Language-Independent Sentiment Analysis with Surrounding Context Extension

Tomáš Kincl[1(✉)], Michal Novák[1], Jiří Přibil[1], and Pavel Štrach[2]

[1] University of Economics, Prague, Czech Republic
{tomas.kincl,michal.novak,jiri.pribil}@vse.cz
[2] Upper Austria University of Applied Sciences, Steyr, Austria
pavel.strach@fh-steyr.at

Abstract. Expressing attitudes and opinions towards various entities (i.e. products, companies, people and events) has become pervasive with the recent proliferation of social media. Monitoring of what customers think is a key task for marketing research and opinion surveys, while measuring customers' preferences or media monitoring have become a fundamental part of corporate activities. Most experiments on automated sentiment analysis focus on major languages (English, but also Chinese); minor or morphologically rich languages are addressed rather sparsely. Moreover, to improve the performance of machine-learning based classifiers, the models are often complemented with language-dependent components (i.e. sentiment lexicons). Such combined approaches provide a high level of accuracy but are limited to a single language or a single thematic domain.

This paper aims to contribute to this field and introduces an experiment utilizing a language– and domain– independent model for sentiment analysis. The model has been previously tested on multiple corpora, providing a trade-off between generality and the classification performance of the model. In this paper, we suggest a further extension of the model utilizing the surrounding context of the classified documents.

Keywords: Sentiment analysis · Cross-domain · Cross-language · Document surrounding context

1 Introduction

Expressing attitudes and opinions towards various entities such as products, companies, people and events has become an instant phenomenon with the proliferation of social media. Nowadays, the monitoring of what customers think is a key task of marketing research. Opinion surveys, measuring customers' preferences or media monitoring have become a fundamental part of corporate activities [1]. Recognizing opinion polarity and finding out about people's attitudes has become a challenge, which is addressed by (automated) sentiment analysis [2]. The literature highlights two main approaches to this issue [3]. The first approach is based on utilizing a dictionary of words (opinion lexicon) to recognize the sentiment polarity (lexicon-based approaches [4]). The second group of approaches to sentiment analysis is based on (supervised) machine learning [5]. Such methods usually require labeled training set to build the classifier [6].

G. Meiselwitz (Ed.): SCSM 2015, LNCS 9182, pp. 158–168, 2015.
DOI: 10.1007/978-3-319-20367-6_17

Most experiments on automated sentiment analysis focus on major languages (English, but also Chinese); minor or morphologically rich languages are rarely addressed [7]. Moreover, to improve the performance of machine-learning based classifiers, the models are often complemented with language-dependent components (i.e. sentiment lexicons). Such combined approaches provide a high level of accuracy but are also limited to a single language or a single thematic domain.

This paper aims to contribute in this field and introduces an experiment utilizing a language– and domain– independent model for sentiment analysis. The model has been previously tested on multiple corpora [8], providing a trade-off between generality and the classification performance of the model. In this paper, we suggest a further extension of the model utilizing the surrounding context of the classified documents.

2 The Surrounding Context Correction

From the marketing communications perspective, negative comments are more important than positive ones. This is not to say that the positive remarks from customers are not important—i.e. when spread by word-of-mouth to encourage potential customers—although these do not usually require any immediate (re-)action from the company, as is in the case of negative comments. Moreover, customers tend to share a negative experience which often initiates further activity of other disappointed or frustrated users, thereby creating a snowball effect. Negative emotions not only strengthen and prolong discussions but can also influence or even disrupt communication between community members and/or a company [9]. Therefore discovering a negative comment or an expression of negative emotions could enable a faster and more appropriate reaction by the company to a potentially emerging problem.

Common approaches to the sentiment analysis problem usually classify the comments (or documents in general) separately [2]. However, incorporating the sentiment of surrounding comments could provide additional information to further improve classification accuracy. When the classifier cannot detect the sentiment of the analyzed comment with high confidence, it could look around for the prevailing sentiment of surrounding comments. If there are more surrounding negative comments, this could indicate that the sentiment of the analyzed comment is also rather negative [10]. The surrounding context could be either local (i.e. the comments are in the same discussion or thread) or chronological (i.e. the comments addressing a similar issue appear in the same time frame but on different locations).

Therefore the model includes a correction to further improve the accuracy of the analyzed comments. For each classified comment, the classifier computes two values c_{neg}, $c_{pos} \in [0, 1]$; $c_{neg} = 1 - c_{pos}$ which express the confidence that the classified comment belongs to the given class of sentiment [11]. If the confidence of the classifier is high (c_{neg}, c_{pos} values are far from 0.5), the correction plays only a marginal to no role in the classification. However, when the classifier has low confidence in the sentiment of the analyzed comment (c_{neg}, c_{pos} values are very close to 0.5), the correction is applied. The extent of the correction applied according to the c_{neg} value is displayed in Fig. 1.

If the c_{neg} value is close to the value 0.5 (the level of the classifier's confidence in assigning a comment to a positive/negative class is low), then the correction (*corr*) is applied to the full extent. However with an increasing distance from $c_{neg} = 0.5$, the amount of correction applied declines rapidly. The z parameter influences the extent of which is the correction applied with the increasing distance from the point where $c_{neg} = 0.5$. In the current setting described in Fig. 1, the correction doesn't play any important role for $c_{neg} \notin (0.4, 0.6)$. The z parameter has been experimentally set to $z = 500$. However, such an approach is only heuristic and the value can be set up differently according to the classification task.

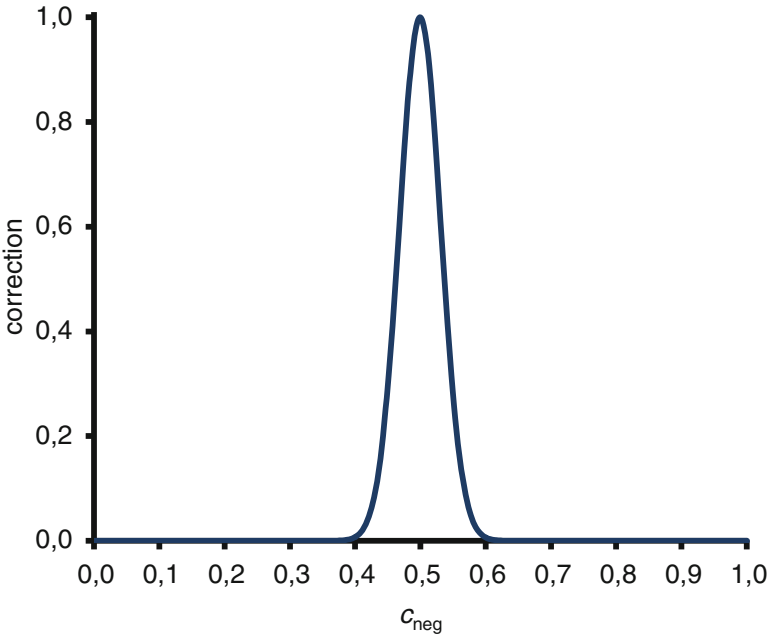

Fig. 1. The extent of the correction applied according to the c_{neg} value

The correction (*corr*) can be expressed as in the Eq. (1):

$$corr = e^{-z \cdot (c_{neg} - 0{,}5)^2} \tag{1}$$

The correction mechanism has been designed to take into account two various effects. The first effect (E_1) represents the predominant sentiment of related comments published from the when the topic first appeared. The related comments can be represented by a discussion in one thread in a discussion group or on a social network, or by comments published on various sites but still discussing the same topic (identified i.e. by topic keywords).

The second effect (E_2) includes the current trend of sentiment polarity. This exclusively takes into account the sentiment of recently published comments. The number of comments representing the trend is $n - k$ and can be set up differently according to the classification task. Again, the comments can be represented either by a discussion in one thread or by comments addressing the same topic. Both effects are demonstrated in Fig. 2.

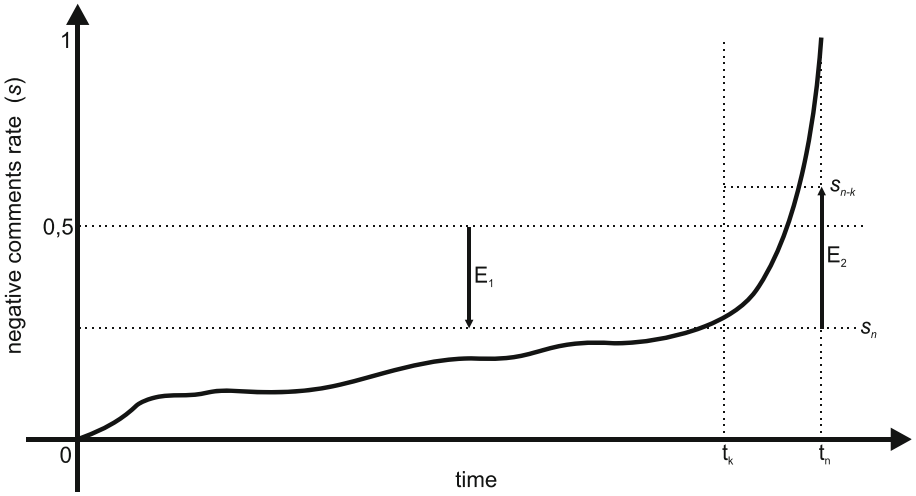

Fig. 2. An example of both sentiment correction components E_1 and E_2

The horizontal axis represents the time sequence of the comments published. The vertical axis represents the relative share of comments (s) that were assigned to the negative class by the classifier. Figure 2 illustrates an example where the positive comments are prevailing at the beginning of the discussion (the share of negative comments is less than 0.5). Subsequently, the overall mood in the discussion shifts and many negative comments are made (comments k to n). The task is to classify the $n + 1$ comment.

First, the classifier evaluates the $n + 1$ comment regardless of the surrounding context. If the c_{neg} is far from the 0.5 value (the classifier has high confidence in the polarity of the analyzed comment), the correction doesn't apply. However, if the c_{neg} is close to 0.5, the correction (including effects E_1 and E_2) is applied to the extent according to Eq. (1).

The first E_1 (prevailing surrounding sentiment) effect is included in the model as expressed in the Eq. (2):

$$E_1 = s_n - 0,5 \tag{2}$$

The first E_1 effect is the relative share of negative comments (s_n) from the beginning of the discussion until now, minus 0.5. For $E_1 < 0$, the predominant sentiment of analyzed

comments is positive (the relative share of negative comments in the discussion is less than 0.5) and the $E1$ effect will be reflected in lowering the c_{neg}. For $E_1 > 0$, the corrected c_{neg} will be higher, since the negative sentiment is prevailing (the relative share of negative comments s_{n} is more than 0.5).

The second effect compares the relative share of negative comments s_{n-k} in last $n - k$ cases according to the cumulative share of negative comments for all analyzed cases s_{n}. The E_2 effect is included in the model as expressed in the Eq. (3):

$$E_2 = \frac{\frac{n \cdot s_n - k \cdot s_k}{n-k}}{s_n} - 1 \tag{3}$$

For $E_2 > 0$ is the relative share of negative comments s_{n-k} for last $n - k$ cases is higher than for all analyzed comments s_{n}. Therefore it could be expected that the share of negative comments will continue to increase in the near future. The effect will contribute to the increase of c_{neg} for the classified comment. For $E_2 < 0$, the corrected c_{neg} will be lower since the relative share of negative comments is lower for the last $n - k$ cases. The E_2 effect could even rise beyond value 1 (in case that a recent rise in negative comments is highly significant). If so, the effect value counts as 1.

The correction mechanism also includes weights for both effects. By default, the weights were chosen identically for both effects as $w_1 = w_2 = 0.5$. The corrected c_{neg} – $ccor_{\text{neg}}$ – can therefore be expressed as in the Eq. (4).

$$ccor_{\text{neg}} = c_{\text{neg}} + corr \cdot (w_1 \cdot E_1 + w_2 \cdot E_2) \tag{4}$$

3 Results

Several datasets were used to validate the model. The first corpus data were obtained from the Czech-Slovak Film Database available at http://www.csfd.cz. CSFD is the largest community-driven online database of reviews and information in the Czech (Slovak) language related to films, television programs, including cast, production crew and biographies. CSFD is often considered as a local alternative to the Internet Movie Database (IMDb). This dataset has been chosen to compare the results on local language with studies conducted on similar (i.e. IMDb) corpora. The second dataset was obtained from mall.cz, the second largest [12] Czech e-shop. Mall.cz offers various products ranging from electronics and home appliances to sports gear and supplies for hobbies and pets. This dataset has been chosen to validate how the model deals with cross-domain sentiment classification [2]. The third dataset was the Large Movie Review Dataset [13] based on the IMDb data (movie reviews). This dataset has been chosen not only to compare the results on CSFD corpus, but also to compare the model performance with other studies (since the corpus is publicly available and often utilized as a sentiment analysis dataset).

The last dataset contained reader reviews retrieved from Amazon websites in multiple languages. All reviews related to the 2012 bestselling book, Fifty Shades of Grey by E.L. James (all versions – hardcover, paperback, kindle or audio version) were obtained.

The languages included English (amazon.co.uk; 7,255 reviews), German (amazon.de; 4,154 reviews) and French (amazon.fr; 1,258 reviews). The dataset has been chosen to test how the model deals with cross-language sentiment classification [2].

3.1 Czech-Slovak Film Database Results

The training set contained 3,000 positive (4 - and 5-star reviews) and 3,000 negative comments (0-, 1- and 2-star reviews). Then the trained model was validated on a randomly selected set of 120,000 comments (20,000 from each star-rating). The results of the classification (without the correction applied) are summarized in Table 1.

Table 1. Classification results on CSFD validation set

Star rating	Predicted negative	Predicted positive
0	89.57 %	10.44 %
1	86.17 %	13.83 %
2	74.60 %	25.40 %
3	46.44 %	53.56 %
4	21.98 %	78.03 %
5	15.73 %	84.28 %

The classification accuracy without the correction mechanism applied (computed on 0-, 1-, 2-, 4-, and 5-star comments since 3-star comments were considered as neutral and therefore without any sentiment) reached 82.53 %. Subsequently, the correction including the surrounding context has been applied. The context to the CSDF data was considered as the previously published comments to the same movie. There was more than one comment about 12,000 movies in the test dataset. The highest number of comments about one single movie was 175. The dataset contained various movies, from The Shawshank Redemption (2^{nd} most popular movie on CSFD) to Playgirls (considered as the 14^{th} worst movie of all time). The classification results with the suggested correction applied are summarized in Table 2.

Table 2. Classification results on CSFD validation set (correction applied)

Star rating	Predicted negative	Predicted positive
0	91.58 %	8.42 %
1	87.87 %	12.14 %
2	76.22 %	23.78 %
3	47.31 %	52.69 %
4	21.87 %	78.14 %
5	15.11 %	84.89 %

The classification accuracy with the correction mechanism applied (computed on 0-, 1-, 2-, 4-, and 5-star comments since 3-star comments were considered as neutral and therefore without any sentiment) reached 83.74 %. Therefore the correction applied improves the classification accuracy by 1.21 %.

3.2 Large Movie Review Dataset (IMDb) Results

The IMDb allows users to review the movies on a scale ranging from 1- to 10-stars. The training set from [13] contained 12,500 positive (7- to 10-star reviews) and 12,500 negative comments (1- to 4-star reviews). The validation set contained the same amount of comments – 12,500 positive and 12,500 negative reviews. The results of the classification (without the correction applied) are summarized in Table 3.

Table 3. Classification results on IMDb validation set

Star rating	Predicted negative	Predicted positive
1	92.49 %	7.51 %
2	88.79 %	11.21 %
3	82.88 %	17.12 %
4	74.80 %	25.20 %
5	23.54 %	76.46 %
6	13.96 %	86.04 %
7	11.31 %	88.69 %
8	8.98 %	91.02 %
9	92.49 %	7.51 %
10	88.79 %	11.21 %

The classification accuracy without the correction mechanism applied (computed on 1-to 4- and 7- to 10-star comments) reached 86.44 %. Subsequently, the correction including the surrounding context has been applied. Similarly to the CSFD data, the comments to the same movie were considered as a context to apply the correction. The classification results with the suggested correction applied are summarized in Table 4.

Table 4. Classification results on IMDb validation set (correction applied)

Star rating	Predicted negative	Predicted positive
1	93.65 %	6.35 %
2	90.83 %	9.17 %
3	84.65 %	15.35 %
4	77.00 %	23.00 %
5	21.20 %	78.80 %
6	12.32 %	87.68 %
7	9.26 %	90.74 %
8	8.04 %	91.96 %
9	93.65 %	6.35 %
10	90.83 %	9.17 %

The classification accuracy with the correction mechanism applied reached 87.88 %. Therefore, the correction applied improves the classification accuracy by 1.44 %. Studies conducted on the same dataset reached similar performance (i.e. 88.33 % in [13]). However the model suggested in our experiment does not utilize any language- or domain-dependent components (i.e. sentiment vocabularies or ontologies) and therefore could be used to address cross-language or cross-domain sentiment classification.

3.3 Mall.Cz Results

The mall.cz dataset has been analyzed to validate the model on data from multiple domains (product reviews from various product categories). The training set contained 3,000 positive (4- and 5-star reviews) and 3,000 negative comments (1- and 2-star reviews). Then the trained model was validated on a randomly selected set of 100,000 comments (20,000 from each star-rating). The results of the classification (without the correction applied) are summarized in Table 5.

Table 5. Classification results on mall.cz validation set

Star rating	Predicted negative	Predicted positive
1	81.38 %	18.62 %
2	73.67 %	26.33 %
3	57.09 %	42.91 %
4	32.58 %	67.42 %
5	20.50 %	79.50 %

The classification accuracy without the correction mechanism applied (computed on 1, 2-, 4- and 5-star comments) reached 76.71 %. Subsequently, the correction including the surrounding context has been applied. As a context, reviews previously commenting on the same product were considered. The classification results with the suggested correction applied are summarized in Table 6.

Table 6. Classification results on mall.cz validation set (correction applied)

Star rating	Predicted negative	Predicted positive
1	82.89 %	17.11 %
2	75.51 %	24.49 %
3	50.81 %	49.19 %
4	27.42 %	72.58 %
5	17.04 %	82.96 %

The classification accuracy with the correction mechanism applied reached 78.95 %. Therefore, the correction applied improves the classification accuracy by 2.24 %. The performance of the model on mall.cz dataset did not reach such values as

in the previous cases. However, the model still performed on a satisfactory level considering that the data came from multiple domains, Czech is a morphologically rich language and the model does not utilize any language-dependent component to improve the classification.

3.4 Amazon Results

For the experiment on Amazon data, 100 positive (4- and 5-star reviews) and 100 negative (1- and 2-star reviews) comments from each language were used to train the model. Even though the amount of data obtained from the Amazon websites was much higher, there were only a limited number of French negative comments collected. This was also the reason why a separate training and validation set could not be used. In this case, the model utilized the 10-cross fold validation. All three languages were merged together into one and treated as one single dataset. The aim of this experiment was to discover how the model deals with cross-language sentiment classification. The results of the classification (without the correction applied) are summarized in Table 7.

Table 7. Classification results on Amazon multilingual set

Star rating	Predicted negative	Predicted positive
1	89.00 %	11.00 %
2	88.00 %	12.00 %
3	57.33 %	42.67 %
4	19.00 %	81.00 %
5	7.67 %	92.33 %

The classification accuracy without the correction mechanism applied (computed on 1, 2-, 4- and 5-star comments) reached 87.58 %. Subsequently, the correction including the surrounding context has been applied. The classification results with the suggested correction applied are summarized in Table 8.

Table 8. Classification results on Amazon multilingual set (correction applied)

Star rating	Predicted negative	Predicted positive
1	90.33 %	9.67 %
2	89.67 %	10.33 %
3	59.00 %	41.00 %
4	19.67 %	80.33 %
5	8.00 %	92.00 %

The classification accuracy with the correction mechanism applied reached 88.08 %. Therefore the correction applied improves the classification accuracy by 0.50 %. Even though the model does not utilize any language-dependent components,

it performs very well on multilingual data. Moreover, the correction we suggest was able to further improve the classification accuracy of the model.

4 Discussion and Conclusion

The results indicate that the performance of the sentiment classification model could be improved even without utilizing language-dependent components. The paper suggests a correction based on incorporating the surrounding context of the analyzed document. This context is either local (i.e. the comments are in the same discussion or thread) or chronological (i.e. the comments addressing a similar issue appear in the same time frame but on different locations). The extended model has been tested on several corpora to reveal how the suggested approach deals with the usual sentiment analysis problems (i.e. cross-domain, cross-language sentiment analysis or sentiment analysis on morphologically rich or minor languages). The results of the experiments are encouraging and summarized in Table 9.

Table 9. Results summary

Corpus	Corpus characteristics	Model performance
Mall.cz	Czech (morphologically rich) language, multiple thematic domains (product reviews in various categories).	76.71 % correctly classified cases without the correction applied. With the knowledge of the surrounding context, the model reached 78.95 %. 2.24 % overall improvement
CSFD	Czech (morphologically rich) language. One thematic domain (movie reviews)	82.53 % correctly classified cases without the correction applied. With the knowledge of the surrounding context, the model reached 83.74 %. 1.21 % overall improvement.
IMDb	English language. One thematic domain (movie reviews).	86.44 % correctly classified cases without the correction applied. With the knowledge of the surrounding context, the model reached 88.33 %. 1.44 % overall improvement.
Amazon	Multiple (English, German, French) languages. One thematic domain (book reviews).	87.58 % correctly classified cases without the correction applied. With the knowledge of the surrounding context, the model reached 88.08 %. 0.55 % overall improvement.

The results suggest that the model performs well on multiple languages even though it does not utilize any language-dependent component (for a further description of the model please see [8]). It performs well on major languages (English, German or French) as well as on a morphologically rich language (Czech). The performance of the model could be easily improved by including the surrounding context of the analyzed

document. The model also achieves comparable results with studies conducted on similar datasets [13].

It may represent an opportunity for a subsequent research inquiry to address the limitation of the effect of the previously reached performance and on the nature of the analyzed data on the magnitude of correction outcomes. The higher the previously reached performance, the lesser the correction contributes to further classification improvement.

References

1. CyberAlert. http://www.cyberalert.com/downloads/media_monitoring_whitepaper.pdf
2. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**, 1–167 (2012)
3. Balahur, A.: Sentiment analysis in social media texts. In: WASSA 2013, p. 120 (2013)
4. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**, 82–89 (2013)
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
6. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC, pp. 1320-1326 (2010)
7. Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., Tounsi, L.: Statistical parsing of morphologically rich languages (SPMRL): what, how and whither. In: Proceedings of the First Workshop on Statistical Parsing of Morphologically-Rich Languages, NAACL HLT 2010, pp. 1–12. Association for Computational Linguistics (2010)
8. Kincl, T., Novák, M., Přibil, J.: Getting inside the minds of the customers: automated sentiment analysis. In: European Conference on Management Leadership and Governance ECMLG 2013, pp. 122–129. Alpen-Adria Universität Klagenfurt, Austria (2013)
9. Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., Hołyst, J.A.: Collective emotions online and their influence on community life. PLoS ONE **6**, e22207 (2011)
10. Brychcín, T., Habernal, I.: Unsupervised improving of sentiment analysis using global target context. In: International Conference Recent Advances in Natural Language Processing (RANLP 2013) (2013)
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intel. Syst. Technol. (TIST) **2**, 1–39 (2011)
12. http://byznys.ihned.cz/c1–54991650-cesko-je-e-shopovou-velmoci-internetove-obchody-vygeneruji-37-miliard-obratu
13. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)