Conflict Monitoring

Samantha Kent
Fraunhofer FKIE
GERMANY

samantha.kent@fkie.fraunhofer.de

Theresa Krumbiegel
Fraunhofer FKIE
GERMANY

theresa.krumbiegel@fkie.fraunhofer.de

Albert Pritzkau
Fraunhofer FKIE
GERMANY

albert.pritzkau@fkie.fraunhofer.de

Hans-Christian Schmitz
Fraunhofer FKIE
GERMANY

hans-christian.schmitz@fkie.fraunhofer.de

ABSTRACT

Hybrid warfare includes the fuelling of conflicts in order to weaken the opponent. Respective operations take place both in the physical world and within the media space (which is often called "information space"). Defense against hybrid warfare demands comprehensive situational awareness which requires intelligence in both domains, i.e., the physical and the media. To this end, Open Source Intelligence (OSInt) is tasked with analysing openly accessible information from the media space. As the media space is very large and continuously growing, OSInt requires technological support. In this paper, we will describe the automatic detection and extraction of events in the physical world as well as media events. We will discuss how event representations of different types can be related to each other, and how a network of event representations can contribute to situational awareness.

1.0 INTRODUCTION

Open Source Intelligence (OSInt) is tasked with exploring and analysing the openly accessible media space in order to gather information on (potential) conflicts, among other topics. With the term "media space", we refer to the very large, fast and continuously growing multilingual corpus of texts, images, video and audio data that are distributed via traditional media, such as television, radio and newspaper, as well as social media, including various web blogs. Social media are mostly platform bound. Platforms include YouTube, Twitter, Facebook, Instagram, and others [1,2]. To a large extent, the media space is accessible via the Internet. Very many parts are open to the public. However, there also exist semi-open areas with potentially valuable information that are not intended to be accessible to everyone, e.g., Telegram groups and Facebook pages.

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? It reacts very quickly to events in the physical world, i.e., information is provided almost immediately [3]. Therefore, the media space seems to be a promising "sensor" for events in the physical world. However, it remains a challenge to retrieve particularly relevant information from the overwhelmingly large body of information, as by far most information provided is entirely irrelevant, at least to the military. Moreover, the media space is not consistent – it includes both true and false information and, thus, fact checking is a further challenge.

Besides being a sensor for the physical world, the media space is a forum for ideologies, opinions and values. It is an important space for the negotiation of what a society considers to be permissible, prescribed or forbidden, and for acting out sentiment and bias. As such, it has become a theatre of operations for hybrid warfare, that is, for operations conducted with the aim "to disrupt, undermine or damage the target's political system and cohesion through a combination of violence, control, subversion, manipulation and dissemination of (mis-) information" ([4], p.2). (Mis-) Information operations cause what we call "media events". Media events can be triggered in order to influence sentiments, ideology and the public view on the physical world.

Thus, on the one hand, the media space provides representations of the physical world that can be evaluated for gaining situational awareness. On the other hand, the media space is an area of operations in itself. It constitutes a domain of ideas for which situational awareness has to be achieved too.

The media space is huge and grows continuously. For human analysts it is impossible to view the full amount of potentially relevant information closely. Therefore, methods of distant reading have to be developed. Distant reading requires defining features that are particularly relevant to a specific question, in other words, a collection task. These features are extracted from a corpus of texts or other documents. The features are processed and visualized [5]. This enables the analyst to perceive the relevant information from the corpus without reading its texts closely. Distant reading is a technique originally developed within cultural studies and the digital humanities [6]. In OSInt, it supports the most fundamental task of the analytic spectrum, namely descriptive analysis, both of events in the physical world and of media events. (Within the analytic spectrum, descriptive analysis precedes explanatory analysis, evaluative analysis and, finally, estimative analysis [7].)

How can we enable distant reading? Distant reading requires automatic information reduction. Automatic information reduction includes the extraction and, possibly, transformation of relevant features from a corpus. Appropriate features have to be specified. Then, it has to be determined which of these features are to be actually derived in order to answer a respective question. Information reduction should depend on the collection task. Effective distant reading support should enable users to apply various information reduction techniques that serve the answering of diverse questions. Users should be provided with a respective toolbox.

In this paper, we will discuss the extraction of information about physical real-world events from the media space. To this end, we will define structured event representations and propose a pipeline for automatic event recognition and extraction (Section 2). Then, we will discuss media events and the features needed to identify and represent them (Section 3). Stakeholders, in particular agents and patients, can be considered a glue for bringing together different event types, so that a broader picture of the events in multiple domains can be formed. Events can be related via their stakeholders through social networks. Therefore, we will discuss social network analysis in Section 4. Finally, we will tentatively bring together the various analytic approaches and give an outlook on future work.

2.0 EVENT EXTRACTION

The most basic representation of a physical world event comprises its type – e.g., "IED event", "suicide bomb", "peaceful protest", "attack" – as well as its location and time. Types are pre-defined and locations and times can be defined with varying accuracy – e.g., locations can be specified as cities, by addresses, or by coordinates. Type, time and location are the fundamental features needed to classify the event and represent it on a map or in a timeline. Beyond the basic information, additional information on the actor, the affected and other stakeholders as well as used materiel can be provided, if available. Event representations can be linked to other data, such as responsibility claims, photos or similar. Furthermore, metadata can specify the information source as well as sentiments of the reporter, regarding the event.

We start with the very basic representation of an event.

In order to achieve the goal of coherent event recognition and extraction, we invoke the concept of a processing pipeline, which includes several sub-tasks to aid in event extraction. The first step of this processing pipeline is data collection. Here, a wrapper is to be implemented that extracts data from the media space. The collection of data can either be specifically triggered by a keyword or can 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 source (e.g., of the last two hours) are considered relevant and are included in the corpus accordingly. Based on the specified keyword(s), the subject area of the corpus can already be greatly narrowed down.

To further limit the database, an initial classification is performed. A binary classification model is used to determine whether an entry in the corpus contains an event description or not. This classification task can be implemented on the document level and/or on the sentence level. Therefore, it can be seen as a two-step process. Firstly, larger entries within the corpus, i.e., entire documents are analysed and marked as including an event description or not. Secondly, specific sentences of the documents that include event descriptions are extracted. The extracted sentences are the ones holding the actual event information. Due to the inherent multilinguality of the information space, classification models should make use of multilingual language models accordingly. In a first implementation, we use one hundred separately trained densely connected neural nets for the document classification task and a single net for the classification on the sentence level [8]. Both approaches make use of document embeddings that are generated by using the pre-trained multilingual cased BERT model [9]. Our models were tested on multilingual datasets 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 conducting further experiments in the future.

Documents and sentences that include an event description are processed further. Metadata of the relevant corpus entries are read out. Metadata of interest are, e.g., the author of the text as well as place and time of creation. The metadata obtained is 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. Additionally, since the data evaluation is often implemented quickly, especially when it comes to the analysis of current crisis situations, it is advisable to reduce the amount of data to a minimum as early as possible.

Next, co-reference resolution is performed on the resulting filtered corpus. It needs to be determined which documents and/or sentences describe the same event. The task of co-reference resolution can be treated as a clustering problem. 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 could look as follows:

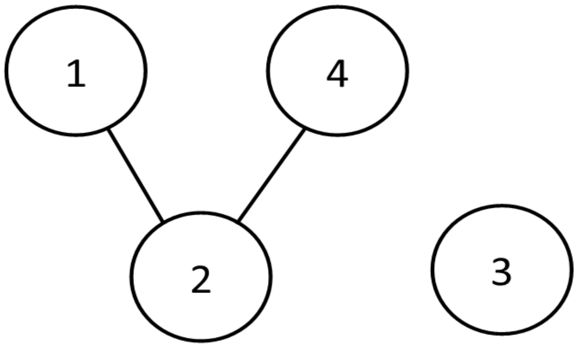

Figure 1: Example of a possible graph

In this simplified example graph, four documents including an event 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.

The next step in our proposed processing pipeline is the extraction of event arguments. This means that the answers to the "5W1H" questions ("who, when, where, what, why" and "how") have to be extracted. The "what", i.e., the event type, can be determined by using a fine-grained classification model. One database that could be used as training data for such a classifier is the Armed Conflict Location and Event Data Project (ACLED) database [10]. 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, a demonstration or non-violent actions. The six event types are "battles", "explosions/remote violence", "violence against civilians", "protests", "riots" and "strategic developments". The sub event types define these event types in more detail and include, among others, "armed clash", "suicide bomb", "attack", "grenade", "mob violence", "arrests", "peaceful protest" and "agreement". Our current approach uses a fine-tuned RoBERTa transformer model [11] with document embeddings and ACLED data as its basis [12]. 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 average 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.

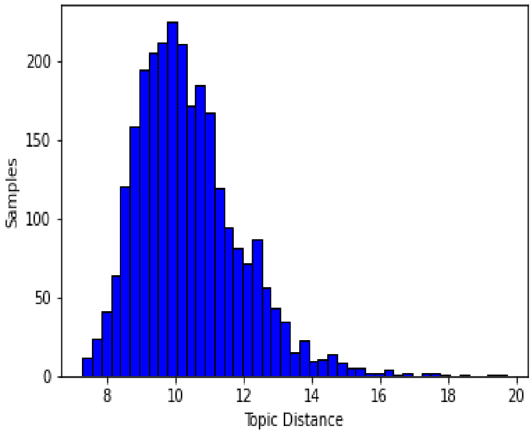

Figure 2: Topic Distance “other”

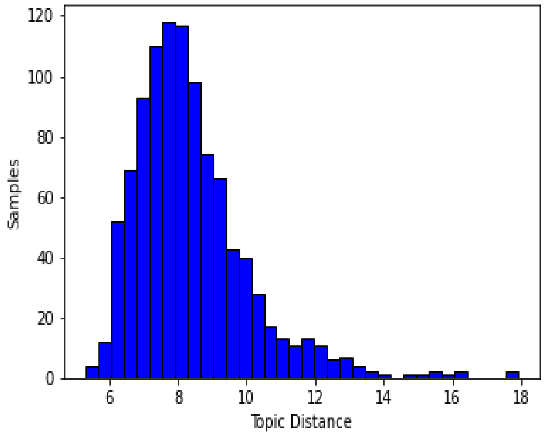

Figure 3: Topic Distance “suicide bomb”

In addition to the event types, i.e. the “What”, the “Where” and “When” of an event are crucial in order to create situational awareness. This means that the location and time of an event 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 time/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 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 [13] is described below. To depict the results, we use an entry for the sub event type “suicide bomb” from the ACLED database.

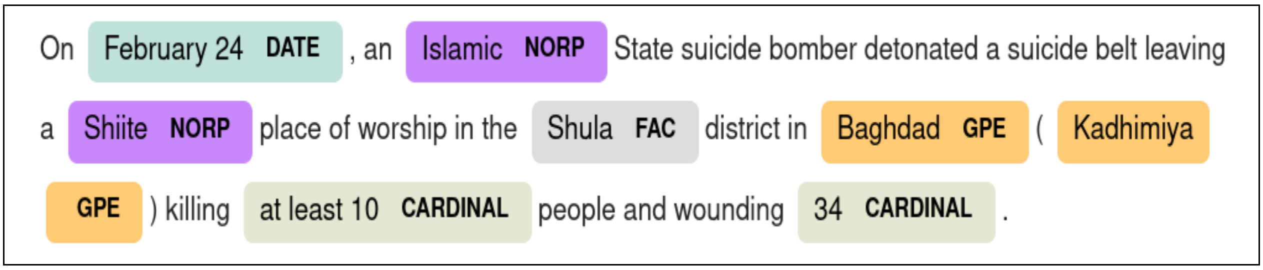

Figure 4: Example of NER

We can see that the NER model finds the time (DATE) as well as the location (FAC, GPE) of the event. Additionally, nationalities and religious groups (NORP) are identified. To use the information provided by the tags NORP and CARDINAL – e.g., for determining the actor (“Islamic State suicide bomber”) and the affected place (“Shiite place of worship in the Shula district in Bagdad”) – further processing would have to be implemented, as these entities have to be seen in the correct context.

Now, the event 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.

In a final step, the standardized representation that was creating using the template is transferred into a symbol library that is based on the APP 6C's Stability and Civil Support Activities Symbols [14]. A viewer service can draw the selected symbol on a map. The goal to depict events as symbols on a map strengthens the necessity for identity management based on event information, as each event should only be displayed once.

In summary, our proposed processing pipeline enables the analysis of reports about events in the media space and connects the gained information with a picture of the physical world, thereby creating situational awareness. The implementation of several classification and identity management steps helps to reduce the amount of data that needs to be analysed.

3.0 MEDIA EVENT EXTRACTION

We propose a broad definition of media events: 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 and other media events. Topics and events can be hierarchically structured. That is, topics can contain sub-topics and media events can consist of several sub-events. It is not impossible that, at some point in time, a media event recursively refers to itself.

Propaganda and operations of (mis-) information create media events with the aim to influence sentiment, opinion and, ultimately, behaviour. Such operations are a difficult business, as success is far from guaranteed [15]. However, social media seem to provide opportunities [3]. Institutional burdens to participate in the media space are reduced, if not abolished. Fragmentation of public discourse can be exploited: it becomes possible to target selected audiences, adapt to their specific expectations and biases and, thereby, aim at influencing them without making other audiences aware of the influencing. As there is in principle no quantitative limit on contributions to the media space, sections of this space can be flooded with (mis-) information. The goal can be to convince the audience of a certain position. However, it can also be just to sow doubt or insert large amounts of noise in order to overlay other information. To this end, algorithmic methods of content generation and distribution can be applied. Finally, recommendation engines of social media platforms do their part in supporting the spread of (mis-) information and the creation of "hypes" [1].

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, thus, the distant reading of communication threats. In our analysis we focus on selected social media platforms, e.g., Twitter. We apply techniques of content analysis [16], including automatic text analysis/ text mining. Content analysis is a scientific method used to systematically and reliably quantify communication based on the symbols used. Methodical approaches to content analysis, text analysis and text mining examine artefacts of social communication, typically given by written documents or transcripts, to enable inferences about a subject or topic of interest, the framing of conversations and many other characteristics of messages.

The analysis of text collections involves breaking down the content of documents in order to extract meaningful units and relations. To this end, natural language processing (NLP) techniques are being applied. For further analysis, a range of methods of automatic text analysis, particularly useful in web data analysis, take advantage of association rules and sequential patterns, as well as supervised and unsupervised learning. Increasingly, statistical algorithms from machine learning are finding their way into the standard methods of automated text mining. The development of unsupervised or semi-supervised learning methods for natural language processing seems to be particularly promising, since traditional supervised methods are unlikely to scale to the vast amounts of possibly relevant data, which is available in the media space. The primary advantage of these methods is that they do not require annotated data, as training data is often either limited or non-existent for most structures.

Unsupervised machine learning methods, such as document clustering [17] and topic modelling, can significantly contribute to the exploratory analysis of a text corpus. By identifying homogenous groups of documents/messages, these methods are able to reveal hidden patterns as well as the underlying topic structure. As an example, let us consider Latent Dirichlet Allocation (LDA). LDA is a generative statistical model proposed by Blei, Ng and Jordan [18]: each document of a collection is modelled as a finite mixture over an underlying set of topics, resulting in a three-level hierarchical Bayesian model. Clustering of a text corpus into a set of topics based on the keywords of the individual documents is an example of LDA usage where each document is described as a distribution of topics, and each topic as a distribution of words. We applied LDA to a collection of news articles. The interpretation of a selected topic in Figure 5 results from the term frequency of corresponding terms (red). The trained model provides a soft clustering of the documents. Converting this to a partitioning grouping is done by assigning topic with highest probability to each document.

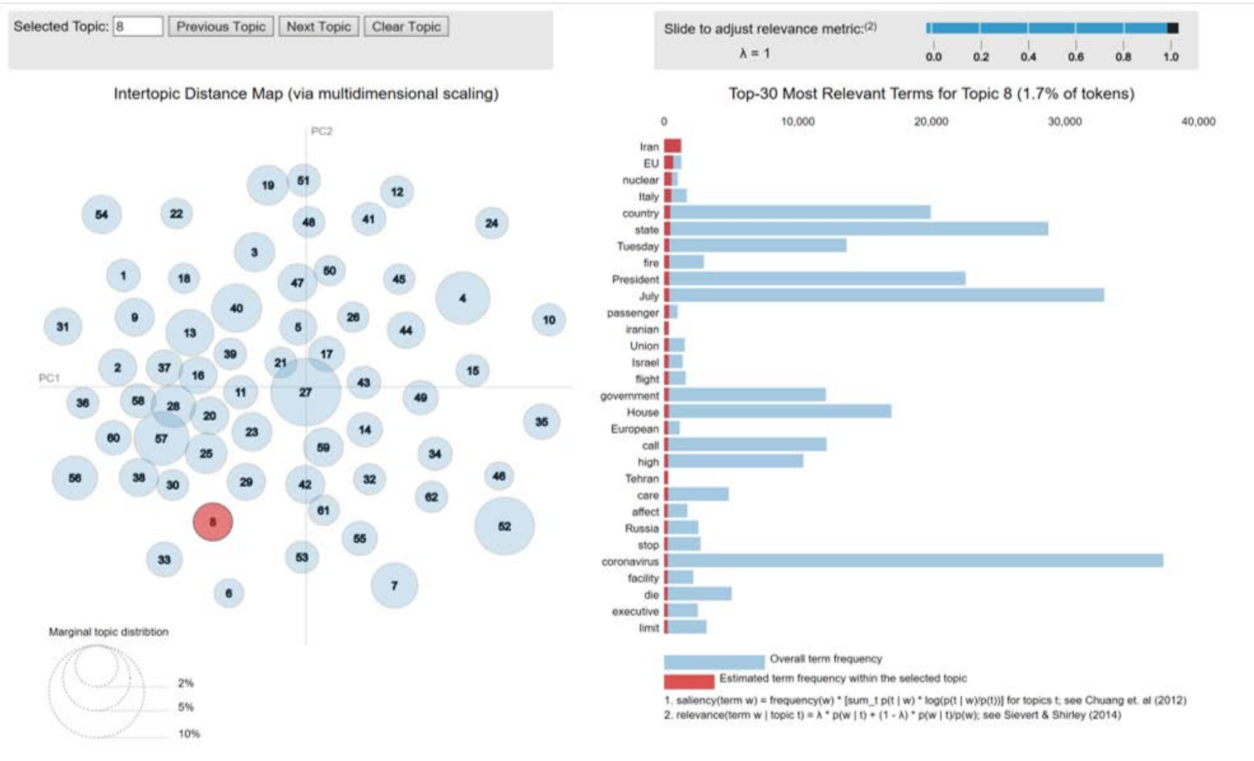

Figure 5: Topic visualisation on news articles

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. Consider Figure 6: the diagram shows the number of messages sent by two different outlets, distinguished by their colours (read vs blue). In our example, the outlets are Twitter accounts. Strong spikes of the red outlet may indicate algorithmic support, i.e., the automatic generation and distribution of messages. The distinction between the different outlets also enables the analysis of their different roles: while the red outlet sets the theme, the blue outlet takes a merely reactive role.

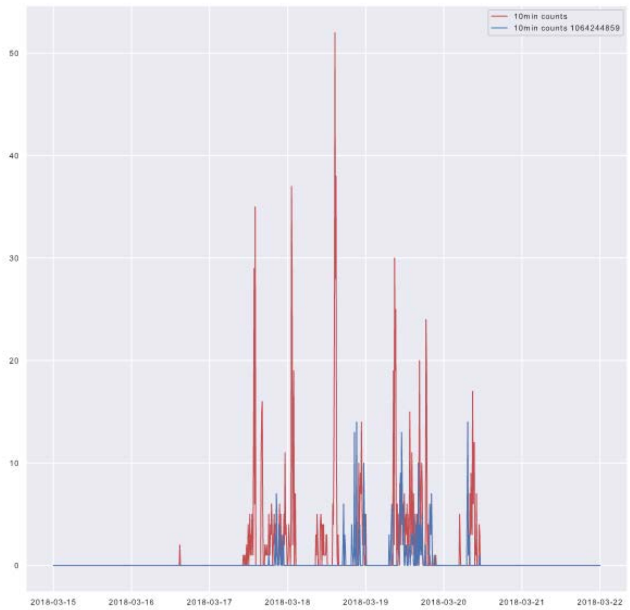

Figure 6: Temporal distribution of two different outlets on a specific topic

4.0 SOCIAL NETWORK ANALYSIS

Media events include persons. Persons can be involved in topics, either actively or as being affected. They can even *be* topics themselves. Persons are related to each other by occurring within the same topics, possibly in different roles. Moreover, persons act as authors, recipients and/or intermediaries in the exchange and distribution of information. They are related to others via the information distribution network underlying the media space. Just like persons appear in media events, they also appear in real world events, again taking various roles. They are connected both via the events themselves and via their reporting, i.e., the respective media events that take real world events as topics.

We can create graphs with persons as vertices and their event induced relations as edges. These graphs represent the social dimension of the event space and the media space, in particular. Events can be added as extra vertices, thereby connecting events via the social network. Events and their media sources can be clustered regarding their appearance in the social dimension. Thereby, an additional approach to distant reading of sources is enabled, and the media space can be browsed through its social dimension. Moreover, social network representations can serve information retrieval by helping to detect the information that is crucial for creating communities and, thus, should be read closely.

An enormous amount of research has been performed in recent years on exploiting data collected from social media services. Meanwhile, technology is mature and easy to apply [19,20]. In particular, it has been observed that the correlation of the linkage structure as an intrinsic and discriminating feature of these services with different kinds of data can increase the performance of mining and searching in information networks. When trying to assess anomalous patterns, events and changes, it is crucial to understand the underlying social network. Thus, mining the attributes in addition to the contents of social media services provides the opportunity to discover social structure characteristics, and analyse activity patterns both qualitatively and quantitatively. Emerging approaches seek to explain and predict social interactions through network structures by examining patterns of links among vertices. They favour the structural characteristics of the network (i.e., the patterns of connections among vertices) over individual attributes of the vertices. This shift in social analysis away from atomistic approaches to contextual and relational ones led to the ability to understand and explain the behaviours of multiple social formations such as technology-mediated communities.

We envision that the whole phenomenon of social networks will continue to evolve quickly as digital technology increasingly penetrates the realm of the physical world, providing new research challenges for information systems. Since most current social network services usually implement only very simple models of a social network, it should be noted that these models cannot mirror the richness of real world complexity. However, even these abstract representations of social dynamics have proven to be useful in the acquisition of knowledge for decision making [21].

5.0 SYNTHESIS AND OUTLOOK

To conclude: conflict monitoring shall provide us with situational overviews of both the physical space and the media space. OSInt contributes to these situational overviews by observing and analysing the openly accessibly part of the media space. OSInt requires technological support in order to deal with the vast amount of information that is available. Basic operations of information retrieval, extraction and distillation should be automatized, so that human analysts can better focus on the more challenging tasks of the analytic spectrum, namely explanation, evaluation and estimation.

We have proposed a pipeline for extracting information on real world events from texts. The output of this pipeline are structured event representations. These representations can contribute to the situational picture: respective symbols can be selected and drawn on a map (Section 2). Moreover, we have proposed an approach to the identification and description of media events. We assume that media events can be correlated with topics under discussion in the media space. Therefore, topic modelling can provide us with an overview over media events. Further insight can be gained by analysing topic changes and the dynamics of information exchange within topics (Section 3).

Real world events are represented by structured information while topics and, thus, media events are represented by mere distributions of content words. These representations differ in type and are not easily compatible. It remains a challenge to integrate situational pictures of the physical world and the media space in order to facilitate the answering of questions like to following: what is the discursive context of events in the physical world? How does the media space react to such events? How are physical events being exploited for media events? And ultimately: are there indications within the media space that enable the prediction of real world events? Linking the physical domain to the information space is not a trivial task. This is due to the fact that information operations in particular, as media events, do not presume this link to the physical domain. With respect to physical events, one can of course attempt to determine its media exploitation. We assume the relation in the opposite direction to be much more difficult or even is not feasible in most cases. A nearby attempt to integrate the physical and the media domain is by linking events of various types via the involved persons, as we proposed in Section 4.

Further work is on improving the models for (media) event extraction and for evaluating these models with various data sets. It is important to note, that within the media space – in particular within that part which is relevant for conflict monitoring – English is not the only language. Thus, improved solutions must provide multilingual support, also for underrepresented languages.

REFERENCES

[1] Sinan Aral (2020): *The Hype Machine. How Social Media Disrupts Our Elections, Our Economy and Our Health – and How We Must Adapt*, Harper Collins Publishers: London.

[2] Martin Andree & Timo Thomsen (2020): *Atlas der Digitalen Welt*, Campus: Frankfurt, New York.

[3] Peter W. Singer & Emerson T. Brooking (2019): *Like War. The Weaponization of Social Media*, First Mariner Books: Boston, New York.

[4] Mikael Weissmann, Niklas Nilsson, Björn Palmertz & Per Thunholm (eds.) (2021): *Hybrid Warfare. Security and Asymmetric Conflict in International Relations*, I.B. Tauris: London et al.

[5] Lev Manovich (2020): *Cultural Analytics*, MIT Press: Cambridge/ Mass., London.

[6] Franco Moretti (2013): *Distant Reading*, Verso: London.

[7] Katherine H. Pherson & Randolph H. Pherson (2020): *Critical Thinking for Strategic Analysis*, 3rd ed., CQ Press: Thousand Oaks, London.

[8] Nils Becker & Theresa Krumbiegel (2021): FKIE_itf_2021 at CASE 2021 Task 1: Using Small Densely Fully Connected Neural Nets for Event Detection and Clustering. In *Proceedings of the Workshop on Challenges and Applications of Automated Text Extraction of Socio-Political Event from Text (CASE 2021), co-located with the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 113–119, online. Association for Computational Linguistics (ACL).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics (ACL).

[10] Clionadh Raleigh, Andrew Linke, Havard Hegre, and Joakim Karlsen (2010): Introducing acled: An armed conflict location and event dataset: Special data feature. In: *Journal of Peace Research*, 47(5):651–660.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019): Roberta: A robustly optimized bert pretraining approach. Cite arxiv: 1907.11692.

[12] Samantha Kent & Theresa Krumbiegel (2021): CASE 2021 Task 2: Socio-political Fine-grained Event Classification using Fine-tuned RoBERTa Document Embeddings. In *Proceedings of the Workshop on Challenges and Applications of Automated Text Extraction of Socio-Political Event from Text (CASE 2021), co-located with the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 208–217, online. Association for Computational Linguistics (ACL).

[13] Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

[14] North Atlantic Treaty Organization (2017): *NATO STANDARD APP-6 NATO JOINT MILITARY SYMBOLOGY*. NATO Standardization Office (NSO).

[15] Philip M. Taylor (2003): *Munitions of the Mind: A History of Propaganda from the Ancient World to the Present Era*, 3rd ed., Manchester University Press: Manchester.

[16] Klaus Krippendorff, K. (2018): *Content analysis: An introduction to its methodology*, 4th ed., Sage: Los Angeles et al.

[17] David C. Anastasiu, Andrea Tagarelli & George Karypis (2018): Document Clustering: The Next Frontier, in Charu C. Aggarwal & Chandan K. Reddy (eds.), *Data Clustering. Algorithms and Applications*, Chapman and Hall/CRC: Boca Raton/Fl., 305-338).

[18] David Blei, Andrew Y. Ng, & Michael I. Jordan (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

[19] Krishna Raj P.M., Ankith Mohan & K.G. Srinivasa (2019): *Practical Social Network Analysis with Python*, Springer: Heidelberg.

[20] Marcello Trovati (015): Mining Social Media: Architecture, Tools, and Approaches to Detecting Criminal Activity, in Babak Akhgar, Gregory B. Saathoff, Hamid R. Arabnia, Richard Hill, Andrew Staniforth & Petra Saskia Bayerl (eds.), *Application of Big Data for National Security. A Practitioner's Guide to Emerging Technologies*, Elsevier: Amsterdam et al.

[21] Albert Pritzkau (2016): Sensemaking for Intelligence Analysis on Social Media Data, in *Proceedings of SPIE 9851, Next-Generation Analyst IV, 98510J* (12 May 2016), https://doi.org/10.1117/12.2242537.