ECO: Using AI for Everyday Armed Conflict Analysis

Anusua Trivedi[1*], **Kate Keator**[2], **Avirishu Verma**[1], **Rahul Dodhia**[1], **Juan Lavista Ferres**[1]
[1]Microsoft
[2]The Carter Center

Abstract

Conflict resolution practitioners consistently struggle with access to structured armed conflict data, a dataset already rife with uncertainty, inconsistency, and politicization. Due to the lack of a standardized approach to collating conflict data, publicly available armed conflict datasets often require manipulation depending upon the needs of end users. Transformation of armed conflict data tends to be a manual, time consuming task that nonprofits with limited budgets struggle to keep up with. In this paper, we explore the use of a deep natural language processing (NLP) model to aid the transformation of armed conflict data for conflict analysis. Our model drastically reduces the time spent on manual data transformations and improves armed conflict event classification by identifying multiple incidence types. This minimizes the human supervision cost and allows nonprofits to access a broader range of conflict data sources to reduce reporting bias. Thus our model contributes to the incorporation of technology in the peace building and conflict resolution sector.

1 Introduction

Just as information and communication technologies (ICTs) have changed the way people interact with each other (online chat rooms, Reddit, Facebook, Twitter etc.), ICTs have also changed how conflict resolution actors conduct conflict analysis and stakeholder mapping [Tufekci and Wilson, 2012]. Organizations such as the Carter Center and the Armed Conflict Location and Event Data (ACLED) use publicly available information shared by social media and online news sources to document conflict activity in support of conflict resolution actors, humanitarians, and researchers. Yet the classification of conflict data is highly contextual, let alone political, with organizations classifying conflict data according to their own needs. In addition, conflict data collection is often a largely manual task due to the ubiquitous terms defining conflict. Analysts often must make subjective decisions such as what constitutes an armed clash versus mutual shelling, or what differentiates an insurgent from a rebel, or which violent and non-violent incidences to collect. This combined with the deluge of information can make the conflict data collection difficult to keep up with. Due to this information overflow, many practitioners and academics working with conflict data often turn to resources such as the Armed Conflict Location and Event Data (ACLED) Project [Raleigh *et al.*, 2010] as a data source due to its public availability and structured datasets [ACLED, 2019], even though its terminology or collection methodology may not align with a team's particular needs. Consequently, some teams spend hours manually transforming conflict data to fit their needs, thereby reducing time spent on analysis and other important tasks. According to research from the Peace Research Institute Oslo [Dupuy and Rustad, 2018], interstate conflict has gradually declined in the post-Cold War era while intrastate conflict is on the rise, many of which involve contribution of troops from external states. These internationalized conflicts are on average more violent, more difficult to solve, and tend to last longer as exemplified by the ongoing Syrian conflict. Such realities require conflict resolution practitioners to find ways to handle the inevitable data overload and complexity if they are to leverage the opportunities of our digital world.

*Contact Author

2 Data Predicament and AI

Since 2013 the Syria Conflict Mapping Project (SCMP) at the Carter Center has used publicly available, open source information to conduct detailed analysis on the Syrian conflict, in support of conflict resolution and humanitarian actors. This includes a database of 122,000 reported incidences of violence throughout Syria collected between 2013 and 2019 from YouTube, Twitter, Facebook, activist reporting websites, and local partner organizations. The Center was logging on average 400 events per week in 2015. Unable to maintain the cost of manually collecting and organizing conflict data for a high-volume area such as Syria, the Center began using ACLED's dataset in 2018 even though ACLED's data structure does not seamlessly mesh with its pre-existing conflict database.

Manual transformation of ACLED's Syria-related data, released on a weekly basis, takes on average 4 hours per week due to the contextual nature of the data . An individual must be knowledgeable of the actors in the Syrian conflict, familiar with ACLED's collection methodology and resulting data structure, and well-versed in the Carter Center's conflict incident classification in order to understand what needs to be changed and how. As these parameters are replicated in other organizations across a variety of contexts, artificial intelligence, particularly supervised machine learning, has caught the attention of organizations working in the peace and security field. While natural language processing (NLP) is commonly used in marketing and advertising (such as crimson hexagon [Hitlin, 2015]), training NLP models to work with conflict data, especially information translated from one language to another, proves to be much more difficult. In NLP, domain adaptation has traditionally been an important topic for syntactic parsing [McClosky *et al.*, 2010] and named entity recognition [Chiticariu *et al.*, 2010] among others. With the popularity of distributed representation, pre-trained word embedding models such as word2vec [Mikolov *et al.*, 2013] and glove [Pennington *et al.*, 2014] are also widely used for natural language tasks. For human beings, text comprehension is a basic task, performed daily. As early as in elementary school, we can read an article, and conclude about its key ideas and details. But for AI, full text comprehension is still an elusive goal. Therefore, building machines that can perform text comprehension is of great interest. In this paper, we will focus on application of Deep Bidirectional Transformers for Language Understanding (BERT) [Devlin *et al.*, 2018] to the problem of multi-label text classification.

Conflict resolution actors, conflict analysts, and others in the humanitarian field are unable to overcome the intensive time required to use publicly available conflict data. By automating this event classification process, we can significantly reduce time spent on data preparation . Thus, conflict researchers would now be able to spend more time understanding the trends and key insights needed to support their peace building and humanitarian efforts.

3 Data for this work

The first version of the model is trained on Syria conflict data from ACLED between 2018 and 2019, totaling 7,847 events. Since a single conflict event can contain multiple incident types, authors manually created a new instance for each incident type while maintaining the unique ID of each conflict event. This clearly identified the key words used to describe each incident type. This manual work resulted in 8,942 conflict events detailing 11 different incident types, from which we used three (shelling, clashes, and strategic development) to train the NLP model.

4 Method

Text Classification is a long-standing challenge in NLP, and the community has introduced several paradigms and datasets for the task over the past few years. These paradigms differ from each other in the type of text and labels, and the size of the training data, from a few hundreds to millions of examples. In this article, we are particularly interested in the context-aware text classification paradigm, where the label for each text snippet can be obtained by referring to its accompanying context (paragraph or a list of sentences). Research in the field of using pre-trained models have resulted

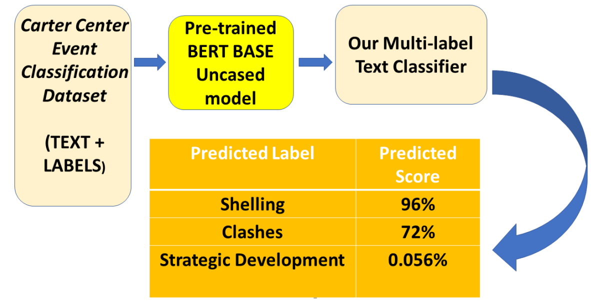

Figure 1: BERT Multi-Label Classification for Carter Center Event Classification

in massive leap in state-of-the-art results for text classification. Some of the key milestones have been ELMo [Peters *et al.*, 2018], ULMFiT [Howard and Ruder, 2018] and OpenAI Transformer [Radford *et al.*, 2018]. All these approaches allow us to pre-train an unsupervised language model on large corpus of data such as all Wikipedia articles, and then fine-tune these pre-trained models on downstream tasks. However, the release of BERT [Devlin *et al.*, 2018], a multilingual transformer-based model, has achieved state-of-the-art results, outperforming all the other models.

4.1 Fine-tuning BERT for Multilabel Armed Conflict Event Classification

Modern machine learning models, especially deep neural networks, often significantly benefit from transfer learning. In computer vision, deep convolutional neural network (CNN) [Krizhevsky *et al.*, 2012] trained on a large image classification dataset such as ImageNet [Deng *et al.*, 2009] have proved to be useful for initializing models on other vision tasks [Yosinski *et al.*, 2014]. Researchers have shown the value of transfer learning using the trained neural network as the basis of a new purpose-specific model. In recent years, researchers have been showing that a similar technique can be useful in many natural language tasks. We use the smaller BERT Base, uncased model as the base model for this paper. The BERT Base model has 12 attention layers and uses the word-piece-tokenizer [Wu *et al.*, 2016], which converts all text to lowercase. We modify the BertForSequenceClassification class in BERT GitHub [Hugging-Face, 2019] for multi-label classification. We use binary-cross-entropy-with-logits [Gomez, 2018] as loss function for multilabel classification task instead of the standard cross-entropy loss used by BERT model. Binary cross-entropy loss allows our model to assign independent probabilities to the labels. Figure 1 explains our multilabel classification pipeline. The training loop is identical to the one provided in run_classifier.py in [Hugging-Face, 2019]. We train the model for 4 epochs with batch size of 16 and sequence length as 256. The learning rate is kept to 3e-5, as recommended for fine-tuning in the original BERT paper. We do not use the precision FP16 technique as binary-cross-entropy-with-logits loss function does not support FP16 processing.

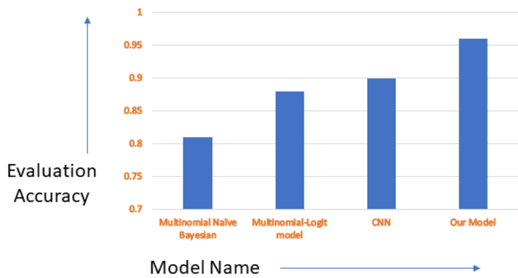

Figure 2: Comparing evaluation accuracy of prior work models and our model

5 Evaluation of our Model

Out of the 8,942 conflict events, we used 80% for training and 20% for evaluating the model. The evaluation comprised of 1788 events from the dataset and had an accuracy of 96% accuracy (roc_auc_micro = 0.97). Figure 2 compares evaluation accuracy of all prior work models vs. our model.

5.1 Out-of-sample Evaluation of our Model

The out-of-sample data consisted of 57,253 Syria conflict events from ACLED between 2017 and 2019, thereby including 31,393 conflict events from 2017 completely unseen by our model. Despite the variety of incident types in the test data set, our model had an accuracy rate of 90%. Experts manually reviewed the results of the out-of-sample data. For out of sample evaluation, experts manually divided our results into four thresholds:

- incidences given a probability higher than 85% of being the incident type identified by ACLED
- incidences given probabilities between 75% and 85%
- incidences less than 75%
- incidences given a probability higher than 85% for an incident type not identified by ACLED.

These thresholds were used to group results by incident type (as shown in Figure 3) for a human analyst to review. Figure 3(a),(b),(c) shows the breakdown of our model's classification of out-of-sample 57,000 conflict events based on the three incident types it was trained on: shelling, clashes, and strategic developments. For example, in Figure 3(a), the largest grouping of results for the incident types shelling was those identified as shelling by ACLED and given an 85% probability by our model as being shelling. Figure 3(d),(e),(f) shows the breakdown of our model's classification on 2017 conflict events only, based on the three incident types it was trained on: shelling, clashes, and strategic developments. For example, in Figure 3(d), the largest grouping of results for the incident types shelling was those identified as shelling by ACLED and given an 85% probability by our model as being shelling.

6 Discussion

Results for an expected outcome were reviewed for 1000 samples, while abnormal results were reviewed more thoroughly by expert researchers for evaluation. Overall, our model identified conflict incidences that should have more than one incident type. For example, our model gave 270 conflict events identified by ACLED as clashes a probability of 75% or less of being clashes. Rather than this being an inaccuracy, our model accurately identified that 105 of these events were also shelling. A consistent trend in conflict events in Syria are that clashes and shelling often accompany each other, a trend that Carter Center had started manually documenting by inputting multiple incident types into one conflict event. Conflict activity composed of other incident types that was not trained in the model, such as air/drone attack, were often classified as shelling or clashes. For example, of the 4,782 incidents given an 85% probability or higher by our model as being shelling, despite not identified as shelling by ACLED, 4,388 of these were air/drone attacks according to ACLED's typology. In searching through these, the authors found that 1,374 of incidents contain the word "shell" pointing to the likelihood that many of the conflict incidents are multi-incident types. This aligns with the trends of the conflict as air/drone attacks are often accompanied by on the ground shelling as well as aerial shelling. Often language used to describe an air/drone attack includes words such as "targeted" and "artillery", which is like descriptions of conflict activity involving shelling. Our model has successfully been able to identify the above trends from the context hidden in text data. Thus, we can conclude that use of this model would be able to save Carter Center many hours of manual data preprocessing task. Sample code and data for this work is provided in footnote below[1]

Open source conflict reporting is often not structured in a way to facilitate automation, making visualization and analysis of conflict data a time intensive process. While ACLED is among the leaders of publicly available, structured conflict data, the contextual nature of this type of information means there is no one size that fits all. Using automation to speed up the manual transformation of conflict data gives the nonprofit practitioners more time to conduct the analysis essential to their work and access an array of conflict datasets, which can lower reporting bias through a diversification of data sources. Lastly, and more broadly, our model contributes to the gradual trend of integrating technology and the peace building and conflict resolution sector as practitioners recognize the potential impact of the digital age on their work.

[1] https://tinyurl.com/y5cs65dx

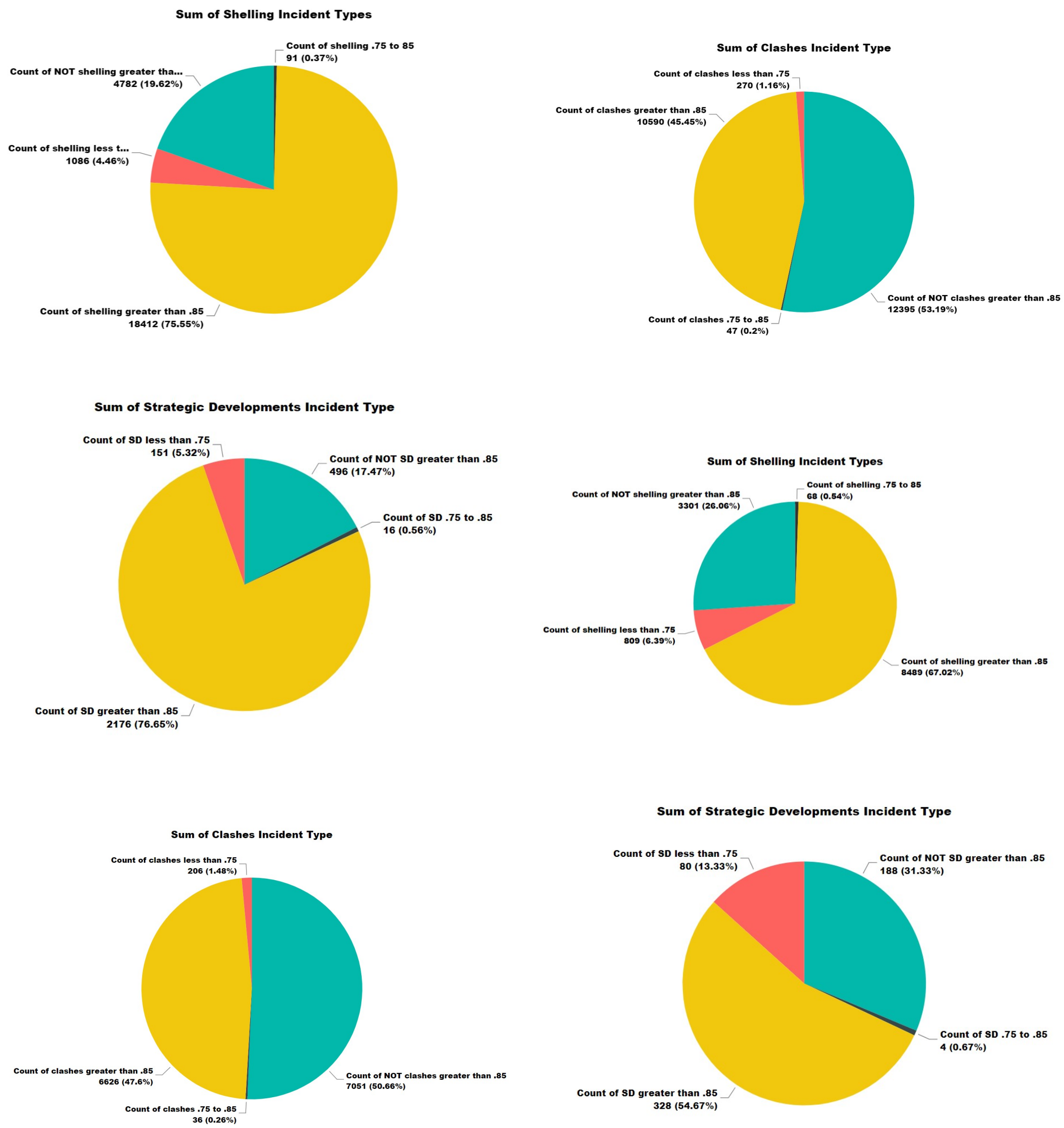

Figure 3: Evaluation of our model: Incident Types Breakdown

References

[ACLED, 2019] ACLED. https://www.acleddata.com/data/, 2019.

[Chiticariu *et al.*, 2010] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1002–1012, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805, Oct 2018.

[Dupuy and Rustad, 2018] K. Dupuy and S.A. Rustad. Trends in Armed Conflict. 2018.

[Gomez, 2018] Raul Gomez. Understanding categorical cross-entropy loss, binary cross-entropy loss, softmax loss, logistic loss, focal loss and all those confusing names. https://gombru.github.io/2018/05/23/cr, 2018.

[Hitlin, 2015] Paul Hitlin. Methodology: How crimson hexagon works, 2015.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text

Classification. *arXiv e-prints*, page arXiv:1801.06146, Jan 2018.

[Hugging-Face, 2019] Hugging-Face. Bert pytorch https://github.com/huggingface/pytorch-transformers, 2019.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[McClosky *et al.*, 2010] David McClosky, Eugene Charniak, and Mark Johnson. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California, June 2010. Association for Computational Linguistics.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, page arXiv:1301.3781, Jan 2013.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543, 2014.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv e-prints*, page arXiv:1802.05365, Feb 2018.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Forthcoming*, 2018.

[Raleigh *et al.*, 2010] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47:651–660, 09 2010.

[Tufekci and Wilson, 2012] Z. Tufekci and C. Wilson. Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. 2012.

[Wu *et al.*, 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv e-prints*, page arXiv:1609.08144, Sep 2016.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 28th Conference on Neural Information Processing Systems*, 2014.