SACAM
A Model for Describing and Classifying Sentiment Analysis Methods

Aleksander Waloszek and Wojciech Waloszek
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland,

Keywords: Sentiment Analysis, Opinion Mining, Knowledge Management, Ontologies.

Abstract: In this paper we introduce SACAM — a model for describing and classifying sentiment analysis (SA) methods. The model focuses on the knowledge used during processing textual opinions. SACAM was designed to create informative descriptions of SA methods (or classes of SA methods) and is strongly integrated with its accompanying graphical notation suited for presenting the descriptions in diagrammatical form. The paper discusses applications of SACAM and shows directions of its further development.

1 INTRODUCTION

This paper presents a novel model of describing methods of sentiment analysis. Sentiment analysis is a very quickly evolving field of research, and focuses on assessing emotional attitudes expressed in textual opinions contained in various documents.

The model has been developed within a research project conducted by academia in cooperation with industry and placed in the field of sentiment analysis and knowledge management, representation and reasoning. The main goal of the project was to bring closer those two fields of research and to use formal ontologies in sentiment analysis.

Planning of project tasks and designing new methods of sentiment analysis involve frequent references to existing methods. To facilitate the task and to make its result easier to present and examine, we developed a novel model of describing and classifying sentiment analysis methods. The method is called SACAM, *Sentiment Analysis Content Awareness Model*.

The purpose of the model is to provide means for concise graphical description of sentiment analysis methods. The model is well-suited for describing both particular methods and classes of them. Graphical notation provided with the model is designed to underline the crucial aspects from the point of view of the managing knowledge during analysis: knowledge repositories used within the method, the methods of building or augmenting such repositories, stages and means of the processing.

Use of the model allowed us for easier navigation in the hard field of sentiment analysis, whose rapid evolvement results with at least several hundreds of notable papers appearing every year. It also permitted us to precisely pinpoint the area in which we place our future efforts, and to plan further actions within the project.

The paper is focused on presenting SACAM model, and while the next Section provides some essential information about sentiment analysis, the paper should not be treated as a survey in the field. This role is fulfilled by some excellent existing papers and books like (Pang and Lee, 2008), (Cambria et al., 2013) and (Liu, 2012).

The rest of the paper introduces the model, shows examples of its use, and discusses its potential applications and directions of development.

2 SENTIMENT ANALYSIS

This Section provides a short introduction to sentiment analysis field (Section 2.1) and the existing approaches to classifying sentiment analysis methods (Section 2.2). It draws a background for introducing the SACAM model.

2.1 Introduction to SA

Sentiment analysis (SA) evolved and separated itself from the fields of Natural Language Processing and Affective Computing in early 2000. The term itself

Waloszek A. and Waloszek W.
SACAM - A Model for Describing and Classifying Sentiment Analysis Methods.
DOI: 10.5220/0006199901960206
In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017)*, pages 196-206
ISBN: 978-989-758-220-2

appeared in 2003 in (Nasukawa and Yi, 2003), in the same year another prominent name for the domain *opinion mining* was proposed in (Dave et al., 2003). Both the terms *sentiment analysis* and *opinion mining* are used in the literature, along with many related, though a bit more specific, terms like *review mining*, *opinion affect analysis*, *sentiment mining* or *emotion analysis* (Liu, 2012).

Sentiment analysis from its beginning focuses on extracting information about users' emotional attitudes from large corpora of documents, especially from social media. In comparison with its ancestor fields (NLP and affective computing) the problems here are approached more directly and specifically. The researchers do not focus on creating methods for perfect understanding of texts being analyzed. The texts are very often being treated as bags-of-words exposing some features based on presence or absence of specific words (or their co-presence expressed as *n*-grams—pairs, triples, etc. of words). Also the emotions exposed in the texts are typically not identified very comprehensively. The usual outcome of the analysis is bipolar: emotions are identified as positive or negative (sometimes neutral).

The range of phenomena being analyzed is quite broad, and includes sentiments, emotions, evaluations, and attitudes towards products, services, organizations, persons, events, news etc. However, there exists no tool suitable for handling all those phenomena universally, and most of algorithms developed in the field focus on a single problem: specific type of text, like microblog entry, and specific object being evaluated, like a tablet or a mobile phone.

The strength of methods of sentiment analysis most frequently stems from their statistical character. For instance, presence of the word "*excellent*" in a text may be treated as a sign of the text bearing positive opinion. This rule, while in some (perhaps many) cases not true, when applied to a large corpora of texts may turn out to be feasible enough to positively contribute to extracted information about expressed opinions.

Attention drawn by the subject of sentiment analysis increased rapidly, from purely scientific interest, towards many applied methods. Currently most of the companies involved in business intelligence (like Microsoft or SAS) offer their own solutions for opinion mining. One of the reasons is very broad range of potential applications: sentiment analysis has been used for assessing sales volume (Liu et al., 2007), ranking sellers and products (McGlohon et al., 2010), prognosis of movie box office (Asur and Hubeman, 2010) or assessing attitudes of stock exchange investors ((Bollen et al., 2011) on the basis of tweets, (Bar-Haim et al. 2011) using posts in expert microblogs). Semantic analysis found its applications also in political debate (Tumasjan et al. 2010, Chen et al. 2010) to predict the results of presidential vote in the USA.

2.2 Classifying SA Methods

The field of sentiment analysis is very rich and many papers in the domain contain proposals of classification schemes for SA methods. Such proposals are most frequently presented in survey papers and books reviewing the field, and can be used to underpin some of the most important characteristics of methods being classified.

One of the most classic decompositions of the methods in the field was presented in (Feldman, 2013). The methods are classified along two dimensions. The first dimension is about granularity, i.e. the degree into which a method investigates the contents of a document. While not precisely distinguished in (Feldman, 2013), one can order those degrees into the following hierarchy:

- Document-level analysis,
- Sentence-level analysis,
- Aspect-level analysis.

Analysis at a document level is the most straightforward way of assessing sentiment. Methods at this level assign sentiment orientation to whole documents, most frequently in the bipolar form of positive/negative score. Sentence-level analysis consists in assigning orientation to subsequent sentences. Working at this level might be helpful in detecting mixed opinions about the object of sentiment, and is also useful when some special kinds of sentences should be treated in a special way (like, simply, filtering out some sentences, say sarcastic ones). Aspect-level analysis allows for assigning sentiment not only directly to the object being assessed but also to its "parts", known as aspects. Aspects need not to be necessarily physical parts of the object, they may also refer to its features (like "display quality"). Assessing at aspect level allows for assigning sentiment score to parts and features of the object and, consequently, allows to extract interesting information also from mixed opinions. At the end of such analysis user may be presented with more detailed report with score for each of the aspects.

Analysis-level dimension is augmented by the division of methods by the learning technique applied. We distinguish here *supervised* and *unsupervised* methods. In supervised learning we

assume that we have at our disposal a training set that has been already labeled (e.g. by a human expert). For instance in the training set we may have examples, each representing a single review, with features like *n*-grams and labels in the form of positive/negative. From this set, a machine learning algorithm derives rules of assigning each example (and new examples) to the distinguished classes. Many data mining methods are suitable for this purpose, like Decision Trees, Naïve Bayes Classifiers, Support Vector Machines etc.

In unsupervised methods we do not have any labeling ready. Instead the algorithm, according to some rules, has to extract some features or parts of the document. An example may be syntactic rules for finding phrases in Turney's method (see Section 3.2). This kind of learning usually requires the existence of some kind of sentiment lexicon, i.e. a set of reference words with assigned information about strength and polarization of the sentiment expressed by them.

The two mentioned dimensions are most frequently used in classification. Many other methods of classification are based on those dimensions and extend them by new ones. An example is (Cambria et al., 2013) whose authors (apart from adapting granularity dimension) propose two new dimensions: discourse, and conceptual.

Discourse dimension refers to the level of awareness of the structure of discourse presented throughout the document. Most of the methods ignore this structure, e.g. by treating all the phrases equally regardless of the context of their appearance and their position in the document. Simple mechanisms that exhibit some discourse awareness may, for instance, detect summaries presented at the beginning and at the end of the document, and treat phrases in these sections differently (e.g. with higher weight). Most advanced mechanisms should be able to detect citations, quoted opinions, sarcastic answers or examples, in order to properly interpret the sentiment of specific phrases of sentences, or maybe even just to exclude some kinds of them from further analysis.

Conceptual dimension refers to mechanisms used for extract sentiment, and order them in accordance to the ability of extracting the meaning of word and phrases. This conceptual dimension is also very much connected to the ability of interpreting context of the words and phrases being examined. While simple methods rely on a keyword list with sentiment explicitly assigned, more advanced methods may treat sentiment carried by a word or a phrase more cautiously, for example by assigning them probabilities of expressing positive and negative opinion. Even more advanced mechanisms may take into consideration co-occurrence of the terms and their position in the sentence (e.g. by considering also punctuation). Most advanced methods along this dimension can be equipped with a knowledge base about concepts in the domain of interest (like the knowledge about the construction of a mobile phone) and may be able to use this knowledge for in-depth analysis of the sentences. Quoting (Cambria et al., 2013): "Concept-based approaches can analyze multi-word expressions that don't explicitly convey emotion, but are related to concepts that do".

Classification schemes very often refer also to the particular mechanisms included in the method (Liu, 2012, Cambria et al., 2013, Feldman, 2013). The list of the mechanisms is more or less predetermined and often includes features like presence or absence of sentiment lexicon (set of word with assigned information about sentiment carried), aspect lexicon (expressions about feature of parts of object being assessed), ability to identify special sentences (like sarcastic ones, or objective sentences telling us about facts), ability to identify special entities (like proper names) etc.

Apart from dimensions and mechanisms (Feldman, 2013) also describes the general architecture of a sentiment analysis method, and treats this as a useful resource during classification. The architectural schema proposed by (Feldman, 2013; with minor changes mostly due to other graphical layout of the paper) is presented in Figure 1.

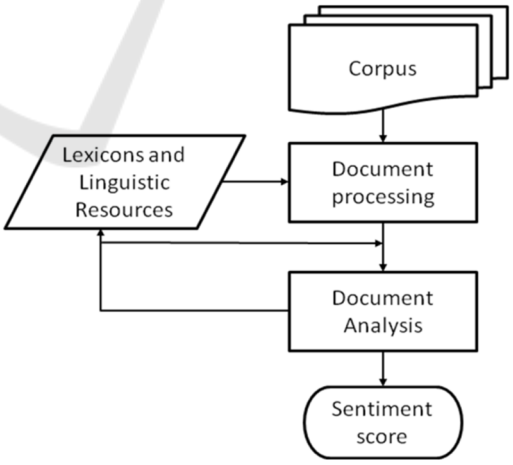

Figure 1: A general architecture of a sentiment analysis method on the basis of (Feldman, 2013).

Input for a method is a corpus of documents, which can be pre-processed in order to create or refine the contents of additional resources (e.g. lexicons). The resources are then used for proper document

analysis and to produce sentiment score. Important thing worth noting in the picture is the recurrent character of the flow, which may consist of stages whose execution might be repeated during the course of the method.

3 SACAM

This Section describes the SACAM model (Section 3.1) and gives an example of its use, based on one of the very widely known Turney's method (Turney, 2002) for sentiment analysis (Section 3.2).

3.1 General Rules of SACAM

The classification schemes described in Section 2.2 certainly underpin some of their interesting of SA methods being classified, and can be treated as base for creating their formal and abstract descriptions. Considering, however, the discrepancies between various classification schemes, and the fact that no survey or book proposed a detailed method of such formal description, we decided to build a new model, SACAM, on the top of the selected existing classification schemes.

Driven by the requirements of the project, mentioned in the Introduction, we strived for a model for describing sentiment analysis methods particularly focusing on processing knowledge during the analysis of textual opinions. All the methods of sentiment analysis base on knowledge about the content being processed. This knowledge can be subdivided into several areas: knowledge about grammatical structure of the texts being analyzed, knowledge about meaning of specific words, and knowledge about the domain of interest, i.e. the object being assessed. In our classification we wanted to focus on how the knowledge in these (and perhaps additional) areas is used, stored and processed. This is the reason why we called the new method *Sentiment Analysis Content Awareness Model* (SACAM).

What we wanted to achieve was the model for describing methods of sentiment analysis. The descriptions prepared in accordance to the model should give clear and immediate clues about the way of processing knowledge during the analysis, especially should answer the questions like: whether the knowledge in aforementioned areas is taken into consideration at all, is it statically programmed into method, or maybe prepared in some preliminary stages, is it expressed explicitly or is implicitly contained in some components. As it can be seen from this description, the structure of the process itself was also very important to us. Apart from these properties, we also expected the method to expose such characteristics as how much human expert work is needed.

Another very important requirement for the model was the ability to describe not only single specific methods of sentiment analysis, but also whole families of methods. In this way we could also obtain the tools for classifying methods and to compare them on the more aggregated basis.

Taking these requirements into consideration we developed a graphical notation being the core of the SACAM model. Using this notation one can prepare a diagram being a description for a method or a family of methods of sentiment analysis. The notation is based on standard block diagrams being used to depict business processes in an organization.
Each diagram consists of standard elements depicted in Figure 2. We drew inspiration here from the work (Feldman, 2013) which also borrowed such elements to describe the general process of sentiment analysis (see Figure 1). The meaning of each of the elements was slightly changed in order to express specifics of sentiment analysis.

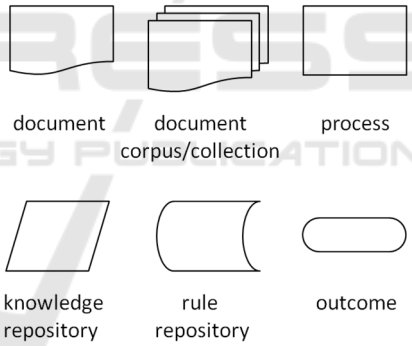

Figure 2: Basic elements of a SACAM diagram.

The elements from Figure 2 can be combined into a description of a flow. This flow depicts the main steps taken in the described method. The flow may be divided into several stages, differentiated by (not necessarily disjoint) frames. It should end with the terminal symbol, representing the final outcome of the algorithm.

The SACAM diagrams come in two flavors. A diagram may depict a single method (*particular diagram*), or a family of methods (*generic diagram*). In the first case, the elements of the diagram refer to repositories used and algorithms executed by the particular method. In the second case the elements represent more generic repositories and steps, and are a placeholder for filling by a more specific family (or

a specific method). An assumption is made here, that this is not necessary (for the more specific family or method) to fill all the placeholders.

Such an approach allows us to create a diagram that captures the broadest family of almost all typical methods of sentiment analysis. This general diagram (called a *root diagram*) is shown in Figure 3.

The root diagram shows the three kinds of knowledge repositories used by typical methods of sentiment analysis: repository of processing rules (including use of grammar, e.g. mechanisms like POS-taggers), repository of knowledge about sentiment words (sentiment lexicons) and repository of knowledge about the object of assessment (aspect lexicons). The three repositories are depicted by three vertical flows (diagram "columns").

Sentiment lexicons and aspect lexicons (SL, AL) are traditionally understood repositories, depicted by an appropriate symbol. For such repositories there often might be a special process of their building (learning stage). Most commonly some preliminary form of a lexicon is given (called seed lexicon; SSL, SAL). Seed lexicon is being iteratively extended during the process. The seed forms of lexicons are depicted in the first row of the diagram, and the iterative process of building the final form of the repositories is illustrated by the upper frame. The frame is "plural", which is indicated by additional incomplete rectangles extending beyond the main frame. The plural form of the frame indicates that the stage may be repeated a number of times.

The knowledge about processing is represented in the form of repository of rules (PR). It plays the major role in forming the main loop of an algorithm being described. This kind of knowledge is "operational", as it constitutes the flow in the process of generating the outcome (final assessment). However, this does not necessarily mean that the exact form of rules needs to be known in advance: the processing rules may be also generated or made more specific during the learning stages. As those rules form the low-level knowledge about processing documents in general, they might be also used (sometimes in a modified form) to generate sentiment and aspect lexicons (SPR, APR) and sometimes to refine the processing rules themselves (RPR; PR generally consists of RPR, SPR and RPR which is represented in the diagram with a dotted line). Processing rule very often embrace basic or more advanced knowledge about grammar, as many of the sentiment analysis

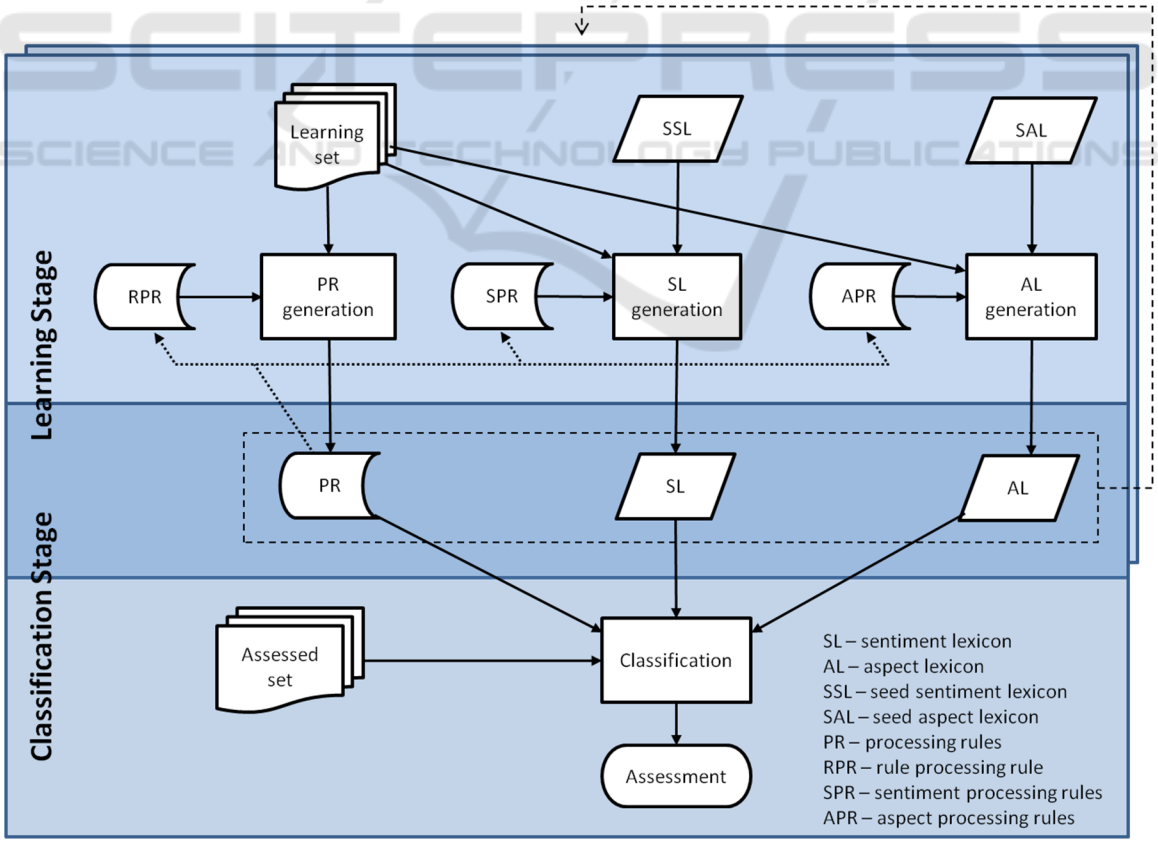

Figure 3: The root diagram of SACAM model, with the legend for abbreviations in the bottom right corner.

methods perform various actions on the basis of grammatical properties of encountered words or sequences of words (like Part-of-Speech, PoS).

Solid arrows show flow of knowledge in the method. Arrows directed towards processes show that a process uses a repository (documents), arrows directed outside processes show its outcome (final outcome or contribution to a repository). Dashed line show a feedback loop within a plural stage. As mentioned above, dotted line show that some repository may consists of other repositories.

Root diagram is a general SACAM diagram, which means it can be specialized to show a specific method, or a narrower family of methods. When specializing a general diagram, the more specific diagram reuses the subset of the elements from the more general one, and augments them with comments. The comments are placed within the elements and for the processes they should normally concern:

- If the process involves human work;
- If the process requires external applications and/or services;
- What algorithms are used.

While for repositories they should indicate:

- What does the repository contain;
- How the contents are represented.

A person creating the diagram naturally has some freedom in selecting the most important features to be described, but they should generally follow the main principle of SACAM: to depict the flow of knowledge. Therefore, all the information concerning knowledge acquisition, representation and processing should be a priority.

3.2 Creating a Particular SACAM Diagram for Turney's Method

In this Section we show how to construct a particular SACAM diagram, taking as an example one of the most widely known SA methods, Turney's method, described in (Turney, 2002). In general the method uses unsupervised learning for analyzing texts on the basis of syntactical patterns for expressing opinions.

The syntactical patterns used in the algorithm reflect the observation that different parts of speech have different influence on expressed opinions. The most important (i.e. those whose sentiment is mostly correlated with the overall sentiment) are adjectives and adverbs.

The syntactical patterns are used in Turney's method to find phrases of interest. The patterns (and phrases) consist of two consecutive words (sometimes with a third following word not included into the phrase) of appropriate part-of-speech. An example of a pattern expressed with tags (Penn Treebank Project, 2016) is: JJ NN/NNS, which means that it matches adjectives followed by nouns in plural or singular form.

The main algorithm accepts a set of reviews S and consists of three stages. In the first stage the algorithm looks for matching phrases from S. In the second stage sentiment orientation of each phrase is determined. For this task a measure called pointwise mutual information (PMI) is used, which, for two different phrases, is calculated with the equation:

$$\mathrm{PMI}(phrase_1, phrase_2) = \log_2 \frac{p(phrase_1 \wedge phrase_2)}{p(phrase_1) \cdot (phrase_2)} \quad (1)$$

where $p(phrase_1 \wedge phrase_2)$ is the probability of co-occurrence of the two phrases, where $p(phrase_1) \cdot (phrase_2)$ is the product of probabilities of occurrence of the two phrases (or also co-occurrence, if they are independent).

The algorithm uses this equation to determine "the distance" between the matched phrase ($phrase_1$) and the two predetermined phrases "poor" and "excellent". The sentiment orientation (SO) of the phrase is then calculated as:

$$\mathrm{SO}(phrase_1) = \mathrm{PMI}(phrase_1, \text{excellent}) - \mathrm{PMI}(phrase_1, \text{poor}) \quad (2)$$

which can be transformed into:

$$\mathrm{SO}(phrase_1) = \log_2 \frac{p(phrase_1 \wedge \text{excellent}) \cdot p(\text{poor})}{p(phrase_1 \wedge \text{poor}) \cdot p(\text{excellent})} \quad (3)$$

The probabilities of occurrence and co-occurrence of phrases are unknown. But in Turney's method they are estimated with clever use of Internet search engine (Turney used AltaVista): $p(phrase)$ is estimated as proportional to number of hits for the phrase, while $p(phrase_1 \wedge phrase_2)$ as proportional to number of hits for the query $phrase_1$ NEAR $phrase_2$. This allows for further transformation of the equation (3):

$$\mathrm{SO}(phrase_1) = \log_2 \frac{hits(phrase_1 \text{ NEAR excellent}) \cdot hits(\text{poor})}{hits(phrase_1 \text{ NEAR poor}) \cdot hits(\text{excellent})} \quad (4)$$

where *hits* denotes the number of hits returned by the search engine.

In the third and final stage SO of each review from the set S is determined as the average of SO of all phrases extracted from this review. Positive number indicates a positive sentiment orientation.

Fig. 4 contains the description of Turney's method in the form of a particular SACAM diagram (as a

specialization of the root diagram). The *primary features* of the diagram are expressed with its shape. Along the horizontal axis we can see that the Turney's method exploits two repositories: of grammatical (processing) rules and a sentiment lexicon.

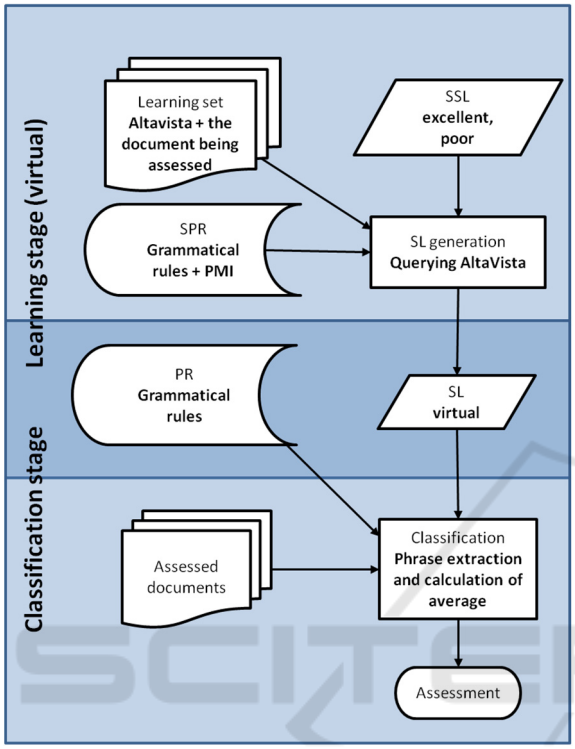

Figure 4: Turney's method described with a SACAM diagram.

Along the vertical axis we can see that the sentiment lexicon needs to be prepared, and is used for the proper assessment. The process of creation the sentiment lexicon is virtual, which means that the lexicon is not materialized, and the learning stage is in fact intertwined with the process of classification. Nevertheless, the knowledge about sentiment of specific words is very strongly present in Turney's method. In addition, it is based on large corpora of documents indexed by AltaVista, which is shown by the *secondary features*, i.e. features that are described in the comments.

4 APPLICATIONS OF SACAM

In this Section we present applications of SACAM, including already performed by us (Sections 4.1, 4.2, and 4.3) and further possible applications (Section 4.4).

4.1 Roadmap of SA Methods

One of the most straightforward uses of SACAM is to depict classes of existing SA methods. Within our project we undertook such a task and created general diagrams for top-level classes of methods, basing mainly on the division presented in (Liu, 2012).

Created diagrams, among others, depicted:

- Supervised and unsupervised SA methods;
- Methods involving detection of subjective sentences;
- Methods involving creation or use of sentiment lexicon,
- Methods involving creation or use of aspect lexicon.

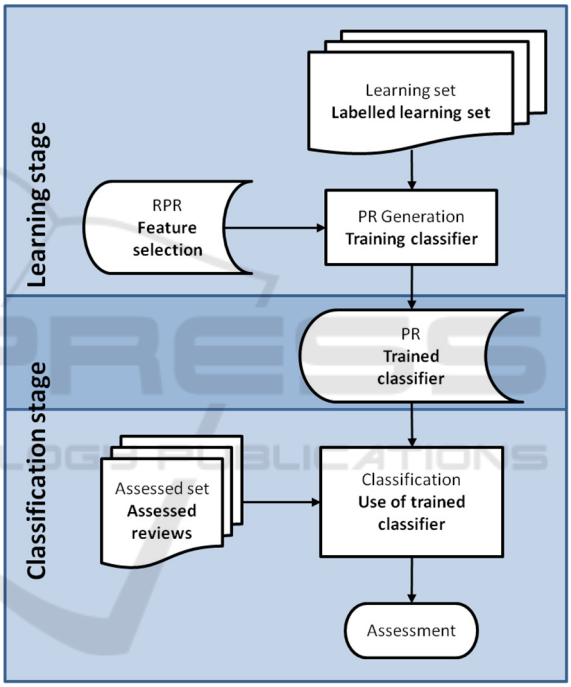

Figure 5: SACAM diagram for supervised SA methods.

Due to constrained space for the paper we only reproduce here diagrams for supervised methods (Figure 5) and for methods involving creation of sentiment lexicon (Figure 6).

As diagrams like those presented in the aforementioned figures tend to increase in size, we also introduced the notion of a partial SACAM diagram, which highlights some of the fragments of the flow within a method or a class of methods. An example of such diagram is shown in Figure 7, where a mechanism for creation of sentiment lexicon with use of WordNet (Miller, 1995) is highlighted.

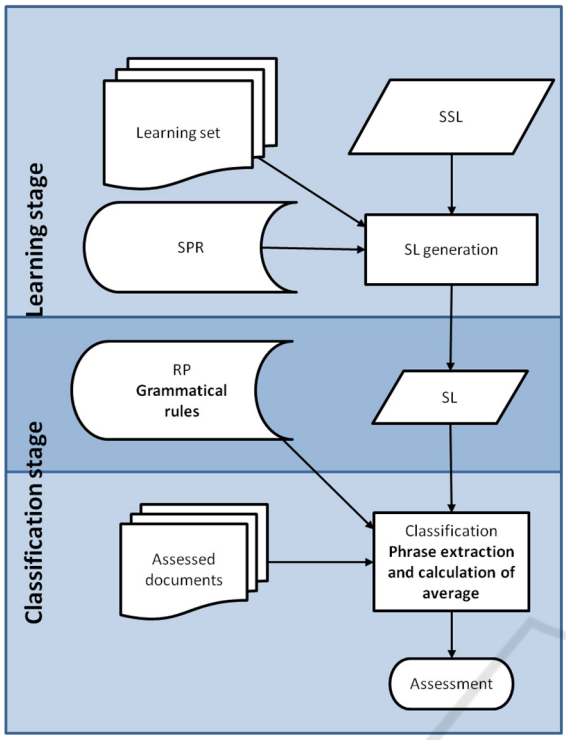

Figure 6: General SACAM diagram for unsupervised methods.

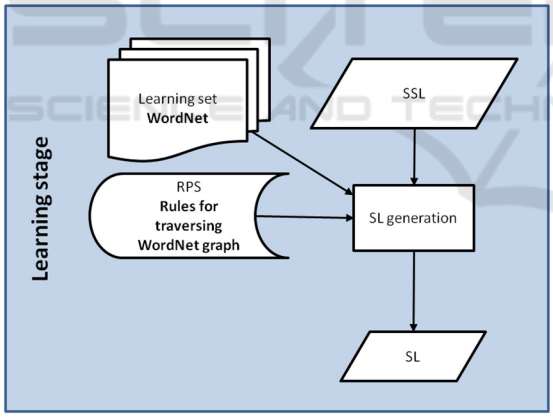

Figure 7: Partial SACAM diagram showing sentiment lexicon creation with use of WordNet.

Diagrams like this can be useful to create a general map of the field, depicting main classes of methods and pointing out similarities and differences between them. Individual methods can be matched against general diagrams. For instance, Turney's method (traditionally counted in unsupervised methods) fits almost perfectly into the schema for methods involving sentiment lexicon, obviously with the (very) notable exception that the lexicon in Turney's method is of purely virtual character.

Such a collection of general SACAM diagrams can easily be maintained and augmented with the appearance of new classes of SA methods.

4.2 Designing New SA Methods and Comparing SA Methods

Particular SACAM diagrams are also a very efficient tool of designing new SA methods and to compare (newly designed or existing) methods to each other. Consider an exemplary new method with the following characteristics. The method consists of three preliminary stages and two main steps. During the preliminary stages the data for proper classifications of reviews is prepared. We assume that a given set of reviews of a product (any product, but we may think of a mobile phone here) is available.

In the first preliminary stage, the topics of the assessed reviews are automatically predetermined. Latent Dirichlet Allocation (LDA; Blei et al. 2003), revealing words most characteristic for each topic, and in this way, indirectly, determine the set of relevant topics. In the second preliminary stage the collection of topics extracted during the first stage is reviewed by experts. The experts identify phrases which refer to qualitative aspects of a product (like "display quality", "memory size", "energy consumption") and identify the aspects as "positive" (like "display quality"), which means that larger amount of this quality is desirable, or "negative" (like "energy consumption"), which means that larger amount of this quality is undesirable. In the third preliminary stage the experts identify the words (intensity words) that indicate higher amount or intensity for each of aspects (like "high" for both "display quality" and "energy consumption" and "unacceptable" only for "energy consumption") along with words for lower intensity for each of aspects.

In the first main step, phrases that are combinations of intensity words and aspects are identified, and the sentiment orientation of each phrase is determined on the basis of whether the qualitative aspect itself is positive of negative and whether the intensity word indicates high or low intensity/amount of this aspect. In the second main step the assessment of a single document is calculated as a difference between the number of "positive" phrases and "negative" phrases.

These assumptions are sufficient to allow us to easily create a SACAM diagram for the method (see Figure 8). The method can be easily compared to other methods, like Turney's, by checking the primary and secondary features of the diagrams.

The differences can immediately be seen in primary features: the new method uses an aspect lexicon. The secondary features reveal more distinctions: primarily the amount of experts' work needed in the alternative method, and concrete (non-virtual) character of lexicons prepared there.

Naturally, there are also similarities between the two methods. Both of them use knowledge about sentiment carried by specific words (sentiment lexicons), both need the lexicons to be prepared, and both calculate the final score for the documents as an aggregated score of extracted phrases.

4.3 Assessing SA Methods

SACAM diagrams may be used as a basis for creating formal or semi-formal measures of fitness of particular methods or classes of methods for a specific task.

Within our project we used this approach to estimate the best directions for integrating ontological knowledge management with traditional sentiment analysis. To this end we created a semi-formal measure of semantic potential, whose value was determined on the basis of presence or absence of specific constructs in the SACAM diagrams.

While the detailed description of the measure is outside of the scope of this paper, the basic factors influencing the measure (positively) were the number of knowledge repositories, presence of procedures for building or augmenting the repositories, and interactions between repositories, i.e. using the contents of one of the repositories to refine the contents of another repository. The last effect is present, for instance, in the methods which build their aspects lexicon on the basis of identification of sentiment words (like e.g. (Blair-Goldensohn, 2008) or (Somasundaran and Wiebe, 2009)), which mechanism is depicted in SACAM in Figure 9.

Use of the aforementioned measure allowed us to direct our interest towards methods of this level of interaction between repositories. Ultimately, we chose a bit different course of actions and decided to use modular ontological knowledge bases to support the process of creating customized sentiment lexicons for each of the identified aspects. The measure, developed using SACAM, allowed us to make a more informed decision in this subject.

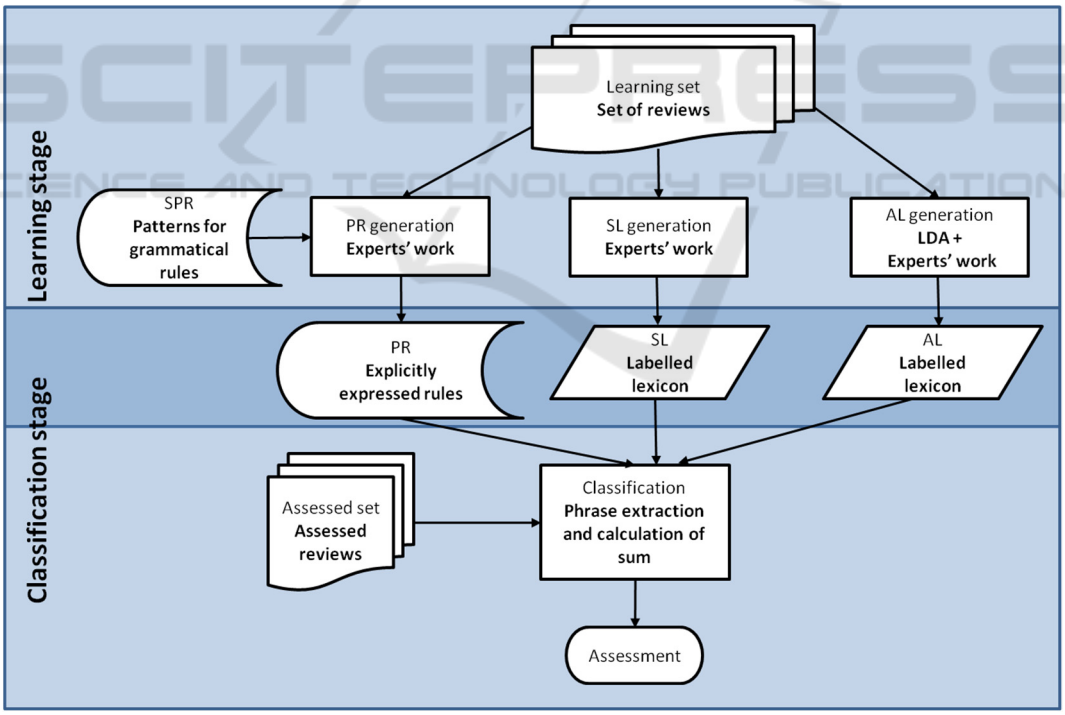

Figure 8: The root diagram of SACAM model, with the legend for abbreviations in the bottom right corner.

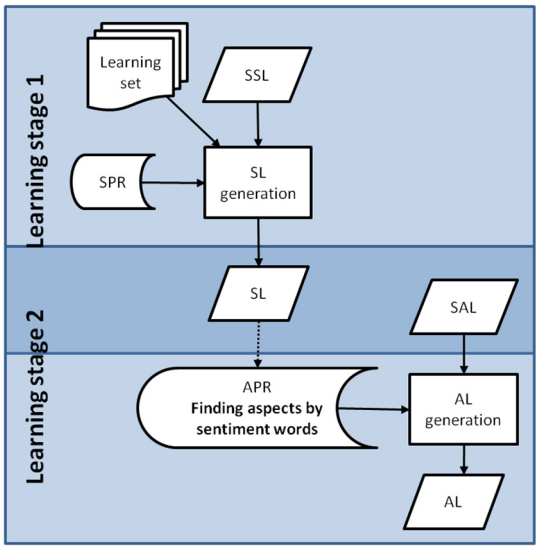

Figure 9: Partial SACAM diagram showing use of sentiment lexicon in building aspect lexicon.

4.4 Further Applications of SACAM

Another direction for development of SACAM consists in laying more solid foundations for its formalization. One of possible approaches here is to create an ontology for the model. Ontologies are formal specifications of conceptualizations in various fields (Gruber, 1993). As such, they constitute a very good instrument for formalizing all kinds of descriptions. The state-of-the art language for creating ontologies is OWL 2 (OWL 2, 2012).

Creation of a formal OWL ontology for SACAM would create numerous possibilities. Firstly, it would allow for automated verification and comparison of SACAM descriptions. Secondly, it would be possible to express in the ontology not only the descriptions of methods and classes of methods but, for example, the requirements for a method for a specific project. This would allow for developing measures for fitness of specific methods for specific tasks and, possibly, automated search of methods.

5 SUMMARY

In this paper we presented the SACAM model for describing methods and families of methods of sentiment analysis. The core of the model is a graphical notation used to depict the flow of knowledge during such analysis.

The graphical notation adopts the elements of block diagrams used for describing business process. Shape of each SACAM diagram expresses primary features of a method (or family of methods), which are the knowledge repositories used and the required stages of processing. The secondary features, like involvement of experts, way of knowledge representation, and use of specific tools for knowledge processing are revealed in the comments present in each block.

The applications of SACAM are numerous. Within our project SACAM proved itself a helpful tool. It has been successfully used to describe both various classes of methods and particular methods and to compare them to each other. Practitioners in the field may also find it a useful tool for designing new methods. On the basis of the presence or absence of some specific construct in the diagram one can derive measures of choice for more detailed assessment of SA methods and algorithms.

One of the most promising directions of use and development of the model is to formalize it with use of ontologies. It might lead to new uses of SACAM, like automated verification and assessment of sentiment analysis methods.

ACKNOWLEDGEMENTS

This work described in the paper was partially supported by the Polish National Centre for Research and Development (NCBiR) under Grant No. PBS3/B3/35/2015, project “Structuring and classification of Internet contents with prediction of its dynamics” (Polish title: “Strukturyzacja i klasyfikacja treści internetowych wraz z predykcją ich dynamiki”).

REFERENCES

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

Cambria, E., Schuller, B., Xia, Y., Havasi, C., 2013. New Avenues in Opinion Mining and Sentiment Analysis, in *IEEE Intelligent Systems* 28 (2), pp. 15–21, 2013.

Liu, B., 2012. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool.

Nasukawa, T., Yi, J., 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the KCAP-03*.

Dave, K., Lawrence, S., Pennock, D. M., 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of International Conference on World Wide Web (WWW-2003)*.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., Zhou, M., 2007.

Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*.

McGlohon, M., Glance, N., Reiter, Z., 2010. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)*.

Asur, S., Huberman, B. A., 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

Bollen, J., Mao, H., Zeng, X-J., 2011. Twitter mood predicts the stock market. In *Journal of Computational Science*.

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., Goldstein, G., 2011. Identifying and Following Expert Investors in Stock Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M., 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. Im *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)*.

Chen, B., Zhu, L., Kifer, D., Lee, D., 2010. What is an opinion about? Exploring political standpoints using opinion scoring model. In *Proceeedings of AAAI Conference on Artificial Intelligence (AAAI-2010)*.

Feldman, R., 2013. Techniques and Applications for Sentiment Analysis. In *Communications of the ACM*, vol. 56(4).

Turney, P. D., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002)*.

Penn Treebank Project, 2016. Alphabetical list of part-of-speech tags, https://www.ling.upenn.edu/, Accessed May 2016.

Miller, G. A., 1995: WordNet: A Lexical Database for English. In *Communications of the ACM* Vol. 38, No. 11: 39-41.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. In *The Journal of Machine Learning Research*.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., Reynar, J., 2008. Building a sentiment summarizer for local service reviews. In *Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era*.

Somasundaran, S., Wiebe, J., 2009. Recognizing stances in online debates. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*.

Gruber, T. R., 1993. A Translation Approach to Portable Ontologies. In *Knowledge Acquisition*, 5(2):199–220.

OWL 2, 2012. *OWL 2 Web Ontology Language Primer*, 2nd Edition, W3C Recommendation.