The final publication is available at Springer
via http://link.springer.com/chapter/10.1007/978-3-319-22593-7_19

SENTIMENT CLASSIFICATION IN MULTIPLE LANGUAGES – FIFTY SHADES OF CUSTOMER OPINIONS

TOMÁŠ KINCL
Department of Exact Methods
Faculty of management, University of Economics, Prague
Czech Republic
kincl.tomas@gmail.com

MICHAL NOVÁK
Department of Exact Methods
Faculty of management, University of Economics, Prague
Czech Republic
novak@michalnovak.eu

JIŘÍ PŘIBIL
Department of Exact Methods
Faculty of management, University of Economics, Prague
Czech Republic
jiri@pribil.cz

Abstract: Sentiment analysis is a natural language processing task where the goal is to classify the sentiment polarity of the expressed opinions, although the aim to achieve the highest accuracy in sentiment classification for one particular language, does not truly reflect the needs of business. Sentiment analysis is often used by multinational companies operating on multiple markets. Such companies are interested in consumer opinions about their products and services in different countries (thus in different languages). However, most of the research in multi-language sentiment classification simply utilizes automated translation from minor languages to English (and then conducting sentiment analysis for English). This paper aims to contribute to the multi-language sentiment classification problem and proposes a language independent approach which could provide a good level of classification accuracy in multiple languages without using automated translations or language-dependent components (i.e. lexicons). The results indicate that the proposed approach could provide a high level of sentiment classification accuracy, even for multiple languages and without the language dependent components.

Keywords: SENTIMENT ANALYSIS, LANGUAGE-INDEPENDENT, CUSTOMER REVIEWS, OPINION MINING, MARKETING

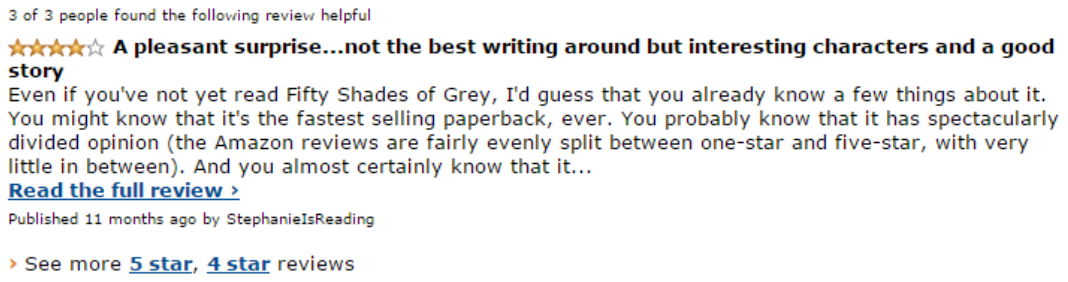
1 Introduction

Sentiment analysis or opinion mining seems to be a hot and recent topic related to the emergence of digital media and communication technologies although people have always listened to their relatives and people they respect (Anderson, 1998, Goldenberg *et al.*, 2001). Monitoring what customers think about companies and their products is a key marketing research task. Opinion surveys, measuring customers' preferences or media monitoring have become an integral part of corporate activities (Comcowich, 2010).
However, media monitoring has changed rapidly with the emergence of new technologies. The content is digitalized and available online; linguistic and geographic barriers no longer exist. Moreover, the online environment is highly fragmented and new information resources are continuously emerging and vanishing. Almost anyone can become an influencer or opinion leader – many of today's most influential blogs (i.e. The Huffington Post, Techcrunch, Engadget, Mashable) began a few years ago as a personal website, quickly growing into widely recognized and respected media (Aldred *et al.*, 2008). More than 80 % of US internet users

have previously conducted online research about a product. More than 75% of online-hooked customers admit that reviews significantly influence their purchase intentions and that they are willing to pay more for a product with better customer reviews (Horrigan, 2008).
Recognizing sentiment and determining people's attitudes becomes a challenge in such a highly fragmented and chaotic environment. Nevertheless, computer-based processing and modeling allows for automated sentiment analysis (Liu, 2012). The goal of sentiment analysis is to (automatically) identify and extract subjective information, generally to classify the polarity (usually positive or negative) of expressed opinion (Pang and Lee, 2008).The analysis can be made of the document, sentence, or feature/aspect level and there are two major approaches to sentiment analysis (Zhang *et al.*, 2011). The lexicon-based approaches usually utilize prepared dictionaries of sentiment words and phrases with associated orientations and strength (Taboada *et al.*, 2011). The second group is based on (supervised) machine learning. Such methods usually require labeled training set to build the classifier (Pak and Paroubek, 2010). However most experiments focus on major languages (English, but also Asian languages) while minor or morphologically rich languages are rarely addressed (Tsarfaty *et al.*, 2010).
Only a few studies have addressed cross-domain or cross-language (sub)tasks, where the data comes from multiple thematic domains or various languages (Liu, 2012). We believe that domain or language dependent models do not truly reflect the business needs. The companies will seldom use a specific tool or model for each market or country they operate in. The aim of this paper is to contribute to this field and develop a language-independent model for sentiment classification, which would provide a good performance (classifier accuracy) among multiple languages. Such a task might be approached as a domain adaptation problem (Cambria *et al.*, 2013). Even though it is widely accepted that a classifier trained on the set from one domain often performs poorly on data from another domain (Liu, 2012), there are studies suggesting that a simple "all-in-one" classifier outperforms models utilizing training sets solely from a single domain (Mansour *et al.*, 2013) . Our suggested approach analyzes a model trained on a multiple-language set.

2 Methodology

The dataset contains reader reviews retrieved from Amazon websites in July 2014. All reviews related to the 2012 bestselling book, Fifty Shades of Grey by E.L. James (all versions – hardcover, paperback, kindle edition or audio version – included) were downloaded. The dataset contains readers' comments in three languages – English (amazon.co.uk; 7,255 reviews at the time of data retrieval), German (amazon.de; 4,154 reviews at the time of data retrieval) and French (amazon.fr; 1,258 reviews at the time of data retrieval). Each review has been saved as a single text file. An example of such reader review is on the Figure 1.


3 of 3 people found the following review helpful

A pleasant surprise...not the best writing around but interesting characters and a good story

Even if you've not yet read Fifty Shades of Grey, I'd guess that you already know a few things about it. You might know that it's the fastest selling paperback, ever. You probably know that it has spectacularly divided opinion (the Amazon reviews are fairly evenly split between one-star and five-star, with very little in between). And you almost certainly know that it...

Read the full review ›

Published 11 months ago by StephanieIsReading

› See more 5 star, 4 star reviews

Figure 1 An Example of Positive and Negative Reader Review at Amazon.co.uk
Source: http://www.amazon.co.uk/product-reviews/B007L3BMGA/

The number of stars assigned by the reader has also been saved to provide training data for the model. The 1- and 2-star reviews were considered as negative; the 4- 5-star reviews as positive. The 3-star reviews were regarded as neutral and thus omitted from the training set. The length of reviews obtained varies from several words (~3) up to highly comprehensive and detailed reviews (almost 1.500 words).

RapidMiner software (http://rapid-i.com/) was used for the analysis and the preprocessing procedure included case transformation (all lowercase), tokenization (splitting the document into a sequence of tokens – words) and generating character n-grams. The difference between word and character n-grams is depicted by the following example (for n=3).The sequence “to be or not to be” is represented as the following word 3-grams: to be or, be or not, or not to, not to be. Similarly for character 3-grams: to_, o_b, _be, be_, e_o, _or, or_, r_n, _no, not, ot_, t_t, _to, to_, o_b, _be (spaces are visualized as underlines). Such representation offers several advantages over the word based n-grams.

Social network posts and comments often contain informal expressions, typos, errors, emoticons or other unknown or hardly recognizable terms (Ptaszynski *et al.*, 2011). Existing Natural Language Processing (NLP) frameworks often struggle to address such phenomena, which is a problem for lexicon-based sentiment analysis approaches (Ritter *et al.*, 2011). As a result, ,the model employs generating character n-grams which has previously been proven to be effective for various NLP tasks, i.e. spam filtering (Kanaris *et al.*, 2007) or authorship attribution (Escalante *et al.*, 2011) The word-based models in sentiment analysis often suffer from disadvantages such as word identification (especially for Asian languages) or require specific and language-dependent knowledge. However, there is only limited evidence regarding the use of character n-grams even though some research has suggested that character n-grams could reach state of the art or even better performance (Peng, 2003) and could be easier to implement (Blamey *et al.*, 2012). This has also been confirmed for some sentiment analysis experiments (Rybina, 2012, Raaijmakers and Kraaij, 2008) although other results offer more mixed findings (Ye *et al.*, 2009). Since the sentiment analysis in multiple languages can’t rely on language-dependent components, our model suggests character n-grams for feature selection. The features weighting scheme utilizes tf-idf weighting, which can provide a significant increase in classification accuracy (Maas *et al.*, 2011).

The model utilizes the Support Vector Machines (SVM) classifier (Cortes and Vapnik, 1995) which is (along with Naïve Bayes and Maximum Entropy) the most frequently used within supervised learning models for sentiment analysis (Liu, 2012). An SVM classifier constructs a hyperplane (or a set of hyperplanes), which separates examples (represented as points in a space) of different categories by a maximized gap (Chang and Lin, 2011). Moreover, the character n-gram feature selection together with SVM has previously provided good results (Aisopos *et al.*, 2012). To evaluate the model performance, the 10-fold cross-validation has been applied. First, the model was built for each language separately. Then a multiple-languages model was created with the subsequent comparison of the performance of the models.

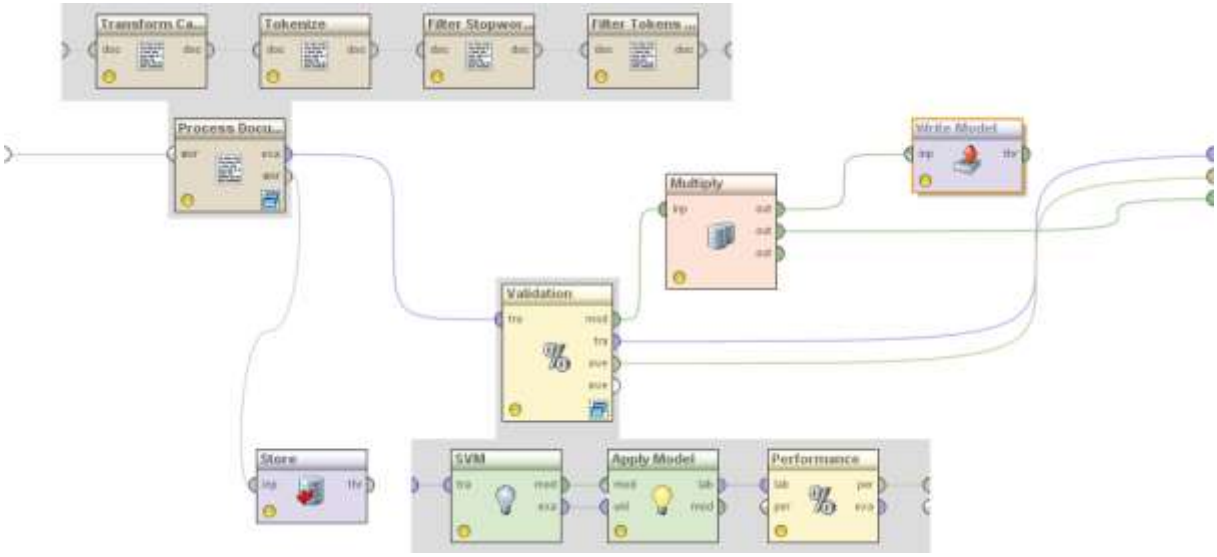


Figure 2 A classification model in RapidMiner Software

The multi-lingual sentiment analysis task can be approached as a domain adaptation problem (Cambria *et al.*, 2013). Aue and Gamon (2005) suggest four different approaches to overcome a domain-specificity problem. The suggested model in this paper utilizes training on a mixture of labeled data from all domains (languages) since such data was available.

3 Results

From each language, 200 positive (100 4-star and 100 5-star) and 200 negative (100 1-star and 100 2-star) reviews were randomly selected as a training set. It was first though that the number of downloaded reviews could be selected as further examples for a training set, however the number of 1- and 2- star French reviews were slightly over 100. The size of the training sets was therefore limited to allow a comparison of the model performance for different languages.

For the English language, the model reached an accuracy of 83.50%, precision: 84.17% and recall 84.00%. Some studies have shown better performance previously for English language (Liu, 2012), however they often utilized a language-dependent component (i.e. sentiment lexicons) and used a much larger training set, which also influences the performance (Brooke *et al.*, 2009). The performance of the model for the English language is depicted in Table 1.

Table 1 Performance of the classifier for the English language (10-fold cross-validation)

	true neg	true pos	class precision
pred. neg	166	32	83.84%
pred. pos	34	168	83.17%
class recall	83.00%	84.00%	

For the German language, the model performed very similarly. The accuracy was 83.50%, precision 85.26% and recall 82.50%. The performance of the model is summarized in Table 2.

Table 2 Performance of the classifier for the German language (10-fold cross-validation)

	true neg	true pos	class precision
pred. neg	169	35	82.84%
pred. pos	31	165	84.18%
class recall	84.50%	82.50%	

For the French language the model reached an accuracy of 84.75%, precision 84.36% and recall 86.00%. The performance of the model is depicted in Table 3.

Table 3 Performance of the classifier for the French language (10-fold cross-validation)

	true neg	true pos	class precision
pred. neg	167	28	85.64%
pred. pos	33	172	83.90%
class recall	83.50%	86.00%	

3.1 Results for all-in-one multilingual model

For the multilingual model, the same randomly selected comments from each language have been used as a training set. Therefore, the training set consists of 600 positive (100 4- and 100 5-star reviews from each of the three languages) and 600 negative (100 1- and 100 2-star comments from each of the three languages) reviews. The classifier accuracy reached 85.33%, precision 86.01% and recall 84.83%. The performance of the model is summarized in Table 4.

Table 4 Performance of the classifier for multiple languages (10-fold cross-validation)

	true neg	true pos	class precision
pred. neg	515	91	84.98%
pred. pos	85	509	85.69%
class recall	85.83%	84.83%	

Even though the classifier was trained on mixed data, the performance has slightly improved (by 1.83 % in case of English and German data and by 0.58 % for French language). This supports the theory that multilingual feature spaces could be outperforming models trained solely on individual languages (Banea *et al.*, 2010).

4 Conclusion and Discussion

The results indicate that there might be an alternative to common approaches addressing the cross-language sentiment classification issue. Such common approaches often utilize language-dependent components (Liu, 2012), i.e. sentiment lexicons or automated translators and thus are dependent on the quality of such resources. This provides a good level of classifier performance for major languages (i.e. Spanish, French or German), however not many studies have addressed sentiment analysis for morphologically rich or minor languages i.e. Arabian, Hebrew or Czech (Tsarfaty *et al.*, 2010, Habernal *et al.*, 2013).

The suggested model based on multilingual labeled data from three Amazon local websites (UK – English, German and French) outperforms the models that utilize the training set only from a single language. Such a conclusion is supported by previous experiments in related fields – i.e. Mansour *et al.* (2013) addresses a domain adaptation issue and suggests a simple "all-in-one" classifier (utilizing all available training data) that outperforms traditional approaches.

The classifier performance improved for English and German by 1.83% and for the French language by 0.58%. The model benefits from the fact that many languages share a significant amount of similar (or identical) words with the same (or close) meaning. A comprehensive list of such words can be found in Anon (n.d) for English and German languages or in Anon (n.d.) for English and French languages. There are even more words from various languages where the spelling is only partially identical – this is where the character n-grams come into play. An identical character n-gram from a similar word (but from different languages) with the same sentiment polarity improves the classifiers performance, since the model has more supportive data to distinguish between cases. Moreover, Anglicisms have intensively proliferated into a

wide variety of languages with the extensive use of digital media and social networks in recent years (Berns *et al.*, 2007).
Even though the results are promising, they must be considered carefully. Further studies including more languages and more domains should be conducted. Regardless of this, the authors have conducted several experiments using a similar model addressing the cross-domain sentiment classification task with similarly promising results.

5 References

Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T., 2012. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In: Proceedings of the 23rd ACM conference on Hypertext and social media, Milwaukee, WI. USA. June 25-28, 2012. ACM New York, NY, USA, pp. 187-196.

Aldred, J., Astell, A., Behr, R., Cochrane, L., Hind, J., Pickard, A., Potter, L., Wignall, A., and Wiseman, E., 2008. The world's 50 most powerful blogs, The Guardian [Online]. Available at: <http://www.guardian.co.uk/technology/2008/mar/09/blogs> [Accessed 6/4/2013].

Anderson, E.W., 1998. Customer Satisfaction and Word of Mouth. Journal of Service Research, 1(1), pp. 5-17.

Anon, n.d. Ähnliche Wörter Englisch–Deutsch, Wiktionary [Online]. Available at: <http://de.wiktionary.org/wiki/Verzeichnis:Englisch/%C3%84hnliche_W%C3%B6rter_Englisch%E2%80%93Deutsch> [Accessed 19/08/2014].

Anon, n.d. English-French relations, Wiktionary [Online]. Available at: <http://en.wiktionary.org/wiki/Appendix:English-French_relations> [Accessed 19/08/2014].

Aue, A., and Gamon, M., 2005. Customizing sentiment classifiers to new domains: A case study. In: Proceedings of the Recent Advances in Natural Language Processing RANLP 2005, Borovets, Bulgaria, September 21-23, 2005. Microsoft Research, pp. 1-7.

Banea, C., Mihalcea, R., and Wiebe, J., 2010. Multilingual subjectivity: are more languages better? In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, August 23-27, 2010. Association for Computational Linguistics, pp. 28-36.

Berns, M., De Bot, K., and Hasebrink, U., 2007. In the Presence of English: Media and European Youth: Media and European Youth. Springer.

Blamey, B., Crick, T., and Oatley, G., 2012. RU:-) or:-(? character-vs. word-gram feature selection for sentiment classification of OSN corpora. In: Proceedings of the Thirty-second SGAI International Conference on Artificial Intelligence Cambridge, UK, December 11-13, 2012. Springer, pp. 207-2012.

Brooke, J., Tofiloski, M., and Taboada, M., 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In: Proceedings of the Recent Advances in Natural Language Processing RANLP 2005, Borovets, Bulgaria, September 14-16, 2009. pp. 50-54.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. Intelligent Systems, 28(2), pp. 15-21.

Comcowich, W.J., 2010. Media Monitoring: The Complete Guide, CyberAlert [Online]. Available at: <http://www.cyberalert.com/downloads/media_monitoring_whitepaper.pdf> [Accessed 8/8/2013].

Cortes, C., and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp. 273-

297.

Escalante, H.J., Solorio, T., and Montes-Y-Gómez, M., 2011. Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, June 19-24, 2011. Association for Computational Linguistics, pp. 288-298.

Goldenberg, J., Libai, B., and Muller, E., 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters, 12(3), pp. 211-223.

Habernal, I., Ptácek, T., and Steinberger, J., 2013. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Atlanta, GA, June 14, 2013. pp. 65-74.

Horrigan, J.B., 2008. Online Shopping, Pew Internet & American Life Project [Online]. Washington, D.C. Available at: <http://www.pewinternet.org/Reports/2008/Online-Shopping/01-Summary-of-Findings.aspx> [Accessed 8/8/2014].

Chang, C.C., and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), pp. 1-39.

Kanaris, I., Kanaris, K., Houvardas, I., and Stamatatos, E., 2007. Words versus character n-grams for anti-spam filtering. International Journal on Artificial Intelligence Tools, 16(06), pp. 1047-1067.

Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), pp. 1-167.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., and Potts, C., 2011. Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, June 19-24, 2011. Association for Computational Linguistics, pp. 142-150.

Mansour, R., Refaei, N., Gamon, M., Abdul-Hamid, A., and Sami, K., 2013. Revisiting The Old Kitchen Sink: Do We Need Sentiment Domain Adaptation? In: Proceedings of the Recent Advances in Natural Language Processing, RANLP 2013, Hissar, Bulgaria, September 9-11, 2013. pp. 420-427.

Pak, A., and Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC, 2010, Valletta, Malta, May, 17-23, 2010. pp. 1320-1326.

Pang, B., and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), pp. 1-135.

Peng, F., Schuurmans, D. Wang, S., 2003. Language and task independent text categorization with simple language models. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL '03, Edmonton, Canada, May 27 - June 1, 2003. Association for Computational Linguistics, pp. 110-117.

Ptaszynski, M., Rzepka, R., Araki, K., and Momouchi, Y., 2011. Research on emoticons: review of the field and proposal of research framework. In: Proceedings of the Seventeenth Annual Meeting of the Association for Natural Language Processing (NLP-2011) Toyohashi, Japan, March 7-11, 2011. The Association for Natural Language Processing, pp. 1159-1162.

Raaijmakers, S., and Kraaij, W., 2008. A Shallow Approach to Subjectivity Classification. In: Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008, Seattle, WA, USA, March 30-April 2, 2008. Association for the Advancement of Artificial Intelligence, pp. 216-217.

Ritter, A., Clark, S., and Etzioni, O., 2011. Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, UK, July, 27-31, 2011. Association for Computational Linguistics, pp. 1524-1534.

Rybina, K., 2012. Sentiment analysis of contexts around query terms in documents. Master's thesis, Technische Universität Dresden.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp. 267-307.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., and Tounsi, L., 2010. 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, Los Angeles, CA, USA, June 5, 2010. Association for Computational Linguistics, pp. 1-12.

Ye, Q., Zhang, Z., and Law, R., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, 36(3), pp. 6527-6535.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B., 2011. Combining lexiconbased and learning-based methods for twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011-89.