, several media analysts have to read the collected news, select relevant statements from the articles and assign a sentiment for each statement. This means in effect, a MRA needs a big human effort. At the same time, the internet contains more and more potentially relevant articles. As a consequence, media monitoring services require more machine-aided methods. Opinions are not stated so clearly in newspaper articles [1]. In the news, some special features are important for the sentiment, so that an only-word-based method cannot solve this problem.

Formal Task Definition: *Given a statement s which consists of the words w_i with $i \in \{1, ..., s_n\}$. The task is to find the polarity of sentiment y for the statement s:*

$$f : s = (w_1, ..., w_{s_n}) \mapsto y \in \{pos, neg\} \tag{1}$$

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 272–277, 2013.

2 Related Work

Research in Opinion Mining is far-reaching [7], however the most techniques tackle this problem in the domain of customer reviews [7]. Many approaches for Opinion Mining in reviews collect sentiment-bearing words [6]. There are methods [4] which try to handle linguistic or contextual sentiment such as negations. The negation as the maybe most important linguistic factor is often treated by heuristic rules [4], which reverse the polarity of sentiment words. Interesting techniques for the effects of negations have been introduced by Jia et al. [5]. Here, the scope of negations are derived from different rules. In addition, we are interested in a linguistic and grammatical context as in Zhou et al. [13]. They show that conjunctions can be used to avoid ambiguities within sentences. In the news domain, many approaches on this topic only work with reported speech objects [1]. News articles are less subjective [1], but quotations in newspaper articles are often the place where more subjective text and opinions can be found [1]. However, only opinions, which are part of a reported speech object, can be analysed by this method. An analysis [9] shows that in a MRA less than 22% of the opinion-bearing text contain quoted text and only in less than 5% the area of quoted text is larger than 50% of the whole relevant opinion.

3 Determination of Sentiment Polarity

Our approach calculates four basic sentiment features (**Basic Sentiment Features** α) first. These features are based on the four word categories adverbs, adjectives, nouns, and verbs, which are the most important word classes for the polarity of sentiment [8]. We use existing methods such as chi-square [6], the PMI-method [3,6], the entropy-based method [11], the method of information gain [11], and the German sentiment lexicon SentiWS [8] for the weighting of the polarity (our sentiment score σ). We compute four sentiment features for one statement (**Basic Sentiment Features** α). Every feature is the average of the sentiment scores in one category: The first feature is the average of the scores of all the statement's adjectives ($f_{\alpha_1}(s) = \sigma_{Adj}(s)$), the second of all nouns ($f_{\alpha_2}(s) = \sigma_{No}(s)$), and so on.

$$\sigma_{cat}(s) = \frac{1}{|s_{cat}|} \sum_{w \in s_{cat}} \sigma_{method}(w) \tag{2}$$

Here, s_{cat} are only the words in statement s which belong to one of the four important categories (adjectives, nouns, verbs, and adverbs) and σ_{method} is one of the five word based methods.

4 Linguistic and Contextual Features

4.1 Two Techniques for the Effect Measurement

The first technique only measures, whether or not the linguistic effects are present in a given statement and stores it as one feature for every aspect

(**Linguistic Effect Features** β). The second technique tries to capture an area of this effect and it takes the sentiment of the area as the feature value of this aspect (resulting in **Linguistically Influenced Sentiment Features** γ). The feature value is the sum of the sentiment of the influenced words. We implement techniques from Jia et al. [5], who are trying to capture different effect areas for negations. We adapt their *candidate scope* [5] and *delimiter rules* [5] using static and dynamic delimiters for the German language and expand them also for our non negation features: The static *delimiters* [5] remove themselves and all words after them from the scope. Static *delimiters* are words such as "because", "when" or "hence" [5]. A *conditional delimiter* [5] becomes a delimiter if it has the correct POS-tag, is inside a negation scope, and leads to opinion-bearing words. Examples are words such as "who", "where" or "like". In addition, we have designed a second method which creates a scope around an effect word. All words in the scope have a smaller distance to all other effect words (in number of words between them).

4.2 Calculation of the Features

The sentiment of words can change depending on whether the statements concern persons or organisations. So, the first two features represesent the proportion of persons and organisations: In equation 3 for the first two β features, $p(s)$ and $o(s)$ are the number of persons and organisations, respectively, in the statement s. For the two type γ features, P_w and O_w are the set of words which belongs to persons' and organisations' scope (second method, cf. section 4.1), respectively.

$$f_{\beta_1}(s) = \frac{p(s)}{p(s) + o(s)} \qquad f_{\beta_2}(s) = \frac{o(s)}{p(s) + o(s)} \tag{3}$$

$$f_{\gamma_1}(s) = \sum_{w \in P_w} \sigma(w) \qquad f_{\gamma_2}(s) = \sum_{w \in O_w} \sigma(w) \tag{4}$$

The negation feature shows, whenever a negation is present in statement s. N_w are the affected words. At this point, the area of affected words is determined by the *candidate scope* [5] and *delimiter rules* [5].

$$f_{\beta_3}(s) = \begin{cases} 1.0 \text{ if } \exists w \in s : w \text{ is a negation} \\ 0.0 \text{ otherwise} \end{cases} \qquad f_{\gamma_3}(s) = \sum_{w \in N_w} \sigma(w) \tag{5}$$

The use of conjunctions can also indicate a polarity. We create a test data of 1,600 statements, collect the conjunctions and associate them with a sentiment value ν_c by their appearance in positive and negative statements. Table 1 (left) shows the different conjunctions and their value to influence the sentiment. The type β feature for conjunctions is the sum of all sentiment values ν_c of all conjunctions C_s of the statement s. The conjunction influenced words are C_w. The scope is

Table 1. Left: Conjunctions and sentiment value. Right: Hedging auxiliary verbs.

word	ν_c	word	ν_c	word	ν_c	word	ν_c
whereas	-0.5	as well	1.0	but	-1.0	or	0.5
however	-0.5	though	-1.0	and	1.0	by	1.0

can	may	could	might
would	shall	should	ought to
will	must		

determined by the *candidate scope* [5] and *delimiter rules* [5], but only words after the conjunction are concerned because the conjunction itself is a delimiter. The multiplication with ν_c indicates which type of conjunction influences the affected words. If the conjunction expresses a contrast (e.g. "but" with $\nu_c = -1.0$), the sentiment of the words will be inverted.

$$f_{\beta_4}(s) = \frac{\sum_{c \in C_s} \nu_c}{|C_s|} \qquad f_{\gamma_4}(s) = \sum_{w \in C_w} \nu_c * \sigma(w) \tag{6}$$

A short part of quoted text can be a hint for irony in written texts [2] and a long part can stand for a reported speech object. As a result, a machine learning approach can better differentiate between irony and reported statements, if the length and the affected words of quoted text are measured. $q(s)$ is the part of a statement s, which appears in quotation marks. $l(x)$ is the length (in characters) of a text x. Q_w are the words inside a quotation.

$$f_{\beta_5}(s) = \frac{l(q(s))}{l(s)} \qquad f_{\gamma_5}(s) = \sum_{w \in Q_w} \sigma(w) \tag{7}$$

Modal verbs like "can" or "would" can weaken the strength of the polarity. The full list of auxiliary verbs for hedging expressions is shown in table 1 (right). The method counts how often full verbs are influenced by hedging expressions $h(s)$ in comparison to all full verbs $v(s)$. H_w is the set of words affected by hedging. Here again, the *candidate scope* [5] and *delimiter rules* [5] are used.

$$f_{\beta_6}(s) = \frac{h(s)}{v(s)} \qquad f_{\gamma_6}(s) = \sum_{w \in H_w} \sigma(w) \tag{8}$$

4.3 Machine Learning Technique for Sentiment Classification

For the classification, we use a SVM (Rapidminer[1] standard implementation). The SVM receives the feature sets β and γ as input values for learning, as well as it obtains the **Basic Sentiment Features** α. In this way, our machine learning approach is able to learn from the sentiment features and the linguistic features.

[1] Rapid-I: `http://rapid-i.com/`

5 Evaluation

We evaluate our approach on two different datasets: The first corpus, called **Finance**, represents a real MRA about a financial service provider. It contains 5,500 statements (2,750 are positive, 2,750 are negative) from 3,452 different news articles. The second dataset is the **pressrelations** dataset [10]. We use approx. 30% of the dataset to construct a sentiment dictionary. This means that 1,600 statements (800 are positive, 800 are negative) are used for Finance and 308 statements for the pressrelations dataset. The sentiment dictionaries contain words which are weighted by the methods explained in section 3. We use 20% of the remaining set to train a classification model. The results are depicted in table 2 and show that the features β and γ improve sentiment allocation. The features increased performance of all methods, except the information gain method on pressrelations. However, in all other cases, the methods achieved the best results by using all features. SentiWS, as the dictionary based approach, got the highest improvement (over 7% on finance and over 14% on pressrelations). The entropy-based method with all features got the highest accuracy with 75.28% on Finance, which is an improvement of over 5% to the baseline.

Table 2. Results of the linguistic features

	Finance dataset				pressrelations dataset			
Method	α	$\alpha+\beta$	$\alpha+\gamma$	all	α	$\alpha+\beta$	$\alpha+\gamma$	all
SentiWS	0.6036	0.6590	0.6311	**0.6792**	0.5526	0.5604	0.615	**0.6943**
PMI	0.6174	0.6586	0.6317	**0.6881**	0.6245	0.6057	0.634	**0.6887**
χ^2	0.6872	0.7071	0.6981	**0.7234**	0.6453	0.6453	0.6717	**0.6868**
Entropy	0.7006	0.7221	0.7428	**0.7528**	0.6642	0.6604	0.6774	**0.6943**
Information Gain	0.6955	0.7186	0.7243	**0.7349**	**0.6912**	0.6761	0.6811	0.6828

By comparing all results, the influence of feature set β seems to be bigger than the influence of feature set γ on Finance, while it is the other way around on the pressrelations dataset. The reason for this is the nature of the two domains. The political texts are more complicated so that a deeper analysis, which exploits values of the influenced sentiment-bearing words, provides more benefit. Nevertheless, except the for information gain method, the combination of all linguistic features achieved an increase to the baselines of at least over 3%.

6 Conclusion

In conclusion, linguistic features are very useful for Opinion Mining in newspaper articles. The evaluation shows that the linguistic features can be integrated into existing solutions and thereby improve the computation of sentiment. The improvement is especially large and therefore interesting for dictionary based approaches. Moreover, this approach achieved high accuracies of over 70% and in one case an accuracy of over 75%.

Acknowledgments. This work is funded by the German Federal Ministry of Economics and Technology under the ZIM-program (Grant No. KF2846501ED1).

References

1. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment analysis in the news. In: Proc. of the 7th Intl. Conf. on Language Resources and Evaluation, LREC 2010 (2010)
2. Carvalho, P., Sarmento, L., Silva, M.J., de Oliveira, E.: Clues for detecting irony in user-generated contents: oh..!! it's "so easy";-). In: Proc. of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, TSA 2009, pp. 53–56 (2009)
3. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. In: Proc. of the 27th Annual Meeting on Association for Computational Linguistics, ACL 1989, pp. 76–83 (1989)
4. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proc. of the Intl. Conf. on Web Search and Web Data Mining, WSDM 2008, pp. 231–240 (2008)
5. Jia, L., Yu, C., Meng, W.: The effect of negation on sentiment analysis and retrieval effectiveness. In: Proc. of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1827–1830 (2009)
6. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of html documents. In: Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL (2007)
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
8. Remus, R., Quasthoff, U., Heyer, G.: SentiWS – a publicly available german-language resource for sentiment analysis. In: Proc. of the 7th Intl. Conf. on Language Resources and Evaluation, LREC 2010 (2010)
9. Scholz, T., Conrad, S.: Integrating viewpoints into newspaper opinion mining for a media response analysis. In: Proc. of the 11th Conf. on Natural Language Processing, KONVENS 2012 (2012)
10. Scholz, T., Conrad, S., Hillekamps, L.: Opinion mining on a german corpus of a media response analysis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 39–46. Springer, Heidelberg (2012)
11. Scholz, T., Conrad, S., Wolters, I.: Comparing different methods for opinion mining in newspaper articles. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (eds.) NLDB 2012. LNCS, vol. 7337, pp. 259–264. Springer, Heidelberg (2012)
12. Watson, T., Noble, P.: Evaluating public relations: a best practice guide to public relations planning, research & evaluation. PR in practice series, ch. 6, pp. 107–138. Kogan Page (2007)
13. Zhou, L., Li, B., Gao, W., Wei, Z., Wong, K.-F.: Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 162–171 (2011)