
# Cross-modal Transfer Between Vision and Language for Protest Detection

# Cross-modal Transfer Between Vision and Language for Protest Detection

**Ria Raj** [*1,2] **Kajsa Andreasson** [*1,2] **Tobias Norlund**[1,2]
**Richard Johansson**[1,3] **Aron Lagerberg**[2]
[1]Department of Computer Science and Engineering, Chalmers University of Technology,
[2]Recorded Future, [3]University of Gothenburg
{firstname.lastname}@recordedfuture.com,
richard.johansson@gu.se

## Abstract

Most of today's systems for socio-political event detection are text-based, while an increasing amount of information published on the web is multi-modal. We seek to bridge this gap by proposing a method that utilizes existing annotated unimodal data to perform event detection in another data modality, zero-shot. Specifically, we focus on protest detection in text and images, and show that a pretrained vision-and-language alignment model (CLIP) can be leveraged towards this end. In particular, our results suggest that annotated protest *text* data can act supplementarily for detecting protests in images, but significant transfer is demonstrated in the opposite direction as well.

## 1 Introduction

Information published on the web, and in particular social media, has become a crucial source for understanding the world and how it develops. Systems for the automatic detection and extraction of socio-political events are an important tool for processing this stream of information at scale. Traditionally, these systems are primarily designed to process information in the form of text, but with the growing use of multimedia content (such as images and video) on the web and social media especially, there is a great potential for extending the analysis to additional data modalities as well (Joo and Steinert-Threlkeld, 2018). A question is however how this can be done in the most efficient manner, and whether existing data in one modality can be reused for extending analysis to another.

In this work, we take a focused look at the task of *protest detection*, and investigate whether data from different modalities can act both *supplementarily* as well as *complementarily* for this task. We do so by seeking to answer the following research questions:

RQ1 To which extent can the performance of a unimodal protest detection model transfer from one modality to another?

RQ2 Can unimodal detection of protests be improved by using a multi-modal protest detection model?

Considering the natural way text and images complement each other, the hypothesis is that a multi-modal model trained on both text and images would have a broader understanding of the concept of protests.

The investigation has been carried out by combining two existing open datasets for protest event detection, namely the textual CLEF 2019 Protest News dataset (Hürriyetoğlu et al., 2019) and the UCLA Protest Image dataset (Won et al., 2017).

Our contributions are:

1. We propose a modality-agnostic setup for socio-political event detection, where annotated data in one modality can be leveraged to detect the same event in another modality.
2. We demonstrate significant zero-shot protest detection performance when applying a model on a modality not observed during training.
3. Whereas we show protest text and image data to act supplementarily, our results do not support the hypothesis that the data can act complementarily to the same degree.

## 2 Datasets

We took use of two open-source datasets: the UCLA Protest Image dataset (Won et al., 2017) for images and CLEF 2019 Protest News (Hürriyetoğlu et al., 2019) for texts. The UCLA dataset consists of a training set of 32,612 images and a test set of 8,154. In our experiments we only consider the binary protest/not protest prediction task. Meanwhile, for the Protest News dataset we only

[*]Equal contribution.



*Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 56 - 60

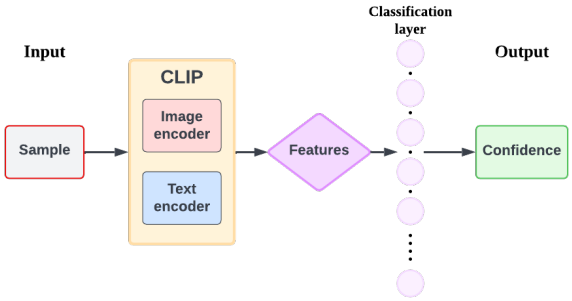


Figure 1: Our model setup. A sample is fed through its respective CLIP encoder and the resulting feature vector is fed through a classification layer that outputs a binary prediction score.

consider the binary sentence-level classification task in English. It comprises 22,825 sentences in total, retrieved from news articles, which was split into a training and test set with a 75-25 ratio.

## 3 Experiments

We begin by exploring RQ1, that is, whether text and image representations can be interchanged in the task of protest detection. In practice, this would mean that a classifier trained on protest *images* is tested on protest *texts*, and vice versa. This is made possible by using a pretrained encoder that is able to represent both modalities in a common feature space. We denote such experiments as *cross-modal*, where training and test data are of different modalities, meaning zero-shot classification. The extent to which this capability can be transferred between modalities can then be evaluated by comparing to a *unimodal* baseline, which essentially means we train and test on the same modality. To get a lower bound we also compare against a random classifier baseline. In all experiments we evaluate using the AUC-PR metric.

To adress RQ2, we consider the case where training data for both text and images are available and investigate whether these datasets can synergistically complement each other. Specifically, we explore training a model jointly on the Protest News and UCLA datasets, but evaluate on each modality separately, similarly to the above experiments. We denote this experiment *multi-modal*, because it is trained on both modalities. This experiment is implemented by combining sentences and images in each training batch. To make use of all the image data, each batch contained 65% images and the remaining 35% sentences. These results can then be also compared to the unimodal baselines explained above.

| | Model | | | |
|---|---|---|---|---|
| **Test set** | IM | TXT | MM | Random |
| Image | **0.962** | 0.687 | 0.957 | 0.290 |
| Text | 0.458 | **0.734** | 0.707 | 0.187 |

Table 1: AUC-PR scores for the models trained on different regimes, when testing over the different modalities.

## 4 Model

CLIP (Radford et al., 2021) was used to generate feature representations of each text and image in the datasets. CLIP is a pretrained visual-and-language model that has been trained to align text and images in a common feature space where samples containing similar textual or visual concepts are pulled together while nonsimilar concepts are pushed apart. Models like CLIP are suitable for the investigation in this work since it should create similar feature representations of protests regardless of the modality. While little information is provided about the pretraining data of CLIP, we hypothesize news, including protests, to be represented to some extent. In such case, the representations of CLIP should be somewhat aligned for this type of data. The features generated by CLIP were used as input to a linear classification layer, which was trained to classify text or image samples as protest or non-protest. This is visualized in Figure 1. We train only the linear classification layer weights, and keep the pretrained CLIP weights frozen during training. This is to be able to swap the encoder at test time in the cross-modal experiments.

The training was done on three different datasets, as described in Section 2, resulting in three trained classifiers: one trained on the pure image dataset (henceforth refered to as IM), one on the pure text dataset (henceforth refered to as TXT) and a third trained jointly on both, i.e. the multi-modal mixed dataset (henceforth refered to as MM).

For IM, the learning rate (LR) was set to $0.01$, and for both TXT and MM it was set to $0.001$. The LR-scheduler for IM and TXT was a lambda decay, with $\lambda = 0.95$, and none for MM. In addition to these hyperparameters, the Adam optimizer and batch size of 128 were used for all three models.

(a) IM model score: 0.99
TXT model score: 0.91
MM model score: 0.98
Label: protest


(b) IM model score: 0.99
TXT model score: 0.21
MM model score: 0.92
Label: protest

(c) IM model score: 0.097
TXT model score: 0.96
MM model score: 0.43
Label: protest

(d) IM model score: 0.0052
TXT model score: 0.51
MM model score: 0.097
Label: not protest

Figure 2: Three randomly chosen positive examples and one negative from the UCLA test data along with the three models' prediction scores. Subfigure 2a shows an example when the scores of the IM and TXT models coincide and Subfigure 2b when the IM model scores high, but the TXT model does not. Subfigure 2c shows an example in which the TXT model scores high and the IM model does not and Subfigure 2d shows a negative (i.e. not protest) example.

## 5 Results and Discussion

As seen from Table 1, the image and text baselines both perform better than their cross-modal counterparts, where the models are tested on the opposite modality than they are trained on. However, the cross-modal performance is significantly better than the random baseline, which indicates that the protest detection ability indeed can be transferred between modalities to some extent. One interesting result is that the TXT model performs almost equally well on images compared to when testing on text. This indicates that the TXT model's understanding of protests can almost fully be transferred to images, since the performance between the modalities only differs by a score of $\sim 0.05$. In contrast, the performance of the IM model decreases by more than half when testing on text compared to images. Considering these two outcomes, it seems reasonable to conclude that training a classifier on texts provides a more general understanding of protests, which can be transferred to images, while training on images gives the model a way of interpreting protests that cannot be found to the same extent in the texts used for testing.

When comparing the unimodal baselines, it is clear that the IM model performs much better with an AUC-PR score of 0.962 compared to the TXT baseline of 0.734. This could be a consequence of the image data being more homogenous in terms of how they represent protests. This also follows the reasoning above: that the texts contain a wider range of representations of protests.

An aspect that would be interesting to further investigate is which characteristics in the data that are significant for the separate models when classifying protests, by carrying out an even more thorough data analysis. When inspecting some samples that the IM model scores high on, see Figure 2, many of them contains concepts such as banners and placards, full-body humans, roads and cities as well as buildings. As for the text-model, some words that often occured in samples that recieved high scores include protest, traffic, roads, bodies of power (ie. government, police), bomb, students, injured, crowd. Neither the list of visual concepts or words are exhaustive, but they could give an

| **Fragment** | $P_{\text{IM}}$ | $P_{\text{MM}}$ | $P_{\text{TXT}}$ | Label |
|---|---|---|---|---|
| "Taxi operators marching in protest against the government's taxi recapitalisation scheme reached the Union Buildings in Pretoria on Friday." | 0.87 | 0.82 | 0.93 | protest |
| "(SUBS: Pics will be available later on www.sapapics.co.za) South African rape laws still blame the survivor of rape, People Opposing Woman Abuse (Powa) said on Friday at a protest outside the Johannesburg High Court." | 0.85 | 0.77 | 0.63 | protest |
| "Workers at the company's Zondereinde mine, near Amandelbult in Limpopo, went on strike on November 3." | 0.30 | 0.67 | 0.87 | protest |

Table 2: Three randomly chosen positive examples from the Protest News test data. The first row shows an example for which both IM and TXT give high scores of it being a protest. The second row shows an example that recieves high scores from IM, but not from TXT. The last row shows an example that recieves a high score from TXT, but not from IM.

indication of what the models recognize when identifying protests. These lists also show that different aspects of protests are captured in the data due to the nature of the modalities and the sources of the data, which could be an aspect that affects the performance. Figure 2d shows one image with a negative label that recieves low scores from the IM model, despite the fact that it pictures a crowd. It does however lack placards and several of the other characteristics that are described above as possible factors that trigger the IM model to give high scores. The TXT model on the other hand, gives an intermediate score which indicates that the model is inferior at distinguishing between casual, friendly crowds and protest related crowds in images. This behaviour is seen for multiple similar samples that aren't displayed here.

When it comes to the case of the multi-modal (MM) model we see two things. Firstly, when comparing performance to the unimodal baselines it is clear that the MM model performs slightly worse in both cases. When testing on images, the difference in performance is $\sim$0.013, whereas for text $\sim$0.034. In contrast to our hypothesis, this indicates that training on both modalities does *not* provide the model with a broader understanding of the concept of protests, and consequently the performance on unimodal test sets is not improved. We speculate this is partly due to the fact that texts and images come from different source types (mainstream news vs social media), whereas CLIP has been trained to align text and image pairs from the *same source*. There is a possibility that the results would be different if the data used for testing was collected from a wider range of sources than the data used for training, since new data sources may represent protest in slightly different ways.

What one does see however, is that when comparing the MM model to both the IM and TXT models in the cross-modal set up, the MM model performs noticeably better. This is an indication that the multi-modal model in fact learns a representation of protests that succesfully incorporates information from both modalities.

## 6 Related work

Our work is a contribution to the field of event detection, that is, identifying mentions of whether a certain event has occurred. Early data-driven approaches to this task based on machine learning relied heavily on hand-designed lexical and syntactic features e.g. (Li et al., 2013; Patwardhan and Riloff, 2007). However, since then approaches based on deep learning indicate better performance can be achieved using less feature engineering by training on "raw" (textual) data (Nguyen et al., 2016; Chen et al., 2015; Boros, 2018). Specifically for the extraction and detection of socio-political events such as protests, some recent works have taken a pure visual approach. For example, Joo and Steinert-Threlkeld (2018) demonstrated that a visual analysis can contribute protest related features that might be harder to extract from pure text, such as violence, crowd size as well as demographic composition. Won et al. (2017) further investigated the ability to extract protest related information from images, where the UCLA Protest Image dataset is presented along with experiments for the detection of protests and related attributes.

Previous works taking a multi-modal approach to socio-political event detection also exist. Petkos et al. (2012) used a clustering method of textual as well as visual features to discover events in social media data. Qian et al. (2015) proposed a boosted multi-modal extension to LDA for training a supervised event classification model. More recently, Zhang and Pan (2019) take a deep learning approach to the detection of collective action events based on text and potentially an image from social media posts in China. Similarly to CLIP, they use a late-fusion dual encoder for the processing of text and image modalities.

Our work differs in that we investigate using non-parallel data, e.g. where protest texts and images are labeled and classified individually. We also differ in that we use data from different sources (images from social media and text from mainstream news), as well as using state-of-the-art pretrained visual-and-language representations.

## 7 Conclusions

From the results and discussion carried out, we can conclude that the performance of a unimodal protest detection model trained on text can transfer almost fully to do zero-shot classification of protests in images. This means that a protest classifier trained on texts can be used directly on images without any further training or fine-tuning involved, and without significant decrease in performance. The benefit of this would naturally be that an image protest classifier can be put in use without the need of annotating any image data. On the other

hand, we observe that the transfer from images to text implies a loss of performance while it is still significant compared to the random baseline. Furthermore, the investigation shows that multi-modal training for protest detection can be used almost interchangeably to a unimodaly trained model, as performance does not differ substantially.

## Ethical statement

Socio-political analysis is important for understanding society at large, and to be able to report on how it develops. It is however of utmost importance that the development of tools and methods is performed with ethical considerations in mind. For example, risks include misuse for large scale surveillance by authoritarian regimes as well as discriminatory performance against minorities due to hidden system biases.

The underlying data used for training protest detection models will inevitably contain spurious correlations that the model might learn to base a protest/not protest decision on. For text based detection, this could be names of organizations, geographical locations or other entities prominent in protests occuring when the data was collected. For image based detection, visual traits such as the etnicity of individual protestors might also be a source of bias.

While these aspects of the model were not explicitly addressed by our research questions in this work, they are important to investigate further as a prerequisite for application of these systems.

## Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

## References

Emanuela Boros. 2018. *Neural methods for event extraction*. Ph.D. thesis, Université Paris Saclay (COmUE).

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.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019. 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.

Jungseock Joo and Zachary C. Steinert-Threlkeld. 2018. Image as data: Automated visual content analysis for political science.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 717–727.

Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, New York, NY, USA. Association for Computing Machinery.

Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and M. Shamim Hossain. 2015. Social event classification via boosted multimodal supervised latent dirichlet allocation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(2).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images.

Han Zhang and Jennifer Pan. 2019. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.