Computers in Industry 78 (2016) 96–107

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Review

Text analytics in industry: Challenges, desiderata and trends

Ashwin Ittoo [a,*], Le Minh Nguyen [b,c,d,*], Antal van den Bosch [e]

[a] Department of Operations, University of Liège, Liège, Belgium
[b] Division of Data Science, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[c] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[d] School of Information Science, JAIST, Japan
[e] Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 27 November 2015
Accepted 1 December 2015
Available online 30 December 2015

Keywords:
Text analytics
Text data mining
Natural language processing
Industrial applications
Trends

ABSTRACT

The recent decades have witnessed an unprecedented expansion in the volume of unstructured data in digital textual formats. Companies are now starting to recognize the potential economic value lying untapped in their text data repositories and sources, including external ones, such as social media platforms, and internal ones, such as safety reports and other company-specific document collections. Information extracted from these textual data sources is valuable for a range of enterprise application and for informed decision making. In this article we provide a systematic review of the current state of the art in the application of text analytics in industry. Our review is structured along three dimensions: the application context, the methods and techniques utilized, and the evaluation procedure. Based on the review, we identify the different challenges and constraints that an real-world, industrial environment imposes on text analytics techniques, as opposed to their deployment in more controlled, research environments. In addition, we formulate a set of desiderata that text analytics techniques should satisfy in order to alleviate these challenges and to ensure their successful deployment in industry. Furthermore, we discuss future trends in text analytics and their potential application in industry.

Contents

* Corresponding authors.
E-mail addresses: ashwin.ittoo@ulg.ac.be (A. Ittoo), nguyenleminh@tdt.edu.vn, nguyenml@jaist.ac.jp (L.M. Nguyen).

http://dx.doi.org/10.1016/j.compind.2015.12.001
0166-3615/

1. Introduction

In recent years, the world has witnessed a surge in the volume of data in unstructured, textual format. This proliferation can be attributed to several factors, including the ubiquity and continuous growth of ICT and social media platforms. For instance, society at large can now express itself via messages posted online; similarly, individual users can describe their positive or negative experiences with products or services in dedicated forums.

This rapid expansion in text data has also been witnessed in the enterprise ICT realm, which has evolved from a landscape dominated by structured data, organized in relational database management systems and data warehouses, to one dominated by semi-structured and unstructured data, including text data. Buried within these huge volumes of texts are valuable information nuggets, which, if correctly identified and extracted, can be exploited for informed decision-making and for supporting a wide range of enterprise activities. These include mining customer opinions (sentiments) on products in order to enhance customer satisfaction and improve product quality, and increasing the efficiency of certain tasks to optimize workflows, such as document classification. The need to mine useful information from these textual data sources led to the emergence of text mining techniques, which can be considered as an extension of classical data mining (knowledge discovery in databases) intended for traditional structured and unstructured non-textual data.

A recent report of the McKinsey Global Institute [1] highlights the economic potential of text (and data) mining techniques. For instance, a better exploitation of data using appropriate data and text mining techniques would enable US Health Care to create more than $300 billion in value. Similarly, the efficient use of insights garnered from data to improve operations and to detect frauds could generate up to $250 million of potential annual value to Europe's public sector administration.

Thus, text mining is poised to play an increasingly important role in industry in the coming years. In fact, many companies have acquired a competitive advantage by exploiting text mining technologies. A case in point is Netflix: according to its director of Streaming Science and Algorithms, one of the most interesting ways to exploit the huge volumes of data collected from members is in the use of text mining to analyze these data to improve the actual quality of the movies and shows [2,3]. In a similar vein, the IBM Watson Question-Answering system [4–6] has opened up new opportunities to improve the quality of service in various domains and industries, including health care and finance.

Over the years, text mining techniques have evolved in sophistication, incorporating more and more elements originating from computational linguistics research. This evolution has nuanced the distinction between text mining and other related research disciplines, such as natural language processing. Consequently, in this article, we use the term *text analytics* (TA) to refer to both text mining and natural language processing (NLP), unless otherwise stated. Another motivation for preferring TA stems from it being more popular in industry than TM or NLP.

Our aim with this article is to provide an overview of industrial TA applications. Specifically, we will review several state-of-the-art TA techniques, developed and evaluated within the context of academic research, and applied to address real-life problems in industry. Thus, an important point to note is that our article is neither a review of commercial solutions, such as TA packages from vendors like SAS or IBM, nor a review of current TA/NLP research efforts, such as those published in scientific outlets as the ACL conferences and journals.[1]

Our article is structured as follows. In Section 2 we give examples of popular companies that are developing or adopting TA solutions for various applications. Then, in Section 3, we present our actual review and describe state-of-the-art TA techniques developed in the scientific community and applied to solve real-world industrial problems. We present each technique systematically by structuring our review along three dimensions: Overall Context, Methods and Techniques, and Evaluation. Based on our review in Section 3 we identify a set of challenges and constraints that industrial applications pose to TA techniques in Section 4. Next, in Section 5 we formulate a set of desiderata that TA techniques should satisfy in order to ensure their efficient deployment within industrial environments. Finally, Section 6 discusses future trends. We discuss current innovations in the field of TA, such as Deep Learning, Word Embeddings, and TA and Vision, and their applications in industrial contexts to solve real-world problems.

A note on the terminology: as mentioned before, we will use the term *text analytics* (TA) to refer to text mining (TM) and natural language processing (NLP). Also, we will use the term *technique* and *algorithm* interchangeably, wherever deemed more appropriate.

2. Popular text analytics applications in industry

Several well-known companies are adopting TA to support and improve their core activities. We give 3 examples, which were chosen to reflect the added value that TA brings to diverse application domains.

Netflix: Netflix makes heavy use of TA techniques to analyze feedback and comments that its members post on various online sources, including social media networks. The information mined from these sources are then used for a plethora of activities. For instance, information on members' preferences is useful in recommendation systems and in providing more personalized contents. This information also provides valuable information on contents that need to be replaced as they fail to satisfy the demands of the members, and on inaccuracies in captions and subtitles. Thus, at a more strategic level, TA

[1] https://aclweb.org/anthology/.

techniques are at the core of powerful models that enable Netflix to deliver better and enriched contents, closing the loop on quality and enhancing the members' overall experience [2].
Bank of England: The bank's Advanced Analytics division is adopting TA techniques to address several potential applications across central banking and to build more agile and wide-ranging data analysis capabilities for the future. In one exemplar application, they searched for tweets containing phrases or terms that could indicate that depositors were preparing to withdraw their money from Scottish financial institutions in the run up and immediate aftermath of the Scottish independence referendum in September 2014 [7].
IBM Watson Question-Answering System: The Watson question-answering system is often touted as a flagship TA technology from IBM. In essence, a question-answering system (QA) system takes as input questions expressed in natural language and looks up or reasons about the answers based on relevant information extracted from various sources. The input questions range from simple definition or "what is" questions, such as "what is the capital of Belgium?" to more complex formulations, such as riddles or "why" and "how" questions. Health care is often considered as one of the most promising application domains for QA systems [5]. For instance, Watson's QA capabilities have been augmented with speech recognition technologies for use as an aid to medical diagnosis in a joint project involving IBM, Nuance Communications, Columbia University Medical Center, and University of Maryland University of Medicine [5]. Also in the healthcare domain, Wellpoint insurers and Memorial Sloan-Kettering Cancer Center (MSKCC) relied on Watson technology to read and interpret massive volumes of data in text format, including medical literature, patients' treatments and family histories and clinical trials. Information from these text documents were used to assist oncologist to recommend courses of treatments [8]. According to IBM, at a more general level, Watson and its TA capabilities will change the paradigm in which we work with computers and will transform many industries. Watson's ability to ingest, process and understand large volumes of data faster than human counterparts extends its applications over a whole range of information-sensitive industries, such as in the legal domain where lawyers have to sift through large piles of evidence materials as well as in the insurance and banking domains, which require processing massive amounts of text data in the form of claims, emails and financial statements and reports [9–11].

The aforementioned applications tend to rely on commercial TA solutions, often marketed as part of enterprise application software. As mentioned earlier, our aim with this article is to provide an overview of state-of-the-art TA techniques, developed in the context of scientific research, and applied to address real-world, industrial problems. This will be the theme of the next section.

3. Text analytics research and applications in industry

Our review of the current state-of-the-art will be based on the scientific articles published in this special issue of the Computers in Industry journal[2] on text analytics in industry, namely [12–18]. Our rationale for focusing on these articles are as follows.

Recency: by virtue of their recency, these articles reflect the current state-of-the-art and trends in the field. Thus, they serve as an up-to-date reference for grounding our analysis.
Heterogeneity of Application Domain: the TA techniques presented in the articles are applied to a range of diverse domains, from classical application areas such as health care/bio-medical to more novel ones such as road traffic and aviation. Thus, they highlight the versatility of TA techniques and demonstrate their potential and the promising role they are expected to play in virtually all industry segments.
Scientifically Sound Methods Applied to Industry: the applications presented in the articles illustrate how scientific techniques could be applied to address real-life, industrial problems. Thus, they reflect a good mix of "theory and practice".
Quality/Journal Reputation: the Computers in Industry journal has a good reputation in publishing original, high-quality, application-oriented research papers that demonstrate the industrial use of new or existing technologies in knowledge intensive domains.

We will start in Section 3.1 by reviewing the TA techniques and applications in each article individually.[3] To facilitate our discussion, each review will be structured along three main dimensions, viz:

1. *Overall context*: providing general information (for e.g. application domain).
2. *Methods and techniques*: describing the main methods and techniques employed as well as their implementations (for e.g. machine learning and natural language processing libraries or toolkits that were used).
3. *Evaluation*: describing the experiments performed to assess the performance of the proposed techniques.

Afterwards, in Section 3.2, we provide a more general review, encompassing the different TA techniques, and we highlight the salient aspects across them.

3.1. Text analytics in industrial applications review

3.1.1. A distributional approach to open questions in market research

3.1.1.1. Overall context. This article proposes a system called The Klugator Engine (TKE), for analyzing responses to open-ended survey questions. The responses are expressed in freely formed natural language texts in English and German. The proposed system was developed in a collaborative effort between academic and industrial partners. The system is already in commercial use as the core engine of the Rogator Text Clustering Solution (RogTCS), distributed by German market research company Rogator AG.

3.1.1.2. Methods and techniques. The proposed system starts with the classical step of pre-processing, in which the input texts (documents) are successively analyzed by different components of a standard (NLP) pipeline. This involves sentence splitting and tokenization with a customized version of NTLK [19], and word stemming with the Snowball stemmer [20].

Documents are represented according to the basic bag-of-words (BoW) model, yielding a term-document matrix (vectors). To alleviate the issue of sparsity in the matrix, Singular Value Decomposition (SVD) [21] is applied in order to identify the most important matrix "dimensions".

Then, clusters of similar documents are identified using the traditional *K*-Means algorithms. Similarity between documents is computed as the cosine angle between their respective vectors. The number of clusters to be discovered, i.e. the parameter *K*, is set to 20, a value found to yield adequately informative clusters for the domain.

The sentences are also assigned a sentiment score, ranging from −1 (extremely negative) to +1 (extremely positive), reflecting their polarity, with a simplified version of the SentiKlue [22] system.

[2] http://www.journals.elsevier.com/computers-in-industry/.

[3] The articles are not presented in any specific order.

Finally, the results of the text analysis are displayed in the form of a semantic map. In this map, clusters of similar documents are depicted as nodes, whose area is proportional to the clusters' size. Node are labeled with corresponding topics, which are salient words identified from the documents using the log-likelihood ratio [23]. The nodes are colored according to the overall sentiment of the documents in the clusters. Clusters with documents that expressed a positive sentiment are displayed as green nodes. Conversely, red nodes denoted a negative sentiment, while yellow is used for neutral sentiment. Furthermore, the nodes were positioned such that their distances reflected the similarity between their topics.

3.1.1.3. Evaluation. The evaluation focuses on the systems' two main components, namely, text clustering and sentiment analysis. Evaluation proved to be a rather challenging process given the lack of clearly defined gold-standards in industrial applications, which is further compounded by the high level of subjectivity involved in tasks such as clustering and sentiment analysis. The experiments show that even human judges found it difficult to assess and agree on the correctness of the clusters and sentiments.

Concerning the text clustering task, the performance is evaluated using the standard purity measure on different "configurations". The best configuration was one in which the 20 clusters generated automatically by *K*-Means were refined by human intervention (for e.g. merging and splitting clusters), resulting in 17 clusters, with an overall purity score of 0.66.

Concerning the evaluation of the sentiment analysis task, the system favors precision over recall. The rationale in doing so is to ensure that a sentence is labeled as positive or negative only if its polarity could be inferred (as positive or negative) with high confidence (i.e. precision), and to avoid giving misleading impressions on the sentiment polarity. The best precision scores reported are 0.79 (for positive polarity) and 0.90 (for negative polarity). It is also worth noting that the system achieves an overall F1-score of 0.69, comparing relatively well with the F1-score of 0.76 of a human judge.

3.1.2. Analysing and evaluating the task of automatic tweet generation: knowledge to business

3.1.2.1. Overall context. This article deals with the issue of automatic tweet generation in English and Spanish. An interesting aspect is that tweet generation is addressed not from a natural language generation (NLG) perspective, but rather as a text summarization problem. In essence, text summarization techniques are applied to news articles, yielding ultra-concise summaries, which then serve as tweets. A potential application of these automatically generated tweets is in supporting the marketing and communication activities of companies. Therefore, it is important to assess the tweets' quality, especially taking into account the fact that they are created without human intervention.

The tweets' quality is assessed in terms of 2 dimensions, viz:

1. *Interestingness*: to what extent users would be interested in knowing more about the tweets' contents.
2. *Informativeness*: whether the tweets accurately reflected the news articles from which they were generated.

Furthermore, this study also attempted at identifying linguistic features contributing to the interestingness and informativeness of tweets.

3.1.2.2. Methods and techniques. Seven state-of-the-art text summarization techniques, described in [24], were applied to 201 news articles (101 in English and 100 in Spanish), resulting into 1407 summaries, which were considered as tweets. In addition, 201 tweets were also created manually, i.e. by manually summarizing 201 news articles. The tweets were processed using FreeLing [25], an TA toolkit for English and Spanish, to extract various types of linguistic features from their contents, including:

- Lexical features, such as Parts-of-Speech (POS) and number of words
- Syntactic features, such as noun phrases and verb phrases
- Semantic features, such as named entities (e.g. persons, locations)

3.1.2.3. Evaluation. The tweets' quality (interestingness, informativeness) was evaluated manually by English and Spanish native speakers. These judges were not informed on the tweets' provenance, i.e. whether they had been automatically or manually generated. They were, however, given the news articles from which the tweets had been generated. Each tweet was evaluated by the judges on a 3-level Likert scale for their interestingness and informativeness. These scores were subsequently used to calculate a mean interestingness and mean informativeness score.

An analysis of the evaluation results revealed that for the English tweets, there was no significant difference in the informativeness score of automatic and manual tweets, i.e. users considered the automatically generated tweets to be as informative as their manually generated counterparts. Conversely, for Spanish, users considered the manual tweets as being more informative.

As for the interestingness, users perceived the automatically generated tweets in English as being more interesting than the manually created ones. Conversely, for Spanish, manually generated tweets were considered as more interesting.

With regards to the linguistic features, syntactic and semantic features were found to contribute the most to the quality of English tweets. However, for Spanish tweets, no particular features that were archetypal of good quality tweets could be identified.

3.1.3. A methodology for traffic-related twitter messages interpretation

3.1.3.1. Overall context. This article presents a system for automatically interpreting traffic-related Twitter messages in the Portuguese language. Interpretation in this case refers to the transformation of the tweets' unstructured textual contents into a more structured formalism, namely as RDF-triples.

The system is deployed as a prototype for monitoring the truck fleet of a gas transportation and fuel transportation company. Several benefits are reported as a result of the system's application in these companies, such as cost reduction, more efficient fleet management, and improved customer satisfaction due to a more accurate prediction of delivery time.

3.1.3.2. Methods and techniques. An ontology, called TEDO (Traffic Event Domain Ontology), specific to the traffic domain was manually constructed. It is a relatively simple ontology, comprising of nine classes, seven object properties, and eight datatypes properties (relations). Example concepts in TEDO are Location and WeatherCondition, while example relations include "causes" and "hasPrimaryLocation". The ontology was subsequently used to convert tweets into corresponding RDF triples.

The first step of this conversion was to preprocess the tweets' textual contents using a standard pipeline, which involved tokenizing the texts and annotating the tokens with morpho-syntactic information, such as POS and stems. These pre-processing tasks were executed using the F-EXT web-service [26].

The tokens are then annotated with particular tags using named entity recognition, implemented with a sequential minimal

optimization support vector machine (SMO SVM) [27] from the Weka machine learning library [28]. Example tags in the application include "interdiction", "collision". Each tag corresponds to one ontology term. Furthermore, the coordinates of entities tagged as Location in the tweets are determined using the SmartGeocode algorithm [29] for geocoding.

The next task in the interpretation of tweets is that of detecting relations between entities. It involves inferring the dependency tree between the entities and learning the possible relations that could exist between them using a large margin structured perceptron [30].

Finally, the tweets are converted into RDF-triples. This step relied on the tagged tweets, the TEDO ontology (each tag/entity in a tweet corresponded to an ontology term), and on the relations between the entities in order to simplify the translation of the natural language texts into RDF-triples.

3.1.3.3. Evaluation. The evaluation is performed on 690 Portuguese tweets describing traffic-related events, and focuses on the tasks of named entity recognition and relation extraction. The respective performances of these 2 latter tasks, measured with the F1 metric, are both around 0.7.

3.1.4. Text classification based filters for a domain-specific search engine

3.1.4.1. Overall context. The system presented in this article attempts to automatically predict filters for refining the search results of domain-specific search-engines. The targeted domain is that of job search. Example filters for this domain are the "region" to which a job advertisement is applicable, or the job type (full-time or part-time). The issue of predicting filters is addressed as a classification task, where the aim is to assign the documents to predefined categories (e.g. "full-time", "part-time"), which actually act as filters. The proposed system for filter prediction has been deployed the in Kimeta application,[4] a real-life (domain-specific) search-engine for job searches in the German language.

3.1.4.2. Methods and techniques. Filter prediction (document categorization) is performed with the CENFA system [31], which incorporates an ensemble of support vector machines (SVM) in an active learning configuration. First, a base classifier attempts to predict the filters to be assigned to the documents. Those documents that could not be conclusively categorized, based on a threshold score, are categorized by human annotators (domain experts). These manually categorized documents are then used to train another classifier ensemble, known as the specialized classifier. This iterative procedure of manually annotating ambiguous documents, and then using them to train the specialized classifier is repeated until the final classification results are deemed to be satisfactory by human experts. In this active learning approach, the base classifier is trained only once, while the specialized classifier is trained iteratively but with a smaller number of documents in each subsequent iterations. Such a configuration considerably reduces the overall training time.

3.1.4.3. Evaluation. The system is evaluated by measuring its performance in predicting 3 specific filters, namely "R&D", "Full-Time" and "Service/Customer Support", on a corpus of 300 documents. The performance metrics used were the (i) area under curve (AUC), which provides a more robust estimate when the distribution of documents per category is imbalanced, and (ii) weighted average precision since the aim of a search-engine was to provide results with the highest accuracy and recall was considered as less important. The reported AUC achieved by the system in predicting document categories ranges from 0.7 to 0.8 (for different categories), while the mean average precision ranges from 0.8 to 0.85. The experiments also show that the active learning approach is useful in gradually reducing the size of the training set, and thus, the training time. The performance of the system also increased during each iteration of the active learning process.

3.1.5. Natural language processing for aviation safety reports: from classification to interactive analysis

3.1.5.1. Overall context. In this article, a system for automatically analyzing safety reports in the aviation domain is proposed. Reports are expressed in French and in English. Analysis in this application involves classifying the reports according to safety-related categories, defined in the ADREP.[5] The categorized reports can then be used for various activities, e.g. they can be queried and the evolution of existing safety events (categories) can be monitored over time. However, text categorization does not enable the identification of emerging trends and topics. For this task, the proposed system relies on topic modeling. In addition, the system also incorporates an information retrieval (IR) tool that enables users to explore the report corpus interactively by searching for reports describing similar incidents.

The system is developed by a joint academia-industrial collaboration. Some modules have been deployed for real-life usage by the French Direction Gnrale de lAviation Civile[6] and are at an advanced testing phase at a French airline company before their potential integration in their safety management system.

3.1.5.2. Methods and techniques. Analysis of the aviation safety reports starts with a pre-processing phase. During pre-processing, terminological variations (e.g. "take-off", "T/O") are conflated to a single representation (e.g. "take off"). This normalization is achieved using a rule-based approach. Then, words are stemmed with the Snowball stemmer [20] and annotated with their respective POS tags and lemmatized using TreeTagger [32]. Relevant words are identified using the term frequency-inverse document frequency score (TF-IDF).

As mentioned earlier, the categories used for classifying the reports are from the ADREP taxonomy, which, with 37 categories, offers a fine-grained scheme for incident classification. In the proposed system, the reports classification is performed using SVM, implemented in the Liblinear Java Library [33]. Features include word stems, character n-grams, and stem-grams. One of the main challenges in the classification task is that a single report could belong to multiple categories. This issue is addressed by training one independent binary SVM per ADREP category (i.e. 37 in all) to predict whether a report belongs to a particular category or not.

The reports are also analyzed in order to detect latent topics from their contents. This is achieved using topic modeling with Gibbs sampling, implemented as part of the Gensim library [34]. The number of topics to be discovered is set to 50, chosen as a value as it enables a satisfactory interpretation of the discovered topics.

The final analysis task is that of information retrieval, implemented in the form of a search engine, which enables users to identify reports describing similar safety events. The reports are represented using the traditional BoW model as vectors, and similarity is computed as the cosine of the angle between vectors.

[4] http://www.kimeta.de/.

[5] Accident/incident Data REPorting; http://www.icao.int/safety/airnavigation/aig/pages/adrep-taxonomies.aspx.

[6] DGAC, French Civil Aviation Authority, http://www.developpement-durable.gouv.fr/-Secteur-Aerien,1633-.html.

Search results, i.e. similar documents, are plotted along a temporal axis on a 2-D scatter-plot for easier visualization. Reports are displayed a dots on the plot; the higher the dots are on the plot, the more similarity they exhibit with the input query/source report. Such a temporal display is useful to monitor whether incidents are isolated or constitute an emerging threat, and to determine the risk level associated with incidents.

The search-engine solution is used together with the text categorization module in an active learning configuration. In this configuration, human experts start with a rough estimation of the reports belonging to a particular incident category. This initial set of reports can be obtained via the search engine, and is used to train an SVM. Borderline cases, i.e. reports that could not be conclusively classified by the SVM, are presented to the experts for manual classification, and are used to train the SVM. This iterative process of manually annotating borderline cases and using them to train the classifier is repeated until the experts are satisfied with the results.

3.1.5.3. Evaluation. The different components of the system are evaluated separately. The performance on report classification is estimated using 10-fold cross-validation over a corpus of 136,681 reports. The overall F1-score, computed over all the categories, is 0.79. As expected, some variations are observed in the performance across different categories, with higher performance for narrowly defined categories and lower performance for loosely defined ones.

Concerning the topic modeling task, 15 representative words for each of the 50 topics generated were presented to a domain expert. The latter then assessed whether each of the 15 word-set (per topic) accurately described a unique theme (topic). Based on this evaluation procedure, 43 (out of the 50) word-set could indeed be assigned to a particular theme. As another means to evaluate the topics' quality, the Pearson correlation statistics is computed between the 15 word-sets and the meta-data of corresponding reports. The results are mitigated with a high degree of correlation of $\sim$0.7 for some topics, and a much lower correlation of $\sim$0.11 for others.

3.1.6. Integrating a semantic-based retrieval agent into case-based reasoning systems: a case study of an online bookstore

3.1.6.1. Overall context. The system presented in this article employs Case-Based-Reasoning (CBR) to enhance the search experience of end users in business-to-consumer websites. Specifically, it accepts queries as inputs, which could be formulated either as natural language questions or as keywords. It then consults a case base consisting of target problems (e.g. questions) and their corresponding solutions (answers). The best matching solution in the case is then returned as answer to the input query.

3.1.6.2. Methods and techniques. A two-step mechanism is used to find matching cases in response to an input query. First, techniques based on recognizing textual entailment (RTE) determine which cases in the case-base are most representative of the input case. If a representative case is found, based on a threshold score, its solution is returned in response to the input. Else, Short-Text Semantic Similarity (STSS) techniques are used to identify potentially matching cases, which may provide a solution to the target problem.

Similarity between the input query and the cases in the case-base is computed by various semantic similarity measures, such as the PATH method [35] and other traditional WordNet-based measures such as [36] and [37].

3.1.6.3. Evaluation. Two types of evaluations are performed, namely (1) a scientific evaluation against standard baselines and on standard corpora, as typically done in TA research, and (2) a case-study evaluation in an online bookstore.

For the scientific evaluation, the performance of the RTE component was assessed over the Microsoft Paraphrase Corpus [38], consisting of 4076 training and 1725 testing sentence pairs, with each pair labeled as 0 or 1 to indicate whether one is a paraphrase of the other. The proposed system achieves an average similarity score (between sentences) of 0.73, comparing well to other benchmarks such as [39] and [40]. The performance on the STSS task was assessed over the corpus presented in Li et al. [41], consisting of 65 sentence pairs. The proposed system achieves a score of 0.83, as measured using the Pearson correlation coefficient. Thus, it outperforms other STSS approaches, such as [42] and [43]. It should be noted that among all the similarity measures evaluated in the proposed system, the best performance for both the RTE and STSS tasks is achieved by the PATH method [35].

Concerning the application-type evaluation, the system's performance is assessed over a case-base consisting of 1200 books from the Amazon online bookstore. Queries to the case-base were then issued, and the performance was measured using the mean average precision (MAP). This experiment shows that the proposed system achieves an MAP of 0.93, markedly outperforming two baselines against which the system was compared.

3.1.7. Turning user generated health-related content into actionable knowledge through text analytics services

3.1.7.1. Overall context. This article proposes a text analytics system for pharmacovigilance. In essence, the proposed system extracts information pertaining to Adverse Drug Reactions (ADR) and their relationships to medications from Spanish social media websites.

3.1.7.2. Methods and techniques. The system mines ADR information from two main social media sources, namely:

1. Twitter: tweets were harvested based on specific keywords in Spanish
2. Saluspot[7]: an online forum that allows public users to make inquiries on health-related issues and receive answers from medical practitioners.

Texts from these sources are harvested and stored in a data warehouse, implemented using ElasticSearch.[8] These texts are then pre-processed by lemmatization and POS tagging using relevant APIs from MeaningCloud.[9] Word lemmas and POS are employed in a rule-based approach to word sense disambiguation. Next, relevant entities are identified from the texts with NER (which the article refers to as Topic Extraction). The NER step relies on several tools and resources, including APIs from MeaningCloud, the annotation pipeline from GATE [44] and domain-specific dictionaries, such as the Spanish-DrugEffectDB and UMLS-SNOWMED. Named entities in the domain typically correspond to drugs, effects, and diseases. Finally, the semantic relationships between these entities are labeled as one of three types, viz, (1) "adverse effect", (2) "indication" and (3) "possible". The first two relationships are detected based on the SpanishDrugEffectDB, while the 3rd type corresponds to undocumented relationships.

3.1.7.3. Evaluation. The different components of the system are evaluated separately. For the NER task pertaining to the detection of drugs, the main source of false negatives is abbreviated drug

[7] http://www.saluspot.com.
[8] http://elasticsearch.org/.
[9] https://www.meaningcloud.com/.

names, while false positives are mostly adjectives that are also used as drug names (for e.g. "sedatives"). The reported F1-score is 0.76. A baseline NER that relies on a gazetteer populated from domain-specific dictionaries achieves a much lower F1-score of 0.43. Concerning the NER task pertaining to the detection of an effect, the F1-score of the proposed system is estimated at 0.6. In this case, false negatives are mostly due to colloquial expressions (for e.g. "my head is ringing"), while false positives are mostly due to ambiguous terms. With regards to the relation extraction task, the system achieves an F1-score of 0.72; false negatives are mainly attributed to the limited coverage of dictionaries for labeling the relations, while false positives are due to the difficulties in identifying relations from complex sentences.

3.2. Overall review

After having reviewed the applications individually, we now provide a more overall review, encompassing the different applications and highlighting their common characteristics along the following dimensions:

1. Methods, techniques and applications;
2. Implementation;
3. Evaluation and metrics.

We focus on these dimensions as they are commonly explicated across the different applications described in the previous section. Also, these dimensions correspond to core issues that require consideration when developing and implementing text analytics (TA) applications, both in the scientific and corporate realms.

3.2.1. Methods, techniques and applications

The articles we reviewed cover a wide range of application domains, ranging from document classification to traffic monitoring. We can distinguish three main groups of applications. First are those aimed at alleviating human effort in tedious and time-consuming tasks, thereby enabling human experts to focus on more critical or value-added activities. A typical example of such applications is that of text classification, as presented in [15,16]. Second are those applications that aim at facilitating and enhancing the quality of the information-seeking process of end users and the results, such as the CBR application that accepts natural language question from users [17] and the information retrieval component of the application in [16]. Finally, applications in our third group are those that assist in decision making. These include the traffic monitoring application developed in [14] for fleet management. Naturally, such a wide set of applications calls for an equally diverse set of TA techniques (and methods), as will be briefly described next.

3.2.1.1. Text classification. Two of the applications rely on *text classification* techniques for:

1. Categorizing incident reports in the aviation sector according to the types of safety events they described (i.e. safety categories) [16].
2. Categorizing job offers according to the jobs' characteristics (e.g. part-time, full-time) to support a domain-specific search engine [15].

In both applications, support vector machines are used in an active learning configuration in order to enhance the classification performance.

3.2.1.2. Text similarity. Methods for estimating *text similarity* constitute another set of techniques at the core of many applications in the articles reviewed. These techniques measure the degree of similarity between texts at the lexical level as well as at the semantic level. In all the applications surveyed, similarity at the lexical level is computed by projecting the texts (documents) into a multi-dimensional (vector) space and computing the cosine angle between the corresponding document vectors, commonly referred to as cosine similarity. For instance, the IR tool of [16] relies on the cosine similarity measure to retrieve similar documents in response to a query. Likewise, the *K-Means clustering* algorithm of [12] identifies groups of similar documents based on their cosine similarity. Similarity at the semantic level is estimated using the traditional WordNet-based measures [17], i.e. the distance (e.g. path length) separating two words in the WordNet lexico-semantic dictionary.

3.2.1.3. Text summarization. Various techniques for text summarization were employed in [13].

3.2.1.4. Sentiment analysis, topic modeling. Our review reveals that more recent TA trends are also being deployed in industrial applications. In particular, *sentiment analysis* is employed in [12] to determine the polarity of survey response, and *topic modeling* is used in [16] to automatically infer topics from aviation safety reports.

3.2.2. Implementation

The various applications reviewed are implemented following a modular architecture; i.e. the application is decomposed into a sequence of tasks and each task is deployed as a separate module. In essence, the modules constituted a pipeline (commonly known as the "NLP Pipeline"), composed of upstream and downstream tasks. Upstream tasks are typically concerned with pre-processing, such as sentence splitting, tokenization, stemming or lemmatization, NER and PoS tagging. The aim of these pre-processing tasks is to make the input texts more amenable to further analysis by downstream tasks. Downstream tasks often corresponded to the actual application itself, for e.g. text classification or sentiment analysis. Such a modular approach is often adopted in commercial software engineering projects, whereby the different functionalities of the final software are developed separately by potentially disparate development teams. A modular implementation brings about numerous benefits. Most notably, for industrial applications, modularization facilitates maintenance; selected modules of the application can be replaced with minimal disruption to the overall functionality and without the need to re-write and re-test substantial portions of the source-code.

From our survey, we also noted that the various industrial applications leverage extensively upon open-source tools (and technologies) to implement the functionalities of the different modules. For e.g. the sentence splitter of the NLTK toolkit [19] is employed in the text clustering and sentiment analysis application of [12]. The application for analyzing aviation safety reports presented in [16] relies on the TreeTagger toolkit [32] for PoS-tagging and lemmatization, while the Freeling toolkit [25] was used in [13] to process tweets. The application for processing health messages on social media in [18] made extensive use of the annotation pipeline from the GATE toolkit [44]. To classify texts, [16] and [14] respectively use the SVM implementation provided by the Liblinear [33] and Weka [28] libraries. These open-source tools are more traditionally associated with the development of proof-of-concepts and experimental systems in the research community.

Another interesting observation is the exploitation of webservices for performing various tasks. For e.g. [14] relied on the F-EXT webservices [26] for processing tweets, while [18] relied on webservices from MeaningCloud to process social media messages. The adoption of webservices in the TA community reflects a similar trend in the enterprise arena, where webservices are expected to play an increasingly important role. This increasing trend can be

attributed to the growing number of enterprises realizing the benefits of cloud computing and adhering to the Software-as-a-Service (SaaS) philosophy [45,46].

3.2.3. Evaluation and metrics

The performance of the various applications is estimated using several evaluation metrics, traditionally employed in the context of TA research. For instance, the clustering quality in [12] is evaluated using the *purity* metric. In [17], the accuracy of the solutions, retrieved from a case-base, in response to a query is estimated according to the average word *similarity score* and the *mean average precision (MAP)*. The performance of other applications, such as text categorization [16,15], sentiment analysis [12], named entity recognition [14], is estimated using the standard measures of *precision, recall and F1-score*, the latter being a harmonic mean that balances precision and recall [47]. However, it is worth noting that in some applications, for e.g. the sentiment analysis component of [12], precision is favored over recall. Also, the area under curve (AUC) was preferred to the F1-score in [14] as it provides a more unbiased measure of the performance (in this case, text categorization) when the classes are imbalanced.

The evaluation procedure of [13] deserves particular attention. In this study, the authors evaluate the application's performance by measuring how users perceive its end results. Specifically, they gauge the degree to which users perceive automatically generated tweets as informative and interesting. This evaluation procedure is somewhat atypical of the TA community, whereby evaluation tends to focus on the efficiency with which the proposed TA techniques performed a given task, such as text classification. Users' perceptions of the results and the usefulness are, in most cases, overlooked, but see [48] for an example of a related natural language generation task in an academic experimental setting, using test subjects to rate the output of rivaling systems.

4. Challenges in industrial applications

We now discuss the different challenges and constraints that industrial environments impose on TA techniques, as opposed to deploying these techniques in a more controlled, experimental research environment. Presenting an exhaustive or comprehensive list of constraints and challenges is a formidable, almost impossible, endeavor, requiring a survey of all TA techniques deployed in industry. Therefore, we will base our discussion on the applications described in the previous section.

4.1. Challenge 1: Variety/handling heterogeneous data sources

Texts to be processed in TA applications will typically originate from a variety of sources, resulting in corpora of largely heterogeneous documents. This heterogeneity can be manifested in texts in many different ways, e.g.

- The texts can be encoded in different formats or rendered in different layouts, such as webpages;
- The texts' length can also be different; tweets for instance will be much shorter than other types of texts;
- They can be expressed in multiple languages. For instance, in many companies, documents for the higher management is often expressed in English. Conversely, at the operational level, communication predominantly takes place in the local language.

4.2. Challenge 2: Text and genre artifacts

Texts in real-life, industrial environments often exhibit certain peculiarities that complicate their automatic processing by TA techniques. An overview of the main peculiarities is presented below.

- TA techniques in many application domains have to deal with short and sparse texts. These texts do not provide sufficient statistical redundancy, as opposed to larger, standard corpora employed in academic research efforts. Consequently, TA techniques are unable to acquire reliable statistical evidence from the contents of such texts, hindering their analysis. The issue of sparsity is further exacerbated by the heavy usage of variants, such as "TO" "takeoff" and "take-off" to represent a single term or concept;
- It is also common for these texts to be expressed in an informal style, rife with ill-formed grammatical constructs. Thus, it is hard for TA techniques to acquire reliable linguistic information from the contents of such texts, impeding their automatic analysis;
- Another challenge, especially for text categorization applications, is that of class imbalance, i.e. the training instances (text documents) are unevenly distributed across the different classes. For e.g. the aim of the application presented in [16] is to classify aviation incident reports using a classification scheme consisting of 37 categories. However, the training data distribution was such that a single category contained around 41% of all reports, while 25 categories together covered less than 1% of all reports. The difficulty when dealing with such class imbalance problems is that of learning a model that accurately predicts the majority as well as the minority classes;
- As related issue as described in [16] is that some classification schemes in real-life applications are too unwieldy and complex for TA techniques. For instance, a classification scheme consisting of 1600 categories, hierarchically organized across three levels, was deemed "out of reach" by the text classification application of [16].

4.3. Challenge 3: Lack of gold-standards and annotated data

Evaluating the results of TA techniques can be particularly challenging due to subjectivity inherent in certain applications, such as sentiment analysis, text classification and text clustering. In the context of academic research, such difficulties are alleviated by relying on gold-standard datasets and algorithms from previous research for comparing, benchmarking and assessing the results. However, gold-standard data are rarely available in industrial settings, compounding the difficulties in evaluating the performance of the developed TA techniques. A related and more general issue is the lack of annotated data for training supervised machine learning algorithms in applications such as text categorization and sentiment analysis.

The creation of gold-standard datasets and the annotation of training data is a tedious, time-consuming and complex process. For instance, human experts took around five hours merely for defining the annotation guidelines to create training data for the text classification application in [15]. Furthermore, as described in [12], it is hard even for human experts to reach a consensus in defining annotation guidelines.

4.4. Challenge 4: Quality of results

Given the current state-of-the-art in research, it is unreasonable to expect TA techniques to outperform humans or even to yield results that are close to perfection, i.e. 100% accuracy. This is particularly relevant for industrial applications, where extensive domain-knowledge and expertise are crucial for the successful execution of certain tasks, such as classifying aviation incident reports into safety categories or finding clusters of similar texts. Furthermore, as discussed before, these tasks prove to be

extremely challenging even for human experts due to the subjectivity involved. For instance, one of the classification schemes in the aviation sector consisted of more than 1600 categories [16]. As can be expected, it was considered to be too overwhelming for use by TA (text categorization) applications for several reasons. Intuitively, accurately discriminating between such a large number of classes is challenging for automatic TA techniques and even for human experts. Then, there are there are the issue of sparseness, i.e. insufficient training examples per class, as well as that of class imbalance. Another issue pertains to the fact that some categories correspond to broad, vague concepts. They are inherently hard to formalize. For example, the category "system component failure" represents all failures that affect the large number of components in an aircraft. Precisely defining such a category is hard. Furthermore, such a category is often described in terms of extraneous information, such as crews declaring an emergency or troubleshooting. Consequently, TA techniques are unable to learn the salient features that accurately discriminate the category for classifying incident reports [16].

The limitations of TA techniques for certain domain-specific tasks and the added-value of expert knowledge were also highlighted in the study of [18]. Specifically, the proposed application fails in detecting drug names that are lexically realized as adjectives, for e.g. "antidepressant". It was also unable to detect drug effects realized using colloquial expressions, for e.g. "my head is ringing".

A growing trend in many TA applications, particularly those for text classification [15,16], is to make use of active learning to improve the overall TA process and the results. However, even though active learning has been shown to be a promising strategy, its usage in industrial applications also gives rise to several pertinent questions. For instance, an important issue is the choice of the initial document subset, which tends to have a significant impact on the final classification accuracy. Another issue is that of determining an optimal threshold for identifying those borderline/ambiguous cases (documents) to be submitted to the experts' judgments. Related to the previous issue is the problem of determining an appropriate number of documents to be presented to the experts for further evaluation in each iteration.

4.5. Challenge 5: Velocity

"Velocity" (speed) is a major factor in industrial applications. It is manifested in two primary ways. First, some industrial applications, particularly those that support decision making in mission-critical activities, are required to process data (texts) and generate results in a timely and efficient manner. Second, is the rate at which certain sources, especially social media networks, produce data (text streams). For instance, around 6000 tweets are generated on average every second.[10] Such high-paced text streams demand a timely analysis so that relevant and meaningful information are extracted from their contents.

5. Desiderata of industrial applications

We now formulate a set of desiderata that TA techniques should satisfy in order to ensure their successful deployment in industrial applications. As will be seen in our discussion, some of the desiderata stem directly from the aforementioned challenges, while others are more general.

5.1. Desideratum 1: Flexibility – data

As mentioned earlier, texts to be processed in industrial TA applications will typically originate from a variety of sources. For instance, documents (job offers) that are fed to the domain-specific search-engine in [15] are in heterogeneous formats and layouts. The search-engine should be able to reliably pre-process all these different types of documents and index their contents. Similarly, the application for detecting ADR presented in [18] has to mine relevant information from different social media sources, such as Twitter and web forums. Thus, industrial TA applications should be sufficiently flexible so as to support and process texts in a wide variety of formats and from multiple sources.

5.2. Desideratum 2: Flexibility – language

In addition, industrial TA applications should be flexible with regards to the language. Several of the applications we reviewed are targeted at texts in languages other than English. For instance, the text clustering and sentiment analysis application in [12] process texts in German and English; the traffic monitoring application in [14] analyzes tweets in Portuguese. The domain-specific search-engine and text classification system in [15] operates on German text, the application in [18] extracts information on ADR from Spanish texts, while [16] perform topic modeling, text classification and information retrieval on texts predominantly expressed in French. This reflects a trend also observed in research, characterized by an increasing number of TA systems for a range of widely spoken languages other than English, such as Chinese, Arabic, Russian, and Turkish.

5.3. Desideratum 3: Robustness

As discussed in the previous section, texts generated in real-life, industrial environments exhibit several intricacies, such as sparsity and ill-formed grammatical constructs. Classical TA techniques, developed within the realm of scientific/academic research, often face several difficulties and are too brittle for analyzing these types of texts [49–51]. Therefore, in industrial applications, the robustness of TA techniques tends to be favored over their sophistication. For instance, latent Dirichlet allocation (LDA) is considered as the current state of the art and one of the most successful techniques for topic detection from corpora. However, the experiments in [16] and [12], conducted over real-life texts, reveal that LDA does not perform well. Consequently, *K*-Means clustering was employed for topic detection as it was more robust and achieved a high clustering quality and high efficiency [12]. Similarly, many of the applications reviewed [12,14,16] preferred stemming over lemmatization as the former was considered to be more robust and less dependent on the language.

5.4. Desideratum 3: Minimal supervision

Annotated corpora for training supervised techniques are often unavailable or difficult to obtain in industrial environments, and their creation entails several challenges as discussed before. To overcome this difficulty, industrial applications could rely on minimally supervised techniques or on techniques based on distant supervision. These techniques have already been applied to various TA tasks, such as term and relation extraction [49–51] and sentiment analysis [52]. They have been shown to achieve promising results while eschewing the need for annotated training data.

5.5. Desideratum 4: Human intervention

It is unreasonable to expect TA techniques to produce results of the same quality as human experts, especially in domains requiring human knowledge and expertise. Despite the increasing adoption of TA techniques, there is still a large number of

[10] http://www.internetlivestats.com/twitter-statistics/.

applications in which human intervention is crucial. As described in [16], humans should be placed at the heart of the (TA) process. In this regard, the study of [12] reveals that users preferred applications that allowed them to interactively experiment with and explore the results and datasets.

Based on the applications that we reviewed, we could classify "user interaction" into three main types, viz.

1. For finetuning the TA process and results, such as merging and splitting clusters and varying thresholds to control the number of topics and number of clusters as well as to select relevant keywords;
2. For validating results, especially in mission critical applications, such as those pertaining to aviation safety [16]. For instance, [16] proposes a configuration whereby no human intervention is required for aviation safety reports that could be classified (automatically) with a precision of at least 0.95 by the application. Conversely, those reports with a lower precision could be dispatched to a human expert for further evaluation and validation. Such a configuration enables the experts to focus on specific cases that cannot be satisfactorily classified by the application, thereby making efficient use of their time;
3. For active learning in text classification applications. In this configuration, documents that cannot be conclusively classified by the application are annotated by a human expert and used for training the text classification application in later iterations. There are several benefits associated with an active learning strategy. First, it optimizes the experts' time by submitting to their judgment only those ambiguous documents that cannot be accurately classified. Second, it alleviates the need for large amounts of training data, which are often not available in industrial contexts.

A common method to realize the desired interactivity is via parameters [12,16]. Furthermore, since users appreciated having a degree of control to experiment with parameter values, no attempts were made in the applications to learn, predict or adjust these values automatically.

5.6. Desideratum 5: Ease of use

An important desiderata, fundamental to achieve the desired level of human intervention, is that TA applications should be easy to use. This desiderata is particularly important considering that lay users in industrial settings will not be TA specialists. Thus, they should be presented with applications that are simple and easy to use, without requiring them to tamper with the applications' internal mechanics. This is not likely in academic TA research, whereby considerations pertaining to user interface design and ease of use are often secondary. Research in human–computer interaction and ergonomics proposes several methods for enhancing the ease of use of systems [53,54]. In the applications we reviewed, we noted 2 main ways to make the proposed TA applications easier to use:

1. Via a Graphical User Interface (GUI), enabling users to interact with the applications in a straightforward manner, for e.g. in tuning and experimenting with various parameter settings;
2. Through visualization (of results), enabling users to analyze (for e.g. drill-down) the information and results generated by the applications. For instance, in the application developed in [12], clusters are rendered as nodes in a semantic map, and are labeled with their corresponding topics. Also, the nodes' size is proportional to the number of documents contained within the clusters, and the nodes are positioned on the map such that the distance between them corresponds to the similarity between their respective topics. Furthermore, nodes are colored according to the sentiments expressed by the texts they contain. Similarly, the IR tool in [16] displays groups of similar reports chronologically along a temporal axis to facilitate the monitoring of trends.

5.7. Desideratum 6: Velocity – fast analysis

TA is often employed in industrial applications to support decision-making at all levels, i.e. strategic, tactical and operational, of an enterprise. Therefore, an essential desideratum is the ability to analyze potentially large volumes of data and generate reliable results in a timely manner. The need for timely response is critical to ensure a high degree of information freshness, which has a significant impact on the practical usefulness, relevancy and veracity of the results produced by the application [55,56]. Failure to satisfy this desideratum can lead to the production of inaccurate and outdated results, which are not interesting to users and can even be detrimental to decision-making. For e.g. aviation incident reports in [16] have to be analyzed in the least time possible so that safety issues can be detected and corrective measures implemented quickly. Other applications may even call for real-time or near real-time processing. For e.g. the search-engine in [15] has to analyze, index and classify job offers from various sources in a near real-time manner in order to provide the most accurate and up-to-date information to users. Similarly, the fleet management application in [14] has to process tweets in real-time in order to acquire the latest traffic information.

5.8. Desideratum 7: Extrinsic evaluation

The various applications that we reviewed are evaluated using widely accepted standard metrics, originating from TA (and machine learning) research. Examples included the traditional precision, recall and F-score [47]. These metrics, no doubt, provide a reliable estimate of performance. They measure the intrinsic performance, i.e. how accurate the TA application is or how efficient it is in performing its task, e.g. text classification? However, measuring the extrinsic performance has been an underrated issue in TA, both in research and practice. Unlike their intrinsic counterparts, which focus on efficiency, extrinsic evaluation measures focus on the effectiveness or usefulness of the application in its usage environment. Thus, a major desideratum for industrial applications would be the definition of metrics for extrinsic evaluation. One solution to fill this lacuna could be the adoption of information systems (IS) success frameworks proposed in organizational management (information systems) research. One such framework is the well-known DeLone and McLean IS model [57]. It posits that IS success can be measured using six interdependent variables as proxies, namely: System Quality, Information Quality, Service Quality, (Intention to) Use, User Satisfaction and Net Benefits. The challenge is now how to accurately measure these variables, given their subjective and qualitative nature, and to determine whether they should all be given the same importance (weights) in different TA applications.

6. Future trends

Our review has provided some evidence that classical text analytics techniques are still at the heart of many industrial applications. These include text classification, text clustering, text summarization, and measures for text similarity. The main strength of these techniques for industrial applications is that they are now considered as mature technologies by virtue of their longstanding history in the TA community and of the significant attention that they have received over the years in various research efforts.

Furthermore, they are considered sufficiently robust to deal with the intricacies of texts generated in real-world industrial applications as discussed in Section 4. In addition, another strength of these classical techniques is that they generate results that are easy to interpret and that can directly be exploited for informed decision-making.

However, beside the development of the classical techniques, some of which were already well understood and developed two decades ago, many new developments have emerged that have started to play a role in industrial applications, or could play a role in the near future. In particular, recent years have witnessed a revival in neural-network-based methods, such as deep-learning neural networks [58,59] that have shown tremendous potential in the processing of streaming media (speech, video) as well as still images. All major commercial speech recognition systems (e.g. Microsoft Cortana, Xbox, Skype Translator, Google Now, Apple Siri, Baidu and iFlyTek voice search, and a range of Nuance speech products, etc.) nowadays are based on deep learning.

Work on deep learning of text processing is lagging behind, possibly because deep-learning methods are less suited to symbolic data than they are for sub-symbolic input streams. Initial successes have been reported, however. For example, Chen et al. [60] propose a recurrent deep learning method for dependency parsing. Sutskever et al. [61] demonstrate how sequence-to-sequence problems (such as machine translation or paraphrasing) in NLP can be formulated and learned through deep learning methods. Schocher et al. [62] show that sentiment classification can be learned with deep learning models by using word embeddings as an input layer to a recursive neural network (RNN). In the field of natural language generation, deep learning has also been shown to be successful through the modelling of semantic constraints [63]. Ma et al. [64] propose a deep learning approach for sentence embedding, and show that their method can achieve state-of-the-art results for many sentence classification problems such as sentiment classification and question classification. Also in the domain of information retrieval, recent work has demonstrated how deep learning models can be applied to learning a similarity score between two text segments [65,66].

Another recent development in the NLP field is to improve NLP or text analytics systems by using so-called word embeddings, distributed word representation produced by dimension-reduction techniques most of which have a neural-network interpretation or implementation, such as word2vec [67] or GloVe [68]. The relative ease of using word embeddings software and output, in contrast to the relatively demanding infrastructure required for deep learning (e.g. high-performance computing, GPU hardware), makes the method easy to integrate into industrial applications; in terms of our desiderata, they are easy to use, require minimal supervision, are robust, and are flexible with respect to language and data. One example is provided by Tosik et al. [69] in the context of information extraction from curriculum vitae documents in German. Word embeddings, replacing lexical features in a machine-learning-based sequence parser, cause the parser to be more accurate in detecting domain-specific named entities, especially in out-of-sample data.

Word embeddings and deep-learning methods are also central to a new trend in NLP to jointly model object recognition in images and generating texts that describe the scene in those images [71]. As many industries work with images and video streams, often combined with speech or text in simultaneous streams, the automatization of generating keywords, captions, or subtitles along with image data offers interesting outlooks for serving enriched content to users.

Yet, in industry the uptake of word embeddings and deep-learning methods is still in its infancy. Early successes such as those reported in [69] are likely to seed further and potentially widespread deployment in real-life industrial applications. On the other hand, typical industrial demands for ease of use of the methods may limit the uptake of these methods, which rely on big data and partly on high-performance computing, of which the operation is far from trivial, and which are not transparent in their internal workings. Furthermore, classical methods offer strong baselines that these new methods are not guaranteed to surpass. Key in this development will likely be the research and development departments of the growing text analytics industry.

References

[1] S. Filippov, Mapping Text and Data Mining in Academic and Research Communities in Europe, Technical Report, 2014 (accessed 19.08.15).
[2] D. Harris, Netflix uses data for a lot more than just recommendations, 2014, https://gigaom.com/2014/06/12/netflix-uses-data-for-a-lot-more-than-just-recommendations/ (accessed 19.08.15).
[3] X. Amatriain, Netflix Recommendations: Beyond the 5 Stars, 2012, http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html/ (accessed 19.08.15).
[4] J. Best, IBM Watson: the inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next, 2014, http://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/ (accessed 19.08.15).
[5] L. Dignan, IBM Watson's next adventure: healthcare with Nuance, 2011, http://www.zdnet.com/article/ibm-watsons-next-adventure-healthcare-with-nuance/ (accessed 19.08.15).
[6] E. Guizzo, IBM's Watson Jeopardy Computer Shuts Down Humans in Final Game, 2011, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/ibm-watson-jeopardy-computer-shuts-down-humans/ (accessed 19.08.15).
[7] The Bank of England mines social media in bid to predict economic events ahead of time, 2015, http://www.out-law.com/en/articles/2015/august/the-bank-of-england-mines-social-media-in-bid-to-predict-economic-events-ahead-of-time/ (accessed 19.08.15).
[8] IBM Watson Hard At Work: New Breakthroughs Transform Quality Care for Patients, 2013, https://www.mskcc.org/press-releases/ibm-watson-hard-work-new-breakthroughs-transform-quality-care-patients/ (accessed 19.08.15).
[9] Watson in Finances, http://www-05.ibm.com/innovation/uk/watson/watson_in_finance.shtml/ (accessed 19.08.15).
[10] DBS Bank Engages IBM's Watson to Achieve Next Generation Client Experience, https://www-03.ibm.com/press/us/en/pressrelease/42868.wss/ (accessed 19.08.15).
[11] Standard Bank deploys IBM's Watson to crunch customer data, 2014, http://www.finextra.com/news/fullstory.aspx?newsitemid=26624/ (accessed 19.08.15).
[12] P. Greiner, S. Evert, F. Baigger, B. Lang, A distributional approach to open questions in market research, Comput. Ind. (2015).
[13] E. Lloret, M. Palomar, Analysing and evaluating the task of automatic tweet generation: knowledge to business, Comput. Ind. (2015).
[14] F.C. Albuquerque, M.A. Casanova, H. Lopes, L.R. Redlich, J.A.F. de Macedo, M. Lemos, M.T.M. de Carvalho, C. Renso, A methodology for traffic-related twitter messages interpretation, Comput. Ind. (2015).
[15] S. Schmidt, S. Schnitzer, C. Rensing, Text classification based filters for a domain-specific search engine, Comput. Ind. (2015).
[16] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, C. Raynal, Natural language processing for aviation safety reports: from classification to interactive analysis, Comput. Ind. (2015).
[17] J.W. Chang, M.C. Lee, T.I. Wang, Integrating a semantic-based retrieval agent into case-based reasoning systems: a case study of an online bookstore, Comput. Ind. (2015).
[18] P. Martínez, J.L. Martínez, I. Segura-Bedmar, J. Moreno-Schneider, A. Luna, R. Revert, Turning user generated health-related content into actionable knowledge through text analytics services, Comput. Ind. (2015).
[19] S. Bird, NLTK: the natural language toolkit, in: Proceedings of the COLING/ACL on Interactive Presentation Sessions, Association for Computational Linguistics, 2006, pp. 69–72.
[20] M.F. Porter, Snowball: A Language for Stemming Algorithms, 2001.
[21] J.R. Bellegarda, Latent semantic mapping: principles & applications, Synthesis Lectures on Speech and Audio Processing, vol. 3, 2007, pp. 1–101.
[22] S. Evert, T. Proisl, P. Greiner, B. Kabashi, SentiKLUE: updating a polarity classifier in 48 hours, in: SemEval 2014, 2014, p. 551.
[23] T. Dunning, Accurate methods for the statistics of surprise and coincidence, Comput. Linguist. 19 (1993) 61–74.
[24] E. Lloret, M. Palomar, Towards automatic tweet generation: a comparative study from the text summarization perspective in the journalism genre, Expert Syst. Appl. 40 (2013) 6624–6630.
[25] FreeLing 3.1, http://nlp.lsi.upc.edu/freeling/ (accessed 19.08.15).
[26] E. Motta, E. Fernandes, R. Milidiú, F-ext-2.0: a web service for natural language processing, in: PROPOR, 2010, 27–30.
[27] J. Platt, et al., Fast training of support vector machines using sequential minimal optimization, Advances in Kernel Methods: Support Vector Learning, vol. 3, 1999.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The Weka data mining software: an update, ACM SIGKDD Explorations Newsletter 11 (2009) 10–18.
[29] F. da Costa Albuquerque, M. Casanova, J.A.F. de Macedo, M.T.M. de Carvalho, C. Renso, et al., A proactive application to monitor truck fleets, in: 2013 IEEE 14th International Conference on Mobile Data Management (MDM), vol. 1, IEEE, 2013, pp. 301–304.
[30] E.R. Fernandes, C.N. Dos Santos, R.L. Milidiú, Latent structure perceptron with feature induction for unrestricted coreference resolution, in: Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 41–48.
[31] S. Schnitzer, S. Schmidt, C. Rensing, B. Harriehausen-Mühlbauer, Combining active and ensemble learning for efficient classification of web documents, Polibits 49 (2014) 39–45.
[32] Treetagger, http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/ (accessed 03.09.15).
[33] liblinear-java, http://liblinear.bwaldvogel.de/ (accessed 01.09.15).
[34] R. Rehurek, Gensim, https://radimrehurek.com/gensim/ (accessed 03.09.15).
[35] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, Systems, IEEE Trans. Man Cybernet. 19 (1989) 17–30.
[36] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1994, pp. 133–138.
[37] C. Leacock, M. Chodorow, Combining local context and wordnet similarity for word sense identification, WordNet: An Electronic Lexical Database, vol. 49, 1998, pp. 265–283.
[38] W.B. Dolan, C. Brockett, Automatically constructing a corpus of sentential paraphrases, in: Proc. of IWP, 2005.
[39] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, in: AAAI, vol. 6, (2006), pp. 775–780.
[40] L. Qiu, M.-Y. Kan, T.-S. Chua, Paraphrase recognition via dissimilarity significance classification, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 18–26.
[41] Y. Li, D. McLean, Z. Bandar, J.D. O'shea, K. Crockett, et al., Sentence similarity based on semantic nets and corpus statistics, IEEE Trans. Knowl. Data Eng. 18 (2006) 1138–1150.
[42] J. Oliva, J.I. Serrano, M.D. del Castillo, Á. Iglesias, SyMSS: a syntax-based measure for short-text semantic similarity, Data Knowl. Eng. 70 (2011) 390–405.
[43] G. Tsatsaronis, I. Varlamis, M. Vazirgiannis, Text relatedness based on a word thesaurus, J. Artif. Intell. Res. 37 (2010) 1–40.
[44] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: an architecture for development of robust HLT applications, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 168–175.
[45] N. Gold, A. Mohan, C. Knight, M. Munro, Understanding service-oriented software, IEEE Softw. 21 (2004) 71–77.
[46] P. Helo, M. Suorsa, Y. Hao, P. Anussornnitisarn, Toward a cloud-based manufacturing execution system for distributed manufacturing, Comput. Ind. 65 (2014) 646–656.
[47] C. Van Rijsbergen, Information Retrieval, Buttersworth, London, 1979.
[48] S. Wubben, E. Krahmer, A. Van den Bosch, Sentence simplification by monolingual machine translation, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL, New Brunswick, NJ, (2012), pp. 1015–1024.
[49] A. Ittoo, G. Bouma, Term extraction from sparse, ungrammatical domain-specific documents, Expert Syst. Appl. 40 (2013) 2530–2540.
[50] A. Ittoo, G. Bouma, Minimally-supervised extraction of domain-specific part-whole relations using Wikipedia as knowledge-base, Data Knowl. Eng. 85 (2013) 57–79.
[51] A. Ittoo, G. Bouma, Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base, Data Knowl. Eng. 88 (2013) 142–163.
[52] M. Purver, S. Battersby, Experimenting with distant supervision for emotion classification, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 482–491.
[53] V. Venkatesh, Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model, Inf. Syst. Res. 11 (2000) 342–365.
[54] R.L. Kumar, M.A. Smith, S. Bannerjee, User interface features influencing overall ease of use and personalization, Inf. Manage. 41 (2004) 289–302.
[55] M. Bouzeghoub, A framework for analysis of data freshness, in: Proceedings of the 2004 International Workshop on Information Quality in Information Systems, ACM, 2004, pp. 59–67.
[56] C.A. Meadows, C. Fernandez-Gago, 7th International Workshop on Security and Trust Management, STM 2011, Copenhagen, Denmark, June 27–28, 2011, Revised Selected Papers, vol. 7170, Springer, 2012.
[57] S. Petter, W. DeLone, E. McLean, Measuring information systems success: models, dimensions, measures, and interrelationships, Eur. J. Inf. Syst. 17 (2008) 236–263.
[58] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[59] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[60] D. Chen, C.D. Manning, A fast and accurate dependency parser using neural networks, in: Empirical Methods in Natural Language Processing (EMNLP), 2014.
[61] Q.L.I. Sutskever, O. Vinyals, Sequence to sequence learning with neural network, in: Proceedings of the NIPS 2014 Conference, 2014.
[62] R. Socher, Recursive deep models for semantic compositionality over a sentiment Treebank, in: Proceedings of the EMNLP 2013 Conference, 2013.
[63] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, S. Young, Semantically conditioned LSTM-based natural language generation for spoken dialogue systems, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1711–1721.
[64] M. Ma, L. Huang, B. Xiang, B. Zhou, Dependency-based convolutional neural networks for sentence embedding, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Beijing, China, 2015, pp. 174–179.
[65] J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, Modeling interestingness with deep neural networks, in: Proceedings of the EMNLP 2014 Conference, 2014.
[66] J. Li, M.-T. Luong, D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, in: Proceedings of the ACL, 2015.
[67] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751.
[68] J. Pennington, R. Socher, C.D. Manning, GloVe: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014, pp. 1532–1543.
[69] M. Tosik, C. Lygteskov Hansen, G. Goossen, M. Rotaru, Word embeddings vs word types for sequence labeling: the curious case of CV parsing, in: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 123–128.
[71] A. Belz, L. Coheur, V. Ferrari, M.-F. Moens, K. Pastra, I. Vuli (Eds.), Proceedings of the Fourth Workshop on Vision and Language, Association for Computational Linguistics, Lisbon, Portugal, 2015.

Since 2013, **Ashwin Ittoo** is an Asst-Professor in Information Systems at the HEC Management School, University of Liège, Belgium. His research interests are minimally-supervised learning techniques for Machine Learning and Natural Language Processing and in the application of these techniques for measuring socio-economic indicators. He received his Ph.D. degree from the University of Groningen, The Netherlands in 2012 and his Bachelors and Masters from the National University of Singapore and the Nanyang Technological University, Singapore.

Le Minh Nguyen is currently an Associate Professor of School of Information Science, JAIST. He leads the lab on Machine Learning and Natural Language Understanding at JAIST. He received his B.Sc. degree in information technology from Hanoi University of Science, and M.Sc. degree in information technology from Vietnam National University, Hanoi in 1998 and 2001, respectively. He received his Ph.D. degree in information science from School of Information Science, Japan Advanced Institute of Science and Technology (JAIST) in 2004. He was an assistant professor at School of Information Science, JAIST from 2008 to 2013. His research interests include machine learning, text summarization, machine translation, natural language processing, and information retrieval.

Antal van den Bosch (Ph.D. 1997, Universiteit Maastricht) is professor of language and speech technology at the Centre for Language Studies at Radboud University, Nijmegen, the Netherlands. His research interests include memory-based and exemplar-based natural language modeling, text analytics applied to historical texts and social media, and proofing tools. He is a member of the Netherlands Royal Academy of Arts and Sciences and ECCAI Fellow.