An Advanced Press Review System Combining Deep News Analysis and Machine Learning Algorithms

Danuta Ploch, Andreas Lommatzsch, and Florian Schultze
DAI-Labor, Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{danuta.ploch, andreas.lommatzsch}@dai-labor.de,
florian.schultze@campus.tu-berlin.de

Abstract

In our media-driven world the perception of companies and institutions in the media is of major importance. The creation of press reviews analyzing the media response to company-related events is a complex and time-consuming task. In this demo we present a system that combines advanced text mining and machine learning approaches in an extensible press review system. The system collects documents from heterogeneous sources and enriches the documents applying different mining, filtering, classification, and aggregation algorithms. We present a system tailored to the needs of the press department of a major German University. We explain how the different components have been trained and evaluated. The system enables us demonstrating the live analyzes of news and social media streams as well as the strengths of advanced text mining algorithms for creating a comprehensive media analysis.

1 Introduction

The analysis of news related to companies and institution is a complex task often performed by human experts. Due to the growing amount of news and articles published in social media, large collections of data must be analyzed. In order to support the efficient creation of press reviews powerful tools are needed for the automatic aggregation and deep analysis of news. This motivates us to develop an extensible framework allowing us to combine advanced machine learning algorithms for filtering, extracting, and visualizing relevant information.

1.1 The analyzed Scenario

In this work we present a system developed for the press department of the Berlin Institute of Technology (TUB). The system should be able to collect news as well as social media articles related to the TUB or to any other of the Berlin's universities. The system is subject to a collection of requirements: The system should detect duplicates or texts with minor variations. Persons and faculties mentioned in news articles are of special interest for a fine-grained analysis. The system should detect known entities and create detailed statistics. Events drive the news. The system should detect and follow news related to the Berlin's universities. Readers of news often drain in information. The system should aggregate and visualize relevant documents in a concise way by computing key figures (e.g. describing the sentiment score for news) and calculating statistics giving a quick overview on the characteristics of the news stream. The results of the news analysis should be accessible in a web application.

1.2 Challenges

The automatic creation of press reviews leads to several challenges. The system has to integrate all important sources and to filter irrelevant documents. A specific challenge is that abbreviations are often used for institutions having a long name. In our scenario the "Technische Universität Berlin" is frequently called "TUB" or "TU". The press review system must infer from the context whether an article is relevant or not. The automatic analysis and enrichment requires a variety of algorithms, including duplicates detection, identification and disambiguation of named entities, and sentiment analysis. The sentiment analysis for news articles is a hard challenge since most journalists seek to write objectively. Nevertheless, news induces emotions relevant in the automatic analysis of news documents. Since the system has

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 109–114,
Berlin, Germany, August 7-12, 2016.

been developed for a major German university, the language analysis focuses on German texts.

1.3 Structure of the Work

The remaining work is structured as follows. In Section 2 we explain the basics of text mining algorithms and discuss existing press review systems. The architecture and the implemented algorithms are presented in Section 3. In Section 4 the visualization of the elicited data is described in further detail. Section 5 explains the most important use cases and presents the evaluation results with respect to the functionality of the press review portal. A conclusion and an outlook to future work is given in Section 6.

2 Related Work

We review advanced text mining algorithms and existing press review systems. There are a lot of commercial press review systems such as `http://www.bluereport.net/de/`, `http://www.pressemonitor.de/` or `https://www.ausschnitt.de/`. The systems focus on printed newspapers but also provide press reviews for online published articles. Traditionally, the systems provide excerpts related to predefined search terms. In general, the companies offer a wide variety of analysis services but the applied algorithms are neither open nor explained. With the pricing models in mind a lot of work is still performed by human experts. Based on the companies' information policy and the marketing language on the websites it is unclear to what extent machine learning or text mining algorithms are used.

Several research-oriented systems complement commercial press review systems. An exemplary application for large scale news analysis is LYDIA. LYDIA focuses on named entity detection. Its key feature is answering questions such as "who is being talked about, by whom, when, and where?" (Lloyd et al., 2005). The SEMANTIC PRESS system (Picchi et al., 2008) uses an alternative approach. It presents the most discussed themes in the Italian-spoken web.
An example for German media resonance analysis is the system explained by (Scholz, 2011). It focuses on entity extraction and sentiment analysis. Similar researches were done by (Hanjalic et al., 1998) and (Zhang et al., 2009).

3 Approach

We develop an open framework enabling us integrating different information sources and machine learning algorithms. The system allows us considering news portals, search engines, RSS feeds, and messages published on TWITTER. We deploy a flexible processing pipeline enriching freshly crawled documents as well as a batch engine used for clustering and generating newsletters. Our framework is open for the integration of new sources and algorithms allowing us incrementally extending and improving our system.

3.1 System Architecture

The structure of the developed system is shown in Fig. 1. The system consists of four major building blocks. The crawler component collects potentially relevant documents and tweets. The documents are persisted in a database. The processing components enrich the crawled documents and run several different machine learning algorithms. Based on the meta-data and the computed annotations the relevance of documents is computed and near duplicates are identified. The batch processing pipeline is a second pipeline used for processing documents from the database in predefined intervals. Both processing pipelines can be easily extended. The use of a database decouples the crawling from the processing allowing an efficient and concurrent computation of annotations.

The enriched documents are presented to the user in a web application and summarized in a periodically created newsletter.

3.2 Text Mining Components

In this Section we present the algorithms implemented for the different components in detail and discuss specific adaptations.

Validation

Several crawlers collect potentially relevant documents subsequently analyzed by the validation component. The crawlers use APIs of major search engines and the TWITTER streaming API. We define for each source a component optimizing the queries in order to ensure that all relevant documents are crawled (taking into account the limits of the sources). Due to the fact that several sources only support simple term queries (instead of phrase queries), an additional filtering is required. For this purpose we manually labeled

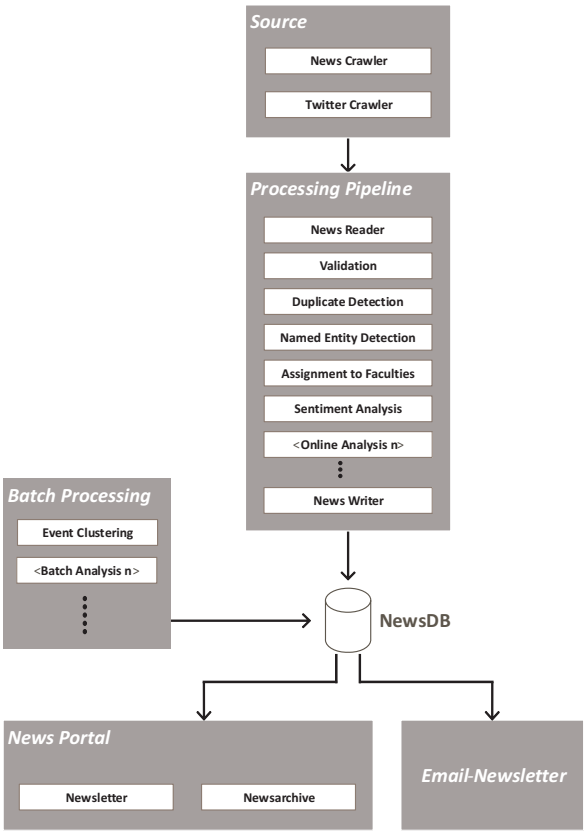

Figure 1: The figure visualizes the system architecture. Potentially relevant news documents and twitter messages collected by the *crawlers* are stored in the *news database*. The *Processing Pipeline* applies different text mining algorithms to each new document. The *Batch Processing* is executed on all documents periodically. The elicited data and documents are represented by a second application, the *News-Portal* which is built with GROOVY and GRAILS.

the documents crawled in a time frame of 2 weeks as relevant or irrelevant. Based on this dataset we trained a rule-based classifier considering phrases and context data for filtering out irrelevant documents. The filtering is especially important for handling abbreviations frequently used for referring to Berlin's universities.

Deduplication

Due to the applied architecture for collecting documents, similar documents might be crawled multiple times from different sources. Hence, we need to integrate a deduplication component identifying (near) duplicates. For ensuring an efficient processing of large text collections we implemented the *Rabin fingerprint algorithm* (Rabin and others, 1981). The algorithm randomly selects a predefined number of text shingles and computes the hash codes. Duplicates are identified by counting the fraction of identical shingles in two documents. We adjust the optimal parameter settings (shingle size, number of considered shingles) on a validation dataset.

Named Entity Detection

The named entity detection component recognizes and disambiguates persons mentioned in news articles. For the recognition part it uses several components from "Stanford CoreNLP" which are explained in (Manning et al., 2014). In particular, the component applies the parser, POS-tagger, and Named Entity Recognizer (NER) to detect mentions of professors, researchers and other university-related experts mentioned in the news articles. Based on the output of the "Stanford CoreNLP" tools the module enriches each identified person with their titles and associated organization, provided the news article contains the necessary information within a window of n words. In addition, the person's name is decomposed into a given name and a surname. In order to identify person mentions unambiguously the module applies local and global disambiguation strategy. The local disambiguation resolves co-referent person mentions within one news article. It assembles a representation of each person as complete as possible. The global disambiguation performs a cross-document co-reference resolution. It considers all person attributes and words calculated from the text surrounding a person mention. Each person from a news article is compared to entries already stored in the database. In the course of similarity calculation all types of information (person attributes and bag-of-word) are weighted differently. If the similarity between a newly detected person and a person from the database exceeds a predefined threshold, the persons are merged in the database. Otherwise, a new person entry is created.

Assignment of Faculties

Usually, universities are structured in several faculties. The presence of single faculties in the media may be an important quality indicator for the universities. Our approach to assigning news articles to a faculty is person-based. Therefore, we first gather the names of all employees from the faculty websites. In this way we create a register of person names aligned with faculty affiliation.

In order to measure the media response of a specific faculty, the implemented component analyzes news articles according to mentions of persons related to the faculty. The implementation of our approach bases on an inverted index containing each document's full text. We search the documents for person names from our register. If our algorithm identifies a faculty-related person, it assigns the news article to the corresponding faculty.

Event Detection

The event detection component clusters news articles dealing with one concrete news event such as the *Queen's Lecture* or the *Long Night of the Sciences* in Berlin. Our approach uses a combination of the *Canopy* and the *k-means* algorithm for clustering which is described by (McCallum et al., 2000). In order to improve the accuracy of the clustering we enable a part-of-speech tagger. We exclude all words that do not contribute to the content like articles, conjunctions, and prepositions; we proceed with the resulting subset of the text. Since the k-means algorithm needs to be initialized with a fixed number of clusters k our component performs two stages. First, the component estimates the number of clusters by applying Canopy. We adjust Canopy's hyper-parameters on a manually annotated validation dataset. Then, the calculated canopies serve as input centroids for the second step, the k-means clustering. Finally, each cluster corresponds to a real-life event.

Sentiment Analysis

Despite of the objective nature of news articles, they are still a valuable source of sentiment information. They may express opinions of cited entities or may contain content influencing the reader's perception regarding a university. Our system incorporates two sentiment analysis components.

The first component implements a lexicon-based approach. It uses the SentiWS sentiment dictionary (Remus et al., 2010) containing positively and negatively connoted words with positive and negative scores respectively. In order to calculate the sentiment score of an entire news article it counts the values of positive and negative words occurring in the news article. The component takes into account negation by exploiting a list of inverting words. If an inverting word precedes a positive or negative connoted word, it changes its polarity.

The second approach uses machine learning techniques. We build a training dataset with about 2,400 randomly selected sentences from crawled documents. We annotate the sentences to have a positive, negative, or neutral sentiment. For the annotation we use the rules from (Clematide et al., 2012). Based on the created dataset we train a Multi-nominal Naive Bayes classifier able to classify each sentence of a news article into one of the three sentiment classes. We represent each sentence in the vector space model applying common text preprocessing steps. Beside unigrams we also include bigrams into the vectors to cover sentiment-related expressions such as "very good". The overall sentiment of a news article is computed based on all single sentence classifications. The classifier achieves promising results providing deep insights into the sentiment distribution within a news article. A more detailed explanation can be found in (Bütow et al., 2016).

4 Visualization

We implemented a web-based user interface visualizing the collected and annotated documents and tweets. The web portal provides two major views. (1) The *Newsletter* or *live* view shows the most recently collected news. (2) The *Newsarchive* view aggregates documents collected in the past and allows the creation of statistics as well as the visualization of events identified by clustering news related to one topic.

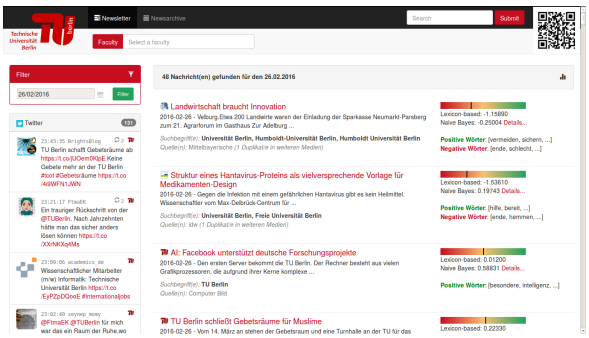

Figure 2: The Figure shows the front page of the system. It visualizes the Twitter messages on the left sidebar, the news articles in the middle and the corresponding sentiments on the right-hand side. The information for each displayed document are the title, a snippet, the date, the keywords used by the crawler and the source. A filter box is placed above the tweets allowing users filtering tweets by date and universities.

The *live* view shown in Figure 2 helps to ex-

plore the news on a daily basis. It gives users a fast overview of the most recently published news articles, shows which sources publish news related to the Berlins universities and visualizes the most important key figures. The view presents the documents as a list, each document provided with the extracted meta-data, such as the corresponding universities. If a document deals with the Berlin Institute of Technology, the faculty connected with the news item is also listed. In addition, the computed sentiment score and a short explanation for the sentiment score are displayed. A statistic showing the aggregated sentiment scores for one day for the major Berlins universities is presented in Figure 4.

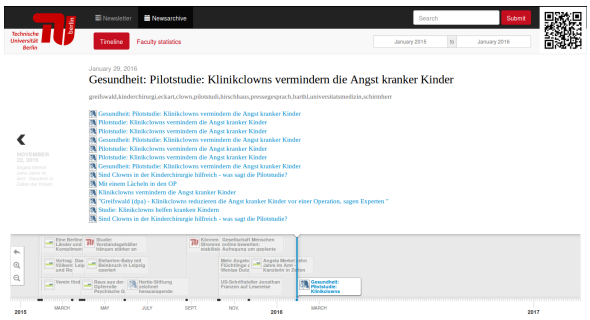

Figure 3: The Figure shows the time-line as one part of the news archive. At the bottom the time-line arranges different clusters of messages. Each cluster has an icon assigned to the related university and a title derived a document of the cluster. A selected cluster appears above the time-line with its corresponding news articles, the date and the ten most frequent terms in that cluster.

The *archive* view allows users analyzing the documents collected in the past. Users can search for documents or analyze the stream of news in detail. A powerful tool helping users to identify the most important events is the view that groups news documents by events on a timeline (Figure 3). The view lists all articles related to the selected events and shows the related institutions. The archive view also provides statistics. Figure 6 visualizes the number of documents related to the faculties of the TUB in a predefined time frame. These diagram supports a quick comparison of different faculties.

In addition to the statistics aggregating information collected over a timeframe, the systems provides views giving insights into single news articles. As discussed, we implemented a sentiment classifier working based on sentences. The sentiments scores computed for each sentence are visualized in Figure 5.

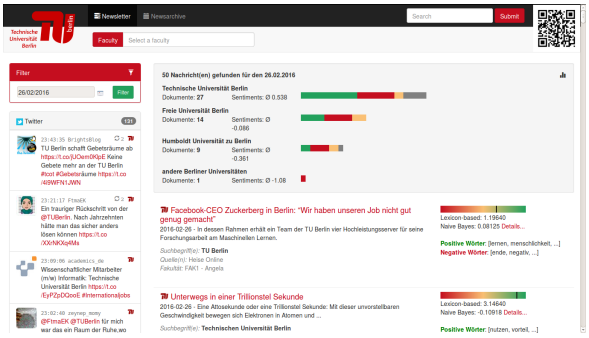

Figure 4: The Figure visualizes a diagram summarizing the results of the news aggregation and sentiment analysis for the current day. The diagram shows the distribution of the positive, negative, and neutral documents assigned to the corresponding university. In addition, the number of collected documents and the average sentiment scores are shown in this panel.

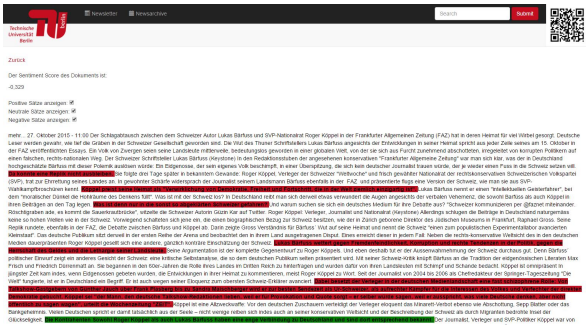

Figure 5: The Figure visualizes the sentiments computed for each sentence in a document. Sentences classified as negative are shaded in red; sentences classified as positive are shaded in green. The checkbox above the full text allows users to hide positive, negative, and neutral sentences.

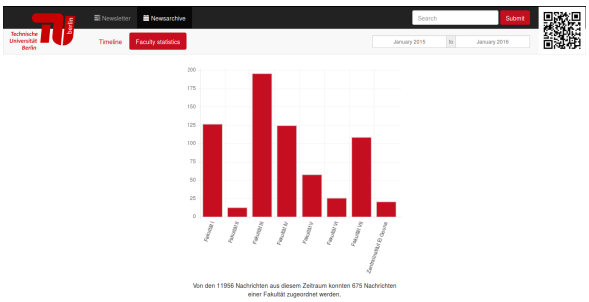

Figure 6: The Figure visualizes the number of news articles related to the “Technische Universiät Berlin” depending on the different faculties.

5 The Demonstration

The demo is accessible at `http://presse.dai-labor.de/pressreview/` with the following credentials: username `demo` and password `pressespiegel`.

The system follows the live news stream and allows users discovering the most recent news as well analyzing documents collected in the past. The web applications provides views for "regular" users but also detailed information and statistics for experts giving more fine-grained insights in the applied methods.

6 Conclusion and Future Work

We developed a powerful system that fulfills the requirements of a press review in the context of Berlin's universities. The system combines several different text mining algorithms and incorporates various visualizations helping users understanding the news and social media contributions. The system is open (upon request). It allows accessing the documents and their annotations by querying the database. The system can be extended by adding new modules to the processing pipelines. Hence, the system can be easily adapted for the specific requirements of other companies and for computing additional metrics. As future work we plan to conduct comprehensive user studies in order to optimize the algorithms to the needs of our users. We continuously work on adding blogs and RSS feeds providing information potentially relevant for our use case. We also plan an improved support for documents in other languages. Considering the identification of relevant persons, we aim to create an extended entity dataset and train a deep neural network. Furthermore, we plan the integration of additional machine learning algorithms for summarizing multiple documents related to events as well as algorithms for tracking the evolution of topics and sentiments over longer time frames.

Acknowledgments

The research leading to these results was performed in the CrowdRec project, which has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement No. 610594.

References

Florian Bütow, Florian Schultze, and Leopold Strauch. 2016. Sentiment Analysis with Machine Learning Algorithms on German News Articles. Technical report, Berlin Institute of Technology, AOT. http://www.dai-labor.de/publikationen/1052.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA – A Multi-layered Reference Corpus for German Sentiment Analysis. In *Procs. of the 8th Intl. Conf. on Lang. Res. and Evaluation*, pages 3551–3556.

Alan Hanjalic, Reginald L. Lagendijk, and Jan Biemond. 1998. Semiautomatic news analysis, indexing, and classification system based on topic preselection. *Proc. SPIE*, 3656:86–97.

Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. 2005. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval*, pages 161–166. Springer.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Procs. of the 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '00, pages 169–178, NY, USA. ACM.

Eugenio Picchi, S Cucurullo, E Sassolini, and Francesca Bertagna. 2008. Mining the news with semantic press. *Procs. of the 8th Intl. Conf. on Language Resources and Evaluation*, pages 141–144.

Michael O Rabin et al. 1981. *Fingerprinting by random polynomials*. Center for Research in Computing Techn., Aiken Computation Laboratory, Univ.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In *Proc. of the Intl. Conf. on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Thomas Scholz. 2011. Ein Ansatz zu Opinion Mining und Themenverfolgung für eine Medienresonanzanalyse. In *Procs. of the 23rd GI-WS Grundlagen von Datenbanken*, pages 7–12. issn: 1613-0073.

Yulei Zhang, Yan Dang, Hsinchun Chen, Mark Thurmond, and Cathy Larson. 2009. Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems*, 47(4):508 – 517. Smart Business Networks: Concepts and Empirical Evidence.