On Text Preprocessing for Opinion Mining Outside of Laboratory Environments

Gerald Petz[1], Michał Karpowicz[1], Harald Fürschuß[1], Andreas Auinger[1], Stephan M. Winkler[2], Susanne Schaller[2], and Andreas Holzinger[3]

[1] University of Applied Sciences Upper Austria, Campus Steyr, Austria
{gerald.petz,harald.fuerschuss,michal.karpowicz, andreas.auinger}@fh-steyr.at
[2] University of Applied Sciences Upper Austria, Campus Hagenberg, Austria
{stephan.winkler,susanne.schaller}@fh-hagenberg.at
[3] Medical University Graz, Medical Informatics, Statistics and Documentation, Austria
andreas.holzinger@medunigraz.at

Abstract. Opinion mining deals with scientific methods in order to find, extract and systematically analyze subjective information. When performing opinion mining to analyze content on the Web, challenges arise that usually do not occur in laboratory environments where prepared and preprocessed texts are used. This paper discusses preprocessing approaches that help coping with the emerging problems of sentiment analysis in real world situations. After outlining the identified shortcomings and presenting a general process model for opinion mining, promising solutions for language identification, content extraction and dealing with Internet slang are discussed.

Keywords: Opinion mining, sentiment analysis, text mining, content extraction, language detection, Internet slang, Web analytics.

1 Introduction and Motivation for Research

The rise of Web 2.0 led to many changes of Internet use; the shift in communication processes allow users to participate and contribute [1], [2]. Neither private users nor companies can avoid this new form of communication, so it is important to pay attention to these communication processes: Content created by users of Web 2.0 applications contains a multitude of market research information and opinions concerning products, brands and corporations [3], through which economic opportunities and risks can be recognized at a very early stage. New strategic, tactical and operational plans and measures can be embraced and new brand messages can be created [4]. One of the biggest challenges for qualitative market research in the Web 2.0 is to handle the variety of information [5] and the handling of the diversity of data, which are mostly weakly structured and non-standardized [6]. Since manual processing of this big data is not manageable, the need for the support of human intelligence by machine intelligence is necessary [7]. Such semi-automatic solutions can be found in the field

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 618–629, 2012.

of opinion mining, which deals with the identification and extraction of subjective opinions about certain topics, e.g. products or services [3].
While performing well in a laboratory environment, where specifically prepared texts are used [8], several problems arise, when using the prototype in real-world situations. These require additional processing to be performed in order to improve the quality of the opinion mining process. Three crucial shortcomings were recognized:

- Contrary to laboratory environments, the content acquired in real-world situations often features a multitude of languages [9, 10]. In the field of opinion mining, where language-specific tools and models are frequently utilized, this is a major problem, since the application of improper methods leads to incorrect or worse sentiment analysis results.
- While social networks, such as *Facebook* or *Twitter*, typically provide programming interfaces, which can be used to acquire clean data, the main content on webpages is almost always surrounded by advertisements, navigational components and other context-irrelevant elements. This so-called *boilerplate* should not be considered, since it provides no valuable information and can decrease the classification accuracy (as described in the context of web mining in [11]).
- User-generated content often includes so called Internet slang, such as emoticons or abbreviations, which is rarely considered in the context of opinion mining. By neglecting this growing phenomenon, a lot of hidden information about the author's sentiment gets lost [12].

The objective of this paper is to discuss techniques that can help coping with the challenges that arise when performing sentiment analysis outside of laboratory environments. In order to achieve this, the utilized prototype and its background are described in the following section. The subsequent section focuses on outlining the identified problems and providing promising solutions to overcome these shortcomings when performing opinion mining in real-world situations.

2 Background and Related Work

2.1 Opinion Mining

In general, opinion mining deals with scientific methods in order to find, extract and systematically analyze views on certain topics, e.g. products, companies or brands. Opinion mining has been studied by many researchers in recent years; one can identify several main research directions [13]: (1) *Sentiment classification*: The goal of sentiment classification is the classification of content according to its sentiment (positive, neutral or negative) about objects (e.g. [14, 15]). (2) *Feature-based opinion mining*: In feature-based opinion mining the sentiment about certain properties of objects (e.g. technical features of a digital camera) is analyzed at sentence level (e.g. [16]). (3) *Comparison-based opinion mining*: Another focus is the comparison-based opinion mining, dealing with the finding of sentences, in which comparisons of similar objects are made (e.g. [17]).

Different technical approaches for opinion mining can be identified: Sentiment classification based on document level and based on sentence level or on feature level. Most classification methods – both at document level and sentence level – are based on the identification of opinion words or phrases. Two approaches are usually used: corpus-based approaches (e.g. [18], [19], [20]) and dictionary-based / lexicon-based approaches (e.g. [16], [14], [21], [22]).

A multitude of algorithms can be used to build models that are able to classify the sentiment of sentences / documents. This includes common approaches such as support vector machines (SVM) [23], linear regression (LR) [24, 25], decision trees (DT) [26], artificial neural networks (ANN) [2], Latent Dirichlet Allocation (LDA) [27], Pointwise Mutual Information (PMI) [27], and genetic programming (GP) [29, 30]. A variety of other approaches and algorithms have been studied: e.g. Markov blanket classification [31], or joint sentiment / topic models (JST) [32].

2.2 Prototype

In the research projects *OPMIN 2.0* and *SENOMWEB*, a software prototype for an opinion mining using a feature-based approach was implemented. This implementation is based on a process model, shown in figure 1, which illustrates the process steps taken in order to find, extract and analyze web data with regard to their sentiment orientation ([33], [34]).

In the first two steps the relevant data sources, such as social networks, forums, websites or RSS-feeds, and appropriate methods for performing the analysis are chosen. The focus of the preprocessing step is the preparation and pre-structuring of the extracted content in order to be able to transform it into a processible structure using methods from the fields of natural language processing and information retrieval in the next phase. In the fifth step the sentiment analysis is performed to determine if a sentence is positive, negative or neutral (cf. section 4). Lastly an evaluation is done to examine the quality of the methods used. The prototype utilizes different kinds of text preprocessing algorithms as well as classification algorithms (e.g. binary classifiers as well as multi-class models [35]). Multi-class classification models assign one specific class (from a set of available classes) to each sample whereas binary classifiers are trained to label samples as belonging to a given class or not.

When using the prototype based on the process model outside of the laboratory environment, several shortcomings could be identified. Since the later steps, especially the analysis, rely on clean and correct content and make use of language-specific models, the recognized problems need to be solved at an early stage. Therefore the techniques and approaches that aim to overcome these issues need to be integrated in the preprocessing phase. Before the processing of Internet slang in the general context of preprocessing, the language must be identified and the relevant content has to be extracted. Thus techniques from the well-established research fields of language detection and content extraction are presented in the sections 3.1 and 3.2, while the subsequent preprocessing of the text is described in section 3.3.

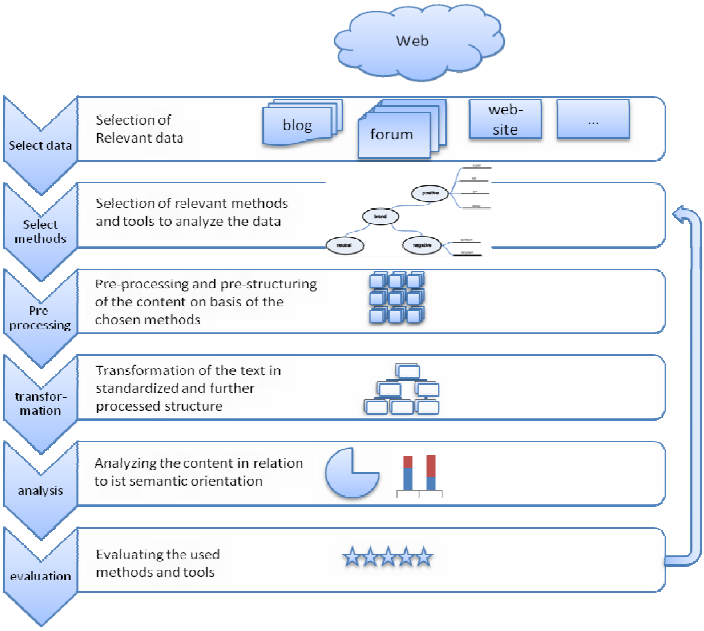

Fig. 1. Process model for opinion mining [34]

3 Approaches for Improving Sentiment Analysis

3.1 Language Identification

As stated above, opinion mining in Web 2.0 has to deal with different languages [10]. In order to perform multilingual opinion mining using language-specific models, that can be obtained by language-specific training or by using methods that automatically generate corpora for new languages using available resources (e.g. [36]), techniques for automatic language identification are required. Being a well-established field of research many possible solutions for determining the language of written text have emerged, ranging from approaches based on data-compression techniques [37] to methods that use language-specific short common words, such as determiners, conjunctions or prepositions [38]. A set of techniques, which are widely utilized in the field of language identification [39], are based on character *n-grams*, i.e. sequences of characters. A representative approach using n-grams was proposed in [40]: First training data is used to create language-specific models, which describe the frequency of the occurring character sequences. In order to classify a text, the frequencies of the contained n-grams are calculated and compared to the ones in the available language models. This way the likelihood of a text to be written in a certain language can be determined, e.g. a document containing multiple occurrences of the character sequence `CZE` is more likely to be Polish than English.

This approach has numerous advantages: Not only is it easy to implement, it also is computationally cheap and has very small language models, which don't require a lot of training samples to construct [40]. While providing high accuracy on long texts [41], the technique doesn't perform well on short fragments (e.g. sentences) [38], especially when trying to detect similar languages. This deficiency is problematic in the context of opinion mining: Since user-generated content often contains multiple languages, e.g. sentences using English Internet slang between text in a user's native language, the identification needs to be performed on short texts.

A promising approach that performs better on short text fragments, is presented in [39]: Using dictionary language models obtained by training, each word gets a score based on how specific it is for a language. This value, referred to as relevance, is calculated by comparing the language-specific frequencies of a word with its frequency in a general, language neutral background corpus, which can be approximated by merging all language-specific corpora.

While requiring larger dictionary models and more training compared to approaches based on n-grams, the described dictionary method provides a number of benefits, which can be useful in the opinion mining process. In contrast to n-gram-based techniques, that only return the most likely language, the dictionary approach can return a set of possible languages as well as no language at all in case the identification was not successful. In the context of an opinion mining tool this would enable the involvement of a human to perform the classification, instead of using the most likely but wrong language. Additionally the authors of the dictionary approach describe a fast method for language segmentation, which can be used to identify single-language blocks in a text. This segmentation can improve the precision of language-specific algorithms, increasing the overall quality of multilingual opinion mining.

3.2 Content Extraction

The loss of precision caused by non-relevant data surrounding the main content is a major problem when performing opinion mining on webpages. While human readers can easily distinguish between the relevant parts of a page and the surrounding *boilerplate* (such as navigation components, advertisements or references to related articles), the automatic identification of the main content is a nontrivial task keeping researchers busy since the middle of the 1990s [3]. Because of the possible benefits that content extraction algorithms can provide to many domains – e.g. by improving the precision of search engines or enhancing the readability of web articles on mobile devices – numerous approaches have been developed. We evaluated the following approaches by using a German-language data set.

Techniques. The earliest approaches to solve the problem of extracting the main content of webpages were based on *wrappers*: programs that gather relevant data from a page according to a site-specific set of rules, which are either handwritten or automatically derived from a number of training documents [42]. Since the aforementioned rules need to be established for every website (resp. website template) specifically, wrapper-based solutions not only require a set of example pages or a developer

manually defining extraction criteria, but also constant maintenance (in case of design changes etc.). Over the years much progress was made in the field of automatic content extraction: Many approaches have been developed, that resolve the shortcomings of wrapper-based techniques. Because of their multitude, an extensive evaluation of all methods would go beyond the scope of this section. Instead criteria are defined, that content extraction techniques have to meet to be eligible for the use in the opinion mining process:

- Techniques used in the opinion mining process should require little to no manual effort.
- In order to be able to process large numbers of documents, content extraction methods should not be costly regarding performance and resources.
- Eligible techniques must not be specific to certain websites. An approach should not need site-specific training, since the availability of example pages can't be guaranteed in the opinion mining process.
- The approaches should not be based on assumptions regarding structural information. Since web development is an ever-changing technology, no clues regarding the main content should be taken from certain HTML-elements.

According to these criteria three promising approaches have been chosen for further analysis and evaluation.

Content Extraction via Tag Ratios (CETR). [43] describes an approach for automatic content extraction, that is based on a simple observation: While typical boilerplate, such as navigation components or advertisements, contains much code and only few text, content sections mostly consist of text. According to this observation, document lines with a high text to HTML-tag ratio are likely to be part of the relevant content and vice-versa. In order to identify the main text, a histogram with the text-to-tag ratios of a document is constructed and smoothed, which is necessary to prevent the loss of content lines at the edges. [44] propose three methods to obtain the content from a document using the smoothed histogram: applying a threshold using standard deviation to determine if lines are content or boilerplate (*CETR-TM*), utilizing k-means clustering (*CETR-KM*) or constructing a 2D model using the histogram and the absolute smoothed derivatives of the tag ratios to perform 2D clustering (*CETR*).

Maximum Subsequence Segmentation (MSS). [45] proposes the use of maximum subsequence optimization to identify a document's main content. For this approach, a page is first tokenized (broken up into tags, words and symbols). A token-level classifier is then used to find a score for each token, which is based on trigrams (the token and its two successors) and its parent tag. Since the scores indicate the tokens' likelihood to be boilerplate, the relevant content can be identified by finding the maximum subsequence. In order to enable the determination of token scores, the classifier requires training on labeled example documents.

Boilerpipe. [46] analyzed language- and domain-independent boilerplate detection using a number of shallow text features, such as average word or sentence length.

They found out, that, after segmenting a document into atomic text blocks, high accuracy in non-content identification can be achieved through the use of solely two text features: *number of words* and *link density* per block. Based on this research *boilerpipe*, a library for automatic content extraction, has been developed by one of the authors of the original paper.

Evaluation. In this part the evaluation setup, including the data set, implementation and metrics used, as well as the results are presented.

Data set and implementation. For training and evaluation purposes a data set consisting of 700 German-language documents from 205 different sources was prepared. The example pages were arbitrary collected from German and Austrian news sites, whereby no distinction was made regarding the popularity and size of the sites or the topics covered. All documents in the data set were hand-labeled by a human annotator using special markers to flag the relevant content.

In order to perform automatic content extraction on the documents in the data set the respective authors' implementations[1] were obtained. While the *boilerpipe* library didn't need any configuration and was used with the extractor types *ArticleExtractor* (specialized for the use on web articles) and *DefaultExtractor*, the parameters needed for *CETR* were determined empirically: The best results could be achieved when using two clusters (*CETR* and *CETR-KM*) and setting the threshold-coefficient to 2.0 (*CETR-TM*). The *MSS* classifier was trained on two-thirds of the data set, leaving one third for the evaluation. In order to increase the accuracy of the evaluation, ten training-evaluation cycles were performed using randomly selected subsets of the whole data set, the results of which were averaged.

Metrics. To evaluate the techniques, various evaluation metrics are used, each of which illustrates a different aspect of the presented approaches and their performance.

Precision and recall - common measures used in the context of information retrieval [47] - can be used to evaluate content extraction techniques in order to depict the relevance of extracted content. In this paper these measures are calculated on token-level using a bag-of-words approach, where precision, recall and their harmonic mean (F_1*-score*) can be defined as

$$P = \frac{|W_h \cap W_a|}{|W_a|}, R = \frac{|W_h \cap W_a|}{|W_h|}, F_1 = 2 \times \frac{P \times R}{P+R} \quad (1)$$

with W_h being the words found by a human annotator and W_a being the words extracted automatically.

Other metrics that can be used when evaluating content extraction techniques are the Levenshtein edit distance, which represents the number of edit operations (insertions, deletions and substitutions) required to transform one sequence into another, as

[1] *CETR:*`http://www.cs.illinois.edu/~weninge1/cetr/`,
MSS:`http://jeffreypasternack.com/demos.aspx`,
boilerpipe: `http://code. google.com/p/boilerpipe/`.

well as the alignment length, which additionally counts the required align operations. For the evaluation in this paper modified versions - working on token-level with no substitution operations allowed - of these metrics are utilized, analogous to the scoring system employed in the *CleanEval* shared task for cleaning webpages [48], that considers a text-only score [45] defined by

$$score(a,b) = 1 - \frac{editDistance(a,b)}{alignmentLength(a,b)} \tag{2}$$

which, while being expensive to compute, provides a good measure for the quality of the extracted content. Since the described techniques should come into use in the opinion mining process, time is an important factor. To allow a comparison of the approaches, their computational time is measured using a randomly assembled set of 100 documents on which content extraction is performed using a *Dell Latitude E6520* laptop powered by an *Intel Core i5-2520M* processor with 4GB RAM.

Table 1. Results of the evaluation

Technique	Precision	Recall	F_1-Score	Text-only score	Time
Boilerpipe Art.	87.94	92.86	90.33	78.79	1.62 s
Boilerpipe Def.	72.49	90.65	80.56	61.95	0.7 s
CETR	69.26	91.77	78.94	60.05	1.97 s
CETR-KM	70.22	90.35	79.02	60.24	1.19 s
CETR-TM	70.44	90.72	79.30	61.05	0.76 s
MSS	91.29	93.56	92.41	81.72	4.1 s

Results. The results, presented in table 1, show that the extraction using *MSS* with a trained model performs best on the German data set. While having slightly worse scores, the *boilerpipe* article extractor not only runs more than twice as fast, it also requires no language-specific training or modifications, which makes this particular technique very suitable for multilingual opinion mining.

Although performing worst in this evaluation, the approaches based on tag-ratios can be a good and fast option when high recall is top priority. By adjusting the parameters, which were chosen empirically in order to obtain high *CleanEval* text-only scores in this evaluation, the recall rate can be enhanced at the cost of precision. This can be achieved by increasing the number of clusters (*CETR* and *CETR-KM*) or decreasing the threshold coefficient (*CETR-TM*).

3.3 Text Preprocessing

The frequent lack of terminal punctuation and grammar in user-generated content poses a problem for sentence-based sentiment analysis. Therefore text preprocessing plays a major role in opinion mining and further analysis approaches.

In order to obtain a satisfactory sentiment analysis result, the following text preprocessing steps are applied:

1. Splitting into sentences and words
2. Replacing acronyms, symbols and emoticons
3. Stemming

In the first step, text blocks are splitted into their sentences and furthermore each sentence into its words for better handling. Next, to make symbols and emoticons usable for sentiment analysis, a corpus with over 500 symbols and emoticons is utilized in order to make symbols and emoticons usable for sentiment analysis. The meaning of these symbols and emoticons was determined manually. All symbols and emoticons that are detected in the sentences are replaced by significant words that enable a proper sentiment analysis result (e.g. the emoticon :-) is replaced by the word *funny*). Additionally, in order to process acronyms in a useful way, a dictionary which includes the most commonly used abbreviations and its synonyms was built and is applied on each sentence. Finally a stemming tool called *TreeTagger* [49] is used to tag every single word and put it in its principal form. Therefore, the sentences are not only preprocessed and pre-structured, but also contain a unified structure that is used for building a matrix, which conduces as input for machine learning algorithms that are utilized for the analysis.

4 Conclusion and Further Research

This paper discusses problems and promising solutions that could be identified when using a prototypical opinion mining tool outside of laboratory environments. The presented approaches and techniques were put into the context of opinion mining and integrated into a general process model as preprocessing tasks. Further research is needed in order to fully assess the benefits added by utilizing these approaches. Therefore an extensive evaluation of an enhanced opinion mining tool being put to use in real-world situations is planned.

Acknowledgements. This work emerged from the research projects *OPMIN 2.0* and *SENOMWEB*. The project *SENOMWEB* is funded by the European Regional Development fund (*EFRE*, *Regio 13*). *OPMIN 2.0* is funded under the program *COIN – Cooperation & Innovation*, a joint initiative launched by the Austrian Ministry for Transport, Innovation and Technology (*BMVIT*) & the Ministry of Economy, Family and Youth (*BMWFJ*).

References

1. Alby, T.: Web 2.0. Konzepte, Anwendungen, Technologien, 3rd edn. Hanser, München (2008)
2. Nelles, O.: Nonlinear system identification: from classical approaches to neural networks and fuzzy models. Springer (2001)
3. Liu, B.: Web data mining. Exploring hyperlinks, contents, and usage data, 2nd edn. Data-centric systems and applications. Springer, Berlin (2008)

4. Steinecke, U., Straub, W.: Unstrukturierte Daten im Business Intelligence. Vorgehen, Ergebnisse und Erfahrungen in der praktischen Umsetzung. HMD - Praxis der Wirtschaftsinformatik 47(271), 91–101 (2010)
5. Guozheng, Z., Faming, Z., Fang, W., Jian, L.: Knowledge Creation in Marketing Based on Data Mining. In: International Conference on Intelligent Computation Technology and Automation (ICICTA), vol. 1, pp. 782–786 (2008)
6. Holzinger, A.: Weakly Structured Data in Health-Informatics. In: Proceedings of INTERACT 2011 International Conference on Human-Computer Interaction, Workshop: Promoting and Supporting Healthy Living by Design, pp. 5–7 (2011)
7. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data. In: Proceedings of the 9th International Joint Conference on e-Business and Telecommunications (ICETE 2012), pp. IS9–IS20 (2012)
8. Holzinger, A., Geierhofer, R., Modritscher, F., Tatzl, R.: Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses. Journal of Universal Computer Science 14(22), 3781–3795 (2008)
9. Maynard, D., Bontcheva, K., Rout, D.: Challenges in developing opinion mining tools for social media. In: Proceedings of @NLP can u tag #user_generated_content?! Workshop at LREC 2012, Istanbul, Turkey (May 2012)
10. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst. 26(3), 12:1–12:34 (2008)
11. Yi, L., Liu, B.: Web page cleaning for web mining through feature weighting. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 43–48. Morgan Kaufmann Publishers Inc., San Francisco (2003)
12. Gamon, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
13. Kaiser, C.: Opinion Mining im Web 2.0 – Konzept und Fallbeispiel. HMD - Praxis der Wirtschaftsinformatik 46(268), 90–99 (2009)
14. Kim, S.-M., Hovy, E.: Determining the Sentiment of Opinions. In: Proceedings of 20th International Conference on Computational Linguistics, Geneva, Switzerland, pp. 1367–1373 (2004)
15. Nadali, S., Masrah, A.A.M., Rabiah, A.K.: Sentiment Classification of Customer Reviews Based on Fuzzy logic. In: Mahmood, A.K. (ed.) International Symposium in Information Technology (ITSim), pp. 1037–1044. IEEE (2010)
16. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
17. Jindal, N., Liu, B.: Mining comparative sentences and relations. In: Proceedings of the 21st National Conference on Artificial Intelligence, vol. 2, pp. 1331–1336. AAAI Press (2006)
18. Hatzivassiloglou, V., Wiebe, J.: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: Proceedings of the 18th Conference on Computational Linguistics, pp. 299–305 (2000)
19. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, pp. 417–424 (2002)
20. Wiebe, J., Mihalcea, R.: Word Sense and Subjectivity. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 1065–1072 (2006)

21. Ding, X., Liu, B., Yu, P.S.: A Holistic Lexicon-Based Approach to Opinion Mining. In: International Conference on Web Search & Data Mining, Palo Alto, California, February 11-12. ACM, New York (2008)
22. Popescu, A.-M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of Human Language Technology Conference, pp. 339–346 (2005)
23. Tong, S., Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research 2, 45–66 (2001)
24. Weisberg, S.: Applied linear regression, vol. 528. Wiley (2005)
25. Vapnik, V.: The nature of statistical learning theory. Springer (2000)
26. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
27. Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.M., Holzinger, A.: A Comparison of Different Retrieval Strategies Working on Medical Free Texts. Journal of Universal Computer Science 17(7), 1109–1133 (2011)
28. Holzinger, A., Simonic, K.M., Yildirim, P.: Disease-disease relationships for rheumatic diseases Web-based biomedical textmining and knowledge discovery to assist medical decision making. In: IEEE COMPSAC, pp. 573–580 (2012)
29. Koza, J.: Genetic programming II: automatic discovery of reusable programs (1994)
30. Affenzeller, M., Wagner, S., Winkler, S.: Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. Numerical Insights. Taylor & Francis (2009)
31. Bai, X.: Predicting consumer sentiments from online text. Decision Support Systems 50(4), 732–742 (2011)
32. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384. ACM, New York (2009)
33. Faschang, P., Petz, G., Dorfer, V., Kern, T., Winkler, S.M.: An Approach to Mining Consumer's Opinion on the Web. In: 13th International Conference on Computer Aided Systems Theory, Eurocast 2011, pp. 37–39 (2011)
34. Faschang, P., Petz, G., Wimmer, M., Dorfer, V., Winkler, S.M.: Evaluation of Tools for Opinion Mining. In: EEE (ed.) Proceedings of the 2011 International Conference on E-Learning, E-Business, Enterprise Information Systems & E-Government, Las Vegas, pp. 3–9 (2011)
35. Schaller, S., Winkler, S.M., Dorfer, V., Petz, G., Fürschuß, H.: A Machine Learning Suite for Opinion Mining in Web. In: Proceedings of the 14th International Asia Pacific Conference on Computer Aided System Theory, IEEE APCast (2012)
36. Mihalcea, R., Banea, C., Wiebe, J.: Learning Multilingual Subjective Language via Cross-Lingual Projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 976–983 (2007)
37. Benedetto, D., Caglioti, E., Loreto, V.: Language Trees and Zipping. Phys. Rev. Lett. 88(4), 48702 (2002), doi:10.1103/PhysRevLett.88.048702
38. Grefenstette, G.: Comparing two language identification schemes. In: Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT 1995), pp. 263–268 (1995)
39. Řehůřek, R., Kolkus, M.: Language Identification on the Web: Extending the Dictionary Method. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 357–368. Springer, Heidelberg (2009)
40. Cavnar, W.B., Trenkle, J.M.: Trenkle: N-Gram-Based Text Categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175 (1994)

41. Dunning, T.: Statistical Identification of Language (1994)
42. Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S., Teixeira, J.S.: A brief survey of web data extraction tools. SIGMOD Rec. 31(2), 84–93 (2002), doi:10.1145/565117.565137
43. Weninger, T., Hsu, W.H.: Text Extraction from the Web via Text-to-Tag Ratio. In: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application, pp. 23–28. IEEE Computer Society, Washington, DC (2008), doi:10.1109/DEXA.2008.12
44. Weninger, T., Hsu, W.H., Han, J.: CETR: content extraction via tag ratios. In: Proceedings of the 19th International Conference on World Wide Web, pp. 971–980. ACM, New York (2010), doi:10.1145/1772690.1772789
45. Pasternack, J., Roth, D.: Extracting article text from the web with maximum subsequence segmentation. In: Proceedings of the 18th International Conference on World Wide Web, pp. 971–980 (2009)
46. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 441–450. ACM, New York (2010)
47. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton, MA, USA (1979)
48. Baroni, M., Chantree, F., Kilgarriff, A., Sharoff, S.: Cleaneval: a Competition for Cleaning Web Pages
49. Schmid, H.: TreeTagger - a language independent part-of-speech tagger, `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` (accessed March 10, 2011)