# Multidimensional Knowledge Representation of Text Analytics Results in Knowledge Bases

Johannes Zenkert and Madjid Fathi

University of Siegen
Institute of Knowledge Based Systems and Knowledge Management
Department of Electrical Engineering and Computer Science
Germany
johannes.zenkert@uni-siegen.de, fathi@informatik.uni-siegen.de

***Abstract*—In the age of digitization, intelligent systems have to cope with an ever-growing amount of data. Therefore, knowledge representation plays a key-role for applications to handle continuously created data and to enable an access on flexible and extensively well-described data structures. This paper introduces a knowledge base design which has the capability of dimensional structuring of semantically-related data and explains how text analytic results can be integrated into a knowledge base. The paper discusses the main advantages of this design and shows how the data can be arranged in the knowledge base. The multidimensional structure of the knowledge base helps to resolve one of the main challenges of knowledge discovery which is the extraction of meaningful information from data in a context.**

***Keywords*—Knowledge Base, Knowledge Representation, Text Analysis**

## I. Introduction

Nowadays, on the way to the Industry 4.0, it is very common in organizations to produce big data in a massive scale. Nearly everything observable in organizational processes is tracked, stored in databases and kept for further analysis. The concept of smart factories and innovative ideas towards the Internet of Things (IoT) will facilitate the data-driven decision making and real-time analysis. In future scenarios, structured data will enable production machines to communicate with each other and will result in more flexible production processes.

Since a large share of the created data in organizations is considered as being unstructured data, efficient ways to recognize, extract and store meaningful information are needed. One of the main challenges in knowledge discovery is the interpretation of data. Data itself can be meaningless, but through interpretation it can be considered as information or, moreover, even knowledge. As a consequence, unstructured data needs to be interpreted and disambiguated. Following this hypothesis, the creation of knowledge and information extraction from data can only be reached if the context and meta data of unstructured data is considered.

Intelligent systems with provided knowledge bases are seen as a possible solution for the disambiguation and interpretation task. An innovative approach for using unstructured textual resources in a knowledge base is to identify information for multidimensional structuring in each text resource itself. Dimensions can be considered as flexible categories, in which parts of the data are inserted and the identified relations are assigned with.

For example, the time information and the main keywords, extracted from documents, can assist in organizing a large collection of documents. In this basic example, the time and keywords can be considered as dimensional information. With a deeper analysis on the document collection and the text information within, more dimensional information can be identified, associations can be provided, relations can be created and as a result, a well-structured knowledge base for documents can be achieved. In this way, the querying of information on the knowledge base with multiple dimensions will provide more accurate search results.

In the aforementioned example, text mining methods are applied on documents and their textual content. In general, Natural Language Processing (NLP) provides the methodologies to create interpretable data from unstructured data for further analysis. Part-of-Speech (POS) Tagging and Named Entity Recognition (NER) are used to enable Sentiment Analysis, Topic Detection and the Concept of the Imitation of the Mental Ability of Word Association (CIMAWA) to further extract dimensional information from a textual resource. With the text analysis results, knowledge can be enriched by adding new relations and associations to existing or new dimensions in the knowledge base.

In the following, the extraction of meaningful information is explained and the used text mining methods are shortly described in Section 2. In Section 3, possible dimensions are introduced to provide a general overview of the analysis possibilities from the presented approach. Furthermore, the knowledge base design is described and the main benefits from the way of dimensional structuring are given in Section 4. In the conclusion, the paper results are summarized and future work in related area is mentioned.

## II. Extraction of meaningful information

For the extraction of meaningful information from textual resources, textual content need to be pre-processed by NLP operations. The operations should be handled by the intelligent system after the textual content is provided as data input. The language of the textual content, which is extremely important

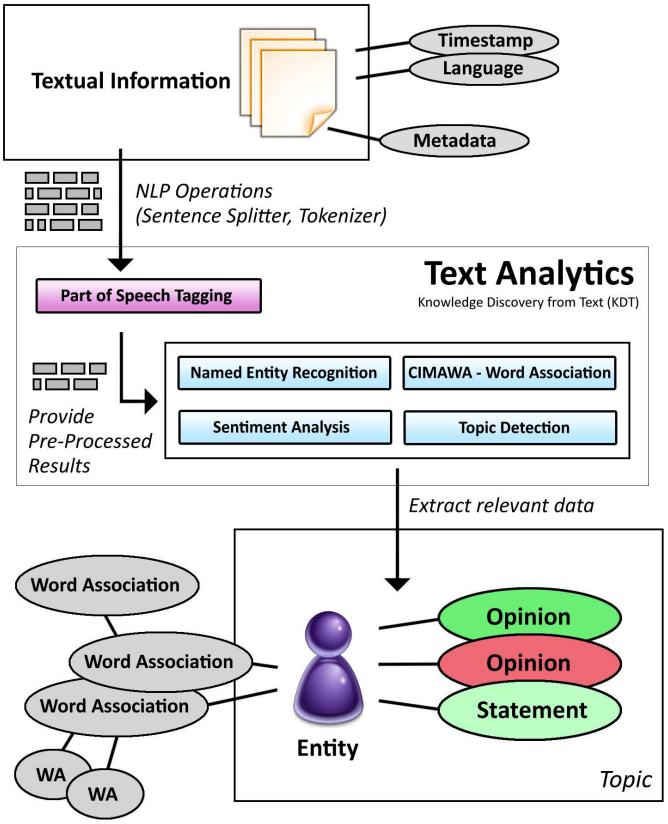


Fig. 1. Extraction of entity information from textual resources with different analysis results

to consider in the analysis, is often provided as meta data, especially within web data. If it is not available, in many cases it can be also derived by the frequent usage of stop words within the text. This approach is also considerable to identify language changes within the textual content. According to the identified language, language-specific models should be selected in the text analysis by the application. The most commonly used NLP pre-processing steps are sentence splitting and tokenizing. Both of them are also applied in the described approach as first steps.

After the pre-processing, different text mining methods can be applied in the extraction of entity information process. The complete process is depicted in Fig. 1. In the following sections, each step of the information extraction process is shortly mentioned with corresponding details.

### A. Pre-processing

Sentence splitting and tokenizing are used to transform documents in a machine-readable form. In the first step, the document is transformed into a list of sentences. Based on the layout of a text, the headline and paragraphs are separated from each other. Afterwards, the text is divided into smaller pieces. The actual content of a document, the document body, is further divided into sentences by the consideration of punctuation. In the next step, the identified sentences are processed with a tokenizer in order to create a list of tokens. Depending on the implementation of the application and further analysis methods, the list of tokens can also be represented internally as a vector of words. The extracted tokens and sentences are used as input for the POS Tagging in the next step. For this input, the punctuation is again very important to highlight and the POS Tagger needs to consider it in order to correctly annotate the POS Tag for each word in a sentence.

### B. Part-of-Speech Tagging

Many approaches for the POS Tagging have been developed in the past. Rules, Maximum Entropy Models and Stochastic Methods are proven methodology for the POS Tagging process [1][2][3]. POS Tagging analyzes lists of tokens from sentences and searches for a probable POS sequence. Tag sets are available in different variations, according to their tag granularity and language differences. For example, in English language, the Penn TreeBank [4] with 45 tags is in common usage, for German language the STTS Tagset [5] with 55 tags is a commonly used tag set. One option for the POS Tagging with stochastic methods is a Hidden Markov Model (HMM). Based on a trained model, the POS Tagger is used to compute the most likely POS sequence based on probability of the next tag. With the annotation labeling of each word in a sentence, grammatical analysis shows the structure of the sentence and entities can be resolved in the next step. Identified part of speech like adverbs and adjectives are also especially useful in the sentiment analysis, since they are considered to have high impact on sentiment intensity [6].

### C. Named Entity Recognition

Named Entity Recognition (NER) is used to identify entities of various types in textual information. Different approaches for NER have been developed in the past. Maximum Entropy Models, Transformation-based Learning, HMM and Robust Risk Minimization have been studied and their performance has been considered as highest [7]. Typical examples for entity types are persons (e.g. person names), organizations (e.g. companies, organizations), locations (e.g. cities, countries) and date or time expressions (e.g. months, years, time). Lists of person names or location names (gazetteers) can be also used in the application to identify entities based on keywords. However, those lists are not always useful, especially if the entities are entitled in variations within the textual content. In this case, NER is focusing on the provided POS tags and makes an estimation based on the probabilistic value of the tag for Named Entity (NE) in a sentence tag sequence.

### D. Sentiment Analysis

Sentiment Analysis is used to identify, if there is any positive, neutral or negative opinion, statement or subjectivity expressed from an entity about another entity. Methods and developments towards opinion-oriented information-seeking systems have been studied in the past [8]. Approaches in the litature are mainly focused on the usage of a corpus, dictionary, keywords or lexicon to decide if the existing textual resources contains terms with a polarity [9]. A well-known lexical

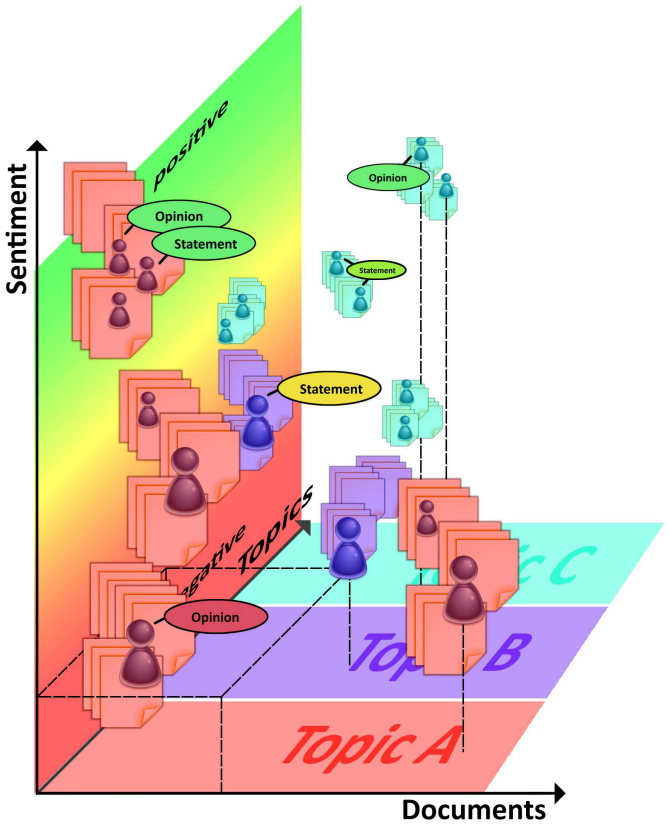


Fig. 2. Conceptual visualization of a three dimensional relation between entities, assigned documents and sentiment evaluation in corresponding topics

resource for opinion mining is SentiWordNet 3.0, which is in widespread usage by more than 300 research groups [10]. A semantic orientation-based approach for sentiment analysis has been developed recently in [11].

Sentiment analysis, as a method to extract information for the multidimensional knowledge base, is focused on the identified entities through NER and therefore searches for terms which highlight and express emotions about other entity information within the text. A classification of terms into types of emotion can be directly used for dimensional information in the knowledge base design of this research.

### E. Word Association

The hybrid word association measurement CIMAWA, introduced in [12], is used to create a numeric value of associative strength between entity keywords and other directly related words. In this way, the human association for each entity can be modeled and provides a list of description features for the knowledge base. By having these association profiles, entities can be compared with each other and the association strength can be frequently analyzed in order to see temporal changes in the knowledge base. Furthermore, word associations can be used in knowledge representation as well as visualization to define distances between entities and dimensions based on their derived numeric value from strength of word association.

### F. Topic Detection

For topic detection, various approaches have been developed in the past. Different levels for the topic detection have been considered. Whole documents have been assigned to topics based on classification, bag-of-words or keyword existence. In Latent Dirichlet Allocation (LDA), each document or textual resource is considered as a mixture of various topics. LDA is generative probabilistic model for collections of discrete data and has been introduced in [13].

Latent Semantic Indexing (LSI) is an indexing and retrieval technique which has been indroduced in [14]. LSI uses Singular Value Decomposition (SVD) to identify patterns in relationships between terms and concepts contained in an unstructured collection of text [14]. LSI has been applied in many researches and applications.

Based on the strengths of word associations, gained from the CIMAWA calculations, the concept of Associative Gravity [15] can also be used for identification of multi-topic structures in textual resources. Associative Gravity utilizes word associations to detect different topics in a text.

## III. Dimensions

Similar to the Semantic Web[1] and the Resource Description Framework (RDF)[2], information from different analyses results should be kept in the knowledge base in a multidimensional structure. With the stated information representation styles as an archetype, the principle of semantic relationship should be considered and modeled in the knowledge base as well. Facts and extracted information, which are normally stored into the knowledge base, will be provided in the form of a triple and represented as the combination of a subject, a predicate and an object. However, these parts are further assigned into their dimensional affiliation.

With the dimensional structuring approach, a combination of different data can be extracted as one information from the knowledge base because the dimensions are considered in the extraction process and offer the necessary context. In this way, queries that request data from different dimensions provide more detailed results and intelligent output from the knowledge base. As an example, opinions from entities about specific topics can be identified within textual resources (e.g. news, social media), analyzed and stored in the knowledge base. By querying the knowledge base, the analysis provides opinions which can be compared from two different dimensional perspectives. Fig. 2 shows the conceptual visualization of one possible scenario.

Further dimensional information can add beneficial analysis potential. In general, dimensions are derived from meta data or are extracted from the textual content with previously mentioned text analysis methods. In the following sections, possible dimensions are shortly introduced. Moreover, interesting application scenarios are given for further explanation of the practical usage from the introduced knowledge base design.

[1]https://www.w3.org/standards/semanticweb/
[2]https://www.w3.org/TR/rdf11-concepts/

### A. Meta data information

Meta data information can be directly used for specific dimensions in the multidimensional knowledge base. Therefore, unstructured data need to be analyzed in search for these information. Various document formats are available and have different opportunities to use provided meta data from them. As an example, in web data, a document created with Hypertext Markup Language (HTML[3]) declares the Document Type Definition (DTD). This meta data information is available in XML format and often used to specify all related properties (e.g. description, author, language, keywords) of the document.

In the comparison of documents, meta data is a useful filter possibility which should be captured and attached to the document itself, corresponding entities, contained topics and other extracted information. Examples of valuable meta data information for the usage in dimensions are given in the following paragraphs.

*1) Time:* Time-stamps are available in the description header of documents. This information can be directly used for the time dimension of extracted information. A distinction must be made in the time dimension between the the date of creation, the date of last access and the date of last modification. By having this separation, the relevance of the knowledge can be derived from the last access or last modification depending on the use case. Parts from the content of the knowledge base, which are never being used could be archived. Newer information and information with a high frequency of usage must be directly kept in the knowledge base and prepared for fast access to avoid long loading duration.

*2) Language:* Language is often a given meta data property on many textual resources. In the description header of documents, the language property is specified and can be used as dimensional information. In this way, the knowledge base can be divided by language specific dimensions. By the availability of different text analysis methods, which are trained on different languages, information from documents can be extracted and assigned to the corresponding language dimension.

*3) Organization Structure:* Knowledge bases in enterprises should cover all information from customers, suppliers, stakeholders, products and business processes to support the overall business strategy. Based on different units or organizational structure, the knowledge base can be modified to cover these specific information for each unit separately through dimensions. In this way, the structure of the knowledge base is directly related to the organizational structure. As an example, each extracted information from textual resources which is relevant only to the marketing division of the enterprise can be assigned to a marketing dimension.

*4) Access Restriction:* In enterprise knowledge bases, a dimensional information for the description of access restrictions is essential. Information needs to be organized in a secure and reliable way. By the utilization of dimensions, a distinction between private and public information can be added easily. Therefore, public information should be labeled as publicly available information in the knowledge base. Furthermore, different user rights can be assigned directly to the dimension in the knowledge base.

*5) Personal Information:* One of the dimensions in a knowledge base should be usable for personal preference and needs. Selected information which needs to be accessed many times should be labeled as highly relevant for the personal work. In this way, time consuming information searches can be avoided by having all related information already prepared in the knowledge base.

### B. Textual Analysis Results

Entity, sentiment, topic and associative content are the introduced dimensions which are derived from the analysis results of the previously mentioned extraction of information process. Examples of analyses results for the usage in dimensions are given in the following paragraphs.

*1) Entity:* Entities which are recognized by NER are considered as a central unit in the knowledge base. Attached and related information for each entity is separated in different dimensions. In this way, a selection of the entity dimension allows a comparison of entities. As an example, different products, persons or other entity types could be directly compared with each other. The entity dimension provides huge analysis potential.

*2) Sentiment:* Extracted textual phrases may contain opinions and emotions which can be classified as positive, neutral or negative. Each expressed opinion should have at least one related entity. Based on the relations from one entity, that mentions an opinion about another entity, a relational connection between the dimensions in the knowledge base is achieved. The results from sentiment analysis have huge analysis potential through the utilization of the multidimensional design of the knowledge base. In the aforementioned organizational use case, the enterprise knowledge base can provide statistically solid answers to interesting market research questions with the analysis of customer feedback to specific products.

*3) Topic:* Identified topics can be assigned to different other textual content in the knowledge base. In this way, a query on the knowledge base that summarizes information related to specific topics is easy to realize. Since topics may change in the knowledge base during the time, temporal analysis can be used for trend analysis. In the use case of the enterprise knowledge base, customer feedback for each specific product is a multidimensional search result and can provide interesting insights for product improvements.

*4) Associative Content:* Based on the associative content dimension, the textual information can be mapped into the knowledge base. As an example, based on association strength from one word to another word, a numeric value from CIMAWA can be used to specify distances between words. As a result, a knowledge map of associated content can be created. By adding further dimensions to the knowledge map further analysis potential is provided. The dimension of associative content can be also used as a basis for enterprise content

[3]https://www.w3.org/TR/html/

management and innovative associative search methods. Given some keywords in a search method, the associative search can directly provide all related content, based on the multi-dimensional mapping in the knowledge base. With additional dimensional filtering, the relevant content can be identified and algorithms for searching the right content in Enterprise Content Management (ECM) Systems can be improved.

## IV. Multidimensional knowledge base design

In many organizations, data is stored in decentralized databases. Different applications have their own databases and don't provide a universal interface to use them as central or distributed knowledge base. Following this approach in the IT strategy has disadvantages. Data is not updated consistently and may have decisive or unnecessary gaps, if companies don't use a uniform database schema. In contrast, if a central knowledge base is provided for the organization, the analysis potential and decision-support for the management and decision makers is increasing with data volume. The introduced knowledge base design is useful for organizational data, as explained with aforementioned uses cases, but is also able to store a wide range of other knowledge from other use cases.

The knowledge base uses the concept of relations between dimensions. In this way, not only the associations or relations can be represented, but also each strength of relation between the represented features in specific dimension can be described.

The characteristics scalability, flexibility, dimensionality and relevance can be considered advantageous in the introduced knowledge base design. In the following sections, these characteristics are described shortly.

### A. Design characteristics

*1) Scalability:* For the scalability of the multidimensional knowledge base, the key question is how knowledge can be represented and visualized. The characteristic has influence on the whole knowledge base design. The integrated data is represented in a relational way between different dimensions. With the relations between data entries, scaling can be directly modeled in the knowledge base. Scaling of data has the requirement for different representation levels to summarize data or specify data in more details. This means, in the knowledge representation, an entity can be represented internally as one entity. However the entity is connected to different other entities through dimensions.

Similar to the main functionality of Online Analytical Processing (OLAP), the knowledge base design provides possibilities to consolidate data (known from roll-up and drill-down) by showing related entities and their information, extract dimensional information (slicing) and search for specific knowledge through multidimensional information (dicing). These possibilities are very beneficial for decision making.

*2) Flexibility:* The multidimensional knowledge base design pays special attention to the various forms of data inputs. Based on the concept for the representation of entity aspects in dimensions, more dimensions can be added to the knowledge base easily to further specify and describe existing knowledge. In addition to the qualitative and quantitative characteristics of data, different formats of unstructured data other than text (e.g. image, audio, video formats) need to be considered for the integration into the knowledge base. The design of the knowledge base allows to integrate different data formats and respects the increasing size of data volume through dimensional structure.

*3) Dimensionality:* The basis for the structure of the knowledge base are dimensions. Dimensions are considered as fully customizable in the knowledge base design. New dimensions can be added to the knowledge base by the creation of new relationships between the object representations in the knowledge base. In this way, the design uses the advantage of a schema-less structure, which adapts automatically to new data input. By the usage of the previously described text mining methods, unstructured text information can be transferred into a dimensional-structured form. Adding and removing additional dimensions can filter results and considerably increases analysis potential.

*4) Relevance:* The relevance of data can be determined by the interpretation of data in a context (e.g. query on the knowledge base for data across multiple dimensions). By consideration of multiple dimensions, contextual information helps to decided whether (new) data must be kept in the knowledge base, must be deleted or even rejected as irrelevant. The relevance characteristic of the knowledge base is also influenced by the included time dimension. New data, or recently modified data is stored for fast access, long-term data is archived accordingly. Moreover, the relevance aspect is largely decisive for the quality assurance of the knowledge base and contained data. The relevance consideration is from high importance for enterprises which often facing the problem to decide whether their information is still up-to-date or reliable.

### B. Continuous update cycle for the knowledge base

The continuous updating process of the knowledge base is visualized in Fig. 3. For the creation of the knowledge base, a first information needs to be created in dimensional structure. Also textual resources need to be provided for applications to perform the analyses. The text corpus is used as an input for training of the text analysis methods for information extraction. At the beginning and initialization of the knowledge base, a user needs to distinguish between correct and incorrect suggestions by the system. The effort of the user and the distinction between correctly and incorrectly extracted information from textual resources can be measured by precision and recall calculations.

This means, the first knowledge in the knowledge base can be a single piece of information. After the insertion of the data in a dimensional context, this relation can be used and extended by the system. As an example, if the knowledge base starts with the information "Berlin is the capital city of Germany", other extracted knowledge can be added as relations to the entities "Berlin" and "Germany" into the

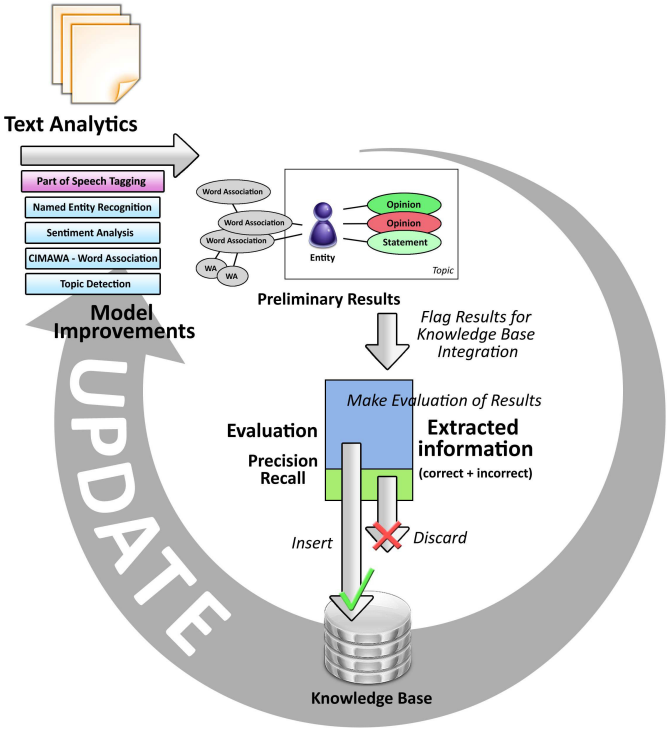


Fig. 3. Update process of the multidimensional knowledge base

knowledge base. In this example, these entities will be easily recognized by location gazetteers from NER. The relations between the added entities don't need to be fully specified since updating the relations and the dimensional structure is always possible in the provided knowledge base design. At this step, the two expressions "Berlin" and "Germany" are also considered as dimensions. Since the word association of "capital" and "city" will push the system in different directions and will add unknown relations and dimensions to the existing knowledge base, the numeric word association strengths will directly provide distance values between the directions and help with contextual differentiation.

Updates and integration of new data into the knowledge base will result in a decreasing effort for users, because more training data will provide better re-trained models as inputs for the applied text mining methods. This hypothesis is true until a better information extraction accuracy is not worth the decreasing system performance from additional rules or exceptions anymore. However, in this knowledge base design with a continuous integration of new data with assigned relations, the over-fitting and over-optimization of models can hardly be avoided. Nevertheless, the optimization process is considered as a specialization in a domain or expertise field.

## V. Conclusion

In this paper, we have presented a new design for knowledge bases, which offers advantages for the integration of extracted information from unstructured textual data into the knowledge base. With an enormous flexibility, the knowledge base design can be used in different use cases and application domains. By the utilization of a dimensional structure, the knowledge representation and knowledge discovery in the knowledge base can be facilitated by the central hypothesis, which is interpretation of data provided contextual information. In addition, the interpretation and analysis potential increases with the data volume and completion of the overall dimensional structure.

With the integration of new extracted information from textual content, the analysis methods can be improved and will result in a lower effort in decision making about the integration of new information in the knowledge base. In future work, the knowledge base design will be implemented and made accessible to applications with the possibility to offer different text mining methods. In this way, the updating process, the frequency of model re-training for text analytics will be further analyzed and the decreasing effort will be shown in a use case.

## References

[1] E. Brill. A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 112-116, 1992.

[2] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Proceedings of the conference on empirical methods in natural language processing, 1, 133-142, 1996.

[3] T. Brants. TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing (ANLC '00). Association for Computational Linguistics, Stroudsburg, PA, USA, 224-231, 2000.

[4] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330, 1993.

[5] A. Schiller, S. Teufel, and C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Manuscript, Universities of Stuttgart and Tübingen, 66, 1995.

[6] F. Benamara, C. Cesarano, A. Picariello, D. R. Recupero, and V. S. Subrahmanian. Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. In ICWSM, 2007.

[7] E. F. Tjong Kim Sang, and Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-147, 2003.

[8] B. Pang and L. Lillian. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1-2 (January 2008), 1-135, 2008.

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

[10] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC, 10, 2200-2204, 2010.

[11] B. Agarwal, and N. Mittal. Semantic Orientation-Based Approach for Sentiment Analysis. In Prominent Feature Extraction for Sentiment Analysis (pp. 77-88), Springer International Publishing, 2016.

[12] P. Uhr, A. Klahold, and M. Fathi. Imitation of the human ability of word association. International Journal of Soft Computing and Software Engineering (JSCSE), 3(3), 2013.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022, 2003.

[14] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88), J. J. O'Hare (Ed.). ACM, New York, NY, USA, 281-285, 1988.

[15] A. Klahold, P. Uhr, F. Ansari, and M. Fathi. Using Word Association to Detect Multitopic Structures in Text Documents, IEEE Intelligent Systems, vol. 29, no. 5, pp. 40-46, Sept.-Oct. 2014.