Mining Large-scale News Articles For Predicting Forced Migration

Sadra Abrishamkar
York University
Toronto, Canada

Forouq Khonsari
York University
Toronto, Canada

ABSTRACT
Many people are being displaced every day from all around the globe. Many of them are forced to leave their homes because of socio-political conflicts, human-made or natural disasters. In order to develop an early warning system for forced migration in the context of humanitarian crisis, it is essential to study the factors that cause forced migration, and build a model to predict the future number of displaced people. In this research, we focus on studying forced migration due to socio-political conflicts for which violence is the main reason. In particular, we investigate whether the degree of violence in a specific region can be detected from news articles related to that region and whether the detected violence scores can be used to improve the prediction accuracy. We investigate three techniques to extract the degree of violence from a corpus of news articles: ED-FE, TD-FE and SWSW. SWSW measures the semantic similarity between documents and a set of seed-words representing violence. ED-FE extracts violent events from news articles, which are the incidents related to attacks or the ones resulting in casualties. TD-FE uses topic modeling techniques to reduce the size of the information for easier analysis and filtering the violent incidents. Experiments indicate that ED-FE and TD-FE provide accurate violence scores which are very effective features for making forced displacement forecasts and using them in prediction models improves the prediction accuracy.

CCS CONCEPTS
• Applied computing
  • Law, social and behavioral sciences;
  • Document management and text processing.

KEYWORDS
Forced Migration Prediction, Violence Detection, News Articles Analysis

1 INTRODUCTION
A state of forced displacement exists when a significant number of people in a given location have been displaced by socio-political conflicts, human-made or natural disasters, economic disturbence, disease or development projects. Examining the causes of forced displacement and building models to detect the number of displaced people are crucial to developing an early warning system for forced displacement.

News agencies produce a vast amount of information about the events happening every day, covering from local to international affairs. The content of news articles ranges from social, political, economic developments, to reports on the environmental events. As a result, the news articles collected from across the globe are a suitable source for analyzing the world events and violent incidents in particular. The objective of this research is to use machine learning and natural language processing techniques to develop a model for predicting the number of displaced people (e.g., refugees and asylum seekers) based on big data of news articles. We particularly focus on displacement in Iraq as a case study. Social scientists who have studied forced migration in Iraq have identified violence and security threat in the region to be the most important factor influencing forced displacement [3].

In this research, we measure the degree of violence by analyzing a corpus of news articles which is an excellent source to extract the latest events in a region. We propose two novel techniques to extract violence scores from news articles called Event Detection based Factor Extraction (ED-FE) and Topic Detection based Factor Extraction (TD-FE). ED-FE extracts violent events from news articles, which are the incidents related to attacks or the ones resulting in casualties. TD-FE uses topic modeling techniques to reduce the size of the information for easier analysis and filtering the violent incidents. We also use another technique called Similarity with Seed-Words (SWSW) introduced by Agrawal et al. [1], which measures the semantic similarity between documents and a set of seed-words representing violence, to assign a degree for violence.

We also apply machine learning techniques to develop prediction models and investigate whether the violence scores extracted from news articles are helpful in predicting forced migration. A number of regression models are used in our application, such as linear regression, neural networks, and random forest. We first build time-series autoregressive models using the above techniques and then add the extracted violence scores as input to the models. The extracted violence scores are evaluated by measuring the improvement observed by adding them to the prediction models. Our experiments indicate that violence...
Our contributions include:

1. We present a novel application of machine learning and natural language processing techniques to prediction of forced migration based on news articles.

2. We propose novel techniques, ED-FE and TD-FE, to extract violence scores from news articles. The quality of these scores and their effectiveness in improving the accuracy of the prediction models are demonstrated through experiments.

3. We incorporate the violence scores detected from news articles into the forced migration prediction model, and demonstrate through experiments that the detected scores are effective in predicting forced migration.

In the following sections, we first discuss the related work. We then describe in section 3 the details of violence extraction techniques. The experimental settings and datasets are explained in section 4 and the final results about building prediction models based on real-world datasets are reported in section 5.

## 2 RELATED WORK

Here we mainly review the work which is the most related to our research presented in this paper. Schmeidl et. al. [14] developed a theoretical model of refugee migration based on the factors with an estimated magnitude. Some of these factors include economical under-development, human rights violation, ethnic and civil conflicts. These factors were then included in a pooled time-series analysis to predict the number of refugees. This research showed that economic and intervening policy variables are less useful for predicting refugee migration than the threat of violence. This work is different from ours regarding the methodology used for extracting the forced migration signals. Their work uses manually generated scores from various resources for the factors mentioned above, while our automatically extracts the scores of the factors of forced migration from news articles.

Predicting forced migration using news articles remains relatively an under-explored field. Factor extraction from text, on the other hand, has been investigated by Agrawal et al. [1]. They introduced a method to extract the magnitude of violence from news articles. This method uses word embedding techniques to embed the words of news articles and then uses similarity measures within the embedding space to compute the similarity between the words of a document and a set of predefined seed words indicating violence. This work only detects the magnitude of violence from news articles and does not discuss prediction models. The quality of the extracted violence scores considerably depends on the quality of the manually generated set of seed words. Furthermore, the accuracy of the extracted violence scores are measured by computing the correlation between the scores and population movements, and no attempt is made towards prediction models. Unlike this research, we thoroughly investigate the quality of our violence scores by building prediction models and investigate the effectiveness of violence scores in improving the accuracy of the prediction models.

## 3 METHODOLOGY

In this section we provide the details of violence score extraction techniques, SWSW, ED-FE and TD-FE. Also the procedure of building prediction models is explained.

### 3.1 Similarity With Seed Words (SWSW)

Agrawal et al. proposed this technique [1], to detect the magnitude of factors of forced migration from news articles. This method relies on a set of manually generated seed-words representing a specific factor (violence in our case). Some examples of seed words are: violence, conflict, fight, killing, battle, massacre, butchery, injury, bombing and explosion. This technique measures the magnitude of violence by detecting the relevance of the documents to violence over time. The relevance between a document and violence is measured by following three steps:

1. Relevance of a word to violence (violence score for a word) is calculated by measuring the semantic similarity between the mentioned word and every word in the set of seed words and finally taking the average of all the similarity values.

2. Relevance of a document to violence (violence score for a document) is calculated by averaging over all the violence scores of all the words in the document.

3. Violence score for a date is calculated by averaging over all the violence scores for all the documents in that day.

Cosine similarity was used to measure the similarity between two words. Before that, embedding techniques are used to embed the words. We used CBOW model introduced by Mikolov et al. [8, 10], for word embedding.

### 3.2 Event Detection based Factor Extraction (ED-FE)

This technique intends to extract the events from news articles and then distinguish the violent events and use them to define a degree of violence for each date. An event is defined as an incident happening in a specific location involving entities, time and location. The main step in the event detection task is event trigger extraction. The event trigger is a word or phrase in the sentence, which evokes an event and carries the most information about it. For example, in the following sentences: "Extremists clashed soldiers in Iraq on Friday", "clashed" is the event trigger. Event extraction task aims to detect whether or not a sentence includes an event trigger and what the trigger word’s type is.

#### 3.2.1 Training data: The ACE2005 corpus is a publicly accessible dataset which includes labels for the event triggers for a large set of sentences and is used for training purposes. The ACE2005 event detection task defines eight event types, and 33
sub-types\footnote{Event types are: life, movement, transaction, business, conflict, contact, personnel, and justice. For sub-types, please visit https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf} for the whole corpus of sentences and assigns one event type to each event trigger for each sentence. For example, in the sentence “200,000 people start protesting in Pakistan.” Protest is labeled as the trigger word with type Conflict and sub-type Demonstrate.

We used the method introduced by Yang et al. [15] to train an event detection model on the ACE2005 dataset and then apply it to our corpus of news articles to extract events. The approach in [15] jointly extracts events and entities within a document context. The learning problem is decomposed into three tractable sub-problems: learning structures for a single event, learning event-event relations, and learning for entity extraction. Two objective functions are defined to solve the first two sub-problems and L-BFGS is used to optimize the training objectives. For entity extraction, a standard linear-chain Conditional Random Field (CRF) is trained. After all, these learned probabilistic models are integrated into a single model to jointly extract events and entities across a document.

3.2.2 Defining ED-FE Violence Scores: After applying the aforementioned method to our corpus of news articles to extract all event triggers, the event triggers of types injure, die and attack were selected. We call these events the violent events. After that, the violence score is computed by dividing the number of violent events for each date by the total number of events for that date. The degree of violence somewhat depends on the scale of other events happening at the same time. Thus, we put the total number of events as the denominator in the formula, to get a scale of the impact of violence on people.

3.3 Topic Detection based Factor Extraction (TD-FE)

Topic detection approach explores the topical space of the news articles to detect violence related factors. It uses topic modeling methods to analyze a large corpus of news articles to extract coherent topics which are further used to compute violence scores for specific dates.

One important dimension of news articles is time. Articles reporting similar incidents with the same publication date, are most likely to report the same incident. Also, news articles usually continue reporting about one incident during the consecutive days. So, articles with same or close publication dates are most likely to form a single coherent topic about one particular subject.

In order to gather all reports of one single incident into an individual topic, first we need to go through a step of classifying the news articles into time-windows. In this step, the news articles are analyzed and processed according to their time-windows. The length of the time-windows could vary from days to weeks. In this research, we chose the monthly time-windows for the topic analysis.

The main reason to use time-windows for topic analysis is that the topics with short time-period may be obscured by the generalized topics observed in the entire collection. Also, time-window based topic analysis enables identifying granular and short-term topics, as well as generalized and long-term topics.

After dividing the original corpus into time-windows, we generate topic models for each time-window. We primarily generate topic models using LDA [2] and Non-negative Matrix Factorization (NMF) [7]. Matrix factorization is a widely used approach for the analysis of high-dimensional data. The objective of NMF is to extract meaningful features from a set of non-negative sparse vectors. The NMF is successfully applied to different applications, such as image processing, hyper-spectral imaging, and text mining. Here we focus on the property of NMF to identify topics in a given set of documents and classify the documents among the underlying topics.

To identify the optimal number of topics \( k \), we generate topics with the number of topics in the range of \( k \in (10, 50) \) with an increment of two. To find the optimal \( k \), we evaluate the quality of generated topics. One way to do so is to measure topic’s semantic coherence. To calculate the topic coherence score, we use TC-W2V [11] and Unify framework [13]. TC-W2V is a distributional semantics measure introduced by [11]. TC-W2V measure is based on the popular word2vec [9] word embedding technique. In this method, the coherence score is the mean pairwise Cosine similarity of the term vectors generated by Skip-gram model.

Another recent work on topic coherence measure is the unifying framework proposed by [13]. The framework represents the coherence measures as a composition of parts, where the objective is to achieve higher correlation with human judgments. The framework has four segments at its core: segmentation of words subsets, probability estimation, confirmation measure, and aggregation.

We finally select the topic model resulting in the highest topic coherence generated by NMF model based on the unify score. At this step, we have generated and evaluated the coherence of the topic models for each time-window. For each of the topics, we generate a topic-document to represent the topic. Figure 1 demonstrates the topic generation pipeline.

![Figure 1: Topic generation process](image-url)

Topic-documents are generated by ranking the top keywords occurring in the topic according to their probability of appearing in the topic. The number of topics per time window may vary since we are only using the topics with best coherence scores in the range of \( k \).

3.3.1 Defining TD-FE Violence Scores: The next step is to label a topic-document with one of these categories: violence/terrorism, relief, economic issues, political conflicts, refugee crisis and environmental issues. This step can be done automatically using a topic labeling method. But we did it manually with the help of the social scientist to ensure the quality of the topic labels. That is because the quality of the labeled data directly affects the
quality of the forced displacement prediction model. The output of this stage is a set of labeled monthly topic-documents.

The violence score for each month is then defined by the total number of violence/terrorism topics for each month divided by the total number of topics for each month.

3.4 Building Prediction Models

We build two types of prediction models for forced migration: one is purely based on time-series analysis and for the other one we add violence scores as input variables into the time series models.

(1) Pure time-series approach: We build auto-regression models on the time-series of forced displaced populations. Auto-regression is a regression model with lagged variables as input features and it is built based on the assumption that the previous values (i.e., lagged values) of a variable might affect the future values. An auto-regression model is built in this approach using the numbers of refugees in the previous time points as input features and it makes predictions without other external source.

(2) Time-series with violence scores: Violence scores and lagged variables are both used as inputs for the predictive regression models. Comparing this approach with the pure time-series approach allows us to investigate the effectiveness of detected violence scores in improving the accuracy of forced migration prediction.

The regression model in the above two approaches can be built using machine learning models such as Ordinary Linear Regression, Stochastic Gradient Decent (SGD), Support Vector Regression (SVR), Random Forest Regression, or advanced neural network models such as multi-layer Perceptron regression (MLP), Long Short Term Memory (LSTM), or Gated Recurrent Unit (GRU). A brief description of these methods is provided below.

**Ordinary Linear Regression:** The most simple and basic linear model, to which we refer as ordinary linear regression, fits a linear model with coefficients \( w = (w_1, \ldots, w_m) \) to minimize a loss function which is the squared-error loss function in this case. Ordinary linear regression minimizes the sum of squared difference between the actual values in the dataset and the predicted values by the linear approximation.

**Stochastic Gradient Decent:** We use the term SGD in our experiments to refer to a linear regression model with Huber loss function, optimized by stochastic gradient decent. Huber loss function is less sensitive to noise and outliers comparing to the squared-error loss function.

**Support Vector Regression:** Support Vector Machines (SVM) are also used for regression problems while keeping all the characteristics of the algorithm such as the maximal margin. The Support Vector Regression (SVR) is an algorithm for regression problems, with the exact same principles as the SVM for classification. Like SVM, SVR is also based on the assumption that data points are linearly separable and attempts to map a linear function (hyperplane) to the input variables while the error is minimized and the margins are maximized. If data are not linearly separable, kernel functions are used to transfer the data points into a higher dimensional feature space in which linear separation could be possible.

**Random Forest for Regression:** Random forests or random decision forests for regression problems are an ensemble learning method. They build multiple decision trees based on the training set, and produce the outputs by computing the average predicted values of the individual trees [5, 6]. Random decision forests use averaging to overcome the over-fitting problem of the trees on the training set. The loss function used for random forest regression model is Square-error loss.

**Multi-Layer Perceptron:** Multi-layer Perceptron (MLP), also known as Neural Networks, can be used for regression problems. We use square error loss as its loss function.

We use an input window for each of the above methods. A window defines how many time-steps to consider as the input of the model. Window size changes from 1 to 5 in our experiments. If the window size is 1, it means that features belonging to time-step \( t \) (features \( \in [\text{lagged variable, violence scores}] \)) are used as input values for the model to predict future time-steps. If the window size is 5, it means that features belonging to time-steps \( t, t-1, t-2, t-3, t-4, t-5 \) are used as input values. A bigger window size allows the method to consider more information from the past and intuitively make more accurate predictions. On the other hand, a bigger window size increases the dimension of the input vector and thus increases the number of the parameters of the model to be estimated.

**Long Short Term Memory:** LSTMs are the most popular type of neural networks that are more capable of capturing long-term dependencies than typical Recurrent neural networks (RNN). LSTM has the same structure as RNN but differs in terms of calculating the hidden state. Plain RNNs have a very simple structured repeating module. However, the repeating module has a different and more complicated structure in LSTMs. This repeating module in LSTM is called memory unit and it is capable of controlling the information flowing through it. Memory unit keeps coded information from the past and by learning its parameters, the network learns how its memory should behave. In other words, the network can decide how much the historical information can affect the memory or how much impact the memory can have on the final output. This a bigger window size allows the method to consider more information from the past and intuitively make more accurate predictions. On the other hand, a bigger window size increases the dimension of the input vector and thus increases the number of the parameters of the model to be estimated.

**Gated Recurrent Unit:** Gated Recurrent Units (GRU) are very much like LSTMs. They too were created to solve the problem of better capturing long term dependencies. Like LSTM, GRU controls the flow of information too, but it does not use a memory unit. GRU has less complex structure compared to LSTM, resulting in more efficiency in terms of computation.
4 EXPERIMENTAL SETTINGS

4.1 EOS dataset
The Expanded Open Source (EOS) \(^3\) collection is a vast unstructured archive of over 700 million media articles gathered over the years by Georgetown University researchers. The news articles used in this research are automatically filtered by our collaborators in Georgetown University in an effort to select articles that are related to Iraq. The data set we used in the experiments consists of 680,456 news articles spanning from January 2012 to May 2017.

4.2 UNHCR Statistical Population Database
We used the UNHCR\(^4\) Refugee population statistics dataset to capture the number of refugees. This dataset contains data about forced displaced populations from 1999 to 2017 on a monthly basis. Information including the status of the population of concern (refugees, asylum seekers, internally displaced persons, etc.) and the origin and destination of the forced displaced persons is provided in the dataset.

For building prediction models we focus on a subset of the UNHCR dataset related to Iraq from 2012 to 2017, as the EOS dataset we have provides the news articles for this period of time. For all the following experiments, we used the EOS data related to Iraq from 2012 to 2017.

4.3 SWSW violence scores
We applied SWSW on EOS news articles to extract violence scores. 300 dimensional word embeddings were trained on the EOS dataset using Word2Vec\(^5\) with negative sampling.

4.4 ED-FE violence scores
As one news article might inform about more than one event, it is important to focus on the main event which is the most important incident reported in the article. According to \([12]\), we can assume that the main event often occurs in the title and the first sentence of the first paragraph of news articles. Thus, we applied the event detection algorithm on the first sentence of the first paragraph of news articles to extract event triggers and event arguments. Table 1 shows some example sentences from the EOS dataset and the detected events. Detected events are in the form of an event trigger which has an event type, and event arguments. Table 1 shows some example sentences from the EOS dataset and the detected events. Detected events are in the form of an event trigger which has an event type, and event arguments. For violent events, event arguments commonly include the place or victim of the event.

Table 1: Some of the detected events from EOS dataset.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Event type</th>
<th>Event trigger</th>
<th>Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fighting continues in Iraq’s Anbar Province as civilians flee.</td>
<td>Attack</td>
<td>Fighting</td>
<td>Place: Anbar_province, Artifact: civilians</td>
</tr>
<tr>
<td>7 soldiers were killed and wounded by a roadside bomb south of Fallujah.</td>
<td>Die</td>
<td>Event trigger: killed</td>
<td>Arguments: Victim: soldiers, Place: Fallujah, Event trigger: wounded</td>
</tr>
<tr>
<td></td>
<td>Injure</td>
<td>Event trigger: injured</td>
<td>Arguments: Victim: soldiers, Event trigger: wounded</td>
</tr>
</tbody>
</table>

In our EOS dataset, the most frequent event triggers include killed, wounded, injured and clashes, indicating that violent events are dominating the events related to Iraq. Also Fallujah, Ramadi and Anbar are observed among the most frequent words referring to the fact that conflicts were generally happening in these cities.

4.5 TD-FE violence scores
We divided the EOS news articles into a set of sequential non-overlapping time-windows \(\{T_1,...,T_t\}\). Each time-window bin includes ordered documents, sorted by their publication date. We chose the length of time-windows to be one month, as UNHCR provides the information about refugee movements on a monthly basis.

After preprocessing news articles and diving them into time-window bins, topics were extracted for each month and 10 words best representing each topic were provided to two annotators who were asked to label topics in one of these categories: violence/terrorism, relief, economic issues, political conflicts, refugee crisis and environmental issues. The annotators labeled each topic separately, and then aggregate their results into the final labels to improve the quality of the labels.

We observed that the NMF topic modeling shows improvements in the topic modeling coherence scores. This result is in agreements with the results discovered in \([4]\). Another significant advantage of the NMF, compared to the probabilistic approaches, is the speed of finding topics. The matrix factorization tends to be faster than its counterpart probabilistic based approaches. Table 2 demonstrates labeled topics extracted from the EOS dataset. To calculate the TD-FE violence scores for a month, the number of violence-related topics in the month is divided by the total number of topics in that month.

Table 2: Topics and keyword representation.

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top ten words sorted by the probability of the word in the topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence/Terrorism</td>
<td>killed, Baghdad, wound, car, attack, bomb, people, suicide, security</td>
</tr>
<tr>
<td>Refugee Crisis</td>
<td>refugee, child, million, Jordan, UNHCR, migrant, humanitarian, people, flee, oil, barrel, export, crude, company, market, energy, price, sanction, say</td>
</tr>
<tr>
<td>Economical Issues</td>
<td>oil, Istanbul, round, Baghdad, Iran, meeting, Jalili, negotiation, pet, Tehran</td>
</tr>
<tr>
<td>Political Issues</td>
<td>oil, Baghdad, round, Baghdad, Iran, meeting, Jalili, negotiation, pet, Tehran</td>
</tr>
</tbody>
</table>

4.6 Experimental Settings for Prediction Models

4.6.1 Baseline: We compare our prediction models described in Section 3.4 to each other and also to a baseline model. The baseline model for predicting forced migration is built using the persistence algorithm (also called naive forecast). The persistence algorithm predicts the value for the future time-step \(t+1\) with the exact value seen at the previous time-step \(t\). For example, the value in January will be used as the predicted value for February.

4.6.2 Prediction Settings: We compare the two types of the prediction models in section 3.4: pure time series regression models and the time series models with the extracted violence scores. In addition, we evaluate each prediction model in three Settings:

- Setting 1: Predicting value in time \(t+1\)
- Setting 2: Predicting value in time \(t+2\)
- Setting 3: Predicting value in time \(t+3\).

where the input to the model contains variables up to time \(t\). This is because intuitively predicting \(t+2\) is harder than \(t+1\), and \(t+3\) is harder than \(t+2\) and so on. We would like to examine...
our prediction models in more difficult settings to be able to make sound conclusions about their prediction quality. Also, more time-steps a model can accurately predict in the future, more useful it will be for early warning systems in terms of preparing for refugee crisis. For example, the value of a model which can accurately (i.e., with an acceptable error rate) predict t+4, allows four months for governments to prepare for refugee crisis.

The UNHCR dataset is split into train set and test set. The training set includes 80% of the data (January 2012 - April 2016) and the test set includes the last 20% observations in the UNHCR dataset (May 2016 - May 2017). The features used when predicting (t+n) time-step, all belong to time-steps before (t) and no information related to the time period of (t) until (t+n) is used, which makes predicting (t+n) harder as n increases.

Root-Mean-Square Error (RMSE) on the test data is reported separately for each Setting. A walk-forward testing is performed when evaluating the prediction results. Each time-step in the test set will be given to the model one at a time, the model predicts a value for the given time-step and then the actual value for that time-step will be accessible to the model to make the next predictions based on it.

5 RESULTS AND EVALUATION

5.1 Violence Scores Extraction

Figure 2 shows the extracted violence scores from the EOS dataset using SWSW, ED-FE and TD-FE techniques. The increasing trend of violence over time is visible for all three sets of scores. The rapid fluctuations in violence scores during early 2015 is due to the missing articles in the EOS dataset during this time.

![Figure 2: Extracted violence scores from EOS dataset using SWSW, ED-FE and TD-FE.](image)

5.2 Prediction Models

To evaluate the effectiveness and quality of the extracted violence scores and to compare them against each other, we compare the predictive errors of the two types of regression models described in section (3.4). If the error decreases when violence scores are added, it can be concluded that the added violence score is an effective element in predicting forced migration. In the following sections, we have tested each set of violence scores individually with 7 different regression models, with the following configurations:

- (1) Random Forest Tree: Number of estimators= 450
- (2) Support Vector Regression (SVR): Kernel= Sigmoid
- (3) MLP Regressor: Learning rate=0.00005, Optimizer= ADAM, Hidden layers=(5,5)
- (4) LSTM: Internal state size=30, Optimizer= ADAM
- (5) GRU: Internal state size=35, Optimizer= ADAM
- (6) Baseline

Table 3 shows the RMSE of the baseline as well as all the regression models in the pure time-series setting. Table4 shows the RMSE of regression models using time series with violence scores. (previously described in section (3.4)). We treated the window size as a parameter to be tuned individually for each model and the tables report the best outcome of each model. The average error over the three settings is reported for each regression model. The decrease in average error with respect to the pure time series approach is marked in table 4 with down arrows.

Table 5 shows the average RMSE of all seven regression models using different violence scores achieved by SWSW, ED-FE and TD-FE techniques. On average, ED-FE and TD-FE violence scores are both showing better performance for all t+1, t+2 and t+3, comparing to SWSW. To check weather the violence scores are able to improve the error of regression models significantly or not, we ran Wilcoxon signed-rank test separately for SWSW, ED-FE and TD-FE violence scores. Wilcoxon signed-rank test tests the null hypothesis that the predictions using the pure time-series approach and the approach with added violence scores come from the same distribution. The test was run three times, between two approaches, with the second approach (i.e., the one with violence scores) using a different set of violence scores (SWSW, ED-FE and TD-FE violence scores) each time. The test’s results indicate that ED-FE and TD-FE violence scores have improved the prediction error significantly (rejecting the null hypothesis with p-value=0.0007 and p-value=0.03 respectively). However, SWSW violence scores...
Table 4: The RMSE of regression models with different violence scores as input features.

<table>
<thead>
<tr>
<th>Regression Models</th>
<th>Other input features besides lagged variables</th>
<th>Predicted time-step</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>t+1</td>
</tr>
<tr>
<td>Ordinary linear regression</td>
<td>SWSW violence</td>
<td>2154</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>2121</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2135</td>
</tr>
<tr>
<td>MLP</td>
<td>SWSW violence</td>
<td>2171</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>1985</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2206</td>
</tr>
<tr>
<td>Random forest</td>
<td>SWSW violence</td>
<td>2398</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>2382</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2553</td>
</tr>
<tr>
<td>SGD</td>
<td>SWSW violence</td>
<td>2094</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>2216</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2410</td>
</tr>
<tr>
<td>SVR</td>
<td>SWSW violence</td>
<td>2094</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>2352</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2510</td>
</tr>
<tr>
<td>LSTM</td>
<td>SWSW violence</td>
<td>3416</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>3202</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>5772</td>
</tr>
<tr>
<td>GRU</td>
<td>SWSW violence</td>
<td>2419</td>
</tr>
<tr>
<td></td>
<td>ED-FE violence</td>
<td>2255</td>
</tr>
<tr>
<td></td>
<td>TD-FE violence</td>
<td>2416</td>
</tr>
</tbody>
</table>

Table 5: The average RMSE of all seven regression models using different violence scores achieved by SWSW, ED-FE and TD-FE techniques.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Predicted time-step</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t+1</td>
</tr>
<tr>
<td>SWSW violence scores, lagged variables</td>
<td>2392</td>
</tr>
<tr>
<td>ED-FE violence scores, lagged variables</td>
<td>2359</td>
</tr>
<tr>
<td>TD-FE violence scores, lagged variables</td>
<td>2828</td>
</tr>
</tbody>
</table>

were not able to significantly improve the prediction models’ error (p-value=0.12).

After all, MLP-regressor with feature set = {Lagged variable, ED-FE-violent scores} was selected as the final model according to the average error over all the Settings (Average (t+1,...,t+3)). This improvement of the final model over the baseline is most visible for Setting 3 (predicting t+3), where the RMSE of the baseline is almost twice the RMSE of our final model.

Figure 3 shows the performance of the final model on the UNCHR dataset, for Setting 1 (predicting t+1). The error rates on both training and test sets are plotted as well as the actual values of the UNCHR dataset.

5.3 Analysis and Discussions

The best prediction model was built using a feature set including Lagged variables and ED-FE-violence scores. According to these results, we conclude that among all the proposed methods for measuring violence, ED-FE outperforms the others. We believe that this is because ED-FE depends on detecting events which are more focused and detailed comparing to topics that the TD-FE violence detection method is based on. Also, SWSW violence detection relies on a predefined set of seed words and the quality of SWSW violence scores directly depends on the quality of these seed words, while ED-FE does not have this disadvantage. Furthermore, the coverage of a violent incident by news agencies could be a good metric for measuring the degree of how violent that incident is. The bigger the size of the incident and its consequences, the more the number of
news articles reporting it. ED-FE violence detection takes this information into account, while TD-FE completely ignores this material by gathering all the reports of a single incident into one unique topic.

It is important to note that going from Setting 1 to Setting 3, the RMSE of the baseline increases gradually. This means that as we try to predict further time-steps in the future, the accuracy of the model decreases. On the other hand, the performance of our final model almost stays the same as we predict further time-steps in the future. This could be explained by the fact that when predicting t+1, the lagged variable is probably the dominant feature, according to the very high auto-correlation of the UNHCR dataset for lagged variable=1 (91%). That is why our model and the baseline almost have similar performance for Setting 1. But as we intend to predict further time-steps in the future (e.g. t+3), the autocorrelation decreases and the model can no longer solely rely on lagged variables as the most effective features. Also, the model faces more missing information and longer gaps, which means that it needs to rely on other features to provide it with extra content to cover for the missing information (due to the gap between the present time-step (t), and the predicted time-step (t+3)). This is why the performance of our final model almost stays the same for all the Settings, while the error of the baseline increases gradually as we predict further time-steps in the future.

These results indicate that ED-FE violence scores are to some extent capable of providing the model with enough supporting information to make accurate predictions for future time-steps when we are facing gaps and missing information. Our intuitive justification for this incident is that our extracted violence scores are representing some sort of triggers for forced migration, meaning that they provide us with early signals of refugee movements. In other words, there is some gap in time between when these triggers take place and when refugee movements actually happen.

Moreover, the average RMSE for our final model is 2.205, which is within the tolerance range, considering that UNHCR time series has a mean of 7.644 and standard deviation of 6.882. After all, the final model was created by MLP-regressor with a window of size 4. The reason that MLP-regressor outperformed which is within the tolerance range, considering that UNHCR when these triggers take place and when refugee movements. In other words, there is some gap in time between when these triggers take place and when refugee movements actually happen. Moreover, the EOS dataset has many missing articles during the first six months of 2015, resulting in negative impact on the quality of our model. This is observable in the sudden drop down in extracted violence scores during 2015 (Figure 2). We believe that violence extraction methods will show more accurate results using a dataset with better coverage.

6 CONCLUSIONS

In this paper, we presented a novel application of machine learning and natural language processing techniques to predicting forced migration based on news articles. We proposed two novel techniques called ED-FE and TD-FE for processing and analyzing news articles to extract the violence scores based on event detection and topic detection, respectively. These scores are used to build a set of prediction models for forecasting future numbers of forced displaced people. We made comparisons between ED-FE, TD-FE and a state-of-the-art method called SWSW. Experiments demonstrate that both ED-FE and TD-FE outperform SWSW. Moreover, ED-FE was identified as the most effective technique for extracting violence scores among the three methods.

Our final prediction model, which was built using ED-FE violence scores and lagged variables as input features, outperformed the baseline with a considerable margin. Results show that adding ED-FE violence scores to the input features of prediction models significantly improves the prediction accuracy indicating that we can rely on violence scores detected from news articles as a useful factor for predicting forced displacement. As part of the future work, this research can be extended to extracting other factors or signals of forced migration from news articles, such as economic instability, environmental threats or negative emotions.

ACKNOWLEDGMENTS

We would like to thank Dr. Aijun An, Jimmy Huang, Susan McGrath and Lisa Singh and for her collaboration in this research.

REFERENCES


