Detecting the Magnitude of Events from News Articles

Ameeta Agrawal, Raghavender Sahdev, Heidar Davoudi, Forouq Khonsari, Aijun An and Susan McGrath*
Department of Electrical Engineering and Computer Science,
*School of Social Work,
York University, Toronto, Canada
{ameeta,sahdev,davoudi,khonsari,aan}@cse.yorku.ca,smcgrath@yorku.ca

***Abstract*—Forced migration is increasingly becoming a global issue of concern. In this paper, we present an effective model of targeted event detection, as an essential step towards the forced migration detection problem. To date, most of the the approaches deal with the event detection in a general setting with the main objective of detecting the presence or onset of an event. However, we focus on analyzing the magnitude of a given event from a collection of text documents such as news articles from multiple sources. We use *violence* as an illustration as it is one of the most critical factors of forced migration. The recent advancements in semantic similarity measures are adopted to obtain relevant violence scores for each word in the vocabulary of news articles in an unsupervised manner. The resulting scores are then used to compute the average daily violence scores over a period of three months. Evaluation of the proposed model against a manually annotated data set yields a Pearson's correlation of 0.8. We also include a case study exploring the relationship between violence and key events.**

***Keywords*—event magnitude detection, semantic similarity, word embedding**

I. Introduction

Forced migration exists when a significant number of people are displaced from their homes by a socio-political conflict, natural or human-made disaster, economic disturbance, disease, and so on. It is increasingly becoming a major topic of concern for not only the affected regions but the whole world. For instance, the refugee crisis currently unfolding in Syria is making daily news headlines worldwide where according to the latest statistics, almost 13.5 million people need humanitarian assistance due to a violent civil war [1], [2], [3]. Since the Syrian war began, 320,000 people have been killed, and about 1.5 million people have been wounded or permanently disabled, as stated by the Syrian Observatory for Human Rights [4]. It is, therefore, highly relevant to develop a model for predicting forced migration.

In order to develop such a prediction model, important factors affecting forced migration need to be identified and analyzed. There are many factors that can affect forced migration; violence is one of the most obvious and critical ones. In a country undergoing civil unrest, violent events occur all the time. Assigning a binary indicator such as violent or non-violent to events is likely to return inadequate results. The degree of violence is what matters for people in making a decision on whether to move. In this paper, we introduce the task of targeted event analysis. Unlike previous event detection techniques, which aim to detect generic new or changing events, our focus is on identifying the magnitude of a targeted event such as violence over time. This, we believe, is a preliminary step in the direction of solving the larger challenging problem of predicting forced migration based on the magnitudes of its events. The result of event magnitude detection can also be used to analyze the correlations between different events.

In this paper we seek to develop a model to efficiently track the magnitude of a target event, such as 'violence' from a corpus of news articles. Although much of the recent research has focused on extracting events from microblogs such as Twitter data [5], [6], [7], this dataset is notorious for being constantly evolving, inconsistent and quite noisy as it often originates from different sources. These characteristics make it an unfavorable choice for studying long-term variations in one factor consistently. Since our primary goal is to develop an effective and sustainable prediction model of forced migration, we choose to use traditional broadcast media, i.e., newspaper articles, which have plenty of advantages such as originating from a few handful of sources, being detailed while also sufficiently succinct, frequent and generally timely.

Our work is part of a larger project for developing a model for predicting forced migration based on big data. In this paper, we present an effective approach to computing the magnitude of an event, such as violence, over a period of time. First, we compute the similarity between words in a document and domain seed words using some measure of semantic similarity. These scores are then used to obtain an aggregate violence score to plot a time-series chart. A ground truth data set built manually by domain experts serves as the evaluation benchmark.

Our main contributions include:

- introducing the task of targeted event detection and proposing an efficient model for computing the magnitude of a target event or factor from news articles in an unsupervised setting;
- evaluating the proposed approach on a manually annotated data set to determine the effectiveness of the

proposed approach as well as the measure of semantic similarity most suitable for a task such as ours;
- and, presenting an interesting case study exploring the relationship between violence and key events.

The rest of this paper is organized as follows: A summary of the relevant work carried out in this field is presented in Section II. We then describe the details of our algorithm for computing the event scores in Section III, while Section IV outlines the models of semantic similarity that are employed in our proposed approach. Section V discusses our experimental setup and the analysis of the results. In Section VI, we present a case study exploring the interesting relationship between violence and key events. Finally, Section VII wraps it up with the conclusions and possible avenues of future work.

II. Related Work

Although event detection has been widely researched in various forms and contexts, most studies aim to detect the presence or onset of a new event [8], [9]. In this paper, we study the problem of targeted event detection, which focuses on specific events and their degree of continuous change over a period of time. We briefly discuss some background work in the context of event detection first before focusing on the area of targeted event detection.

Event detection aims to study the presence or onset of any new emerging event. Li et al. [6] build a model, called TEDAS (Twitter based Event Detection and Analysis System), to detect new events and analyze the spatial and temporal patterns of those events. Stuetzle et al. [10] introduce the problem of finding interesting rare events from a data set which has occasional interesting events. The above-mentioned works do not detect a specific event or the degree of its change. Instead, they detect the existence of any new event within a document stream. Hua et al. [7] detect and visualize events from Twitter in a semi-supervised manner using transfer learning and label propagation to automatically label data. Then they learn a customized text classifier based on mini-clustering, and finally apply fast spatial scan statistics to estimate locations of the events. Chakrabarti and Punera [5] use Hidden Markov Models for summarizing event tweets by learning the hidden state representations of the events.

Wang and McCallum [11] analyse the trends of a particular topic over time using an LDA-based topic model. They present a model which jointly models both word co-occurrences and localization in continuous time. Mathioudakis and Koudas [12] present a system that performs trend detection using the twitter data. Their architecture does trend detection by detecting and grouping Bursty key words. Finally trends are analyzed by employing context extraction algorithms (PCA, SVD, etc.) over the recent history of the trend.

Much recent work in event detection is designed to work with short snippets of text, mostly extracted from twitter data [13], [14], which is assumed to discuss a single topic in a tweet. A news article, on the other hand, tends to be longer in length, thus presenting its own set of challenges such as containing multiple words which may not be related to the overall theme of the article. For example, consider the following snippet from an article:

> "*Iraq's government on Sunday revoked the operating licenses of Al Jazeera and nine other television channels, saying that they were inciting sectarian conflict. All but one of the channels are aligned with Sunni financial backers, and the move was widely perceived as a crackdown on dissent by the Shiite-led government that is facing an increasingly violent Sunni uprising. The decision will not banish the channels from the airwaves: as satellite channels based abroad, they are beyond the reach of the Iraqi government. But it prohibits the channels journalists from reporting inside Iraq*."

While one may agree that this article is predominantly reporting a non-violent news event, it does include a few words (*conflict, crackdown, violent, uprising*) that could be considered as related to violence. The challenge lies in accurately understanding that the degree of violence in this article is extremely low, if not non-existent.

In the context of online news, Allan et al. [8], Yang et al. [15], and Brants et al. [9] propose methods for analyzing events in online news articles. They adopt the *tf-idf* model for comparing each new article with all the previously detected events. If the similarity between the two sets falls under a certain threshold, the document is regarded as discussing a new event. However, all these approaches aim at detecting new or unseen events.

Target event detection usually focuses on a specific event and analyzes its patterns over time. For instance, models of detecting earthquakes [16] and civil unrest [17] have been previously proposed. These targeted approaches typically begin with a keyword vocabulary collected by domain experts, filtering the raw corpus with the domain vocabulary, and using the surge of the number of retained documents in a time window to signify occurring of the target event. Similar to these works, we also begin by compiling a list of domain specific seed words. However, we further employ those seed words to compute semantic similarity with document words, essentially expanding the coverage of the target event, and using that metric of similarity as a proxy for obtaining the magnitude of change.

III. Proposed Approach

Given a contiguous time frame $T = \{t_1, t_2, ..., t_z\}$ and a collection of news articles $\mathbb{D}$ for each date t_s, where each article $d_k \in \mathbb{D}$ is composed of a set of words W, the task of event magnitude detection is to compute the magnitude of an event E on t_s, where magnitude is defined as a value within a fixed range indicating the high or low of the event. Technically, E can be any target event such as violence, humanitarian assistance, disease outbreak and so on, and the magnitude can be a value between $[-1, 1]$ but it can be normalized into other ranges such as $[0, 5]$. Our proposed approach is particularly suitable for an event whose magnitude can be expressed using a set of seed words signalling a *high* magnitude. For instance,

Algorithm 1 Pseudocode of the proposed approach

Input:
A series of dates in a window of time: $T = \{t_1, t_2, ..., t_z\}$
A collection of documents for $t_s \in T$: $\mathbb{D}$
A set of words for $d_k \in \mathbb{D}$: $W = \{w_1, w_2, ..., w_n\}$
A target event: E
A set of m seed words for event E: $Y = \{y_1, y_2, ..., y_m\}$

Output:
A set of event scores ω for T

Procedure:
1. **for** $t_s \in T$ **do**
2. **for** $d_k \in \mathbb{D}$ **do**
3. **for** $w_i \in d_k$ **do**
4. **for** $y_j \in Y$ **do**
5. Similarity w_i and y_j = $Sim\,(w_i, y_j)$
6. Similarity(w_i, E) = average $Sim\,(w_i, y_j)$
7. EventScore d_k = average Similarity(w_i, E)
8. Append average EventScore d_k to ω
9. **return** ω

an essential factor such as disease outbreak can be represented by a few seed words such as infection, epidemic, etc.

The first step includes identifying appropriate seed words. With the help of domain experts, we manually generated a list of seed words related to various events, a sample of which is presented in Table I.

For the purpose of illustration, we choose E to be *violence*. Let $W = \{w_1, w_2, ..., w_n\}$ be a set of n words in an article d_k, where $W \subset d_k$ and $Y = \{y_1, y_2, ..., y_m\}$ be a set of m relevant seed words representing a factor E.

Then, we compute the association between each word w_i of the vocabulary and each seed word y_j representing an event, and average it over all the seed words in Y, as follows:

$$Sim\,(w_i, E) = \frac{1}{\mid Y \mid} \sum_{j=1}^{m} Sim\,(w_i, y_j) \tag{1}$$

to obtain the average word similarity between w_i and E. This ensures that for a word to be considered strongly associated to an event, it needs to have sufficiently high similarity with all the seed words in Y. For *violence* we use the set of seed words described in Table I. To obtain the semantic similarity $Sim\,(w_i, y_j)$ between any two given words, we apply the following three methods (further described in the next section):

1) Normalized Pointwise Mutual Information (NPMI)
2) Continuous Bag-of-Words (CBOW)
3) Skip-Gram (SG).

After computing the similarity $Sim\,(w_i, E)$ between each word in the vocabulary and an event, each news article is assigned an event score by averaging the similarity scores of all the words in that document. For instance, the event score of an article d_k consisting of n words is computed as:

$$EventScore\,(d_k, E) = \frac{1}{n} \sum_{i=1}^{n} Sim\,(w_i, E) \tag{2}$$

Next, to compute the event score for a day t_s, we average the $EventScore\,(d_k, E)$ for all the news articles in set $\mathbb{D}$ for t_s as:

$$EventScore\,(t_s, E) = \frac{1}{\mid \mathbb{D} \mid} \sum EventScore\,(d_k, E) \tag{3}$$

Finally, we plot the trend of the targeted event against days to analyze the change in event over a period of time. Algorithm 1 summarizes the pseudocode of our approach.

IV. Word Semantic Similarity Measures

Semantic similarity measures can be very helpful for classification or organizing words in a language by their semantic relations. Many models of computing word semantic similarity exist, including the more traditional count-based methods such as Pointwise Mutual Information (PMI) to the more recent neural-network-inspired models of word embeddings such as Continuous Bag of Words (CBOW). Although the models are different in the algorithms used, they are fundamentally based on the principle that a word's meaning can be induced by observing its statistical usage across a large sample of language. We describe the models of word similarity that we employ in our approach here.

A. Count-based measures

A popular measure of obtaining word association is Pointwise Mutual Information (PMI) [18]. PMI is defined as the log ratio between two words x and y's joint probability and the product of their marginal probabilities, which can be estimated by:

$$PMI\,(x, y) = \log \frac{p\,(x, y)}{p\,(x)\,p\,(y)} \tag{4}$$

PMI is inspired by the intuition that words that occur closer together are more related. A higher value means the words are more correlated in the same context in the training corpus used. The maximum value of this measure is determined by the minimum value between $-\log\,p\,(x)$ and $-\log\,p\,(y)$ and the minimum value of PMI, which happens when the number of co-occurrences of two words is zero, is $-\infty$.

A well-known shortcoming of PMI is that low frequency events receive relatively high scores. Furthermore, the lack of a fixed upper bound of PMI makes it rather nonintuitive for an association measure. To overcome these shortcomings, Bouma [19] proposed a normalized version of PMI (NPMI), where:

$$NPMI\,(x, y) = \frac{\left(\log \frac{p(x,y)}{p(x)p(y)}\right)}{-\log\,p\,(x, y)} \tag{5}$$

with fixed orientation values as follows: when two words only occur together, NPMI$(x, y) = 1$; when they are distributed as expected under independence, NPMI$(x, y) = 0$; and, when two words occur separately but not together, NPMI$(x, y) = -1$.

TABLE I
SEED WORDS FOR VARIOUS EVENTS

Event	Seed Words
VIOLENCE	war, violence, violent, conflict, fight, kill, battle, massacre, injury, butcher, explosion, bomb, corpse, abduction, ambush, suicide, rape, persecution, assassination, terror, military, attack, assault, gang, crime, theft, clash, mortar, rocket, siege, blockade, shells, force, gun, soldiers, rebel, unrest, troubled, insurgent, revolutionary, dead, crisis, victim, detention, arrest, detainee, prison, indiscriminate, checkpoint, operation, torture, elements, execute, withdrawal, feud, weapons, kidnapping, ransom, ammunition
DISEASE	disaster, outbreak, infectious, diseases, epidemic, contagious, donor, vaccination, polio, cholera, operation, malnutrition, medical, virus, death, immunisation, medicine, dehydration
POLITICS	province, government, authority, mayor, election, rule, law, registration, administration, reform, protest, demonstration, vote, minister, president, corruption, security, coalition, opposition, parliament, policy

B. *Word embedding*

Word embedding is used to refer a group of algorithms used to learn a representation of words in a multidimensional continuous space. Basically, the learned representations come from hidden layers or intermediate weight matrices in a non-linear model. More recently, Mikolov et al. [20], [21] introduced two neural network-inspired models, popularized by the `word2vec`[1] program, to compute word vector representations, also called word embeddings, from large text corpora.

The underlying model is a two layer neural network which takes a text corpus as its input and outputs a set of feature vectors for all the words in that corpus. When the feature vector assigned to a word cannot be used to accurately predict that word's context, the components of the vector are adjusted. Each word's context in the corpus is the teacher sending error signals back to adjust the feature vector. The vectors of words judged similar by their context are nudged closer together by adjusting their numbers in the vector. For example, if a word is said to be encoded as a 300 dimensional vector, it means that 300 numbers are used to describe its location in a continuous multi-dimensional space. Following shows the first five vectors for representation of a sample word, *polluted*, trained from our corpus:

$$\begin{array}{ll} 0.045237543419003486 & 0.0184920988976955 \\ -0.16593559086322784 & 0.0819124802947044 \\ -0.24064932763576508 & 0.5136263370513916 \end{array}$$

Semantically similar word vectors would thus appear in closer vicinity of each other, such as indicated in an example Figure 1. This property is particularly useful in a task such as ours where we need to extract not only the similar words but also obtain a measure of their similarity.

The Continuous Bag-of-Words model (CBOW) ties together the tokens of each context window by representing the context vector as the sum of its words' vectors. It is thus more expressive than the other methods, and has the potential of deriving better word representations. The other model, Skip-Gram (SG), has been known to provide better results for infrequent words on various linguistic tasks. Essentially, both these models are inspired from the feed forward and recurrent neural networks, with the non-linear hidden layer removed

[1]https://code.google.com/archive/p/word2vec/

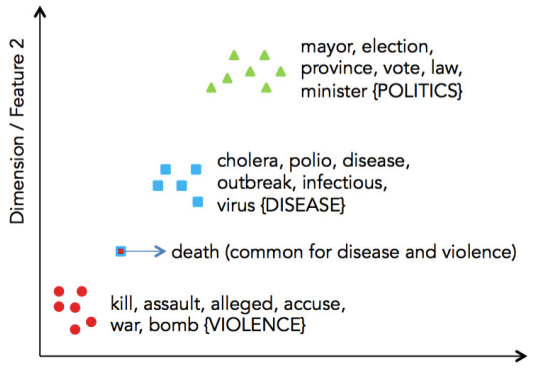

Fig. 1. Sample 2D dimensional vector representation showing closer words appearing in closer vicinity

in order to minimize the computational complexity. The SG model predicts the surrounding context given a target word, whereas the CBOW model relies on the context for predicting a target word.

Unlike NPMI, which outputs the similarity between two words directly, the word embedding models such as CBOW and SG output vector representations of each word in the vocabulary. The similarity between two word representations can be calculated using the cosine similarity within a fixed range of -1 and 1, similar to NPMI.

Table II shows a sample of words from the news articles and their similarity with the factor 'violence', in decreasing order, as computed by various approaches. The word *kill* topped the most violent list for all three measures.

TABLE II
WORDS AND THEIR SIMILARITY WITH 'VIOLENCE'

	NPMI	CBOW	SG
kill	0.299	0.335	0.425
victim	0.240	0.248	0.364
anger	0.164	0.216	0.298
love	0.109	0.045	0.226
spectacular	0.104	0.004	0.148

Furthermore, since all the three models, NPMI, CBOW and SG, output the similarity score between two words within a

neat range of -1 and 1, they are appropriate as a metric for computing the degree of an event. In this paper, we also aim to determine the most suitable model of semantic similarity for a task such as ours. For instance, NPMI, which is based on simple statistics of word frequencies and co-occurrences, yields a simpler model which takes less time and resources to train, while, being more scalable and benefits from using training data with bigger size. On the other hand, although the complex model of word embeddings provides much more information in terms of how similar two words are, it requires considerably longer time and more resources.

V. Evaluation and Analysis

By comparing the results of the proposed approaches against those obtained using baselines and manually annotated data, the effectiveness of our method can be evaluated. We start by describing the process of creating a benchmark annotated data set, then discuss the details of training the algorithms, and finally, present the evaluation analysis.

A. Creating evaluation data set

In this project we consider the Expandable Open Source (EOS) data set, kindly shared by Georgetown University. This archive contains more than 700 million open source media articles, scraped from thousands of local, regional, national and international sources, including mainstream media channels as well as government websites, in 46 languages. Since 2006, it has been actively adding approximately 300,000 articles per day. Although the corpus includes news articles from various countries around the world, we extract only a small subset of it, specifically 367 news articles spanning June to August 2013, those published by Iraqi news agencies. Iraq is chosen as a representative location because it exhibited similar periods of civil unrest prior to the Syrian crisis.

To create the ground truth data set, three annotators manually labelled each news article with a score between 0 and 5, with 0 indicating least or zero violence and 5 indicating extreme violence. The average of the three annotations makes up the final ground truth score for each article.

A sample of labeled articles is shown in Table III, while a frequency word cloud highlighting the most common words in the articles rated as 5 (most violent) and 0 (least violent) is presented in Figure 2.

We measure the inter-annotator agreement in terms of Pearson's correlation, which gives a value between $[-1, 1]$, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. We choose Pearson's correlation as it is suitable for three annotators and ordinal data such as our violence scores in the range of 0 to 5. The inter-annotator agreement between the three annotators stood at $r = 0.793$.

B. Training corpus

A large publicly available text corpus such as Wikipedia[2] is a good resource for deriving word similarity scores. Before using the corpus, the text is cleaned by applying the following pre-processing steps: a) conversion to lowercase b) removal of stopwords and numbers c) stemming, and d) remove tokens occurring less than five times in the corpus. The resulting corpus contains approximately 918 million word tokens in total. The same corpus is used for training all the similarity measures.

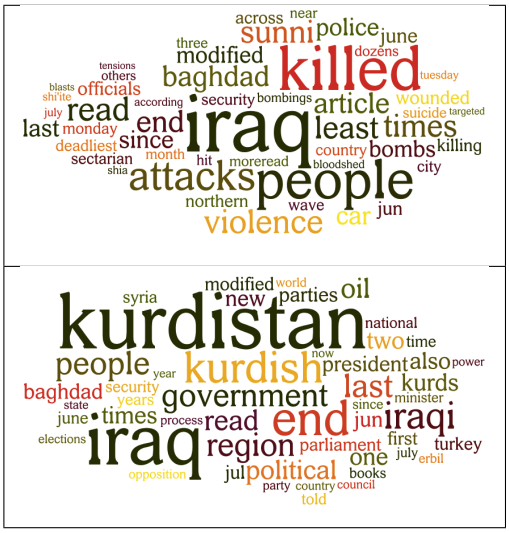

Fig. 2. Word clouds of violent (top) and non-violent (bottom) articles

Some of the parameter settings used for training the algorithms include:

- Word embedding (CBOW and SG): *context window size* = `5`; *dimension size of feature vectors* = `300`; *negative sampling* = `5`;
- PMI: *context window size* = `5`;

Note that the settings chosen for training the word embedding models are the default suggested settings by the `word2vec` model's developers and have known to yield good results [22].

C. Validation

We evaluated our proposed approach using three measures of semantic similarity on the manually annotated evaluation data set. We implemented a simple keyword matching method as a baseline for comparison, where if a word in the article exists in the seed word set, its similarity score is 1; otherwise the score is 0. Pearson's correlation coefficient is used to calculate the correlation between the automatically obtained scores of violence and those assigned manually. The results of evaluating 367 news articles spanning June to August 2013 are summarized in Table IV.

The results indicate that the trend of the degree of violence obtained from our proposed approach strongly correlate with that generated by the domain experts. Moreover, while all the three methods of semantic similarity, NPMI, CBOW and SG, outperform the baseline by a considerable margin, there is little difference between the three themselves. These results are in line with some other recent studies [23], [22] where the simple PMI measure has been known to outperform several other complex algorithms on certain evaluation tasks.

[2]http://download.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

TABLE III
SNIPPETS OF NEW ARTICLES LABELED WITH VIOLENCE SCORE

News article snippet	Violence score
A string of apparently coordinated bombings and a shooting across Iraq on Sunday killed at least 51 and wounded dozens, in a wave of violence that...	5
The United Nations welcomed the launch of the National Environment Strategy and Action Plan in Baghdad today on the occasion of the World Day to Combat Desertification.	0
Fresh statistics point to conflict in Syria as a major new factor in global displacement. More than half the world's refugees came from five countries: Afghanistan, Somalia, Iraq, Syria and Sudan.	1
One of the deadliest attacks came at night when two bombs placed near a market blew up less than a minute apart...	4

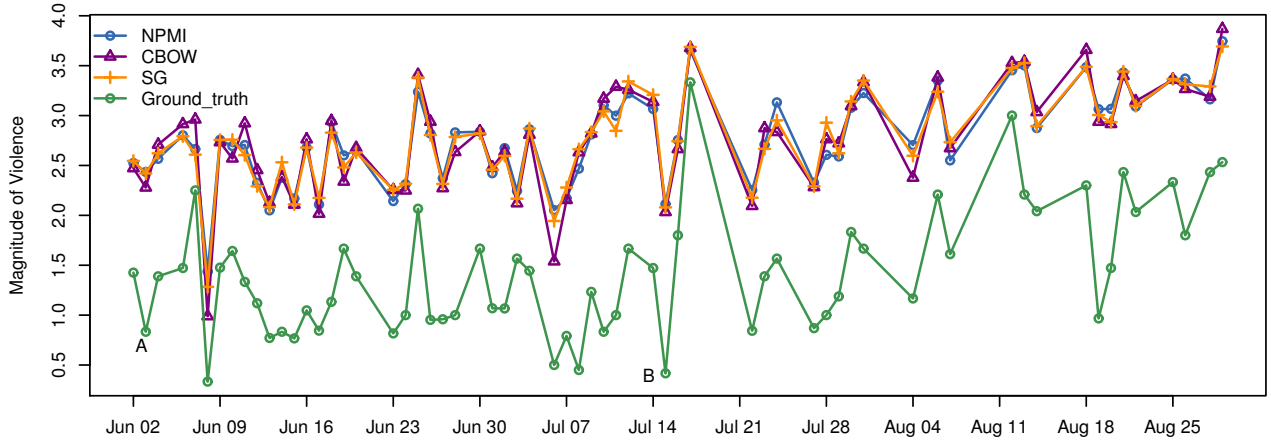

Fig. 3. Magnitude chart spanning June to August 2013

TABLE IV
RESULTS OF OUR PROPOSED APPROACH BY APPLYING THREE DIFFERENT METHODS OF SEMANTIC SIMILARITY

Method	Pearson's correlation
Baseline keyword	0.682
Normalized PMI	0.800
CBOW	0.777
Skip-gram	0.794
Inter-annotator agreement	0.793

The violence scores are plotted as time series data in Figure 3. For comparability, the scores that are output by our approach are converted to a $[0, 5]$ scale so as to be within the same range as the ground truth scores. On further examination of the time series data, it is observed that two data points A and B (indicated on the Figure 3) align with two key events of interest, especially related to forced migration, that were extracted by the domain experts during the same time window. For instance, severe dust storms covered various parts of the country around time A, while event B co-occurs with the Republic day celebration.

To summarize, the experimental results indicate the potential of our proposed approach in effectively detecting the magnitude of violence from news articles. Due to the unsupervised nature of our algorithm, it can be easily generalized and is applicable to computing the trend of any event such as infectious diseases, environmental threat, etc. The biggest limitation of our proposed algorithm includes relying on the domain knowledge for extracting a good quality set of seed words and the various hyper-parameter settings of the individual semantic similarity methods that could affect the performance of the overall algorithm.

VI. CASE STUDY: VIOLENCE AND KEY EVENTS

Encouraged by the strong correlation of results obtained using automatic methods with the human annotated ground truth data in the previous section, we further perform a case study on 2,789 news articles spanning six months (September 2013 to February 2014) with the publication location indicating Iraq, for which we have a corresponding timeline of events compiled by the group of domain experts. As per a United Nations report [24], 2013 was one of the deadliest years for the country since 2008.

An interesting observation made during this case study, as shown in Figure 4, is that the degree of violence seems to hint at a subtle correlation with some cultural and environmental events such as festivals, floods and even an earthquake. For example, observe the noticeable dip in the degree of violence

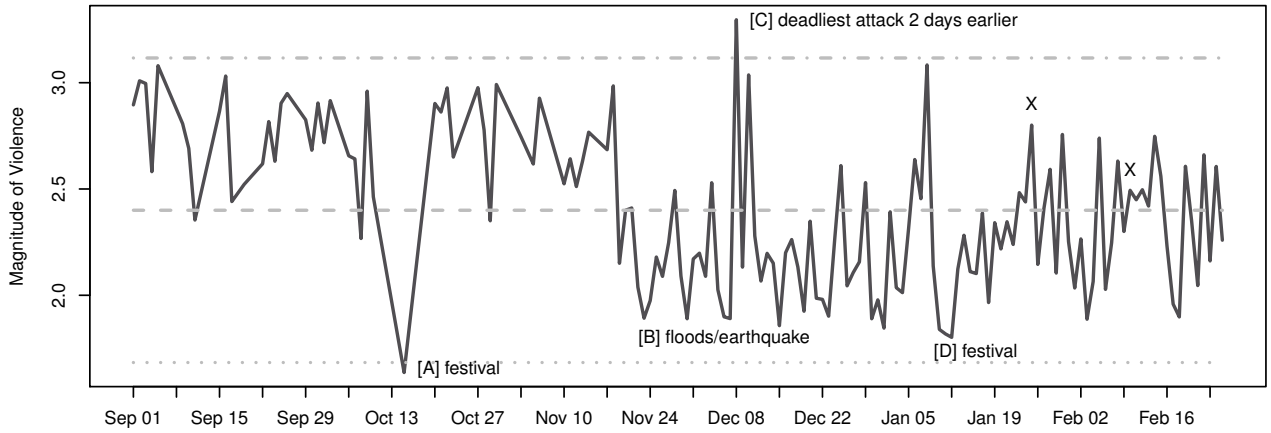

Fig. 4. Case study spanning Sept. 2013 to Feb. 2014. Results obtained by applying the Skip-Gram model of similarity. The dotted grey lines indicate the average and two standard deviations. The two X mark news instances reporting forced migration.

around two important festivals at points A and D. Another notable observation includes event C, which has a sharp peak around 6 Dec. 2013 but the actual attack, one of the deadliest for the year 2013, happened four days ago on 4 Dec. 2013. This lag can be explained due to the delay in reporting usually exhibited by the more traditional media like newspapers. Another trough, event B, is noticeable when the country was hit by heavy floods as well as a small earthquake. These findings, however, should be considered with caution as these results were obtained using our proposed approach applying the NPMI method without any manual labelling and also due to the limited size of the study sample where there may be other hidden contributing latent factors behind these occurrences. Extending this case study to a longer time-frame may shed more light on this phenomenon.

Although our bigger goal is to build a comprehensive model of predicting forced migration, data regarding the number of displaced people are currently inaccessible publicly. However, we extracted some statistics for two data points (indicated by black stars in the chart) using publicly available data. For example, on 24th Jan. 2013, the UN "*warned that residents are fleeing Fallujah and Anbar (two Iraqi provinces) in numbers that rival those from 2009 and believe that over 140,000 have been displaced since the end of last year*". It further claimed that "*over 65,000 have left the two cities in the last week alone*". This trend is hinted at in our chart starting around January as the reporting of violent events becomes more frequent. The Iraqi security forces believed that "*about 80% of the population has fled Fallujah*". About two weeks later, on 10th Feb. 2013, the UN released an estimate that "*up to 300,000 Iraqis have been displaced due to insecurity*".

VII. Conclusion

In this paper, we present an effective approach for detecting the magnitude of an event over a period of time. We use *violence* as an example event as it is one of the most critical factors in the study of forced migration.

We adopted the measures of semantic similarity to build a model for detecting the trend of violence, which we evaluated against our manually annotated data set with the best method obtaining a Pearson's correlation of 0.8. Although all three methods of semantic similarity (NPMI, CBOW and SG) output comparable results, they all outperform the baseline method significantly. This, we believe, is a preliminary step in identifying the magnitude of target events in the study of forced migration.

One avenue of future work includes extending the preliminary case study done in this paper, which indicated a correlation between violence and key cultural and environmental events, to a longer timeframe to substantiate any relationship. We would also like to explore the characteristics of other factors of forced migration. In addition, adapting other models of semantic similarity, such as graph-based methods, is also a part of our future work.

Acknowledgment

The authors are grateful to Lisa Singh of Georgetown University for providing the EOS data set used in this research and also for the many discussions and insights she provided in our joint project on predicting forced migration using big data. We would also like to thank Lisa Yan for her help in annotating the data set used in our experiements and the anonymous reviewers for their valuable feedback. The research is funded in part by Social Sciences and Humanities Research Council of Canada and Natual Sciences and Engineering Research Council of Canada.

References

[1] R. O. for the Syria Crisis, "Syrian arab republic: Humanitarian snapshot," December 2015, [Online; posted 2-December-2015]. [Online]. Available: http://www.reliefweb.int/files/resources/ I

[2] S. R. R. Response. (2016, May) Syria regional refugee response. [Online]. Available: http://data.unhcr.org/syrianrefugees/regional.php I

[3] U. N. O. for the Coordination of Humanitarian Affairs, May 2016. [Online]. Available: http://www.unocha.org/syria I

[4] W. V. Staff, "Syria refugee crisis faq: How the war is affecting children," April 2016, [Online; posted 11-April-2016]. [Online]. Available: https://www.worldvision.org/wv/news/Syria-war-refugee-crisis-FAQ I

[5] D. Chakrabarti and K. Punera, "Event summarization using tweets." *ICWSM*, vol. 11, pp. 66–73, 2011. I, II

[6] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *Data engineering (icde), 2012 ieee 28th international conference on*. IEEE, 2012, pp. 1273–1276. I, II

[7] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan, "Sted: semi-supervised targeted-interest event detectionin in twitter," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1466–1469. I, II

[8] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 37–45. [Online]. Available: http://doi.acm.org/10.1145/290941.290954 II

[9] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03. New York, NY, USA: ACM, 2003, pp. 330–337. [Online]. Available: http://doi.acm.org/10.1145/860435.860495 II

[10] W. Stuetzle, D. B. Percival, and C. Marzban, "Targeted event detection," *arXiv preprint arXiv:1003.2823*, 2010. II

[11] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433. II

[12] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 1155–1158. II

[13] J. Weng and B.-S. Lee, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011. II

[14] A. Ritter, O. Etzioni, S. Clark *et al.*, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1104–1112. II

[15] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 28–36. [Online]. Available: http://doi.acm.org/10.1145/290941.290953 II

[16] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777 II

[17] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. P. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, C. J. Kuhlman, A. Marathe, L. Zhao, T. Hua, F. Chen, C. Lu, B. Huang, A. Srinivasan, K. Trinh, L. Getoor, G. Katz, A. Doyle, C. Ackermann, I. Zavorin, J. Ford, K. M. Summers, Y. Fayed, J. Arredondo, D. Gupta, and D. Mares, "'beating the news' with EMBERS: forecasting civil unrest using open source indicators," *CoRR*, vol. abs/1402.7035, 2014. [Online]. Available: http://arxiv.org/abs/1402.7035 II

[18] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990. IV-A

[19] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009. IV-A

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781 IV-B

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119. IV-B

[22] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015. V-B, V-C

[23] G. Recchia and M. N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behavior Research Methods*, vol. 41, no. 3, pp. 647–656, 2009. [Online]. Available: http://dx.doi.org/10.3758/BRM.41.3.647 V-C

[24] U. D. News. (2013, June) UN Daily News. [Online]. Available: http://www.un.org/news/dh/pdf/english/2013/06062013.pdf VI