Event Detection in the Socio Political Domain

Emmanuel Cartier, Hristo Tanev
European Commission, Joint Research Center
Via Enrico Fermi, 2749
21027 Ispra (VA), Italy
emmanuel.cartier@ec.europa.eu, hristo.tanev@ec.europa.eu

Abstract

In this paper we present two approaches for detection of socio political events: the first is based on manually crafted keyword combinations, and is implemented inside the NEXUS event extraction system, and the second one is based on a BERT classifier. We compare the performance of the two systems on a dataset of socio-political events. We also evaluated only NEXUS on the ACLED event dataset, in order to show the effects of taxonomy mapping and the performance of rule based approaches. Interestingly, both systems demonstrate complementary performance. Both showing their best performance on different event type sets. Nevertheless, an LLM data augmented dataset shows that in this case the transformer-based system improves considerably. We also review in the related work section the most important resources and approaches for event extraction in the recent years.

1. Introduction

1.1. NEXUS event taxonomy

Event extraction started to emerge as a Computational linguistics topic of interest, in relation to the enormous stream of events reported in mainstream media and commented and repeated in the social networks (Kounadi et al., 2015). Event extraction is used in a wide range of applications in diverse domains and has been intensively researched for more than three decades, starting with the seminal works, inspired by the Message Understanding Conferences (Chinchor and Marsh, 1998). It has a large range of applications in policy making, security, disaster management, health, bio medical research, as well as in the domain of business and finances.

In recent years, the significance of event extraction in the socio political domain has garnered considerable attention from the research community. This heightened interest stems from the critical nature of socio-political phenomena and the escalating societal and political tensions witnessed over the past half-decade, attributed to events such as the COVID-19 pandemic, the conflict between Russia and Ukraine, and various other theatres of war, notably in the Middle East. The significance of event extraction technology in the socio-political realm has been underscored in recent workshops such as the CASE (Challenges and Application of Automated Extraction of Socio-political Events) series (Hürriyetoğlu et al., 2021) and other similar venues.

The purpose of this paper is to throw light on the most important approaches and resources for event extraction in the last years, illustrating the two predominant paradigms for event detection: the rule based and the statistical one by evaluating two event detection systems.

The statistical system is based on XLM RoBERTa-base statistical classifiers and the rule-based system Tanev et al. (2008), NEXUS, uses a set of manually curated boolean combinations of keywords. We compare the performance of the two systems on the JRC corpus of crisis events (Atkinson et al., 2017a). We additionally evaluate NEXUS on a subset of the ACLED dataset, which is a standard in the socio-political field, (Raleigh et al., 2010). The purpose of this evaluation was to study how well the NEXUS event types map to the ACLED taxonomy and to explore the effect of taxonomy alignment in the evaluation of event extraction systems.

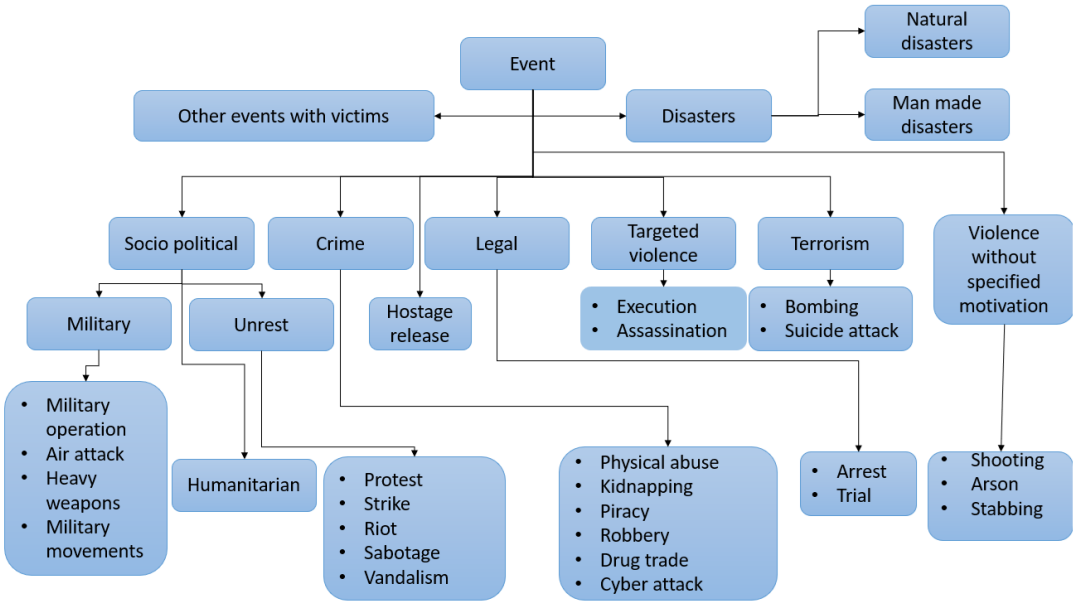
2. Related work

Rule-based event extraction was a predominant paradigm in the early systems in the nineties , as well in the next decade Aone and Ramos-Santacruz (2000); Grishman et al. (2002b,a). However, with the advent of the "big data" paradigm, state-of-the-art research experiments nearly entirely shifted towards the domain of Machine Learning (ML) and Large Language Models (LLM) (Hürriyetoğlu et al., 2021). Nevertheless, rule based systems have been dominating the industrial landscape (Chiticariu et al., 2013) and still provide basis for event detection in the domain of security and media analysis Tanev et al. (2008); Nitschke et al. (2022); Hamborg et al. (2019).

Building machine learning models for event detection was greatly facilitated by the availability of annotated event corpora and event databases. Among the known event corpora, one of the most used one is the ACE corpus (Consortium et al., 2005). Recently, the Joint Research Centre of the

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 12–21

Figure 1: NEXUS event taxonomy



European Commission proposed two corpora, one of them based on the output of the NEXUS event extraction system, (Atkinson et al., 2017b), and the other one containing events related to the COVID pandemic (Piskorski et al., 2023). But this second one has another ontology than the NEXUs system and won't be used.

Security-related event databases (DB) are manually curated, such as ACLED (Raleigh et al., 2010) or automatically created, such as GDELT (Ward et al., 2013). Each DB record describes a security event with its time, location, event type, main actors, their nationality, victims, and optionally a text describing the event. Other well known socio-political databases are POLDEM (Kriesi et al., 2020), POLECAT (Halterman et al., 2023), UCDP data set (Sundberg et al., 2012). An overview of the publicly available event databases is provided in (Olsen et al., 2024).

3. EMM NEXUS

This section briefly describes the real-time event extraction system NEXUS (News cluster Event eXtraction Using language Structures). It is a rule based system, which uses Boolean combinations of keywords for event detection and grammar rules for event argument extraction.

NEXUS is an integral part of the Europe Media Monitor (EMM) and it has been described in details in (Tanev et al., 2008) and (Tanev et al., 2009). Its event taxonomy, see Figure 1 has been used to create the JRC security event corpus, described in (Atkinson et al., 2017a). Moreover, the corpus was created by manually annotating and curating articles, with events pre-detected by the system.

NEXUS uses the clusters of news articles, created by the EMM software (Tanev et al., 2008). Clusters describing various types of crisis events are selected via application of combinations of keywords, manually crafted and expanded with the help of terminology extraction software. The NEXUS system detects and extracts one main crisis event for each news cluster reporting an event of interest.

For each event the system generates a frame, whose main slots are: date and location, number of killed and injured, kidnapped people, actors, and type of event.

Noteworthy, NEXUS processes only the title and three leading sentences for each news article in the news article cluster and then it fuses the event information, extracted from the different articles. The system uses finite state cascade grammar rules over dictionaries of linear grammar patterns **<person> was found dead** or **<person> was stoned to death**. The semi-automatic learning of these patterns and the accompanying lexicon with references to names, professions, organizations, numbers, and other entities, were described in (Tanev et al., 2009).

Event types are detected via a set of keyword based boolean rules. In Table 1 we show excerpts from such rules for the event types *armed conflict*, *riot*, and *air attack*. It is important to consider the following: When processing clusters of news articles, keywords are searched in the title and in the first three sentences of each article in the clus-

Table 1: Samples from the Boolean keyword combinations for event detection

Type	Rule	
riot	AND	"hundreds of angry" OR "demonstration against" OR "mutiny" ...
		"clashes" OR "clashed" OR "burnt" OR "torched" OR "disperse" ...
armed conflict	AND	"troops" OR "soldiers" OR "rebels" OR "insurgents" ...
		"deployed" OR "clashed" OR "battling" OR "returned fire" ...
	AND	"marines" OR "armed forces" OR "troops" ...
		"militants" OR "insurgents" OR "rebels" ...
air attack	AND	"fighter plane" OR "jets" OR "missile" OR "gunship" OR "interceptor" ...
		"damaged" OR "intercepted" OR "pounded" OR "targeted" ...
	-	"helicopter fired" OR "air raid" OR "missile attack" OR "bombing run" ...

ter. Second, each keyword combination has an assigned maximal word proximity. For example, considering the air attack keyword combination, its proximity is defined to be 17 tokens. Consequently, if both the word "jets" and "intercepted" appear in no more than 17 tokens from each other in the first 3 sentences of a news article, the "air attack" event will be triggered.

For several event types, NEXUS requires not only keyword rules to fire, but also the presence of dead or injured victims, detected by the argument extraction grammars. This serves as an additional filter, which increases the precision.

The event type taxonomy, recognized by the NEXUS system reflects the requirements of the Joint Research Centre's Europe Media Monitor (EMM). The event types, recognized by NEXUS are the most frequently reported in the news event classes, referring to crises.

The recognized crisis event types encompass a subset of the security and socio-political events, reported in the news, including man made incidents and natural disasters. Figure 1 shows the taxonomy of the event types, which are a focus of the system. These events can be grouped into several large classes:

1. Socio-political events: They encompass all unrests, protests, military operations, and humanitarian crises. The "unrest" subtypes include violent unrests like *riots*, but also*protests*, *strikes and boycotts*, as well as *sabotages*. Military events involve *armed conflicts*, i.e. battles and sieges performed by military and organized armed groups, *air and missile attacks*, as well as *exploitation of heavy weapons* such as artillery and heavy firing arms. Military events include also *deployment and movements of troops and military vehicles*. Humanitarian events include reports about displacement of people and lack of resources, such as food, water, shelter, and medicines.

2. Crimes: NEXUS recognizes *kidnapping*, *robbery*, *pirate attacks on ships*, *physical abuse*, and *cyber attacks*. Physical abuse events include also cases of sexual abuse. In reality crimes can be part of a terrorist operation, for example kidnapping of a political leader. Similarly, cyber attacks are used as a unconventional warfare and in some cases can also be classified as terrorist attacks. However, given the multifaceted nature of these event classes, they are put in the crime category both for simplicity in classification, as well as because their nature is related to the violation of the law.

3. Legal events: The system detects two legal event types, which are related to the security, namely *arrests* and *trials*.

4. Targeted violence: These are violent events, who are directed towards pre-defined people. The term "targeted violence" is taken from the PLOVER socio political event ontology (Halterman et al., 2023). According to this ontology two event types are considered as targeted violence, namely *execution* and *assassination*.

5. Terrorist attacks: NEXUS recognizes the most common forms of terrorist attacks: namely bombings, including suicide attacks, as well as all violence, explicitly labeled as terrorism or performed by certain armed groups (e.g. Al Qaeda, IRA, etc.).

6. Violence without detected motivation: The three event types of *shooting*, *stabbing*, and *arson* fall in this category, when the system cannot detect the motivation context, which could be crime, terrorism, unrest, or military.

4. Experiments and Evaluation

4.1. Evaluating NEXUS

4.1.1. Evaluation on the JRC event corpus

The NEXUS system has been used, when creating the JRC security event corpus (Piskorski et al., 2023). First, the events were detected by NEXUS,

and then they were manually moderated and errors were corrected. The taxonomy of NEXUS was used when labeling the events from the JRC security corpus.

The JRC corpus contains around 617K events, extracted by NEXUS, of which 17K are manually curated. The authors of the corpus provided also a detailed evaluation of the event type, geolocation, and argument detection accuracy of the NEXUS event detection system. However, they use a very limited gold standard of 16 news clusters. In contrast, we wanted to evaluate the event classification of NEXUS on a proper subset of the manually moderated JRC corpus. In this paper we report on evaluation of the English language part of the manually moderated part of the corpus, which contains 7,934 detected events, each provided with a manually selected event type code, a title, and a text fragment, containing one or two sentences describing the event.

We have run NEXUS on the title and the event describing fragment in each of the 7,934 English language events from the corpus and compared the extracted event types against the ground truth annotation. Then, we have measured the precision, recall and F1 measure. Results are reported in Table 2.

Clearly, the NEXUS system works best for the "Unrest" event type among all socio-political events. The unrest involves all the protests, riots, and violent anti government actions, which are not terrorism. The legal event types, "Arrest" and "Trial" are also among the best performing classes. It was disappointing the low results for the military event types. Notably, we have got very low recall for all the event types "Military operation", "Air attack", and "Heavy weapons". These low results in the military event types clear suggest how to further improve the system.

The system works quite well also on the disaster group of event classes.

4.1.2. Preliminary Evaluation on the ACLED event database

We have conducted an additional evaluation of the NEXUS system on a more standard and widely used event data set.

For this purpose we have chosen the ACLED event dataset (Raleigh et al., 2010). It is one of the largest manually curated event databases. Evaluating against ACLED however was related to the challenge of of mapping NEXUS event types to the ACLED ones.

There are some little differences of the definitions of the event types of ACLED and NEXUS: first, ACLED does not cover incidents and disasters. Second, it does not classify explicitly events as terrorist attacks, but puts part of them in the larger category of "Explosions/Remote violence". Moreover ACLED events encompass also peaceful events, called "Strategic developments" and in NEXUS only one event type, namely "Arrest" is included in this class. Independently of these differences, we have managed to map some of the NEXUS event types into ACLED categories. Mapping was most of the time many to one: many NEXUS categories were mapped to one ACLED class. In Table 3 we show the mapping between the two event classification systems. In Table 4 we report the results from the ACLED evaluation after the mapping took place. What is important is that first, we cover only a small percent of the strategic developments; second, we did not manage to map properly terrorist events, since they are not part of the ACLED taxonomy and they were considered like no events. Therefore, the ACLED evaluation we performed can be considered approximate.

Still, the relations between the performance on different event types show similar trends in both evaluations: The "Protest" event type, which is a subtype of "Unrest", has a relatively high performance in the ACLED evaluation, as its super type "Unrest" has a good performance in the JRC corpus evaluation. Moreover, the system obtains low recall and low F1 score on the ACLED "Battle event", and similarly its corresponding NEXUS "Military operation" shows the same trend in the JRC corpus evaluation. Also, the ACLED Explosion/Remote violence which corresponds to the NEXUS "Heavy weapons", "Air attack" and "Bombing" obtains low recall, as its corresponding NEXUS types in the JRC corpus evaluation.

The conclusions drawn from both evaluations indicate that mapping between event taxonomies poses challenges, such as: partially overlapping event types, one to many event type relations, taxonomy gaps (for example the lack of terrorist attack in ACLED). The evaluation we have conducted on the ACLED data demonstrate these challenges.

On the other hand, this evaluation was also useful, since it confirmed several trends, observed in the JRC corpus evaluation, namely a notable underperformance of event detection rules in identifying military events and relatively high accuracy in modeling "Unrest" and its subtype "Protest".

4.2. Comparing Nexus to a Transformer-based system

So as to assess the respective merits of rule-based and transformer based systems, we fine-tuned a XLM-Roberta-base system on the JRC corpus. As this kind of system is sensible to dataset balance, we first give some general figures on the JRC corpus. Figure 2 shows the unbalanceness of this dataset of 6,892 annotated sentences.

Table 2: Performance of NEXUS on the JRC corpus

Event Type (code)	**Precision**	**Recall**	**F1**
Socio political			
Military operation (ARM)	0.66667	0.25586	0.36979
Air/missile attack (AA)	0.81395	0.30702	0.44586
Heavy weapons (HW)	0.52000	0.36111	0.42623
Terrorist Attack (TA)	0.63071	0.74146	0.68161
Bombing (BO)	0.67164	0.60811	0.63830
Unrest (SP)	0.83877	0.77140	0.80368
Humanitarian (HUM)	0.51485	0.4000	0.44835
Legal			
Arrest (AR)	0.92012	0.62854	0.74688
Trial (TRIAL)	0.92181	0.38063	0.53879
Crimes			
Kidnapping (KD)	0.73810	0.70992	0.72374
Physical abuse (PA)	0.55556	0.31746	0.40404
Non violent			
Hostage Release (RE)	0.83721	0.39560	0.53731
Violence without defined motivation			
Shooting (SH)	0.84834	0.47733	0.61092
Stabbing (ST)	0.73171	0.58824	0.65217
Targeted killing			
Execution (EX)	0.76190	0.64000	0.69565
Accidents and Disasters			
Earthquake (EQ)	0.90278	0.67708	0.77381
Flood (FL)	0.77477	0.74783	0.76106
Winter storm (IR)	1.00000	0.71875	0.83636
Storm (SR)	0.81481	0.62857	0.70968
Tropical storm (TR)	0.84211	0.68571	0.75591
Wild fire (WF)	0.96154	0.71429	0.81967
Landslides (LS)	0.73333	0.47826	0.57895
Man made disaster (MM)	0.86826	0.68289	0.76450
Maritime accident (MT)	0.94000	0.66197	0.77686
Explosion (XP)	0.68519	0.48684	0.56923
Another event type with dead or injured			
Other (NONE)	0.11632	0.75	0.20141
Accuracy			0.56479
Macro Avg	0.691476	0.61701	0.65035

Table 3: Mapping NEXUS to ACLED event types

ACLED category	**ACLED Explained**	**NEXUS**
Protest	Protests which start as peaceful	Protest
Riot	Riot or Mob Violence	Riot
Battle	Battle between organized forces	Military operation
Explosion/Remote violence	Bombings, shellings, air raids	Air Attack; Heavy Weapons ; Bombing; Suicide Attack
Strategic developments	Arrests, agreements, transfer of territories	Arrest
Violence against civilians	Violence against civilians	Physical Attack; Kidnapping

Table 4: Performance of NEXUS on the ACLED corpus

Class	Precision	Recall	F1-score
Battle	0.5995	0.3334	0.4285
Explosion/Remote violence	0.9356	0.2558	0.4018
Protest	0.8709	0.7278	0.7929
Riot	0.6607	0.1109	0.1899
Strategic developments	0.1794	0.3102	0.2273
Violence against civilians	0.8253	0.0695	0.1282
Accuracy			0.4153
Macro Avg	0.5816	0.2582	0.3098
Weighted Avg	0.7856	0.4153	0.4978

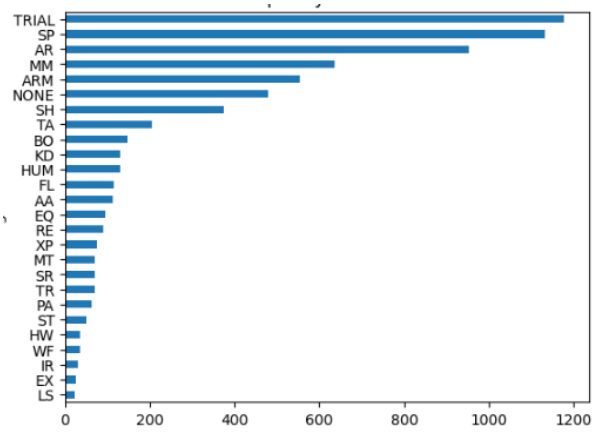


Figure 2: Distribution of classes in JRC news Dataset

We divided this dataset into the traditional training-development-test split: 80-10-10. it was done with the dataset Huggingface module thus respecting the distribution of the overall dataset. We train the model for 15 epochs, learning rate 2e-5, batch size 16. Table 5 shows the performance XLM Roberta achieves on the test set of EMM News. The overall accuracy is 0.67 after 15 epochs (after 8 epochs, we reached 0.70) surpassing the rule-based model by a large margin, especially when enough learning data are available. On this respect, the motto "more data, better results" is easily confirmed, pushing us to augment the EMM data.

4.3. Evaluating a Transformer-based system with augmented data

The last experiment we undertook consisted in balancing the fine-tuning dataset. Among several techniques to do so (REF), we decided to use LLM data augmentation techniques, as generative Language models reveal to be quite efficient in reformulating sentences (see e.g. (**?**)). As seen in the previous experiment, fine-tuning a model requires a lower-bound number of examples. To balance the dataset for under-represented classes, we used the following prompt:

> Your task is to generate {number} sentences, denoting the following type of event: {label}. As a help, the following sentence denote this type of event. To generate these sentences, please try to mimic a headline style, describing the facts and circumstances of the event. Generate these sentences in English, and rephrase the original sentence with several techniques, like synonym substitution, adverb insertion, paraphrasing and other distributional operations enabling to preserve the overall meaning while changing wording and phrasing. As output, please generate the sentences one per line. Be the most concise you can. Example sentence: {sentence}

where {number} represents the number of sentences to generate for every given class source example, calculated by the number of examples of the most represented class (Trial, 1,177) divided by the number of examples of the given class, rounded to the ceil. For example, for the class Arrest, 2 sentences will be generated for each source example (1,177 / 953 = 1.23 == 2); {label} represents the class, e.g. Arrest, and {sentence} represents the given example to rephrase, eg. *Nine held in Eta anti-terror raids*. Table 6 shows a few examples of paraphrases generated by GPT4 (OpenAI's June version with a context length of 8,192 tokens).

Figure 3 gives the distribution of samples per classes after data augmentation with a total of 35,583 example titles and on average more than 1,250 examples per class.

We then fine-tuned, with the same parameters as in the previous experiment, a language model. Table 7 shows the results on the EMM source titles for the sake of comparison with the Nexus system. As can be seen, the results are very promising, even if they need to be further confirmed on a totally new dataset. We also performed an error analysis, from the dispersion plot fig:displot.augmented.

Table 5: Evaluation results on JRC news dataset, fine-tuned XLM-Roberta-base model

event category	precision	recall	f1-score	support
Military operation (ARM)	0.52830	0.5	0.51376	56
Air/missile attack (AA)	0.64285	0.75	0.69230	12
Heavy weapons (HW)	0.0	0.0	0.0	3
Terrorist Attack (TA)	0.48	0.57142	0.52173	21
Bombing (BO)	0.33333	0.133333	0.190477	15
Unrest (SP)	0.75862	0.77192	0.76521	114
Humanitarian(HUM)	0.54545	0.461536	0.5	13
Arrest(AR)	0.71153	0.77083	0.74	96
TRIAL	0.78703	0.72033	0.75221	118
Kidnapping(KD)	0.6	0.461536	0.52173	13
Physical abuse(PA)	0.33333	0.33333	0.33333	6
Hostage release (RE)	0.54545	0.66666	0.6	9
Shooting (SH)	0.61702	0.76315	0.68235	38
Stabbing (ST)	0.25	0.2	0.22222	5
Execution(EX)	1.0	0.5	0.66666	2
Earthquake(EQ)	0.61538	0.88888	0.72727	9
Flood(FL)	0.69230	0.75	0.72	12
Winter Storm(IR)	1.0	0.66666	0.8	3
Storm (SR)	0.77777	1.0	0.875	7
Tropical storm(TR)	1.0	0.57142	0.72727	7
Wild fire (WF)	0.5	0.33333	0.4	3
Landslides(LS)	1.0	0.5	0.66666	2
Man made disaster (MM)	0.78461	0.796875	0.79069	64
Maritime accident (MT)	0.875	1.0	0.93333	7
Explosion (XP)	0.363635	0.57142	0.44444	7
Other (NONE)	0.413047	0.39583	0.404254	48
accuracy			0.66956	690
macro avg	0.62133	0.57994	0.58426	690
weighted avg	0.66672	0.66956	0.66372	690

Table 6: Example of paraphrase generation from GPT4 (arrest category)

Source sentence	GPT4 paraphrase
Man arrested after planting fake bomb in Chicago (AP)	Individual detained for setting up counterfeit explosive in Chicago (Reuters)
	Chicago law enforcement apprehends man for hoax bomb plant (BBC)
Three arrested over injured rugby player	Trio apprehended linked to wounded rugby athlete
	Three detained in connection with harm inflicted on rugby sportsman
Suspect arrested after television appeal	Individual apprehended following TV plea
	TV appeal leads to suspect's detention

Apart from the already observed size effect (more data, better prediction), a few categories are predicted with a F1 score less than 0.90: Heavy Weapons Fire, Execution and Stabbing have the worst outcome with 0.80. Undefined is at 0.88 and all the other categories are above 0.90 which represents a new state-of-the-art by a large margin. First, if we compare the overall results to the same with fine-tuned model with just the source data, we can clearly see the benefit of data augmentation,

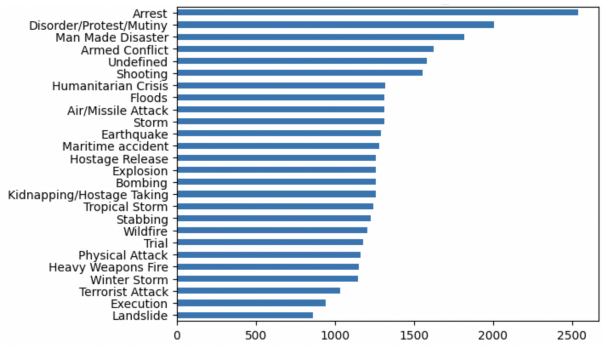


Figure 3: Distribution of classes in JRC news Dataset after GPT4 augmentation

even for the under-represented categories. (see categories with less than 100 support).

5. Conclusions and Perspectives

We have evaluated two event detection system, the first based on rules and the second, based on transformer-based classifiers. We have also experimented with data augmentation, using a state-of-the-art LLM.

Transformer-based classifiers gave overall comparable performance to the rule based system. Both systems show their own advantages: statistical classifiers achieve better classification accuracy (0.67 vs. 0.56). On the other hand, these classifiers show lower average F1 performance, mainly due to the imbalanced training set. This disadvantage was removed with data augmentation and dataset balancing, achieving much higher accuracy (0.93).

Going deeper into the details, statistical classifiers provided a much better F1 score for the classes which are frequent in the corpus, the TRIAL event class: 0.75 vs. 0.54 for NEXUS; Military operation (ARM): 0.51 vs. 0.37, and the NONE class, which is event reporting dead or injury, not belonging to any of the classes in the corpus, 0.40 vs. 0.20. The other case, where statistical classifier notably outperforms NEXUS is for the event type Storm (SR), 0.87 vs. 0.7, and Maritime accident (MT), 0.93 vs. 0.78. On the other hand, the NEXUS system has detected far better the following important event types: Terrorist attack (TA), 0.68 vs. 0.52, Kidnapping (KD), 0.72 vs. 0.52, Bombing (BO), 0.62 vs. 0.19, and Explosion (XP)

For most of the other classes we have similar performance between the two systems with the statistical biased towards more frequent classes and demonstrating much better overall accuracy and the rule based NEXUS with more balanced behaviour, showing a significantly higher macro average F1. Considering that both system approaches have different strong points, delivering a combined model will most likely deliver the most optimal results.

Another conclusion, based on the last experiments is that large language models can help build relevant datasets for fine-tuning transformer models on Event Extraction. Even if it is not possible so far to use LLMs directly for live detection, mainly due to the hardware requirements of such models and secondly due to the currently lower quality of open-sourced models, progress in these two areas should lead us in the future to directly use these models, as they show a amazing ability to learn from few examples. The next step would also be to complement sentence or passage classification with extracting the arguments of the events. For example, instead of just classifying *Two passenger trains collide in Egypt, killing 25* as a "Man made disaster", generate a structured extraction stating the specific disaster (collision), the participants (two passenger trains), the time (unspecified here but can be inferred from the source of the headline), location (Egypt) and the resulting damage (25 human deaths).

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Sixth applied natural language processing conference*, pages 76–83.

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017a. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65, Vancouver, Canada. Association for Computational Linguistics.

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017b. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65.

Nancy Chinchor and Elaine Marsh. 1998. Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pages 359–367.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.

Linguistic Data Consortium et al. 2005. Ace (automatic content extraction) english annotation guidelines for events version 5.4. 3. *ACE*.

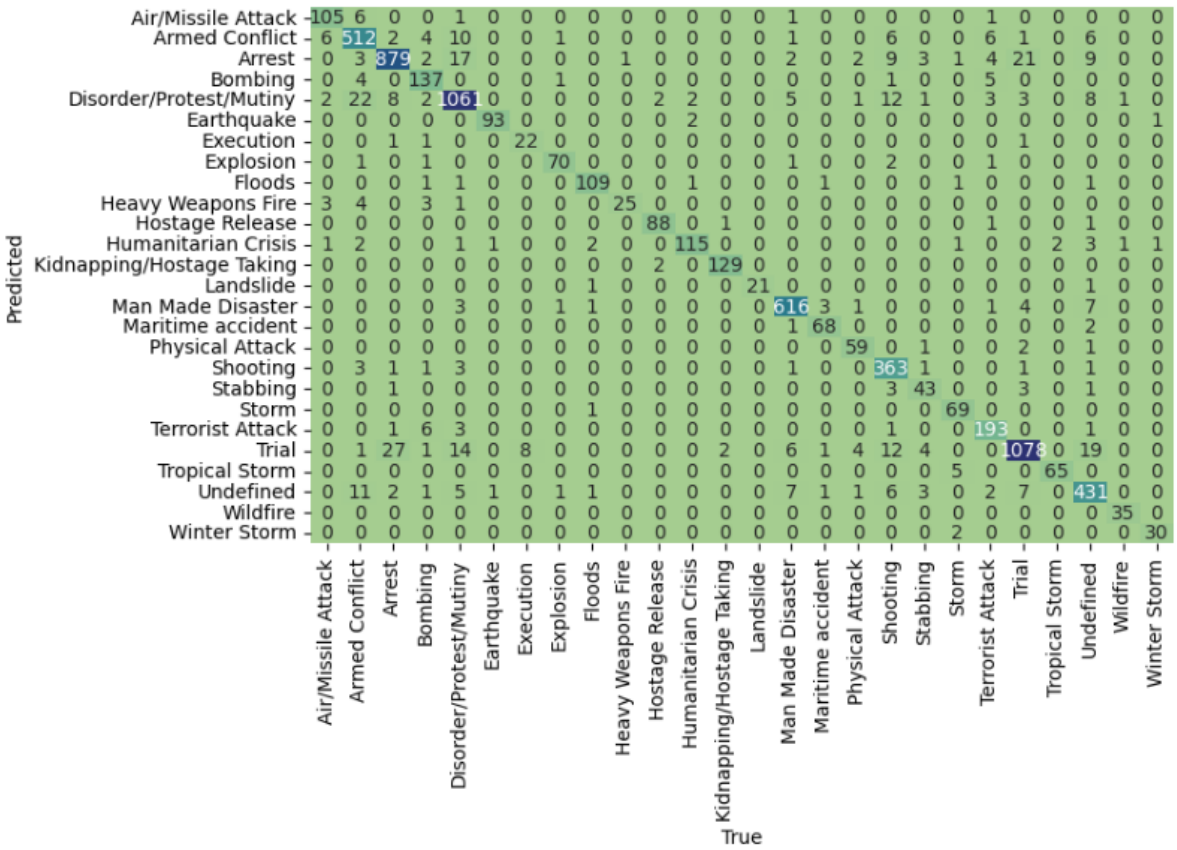


Figure 4: Dispersion plot of results of fined-tuned model on EMM dataset

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002a. Information extraction for enhanced access to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002b. Real-time event extraction for infectious disease outbreaks. In *Proceedings of Human Language Technology Conference (HLT)*, pages 366–369.

Andrew Halterman, Benjamin Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. Plover and polecat: A new political event ontology and dataset.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. In *7th International Workshop on News Recommendation and Analytics*, pages 35–43.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021. Challenges and applications of automated extraction of socio-political events from text (case 2021): Workshop and shared task report. *arXiv preprint arXiv:2108.07865*.

Ourania Kounadi, Thomas J Lampoltshammer, Elizabeth Groff, Izabela Sitko, and Michael Leitner. 2015. Exploring twitter to analyze the public's reaction patterns to recently reported homicides in london. *PloS one*, 10(3):e0121848.

Hanspeter Kriesi, Bruno Wüest, Jasmine Lorenzini, Peter Makarov, Matthias Enggist, Klaus Rothenhäusler, Thomas Kurer, Silja Häusermann, Patrice Wangen, Argyrios Altiparmakis, et al. 2020. Poldem-protest dataset 30 european countries.

Remo Nitschke, Yuwei Wang, Chen Chen, Adarsh Pyarelal, and Rebecca Sharp. 2022. Rule based event extraction for artificial social intelligence. In *Proceedings of the First Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 71–84.

Helene Bøsei Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio Political Events (CASE)*.

Jakub Piskorski, Nicolas Stefanovitch, Brian Doherty, Jens P Linge, Sopho Kharazi, Jas Mantero,

Table 7: Classification report on EMM data, fine-tuned on augmented data

	precision	**recall**	**f1-score**	**support**
Air/Missile Attack	0.89743	0.92105	0.90909	114
Armed Conflict	0.89982	0.92252	0.91103	555
Arrest	0.95336	0.92235	0.9376	953
Bombing	0.85625	0.92567	0.88961	148
Disorder/Protest/Mutiny	0.94732	0.93645	0.94185	1133
Earthquake	0.97894	0.96875	0.97382	96
Execution	0.73333	0.88	0.8	25
Explosion	0.94594	0.92105	0.93333	76
Floods	0.94782	0.94782	0.94782	115
Heavy Weapons Fire	0.96153	0.69444	0.80645	36
Hostage Release	0.95652	0.96703	0.96174	91
Humanitarian Crisis	0.95833	0.88461	0.92	130
Kidnapping/Hostage Taking	0.97727	0.98473	0.98098	131
Landslide	1.0	0.91304	0.95454	23
Man Made Disaster	0.96099	0.96703	0.96400	637
Maritime accident	0.91891	0.95774	0.93793	71
Physical Attack	0.86764	0.93650	0.90076	63
Shooting	0.87469	0.968	0.91898	375
Stabbing	0.76785	0.84313	0.80373	51
Storm	0.87341	0.98571	0.92617	70
Terrorist Attack	0.88940	0.94146	0.91469	205
Trial	0.96164	0.91588	0.93820	1177
Tropical Storm	0.97014	0.92857	0.94890	70
Undefined	0.87601	0.89791	0.88683	480
Wildfire	0.94594	1.0	0.97222	35
Winter Storm	0.9375	0.9375	0.9375	32
accuracy			0.93093	6892
macro avg	0.91761	0.92573	0.91991	6892
weighted avg	0.93250	0.93093	0.93113	6892

Guillaume Jacquet, Alessio Spadaro, and Giulia Teodori. 2023. Multi-label infectious disease news event corpus.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Ralph Sundberg, Kristine Eck, and Joakim Kreutz. 2012. Introducing the ucdp non-state conflict dataset. *Journal of peace research*, 49(2):351–362.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, June 24-27, 2008 Proceedings 13*, pages 207–218. Springer.

Hristo Tanev, Vanni Zavarella, Jens Linge, Mijail Kabadjov, Jakub Piskorski, Martin Atkinson, and Ralf Steinberger. 2009. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamática*, 1(2):55–66.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Analysis*, 21(1):267–297.