

Event Extraction and Discourse Monitoring

Theresa Krumbiegel¹, Albert Pritzkau¹ and Hans-Christian Schmitz¹

¹Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE), Fraunhoferstr. 20, 53343 Wachtberg, Germany

Abstract

Event extraction demands the detection of event descriptions in texts and the transformation of such descriptions into a standardised, structured format. It is framed as a natural language understanding task. In addition, *media event extraction* aims at structuring the media space and giving an overview on the topics being discussed and the dynamics of discourse. We will describe approaches to event extraction and media event extraction and discuss how these approaches can be linked in order to support the analysis of diverging world views in media discourse.¹

Keywords

distant reading, event extraction, media events, media space

1. Introduction

Our aim is to explore the media space, in particular social media in order to investigate the creation of conflicting world views. By “media space”, we refer to the very large, fast and continuously growing multilingual collection of texts, images, video and audio data that are distributed via traditional media as well as social media. Social media include YouTube, Twitter, Facebook, and other platforms [3]. A large part of the media space is accessible via the Internet. It contains huge amounts of “cultural data”, relevant for cultural analysis [4].

The media space provides information on the physical world: what happened? Which events are currently ongoing? What is planned or predicted to happen in the future? [5] Besides being a sensor for the physical world, the media space is a forum for ideologies, opinions and values. It is a space for the negotiation of what a society considers to be permissible, prescribed or forbidden, and for acting out sentiment and bias. As such, the media space is a research object for ideology analysis.

If the entire media space is considered a research object, then exhaustive close reading is plainly impossible. Therefore, media space analysis demands the development of suitable distant reading methods. A distant reader focuses on specific features of texts instead of reading the


¹This paper is based on our papers on “Distant and Reading and Event Extraction” [1] and “Conflict Monitoring” [2]. The former can be considered an extended abstract of the present paper. The latter examines the value of extracting and connecting details of real-world and media events from the media space, with the goal to use this information to enhance situational awareness in crisis management. For the present paper, we re-use text modules from both [1] and [2].

Humanities-Centred AI (CHAI), Workshop at the 44th German Conference on Artificial Intelligence, September 28, 2021, Berlin, Germany

✉ theresa.krumbiegel@fkie.fraunhofer.de (T. Krumbiegel); albert.pritzkau@fkie.fraunhofer.de (A. Pritzkau); hans-christian.schmitz@fkie.fraunhofer.de (H. Schmitz)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

texts completely. By reducing the reading effort in this way, she is enabled to receive large text collections. Focusing on specific features can also support her in finding the most relevant texts or text passages that are to be read closely [6, 7].

Distant reading requires automatic information reduction. Automatic information reduction includes the extraction and, possibly, transformation of relevant features from a corpus. Features can be defined as properties or attributes of the underlying data set. There is no definite path to identify the most relevant ones and only crude guidelines exist. It is very much guided by available data and the properties of the task to be solved [8]. With the objective to extract common and useful patterns from data, feature engineering must be regarded as a very important skill of the researcher [9]. The difficulty that comes with feature engineering and the effort involved are the main reasons for the emergence of algorithms that automatically engineer these feature representations. Feature learning algorithms and tools that find common patterns already exist – based on autoencoders and embeddings – and they are changing every aspect of how relevant features are identified, encoded, and distinguished from irrelevant ones [10].

In the present paper, we will discuss two approaches to extracting and presenting information from large text collections and thereby supporting distant reading. The focus of the paper is on the first of these approaches, that is, event extraction. Event extraction involves the recognition of event descriptions and the transformation of these descriptions into a standardized, structured format. The aim is to collect and display the events that the texts are about. In Section 2, we will propose a processing pipeline for event extraction. Texts cannot only be *about* events, but they can also constitute events themselves. We call such events “media events”. Tentatively, we define a media event as a the coherent reporting, commenting or discussion of a certain topic in the media. Media event extraction aims at structuring the media space and giving an overview on the topics being discussed and the dynamics of discourse. To this end, word-distribution based topic modellings and clustering can be applied. We briefly describe a media event extraction approach in Section 3. Finally, in Section 4, we will give an outlook on the linking of event and media event extraction and their application for ideology analysis.

2. Event Extraction

An important function of texts is to represent states of affairs and courses of actions. A text can be fictional or just false. Still, it describes a world, even if this world does not agree with what we consider to be *real*. In order to further analyse the signified, our aim is to extract event descriptions from text and transform these descriptions into standardized, structured representations. Structured event representations can serve as input for visualization and further investigation.

The most elementary representation of an event includes the event type, time and location. They can be inserted into a template that defines the structure of the event representation. Types are pre-defined. Times and locations can be defined with varying accuracy – e.g., locations can be specified as cities, by addresses, or by coordinates. Besides the basic information, further information on the actor, the affected, the reporter, etc. can be provided, if available. Event extraction has been studied for numerous years. Due to the continuous development in the field of natural language processing, approaches to event extraction are diverse [11].

Our processing pipeline for automatic event recognition, extraction and display consists of the following steps: first, wrappers collect input data. A wrapper can perform key word guided information retrieval and, therefore, function as a first filter. The usage of a keyword narrows down the subject area of the corpus. The collection of data can also happen non-specifically without a keyword. An example of a situation in which the collection is not keyword-triggered is when all recent articles from a specific feed (e.g., of the last two hours) are considered relevant and are included in the corpus accordingly. Which data is used as input data can be adapted depending on the use case. Examples are, among others, social media messages (e.g., obtained from Twitter) as well as texts extracted from online newspapers. We restrict ourselves to text analysis and leave images, videos and audio data out of consideration.

Second, using trained binary classification models, it is decided whether a retrieved document contains information on an event and is, thus, to be further analyzed. This task can be solved on the document-level or on the sentence-level and can therefore be seen as a two-step process. In the first step, larger entries within the corpus, i.e., entire documents are analysed and marked as including an event description or not. In the second step, specific sentences of the documents that include event descriptions are extracted. The extracted sentences are the ones holding the actual event information. The information space is inherently multilingual. Therefore, classification models should make use of multilingual language models. In a first experiment to solve the task of binary classification, we used one hundred separately trained densely connected neural networks for the document classification task and a single network for the classification on the sentence level [12]. Both approaches made use of document embeddings that are generated by using the pre-trained multilingual cased BERT model [13]. Our models were tested on multilingual data sets including English, Spanish, Portuguese and Hindi text instances. For document classification, a macro F1 (F-measure) of 0.65 was reached, for sentence level classification a macro F1 of 0.70. We plan to improve these scores by further developing our models and conducting additional experiments.

Third, document metadata are extracted and saved for further processing. Metadata of interest are, e.g., the author of the text as well as place and time of creation. The metadata obtained are used for data management. A comparison with existing entries in the data collection is performed. Entries that are already included are not incorporated again and are consequently excluded from further processing. This step is optional, as a further comparison of the text based on the extracted events is conducted in the following step. However, it can be useful to reduce the amount of data already at this stage to avoid the unnecessary processing of duplicates.

In the fourth processing step, different mentions of the same events are determined. Resolving co-references to events can be handled as a clustering task. Further analyzing different mentions supports the detection of contradictions between various mentions of the same event. One approach to solving the clustering problem is to train and optimize a simple neural network to compare two documents that both include an event description. This means that the neural network basically acts as a comparison function. The results of this comparison can then be used to build a graph. The graph consists of vertices and edges. The documents/sentences are represented by the vertices. If the network predicts that two documents/sentences belong to the same cluster, an edge is added in the graph between the corresponding vertices, otherwise no edge is added. The resulting graph is analysed with regard to disjoint subgraphs. Each individual subgraph represents an event cluster. Such a graph is shown in Figure 1.

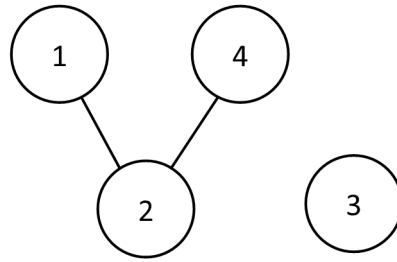


Figure 1: Example of a possible graph

In this simplified example graph, four documents including event descriptions are given. After processing these documents in the way that was described above, two event clusters are found. The first cluster includes documents 1, 2 and 4, the second cluster includes only document 3. We can determine that documents 1, 2 and 4 are about the same event and that document 3 is about another event.

Fifth, a fine-grained event classifier is used to determine the type of the detected event. Additional information, in particular on time and location are extracted. The event type, can be determined by using a fine-grained classification model. One database that can be used as training data for such a classifier is the Armed Conflict Location and Event Data Project (ACLED) database [14]. The ACLED database contains six event types and 25 sub event types. Fine-grained classification of events aims at the detection of the 25 sub event types. In general, all event and sub event types describe either a violent event (e.g. “battles” and “explosions”), a demonstration (e.g. “protests” and “riots”) or non-violent actions (e.g. “strategic developments”). Our current approach uses a fine-tuned RoBERTa transformer model [15] with document embeddings and ACLED data as its basis [16]. The model was evaluated on non-ACLED data in order to assess its robustness and reached a weighted F1 of 0.830. The model scores high for sub event types that can be seen as (semantically) compact, e.g., “suicide bomb” (F1 0.976) and “remote explosion” (F1 0.957) and low for sub event types that are less compact, e.g., the generic sub event type “other” (F1 0.400). These results were to be expected.

To calculate the topic compactness of each sub event type, we first embedded all examples. We then averaged the resulting vectors for each sub type. The average represents the topic centroid. Then, the Euclidean distance of each text vector to the topic centroid is calculated. Figures 2 and 3 show the topic compactness of the event sub types “other” and “suicide bomb” respectively.

In addition to the event types, i.e. the “What”, the “Where” (location) and “When” (time) of an event are crucial elements of an event representation. They need to be extracted. To this end, the text entries that remain after the previously described processing steps can be analysed with a Named Entity Recognition (NER) model. Named Entities are for example locations, organizations, persons and times/dates. A well-functioning NER model can provide relevant information about a given event almost entirely, including information about location and time and even involved actors. If one were to use a NER model for an entire document, problems would arise in terms of selecting the relevant information for a given event. Since in a previous processing step the exact event sentences were determined, meaning the sentences from a

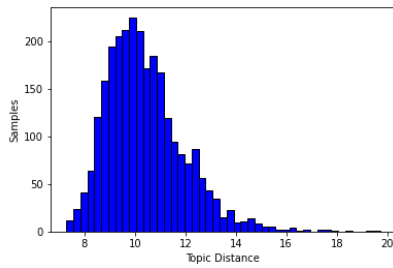


Figure 2: Topic Distance “Other”

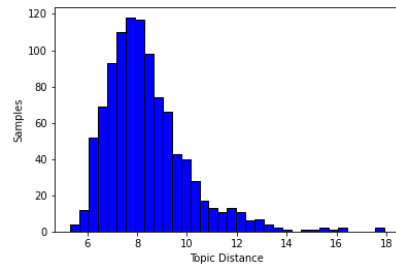


Figure 3: Topic Distance “Suicide Bomb”

document that contain the event arguments, we assume that the extraction of the applicable location, time and involved persons with a NER model can work in our case. An example using the spaCy NER model [17] is described below. To depict the results, we use an entry for the sub event type “peaceful protest” from the ACLED database (Figure 4).

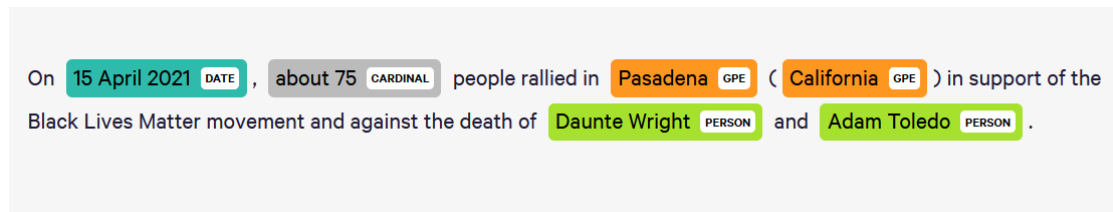


Figure 4: Example of NER

We can see that the NER model finds the time (DATE) as well as the location (GPE) of the event. Additionally, two persons are identified, namely “Daunte Wright” and “Adam Toledo” and a cardinal (“about 75”) is given. While information about the location and date of the event seem to be straightforward, the entities marked as person and cardinal need to be set into the correct context in order to understand their role.

For assigning meaning to the identified entities, Semantic Role Labeling (SRL) can be applied. SRL is “the computational identification and labeling of arguments in text.” It aims at determining “who” did “what” to “whom”, “where”, “when”, and “how” [18]. Thus, the arguments of a text that can be found with the help of SRL are in line with the arguments that we need to define a meaningful representation of an event. A number of annotation sets for SRL exist. We draw on the ones proposed for the Proposition Bank corpus [19]. SRL can help finding the correct context for previously identified entities. These relations then support the correct interpretation of already extracted event arguments. For the example given above, the SRL model by AllenNLP [20] provides the annotations displayed in Figure 5.

We see that the entities for date and location are marked as a temporal and a location modifier, respectively. This means that with regard to their function in the text, no new information is gained by using SRL. This was expected. However, the entities “about 75”, “Daunte Wright”

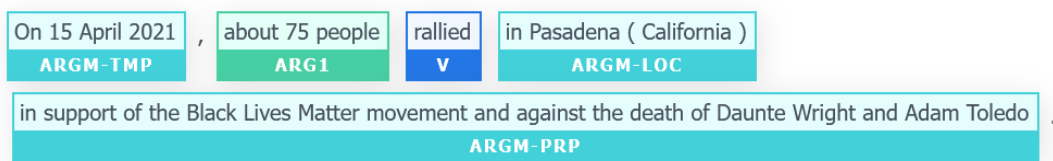


Figure 5: Example of SRL

and “Adam Toledo” are now set into the correct contexts. “About 75 people” which is not been detected as a person entity, is now marked as the agent of the event sentence. “Daunte Wright” and “Adam Toledo” are components of a purpose modifier, showing that they are not agents or patients of the event, but part of the reason the event is happening. These additional information contribute to a better understanding of the named entities and the event in general.

Sixth, all information is inserted into a predefined template, so that a standardized event representation is created. This template includes all of the event arguments mentioned above as well as a unique identifier. On the basis of the completed template, a further comparison of the events contained and identified in the database can be performed. This is necessary because it can be assumed that a specific event is not reported only once within the media space. In order to create a coherent situational picture, reports on identical events, insofar as they do not contain useful new information, do not all have to be processed further. In contrast to the identity management step mentioned earlier, the goal here is to find duplicate events that still may have come from different sources and not to find duplicated sources/reports/articles. Co-reference resolution already addresses the distinction or identification of event representations, however, it can be assumed that a specific event cluster that is found using the described approach for co-reference resolution, may include instances of the same main event but different sub events.

Seventh, this representation is transferred into a symbol from a given library. A viewer service draws the chosen symbol on a map. By extracting events from a text collection and drawing respective symbols on a map or creating time-lines of events, respectively, we represent situations and courses of action as they are described in the text collection.

3. Media Event Extraction

A media event is the coherent reporting, commenting or discussion of a certain topic in the media. Media events can be very short, e.g., when a specific incident is reported only once and then “forgotten”. However, media events can also be much longer, e.g., when information is progressively updated, commented and discussed. In that case, different reporters can contribute to the same media event. Topics of media event can be entities like persons, institutions etc., or other events, among them both physical world events – events as described in the previous Section, including events that never took place – and other media events.

Media event extraction aims at structuring the media space and giving an overview on the topics being discussed and the dynamics of the discussions. In essence, it can be implemented as a

clustering of the media space and the distillation and description of the clusters' features that are deemed relevant. It enables browsing through topics and the distant reading of communication threats.

A paradigmatic use case is the investigation of social media discourse. Topics are modelled as distributions over content words derived from documents. To this end, we apply Latent Dirichlet Allocation (LDA, [21]): based on the vocabulary of a document, topics can be assigned to it with a certain probability. Therefore, topics give rise to a soft, i.e., probabilistic clustering. Documents can also be clustered directly without creating topic models first. Clustering [22] is applied to document representations capturing also contextual information, by making use of automatic feature engineering such as autoencoders [10]. Resulting clusters can be described by their characteristic keywords in a following step.

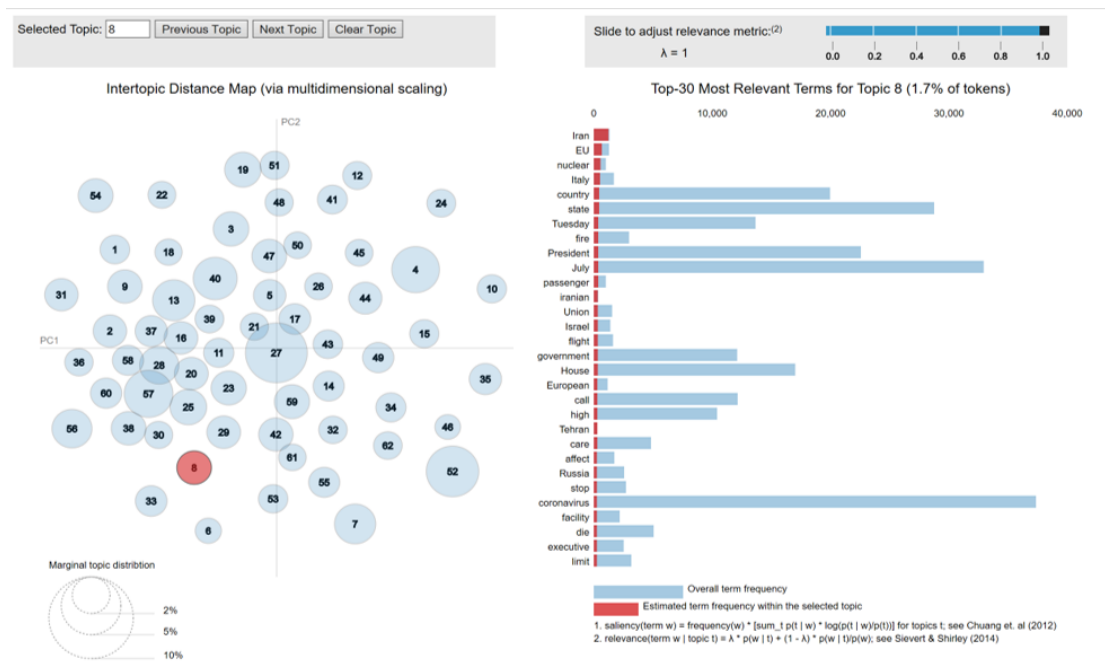


Figure 6: Topic visualisation on news articles

Figure 6 shows a map of topics for a text collection (left) and a key word distribution for a selected topic (right).

Topic models and text clusters give the distant reader an overview on the topics under discussion and enable her to identify relevant clusters that are to be investigated further. Topics that are significantly prevalent in an underlying data set can be considered as correlated with media events. The temporal distribution of the volume flow of a topic can be understood as communication behaviour, that is, the internal dynamics of a media event.

4. Conclusion and Outlook

We have described two different techniques for distilling information from large text collections, namely *event extraction* and *media event extraction*. By event extraction, information on events is identified in texts and transformed into standardized, structured representations. These representations give an overview on the events that are referred to within the texts. They display an essential aspect of the world view conveyed by a text collection. We can compare different text collections regarding the events they describe and, thus, extract the divergences of their world views.

By media event extraction, we identify and describe the topics being under discussion in the media space. If we take the time dimension into consideration, we can describe the dynamics of discourse and topic changes.

It remains an interesting challenge to link event representations and topic/media event representations in order to answer questions on the discursive context of events and on the roles events play in creating world views through texts. An obvious link between the two kinds of representations is that (physical world) events can be topics themselves or occur in topics. Topic representations can then help to describe the narratives around these events.

A further way to link events and topics is via social networks. Persons can be involved in topics, either actively or as being affected. Persons are related to each other by occurring within the same topics, possibly in different roles. Moreover, they act as authors, recipients and/or intermediaries in the exchange and distribution of information. Just like persons appear in media events, they also appear in physical world events, again taking various roles. They are connected both via the events themselves and via their reporting. Thus, networks of persons (social networks) can be connected to both events and media events and thereby be a glue for linking them.

Social network analysis enables us to attribute event descriptions and media events to social groups. We assume that event descriptions and topics are a plausible basis for describing the world views of these groups and analyse potential conflicts between them. Thus, we claim the hypothesis that event extraction, media event extraction and social network analysis can be promising tools for ideology analysis. "Further research is needed."

References

- [1] T. Krumbiegel, A. Pritzkau, H.-C. Schmitz, Distant reading and event extraction (2021). URL: https://www.fdr.uni-hamburg.de/record/9672/files/KI2021_CHAI_submission3.zip?download=1.
- [2] S. Kent, T. Krumbiegel, A. Pritzkau, H.-C. Schmitz, Conflict monitoring, in: Artificial Intelligence, Machine Learning and Big Data for Hybrid Military Operations (AI4HMO), NATO, 2021.
- [3] M. Andree, T. Thomsen, Atlas der Digitalen Welt, Campus, 2020.
- [4] L. Manovich, Cultural Analytics, MIT Press, Cambridge/Mass, 2020.
- [5] P. W. Singer, E. T. Brooking, Like War. The Weaponization of Social Media, First Mariner Books, 2019.

- [6] F. Moretti, *Distant Reading*, Verso, London, 2013.
- [7] S. Jänicke, G. Franzini, M. F. Cheema, G. Scheuermann, On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, in: R. Borgo, F. Ganovelli, I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis) - STARs*, The Eurographics Association, 2015. doi:10.2312/eurovisstar.20151113.
- [8] P. Domingos, A few useful things to know about machine learning, *Commun. ACM* 55 (2012) 78–87. URL: <https://doi.org/10.1145/2347736.2347755>. doi:10.1145/2347736.2347755.
- [9] K. Krippendorff, *Content analysis: An introduction to its methodology*, 4th ed., Sage, Los Angeles, 2018.
- [10] I. Goodfellow, Y. Bengio, C. A., *Deep Learning*, MIT Press, Cambridge/Mass, 2016.
- [11] W. Xiang, B. Wang, A survey of event extraction from text, *IEEE Access* 7 (2019) 173111–173137. doi:10.1109/ACCESS.2019.2956831.
- [12] N. Becker, T. Krumbiegel, FKIE_itf_2021 at CASE 2021 task 1: Using small densely fully connected neural nets for event detection and clustering, in: *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, Association for Computational Linguistics, Online, 2021, pp. 113–119. URL: <https://aclanthology.org/2021.case-1.15>. doi:10.18653/v1/2021.case-1.15.
- [13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR* abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [14] C. Raleigh, A. Linke, H. Hegre, J. Karlsen, Introducing acled: An armed conflict location and event dataset: Special data feature, *Journal of Peace Research* 47 (2010) 651–660. URL: <https://doi.org/10.1177/0022343310378914>. doi:10.1177/0022343310378914. arXiv:<https://doi.org/10.1177/0022343310378914>.
- [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
- [16] S. Kent, T. Krumbiegel, Case 2021 task 2 socio-political fine-grained event classification using fine-tuned roberta document embeddings, 2021, pp. 208–217. doi:10.18653/v1/2021.case-1.26.
- [17] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.
- [18] L. Màrquez, X. Carreras, K. C. Litkowski, S. Stevenson, Semantic Role Labeling: An Introduction to the Special Issue, *Computational Linguistics* 34 (2008) 145–159. URL: <https://doi.org/10.1162/coli.2008.34.2.145>. doi:10.1162/coli.2008.34.2.145.
- [19] M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: A corpus annotated with semantic roles, *Computational Linguistics Journal* (2005).
- [20] P. Shi, J. Lin, Simple bert models for relation extraction and semantic role labeling, *ArXiv abs/1904.05255* (2019).
- [21] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022.
- [22] D. Anastasiu, A. Tagarelli, G. Karypis, *Data Clustering. Algorithms and Applications*, 2nd. ed., Chapman and Hall/CRC, Boca Raton/Fl, 2018, pp. 305–338.