Detecting and Using Buzz from Newspapers to Understand Patterns of Movement

Julia Hocket, Yaguang Liu, Yifang Wei, Lisa Singh, Nathan Schneider
Georgetown University
Washington DC
Email:{jlh279,yl947,yw255,lisa.singh,nathan.schneider}@georgetown.edu

***Abstract*—Meaningful leading indicators of mass movement are difficult to discover given the dearth of available data about involuntary movement. As a first step, we propose analyzing whether we can use the changing dynamics of newspaper content as one possible indirect indicator of such displacement. Specifically, we explore whether news media buzz correlates with patterns of migration in Iraq. We consider different methods for detecting buzz and empirically evaluate them on a corpus of 1.4 million articles.**

I. Introduction

Forced migration, or the involuntary movements of refugees or internally displaced persons, is a complex global problem that is very difficult to predict [3]. To address this larger issue, we are interested in determining if we can construct variables from open-source data that can be used as indirect indicators of the movement. Specifically, we want to explore whether patterns of "buzz" extracted from newspaper articles about migration correlate to expected patterns for the topic domain. Here, we informally define "buzz" as the amount of discussion around a specified topic over a particular time window.

More formally, the problem is the following: Given a document collection $\mathbb{D}$ and a topic of interest, $\mathbb{T}$, consisting of one or more seed topic words in lexicon $\mathbb{L}$, determine which vocabulary augmentation strategy produces a lexicon $\mathbb{L}'$, that more accurately captures the amount of buzz, $buzz = [buzz(\tau_0), buzz(\tau_1), ..., buzz(\tau_k)]$, for $\mathbb{T}$ through time.

To this end, this work 1) proposes different methods for detecting buzz, 2) empirically evaluates our proposed methods for detecting buzz on over 1 million articles related to Iraq, and 3) shows that incorporating lexicon expansion leads to meaningful buzz detection correlations to migration patterns. We find that considering word embeddings for vocabulary expansion is a promising direction, and that buzz detection in general is a promising approach for indirectly detecting movement from newspaper data sources.

II. Relevant Literature

While the concept of buzz detection is a new one, there are many similar concepts in the literature, including topic modeling and event detection. The goal of topic modeling is to determine "topics", i.e. sets of overlapping, theme specific words, for a collection of documents. Standard approaches use generative probabilistic modeling [1] or graph-centric approaches [2]. Our interest, however, is not determining the topics that are predominant across a document collection. Instead, we are interested in different ways to define the vocabulary associated with one, single topic and measuring the changing dynamics of that topic, i.e. detecting the topic buzz. The goal of event detection is to extract real world events at a particular time and location from open source data sets [5]. While related, the goals of event detection differ from buzz detection in that the former attempts to identify discrete events, while the latter looks to find the amount of discussion surrounding a particular topic.

III. Calculating Buzz

Algorithm 1 presents our high level approach for computing buzz. Given $\mathbb{D}$ and $\mathbb{L}$ as input, we output $buzz$. We begin by computing the buzz for each document and then combining the buzz scores for each time period. We then use those combined scores to determine an overall level of buzz for each time period (lines 5-7). The component of the algorithm we are analyzing is the buzz calculation ($compute\_buzz$). We present six methods for computing buzz: topic keyword variants (TKV), subject-matter-expert keywords (SME), dictionary expansion (DEX), embedding expansion (EEX), limited embedding expansion (LEEX), dictionary embedding expansion (DEEX), and frequent pattern mining expansion (FPM).

The first two methods (TKV and SME) utilize fixed lexicons, as determined by the topic and by SMEs. The TKV lexicon includes only the words directly and morphologically related to the topic seed words, while the SME lexicon is maually created by experts who identify words relevant to the topic. These two approaches are fairly standard methods that can be viewed as baseline methods.

The next set of methods consider expansions of the base lexicons. We begin with semantic expansion, or the automatic generation of synonyms, of each word in the SME-defined base lexicon. Such an expansion captures a broader understanding of the topic domain, as it includes many more tangential words that can be used to describe that topic. To expand our lexicon then, we obtain all dictionary synonyms for each word (dictionary-expanded lexicon (DEX)).

DEX still neglects capturing morphological variants of each word (e.g. 'migration' to 'migrant'). We therefore turn to a method of expansion that captures both the semantic and morphological variants of each word. We use pre-trained word embeddings to find the words with the most similar vectors to

Algorithm 1 Buzz Computation and Normalization

Input: $\mathbb{D}$, $\mathbb{L}$
Output: *buzz*
Function:
Let $buzz(\tau_k) = 0$ for all $\tau_k \in \tau$
for d **in** $\mathbb{D}$ **do**
    $buzz(\tau_t) \leftarrow buzz(\tau_t)$ + *compute_buzz*(d, $\mathbb{L}$)
end for
return *buzz*

each of the lexicon words. In this work, we use cosine similarity due to its relative simplicity and symmetry. We refer to this strategy as the embedding-expanded lexicon (EEX). We also consider a limited form of the embeddings expansion (LEEX), as the embeddings tend to generate first morphological variants and then semantic variants, of a given word. The LEEX method aims to capture primarily morphological information from the embeddings, without cluttering the expansion with much semantic variation.

Building off of each of the previous methods, we next propose a combination of methods. In the dictionary embedding expansion (DEEX), we first generate synonyms for each SME-defined lexicon word from the thesaurus. Then, we use the embeddings to automatically generate additional words, though we restrict the number of words generated from the embeddings, thus capturing primarily morphological variation and leaving the semantic variants to the thesaurus.

Our final method focuses on identifying words that are not only semantically similar, but also occur frequently with the words that are part of the core lexicon or one of the mentioned augmented lexicons. To accomplish this, for each mentioned strategy, we use Frequent Pattern Mining (FPM) on the data collection to identify context relevant terms. Because this methods tends to generate a large number of frequent words, we post-process the frequent words using TF-IDF.

IV. Considerations When Determining Buzz

While the method for generating the lexicon is the most important consideration when determining buzz, deciding on whether or not to weight the words in the lexicon based on relevance is also a consideration. In this work, we will consider both unweighted and weighted lexicons to understand the impact of weights for buzz detection. An unweighted lexicon means that each word is viewed as equally relevant to our topic. In considering unweighted lexicons, we can calculate buzz to be the number of times a word in the base lexicon appears in a single document (magnitude) or across documents (binary). For the magnitude calculation, we want to calculate a value for a document's buzz as the sum over all words in $\mathbb{L}$ that appear in the document. For the binary calculation, if a word appears one or more times in a document, a value of 1 is returned. Otherwise, a 0 is returned.

We can also consider a weighted variation of the above for embedding expansions. Rather than assuming that each word is equally similar to the given lexicon word, we can weight each added term by its cosine similarity to the lexicon word, yielding a weight between -1.0 and 1.0, inclusive (1.0 being identical, -1.0 being diametrically opposed). If a expanded word (either by thesaurus or by embeddings) is an expansion of two or more lexicon words, its assigned weight is the maximum cosine similarity when compared to each of the words. To calculate buzz using the weighted lexicon, we replace the binary value (0 if the word is not in lexicon, 1 if it is) with a weight representing the relevance of that particular lexicon word. It should be noted that there is no sense of weighting in the binary version of buzz, which indicates the presence of any word in the lexicon.

V. Evaluation and Discussion

In this section, we will show an empirical evaluation of our proposed buzz detection algorithm. The primary dataset used to evaluate each of our methods is the Expandable Open Source (EOS) database, an unstructured archive of over 700 million articles managed by Georgetown University [4]. New articles are added to the archive at the rate of approximately 100,000 per day from over 20,000 Internet sources in 46 languages. We use a subset of more than 1.4 million English-language articles dated between January 2016 and December 2016 that either contained the name of a location in Iraq or were from a news source in Iraq.

In order to evaluate buzz accuracy, we consider the topic "migration" and attempt to determine the buzz detection strategy that most closely maps to actual movement patterns in Iraq. If the correlation is high, it supports the notion that buzz may be a reasonable indirect indicator of migration. To accomplish this, we determine correlations between the buzz values generated from the newspaper data and movement numbers of internally displaced persons released by the International Organization for Migration (IOM). Based on the IOM statistics, we know that a great deal of movement took place in 2016 (see Figure 3).

We begin our empirical evaluation by comparing the buzz detected using the different vocabulary augmentation strategies. Figure 1 shows the comparisons among different semantic augmentation algorithms using the unweighted binary version of the method. Looking at Figure 1, we see that the SME and the TKV methods do not effectively capture the discussion about movement. All the different expansion approaches capture more buzz. Figure 2 compares the different semantic lexicon augmentation strategies with further augmentation using FPM (support = 0.01). We see that using FPM seems to increase the buzz for all the methods except for TKV. The question becomes, which buzz computation is most relevant to the actual migration? We determine this by computing the correlation between each buzz detection method and the IOM movement data (see Table III). We see the highest is 0.726 for the DEX and DEEX expansions.

There are a few take away messages. First, in general, expansion of the lexicon using any method improves the correlations. This is not surprising. Second, using an unweighted

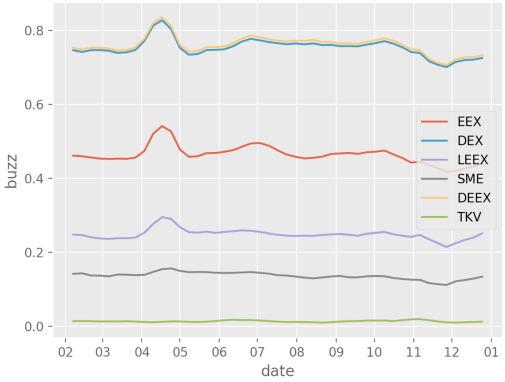

Fig. 1. Buzz for topic 'migration' using unweighted, binary lexicons

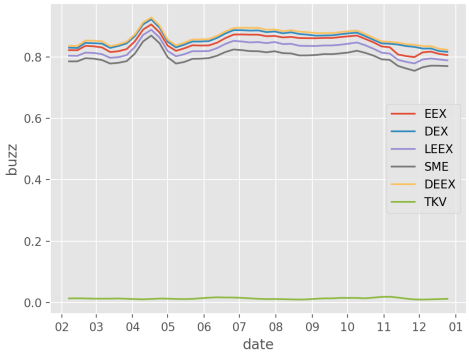

Fig. 2. Buzz for topic 'migration' using lexicons augmented with FPM

binary aggregation of buzz scores across documents is better than considering magnitude (see Table II). We think this results because this document collection has a large number of authors and their word frequency usage varies considerably. So ignoring that seems to lead to a cleaner model. Finally, while FPM methods help in general, some noise is introduced when augmenting the DEX and DEEX expansions, thereby reducing the overall correlations for those two methods.

Finally, Figure 3 compares the ground truth IOM movement data to buzz values generated using DEEX. We see that the trend is similar, but there seems to be a second "bump" in the buzz signal that is not present in the IOM data. This second bump is actually movement of Iraqis to places outside of Iraq, including to Europe. In other words if the IOM data also contained movement to foreign countries, the correlations would likely be higher. Given this, it is reasonable to say that there is a clear value in using buzz as an indirect indicator of movement when other data sources are not available. While we consider buzz as a reasonable indicator for the social phenomena of migration, it is likely not be a sufficient indicator on its own.

TABLE I
CORRELATIONS BETWEEN BUZZ DETECTION STRATEGIES AND IOM DATA

Algorithm	Pearson	FPM Pearson
TKV	-0.063	-0.063
SME	0.658	0.682
DEX	0.726	0.652
EEX	0.677	0.678
LEEX	0.510	0.682
DEEX	0.726	0.646

TABLE II
RESULTS FROM VARYING THE BUZZ COMPUTATION CALCULATION.

Weighting	Computation	Pearson
Unweighted	Binary	0.726
Unweighted	Magnitude	0.670
Weighted	Magnitude	0.633

VI. CONCLUSIONS AND FUTURE WORK

This work defines buzz, identifies and compares seven approaches for computing it, and shows its value on a newspaper data collection for serving as a leading indicator of migration.

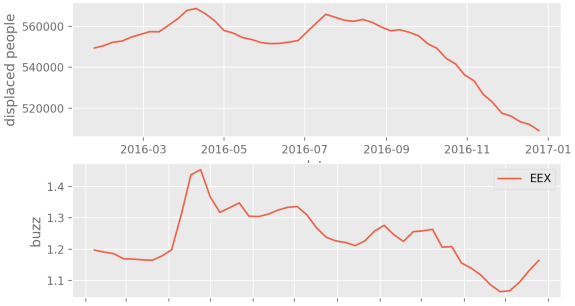

Fig. 3. Detected buzz for topic 'migration' using unweighted lexicons (binary) with IOM migration statistics

The results from the empirical evaluation are promising and indicate that buzz can be captured effectively using newspaper data. We can see that advanced expanding algorithms can contribute a great deal to understanding content in an open-source data set containing documents written by many different authors. The expansions rendered can produce broad representations of the topic that can allow for large amount of variation in the open-source set of articles. Future work will consider buzz more extensively for different domains and social media documents.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation and the Massive Data Institute at Georgetown University.

REFERENCES

[1] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, Apr 2012.

[2] R. Churchill, L. Singh, and C. Kirov. A temporal topic model for noisy mediums. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2018.

[3] S. Martin, S. Weerasinghe, and A. Taylor. *Humanitarian Crises and Migration: Causes, Consequences and Responses*. Routledge, 2014.

[4] L. Singh and R. Pemmaraju. EOS: A multilingual text archive of international newspaper and blog articles. In *IEEE International Conference on Big Data (BigData)*, 2017.

[5] Y. Wei, L. Singh, B. Gallagher, and D. Buttler. Overlapping target event and story line detection of online newspaper articles. In *IEEE International Conference on Data Science and Advanced Analytics*, 2016.