Explainable Event Detection with Event Trigger Identification as Rationale Extraction

Hansi Hettiarachchi
Birmingham City University
Birmingham, UK
hansi.hettiarachchi@bcu.ac.uk

Tharindu Ranasinghe
Aston University
Birmingham, UK
t.ranasinghe@aston.ac.uk

Abstract

Most event detection methods act at the sentence-level and focus on identifying sentences related to a particular event. However, identifying certain parts of a sentence that act as event triggers is also important and more challenging, especially when dealing with limited training data. Previous event detection attempts have considered these two tasks separately and have developed different methods. We hypothesise that similar to humans, successful sentence-level event detection models rely on event triggers to predict sentence-level labels. By exploring feature attribution methods that assign relevance scores to the inputs to explain model predictions, we study the behaviour of state-of-the-art sentence-level event detection models and show that explanations (i.e. rationales) extracted from these models can indeed be used to detect event triggers. We, therefore, (*i*) introduce a novel weakly-supervised method for event trigger detection; and (*ii*) propose to use event triggers as an explainable measure in sentence-level event detection. To the best of our knowledge, this is the first explainable machine learning approach to event trigger identification.

1 Introduction

Every day, numerous socio-political protest events occur worldwide, targeting various decisions made by governments or authorities (Hutter, 2014; Weng and Lee, 2021). These events hold significant importance for political scientists, policymakers, democracy watchdogs, and other stakeholders (Raleigh et al., 2010) due to their potential to provide insights into multiple aspects (Tarrow, 2022). These include analysing the nature, scope, and magnitude of such events, shaping public opinion regarding different causes, assessing the status of freedom and democracy in different nations, and more (Hürriyetoğlu et al., 2021b).

Due to the continuous and abundant data flow over time, news media outlets serve as invaluable sources for social and political scientists who seek to establish comprehensive knowledge bases of protest events (Chenoweth and Lewis, 2013). Early approaches to creating these knowledge bases relied on manual event detection methods (Wang et al., 2016), which can be expensive and slow. Therefore, to cope with the volume of news media, researchers have experimented with automatic event detection methods (Leetaru and Schrodt, 2013). The organisation of the recent shared tasks such as CASE: Challenges and Applications of Automated Extraction of Socio-political Events from Text (Hürriyetoğlu et al., 2021a, 2022) has promoted automatic event detection research within the natural language processing (NLP) community.

Automated event detection tools are designed as pipelines that receive news articles and yield records of events. The first step of these pipelines is discriminating between relevant and irrelevant sentences (Croicu and Weidmann, 2015). In this research, we refer to this as sentence-level event detection. Once event-related sentences are determined, the next task is to extract event information on the token level (Doddington et al., 2004). One such key information is **Event trigger**, defined as the main word that most clearly expresses an event occurrence (Hettiarachchi et al., 2023a). While the sentence-level event detection methods have achieved excellent results recently, the accuracy of word-level predictions still leaves room for improvement. This is partly due to the limited amount of training data, as word-level annotation is time-consuming and expensive. In this research, we introduce a new weakly-supervised approach to event trigger detection that removes the need for training data at the word-level. To achieve this, we propose addressing event trigger detection as a rationale extraction task (Lei et al., 2016).

Proceedings of Recent Advances in Natural Language Processing, pages 507–518
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_056

The domain of explainability encompasses a wide range of techniques focused on explaining the predictions made by machine learning models (Lipton, 2018). Among these techniques, rationale extraction methods aim to identify and select specific portions of the input data that justify the model's output for a given data point. In manual event detection, human perception of sentence-level annotations is guided by the presence of event triggers (Doddington et al., 2004). We hypothesise that sentence-level event detection models also rely on event triggers to make predictions. If that is the case, explanations for sentence-level predictions can be used to detect event triggers, thus removing the need for word-level labelled training data. To extract model explanations, we use post hoc rationale extraction methods (Sundararajan et al., 2017), which try to explain the predictions of a given model.

At the same time, by using event triggers as explanations for sentence-level event detection methods, we introduce a new benchmark for evaluating explainability. In opposition to developing different models for sentence-level and token-level, we propose to train a single model for both tasks.

Our **main contributions** are:

1. We introduce a novel weakly-supervised approach for event trigger detection. We present practical methodologies for leveraging feature attribution methods to extract event triggers from sentence-level event detection models.

2. We provide insights into the behaviour of state-of-the-art sentence-level event detection models by analysing attributions in different learning strategies at sentence-level, monolingual, multilingual and zero-shot.

3. We propose to use event triggers as a new benchmark for evaluating the explainability of sentence-level event detection models. We release the code and the models as the initial baseline for this new benchmark [1].

2 Related Work

2.1 Event Detection

Previous research has proposed different approaches to sentence-level and word-level event detection, which we explain below.

[1] https://github.com/HHansi/XEventMiner

Sentence-level: Sentence-level event detection targets the identification of event-described sentences. Early research widely used linguistic features (e.g. part of speech (POS) tags, Bag of Word (BoW) vectors, token/character n-grams and lemmas) with traditional classification algorithms (e.g. Support Vector Machine (SVM)) for sentence-level detection (Naughton et al., 2010; Lefever and Hoste, 2016). However, following the advances in text embedding models and neural networks, later research focused more on deep learning approaches. Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Convolutional Neural Network (CNN) (Lawrence et al., 1997) were popularly used neural networks for text classification (Luan and Lin, 2019). Following them, various improved architectures such as LSTM-Attention, Convolutional Recurrent Neural Network (CRNN) and CNN-Attention were proposed for sentence-level event detection (Liu et al., 2019a; Huynh et al., 2016). Recently with the success of transformers such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), state-of-the-art sentence-level event detection models are based on transformers (Hu and Stoehr, 2021; Awasthy et al., 2021; Hettiarachchi et al., 2023a) which we also use in this research.

Word-level: Word-level event detection targets the extraction of text spans which describe event details. Word-level methods also show a similar evolution to sentence-level methods. Most of the early work extensively relied on linguistic features due to the complexities of this task (Chen and Ng, 2012; Hong et al., 2011). Later, neural network architectures such as Bidirectional LSTM (Bi-LSTM), Dynamic Multi-pooling CNNs (DMCNNs), Bi-LSTM-DMCNN and multi-attention were proposed for word-level event detection (Nguyen et al., 2016; Feng et al., 2016; Chen et al., 2015; Balali et al., 2020; Ding and Li, 2018). Very recently, similar to the sentence-level, different pre-trained transformers such as BERT and XLM-R were used at word-level (Yang et al., 2019; Huang and Ji, 2020; Awasthy et al., 2021; Hettiarachchi et al., 2023a), setting the state-of-the-art performance (Hürriyetoğlu et al., 2021a, 2022).

In summary, previous research built separate models for sentence and word-level event detection. In both areas, transformer-based approaches hold state-of-the-art performance. Deviating from

Language	Sentence-level			Word-level		
	Sentences	Label Distribution		Sentences	Trigger Distribution	
		1	0		Spans	Tokens
English (En)	21107	2819	18288	3239	4585	6030
Portuguese (Pt)	1095	194	901	87	122	150
Spanish (Es)	2666	375	2291	106	157	216

Table 1: Data statistics in sentence and token-levels. **Label 1** indicates event sentences, and **label 0** indicates non-event sentences. **Spans** are the text spans/ordered sequences of tokens. A trigger can be composed of a span of one or more tokens.

Sentence	Label
Table grape harvesters started protesting about their working conditions in De Doorns last month.	1
There were reports of skirmishes and clashes, including stone pelting, in the area in which two policemen were injured.	1
It is the power to run local affairs as authorised by the central leadership.	0
Fears were that thousands of students, who are writing their National Senior Certificate (matric) exams, could fail to arrive on time.	0

Table 2: Sample event (label=1) and non-event (label=0) sentences. In event sentences, trigger spans are highlighted in yellow.

the common viewpoint, Hettiarachchi et al. (2023a) proposed a transformer-based two-phase learning strategy which captures the interconnections between sentence and word-level tasks for mutual learning. However, as far as we know, no previous work has explored the ability of sentence-level models to predict event words following their learning process.

2.2 Rational Extraction

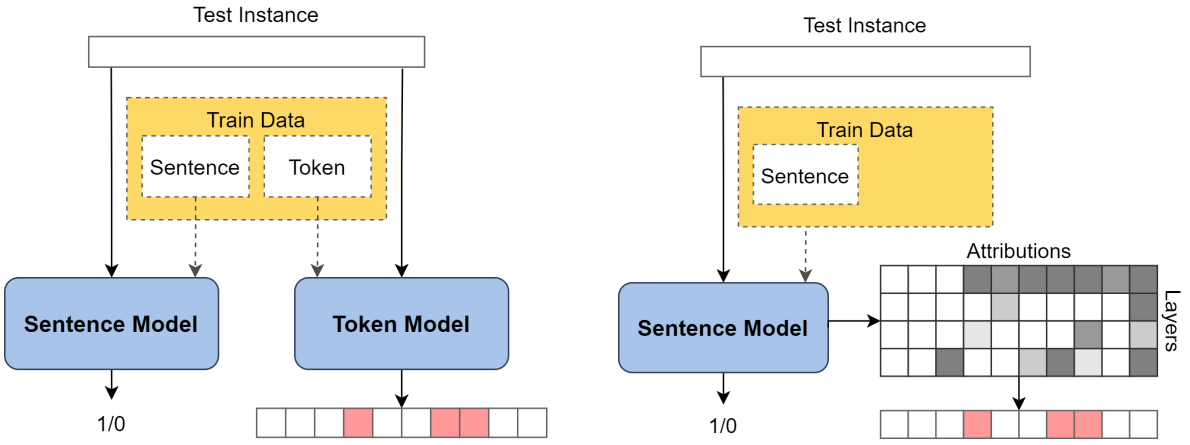
According to Lipton (2018), deep neural network-based NLP models demonstrate impressive performance across diverse tasks, albeit with a trade-off in terms of interpretability. Recent work aims to address this issue by focusing on the explainability of the models (Saeed and Omlin, 2023). Explainability methods typically function by identifying a specific subset of the input that provides a rationale for the model's prediction on an individual data point. This can be achieved through adjustments made to the model architecture (Chalkidis et al., 2021; Yu et al., 2019) or by attempting to explain the predictions generated by a particular model (Schulz et al., 2020) also known as *post hoc*.

Post hoc usually rely on feature attribution methods, which assign an importance value to each input feature of a network (Sundararajan et al., 2017). Feature attribution has a long tradition in image recognition tasks (Vermeire et al., 2022) and has only recently been applied to some NLP tasks (DeYoung et al., 2020). For example, Pavlopoulos et al. (2022) used feature attribution methods such as LIME (Ribeiro et al., 2016) to predict toxic spans in toxic comments. LIME has also been used on offensive token detection in non-English languages such as Sinhala (Ranasinghe et al., 2022) and Korean (Jeong et al., 2022) and has shown that it provides competitive results compared to supervised methods (Ranasinghe and Zampieri, 2021). In translation quality estimation, Fomicheva et al. (2022) used feature attribution to predict word-level errors in the translation.

3 Data

To conduct the experiments, we used the multilingual version of the GLOCON gold standard dataset (Hürriyetoğlu et al., 2021b), which was released by CASE 2021 workshop (Hürriyetoğlu et al., 2021a), considering its recency, open-availability and coverage. This dataset targeted socio-political events covering demonstrations, industrial actions, group clashes, political violence, armed militancy and electoral mobilisations. It contains data at different levels of granularity, document, sentence and word from multiple news sources covering the languages English, Portuguese and Spanish. Considering the scope of this research, we only utilised sentence and word-level data for our experiments from all available languages.

The sentence-level data contained an identifier,



(a) Fully supervised approach

(b) Weakly-supervised approach

Figure 1: Fully supervised word-level event trigger detection (left) and our weakly-supervised word-level event trigger detection as rationale extraction (right). Dashed and solid lines represent training and test time, respectively.

sentence text and binary label, which indicates whether that particular sentence describes/contains an event or not, per instance. For simplicity, we will refer to the event-described sentences as event sentences and others as non-event sentences in the below content. The word-level data were in BIO (Beginning, Inside, Outside) format (Ramshaw and Marcus, 1995), based on event triggers and arguments (i.e. participant, place, target, organiser, event time and facility name).

Data Cleaning: We applied a few techniques to clean the data. Since we aim to evaluate sentence classifiers' ability to recognise event triggers, we removed any sentences shared between sentence and token levels as they could affect the evaluations. Considering the fewer samples available at the word-level, we removed any shared sentence from the sentence-level. Also, following our aim, we only kept the trigger labels at the word-level, excluding event arguments.

The data statistics of cleaned datasets at sentence and token levels covering all three languages are summarised in Table 1. Overall, the sentence-level has more instances/labelled samples than the word-level. Also, there are more non-event sentences than event sentences. Since this imbalance depicts the real scenario and provides more training samples from the targeted domain to the models, we directly experimented with these data without further pruning. Considering the languages, comparatively, English has more instances than others at both granularities explaining its wide usage and data availability. Thus, we consider English as a high-resource language and others as low-resource languages in this research. Additionally, Table 2 provides a few sample sentences in English, covering sentence-level labels and word-level triggers.

4 Methodology

We propose framing weakly-supervised event trigger detection as rationale extraction from sentence-level event detection models. Instead of training a dedicated supervised model for event trigger prediction, we propose deriving word-level scores from a strong sentence-level event detection model by extracting explanations for model predictions (Figure 1). Given a trained sentence-level event detection model and the test data, rationale extraction methods detect the parts of the input that are relevant for model predictions on a sample-by-sample basis. We hypothesise that words with the highest relevance scores should correspond to actual event triggers.

Our methodology has two main steps; (1) Event Sentence Classification (2) Event Trigger Identification, which we describe in the below sections.

4.1 Event Sentence Classification

For the sentence-level models, we used transformer models as they have achieved state-of-the-art results on event sentence classification (Hürriyetoğlu et al., 2022, 2021a; Hettiarachchi et al., 2021). We trained the models on the sentence-level data in the GLOCON gold standard dataset (Hürriyetoğlu et al., 2021b) described in Section 3, where the labels indicate whether that particular sentence de-

scribes/contains an event or not.

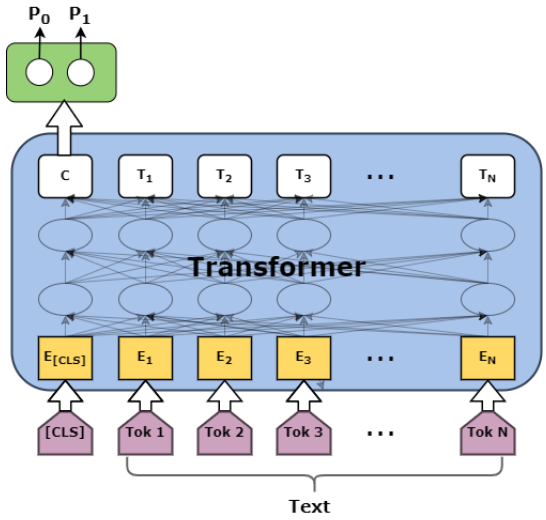


Figure 2: A schematic representation of the transformer models in sentence-level event detection.

From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h})$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels which in our case is two. This architecture is depicted in Figure 2. We employed a batch size of 32, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were evaluated while training using an evaluation set that had one-fifth of the rows in data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs. All the pre-trained transformer models we used for the experiments are available in HuggingFace (Wolf et al., 2020).

We used the following strategies to train sentence-level transformer models.

Monolingual: We trained a separate machine learning model on each of the three languages. We then evaluated the trained model on the test set of the particular language mimicking the `supervised monolingual` setting. For English, we used three popular transformer models; BERT-LARGE-CASED (Devlin et al., 2019), ELECTRA-LARGE-DISCRIMINATOR (Clark et al., 2020) and ROBERTA-LARGE (Liu et al., 2019b). For Spanish, we used BETO-BASE-CASED (José et al., 2020) and BERT-BASE-MULTILINGUAL-CASED (Devlin et al., 2019), while for Portuguese we used BERT-BASE-PORTUGUESE-CASED (Souza et al., 2020) and BERT-BASE-MULTILINGUAL-CASED (Devlin et al., 2019).

All: We concatenated the training sets of all the languages and trained a single machine learning model. We then evaluated the model on each testing set of all three languages mimicking the `supervised multilingual` setting. For this setting we used BERT-BASE-MULTILINGUAL-CASED (Devlin et al., 2019) and XLM-ROBERTA-BASE (Conneau et al., 2020) models. Previous studies have shown that `supervised multilingual` models provide better results than `monolingual` models in event detection (Hettiarachchi et al., 2021).

All-1: We concatenated all training sets except one language and trained a single machine learning model. We then evaluated the model on the test set of that particular dataset that was left out, mimicking the `zero-shot` setting for the left-out language. For this setting also we used BERT-BASE-MULTILINGUAL-CASED (Devlin et al., 2019) and XLM-ROBERTA-BASE (Conneau et al., 2020) models. Previous studies have shown that `zero-shot` setting has provided compatible results that can be useful in low-resource languages where the training data is scarce (Hettiarachchi et al., 2021). We only conducted these experiments for Spanish and Portuguese.

4.2 Event Trigger Identification

For event trigger identification, we propose a weakly-supervised approach by incorporating techniques which explain the predictions of the event sentence classification models. Our focus is mainly influenced by the limitations of annotated data at the word-level due to the annotation complexities and recent advances in the area of model explainability. We use Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) as the classifier explainers in our work, considering their comprehensiveness and dominance in explaining black-box models (Linardatos et al., 2021). More details about these two frameworks are described below.

LIME (Ribeiro et al., 2016): LIME explains the predictions of any classifier by fitting a local in-

Language	Strategy	Model	Event			Not			Weighted Average			F1 Macro
			P	R	F1	P	R	F1	P	R	F1	
English	Monolingual	BERT-LARGE	0.78	0.87	0.82	0.96	0.94	0.95	0.93	0.92	0.93	0.88
		ROBERTA-LARGE	0.82	0.84	0.83	0.96	0.95	0.96	0.93	0.93	0.93	**0.89**
		ELECTRA-LARGE	0.78	0.84	0.81	0.96	0.94	0.95	0.92	0.92	0.92	0.88
	All	XLM-ROBERTA-BASE	0.73	0.88	0.80	0.96	0.91	0.93	0.91	0.90	0.91	0.86
		BERT-MULTILINGUAL	0.73	0.76	0.74	0.93	0.92	0.93	0.89	0.89	0.89	0.83
Spanish	Monolingual	BETO-BASE	0.61	0.69	0.65	0.94	0.91	0.93	0.89	0.88	0.88	0.79
		BERT-MULTILINGUAL	0.59	0.51	0.55	0.91	0.92	0.92	0.86	0.86	0.86	0.73
	All	XLM-ROBERTA-BASE	0.68	0.74	0.71	0.95	0.93	0.94	0.90	0.90	0.90	**0.82**
		BERT-MULTILINGUAL	0.52	0.44	0.48	0.89	0.92	0.91	0.84	0.85	0.84	0.69
	All-1	XLM-ROBERTA-BASE	0.57	0.72	0.63	0.94	0.90	0.92	0.88	0.87	0.87	0.78
		BERT-MULTILINGUAL	0.51	0.48	0.50	0.90	0.91	0.90	0.84	0.84	0.84	0.70
Portuguese	Monolingual	BERT-BASE-PORTUGUESE	0.86	0.76	0.80	0.93	0.96	0.90	0.92	0.92	0.92	**0.88**
		BERT-MULTILINGUAL	0.92	0.52	0.66	0.88	0.98	0.93	0.89	0.8	0.87	0.80
	All	XLM-ROBERTA-BASE	0.73	0.88	0.80	0.96	0.91	0.93	0.91	0.90	0.91	0.86
		BERT-MULTILINGUAL	0.73	0.76	0.74	0.93	0.92	0.93	0.89	0.89	0.89	0.83
	All-1	XLM-ROBERTA-BASE	0.83	0.80	0.81	0.94	0.95	0.95	0.92	0.92	0.92	0.88
		BERT-MULTILINGUAL	0.81	0.52	0.63	0.88	0.96	0.92	0.86	0.87	0.86	0.77

Table 3: Results for sentence-level event detection with different strategies. For each model, Precision (P), Recall (R), and F1 are reported on all classes and weighted averages. Macro-F1 is also listed.

terpretable model. It aims to test the impacts on predictions by varying the input data to the classifier. Per sample, LIME generates a new dataset of perturbed samples and the corresponding predictions of the classifier. Then, it fits a linear model on new data, which results in coefficients per feature as their attribution scores. In this research, each token is considered as a feature and perturbation is achieved by random sampling of tokens in the input text sequence or randomly removing tokens from the input text sequence.

SHAP (Lundberg and Lee, 2017): SHAP explains the predictions of any classifier by following a game theoretic approach. It assigns an importance value to each feature of the input for a particular prediction made by the classifier. The feature importance is calculated using shapely values, a game theory concept that quantifies each feature's contribution to the final prediction. In this research, each token in the input text sequence is considered as a feature while applying SHAP.

As described above, LIME and SHAP return an attribution/importance score per feature (i.e. token) in an input text sequence which explains the sentence classifier's prediction. Theoretically, for a sentence which is classified as an event sentence, high scores depict the tokens which had a high impact on the classifier's prediction or which let the sentence be predicted as an event sentence. Following this assumption, we assign a binary decision of event and non-event to each token based on its corresponding importance score, and we consider event tokens as event triggers. For this assignment, we used a threshold tuned on the ground truth labels (i.e. event triggers) of one-fifth of the word-level data using the Stochastic Gradient Descent algorithm.

5 Results

5.1 Event Sentence Classification

The results of the sentence-level models are shown in Table 3. As the label distribution is highly imbalanced, we evaluate and compare the performance of the different models using the Macro F1-score. We further report per-class Precision (P), Recall (R), F1-score (F1), and weighted average. As can be seen, all the transformer models provided strong results for sentence-level event detection.

For English, ROBERTA-LARGE (Liu et al., 2019b) with the monolingual strategy provided the best Macro F1 score. It should be noted that ***All*** strategy also yields comparable results; however they do not outperform the models with ***Monolingual*** strategy. For Spanish, XLM-ROBERTA-BASE with ***All*** strategy provided the best Macro F1 with a 0.82 score, outperforming ***Monolingual*** strategy. In Portuguese, ***Monolingual*** strategy with BERT-BASE-PORTUGUESE provided the best results. Interestingly, zero-shot ***All-1*** strategy with

Language	Strategy	Model	LIME			SHAP		
			P	R	F1	P	R	F1
English	Monolingual	BERT-LARGE	0.43	0.60	0.50	0.47	0.70	0.56
		ROBERTA-LARGE	0.41	0.66	0.51	0.50	0.69	**0.58**
		ELECTRA-LARGE	0.44	0.66	0.53	0.43	0.76	0.55
	All	XLM-ROBERTA-BASE	0.37	0.64	0.47	0.37	0.66	0.48
		BERT-MULTILINGUAL	0.43	0.57	0.49	0.40	0.61	0.48
Spanish	Monolingual	BETO-BASE	0.13	0.68	0.21	0.55	0.62	**0.58**
		BERT-MULTILINGUAL	0.14	0.72	0.24	0.17	0.68	0.27
	All	XLM-ROBERTA-BASE	0.15	0.64	0.24	0.05	0.99	0.11
		BERT-MULTILINGUAL	0.10	0.64	0.17	0.19	0.49	0.28
	All-1	XLM-ROBERTA-BASE	0.15	0.67	0.24	0.21	0.66	0.32
		BERT-MULTILINGUAL	0.15	0.70	0.24	0.05	0.96	0.11
Portuguese	Monolingual	BERT-BASE-PORTUGUESE	0.22	0.59	0.32	0.47	0.70	**0.56**
		BERT-MULTILINGUAL	0.29	0.61	0.39	0.14	0.64	0.24
	All	XLM-ROBERTA-BASE	0.33	0.69	0.44	0.32	0.71	0.44
		BERT-MULTILINGUAL	0.40	0.43	0.42	0.17	0.63	0.21
	All-1	XLM-ROBERTA-BASE	0.05	0.57	0.10	0.31	0.73	0.44
		BERT-MULTILINGUAL	0.24	0.34	0.28	0.21	0.54	0.30

Table 4: Results for event trigger detection with LIME and SHAP. For each model, Precision (P), Recall (R), and F1 are reported on event trigger words.

XLM-ROBERTA-BASE also provided very close results to the best result.

Overall the results show that transformers provide excellent results for sentence-level event detection. Furthermore, the models and the strategies we used are highly compatible with each other.

5.2 Event Trigger Identification

The results of LIME (Ribeiro et al., 2016), and SHAP (Lundberg and Lee, 2017) with different sentence-level models are shown in Table 4. For the evaluation, we used the precision (P), Recall (R), and F1 score of the event trigger tokens

For English, ROBERTA-LARGE scored 0.58 F1 score with SHAP, for Spanish BETO-BASE scored 0.58 F1 score with SHAP, and for Portuguese, BERT-BASE-PORTUGUESE scored 0.56 F1 score with SHAP. These results suggest that sentence-level event detection models rely on event triggers to make predictions, and our hypothesis is correct. Furthermore, as the weakly-supervised models have provided good results, we can suggest using event triggers as explanations for sentence-level event detection models. The methods that we explored can be considered as a baseline for explainable event detection. In addition, we have the following key observations from the results.

SHAP performs better than LIME: As shown in the results, LIME-based explanations are substantially outperformed by the SHAP-based explanations in all most all the models. This suggests that SHAP create better explanations for sentence-level event detection models.

Strong sentence-level models and explainability: All the models and strategies we experimented with provided compatible sentence-level results with each other. However, these models' weakly-supervised event trigger detection results vary a lot. Several models that had high sentence-level scores provided poor results in event trigger detection. This suggests that stronger sentence-level models do not always guarantee strong explainability.

Multilingual models and explainability: The results in Table 4 shows that multilingual models behave poorly in weakly-supervised event trigger detection. Language-specific transformer models with ***Monolingual*** strategy performed best in all the languages and substantially outperformed multilingual transformer models with ***All*** and ***All-1*** strategies. This result is clear in SHAP and we can assume that SHAP requires language-specific transformers to perform better.

High recall and low precision: As shown in Table 4, all the models result in high recall and low

(a) Ground truth

(b) ROBERTA-LARGE + SHAP

Figure 3: Wordclouds of actual and predicted event triggers in English after removing stop words

precision, which means that our weakly-supervised approach results in many false positives. We manually analyse this scenario with our best English model (i.e. ROBERTA-LARGE) for weak supervision.

> ***Ground truth*** - One of the men who led the strike at Lonmin's platinum mine in August 2012 denied on Monday that he played any part in the fatal attacks that occurred.
>
> ---
>
> ***Our predictions*** - One of the men who led the strike at Lonmin's platinum mine in August 2012 denied on Monday that he played any part in the fatal attacks that occurred.

> ***Ground truth*** - Maharashtra police also overlooked the fact that Naidu was sick as he had observed a day-long fast yesterday and spent over four days without proper facilities.
>
> ---
>
> ***Our predictions*** - Maharashtra police also overlooked the fact that Naidu was sick as he had observed a day-long fast yesterday and spent over four days without proper facilities.

As can be seen in the examples, our weak supervision approach detects words including the stop words around the actual trigger word as triggers. As a result, our approach's precision is low.

Finally, in Figure 3a we show the word cloud of actual event triggers in English data. In Figure 3b we show the word cloud of predicted event triggers by SHAP for English. As can be seen, our approach detects the most common event trigger words correctly.

6 Conclusions

In this work, we proposed a new weakly-supervised approach for event trigger detection by exploring feature attribution methods on sentence-level event detection models. Our results show that sentence-level event detection models rely on event triggers to make predictions, and in turn, event triggers can be used as explanations for sentence-level models. We hope this work will encourage further research on improving the efficiency of event trigger detection models with weakly-supervised methods. Also, we believe our findings will be useful for social media event detection considering the volume and dynamicity of data (Hettiarachchi et al., 2023b). This work presents numerous avenues for future research, ranging from enhancing achieved outcomes through combining different feature attribution methods to investigating alternative underlying architectures and training objectives at the sentence-level.

Acknowledgments

We thank the anonymous RANLP reviewers who have provided us with constructive feedback to improve the quality of this paper.

The computational experiments in this paper were conducted on Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146, Online. Association for Computational Linguistics.

Ali Balali, Masoud Asadpour, Ricardo Campos, and Adam Jatowt. 2020. Joint event extraction along shortest dependency paths using graph convolutional networks. *Knowledge-Based Systems*, 210:106492.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2012. Joint modeling for Chinese event extraction with rich linguistic features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 529–544, Mumbai, India. The COLING 2012 Organizing Committee.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Erica Chenoweth and Orion A Lewis. 2013. Unpacking nonviolent campaigns: Introducing the navco 2.0 dataset. *Journal of Peace Research*, 50(3):415–423.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mihai Croicu and Nils B Weidmann. 2015. Improving the selection of news reports for event coding using ensemble classification. *Research & Politics*, 2(4):2053168015615596.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Ruixue Ding and Zhoujun Li. 2018. Event extraction with deep contextualized word representation and multi-attention layer. In *Advanced Data Mining and Applications*, pages 189–201, Cham. Springer International Publishing.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. DAAI

at CASE 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 120–130, Online. Association for Computational Linguistics.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2023a. TTL: transformer-based two-phase transfer learning for cross-lingual news event detection. *International Journal of Machine Learning and Cybernetics*.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2023b. Whatsup: An event resolution approach for co-occurring events in social media. *Information Sciences*, 625:553–577.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Tiancheng Hu and Niklas Stoehr. 2021. Team "NoConflict" at CASE 2021 task 1: Pretraining for sentence-level protest event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 152–160, Online. Association for Computational Linguistics.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended multilingual protest news detection - shared task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021a. Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021b. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, 3(2):308–335.

Swen Hutter. 2014. 335Protest Event Analysis and Its Offspring. In *Methodological Practices in Social Movement Research*. Oxford University Press.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reaction classification with deep neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 877–887, Osaka, Japan. The COLING 2016 Organizing Committee.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Canete José, Chaperon Gabriel, Fuentes Rodrigo, and Pérez Jorge. 2020. Spanish pre-trained bert model and evaluation data. In *Proceedings of the Workshop on Practical Machine Learning for Developing Countries (PML4DC)*.

S. Lawrence, C.L. Giles, Ah Chung Tsoi, and A.D. Back. 1997. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in Dutch news text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 330–335, Portorož, Slovenia. European Language Resources Association (ELRA).

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1).

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Shulin Liu, Yang Li, Feng Zhang, Tao Yang, and Xinpeng Zhou. 2019a. Event detection without triggers. 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 735–744, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 352–355.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

M. Naughton, N. Stokes, and J. Carthy. 2010. Sentence-level event classification in unstructured texts. *Information Retrieval*, 13(2):132–156.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

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

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2022. Sold: Sinhala offensive language dataset. *arXiv preprint arXiv:2212.00851*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Waddah Saeed and Christian Omlin. 2023. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Sidney Tarrow. 2022. *Power in movement*. Cambridge university press.

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens. 2022. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2):315–335.

Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Jianshu Weng and Bu-Sung Lee. 2021. Event detection in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):401–408.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.