Zero-Shot Ranking Socio-Political Texts with Transformer Language Models to Reduce Close Reading Time

Kiymet Akdemir
Boğaziçi University
kiymet.akdemir@boun.edu.tr

Ali Hürriyetoğlu
KNAW Humanities Cluster DHLab
ali.hurriyetoglu@dh.huc.knaw.nl

Abstract

We approach the classification problem as an entailment problem and apply zero-shot ranking to socio-political texts. Documents that are ranked at the top can be considered positively classified documents and this reduces the close reading time for the information extraction process. We use Transformer Language Models to get the entailment probabilities and investigate different types of queries. We find that DeBERTa achieves higher mean average precision scores than RoBERTa and when declarative form of the class label is used as a query, it outperforms dictionary definition of the class label. We show that one can reduce the close reading time by taking some percentage of the ranked documents that the percentage depends on how much recall they want to achieve. However, our findings also show that percentage of the documents that should be read increases as the topic gets broader.

1 Introduction

For the information retrieval process positively labeled documents in a dataset are important and should not be missed, therefore achieving high recall is extremely important. However, there is generally a large number of documents that are relevant or not to the concerned topic and doing close reading for all documents and annotating them requires lots of time and resources (Hürriyetoğlu et al., 2016; Hürriyetoğlu et al., 2017). Therefore, ranking documents according to relevance to the investigated class may help to reduce close reading time and decrease the likelihood of missing critical information.

Baeza-Yates and Ribeiro-Neto (1999) propose ranking documents in decreasing order of being relevant to a given query to accelerate the information retrieval process. Halterman et al. (2021) apply this method with Natural Language Understanding (NLU) models for binary classification problems using the entailment probabilities of a document and a declarative form of the label. Therefore, to catch a high percentage of positively labeled documents, reading some percentage of documents would be enough since documents that are relevant would be at the top with a high probability. However, their dataset India Police Events focuses on a relatively specific task in information retrieval that is police actions like killing, arresting, failing to intervene, etc. Besides, they apply this method at the sentence level and as they also stated their model suffers from understanding multi-sentence context that increases the false negative rate.

We apply this approach to ProtestNews dataset (Hürriyetoğlu et al., 2021) along with the India Police Events dataset (Halterman et al., 2021) and investigate whether sentence level evaluation or document level evaluation ranks positive documents at the higher level measured by different evaluation metrics. We further investigate whether using the dictionary definition of a class or the declarative form of a class for the query performs this task better. We compare two NLU models DeBERTa-Large-MNLI (He et al., 2020) and RoBERTa-Large-MNLI (Liu et al., 2019) in terms of recall and mean average precision.

We present the related work in Section 2. Next, we introduce two datasets we used in our experiments in Section 3. Then we explain our methodology and list all queries used in this work in Section 4. We detail our experiments for both datasets and present results in Section 5. Finally, we conclude this work in Section 6 and state what can be done as future work in Section 7.

2 Related Work

Protest Event Detection Protest event extraction holds an important place in political social sciences and detection of protest events is generally the first step of the extraction. Due to the cost of manual event extraction, besides the presence of digital news articles and enhancing machine learning

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 124 - 132
December 7-8, 2022 ©2022 Association for Computational Linguistics

methods; automated event extraction comes into play.

Hanna (2017) presents MPEDS, an automated system for protest event extraction that contains an ensemble of shallow machine learning classifiers (SVM, SGD and Logistic Regression) to detect protest-related documents. Caselli et al. (2021) proposes Domain Adaptive Retraining for Transformer Language Models and shows that further training BERT with domain-specific dataset improves the performance. They present PROTEST-ER by retraining pre-trained BERT with protest related data from TREC Washington Post Corpus. Wiedemann et al. (2022) classifies protest related documents in German local news using Pretrained Language Models. They attempt to improve performance and generalizability by eliminating protest-unrelated sentences with keyword search and also by masking named entities with the idea of models may overfit on data by recognizing actors, organizatons and places.

Elsafoury (2019) focuses on both protest events and police actions i.e. protest repression events in Twitter with Machine Learning models with the claim of news articles suffer from bias, censorship and duplication. Won et al. (2017) detects and analyze protest events in geotagged tweets and associated images with Convolutional Neural Networks.

Ranking Documents with Transformer Language Models Yates et al. (2021) presents a comprehensive survey of how BERT (Devlin et al., 2019) works, ranking documents with BERT, retrieve and rerank approach with monoBERT, ranking metrics, etc. One of the most remarkable works in the survey is monoBERT and duoBERT, a multistage ranking approach with transformer language models proposed by Nogueira et al. (2019). The first stage retrieves the candidate documents with BM25 by treating the query as a bag of words and later, documents are reranked with their relevance score with BERT. DuoBERT also takes into account one document being more relevant than the other at a third stage. However, we rank the documents with a language model at one stage.

Halterman et al. (2021) rank documents with RoBERTa-Large-MNLI (Liu et al., 2019) on sentence level by being relevant to a police activity. Yet sentence level evaluation does not take into consideration the relationship between the sentences. Moreover, the task of extracting police events is a relatively specific topic in political event extraction. We apply this method with different document sizes and test on datasets in different topic specificities.

Transformer Language Models DeBERTa and RoBERTa DeBERTa-Large-MNLI (DLM) (He et al., 2020) and RoBERTa-Large-MNLI (RLM) (Liu et al., 2019) are pre-trained language models that improve BERT. Both models are pre-trained on Wikipedia (English Wikipedia dump3; 12GB), BookCorpus (6GB), OPENWEBTEXT (38GB), and STORIES (a subset of CommonCrawl (31GB) and fine-tuned for MNLI task. RLM has a token limitation of 512 whereas DLM has a limitation of theoretically 24,528. We limit the inputs to 512 tokens for both models to be able to compare them fairly. Ye and Manoharan (2021) find that DLM achieves a better performance in different sentence similarity tasks with respect to RLM and BERT. He et al. (2020) also show that DeBERTa outperforms RoBERTa in a variety of NLP tasks even when DeBERTa is trained on half of the training data. Therefore, we use DLM and compare it with RLM for our task.

Transferring Question Answering to Entailment Problem Khot et al. (2018) and Demszky et al. (2018) transfer the question answering problem to the entailment problem by forming the question into a declarative form. Clark et al. (2019) transfer yes/no question answering to entailment problem by training supervised models on entailment datasets and treating entailment probabilities as the probability of the answer being yes. They also use pre-trained ELMo, BERT, and OpenAI GPT as unsupervised models and show that fine-tuning BERT on entailment dataset MultiNLI boosts the performance. The problem of any binary classification can be also transferred to an entailment problem similar to the yes/no question answering, by considering the probability of entailment as the probability of data belonging to the positive class.

3 Data

We carried out the experiments on two different datasets: India Police Events dataset[1] (Halterman et al., 2021) and the ProtestNews dataset of the workshop CASE @ ACL-IJCNLP 2021[2] (Hürriyetoğlu et al., 2021).

[1] Data and code are provided at https://github.com/slanglab/IndiaPoliceEvents.

[2] Information and data are provided at https://github.com/emerging-welfare/case-2021-shared-task.

Event type	Question
kill	Did police kill someone?
arrest	Did police arrest someone?
fail	Did police fail to intervene?
force	Did police use force or violence?
any action	Did police do anything?

Table 1: Question form of each event type.

Event type	Positive Documents
kill	50 (3.98%)
arrest	128 (10.17%)
fail	114 (9.05%)
force	90 (7.15%)
any action	457 (36.24%)

Table 2: Number of positive documents for each event class (India Police Events Dataset).

India Police Events dataset includes 1,257 articles about the Indian state Gujarat from The Times of India and from March 2002. The articles are in English and contain 21,391 sentences in total. Each sentence is classified into 5 different labels regarding police activity: kill, arrest, fail, force, and any action. Question form of the each event type is given in Table 1. A document belongs to a class if any of its sentences belongs to that class. Table 2 illustrates the number of positive documents and the proportion of the positive documents for each event class. Note that one document may belong to one class, several classes or none of them.

ProtestNews dataset includes local news articles of countries India, China, Argentina, and Brazil. These articles are in English, Spanish, Portuguese, and Hindi. For this work, we have only used English articles. There are 9,327 English documents but to equalize data sizes with the India Police Events Dataset we randomly selected 1,257 articles among those. Documents that contain past or ongoing protest events are labeled as positive (Duruşan et al., 2022). Number and proportion of positive documents are given in Table 3.

Dataset	Positive Documents
ProtestNews Dataset	1,912 (20.51%)
ProtestNews Subset	268 (21.32 %)

Table 3: Number of positive documents for ProtestNews Dataset and its subset.

4 Method

First, the probability of entailment for each document and a query is calculated with NLU models from Huggingface[3], and documents are ranked by the decreasing probability of being relevant to the query. Thus we expect the documents that are more relevant are ranked at the top.

Entailment probabilities are evaluated on both sentence and document levels. At sentence level evaluation, entailment probabilities of sentences in a document with the given query are calculated and the largest probability among the sentences is considered as the probability of the document being relevant. For the document level evaluation since RLM is limited to 512 tokens, we divided documents into parts such that each part does not exceed 512 tokens. Similar to the sentence-level approach, probabilities of each part are calculated and the one with the largest probability is considered as the probability of the document. After getting the probabilities for all documents, they are ranked in the decreasing probability.

We compare the results by checking how much recall is achieved when a specified proportion of data is read from the ranked documents following Halterman et. al. (2021) and also by calculating the mean average precision. We release our code publicly[4].

4.1 Models

We focused on the performances of two multilingual NLU models that are RLM[5] (Liu et al., 2019) and DLM[6] (He et al., 2020) which are pre-trained on the same datasets (Wikipedia and BookCorpus). We conduct experiments with both models and compare the results.

4.2 Queries

We have experimented with different types of queries: definitional queries, extended definitional queries and declarative queries.

We used the Cambridge Dictionary[7] and form the definitional queries by using the definitions of the class name (protest, kill, arrest, etc.). Annota-

[3] http://huggingface.co

[4] https://github.com/kiymetakdemir/zero-shot-entailment-ranking

[5] https://huggingface.co/roberta-large-mnli

[6] https://huggingface.co/microsoft/deberta-large-mnli

[7] https://dictionary.cambridge.org

Query type	Query
Declarative query	There is a protest.
Definitional query	There is a strong complaint expressing disagreement, disapproval, or opposition. (definition of protest[9])
Social protest definition (Annotation Manual)	Individuals, groups, or organizations voice their objections, oppositions, demands or grievances to a person or institution of authority.
Contentious politics event definition (Annotation Manual)	There is a politically motivated collective action event.
'protest' + definitional query	Protest, there is a strong complaint expressing disagreement, disapproval, or opposition.
Protest definition + opposition definition	There is a strong complaint expressing disagreement, disapproval, or opposition. Disagreement with something, often by speaking or fighting against it, or (esp. in politics) the people or group who are not in power. (definition of opposition[10])
Protest definition + disapproval definition	There is a strong complaint expressing disagreement, disapproval, or opposition. The feeling of having a negative opinion of someone or something. (definition of disapproval[11])

Table 4: Queries used for the ProtestNews dataset.

Event type	Declarative query	Definitional query
kill	Police killed someone.	Police caused someone or something to die. (definition of kill[12])
arrest	Police arrested someone.	Police used legal authority to catch and take someone to a place where the person may be accused of a crime. (definition of arrest[13])
fail	Police failed to intervene.	Police failed to have an effect. (definition of act[14])
force	Police used violence.	Police used actions or words that are intended to hurt people. (definition of violence[15])
any action	Police did something.	Police have an effect. (definition of act)

Table 5: Queries for the India Police Events dataset.

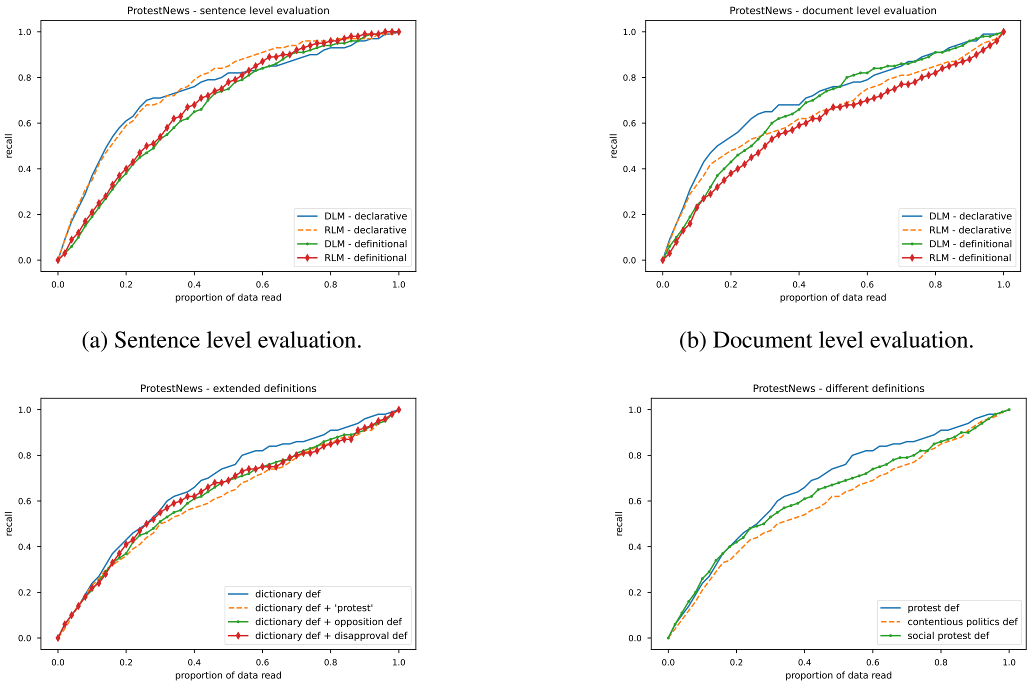
tion manual may possibly be a good resource to find the definition of the investigated class. For this reason, we also experimented with definitions from Annotation Manual[8] (Duruşan et al., 2022). On the other hand, a declarative query is a sentence that simply describes the class. For instance, we use "There is a protest." as the declarative query for the ProtestNews dataset. For the India Police Events dataset, we use declarative queries proposed by Halterman et al. (2021).

We also extended protest dictionary definition by concatenating it with the definitions of words that pass in the query (see last 3 rows in Table 4). In one of the queries, the 'protest' word is added to the beginning of the protest definition. In the other one, definition of opposition is concatenated with the protest definition. In the third one, definitions of protest and definition of disapproval are concatenated and used as a query. Note that we used the definitions of opposition and disapproval since they occur in the protest definition. All queries used for both datasets are listed in Table 4 and 5.

5 Experiments & Results

ProtestNews Dataset is tested with declarative queries, definitional queries and extended definitions on models DLM and RLM and results are presented in Figure 1a for the sentence level evalu-

[8] https://github.com/emerging-welfare/general_info/tree/master/annotation-manuals



(a) Sentence level evaluation.

(b) Document level evaluation.

(c) Extended definitions.

(d) Different definitions.

Figure 1: ProtestNews dataset tested on two models: RLM and DLM.

ation. The x-axis represents what percentage of the data is read and the y-axis represents how much recall is achieved at that stage. One can investigate what percentage of the data should be read to achieve a specified recall. We see that both models yield similar results when the same query is given but positive documents are accumulated at more top with the declarative query compared to the definitional query.

For document level evaluation, Figure 1b illustrates the comparison of the models. DLM achieves higher recall scores than RLM, however, the query type does not affect the performance of the model at the document level significantly.

We compare the extended and Annotation Manual definitions at document level using the DLM model since the DLM achieves higher recall compared to RLM at the document level as in Figure 1b. However, from Figure 1c we see that extending the protest definition performs slightly worse than using the only dictionary definition. Also, Annotation Manual definitions do not perform better than the dictionary definition as we see from Figure 1d.

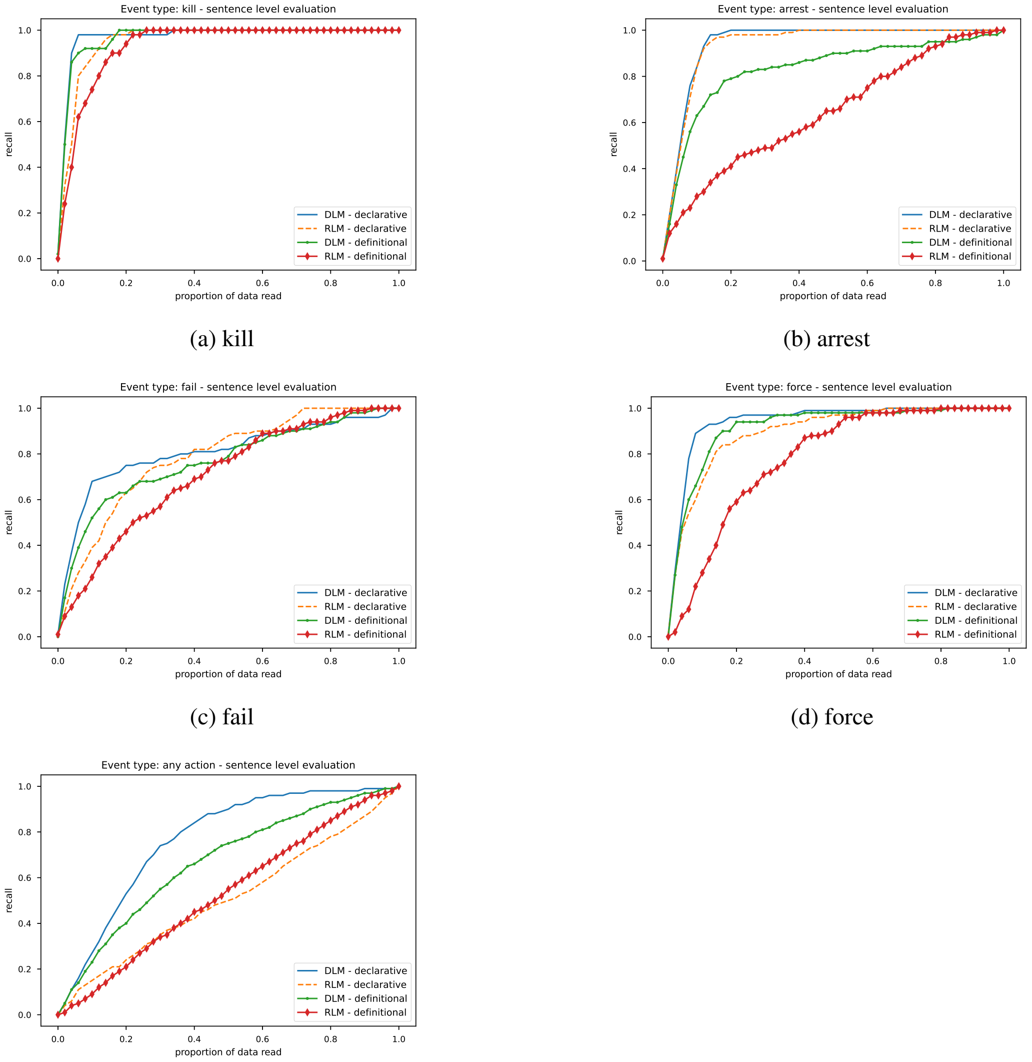
India Police Events Dataset is tested with declarative and definitional queries on RLM and DLM as in ProtestNews dataset. For all event types, we see from Figure 2 and Figure 3 that DLM with declarative query gives the best result that is positive documents are accumulated at more top-level, whereas RLM with a definitional query stays behind other combinations of model and queries.

Mean Average Precision (mAP) is calculated for each ranking and reported in Table 6. Query and document length combination that gives the highest mAP is marked in bold for each dataset and event type.

For the ProtestNews dataset we observe that using models DLM or RLM, and document lengths do not differ significantly. Whereas using a declarative query gives much better mAP than the definitional query. For the India Police Events dataset for all event types DLM and declarative query with the sentence level evaluation yield the highest score rather than the definitional or document level evaluation. Besides, note that there is a large difference with the other combinations. For example for event type force, sentence level evaluation with DLM and the declarative query gives 0.91 mAP whereas document level evaluation with RLM and the definitional query yields 0.11 mAP.
As the topic gets broader, we see that performance gets worse in Table 6. For instance, kill is a more

		ProtestNews	**India Police Events**				
		-	kill	arrest	fail	force	any action
DLM	decl-sent	0.64	**0.96**	**0.94**	**0.65**	**0.91**	**0.89**
DLM	decl-doc	0.60	0.80	0.75	0.25	0.75	0.80
DLM	def-sent	0.35	0.89	0.63	0.47	0.71	0.69
DLM	def-doc	0.41	0.62	0.42	0.21	0.21	0.65
RLM	decl-sent	**0.65**	0.55	0.91	0.34	0.66	0.42
RLM	decl-doc	0.51	0.18	0.44	0.18	0.27	0.36
RLM	def-sent	0.38	0.36	0.26	0.23	0.18	0.38
RLM	def-doc	0.34	0.11	0.15	0.16	0.11	0.37

Table 6: mAP scores for DLM and RLM models with different document lengths and queries.


Event type: kill - sentence level evaluation

Event type: arrest - sentence level evaluation

Event type: fail - sentence level evaluation

Event type: force - sentence level evaluation

Event type: any action - sentence level evaluation

(a) kill

(b) arrest

(c) fail

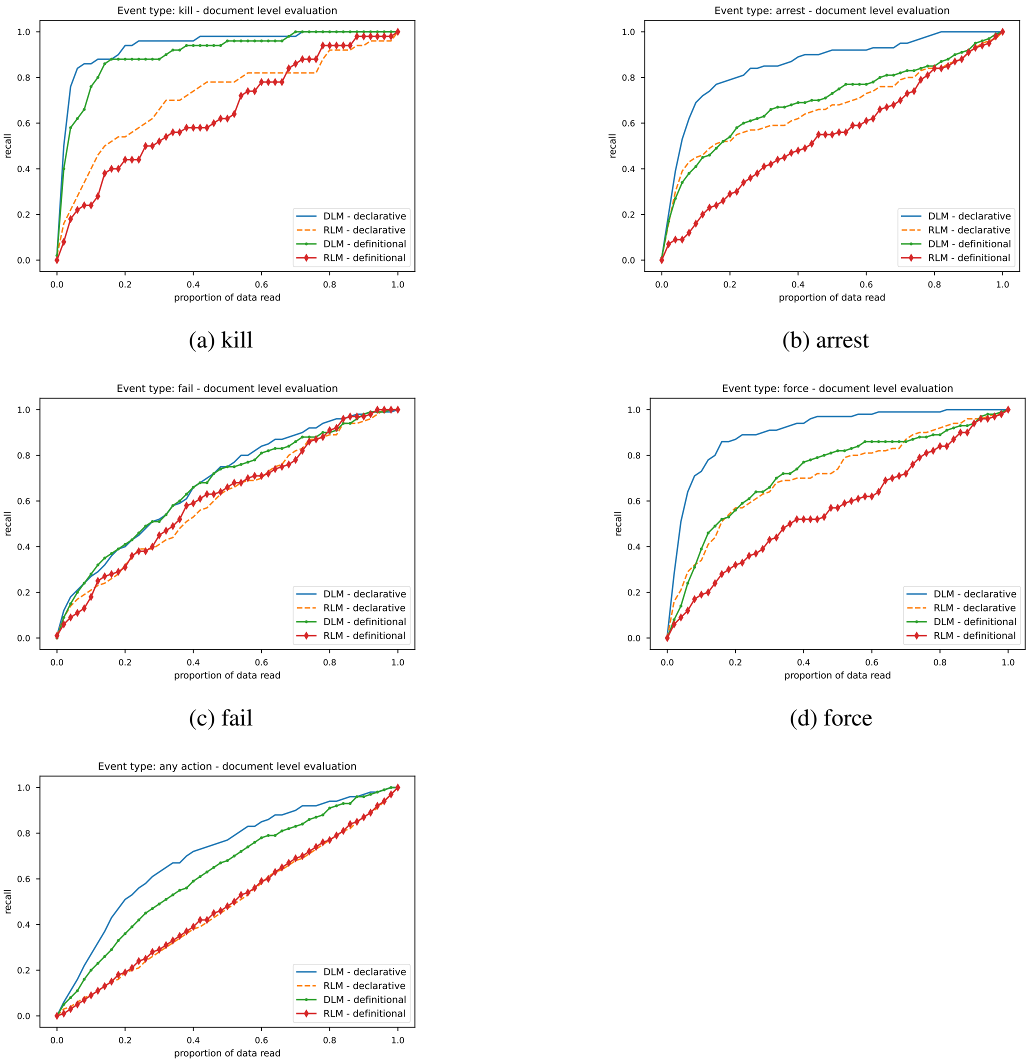
(d) force

(e) any action

Figure 2: India Police Events dataset sentence level evaluation tested on RLM and DLM.

specific topic than any action since any action event type also includes kill events. When 20% of the data read, 90% recall is achieved for event type kill, on the other hand, even 60% recall is not reached for any action.

We take the average sentence and document level mAP scores for each model and present in Table 7. For ProtestNews dataset, sentence or document



(a) kill

(b) arrest

(c) fail

(d) force

(e) any action

Figure 3: India Police Events dataset document level evaluation tested on RLM and DLM.

level does not differ in mAP when DLM is used. However, for India Police Events dataset sentence level evaluation achieves much higher mAP than document level evaluation (0.24 mAP increase for DLM and 0.21 increase for RLM). For both sentence and document level, DLM reaches higher mAP than RLM.

	ProtestNews		**India Police Events**	
	DLM	RLM	DLM	RLM
sent	**0.50**	**0.52**	**0.77**	**0.43**
doc	**0.50**	0.42	0.53	0.22

Table 7: Average mAP on ProtestNews and India Police Events Dataset for all event types.

6 Conclusion

We investigate the performances of two Transformer Language Models (DLM and RLM), different query types (declarative and definitional) in different document lengths (document and sen-

[9] https://dictionary.cambridge.org/dictionary/english/protest

[10] https://dictionary.cambridge.org/dictionary/english/opposition

[11] https://dictionary.cambridge.org/dictionary/english/disapproval

[12] https://dictionary.cambridge.org/dictionary/english/kill

[13] https://dictionary.cambridge.org/dictionary/english/arrest

[14] https://dictionary.cambridge.org/dictionary/english/act

[15] https://dictionary.cambridge.org/dictionary/english/violence

tence level). Our experiments that conclude DLM achieves higher mAP scores than RLM are consistent with the findings of Ye and Manoharan (2021) and He et al. (2020). In general, we find that the combination of DLM with a declarative query in sentence level outperforms other combinations in mAP score. However, scores decrease as the topic or event type gets broader where protest events can be considered broader than specific police actions.

7 Future Work

We plan to analyze results more for example by considering subcategories of protest events for the ProtestNews dataset. Future work can extend this work to a different political event classification dataset and further investigate the association between the broadness of the topic and metric scores. Experiments in languages other than English are also left as future work.

References

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. PROTEST-ER: Retraining BERT for protest event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets.

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.

Fırat Duruşan, Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Çağrı Yoltar, Burak Gürel, and Alvaro Comin. 2022. Global contentious politics database (glocon) annotation manuals.

Fatma Elsafoury. 2019. *Detecting protest repression incidents from tweets*. Ph.D. thesis, University of Glasgow.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253, Online. Association for Computational Linguistics.

Alex Hanna. 2017. Mpeds: Automating the generation of protest event data.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. 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, Nelleke Oostdijk, Mustafa Erkan Başar, and Antal van den Bosch. 2017. *Supporting Experts to Handle Tweet Collections About Significant Events*, pages 138–141. Springer International Publishing, Cham.

Ali Hürriyetoğlu, Christian Gudehus, Nelleke Oostdijk, and Antal van den Bosch. 2016. Relevancer: Finding and labeling relevant information in tweet collections. In *Social Informatics - 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II*, volume 10047 of *Lecture Notes in Computer Science*, pages 210–224.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy J. Lin. 2019. Multi-stage document ranking with bert. *ArXiv*, abs/1910.14424.

Gregor Wiedemann, Jan Matti Dollbaum, Sebastian Haunss, Priska Daphi, and Larissa Daria Meier. 2022. A generalized approach to protest event detection in German local news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3883–3891, Marseille, France. European Language Resources Association.

Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2666–2668, New York, NY, USA. Association for Computing Machinery.

Xinfeng Ye and Sathiamoorthy Manoharan. 2021. Performance comparison of automated essay graders based on various language models. In *2021 IEEE International Conference on Computing (ICOCO)*, pages 152–157.