Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction

Sonja Häffner[1], **Martin Hofer**[1], **Maximilian Nagl**[2] and **Julian Walterskirchen**[1]

[1] *Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung), Universität der Bundeswehr München, Neubiberg, Germany. E-mail: julian.walterskirchen@unibw.de*
[2] *Lehrstuhl für Statistik und Risikomanagement, Universität Regensburg, Regensburg, Germany*

Abstract

Recent advancements in natural language processing (NLP) methods have significantly improved their performance. However, more complex NLP models are more difficult to interpret and computationally expensive. Therefore, we propose an approach to dictionary creation that carefully balances the trade-off between complexity and interpretability. This approach combines a deep neural network architecture with techniques to improve model explainability to automatically build a domain-specific dictionary. As an illustrative use case of our approach, we create an objective dictionary that can infer conflict intensity from text data. We train the neural networks on a corpus of conflict reports and match them with conflict event data. This corpus consists of over 14,000 expert-written International Crisis Group (ICG) CrisisWatch reports between 2003 and 2021. Sensitivity analysis is used to extract the weighted words from the neural network to build the dictionary. In order to evaluate our approach, we compare our results to state-of-the-art deep learning language models, text-scaling methods, as well as standard, nonspecialized, and conflict event dictionary approaches. We are able to show that our approach outperforms other approaches while retaining interpretability.

Keywords: natural language processing, objective dictionaries, deep learning, transformers, conflict dynamics

1 Introduction

Extracting information from text corpora by utilizing natural language processing (NLP) techniques has become widespread in many scientific fields. NLP techniques have significantly improved accordingly, with a move away from more static representations of text, such as dictionaries, to more advanced methods like word embeddings and transformer models. Nonetheless, applying NLP methods and extracting useful information from text sources is not a trivial matter and while approaches have proliferated, these have become increasingly complex and computationally expensive (see, e.g., Sharir, Peleg, and Shoham 2020). Many modern NLP approaches require large amounts of training data, which are not only costly to acquire but also are often proprietary, reducing accessibility and inhibiting efforts to make research easier to reproduce. Furthermore, the computational requirements are often not easily fulfilled. While this is in part due to the large corpora needed,[1] it is also because of the complex and computationally intensive transformations of the text data that NLP techniques need to perform. Hence, access to high-performance computing architecture is necessary to use such approaches.[2] Furthermore, the complexity inherent

1 With large corpora, even less complex methods, such as topic models, can become incredibly computationally costly (see, e.g., Yuan *et al.* 2015).
2 For example, for word embeddings, there are different approaches to deal with the trade-off between precise representation of text data and the computational costs associated (see, e.g., Raunak, Gupta, and Metze 2019; Rodriguez and Spirling 2022; Shu and Nakayama 2017).

Political Analysis (2023)
vol. 31: 481–499
DOI: 10.1017/pan.2023.7

Published
22 March 2023

Corresponding author
Julian Walterskirchen

Edited by
Jeff Gill

in many NLP approaches poses problems to interpretability and transparency. Although this is a common issue that applies to a wide array of modern machine learning approaches, state-of-the-art NLP approaches, like transformer models, are particularly prone to be difficult to interpret (see, e.g., van Aken *et al.* 2019).

Hence, we propose a deep learning approach to create objective dictionaries for domain-specific applications. This approach is designed to alleviate some of the problems associated with modern, complex NLP methods but also improves on traditional and automatic approaches to dictionary creation. Our approach puts a strong emphasis on interpretability, by employing a technique proposed by Horel *et al.* (2018), eliminates most manual labeling and coding costs, and produces data-driven dictionaries. Our approach aims to be computationally cheaper, easier to implement, requiring smaller corpora, and be adaptable to other research fields, even beyond political science, while ensuring transparency and reproducibility. Ultimately, this should enable researchers to create objective dictionaries from their own domain-specific corpora predicting any concept of interest.

As text-as-data approaches proved promising for alleviating data availability concerns, the area of conflict research is particularly appropriate as our test case (see, e.g., Gleditsch 2020). Given the complexities in conflict processes, researchers have long struggled with adequately modeling central aspects, and acquiring fine-grained data continues to prove difficult (de Coning 2020; Weidmann and Ward 2010). Hence, we create a dictionary that can infer conflict intensity from text data. We use Uppsala Conflict Data Program Georeferenced Event Dataset (UCDP GED) monthly fatality numbers (natural logarithm) to measure conflict intensity (Davies, Pettersson, and Öberg 2022; Sundberg and Melander 2013). For our text source, we rely on approximately 14,000 expert-written International Crisis Group (ICG) CrisisWatch reports between 2003 and 2021.

We evaluate and compare the performance of our approach to three widely used NLP approaches. First, its performance is compared to two general purpose dictionaries (Harvard IV-4 sentiment dictionary [Dunphy 1974; Stone, Dunphy, and Smith 1966] and the social media sentiment dictionary by Hutto and Gilbert (2014)) and a conflict event coding dictionary (Norris, Schrodt, and Beieler 2017). Second, we assess its performance against two widely used document scaling techniques: Wordscores (Laver, Benoit, and Garry 2003) and Wordfish (Lo, Proksch, and Slapin 2016). Finally, as transformer architectures are considered the current state-of-the-art technique in NLP (see, e.g., Widmann and Wich 2022), the performance of our dictionary is compared to the performance of the newly released ConfliBERT model (Hu *et al.* 2022). As a measure of performance, we investigate alignment with conflict trends over time and correlations with our variable of interest. In addition, we evaluate how accurately each approach can solve a text regression task.[3]

We find that our approach is well equipped to create a dictionary that adequately measures trends in conflict intensity over time. The results also suggest that our approach is able to consistently outperform the benchmark models in a text regression task while lowering computational costs and costs associated with manually creating a dictionary. These results indicate that our approach can serve as a successful blueprint for future researchers to analyze text data for domain-specific applications inside and outside of conflict research. Hence, the contribution of this paper to the existing research is twofold. The main contribution lies in introducing a novel approach to generating objective dictionaries for domain-specific applications. In addition, it develops a dictionary that accurately infers conflict intensity from text data.[4]

3 A variation of a text classification task, where the outcome variable is continuous rather than binary or ordinal.
4 A web application demonstrating our dictionary can be found at https://kfeapps.github.io/ocodi/.

2 Related Work

2.1 Advancements in Dictionary Creation

While the use of dictionary methods has declined, due to the high costs of manually constructing and maintaining them and the accuracy increase achieved with more complex models, researchers have sought to leverage modern machine learning techniques to reduce these costs and increase their performance. Most of these approaches were intended to improve sentiment analysis. They focus on extending existing dictionaries that are more finely adjusted to their specific task. Jha *et al.* (2018) build a novel sentiment dictionary by training a model on labeled and unlabeled text data from different domains, allowing them to identify sentiment words in the target corpus based on the information learned from the source domain corpus. Sood, Gera, and Kaur (2022) highlight how using different algorithms (Naive Bayes, Stochastic Gradient Descent, Lasso, and Ridge) trained on labeled text documents can be used to build and extend domain-specific dictionaries. Relatedly, Lee, Kim, and Song (2021) build a dictionary using Lasso regression trained on a corpus of hand-labeled product reviews. Carta *et al.* (2020, 2021) build domain-specific dictionaries for stock market forecasting by assigning weights based on a company's performance to words extracted from financial news articles. The dictionaries created based on this approach are then used as features in decision trees to assess the company's future performance. Similarly, Palmer, Roeder, and Muntermann (2021) build a domain-specific dictionary by assigning word polarity through a linear regression linking words and stock returns. De Vries (2022), following an approach introduced by Rheault *et al.* (2016), uses a seed dictionary and a word embedding model to automatically identify additional words and their corresponding tonality. They are able to show considerable improvements of their approach compared with other dictionaries. Similarly, Widmann and Wich (2022) apply a word embedding model and manual coding to extend an existing German-language sentiment dictionary. They compare this dictionary to word embeddings and transformer models, finding that transformer models outperform the other approaches. Li *et al.* (2021) also rely on seed words to build their domain-specific dictionary; however, they combine dictionary-based and corpus-based seed words and use a deep neural network to train a sentiment classifier to assign positive and negative tonality to their word lists. These approaches still rely mostly on existing labeled data, manual creation of seed words, or applying relatively simple weighting methods. Therefore, we propose an approach to automatic dictionary creation that leverages the advantages of deep neural networks in combination with techniques to improve model interpretability.

2.2 Using Text Data in Conflict Research

NLP approaches have long played an important part in conflict research, as some of the first uses of NLP in conflict research were for the purpose of improving data collection efforts. Building on early work for collecting political event data by McClelland (1971) (WEIS) and Azar (1980) (COPDAB), researchers developed dictionaries and rule-based systems to automatically extract events from news articles. The Kansas Event Data System is one such pioneering attempt (Schrodt 2008). It relies on WEIS codes as the basis for its dictionary to extract events in the Middle East, Balkans, and West Africa from English-language news articles. Extending the WEIS framework, the Integrated Data for Events Analysis includes additional, more fine-grained event types and non-state actors (Bond *et al.* 2003). These approaches have consecutively been improved upon, leading to the development of CAMEO (Schrodt *et al.* 2012), TABARI,[5] and PLOVER[6] event dictionaries, which are used in event extraction systems, such as PETRARCH, GDELT, and ICEWS (see, e.g., Leetaru and Schrodt 2013; Norris *et al.* 2017; Shilliday and Lautenschlager 2012). Nonetheless, while these

5 Textual Analysis by Augmented Replacement Instructions, a software developed by Philip Schrodt. For more information, see https://parusanalytics.com/eventdata/software.dir/tabari.html.

6 A new ontology for event data that was intended to supersede CAMEO (https://github.com/openeventdata/PLOVER).

dictionary-based approaches have proved helpful for event extraction, they are heavily reliant on being manually maintained, updated, and extended.[7] Furthermore, although approaches have been suggested to use these dictionaries to classify how conflictual or cooperative relationships are (Goldstein 1992), they were not designed for this purpose.[8]

Researchers have also sought to apply other NLP approaches to increase our understanding of conflict processes. Chadefaux (2014) was able to apply NLP techniques to a large collection of news articles which helped increase the accuracy of interstate and intrastate war prediction. Mueller and Rauh (2018, 2022a,b) were able to show that including features extracted from newspaper articles through a topic modeling approach can support conflict prediction models. Relatedly, Boussalis *et al.* (2022) apply a topic modeling approach to French diplomatic cables to predict French interstate conflicts. There have also been attempts to leverage sentiment analysis. Watanabe (2020), for example, is able to classify newspaper articles for their polarity with regard to the state of the economy with a semi-supervised Latent Semantic Scaling approach. Trubowitz and Watanabe (2021) employed a similar approach. The authors were able to automatically identify how adversarial or friendly the relationship between the United States and other countries is based on *New York Times* news summaries. Greene and Lucas (2020) applied a standard sentiment dictionary in order to shed light on non-state armed group relationships. They are successful in identifying rivalries and alliances between Hezbollah and other non-state armed groups based on Hezbollah-produced and disseminated documents. Also, focusing on non-state armed groups, Macnair and Frank (2018) analyze the tonality of ISIS propaganda magazines to identify changes in rhetoric, including also the level of hostile language toward other non-state armed groups.

All these works have provided invaluable contributions and advanced the study of text as data for conflict processes. Nonetheless, an easy approach that allows researchers to create their own NLP tool, tailored to their requirements, is missing. Consequently, we propose an approach to create domain-specific dictionaries which seek to improve upon these existing approaches. Section 3 will describe the intuition and idea behind this in more detail.

3 An Advanced Approach to Creating Objective Dictionaries

We present an approach that allows future researchers to build domain-specific dictionaries for their respective areas of interest. To illustrate the usefulness of our approach, we create an Objective Conflict Dictionary (OCoDi) that can infer conflict intensity from text data with the goal to show researchers how to build such dictionaries in a transparent and computationally sensitive manner.

The main advantage of our approach, in contrast to traditional approaches to building dictionaries, is that we leverage the power of deep neural networks to extract a list of words associated with conflict intensity. The general idea is to train neural networks on a corpus of ICG CrisisWatch reports and match each text with the monthly number of fatalities that occurred in a given country as reported by the UCDP GED (Davies *et al.* 2022; Sundberg and Melander 2013). Based on the trained neural networks, we can extract words more or less strongly associated with conflict dynamics by using the feature importance analysis suggested by Horel *et al.* (2018). These feature importance scores can then not only be used to identify "positive" and "negative" words, but can also be used to give weights to each word on how strong its association with conflict fatalities is. This allows us to build a dictionary that is not only "objective," in contrast, to manually annotated dictionaries that are subject to human interpretation, but also to measure different strengths of

7 In recent years, there have been attempts to automate this maintenance by combining newer NLP and machine learning approaches. For an example, see Radford (2021).

8 For a discussion of problems associated with applying dictionaries outside their intended purpose, see, e.g., van Atteveldt, Velden, and Boukes 2021; Boukes *et al.* 2020; Loughran and McDonald 2011; Watanabe 2020. Furthermore, as these approaches are geared toward identifying events, it is difficult to extract useful information from text sources that do not necessarily contain event descriptions.

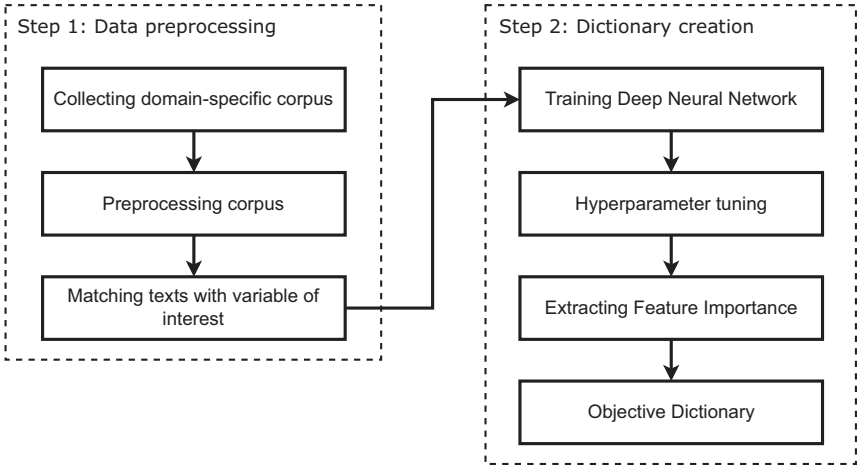


Figure 1. Schematic overview of our advanced dictionary creation approach.

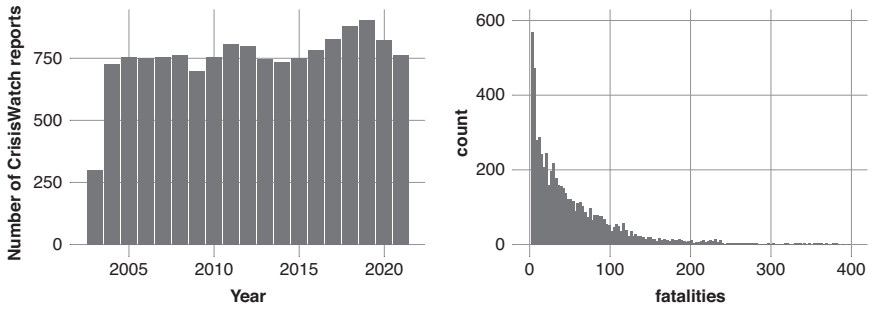
association. These modeling choices were made by carefully considering the trade-off between the complexity of language representation and interpretability. The decision to employ a deep neural network architecture allows for a relatively complex representation of the underlying text data, and combining it with sensitivity analysis to extract features for the dictionary increases explainability and transparency. Hence, this technique offers multiple advantages over other approaches. First, by training the neural networks on observed conflict dynamics, not only can some subjectivity of manual labeling be eliminated, but the resulting words are more directly linked to the concept of interest. Second, this approach can identify words that are reliably associated with positive or negative conflict trends, but that may seem counterintuitive or unsuspicious to an area expert and hence would not be included in a dictionary. Also, many words do not carry an inherent "positive" or "negative" conflict-related connotation as is the case in sentiment analysis. Third, this approach is able to scale each word relative to other words with respect to how influential they are for generating the model results, avoiding oversimplification by dividing words into a dichotomous categorization or arbitrarily weighting terms. With this approach, differences in importance are measurable, which allows for a much more nuanced dictionary. Finally, this technique offers a great deal of flexibility, allowing researchers to create dictionaries that are precisely calibrated to their needs without having to go through an arduous process of, for example, fine-tuning a transformer/BERT model (Devlin *et al.* 2018).

Figure 1 gives a schematic overview and illustrates all steps necessary to create an objective dictionary. We do believe that this approach can serve as a blueprint for other fields and applications as it is computationally inexpensive and provides a fast alternative to creating domain-specific dictionaries in a transparent manner.

Based on these advantages, we expect our approach to outperform general-purpose dictionaries in terms of accuracy and perform comparably with more complex approaches, such as text-scaling approaches or transformer models, while at the same time being more resource efficient. The following sections will give an intuitive description on how the dictionary was created, and we will provide an evaluation of its performance in comparison to other related approaches. For the interested reader, some of the applied concepts will be elaborated in greater detail in the Supplementary Material.

4 Creating the Conflict Dictionary

Our goal is to construct a dictionary that leverages the learning capabilities of a deep feed-forward neural network in combination with model interpretability. On the one hand, we want to build a model that is able to learn highly complex and possibly nonlinear relationships between



(a) Distribution of ICG reports per year. August 2003 to December 2021

(b) Distribution of fatalities. *Source*: UCDP 2003–2021

Figure 2. Distribution of our main data sources.

input features and target. On the other hand, we want this model to be as interpretable and computationally inexpensive as possible. This section will present the dataset, the methodology for creating the objective dictionary, and the evaluation process.

4.1 Data

The core data source for our dictionary is a corpus of English conflict reports. It is based on ICG CrisisWatch reports. The ICG employs a large roster of country and regional experts that compile assessments and outlooks for over 70 crises around the world on a regular basis. Until the end of 2021, CrisisWatch has produced over 14,000 monthly reports on a variety of conflicts since 2003. These reports are freely available online and are an essential tool for policymakers around the world as they focus on improving or deteriorating developments in conflict environments.[9] The amount of CrisisWatch reporting has remained relatively consistent (Figure 2a). For our target variable, we rely on the UCDP GED[10] that records estimates of fatalities for each event (Davies *et al.* 2022; Sundberg and Melander 2013).[11] We aggregate these fatalities on a country-month level as the CrisisWatch reports are also published at a monthly level and log-transform them (natural logarithm). As one can see in Figure 2b, even when omitting months with 0 fatalities, which is the vast majority of cases, the data are quite imbalanced and right-skewed, a common feature of conflict-related data.

4.2 Training Deep Neural Networks on Text Data

We train a deep feed-forward neural network that uses the CrisisWatch text corpus to predict the natural logarithm of fatalities on a country month level. The text corpus is transformed into a document term matrix, where each row corresponds to a document and each column corresponds to a word in the corpus.[12] The feature space was reduced to the most frequent 3,000 words as well as the top 1,000 bi-grams.[13] In order to select those words, the texts were preprocessed using standard NLP practices. It is important to mention that we ensured that the dictionary remains general enough over time by excluding words related to locations (countries, regions, etc.), landmarks (such as names of rivers and mountains), and people (names). Therefore, the dictionary does not simply learn that a country is associated with conflict but rather picks up different patterns. For a detailed description of those preprocessing steps and the document term matrix, the reader is referred to the Supplementary Material.

9 All CrisisWatch reports can be accessed at https://www.crisisgroup.org/crisiswatch.
10 UCDP GED can be accessed at https://ucdp.uu.se/downloads/.
11 Replication code and data for this article are available in Häffner *et al.* (2023) at https://doi.org/10.7910/DVN/Y5INRM.
12 For a discussion on why we work with a document term matrix instead of an long short-term memory (LSTM) model taking into account word dependencies, the reader is referred to the Supplementary Material.
13 We chose the cutoff of 3,000 most frequent words based on the fact that less frequent words appeared fewer than 12 times (0.002%) in the whole corpus.

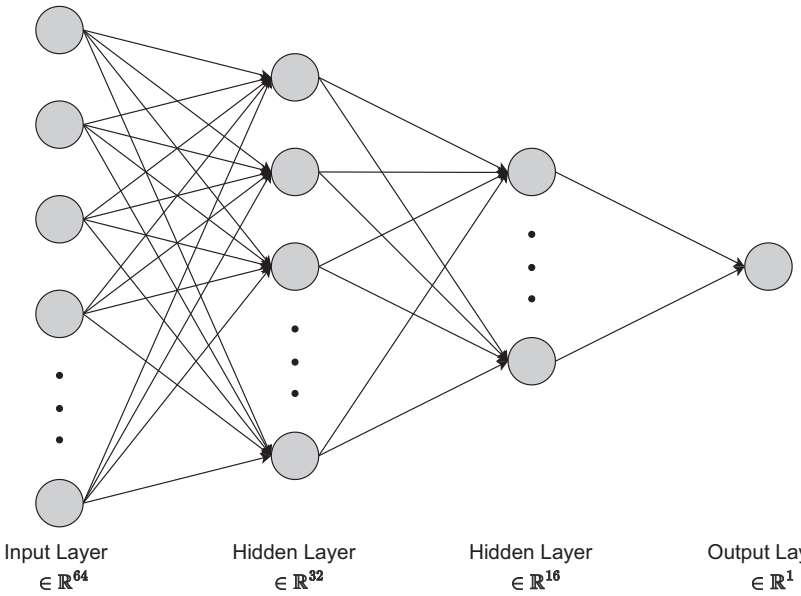


Figure 3. Neural network architecture.

The dataset is split into a train (all observations between 2003 and 2020) and a test set (2021). The train set is further split into the real train set (2003 to first half of 2020) and the validation set (second half of 2020). Figure 3 shows an overview of the final architecture of the neural network. It contains 64 input neurons and one output neuron that outputs the predicted log fatalities on a country-month level. The input and output layers are connected by two hidden layers which are connected by Swish activation functions. The final activation function is a ReLu function and ensures the predicted log fatalities are strictly positive. We use a batch size of 1,024 and 2,000 epochs. The optimizer used to train the neural network is the adaptive moment estimation (Adam) algorithm. It combines the advantages of two other variants of stochastic gradient descent-based optimizers, namely AdaGrad and RMSProp (Kingma and Ba 2017).

One of the most important tasks in machine learning is to build a model with good generalization capacities, meaning that the model needs to perform well on unseen data. In order to avoid overfitting, a kernel regularizer, dropout rates, and early stopping were implemented. All of the abovementioned techniques and the neural network itself entail parameters that are not inherently learned in the training process and need to be specified *a priori*, they are so-called hyperparameters. Intuitively, one might assume that the best neural network also creates the best-performing dictionary. However, due to the in-part random initialization of weights, the performance of the neural networks can vary. According to Goodfellow, Bengio, and Courville (2016), weight initialization that leads to good optimization does not always translate into good generalization capacities. Therefore, as our goal is to obtain a network with good generalization capabilities, we include the dictionary size in the hyperparameter search.[14] We implemented random search with 200 random hyperparameter constellations in the following manner: First, for a given hyperparameter constellation, we estimate 10 neural networks and then extract feature importance scores as described below. We aggregate those scores according to Equation (2) and build a Random Forest model predicting the natural logarithm of fatalities. Finally, the network configuration that produces the best dictionary is chosen as the "best" neural network.

Hyperparameters optimized in the neural network include the learning rate, the number of hidden layers, the number of neurons per layer, the dropout rate, and the lambda parameter for

14 We thank Reviewer 1 for his insightful comment that led to the inclusion of the dictionary in the hyperparameter search which resulted in an improved dictionary.

the $\ell 1$ regularization. The dropout acts as an $\ell 2$ regularization leading to the application of both the $\ell 1$ and $\ell 2$ regularizations in the network. Currently, there is no universal framework on which hyperparameter spaces to choose *a priori*. However, it is agreed upon that the most important hyperparameter in settings with a stochastic gradient descent-based training algorithm is the learning rate (see Bengio 2012; Goodfellow *et al.* 2016). Typical learning rate values range from 10^{-5} to 10^{-1}; therefore, we constructed the search space for the learning rate to be bound between 10^{-6} and 10^{-2}. Regarding the number of neurons in each layer, due to historical implementations, it is common to use (multiples of) the power of 2 constellations. We tested multiples between 1 and 4 of the power of 2 constellations of the following format: 32–16–8–4–2. The search space for hidden layers was restricted to between 0 and 3. Ultimately, the search space for the regularization parameters was bound to be between 0.10 and 0.40 for dropout rates and between 10^{-6} and 10^{-2} for the $\ell 1$ regularization. The following "ideal" parameters were identified: learning rate: 0.0196; hidden layers: 2; dropout rate: 0.3157; lambda: 0.0046; neurons: 64, 32, 16, 1 (for the input, the two hidden, and the output layers, respectively).

After the training of each neural network, the feature importance scores are calculated and then used to test the performance of the resulting dictionary in a Random Forest model. The feature importance scores help to differentiate important from unimportant words and consequently determine the selection of a word into the dictionary. In the following, the concept of feature importance as well as the applied method is introduced.

4.3 Extracting Weighted Dictionary Words

According to Horel *et al.* (2018), sensitivity analysis is an especially suitable approach for assessing the predictions of a neural network. It is intuitive, computationally inexpensive, offers two kinds of model explanations (local and global), and can be applied to many different neural network architectures.[15] In the following, the global aggregation level, as well as a formal notation, will be introduced:

$$\lambda_j = Dir \times \frac{100}{C} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial f(x^i)}{\partial x_j}\right)^2}. \tag{1}$$

The feature importance λ_j of the input feature j is measured in the following way. The derivatives of the output of the neural network are squared to avoid positive and negative values canceling out. Note that the vector of input features of the ith sample used to train the neural network is represented by x^i. Those derivatives are averaged over all training observations, with n being the number of training samples. Although this paper does not use a normalization factor C, it is also possible to normalize feature importance values such that $\sum_{j=1}^{p} \lambda_j = 100$. The number of input features is represented by p. When using a normalization factor C, the size of the feature importance score depends on p. If p is large, values typically are very small and close to zero. In order to distinguish features that are positively (negatively) associated with the target, we employ an indicator function Dir that multiplies the gradient sum by 1 or -1, respectively, if the mean gradient is positive or negative. Consequently, positive values indicate a positive relationship with the target variable and vice versa. Without a normalization factor, feature importance scores are not bound and, in our case, range from around -2.5 to 2.5.

Using this global feature importance metric allows for the differentiation of important from unimportant features. Larger (absolute) feature importance values mean that the considered variable contributes a lot to the model's output sensitivity. Conversely, values close to zero are an indication of the insensitivity of the outcome of a model regarding the respective feature. Based

15 The advantages of sensitivity analysis are discussed in more detail in the Supplementary Material.

on this sensitivity, the most positive and most negative words were selected to form the respective dictionaries. The feature importance scores obtained from this step were then used to weight each word based on its "positivity" or "negativity." The resulting dictionaries were used to construct a conflict intensity index for each document that was used in the evaluation process as described below.

4.4 Evaluation Process

The evaluation process encompasses comparing our dictionary scores to different techniques that are frequently applied in NLP tasks. In order to demonstrate that the use of a deep neural network improves performance compared to more simple methods, we also build a dictionary from a Lasso regression model and compare the performance. In addition to calculating the conflict intensity index for each document utilizing our objective dictionary (OCoDi), we calculate sentiment scores for each document based on two popular sentiment dictionaries (The Harvard IV-4 [HGI4] dictionary and the Valence Aware Dictionary and sEntiment Reasoner [Vader] dictionary). We also analyze our text data with the PETRARCH2 system that employs the conflict-specific CAMEO and TABARI event extraction dictionaries and use the CAMEO conflict-cooperation scale to assign scores to each text (Goldstein 1992).[16] Next, we rely on two different document scaling techniques to infer relative document positions from our evaluation corpus. We also fine-tune a ConfliBERT model on CrisisWatch reports and then directly predict fatalities for the test data. All scores as described above are calculated at the country-month level and matched with monthly aggregated fatalities from the UCDP GED database. These measures, with the exception of the ConfliBERT scores,[17] serve as the input data to several machine learning models predicting fatalities. We employ a Random Forest as well as an eXtremeGradient Boosting (XGBoost) model and optimize both models with regard to their optimal hyperparameters (Random Search with 50 parameter constellations). For the Random Forest model, we treat the number of trees (600–1500) as well as the maximal depth (7–15) as hyperparameters; for the XGBoost models, we treat the learning rate (0.05, 0.1, 0.20), the number of boosting stages (100, 400, 800), the maximal depth (3, 6, 9), and the minimum sum of instance weight (hessian) needed in a child (1, 10, 100) as hyperparameters. A table with the optimal hyperparameters for each model can be found in the Supplementary Material.

4.4.1 Dictionary-Based Approaches. For each text document, we use the abovementioned approaches to calculate a score. For the OCoDi, this score represents which and how many words are contained in a report that have been identified as being associated with higher or lower levels of fatalities by our deep neural network. Different methods exist to calculate those scores ranging from simply counting the emergence of dictionary words per article to more sophisticated word weighting schemes. As our dictionary contains the associated feature importance scores, we utilize this information to extract the weighted conflict intensity index:

$$FI = \frac{\sum(FI\_Score_{pos}) + \sum(FI\_Score_{neg})}{len(doc)}. \tag{2}$$

It is important to account for document length when constructing the scores according to Equation (2). Longer texts potentially contain more words, and the comparison of different scores becomes difficult. Figure 4 shows a CrisisWatch report in its original form, and Figure 5 shows an exemplary CrisisWatch report for Afghanistan after some preprocessing.[18] In Figure 5, we also

16 We would like to thank Reviewer 2 for suggesting to include these dictionaries in our comparison efforts.
17 We use ConfliBERT to directly predict the target variable, rather than extracting document scores and feeding them into a Random Forest or XGBoost model.
18 Only lowercasing, lemmatization, stop-word removal. References to locations and entities were removed in a later step.

Afghanistan, September 2003

Attacks by extremists against U.S. forces, government troops and aid workers continue in south. Four Afghans working for Danish NGO killed on 8 September; two other aid workers killed on 24 September while delivering clean drinking water to village in Helmand province. Growing tension between Kabul and Islamabad: Afghan Government accuses Pakistan of doing too little to prevent militants from regrouping in Pakistan. Both have agreed to reinforce troops on border to monitor crossings. Battles between local commanders in north continue to cause displacement and civilian casualties. Demobilisation and reintegration program delayed by government failure to reform defence ministry. Draft constitution to be unveiled in early October. American special envoy Zalmay Khalilzad named U.S. ambassador. NATO experts to study feasibility of expanding ISAF mandate beyond Kabul; Germany announced readiness to deploy 250-450 troops to northern city of Kunduz. More than 100 Taliban fighters killed since Coalition Operation Mountain Viper launched on 25 August.

Figure 4. Unprocessed CrisisWatch report for Afghanistan, September 2003. *Source*: https://www.crisisgroup.org/crisiswatch.

Afghanistan, September 2003 [+0.24]

attack [1.0] extremist us forces government troop [0.78] aid worker continue south [0.83] work danish ngo kill [1.56] september two aid worker kill [1.56] deliver [0.74] clean drinking water village helmand province [0.66] grow tension [-0.58] islamabad government accuse little [0.72] prevent militant [0.80] regroup agree reinforce troop [0.78] border [-0.77] monitor crossing battle [0.79] local commander [0.75] north continue cause displacement [0.80] civilian [1.28] casualty [0.65] demobilisation reintegration [-0.65] program delay government failure reform defence ministry [0.47] draft constitution [-0.55] unveil early special envoy [0.82; 0.76] name [-0.58] us ambassador expert study [0.78] feasibility expand [0.91] isaf mandate beyond [1.22] announce readiness deploy troop [0.78] northern city [1.16] kunduz fighter [0.83] kill [1.56; 0.63] since [0.84; -0.58] coalition operation [1.45] mountain viper launch

Document length: 90
OCoDi score: 0.24 = (25.46 - 3.72) / 90

Figure 5. Preprocessed CrisisWatch report with dictionary words and scores highlighted.

assigned our feature importance-based weights to each of the words contained in OCoDi and also show how Equation (2) is applied to calculate our OCoDi document-level score. In Figure 5, dictionary words associated with lower levels of fatalities are colored blue, high levels with red, and bi-grams are indicated by an underline.[19]

For each of the other comparison NLP methods, the Supplementary Material provides further information on the respective models as well as on how the respective scores are calculated. Where applicable, we also calculate simple word counts adjusted by document length (unweighted scores) and compare the performance of our OCoDi and the other dictionaries to obtain a more straightforward comparison.

5 Results

In this section, we will discuss the resulting dictionary and how well it performs compared with other NLP approaches. To give a general impression of how the words correspond to feature importance scores, Table 1 gives an overview of the top words associated with more (positive score) and fewer (negative score) fatalities for our dictionary. In total, our dictionary contains 1,100 words. The results shown in this small subsection of words are partly intuitive, whereas others do not seem to be too intuitive. However, as mentioned above, we do not expect to see only intuitive words appear here, but would even consider it a strength of our approach that it is able to identify markers that would usually not be selected.

19 It is noteworthy that words can be in the dictionary by themselves, as well as, as a part of a bi-gram. Furthermore, words can be positive by themselves, but negative in combination with another word.

Table 1. Top 10 most *positive* (more fatalities) and *negative* (fewer fatalities) terms based on feature importance for International Crisis Group reports.

Term ICG	FI score	Term ICG	FI score
Positive		*Negative*	
Cartel	2.25	Battalion	−1.76
Immediate	2.04	Defender	−1.76
Insurgent	1.95	Purge	−1.77
Plateau	1.89	Vessel	−1.82
Wounded	1.88	Identification	−1.86
Allied	1.82	Pacific	−1.92
Forced	1.80	Provocative	−1.98
Generation	1.79	Stab	−2.12
Bad	1.78	Prince	−2.35
Offensive	1.78	Preval	−2.37

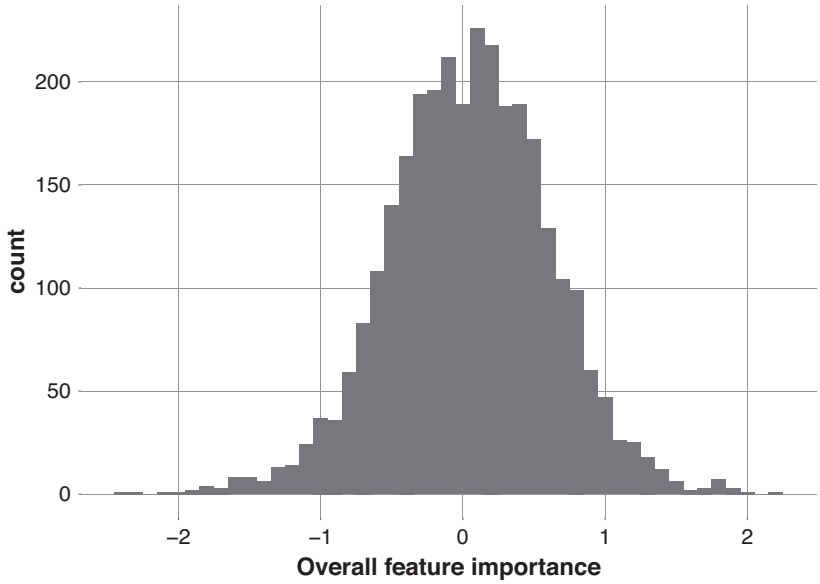


Figure 6. Distribution of feature importance scores for all words.

In Figure 6, we show that the feature importance scores for all words are distributed reasonably well along a normal distribution which is what we would expect to see based on our approach to calculating the scores. Notably, there are fewer words around 0, indicating that our network classifies not as many words as "neutral" or just slightly "positive" or "negative" as compared to a true normal distribution. However, this should not affect its validity. The dictionaries so far seem to produce outputs in line with our expectations. However, in order to assess how well they really capture conflict trends, we evaluate how well each approach is associated with conflict fatalities over time.

To show this, we calculate the difference between the different scores and the actual observed fatalities, aggregated on a yearly level. To make them comparable, the fatalities and scores were scaled and centered. Figure 7 shows the trends for our dictionary (OCoDi) in comparison to the other approaches. The further away each line in Figure 7 is from 0, the less well that approach is aligned with actual observed fatalities. As can be seen, our approach is in general very close to the

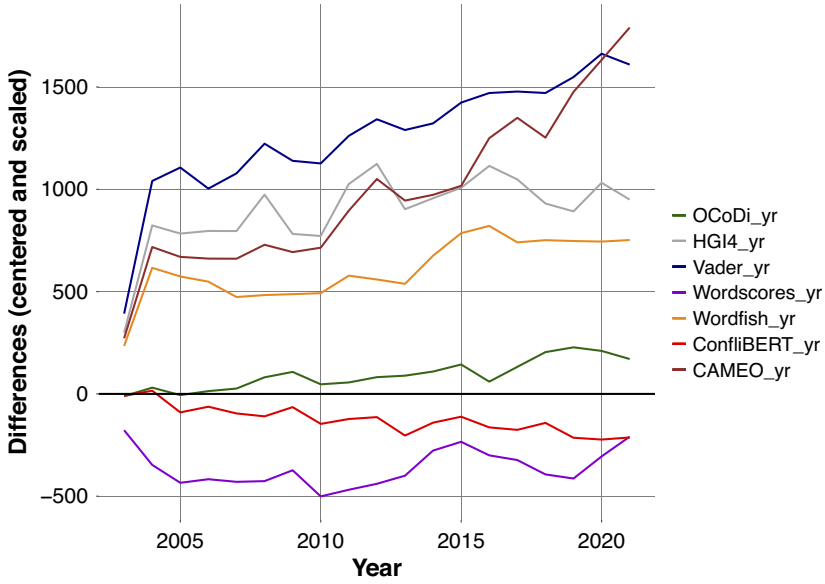


Figure 7. Comparison between different scores and fatalities.

0 baseline, highlighting that our approach is able to capture general trends in conflict dynamics better than traditional approaches. The magnitude of the difference between our approach and the other measurements is surprisingly large, underlining the importance of validating general-purpose approaches when using them in different settings (Bruinsma and Gemenis 2019). ConfliBERT also performs very well in this graph, but, similar to OCoDi, the line increasingly moves away from 0 over time.

However, while visually comparing trends over time gives a good intuition of how these methods behave when highly aggregated, we are more interested in how well they perform on a lower level of aggregation. For this purpose, we also calculate correlation scores between all approaches and the target variable.

As can be seen in Figure 8, OCoDi shows a strong association with the outcome variable. The association between OCoDi and log fatalities is the second highest, with only ConfliBERT having a slightly higher coefficient. It is also an order of magnitude higher than almost all of the other tested approaches, with a correlation score of 0.735 as compared to −0.272 for HGI4, −0.426 for Vader, 0.419 for Wordscores, −0.069 for Wordfish, −0.427 for CAMEO, and 0.792 for ConfliBERT.[20] Furthermore, whereas the correlation between OCoDi/the ConfliBERT model and fatalities is strong and in the expected direction, for Vader, HGI4, and CAMEO, the correlation is lower and negative. While Wordscores reach reasonable levels of correlation, the Wordfish score is close to zero. So far all approaches trained on the specific task of conflict intensity perform much better than general-purpose approaches or unsupervised learning. Additionally, ConfliBERT and OCoDi are performing substantively better than Wordscores.

To further investigate how well our approach performs, we also test its accuracy in solving a text regression task. For this purpose, we train a series of simple Random Forest (Ho 1995) and XGBoost (Chen and Guestrin 2016) models that only take the document scores from the different approaches as predictors. We report mean squared error (MSE) and R^2 for all of our models to assess their performance. The MSE records how far away on average our predictions are from the true values of our target variable, whereas R^2 gives an indication of how much our variables contribute to the predictions compared to a null model. As mentioned before, we also train a ConfliBERT (Hu *et al.* 2022) model, which has been fine-tuned on ICG reports. ConfliBERT can be

20 The corresponding standard errors for all relevant correlation coefficients are given in parentheses.

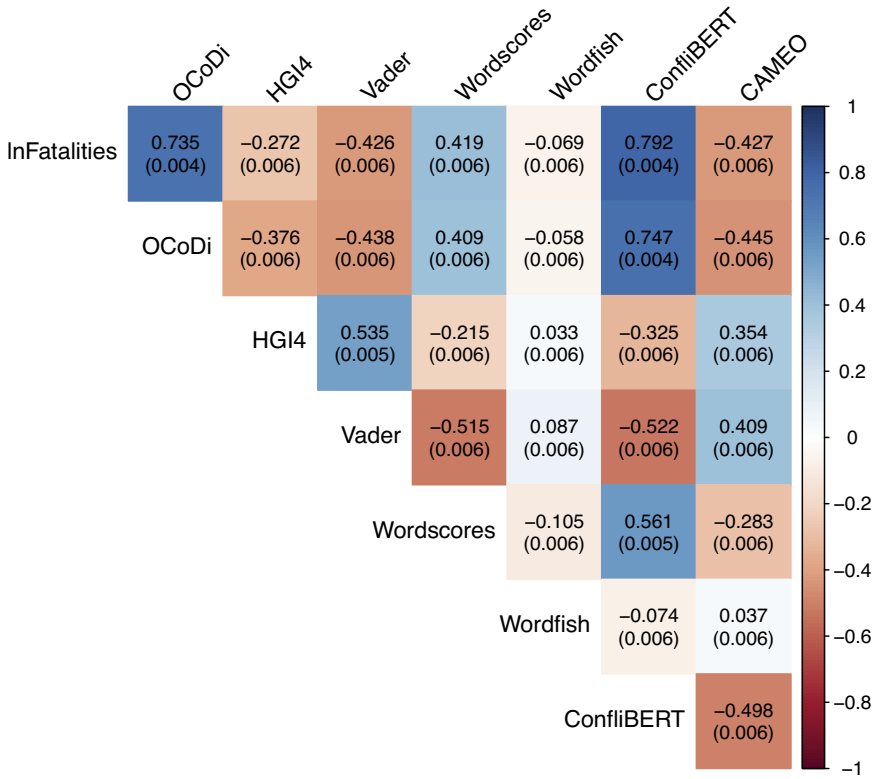


Figure 8. Correlation plot for fatalities and the different scores.

used to directly predict fatalities for our test data. So rather than having to calculate scores for each document and using them in a Random Forest or XGBoost model, we use the more straightforward approach of predicting directly from our ConfliBERT model. Hence, for the ConfliBERT model, we only report one set of performance metrics.

Before presenting the final results of our evaluation process, we want to compare the performance of our dictionary with the results of the neural networks that were used to create the dictionary. The best neural network (out of the 10 neural networks that we estimated in total) reaches an R^2 of 0.65, whereas the dictionary reaches an R^2 of 0.64 (Random Forest) and an R^2 of 0.63 (XGBoost). However, the average performance of the neural networks is very similar to our dictionary (R^2 of 0.64). Some neural networks even have a considerably lower predictive accuracy than the others (R^2 below 0.63). This is in line with Goodfellow *et al.* (2016) who claim that weight initialization that leads to a good optimization does not always translate into good generalization capacities. Therefore, we believe that working with a dictionary as opposed to working directly with the neural network is justified as the results are more stable and there is no decrease in performance.

As can be seen in Table 2, our feature importance-based dictionary approach outperforms all other approaches for both the Random Forest and XGBoost models. The best performing approach is highlighted in bold. It is worthwhile mentioning that the use of a neural network in the dictionary creation process is justified as a dictionary created by a Lasso regression model performs worse than our approach. The results of the Lasso dictionary are reported in the Supplementary Material. Our dictionary also outperforms the other dictionaries when we compare the unweighted feature importance scores. The results of this comparison effort can be found in the Supplementary Material. These results are very encouraging, particularly the comparatively high R^2, which indicates that our model is actually learning a reasonable amount of information that can be used for this text regression task, whereas the other approaches, with the exception of ConfliBERT, reach

Table 2. Results of predicting fatalities with different approaches.

Model	OCoDi	Vader	HGI4	Wordfish	Wordscore	CAMEO
Random Forest						
MSE	**1.59**	2.61	3,13	4,39	2,20	2.68
R^2	**0.64**	0.40	0.29	−0.00	0.50	0.39
XGBoost						
MSE	**1.60**	2.60	2.99	4.40	2.21	2.65
R^2	**0.63**	0.41	0.32	−0.00	0.50	0.39
	ConfliBERT					
MSE	1.75					
R^2	0.60					

considerably lower values. Furthermore, while the MSE is relatively high, underlining that conflict prediction remains a difficult task, it is significantly lower for OCoDi and ConfliBERT than for all other approaches in both model specifications. The good results for ConfliBERT again underline the importance of using domain-specific NLP tools. And, while our approach performs better, these results are impressive, particularly given that ConfliBERT is not pre-trained on this specific task.[21] It is worthwhile mentioning that the R^2 can indeed be negative (Wordfish) as, except the name itself, nothing prevents the R^2 from being negative. The formula for the R^2 for Random Forest and XGBoost states that $R^2 = 1 - \frac{SSR}{SST}$, where SSR refers to the sum of squared residuals of the chosen model and SST to the total sum of squares from the mean model. If the SSR is larger than the SST, the R^2 is negative. This is the case for models that are worse than a model fitting a constant (mean model).

Finally, given the skewed distribution of our target variable, we also investigate how our out-of-sample point predictions compare to the actual observed values (observed value − predicted value = difference). As can be seen in Figure 9, the predictions based on our approach seem to overestimate (negative difference values) some low fatality observations, while increasingly underestimating (positive difference values) observations with higher actual fatalities.[22]

Overall, given the results above, we can conclude that our model is performing reasonably well across all specifications. Furthermore, our approach is outperforming all other approaches on unseen test data. The results also highlight a number of important aspects. First, our approach produces competitive results, even in comparison to more complex approaches. While one should not expect to be able to reach a very high accuracy based on text data alone, it could serve as a viable additional or supplementary indicator that can be made available both at a high temporal resolution and at higher levels of aggregation. Second, conflict prediction remains a difficult endeavor as indicated by the relatively high MSE across all models. This should not come as a huge surprise, as models for conflict prediction, particularly in regression tasks, still often do not reach satisfactory levels of precision (Vesco *et al.* 2022). Third, our approach seems to offer an efficient solution to analyzing text sources in the setting of conflict research. Our approach reduces computational costs, compared with more complex transformer models, while better capturing actual conflict dynamics. Our dictionary also outperformed other dictionary, text-scaling, and transformer-based approaches in the presented text regression task. This further underlines the

21 The pre-training includes tasks such as binary classification, sequence labeling, or named entity recognition.
22 We also investigated which countries are driving these over- and under-estimations. A discussion can be found in the Supplementary Material.

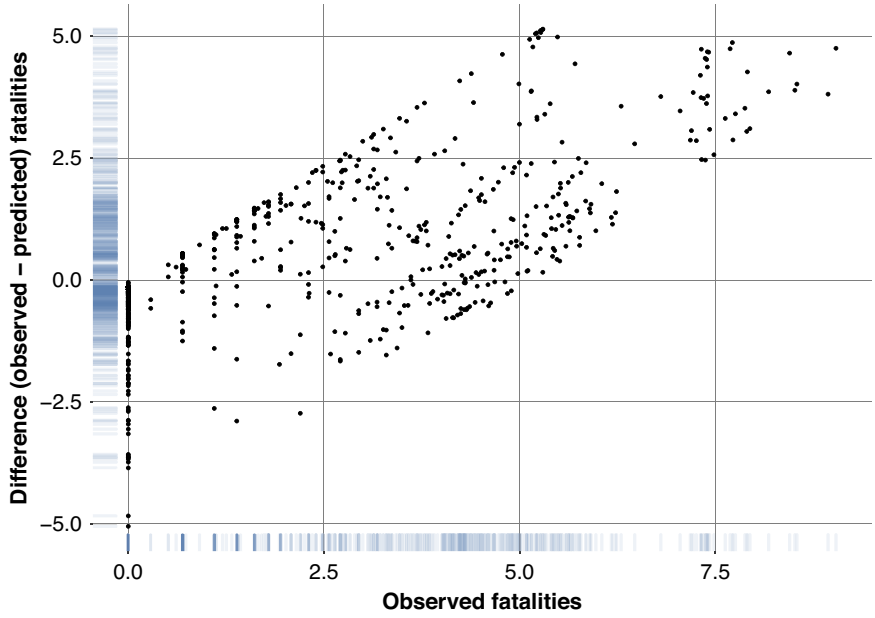


Figure 9. Observed versus difference (sorted by fatalities), Random Forest.

importance to use tools that are domain-specific. Finally, given the flexibility of our approach, it seems worth testing it for different target variables in- and outside of the area of conflict research.

6 Conclusion

Recent advancements in NLP approaches have achieved impressive results. Transformer models particularly have been shown to successfully model complex language patterns. While these improvements have found widespread appeal in many research areas, they do come with their own set of limitations. Modern NLP methods require vast amounts of text data and sophisticated IT infrastructure. As these models rely on very complex representations of the underlying data, they are increasingly difficult to interpret.

In order to alleviate some of these problems, we introduce an approach to domain-specific text analysis that carefully weighs the balance between performance and interpretability. We propose a deep learning approach in combination with techniques to increase explainability to construct objective dictionaries. As an illustrative application of our approach, we build an objective dictionary that can infer conflict intensity from conflict-related reports. The approach relies on training deep neural networks on approximately 14,000 expert-written ICG CrisisWatch reports between 2003 and 2021. In addition, we use fatality numbers (natural logarithm), as reported by the prominent UCDP GED (Davies *et al.* 2022; Sundberg and Melander 2013), as our target variable, linking our dictionary more closely to actual levels of conflict intensity. Consequently, the need for manual annotation or selection of relevant words is eliminated and the subjectivity of the created dictionary is reduced. We then extract words that are indicative of higher or lower levels of fatalities through a feature importance metric (sensitivity analysis) introduced by Horel *et al.* (2018). This sensitivity analysis increases the interpretability of the results of our neural networks while being computationally inexpensive. With these words, we then create our objective dictionary. We are able to show that our dictionary is well equipped to adequately measure trends in conflict intensity over time. We also evaluate how well our dictionary performs at predicting levels of fatalities using Random Forest and XGBoost models. We found that our approach consistently outperforms related approaches, such as general-purpose dictionaries, conflict event coding dictionaries, text scaling, or even BERT models while lowering computational costs.

Overall, we can show that our approach offers a range of advantages over existing approaches. First, our approach can easily be applied to different target variables and text corpora, giving researchers a great deal of flexibility to adapt it to their specific requirements. This also ensures that the dictionaries are better able to capture the relevant concept of interest. Rather than having to rely on a subjective assessment if a word carries an inherent connotation to the concept in question, the deep neural network-based approach guarantees that the words extracted are directly linked to a quantifiable outcome variable. Second, the word list created through our approach is very transparent, particularly compared to BERT models, allowing researchers to validate, assess, reuse, and reproduce it in its entirety. Finally, applying this approach is computationally cheap while outperforming other approaches in a text regression task. It requires much less computing power than state-of-the-art transformer models, and one does not need to go through the laborious effort of creating dictionaries manually. Based on these results, we are confident that our approach can serve as a successful blueprint for future researchers to analyze text data for domain-specific applications.

Nonetheless, there are a number of potential avenues to improve our approach further. Most importantly, building the dictionary on a larger and more diverse text corpus and testing its performance on a different corpus to assess its generalizability (e.g., applying the dictionary to a broader corpus of conflict news) could prove interesting. A comprehensive assessment of how alternative models to neural networks (e.g., Ridge regression or LSTM) perform at dictionary creation across research domains could prove to be a worthwhile endeavor. Testing this approach for additional target variables could offer interesting cases for future applications, as using the number of fatalities, while simple and straightforward, may not be the best or at least not the only operationalization of our target variable. Finally, given the success of our illustrative dictionary in the domain of conflict research, it could prove worthwhile to supplement existing conflict prediction models with features created through our dictionary creation approach to investigate its potential to improve the accuracy of conflict predictions.

Acknowledgments

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any agency of the German government. We gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the Bundeswehr University Munich to fine-tune the BERT models. We would also like to thank the participants of the COMPTEXT Conference held in Dublin, 6-7 May 2022, for their constructive and helpful feedback and comments. Furthermore, we would like to thank Nils Weidmann for his insightful comments during the workshop 'Challenges and Opportunities of Crisis Early Warning' at the Center for Crisis Early Warning (KompZ KFE) in Neubiberg, 22 June 2022. We would like to thank Marje Kaack for supporting the development of the web application. We would especially like to thank Marco N. Binetti, Vanessa Gottwick, Christian Oswald, and Ivana Peric who provided valuable suggestions and corrections to earlier versions of this paper. Finally, we would like to thank Jeff Gill, the Editor of Political Analysis, and the two anonymous reviewers for their extremely helpful and constructive comments.

Funding Statement

The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office.

Conflict of Interest

There is no conflict of interest to disclose.

Data Availability Statement

Replication code and data for this article are available in Häffner *et al.* (2023) at https://doi.org/10.7910/DVN/Y5INRM.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.7.

References

Azar, E. E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24 (1): 143–152. https://doi.org/10.1177/002200278002400106

Bengio, Y. 2012. "Practical Recommendations for Gradient-Based Training of Deep Architectures." In Neural Networks: Tricks of the Trade, edited by G. Montavon, G. B. Orr and K. R. Müller. Lecture Notes in Computer Science, Vol. 7700. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-35289-8_26

Bond, D., J. Bond, C. Oh, J. C. Jenkins, and C. L. Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development." *Journal of Peace Research* 40 6: 733–745.

Boukes, M., B. van de Velde, T. Araujo, and R. Vliegenthart. 2020. "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement between Off-the-Shelf Sentiment Analysis Tools." *Communication Methods and Measures* 14 (2): 83–104. https://doi.org/10.1080/19312458.2019.1671966/SUPPL_FILE/HCMS_A_1671966_SM0283.DOCX

Boussalis, C., T. Chadefaux, A. Salvi, and S. Decadri. 2022. "Public and Private Information in International Crises: Diplomatic Correspondence and Conflict Anticipation." *International Studies Quarterly* 66 4: sqac056. https://doi.org/10.1093/isq/sqac056

Bruinsma, B., and K. Gemenis. 2019. "Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling." *Communication Methods and Measures* 13 (3): 212–227. https://doi.org/10.1080/19312458.2019.1594741

Carta, S. M., S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero. 2020. "Dynamic Industry-Specific Lexicon Generation for Stock Market Forecast." In *Machine Learning, Optimization, and Data Science*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12565, 162–176. Cham: Springer. https://doi.org/10.1007/978-3-030-64583-0_16

Carta, S. M., S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero. 2021. "Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting." *IEEE Access* 9: 30193–30205. https://doi.org/10.1109/ACCESS.2021.3059960

Chadefaux, T. 2014. "Early Warning Signals for War in the News." *Journal of Peace Research* 51 (1): 5–18. https://doi.org/10.1177/0022343313507302

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 785–794. San Francisco: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Davies, S., T. Pettersson, and M. Öberg. 2022. "Organized Violence 1989–2021 and Drone Warfare." *Journal of Peace Research* 59 (4): 593–610. https://doi.org/10.1177/00223433221108428

de Coning, C. 2020. "Insights from Complexity Theory for Peace and Conflict Studies." In *The Palgrave Encyclopedia of Peace and Conflict Studies*, 1–10. Cham: Springer. https://doi.org/10.1007/978-3-030-11795-5_134-1

de Vries, E. 2022. "The Sentiment Is in the Details: A Language-Agnostic Approach to Dictionary Expansion and Sentence-Level Sentiment Analysis in News Media." *Computational Communication Research* 4 (2): 424–462. https://doi.org/10.5117/CCR2022.2.003.VRIE

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "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*, Vol. 1 (Long and Short Papers), 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. https://aclanthology.org/N19-1423

Dunphy, D. C. 1974. *Harvard IV-4 Dictionary General Inquirer Project*. Sydney: University of New South Wales.

Gleditsch, K. S. 2020. "Advances in Data on Conflict and Dissent." In *Computational Conflict Research*, edited by E. Deutschmann, J. Lorenz, L. G. Nardin, D. Natalini, and A. F. X. Wilhelm, 23–41. Cham: Springer. https://doi.org/10.1007/978-3-030-29333-8_2

Goldstein, J. S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *The Journal of Conflict Resolution* 36 (2): 369–385.

Goodfellow, I. J., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge: MIT Press. http://www.deeplearningbook.org

Greene, K. T., and C. Lucas. 2020. "Once More, with Feeling: Using Sentiment Analysis to Improve Models of Relationships between Non-State Actors." *International Interactions* 46 (1): 150–162. https://doi.org/10.1080/03050629.2019.1684913

Häffner, S., M. Hofer, M. Nagl, and J. Walterskirchen. 2023. "Replication Data for: Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/Y5INRM

Ho, T. K. 1995. "Random Decision Forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada*, Vol.1, 278-282. https://doi.org/10.1109/ICDAR.1995.598994

Horel, E., V. Mison, T. Xiong, K. Giesecke, and L. Mangu. 2018. "Sensitivity Based Neural Networks Explanations." Working Paper. http://arxiv.org/abs/1812.01029

Hu, Y., M. S. Hosseini, E. Skorupa Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio. 2022. "ConfliBERT: A Pre-Trained Language Model for Political Conflict and Violence." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5469–5482. Seattle, USA: Association for Computational Linguistics.

Hutto, C., and E. Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media*, 8 (1): 216–225.

Jha, V., R. Savitha, P. D. Shenoy, K. R. Venugopal, and A. K. Sangaiah. 2018. "A Novel Sentiment Aware Dictionary for Multi-Domain Sentiment Classification." *Computers & Electrical Engineering* 69: 585–597. https://doi.org/10.1016/J.COMPELECENG.2017.10.015

Kingma, D. P., and J. Ba. 2017. "Adam: A Method for Stochastic Optimization." Working Paper. Available at https://doi.org/10.48550/ARXIV.1412.6980

Laver, M., K. Benoit, and J. Garry. 2003. "Extracting Policy Positions from Political Texts using Words as Data." *The American Political Science Review* 97 (2): 311–331.

Lee, G. T., C. O. Kim, and M. Song. 2021. "Semisupervised Sentiment Analysis Method for Online Text Reviews." *Journal of Information Science* 47 (3): 387–403. https://doi.org/10.1177/0165551520910032

Leetaru, K., and P. A. Schrodt. 2013. "GDELT: Global Data on Events, Location, and Tone." In ISA annual convention 2 (4): 1–49, San Francisco, USA. International Studies Association.

Li, S., W. Shi, J. Wang, and H. Zhou. 2021. "A Deep Learning-Based Approach to Constructing a Domain Sentiment Lexicon: A Case Study in Financial Distress Prediction." *Information Processing & Management* 58 (5): 102673. https://doi.org/10.1016/J.IPM.2021.102673

Lo, J., S.-O. Proksch, and J. B. Slapin. 2016. "Ideological Clarity in Multiparty Competition: A New Measure and Test using Election Manifestos." *British Journal of Political Science* 46 (3): 591–610. https://doi.org/10.1017/S0007123414000192

Loughran, T., and B. McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1): 35–65.

Macnair, L., and R. Frank. 2018. "Changes and Stabilities in the Language of Islamic State Magazines: A Sentiment Analysis." *Dynamics of Asymmetric Conflict* 11 (2): 109–120. https://doi.org/10.1080/17467586.2018.1470660

McClelland, C. 1971. *The Management and Analysis of International Event Data: A Computerized System for Monitoring and Projecting Event Flows*. Los Angeles: School of International Relations, University of Southern California.

Mueller, H., and C. Rauh. 2018. "Reading between the Lines: Prediction of Political Violence using Newspaper Text." *American Political Science Review* 112 (2): 358–375.

Mueller, H., and C. Rauh. 2022a. "The Hard Problem of Prediction for Conflict Prevention." *Journal of the European Economic Association* 20 (6): 2440–2467. https://doi.org/10.1093/jeea/jvac025

Mueller, H., and C. Rauh. 2022b. "Using Past Violence and Current News to Predict Changes in Violence." *International Interactions* 48 (4): 579–596. https://doi.org/10.1080/03050629.2022.2063853

Norris, C., P. Schrodt, and J. Beieler. 2017. "PETRARCH2: Another Event Coding Program." *The Journal of Open Source Software* 2 (9): 133. https://doi.org/10.21105/joss.00133

Palmer, M., J. Roeder, and J. Muntermann. 2021. "Induction of a Sentiment Dictionary for Financial Analyst Communication: A Data-Driven Approach Balancing Machine Learning and Human Intuition." *Journal of Business Analytics* 5 (1): 8–28. https://doi.org/10.1080/2573234X.2021.1955022

Radford, B. J. 2021. "Automated Dictionary Generation for Political Eventcoding." *Political Science Research and Methods* 9 (1): 157–171. https://doi.org/10.1017/psrm.2019.1

Raunak, V., V. Gupta, and F. Metze. 2019. "Effective Dimensionality Reduction for Word Embeddings." In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP 2019)*, 235–243. Florence: Association for Computational Linguistics. https://doi.org/10.18653/V1/W19-4328

Rheault, L., K. Beelen, C. Cochrane, and G. Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS One* 11 (12): e0168843. https://doi.org/10.1371/JOURNAL.PONE.0168843

Rodriguez, P. L., and A. Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics* 84 (1): 101–115. https://doi.org/10.1086/715162/ASSET/IMAGES/LARGE/FG7.JPEG

Schrodt, P. A. 2008. "Kansas Event Data System (KEDS)." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/EXX5RM

Schrodt, P. A., D. J. Gerner, O. Yilmaz, D. Hermreck, A. Bron, A. Gregory, A. Ingram, M. Jekic, L. McMullen, and L. Prather. 2012. "The CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework." http://eventdata.psu.edu/

Sharir, O., B. Peleg, and Y. Shoham. 2020. "The Cost of Training NLP Models: A Concise Overview." https://doi.org/10.48550/ARXIV.2004.08900

Shilliday, A., and J. Lautenschlager. 2012. "Data for a Worldwide ICEWS and Ongoing Research." In *Advances in Design for Cross-Cultural Activities*, 455–465. Boca Raton: CRC Press.

Shu, R., and H. Nakayama. 2017. "Compressing Word Embeddings via Deep Compositional Code Learning." In *6th International Conference on Learning Representations (ICLR 2018)—Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1711.01068

Sood, M., J. Gera, H. Kaur, 2022. "Creation, Evaluation, and Optimization of a Domain-Based Dictionary." *Journal of Intelligent & Fuzzy Systems* 43 (5): 6123–6136. https://doi.org/10.3233/JIFS-220110

Stone, P. J., D. C. Dunphy, and M. S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–532.

Trubowitz, P., and K. Watanabe. 2021. "The Geopolitical Threat Index: A Text-Based Computational Approach to Identifying Foreign Threats." *International Studies Quarterly* 65 (3): 852–865. https://doi.org/10.1093/ISQ/SQAB029

van Aken, B., B. Winter, A. Löser, and F. A. Gers. 2019. "How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations." In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*, 1823–1832. Beijing: Association for Computing Machinery. https://doi.org/10.1145/3357384.3358028

van Atteveldt, W., M. A. van der Velden, and M. Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15 (2): 121–140. https://doi.org/10.1080/19312458.2020.1869198

Vesco, P., H. Hegre, M. Colaresi, R. B. Jansen, A. Lo, G. Reisch, and N. B. Weidmann. 2022. "United They Stand: Findings from an Escalation Prediction Competition." *International Interactions* 48 (4): 860–896. https://doi.org/10.1080/03050629.2022.2029856

Watanabe, K. 2020. "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages." *Communication Methods and Measures* 15 (2): 81–102. https://doi.org/10.1080/19312458.2020.1832976

Weidmann, N. B., and M. D. Ward. 2010 "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6): 883–901. https://doi.org/10.1177/0022002710371669

Widmann, T., and M. Wich. 2022. "Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text." *Political Analysis*: 1–16. https://doi.org/10.1017/pan.2022.15

Yuan, J., F. Gao, Q. Ho, W. Dai, J. Wei, X. Zheng, E. P. Xing, T. Y. Liu, and W. Y. Ma. 2015. "LightLDA: Big Topic Models on Modest Computer Clusters." *WWW 2015—Proceedings of the 24th International Conference on World Wide Web*, 1351–1361. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741115