IIITT at CASE 2021 Task 1: Leveraging Pretrained Language Models for Multilingual Protest Detection

Pawan Kalyan Jada[1]**, Duddukunta Sashidhar Reddy**[1]**, Adeep Hande**[1]**,**
Ruba Priyadharshini[2]**, Ratnasingam Sakuntharaj**[3]**, Bharathi Raja Chakravarthi**[4]
[1] Indian Institute of Information Technology Tiruchirappalli
[2] ULTRA Arts and Science College, India,   [3]Eastern University, Sri Lanka
[4]Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
pawankj19c@iiitt.ac.in

Abstract

In a world abounding in constant protests resulting from events like a global pandemic, climate change, religious or political conflicts, there has always been a need to detect events/protests before getting amplified by news media or social media. This paper demonstrates our work on the sentence classification subtask of multilingual protest detection in CASE@ACL-IJCNLP 2021. We approached this task by employing various multilingual pre-trained transformer models to classify if any sentence contains information about an event that has transpired or not. Furthermore, we performed soft voting over the models, achieving the best results among the models, accomplishing a macro F1-Score of 0.8291, 0.7578, and 0.7951 in English, Spanish, and Portuguese, respectively. The source codes for our systems are published[1].

1   Introduction

The recent surge in social media users has led many people to express their opinions on various global issues. These opinions travel far and wide within a matter of seconds (Hossny et al., 2018). This can influence many people and may engage public movements (Won et al., 2017a). Therefore, there is a definite need to detect these protests and analyse them to know the significant areas of disinterest.

Being a free and easy to use platform, social media has become a part of our day to day life. It incorporates people of different ages, gender, location, religions, background, and so on. The enormous number of rich and diversified users results in an enormous amount of information being generated, which is helpful in many ways (Kapoor et al., 2018). Some of this even contains private information about the users, which others could misuse. Cases were also found where certain users

---

[1]https://github.com/adeepH/
CASE-2021-Task-1

were being targeted and harassed by people using this platform, a common scenario in cyberbullying (Abaido, 2020).

Social media plays a crucial role in amplifying these protests and movements (Won et al., 2017b). It enables political groups and protesters to organise protest movements and share information. It acts as a platform for the people who are underrepresented by giving a voice to them. It also offers new opportunities for people to engage in activism, political resistance, and protest outside the political groups and civic institutions. Thus, it has a social impact on everyone (Pulido et al., 2018). It is to be noted that social media, similar to news media, plays a vital role in its social and political events worldwide (Holt et al., 2013). For the above reasons, we can state that social media plays a crucial role in most worldwide events.

The English language is widely regarded as the first *Lingua Franca*. Statistically, it is one of the most widely spoken languages globally, having official status in over 53 countries (Crystal, 2008). Over 400 million people speak English as their primary language and widely spoken in the United States and the United Kingdom. *BlackLivesMatter* (Dave et al., 2020), *EarthDay* (Rome, 2010) are some of the major protests that have occurred in these countries. Español commonly referred to as Spanish, is spoken by over 360 million people worldwide, with most of its speakers residing in Mexico, Argentina, Spain. *15-M Movement* (Casero-Ripollés and Feenstra, 2012) and *YoSoy132* (García and Treré, 2014) are some of the recent protests where people have been vocal about in the Spanish language. Portuguese has over 220 million native speakers. Brazil, Portugal, Angola are some of the major countries where this language is spoken. Protests like *Racism Kills*, *May 68* (Ross, 2008) are the recent ones that occurred in the Portuguese language.

The recent upheavals of protests are due to so-

Sentence	Language	Label
Fabius ran against Royal for the presidential nomination in 2007.	English	Event
He planned to start a race war.	English	Event
Metro police intervened and the fire was put out.	English	Not-event
Pero no es ése el mayor problema.	Spanish	Event
La Argentina retrocedería un paso todos los días.	Spanish	Event
Carrió no objetó que se trató de un secuestro.	Spanish	Not-event
Os servidores do Piauí estão em greve há 17 dias.	Portuguese	Not-event
É uma nova experiência mobilizatória.	Portuguese	Event
É decidiram ir às aulas e passar o dia de saia.	Portuguese	Not-event

Table 1: Examples of the dataset indicating events of the past and not-events.

cial media, youth, exaggeration of certain events. (Basile and Caselli, 2020). Any early detection of mass protest detection through social media platforms such as Facebook, Twitter, and Instagram to help minimizing the aftermath of the protests (Wilson, 2017). This has motivated Natural Language Processing (NLP) researchers to develop NLP systems to generalize on data coming from diverse sources to leverage the NLP systems to more realistic environments (Büyüköz et al., 2020). Hence, there is a need to develop NLP systems that could be generalized to any protest/events (Peng et al., 2013), which has motivated us to participate in the shared task for multilingual protest detection (Hürriyetoğlu et al., 2019a, 2021) The objective of the task is to identify if any sentence talks about any mentions of protests or events in three languages, namely, English, Spanish, and Portuguese. Hence, we treat this as a sequence classification task.

The rest of the paper is organized as follows, Section 2 presents previous work on protest detection and analysis. Section 3 entails a comprehensive analysis of the dataset used for our cause. Next, section 4 gives a detailed description of the models used for the multilingual event detection. Finally, section 5 analyses the results obtained, and Section 6 concludes our work while discussing the potential directions for future work.

2 Related Work

The need to detect events that could lead to protests is of prime interest to sociologists and governments (Danilova et al., 2016). There are several active ongoing projects for socio-political event systems such as KEDS (Kansas Event Data System) (Schrodt and Hall, 2006), CAMEO (Conflict and Mediation Event Observation) (Gerner et al., 2002), and several other databases for protest de-

tection systems (Danilova, 2015). These methods have focused on news data as they have traditionally been the most reliant source of events. Protest detection has been one of the major issues in the context of social and political (Ettinger et al., 2017). Papanikolaou and Papageorgiou (2020) presented a computational social science methodology to analyse protests in Greece. Hürriyetoğlu et al. (2021) constructed a corpus of protest events comprising various language sources from various countries. Several systems were submitted to the CLEF ProtestNews Track that consisted of three shared tasks, primarily aimed at identifying and extracting event information spanning to multiple countries (Hürriyetoğlu et al., 2019b, 2020).

3 Dataset

This dataset comprises 26,208 sentences in three languages, namely English, Spanish, and Portuguese. The dataset consists of two classes:

- **Event**: The sentence indicates an event of the past.

- **Not-event**: The sentence does not talk about any event.

The volume of sequences indicating *Not-event* is higher in contrast to that of the *Event* label. Therefore, the dataset distribution is quite imbalanced. We can also notice that the number of English samples exceeds that of Spanish and Portuguese ones. Refer to Table 1 for examples of sentences talking about events and not talking about events displayed in English, Spanish, and Portuguese. The dataset distribution is displayed in Table 2. For our cause, we split the training and validation set in the ratio of 80:20.
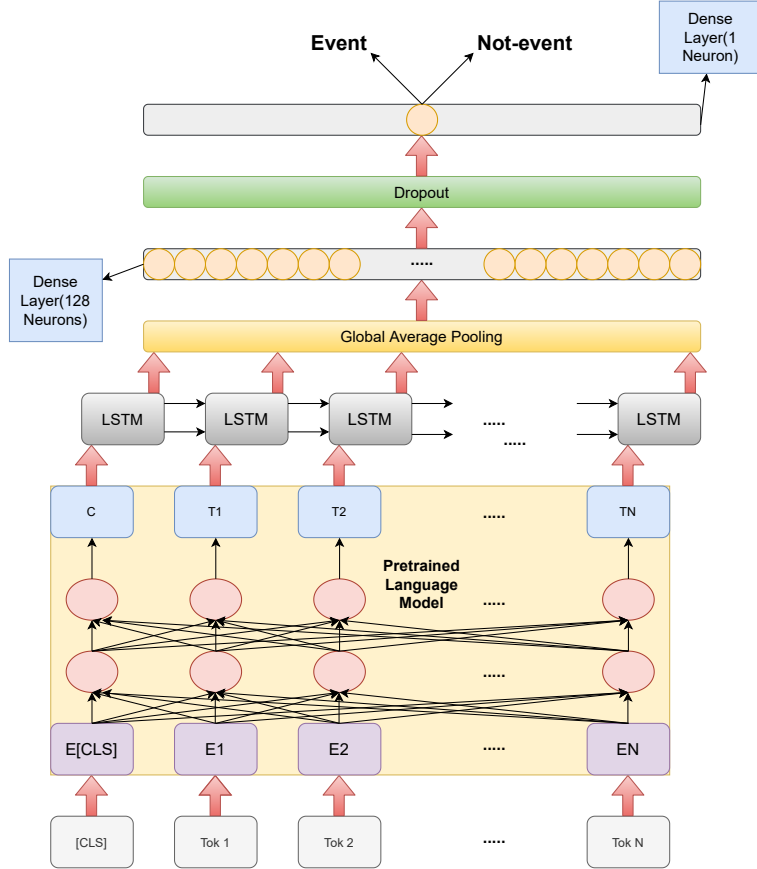
Figure 1: System Architecture based on BERT (Devlin et al., 2019)

Language	English	Spanish	Portuguese
Not-event	18,602	2,291	901
Event	4,223	450	281
Total	22,285	2,741	1,182

Table 2: Classwise distribution of the training set

4 Methodology

We used pretrained transformer-based models for identifying if a sentence talks about an event or not. The models that were used are BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019). Even though there are 3 different languages, we used a single model for all three due to memory constraints and reduced training time. We fine-tuned these models for sequence classification. Soft Voting is done on all these models to produce the respective final outputs for the languages. In soft voting, each classifier predicts that a specific data point belongs to the particular target class. A weighted sum of the predictions is done based on the importance of the classifier (all models have equal weights). The overall prediction is chosen as the target with the greatest sum

of the weighted probability, thus winning the vote (Beyeler, 2017; Hande et al., 2021).

4.1 BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin et al., 2019) is a pretrained language model which was created with the objective that fine-tuning a pretrained model yields better performance. BERT's pretraining phase includes two tasks. Firstly, Masked Language Modeling (MLM) is where certain words are randomly masked in a sequence. About 15% of the words in a sequence is masked. The model then attempts to predict the masked words. Secondly, Next Sentence Prediction (NSP), where the model has an additional loss function, NSP loss, indicates if the second sequence follows the first one. Around 50% of the inputs are a pair, and they randomly chose the other 50. Here, we use a *bert-base-multilingual-cased* (Pires et al., 2019) trained on top of 104 languages in the largest Wikipedia corpus. This model has 12 layers, 12 Attention heads with over 179 million parameters.

	Event			Not-event			Overall			
Model	P	R	F1	P	R	F1	Acc	M(P)	M(R)	M(F1)
mBERT	0.928	0.917	0.922	0.641	0.675	0.658	0.873	0.784	0.796	0.790
DistilmBERT	0.924	0.947	0.936	0.729	0.646	0.685	0.893	0.827	0.797	0.810
RoBERTa	0.910	0.938	0.924	0.670	0.578	0.621	0.873	0.790	0.758	0.772
SoftVoting	0.937	0.939	0.938	0.720	0.713	0.717	**0.899**	**0.829**	**0.826**	**0.827**

Table 3: Precision (P), recall (R), and F1-Score of the models on the validation set; M(P), M(R), and M(F1) are the Macro averages of precision, recall, and F1-Score respectively

4.2 DistilmBERT

DistilBERT (Sanh et al., 2019) is the distilled version of BERT. DistilBERT employs a triple loss language modelling, where it integrates cosine distance loss with knowledge distillation. DistilBERT has 40% fewer parameters than BERT but still promises 97% of the latter's performance. It is also 60% faster than BERT. In this system, we used a cased multilingual DistilBERT model as they are three different languages. For our cause, we fine-tune *distilbert-base-multilingual-cased*, which is distilled from the mBERT checkpoint. The model has 6 layers, 768 dimensions, and 12 Attention heads, totalizing about 134 million parameters.

4.3 RoBERTa

Robustly Optimized BERT (RoBERTa) (Liu et al., 2019) follows the same architecture of BERT while differing in the pretraining strategy. It is pretrained with MLM as its objective where the model tries to predict the masked words. RoBERTa model is trained on the vast English Wikipedia and CC-News datasets. The NSP is not employed as a pretraining strategy, and the tokens are dynamically masked, making the model slightly different to BERT. During tokenization, RoBERTa follows byte-pair encoding (BPE) (Gallé, 2019) as opposed to WordPiece employed in BERT. We use *roberta-base*, a pretrained language model consisting of 12 layers, 768 hidden, 12 attention heads, and 125 million parameters.

4.4 System Description

For our system, we fine-tune the pretrained models discussed in Section 4.1, 4.2, and 4.3. We combine the three datasets as the number of samples for Spanish and Portuguese are quite low. After combining the models, we split the validation set accordingly, maintaining the split's ratio and tabulating the results on the concatenated dataset in Table 3. The embeddings are extracted from

these models to be fed as input to the LSTM layer, (Hochreiter and Schmidhuber, 1997) as shown in Figure 1. The resulting output is fed into a global average pooling layer (Lin et al., 2014) and then passed into fully connected layers, followed by a sigmoid activation function to obtain the resulting probability score for the input sentences. The same parameters are used for all three models. A dropout layer (Srivastava et al., 2014) is also added in between the fully connected layers for regularization. Refer Table 4 for the parameters used in the model.

Parameters	Values
Number of LSTM units	128
Dropout Rate	0.2
Batch Size	16
Max Length	128
Optimizer	Adam
Learning Rate	3e-5
Activation Function	Sigmoid
Loss Function	cross-entropy

Table 4: Parameters used for training the Models

5 Results and Analysis

All pretrained language models are fine-tuned in Google Colab[2] for ten epochs. We use the Tensorflow implementation of the models[3] on the Huggingface transformers library[4]. We compare the macro F1-Scores of our fine-tuned models on the validation set, which were created by splitting the given dataset. The remaining split is the training data. The validation set contains samples from all three languages. It has 4,387 *Not-event* sequences and 963 *Event* sequences making a combined total of 5,350. The results are shown in Table 3.

We fine-tuned BERT, DistilBERT, and RoBERTa models on the training set. We have combined the

---

[2]https://colab.research.google.com/
[3]https://huggingface.co/transformers/pretrained_models.html
[4]https://huggingface.co/

Language	Macro F1-Score
English	0.8291
Spanish	0.7578
Portuguese	0.7951

Table 5: Macro F1-Scores on the Test Set

three language corpora into a single corpus comprising of all the three languages together. The main intention towards using a multilingual model is that the representations learnt during one language's pretraining would help the other. We can observe that DistilBERT achieved a better F1-Score among the models mentioned in the previous sections. RoBERTa gave the lowest score among these. The reason could be that the RoBERTa model was not multilingual, unlike the other two; however, it still managed to get a score very close to the BERT model. It is imperative that performing soft voting on all three models has managed to increase the score. One of the reasons for the poor performance of the models is the imbalance in the distribution of the classes. In the dataset, there are 21,794 *Not-event* sentences and only 4,954 *Event* ones. The models performed very well in the majority class and poorly in the minority class. Having more *Event* samples could have certainly helped the model in distinguishing better among the classes. Based on the performance of soft voting on the validation set, we have used the same for the test set. The results for the test set are shown in Table 5. The reason for relatively low scores of Spanish and Portuguese could be due to the inadequate support of the training set (2,741 and 1,182) instead of English (22,825). We also believe that our approach of combining datasets could have influenced the performance of the low support datasets.

6 Conclusion

The need to develop automated systems to detect any event is an active protest has constantly been increasing because of the escalation of social media users and several platforms to support them. In this paper, we have explored several multilingual language models to classify if a given sentence talks about an event that has happened (*Event*) or not (*Not-event*) in three languages. Our work primarily focuses on fine-tuning language models and feeding them to an architecture we created. We also observe that the problem of class imbalance has had a significant impact on the performance of the

models. The soft voting approach has achieved macro F1-Scores of 0.8291, 0.7578, and 0.7951 for English, Spanish, and Portuguese, respectively. For future work, we intend to explore class weighting techniques and semi-supervised approaches to improve our performance.

References

Ghada Abaido. 2020. Cyberbullying on social media platforms among university students in the united arab emirates. *International journal of adolescence and youth*, 25:407–420.

Angelo Basile and Tommaso Caselli. 2020. Protest event detection: When task-specific models outperform an event-driven method. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 97–111, Cham. Springer International Publishing.

Michael Beyeler. 2017. *Machine Learning for OpenCV*. Packt Publishing Ltd.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Andreu Casero-Ripollés and Ramón Feenstra. 2012. The 15-m movement and the new media: A case study of how new themes were introduced into spanish political discourse. *Media International Australia incorporating Culture and Policy*, 144:68.

David Crystal. 2008. Two thousand million? *English Today*, 24(1):3–6.

Vera Danilova. 2015. A pipeline for multilingual protest event selection and annotation. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 309–313.

Vera Danilova, Svetlana Popova, and Mikhail Alexandrov. 2016. Multilingual protest event data collection with gate. In *Natural Language Processing and Information Systems*, pages 115–126, Cham. Springer International Publishing.

Dhaval M Dave, Andrew I Friedson, Kyutaro Matsuzawa, Joseph J Sabia, and Samuel Safford. 2020. Black lives matter protests and risk avoidance: The case of civil unrest during a pandemic. Working Paper 27408, National Bureau of Economic Research.

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.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. 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 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Rodrigo Gómez García and Emiliano Treré. 2014. The# yosoy132 movement and the struggle for media democratization in mexico. Convergence, 20(4):496–510.

Deborah J Gerner, Philip A Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. 2002. The creation of cameo (conflict and mediation event observations): An event data framework for a post cold war world. In annual meeting of the American Political Science Association, volume 29.

Adeep Hande, Karthik Puranik, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Domain identification of scientific articles using transfer learning and ensembles. In Trends and Applications in Knowledge Discovery and Data Mining, pages 88–97, Cham. Springer International Publishing.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735–1780.

Kristoffer Holt, Adam Shehata, Jesper Strömbäck, and Elisabet Ljungberg. 2013. Age and the effects of news media attention and social media use on political interest and participation: Do social media function as leveller? European Journal of Communication, 28(1):19–34.

Ahmad Hany Hossny, Terry Moschuo, Grant Osborne, Lewis Mitchell, and Nick Lothian. 2018. Enhancing keyword correlation for event detection in social networks using svd and k-means: Twitter case study. Social Network Analysis and Mining, 8(1):1–10.

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, 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), online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, and Osman Mutlu. 2019a. A task set proposal for automatic protest information collection across multiple countries. In Advances in Information Retrieval, pages 316–323, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019b. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In Experimental IR Meets Multilinguality, Multimodality, and Interaction, pages 425–432, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. Data Intelligence, pages 1–28.

Kawal Kapoor, Kuttimani Tamilmani, Nripendra Rana, Pushp Patil, Yogesh Dwivedi, and Sridhar Nerur. 2018. Advances in social media research: Past, present and future. Information Systems Frontiers, 20.

Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in network.

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.

Konstantina Papanikolaou and Haris Papageorgiou. 2020. Protest event analysis: A longitudinal analysis for Greece. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pages 57–62, Marseille, France. European Language Resources Association (ELRA).

Yifan Peng, Manabu Torii, Cathy H. Wu, and K. Vijay-Shanker. 2013. A generalizable nlp framework for fast development of pattern-based biomedical relation extraction systems. BMC Bioinformatics, 15.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Cristina M. Pulido, Gisela Redondo-Sama, Teresa Sordé-Martí, and Ramon Flecha. 2018. Social impact in social media: A new method to evaluate the social impact of research. *PLOS ONE*, 13(8):1–20.

Adam Rome. 2010. The genius of earth day. *Environmental History*, 15(2):194–205.

Kristin Ross. 2008. *May'68 and its Afterlives*. University of Chicago Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Philip A Schrodt and Blake Hall. 2006. Twenty years of the kansas event data system project. *The political methodologist*, 14(1):2–8.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

S. Wilson. 2017. Detecting mass protest through social media. *Social media and society*, 6:5–25.

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

Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017b. Protest activity detection and perceived violence estimation from social media images. *CoRR*, abs/1709.06204.