Vadász Pál

Semantic technologies in sentiment analysis

Abstract

Since the early 2000s a new branch of NLP is gaining ground, sentiment analysis 1 using semantic technologies. Sentiment analysis can be carried out at different levels, document, sentence and the more complex aspect-based2. Sentiment lexicon is a vital part of a sentiment analysis application. An even younger field, emotion mining expands the application field of these computational linguistic disciplines.

Introduction

Due to the development of meaning based computing, semantic technology applications are increasingly capable of performing operations that emulate human communication in an intelligent way. Since the early 2000s semantic technology solutions have been gaining more and more ground in industry and government, as well as in everyday life. [1, pp. 311-331]

Sentiment analysis is widely used in reviewing and evaluating all kinds of social media contents, such as forums, blog sites, message boards, Twitter and Facebook. Sentiment analysis can be applied to the unstructured or semi-structured internal corpora, such as wikis, emails, call centres of companies or organisations. Classification based on opinion, or the emotion it carries, can be useful in information retrieval when one has to filter out certain given types of documents.

The online and real time application of this technology ranges from marketing and PR agencies through reputation analysers, benchmarking specialists, political campaign managers to brand monitors, medical user forums and financial analysts, to mention a few.

Sentiment analysis has been a sizzling topic of scientific literature and commercial development since the early 2000s. More than 7,000 publications have been written in this field and numerous companies are engaged in computational linguistic and statistical solutions such as SAS and IBM/SPSS. It is beyond the scope of this publication to give a

[1] Also called opinion mining, sentiment mining, opinion extraction, review mining.

[2] Also called fragment-based

detailed review of the military and law enforcement applications of sentiment analysis. However, there is doubt that in the field of intelligence, information warfare, psychological operations and CIMIC it is unavoidable to follow the opinion and motivation of the target population.

In the light of the above, the purpose of the present publication is to review and to summarise for the specialists of military and LEA applications the essence of sentiment analysis, its purpose, specifics and applied methods. We shall particularly focus on

- the foundations of sentiment analysis;
- the levels of sophistication or depth of sentiment analysis;
- the role of this sentiment analysis in different cultures, languages with special interest on its limitations.

Foundations of sentiment analysis

In the following section sentiment analysis will be defined, its purpose briefly described and classification methods and semantic dictionaries introduced.

Definition of sentiment analysis

Sentiment analysis3 can be defined as the task of finding opinions of authors about given entities in text corpora and then analysing the polarity of these opinions. [2, pp. 82-89] In other words, sentiment analysis is to determine the attitudes, opinions, appraisals, effects, views, emotions and subjectivity of people to a specific target. Sentiment analysis is considered to be a branch of Natural Language Processing(NLP).

Architecture

Before going into details, it is necessary to understand the basic architecture of a sentiment analysis application.

Input can be any structured text file or files, such as WORD, PDF, HTML, XML etc. These files need to be pre-processed applying tools like stemming, tokenisation, POS tagging or entity extraction. Sentiment lexicons and dictionaries of other relevant language elements, as well as taxonomies can be used to enrich, disambiguate or to perform other linguistic tasks. The heart of the process is the document analysis whereby the pre-pro-

[3] Also called opinion mining, sentiment mining, opinion extraction, review mining.

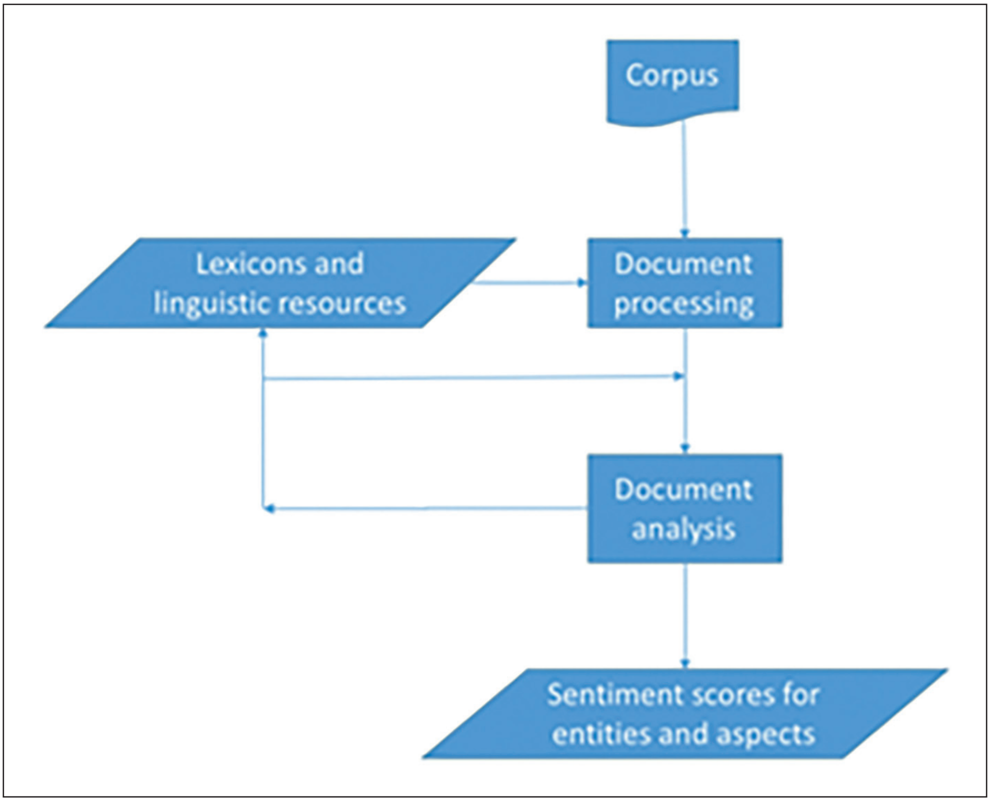
Corpus
Lexicons and linguistic resources
Document processing
Document analysis
Sentiment scores for entities and aspects

Figure 1. Architecture of a generic sentiment analysis system. Source [2]

cessed documents are provided with sentiment annotations. These annotations can be at different levels (document, sentence or aspect-based as described later). The annotated documents supplied with the scores are the output.

Classifying in sentiment analysis

It is crucial to determine, into how many different categories we intend to classify the analysed text. It may be various. Characteristic numbers are 2 (positive or negative), 3 (positive, neutral, negative), 5 (evaluation of a five-star scale) or 10 (ten point scoring scale). In general, an increase in the number of groups decreases the accuracy, increases the difficulty of the task, but the results delivered are much more informative.[3]

Sentiment lexicons

Words carrying sentiments are fundamental to identifying the sentiments in a sentence or document. In an organized structure these words constitute a sentiment lexicon [3, pp. 1-167]. Sentiment lexicons are absolutely crucial for the analysis. They can be created in three different ways: manually, which is extremely tedious and uneconomical, though sometimes unavoidable. Dictionary based, whereby the basic dictionary is enlarged by a lexical database such as WordNet. Corpus-based whereby a basic dictionary is enriched utilizing a large set of domain based documents.

Positive and negative elements and sentiment shifters

A given word or expression may have a different polarity depending on the target or a specific domain, namely the language element is target or domain dependent. A “huge” plate full of pancakes is positive for children, but a “huge” fiasco is negative for the managing director. The consequence is that for the sentiment analysis of a given corpus individual sentiment lexicon may have to be prepared.

Sentiment shifters influence the polarity of the targets. Sentiment values can be strengthened (“extremely interesting”) and weakened (“hardly interesting”) with the help of intensifiers, or negated (“not interesting”). It is worth mentioning that recent research is focusing on this problem with the help of semantic compositional rules.[4]

The depth of SENTIMENT analysis

Depending on the scale of subjectivity and the field of application required, different levels of sentiment analysis can be applied. The simplest way is to look at the document as one unit. In the case of a more heterogeneous corpus, the sentence level is recommended. The most sophisticated method is the fragment-level or aspect-based sentiment analysis that can be very computation intensive. Therefore, in some cases, the separation of objective and subjective sentences is necessary.

Document-level sentiment analysis

Document-level sentiment analysis is the most basic type. It assumes that the author has one type of sentimental approach to each object throughout the entire document. The content is determined to be either positive or negative.

There are two ways of processing the analysis of a document. If training data is available, one can apply the supervised learning method i.e. the classification. Using the training data, the system classifies the hitherto unclassified documents into the given classes with the help of classification algorithms, such as KNN, SVM, naïve Bayes etc. However, if training data is not available, unsupervised learning is to be applied for the grouping. During the process semantic orientation is determined by specific phrases. These phrases are selected either by a lexicon of predefined sentiments or part-of-speech (POS) patterns.

Sentence-level sentiment analysis

Since there can be multiple views of the same object within one document, one needs to go down to sentence level in order to get a sharper picture. Further, to find out the polarity of the sentence, one may have to filter out the objective sentences in order to better determine the polarity of the subjective ones. [5, pp. 271-278] The analysis of objective sentences is more difficult and less fruitful. As discussed in the case of document level analysis, sentence level analysis requires either supervised or unsupervised learning for the classification or grouping. Once the analysis of the sentences is completed, the sentence level result is summarised.

Fragment-level sentiment analysis

The methods discussed above are simplifications of handling real life sentiment expressions. More often than not within one document, or even a sentence, people express their views on one entity describing several attributes or aspects. The field of fragment-level sentiment analysis4 is to detect within a document all minimal fragments containing an entity or one of its aspect and the opinion attributed to the given target separately. Once they have been identified, the polarity of these targets can be established and measured.

Applying an aspect tree is clearly demonstrated by VIRMANI [6, pp. 3262-3266]. The solution is used to get an opinion value for each student based on the remarks given by their teachers. It is outside the scope of the present publication to discuss the process in detail. However, the entity (the student) is clearly visible at the root (top), the hierarchy of the aspects and the weight given to each leaf of the tree.

4 Also called feature-based opinion mining

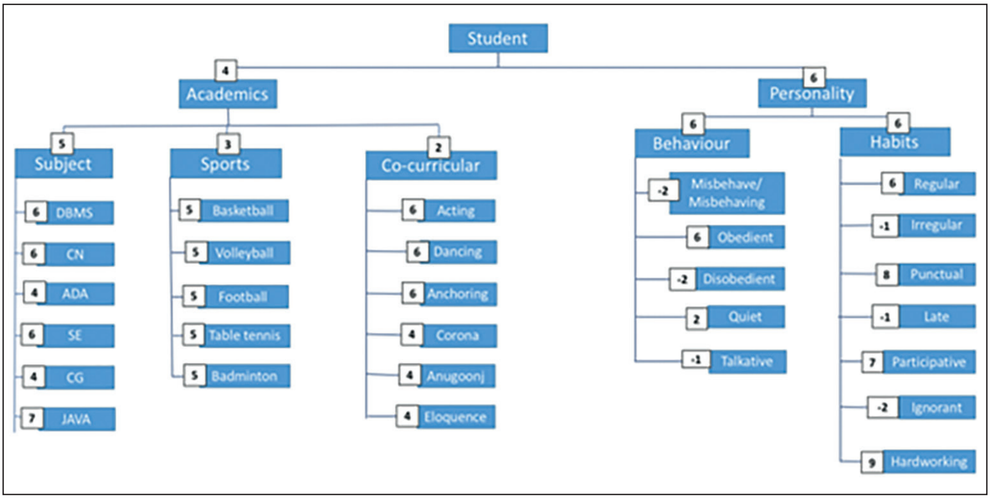

Figure 2. Aspect tree. Source [6] somewhat altered

Some additional Points ON sentiment analysis

It is worth looking outside the English speaking boundaries of the semantic analysis world. While most work is done in English, significant effort are being made in other countries. A quick summary of the field in Hungary cannot be missed. Finally, further tasks and the boundaries of the technology are summarized.

Sentiment analysis in other languages

Due to the dominance of the English language in the web and the fact, that the lingua franca of the scientific research is English, most publications in the field of sentiment analysis target the English language corpora.

There are two ways for sentiment analysis in other languages. The more obvious is to apply a machine translator like Google Translate, if the text to be translated is not sensitive and thus can be processed in the cloud. If the text is classified and therefore cannot be processed in an unsecured environment, it must be translated offline in an environment not connected to the Internet using translation applications such as the product of the French-Korean Systran. Once the translation is complete, all linguistic tools in English are available for analysis. I have found no literature on how perfect such an analysis can be in light of the limited accuracy of the translation programs.

The other way is to perform sentiment analysis in the foreign language. This method obviously assumes that the proper linguistic tools are available in the given language. An example is given by ABDUL-MAGEED [7]. There is no wonder that the larger the population of a language is and the more funds are available for linguistic research, the more sophisticated the applications are. There is one exception. Languages that are of special interest to military and LEA specialist enjoy more attention of organisations that are not necessarily domiciled in the country of the given language.

Sentiment analysis in Hungary and in Hungarian

Since the proportion of the web pages in Hungarian worldwide is about 0,4% and the commercial drive for such solutions in Hungary is just awakening, the scientific literature on this subject is very scarce, nevertheless it is growing. Main research centres are at the Szeged University, MTA SZTAKI, MTA Linguistic Institute. Not all publications refer to sentiment analysis in Hungarian language, some rather to English.

The first publication to be found chronologically in Google Scholar is that of Richárd FARKAS and Gábor BEREND on opinion mining. [8, pp. 408-412] They analyse the sentiment of short forum notes for and against the double citizenship of the portal magyarorszag.hu5. The team tried several methods whose, combination resulted in a 71% correlation with human evaluations.

Sándor SZASZKÓ, Péter SEBŐK and László T. KÓCZY analysed film reviews from port.hu and index.hu. They reached 70-80%, depending on the method. [9]

Martina Katalin SZABÓ published on the experiences gained in the preparation of a sentiment lexicon in Hungarian. [10]

Martina Katalin SZABÓ and Veronika VINCZE published on a deeply annotated sentiment corpus of texts written in Hungarian. [11, pp. 219-226]

Of the commercial ventures, the best known company offering opinion mining services is Neticle Technologies. [12] They announce a growing interest from PR and marketing circles as well as large brand owners.

Analysing public opinion in the social media during election times is discussed by Precognox Ltd. in its well known blog. [13]

Emotion analysis

Though the computational linguistic tools of emotion analysis are similar to those of sentiment analysis, one must make a clear distinction between the two disciplines. While sentiment analysis is exploring the view of authors on certain targets, emotion analysis is endeav-

[5] A government portal

ouring to detect emotional revelations in texts. The granularity used by emotion analysis is frequently based on Paul EKMAN's main categories [14] (joy, sadness, fear, surprise, disgust and anger) or Robert PLUTCHIK's theory of emotion [15], whose wheel of emotions is worth showing. An emotional analysis application is well described by DHAWAN et al. [16, pp. 1145-1153] Social media content is analysed using the Plutchik wheel. See below.

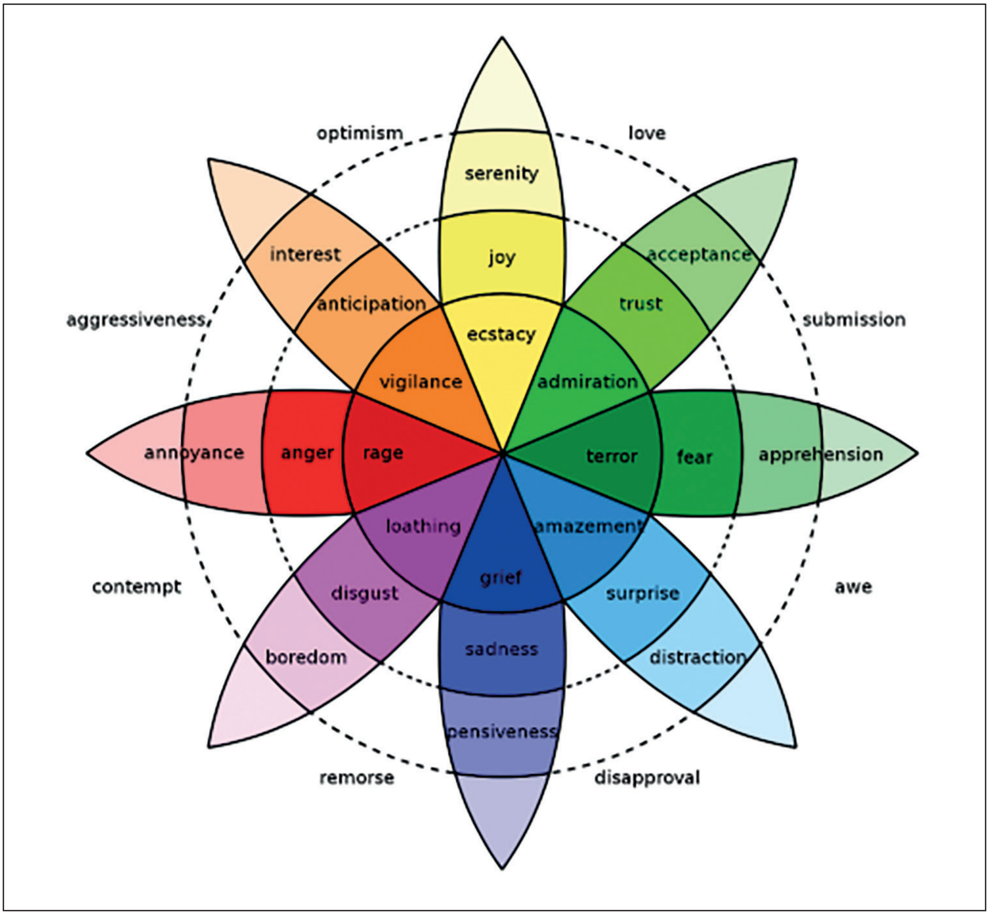

Figure 3. The Plutchik wheel. Source [14, p.17]

The boundaries of the technology

It is extremely difficult to detect sarcasm or irony. For instance, the word "terrific" can mean extreme good or horribly bad.

As to the evaluation of sentiment analysis results one would normally use precision and recall. It is, however, important to note, that it is far from obvious what can be considered a “good” result of a sentiment analysis solution. People are subjective when scoring texts from the point of view of attitudes. 10-20% of people see the sentiments in the very same text differently. [17] If a sentiment analysis program reaches this level, one can call it satisfactory.

Sentiment analysis is not an exact discipline. Still, one can assume that the error rate is more or less constant, so creating a time series of, for example, a cell phone customer review can reveal definite tendencies.

It would be considered normal or even sophisticated to use synonyms for the same entity or its aspects. The “phone” and “device” or “power usage” and “battery life” can only be identified as the same entity by using some kind of taxonomy.

Conclusion

In this paper I have reviewed the significance of sentiment analysis giving a quick insight into the architecture of such an application. Core elements have been briefly discussed. I have given a short description of the levels of sentiment analysis. A short overview of the field outside the English speaking world is given with special emphasis on the Hungarian literature. The application fields of sentiment analysis with particular emphasis on military, intelligence, law enforcement, marketing and pharmaceuticals will be described in a separate publication.

References

[1] MUNK, Sándor: Szemantika az informatikában. – *Hadmérnök*, 2014 (IX.)/2.

[2] FELDMAN, Ronen: Techniques and Applications for Sentiment Analysis, Communications of the ACM, April 2013, Vol. 56, No. 4, ISSN 0001 0782

[3] BING, Liu: Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012, Vol. 5. No.1. ISBN 978 160 845 88 44

[4] RUPPENHOFER, Josef; REHBEIN, Ines: Anchoring sentiment analysis in semantic frames, http://www.uni-hildesheim.de/ruppenhofer/pubs/longversion.pdf

[5] PANG, Bo; LEE, Lillian: Opinion Mining and Sentiment Analysis, Proceedings of the 42nd Annual Meeting on Association for Computer Linguistics, Article No. 271, Foundations and Trends in Information Retrieval, Volume 2, Issue 1-2,

[6] VIRMANI, Deepali; MALHOTRA, Vikrant; TYAGI, Ridhi: Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014

[7] ABDUL-MAGEED, M. et al: A System for Subjectivity and Sentiment Analysis of Arabic Social Media, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics

[8] BEREND, G.; FARKAS, R.: Opinion Mining in Hungarian based textual and graphical clues in Proceedings of the 4th International Symposium on Data Mining and Intelligent Information Processing, Spain, Santander, 2008

[9] SZASZKÓ, Sándor; SEBŐK, Péter, KÓCZY, László: Magyar szövegek véleményanalízise, http://www.inf.u-szeged.hu/projectdirs/mszny2009/MSZNY2009_press_b5_mod_opt.pdf (29.09.2015.)

[10] SZABÓ M.K., 2014. Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai. „Nyelv, kultúra, társadalom” című alkalmazott nyelvészeti konferencia, Budapest

[11] SZABÓ, M.K., VINCZE V. 2015. Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács A., Varga V., Vincze V. (eds) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). Szeged: Szegedi Tudományegyetem.

[12] Neticle Technologies http://www.neticle.hu/#press, http://www.neticle.hu/woi.html, (29.09.2015)

[13] Kereső Világ http://kereses.blog.hu/ (11.10.2015)

[14] EKMAN, Paul ed.: Emotion in the human Face, Malor Books, 2013, Los Altos, California, 2013, ISBN 978 933779 82 9

[15] TURNER, Johnathan H.; STETS, Jan E.: The Sociology of Emotions, Cambridge University Press, Cambridge, UK, 2005, ISBN 978 521 84745 2

[16] DHAWAN, Sanjeev et al.: Emotion Mining Techniques in Social Networking Sites, International Journal of Information & Computation Technology, Volume 4, Number 12, 2014, ISSN 0974 2239

[17] OGNEVA, Maria: How Companies Can Use Sentiment Analysis to Improve Their Business, http://mashable.com/2010/04/19/sentiment-analysis/#Gyr3NMpX6iqk (06.10.2015)

Szemantikus technológiák a hangulatelemzésben

VADÁSZ PÁL

A 2000-es évek eleje óta az NLP új ága nyert teret, a szemantikus technológiákat alkalmazó hangulatelemzés. A hangulatelemzést a szöveg három különböző szintjén lehet végrehajtani, dokumentum-, mondat-, valamint fragmentumszinten egyaránt. A szentimentlexikon létfontosságú része a hangulatelemzésnek. Még újabb terület az emócióelemzés, amely kiterjeszti a számítógépes nyelvészet felhasználási területét.