Blogs as a Collective War Diary

Gloria Mark[1], Mossaab Bagdouri[2], Leysia Palen[2], James Martin[2], Ban Al-Ani[1], Kenneth Anderson[2]

[1]Dept. of Informatics
University of CA, Irvine
Irvine, CA 92697
{gmark, balani}@ics.uci.edu

[2] Dept. of Computer Science
University of Colorado
Boulder, CO 80309
{bagdouri, palen, martin, kena}@cs.colorado.edu

ABSTRACT

Disaster-related research in human-centered computing has typically focused on the shorter-term, emergency period of a disaster event, whereas effects of some crises are long-term, lasting years. Social media archived on the Internet provides researchers the opportunity to examine societal reactions to a disaster over time. In this paper we examine how blogs written during a protracted conflict might reflect a collective view of the event. The sheer amount of data originating from the Internet about a significant event poses a challenge to researchers; we employ topic modeling and pronoun analysis as methods to analyze such large-scale data. First, we discovered that blog war topics temporally tracked the actual, measurable violence in the society suggesting that blog content can be an indicator of the health or state of the affected population. We also found that people exhibited a collective identity when they blogged about war, as evidenced by a higher use of first-person plural pronouns compared to blogging on other topics. Blogging about daily life decreased as violence in the society increased; when violence waned, there was a resurgence of daily life topics, potentially illustrating how a society returns to normalcy.

Author Keywords

Blogs, collective identity, crisis, war, crisis informatics, longitudinal study, topic modeling

ACM Classification Keywords

K.4.3 [Computers and Society]: Organizational Impacts – Computer-supported cooperative work.

INTRODUCTION

Blogging is a global phenomenon. As of this writing, over 162 million blogs worldwide exist with over 800,000 blog posts indexed each hour [6]. Through blogs, people can express their views on local events as they unfold. This information is (mostly) accessible to a global audience, and often provides a range of viewpoints that traditional media do not cover. Blogs may collectively provide a public narrative about an event. In the case of a geographical region in conflict—as we focus on here—blogs provide reports, opinions, and other rich information to others outside the region that can be difficult to obtain, especially if the area is perceived as dangerous for travel.

There may be advantages to understanding the content of blogs when a society experiences a disruptive event. First, we expect that such an analysis of blogs can be used to assess the mood or "pulse" of the nation [cf 26]. Understanding how bloggers—as a subset of a population—respond to a disruption, for example in terms of recovery [24], can help researchers build theories about socio-psychological responses to disaster and crisis. A second reason is that the blog content of thousands of people written during a disruptive event can inform historical accounts about the time period. A third reason is that results can influence national or international policy decisions about the provision of humanitarian aid.

Though traditional media report selective excerpts about a disruption, they often do not adequately capture the response, mood, or opinions of affected residents. Though most wars and conflicts usually have extensive media coverage, such coverage is often influenced by factors such as the "creation" of news, profits, and rivalry, which can lead to media bias [4]. In contrast, bloggers might report on events that they feel are significant or of personal interest to them. Blogs provide an important perspective about a war that may differ from traditional media. For protracted crises like wars, blogs are the modern form of the war diary. In contrast to earlier times when a small number of war diaries took months or years to be published and thus reach a public audience, blogs can inform a global audience in near real-time on conditions of war, as well as on people's responses and adaptations to it.

We collected blog posts from the Iraqi blogosphere during eight years of the Iraqi war. Our broad research question is: what can the blogosphere reveal about how a society responds to war over time? We investigate how blogging about war differs from blogging about daily life, a focus of studies of blogs in non-war settings [22, 36]. However, a significant challenge for analysts using blogs as documentation of events is the sheer amount of material

CSCW 2012, February 11–15, 2012, Seattle, Washington.

bloggers generate. Especially for emergent events such as disasters, disruptive political conflicts, or other mass emergencies, there may be millions of posts each day [3]. This leads to our concomitant research objective: to investigate how data analysis techniques yield insight into the collective expression of large numbers of people about a significant event over time.

LONGITUDINAL RESPONSES TO CRISES

In recent years, research in human-centered computing (HCC) has focused on how citizens coordinate and respond to crises [16, 38, 44, 46, 47] using personal information and communication technology. These studies have examined behavior during the emergency response period, which includes the phases of warning, impact and response, and then endures for about a two-week time frame post-impact.

Crises, however, do not occur as a snapshot in time; rather the experience and effects of a crisis unfold over a long period, sometimes lasting years. Oliver-Smith [37] discusses how responses to and effects of a crisis can last decades. The effects of September 11, 2011 and the 2005 Hurricane Katrina are still felt in the U.S.—as are disasters that have occurred decades before [31]—both by those directly affected as well as by a broader public.

It can be illuminating, then, to capture and analyze data written over the course of the event. Media such as Twitter and Facebook, which have been studied as communication channels during crises [39, 44, 46], do not as easily yield longitudinal data because of the difficulty in collecting the high volume in the brief time it is comprehensively available, or because of Terms of Use restrictions. On the other hand, blogs, unlike Twitter, have no character limit, which enables people to write about opinions, report news and personal experiences at length. Unlike ephemeral media such as mobile phone calls or short-lived data like microblog posts, blog posts are archived automatically (unless deleted by the blogger). Thus, years of reports by bloggers are available and archived chronologically, which makes it possible to understand how opinions, attitudes, and identity evolve in relation to the crisis effects. Further, blogs provide insight into how a group surrounding the blogger responds to events, because a larger public can comment on posts. For these reasons, blogs as data may reveal insights about citizen response to protracted crises.

Some researchers have studied the longitudinal aspects of reaction and recovery from crises. An anthropological study of the Oakland firestorm of 1991 [24] revealed how those affected first experienced a feeling of isolation as victims [15]. Those who lost their homes and belongings initially began to meet, and over time the meetings culminated in a disaster center. The role of place in disaster events has been researched as well by others. In earthquake events, it is not uncommon for places to emerge as public gathering spots to compensate for compromises to the telecommunications and road infrastructures. In the 1999 ChiChi, Taiwan earthquake, shrines became gathering places; during the 2004 Indian Ocean earthquake and tsunami, the UN Humanitarian Center in Sumatra became a public place to receive support from others experiencing the disaster [38].

Online Response in Dangerous, Protracted Events

In some crisis events and geographical areas, however, it is not possible for citizens to find others and meet physically. In Iraq the second Gulf War began in 2003, and is the subject of this investigation. The fall of the Iraqi government in 2003 heralded the beginning of widespread civil disorders and looting [23]. This violent disorder and disruption continued to escalate, heighten, and become more organized as the first free elections in Iraq drew near in 2005 resulting in an increase in the number of fatalities [19, 23]. The level of violence has continued to fluctuate, typically coinciding with political, religious and other events. Iraqis are routinely exposed to acts of violence such as random sniper fire and bombs exploding in public places. Recurring acts of violence lead to curfews, roadblocks and even battles within suburbs [1]. Such conditions have had a significant impact on everyday life [33].

In such an environment, we might expect citizens to experience a feeling of isolation. However, social media can afford residents under such circumstances the possibility to mitigate feelings of isolation by writing about experiences online to share them with a potential audience.

Blogging about war and daily life

Our first research question addresses the nature of blogs written during a war. Most studies of blogs have not addressed the nature of blog writing during a major disruption like war. Studies of blog genres in non-war environments consistently reveal that most blogs are classified as being of a "personal nature" [e.g. 22]. Nardi and colleagues [36] investigated why people write "personal diaries" in such a public forum as the Internet and discovered a range of motives: broadcasting their status, expressing their own views as well as seeking others' views, and articulating thoughts.

We consider that blogs written during wartime may be thought of as online war diaries. The war diary is a genre of writing that can be found as early as the Thirty Years War (1618-1648).[1] War diaries can be highly personal, as with the famous civil war diary of Mary Chesnut, deemed to be written "*for herself and for her eyes alone*" [11, p. 10]. The well-known diary of Anne Frank describes daily life events about her personal life and family in the context of a war.

However, unlike traditional war diaries, blogs during war are written in a highly public forum with the potential for citizens to share their experiences. How then might people express themselves about war in such a public forum? At what point would online blog accounts during this time reflect war in the forefront, and when would blogs rather address other subjects concerning daily life, as bloggers

[1] http://www.mdsz.thulb.uni-jena.de/sz/index.php

typically have been shown to write about [cf. 36]? From a macro-perspective, blog topics might reveal insight into the reactions or even the state of a population experiencing a crisis. Blogging about war might indicate for example, that a segment of society (as represented by the blogosphere) is under stress, whereas blogging about daily life might, for example, indicate that a society has adapted to the war setting. Our research question examines how blog topics change with the course of the war; can we find a pattern over time when citizens blog about daily life events other than war and when they focus on the war? How would such a pattern relate to external war events?

Identity in Wartime Blogs

Identity on the Internet has been a topic of interest since the Internet began to be widely adopted [48]. Our second research question addresses how identity is reflected when people blog during a war. Following the idea that blogs (in non-war environments) are generally found to be of a personal nature, how is identity expressed when people blog about an event that everyone in their region experiences? Is war reflected online more as a personal experience or as a collective experience? We might expect that people blog about war from a personal perspective, e.g. as war diarists such as Mary Chesnut had. We might expect that the reasons for bloggers to write in a self-referential voice include to serve as a catharsis, to relay their status of being safe, or to articulate their thoughts [36].

There are reasons, however, that we might expect citizens to blog about war using a collective voice. Individuals experiencing trauma have been found to use first person plural pronouns to refer to collective trauma [2]. Societies experiencing such trauma often expand "the circle of we," creating solidarity and a sense of collectivity, as have those that have experienced "collective abuse" [2]. Such trauma can lead to a collective identity that transcends the event [32]. In fact, war can even be a means for a group or society to build collective identity [12]. Thus, we may see the emergence of a collective identity in blog data where the topic of war is discussed.

In media written as narratives, psycholinguists explain that pronouns are markers of identity: the use of the personal pronoun "I" reflects self-identity whereas first person plural pronouns of "we" reflects group or collective identity [34, 41]. These markers thus reflect one's referential point: the self or others. Thus, pronouns can be considered to be not only simple descriptors, but also active assertions, with words that represent a specific choice of meaning [34]. In some cases, collective identity formation can also imply an identity-centric partnership of "us against them" [32]. Thus, analysis of pronoun markers is a means to assess the extent to which people focus on the self or others [41].

The Case of Iraq: The First Bloggers

The earliest known Iraqi blogger within Iraq began in 2003 [40]; many Iraqis living in Iraq and abroad began blogging soon after. Their blogs are diverse and express perspectives on a myriad of topics that include the war, daily life and personal interests. It is not possible to determine the total number of active Iraqi bloggers due to the size of the blogosphere. Previous studies (e.g. [1]) have sampled data based on an Iraqi blog that is dedicated to indexing all the blogs created by Iraqis within Iraq and the diaspora. It has been found that people used blogs to form communities of support, to garner emotional support, and to engage in dialogue about the conflict [1]. The current study instead takes a macro-perspective to track how perspectives learned from blogs unfold longitudinally vis-à-vis external events.

RESEARCH METHODS

Though qualitative research can provide a rich understanding of this type of data and domain, the large corpora of blog data can make extended manual analysis intractable. The scope of a qualitative analysis of blogs is limited by the sample size. We therefore employed automatic methods for exploring large-scale document collection content to study reaction to a crisis at the societal level. Specifically, we sought to explore the means by which the blogosphere can be considered more fully through natural language processing techniques.

Methods for analysis of large data sets

We first use topic models, an unsupervised approach to text analysis [5, 20], to reveal the set of underlying topics in the blog data. We then use select temporal and linguistic analyses of these clusters to gain insight into the nature of topics being discussed, their change over time, and what they reveal about collective expression during war.

Topic Modeling

Topic modeling is a computational approach based on generative probabilistic models, where no *a priori* information about the nature of the data is required [5, 20]. The history of these techniques can be traced back to vector-based methods such as Latent Semantic Indexing [12]. These techniques accept a collection of documents as input, and produce as output representations of the latent concepts (topics) that underlie the actual texts. In the case of topic modeling, these latent topics are represented by a probability distribution over terms (i.e. keywords) that are most characteristic of that topic. These distributions in turn can be used to compare the similarity of content of individual documents in the collection (in our case, individual blog posts) to each identified topic. We can then identify what topics appear in a given document at what relative concentrations, e.g. a post is mostly about daily life (75%) but also discusses the war (25%).

Since its introduction in 2003, topic modeling has been applied to a wide variety of domains and applications, many of which include correlations with, and in some cases predictions of, events external to the model. Yano et al. [49] used it to predict the users most likely to comment on a political post and to generate the content of these comments. Topic models have also been used to track the drift of topics over time in research fields [21] as well as in blogs about commercial products [28]. In a preliminary examination, Kireyev, et al. [27] applied topic models to

crisis-related microblog posts. Work has also been done to validate the analysis of social media from external facts [7].

Pronoun Analysis

Word count techniques assume that the use of specific words, such as pronouns, reveal insight into meaning independent of the semantic content [41]. With respect to the use of pronouns, the analysis of the variation in use of referring expressions (noun phrases, names, pronouns, etc.) in spoken and written discourse has a long history in sociolinguistics and social psychology. Such analyses have been applied to collective identity [8], author and gender identification [29], cultural differences [17] and power and dominance relations [9]. Many of these studies have demonstrated the importance of differences in particular pronoun use as well as differences in the use of pronouns versus other referring expressions.

In this research, we first apply topic modeling to blog posts written during the Iraq War. We use the discovered topics to assign individual posts to particular topics (given the concentration of each topic in that post.) Having done this, we show how topics are correlated with external metrics and how the common themes in blogs changed over time in response to external events related to the war, thereby lending validity to the topic identification. We then examine in more detail the commonalities and differences among the clusters of posts by topic. We then apply pronoun analysis to clusters concerning war, comparing this to other clusters to infer personal or collective identity.

DATA COLLECTION AND PREPARATION

We developed an application to crawl a list of Iraqi blogs indexed at http://iraqblogcount.blogspot.com and http://iraqiblogindex.blogspot.com to create our dataset. These two indices contained 272 publicly available blogs, which we captured in their entirety. Since many authors write multiple blogs, a scan of the 272 publicly available blogs led to the discovery of 185 additional blogs written by the same authors. We captured these blogs as well, leading to a final data set of 457 blogs containing 46,828 individual posts and 188,011 comments covering the period 1 May 2002 to 3 February 2011. Our sample contains at least 236 unique blog authors. We cannot ascertain the exact number because the API did not return the ID of some profiles; thus we do not know the owners of 30 blogs. For each post, we extracted and stored the title, author, and publication date.

The posts were then preprocessed to detect the language used to enable comparison of Arabic versus English posts. We found that the actual language of the post did not always match the language indicated in the metadata of the post. In addition, we found that some bloggers tended to mix both languages in the same posts. These findings led us to base calculations on the number of Arabic versus English characters in a post's content to assign its language. A post that had more than 50% of the characters after excluding hyperlinks as Arabic letters was classified as an Arabic post and vice versa. This scheme resulted in 11,668 Arabic posts, 31,246 English posts and 3,914 undetermined. We analyzed only the English and Arabic datasets. We cannot ascertain if all the English bloggers lived in Iraq. Even so, though bloggers may not experience the violence of war directly, people outside of a war region can still experience its impact on the psyche and socio-cultural order [31].

Posts were then tokenized and stemmed. Stemming is an algorithm that aims to find the stem (i.e., the root) of a word by executing a set of operations such as the elimination of prefixes and suffixes. The Snowball stemmer [43] and Buckwalter morphological analyzer [10] were used respectively for English and Arabic posts. Our preliminary work showed that the use of stemming was of some benefit for English, and critical for Arabic posts.

APPLYING TOPIC MODELING

We then applied a widely used topic modeling toolkit [20] to our collection. For the English collection a model was trained on 31,246 posts containing a stemmed vocabulary of size 126,100; the Arabic model was trained on 11,668 posts with a stemmed vocabulary of size 73,250. The models presented here were trained with 10 topics for both English and Arabic.

The topics discovered by the English and Arabic models are illustrated in Tables 1 and 2. Each topic is illustrated with the top words that are most representative of it (the complete representation of each topic consists of a full distribution across all words in the vocabulary). The final column indicates each topic's percentage among all topics.

Topic modeling can enable identification of topics of primary interest, which helps analysts focus on posts predominately concerned with specific topics. For example, Table 2 shows how the model detected a set of posts that contribute to topic TE04, which include the words "Iraq Baghdad kill Iraqi force war attack, etc." Analysts interested in discussions of *country, politics, Iraq and Arabs* can locate those posts that contributed to this extracted topic.

To assess the accuracy of the models, a single researcher fluent in English and Arabic and familiar with this study domain analyzed random samples of Arabic and English posts. Four posts were randomly selected and manually analyzed for every month from 2003-2010 from the English data set, and from 2004-2010 from the Arabic data set (2003 was excluded because it only represented one blogger who posted about 4 posts per topic). The results of the manual analysis were consistent with the topic modeling categorizations.

ID	Topic	Percentage
TE01	People Iraqi Iraq Country American war Saddam live year kill America happen terrorist government	9.25%
TE02	year work company busy develop tea market project system service money provide case program	7.86%
TE03	day time t friend thing I'm feel start house talk go work good told car	9.44%
TE04	Iraq Baghdad kill Iraqi force war attack military police bomb report American soldier bush army	15.20%
TE05	Iraq al Iraqi govern politic elect Sunni party Iran Arab nation vote leader minister Kurdish	9.24%
TE06	god women Muslim love life man Islam men heart person religion woman Christian face eye	9.04%
TE07	oil de time receive model turn high t power radio la circuit set work point	10.51%
TE08	al news year day today call Baghdad ago Abu watch TV video family play story	7.78%
TE09	water picture city place black build white small food long game green open red photo	13.79%
TE10	post read blog write Arab book comment university time blogger music link interest student publish	7.89%

Table 1: Most frequent terms in the English posts by topic (TE = Topic in English)

ID	Topic	Percentage
TA01	one want elapse days same go people day man saying listen know	12.04%
TA02	life times poetry love amiability eye heart death night art soul beautiful spirit face sound	8.07%
TA03	Allah son say peace the Imam pray Islam Mohammed saying kin father worshipper Muslim prophet	12.12%
TA04	Iraq the force government president say council security governorate operation Baghdad visit America	16.58%
TA05	other many book enable write period something person same one do picture time site more	13.06%
TA06	the scholar religion history other Arab saying knowledge idea first origin culture earth own soul	9.24%
TA07	Iraq America Arab people occupation Saddam party war Ba'ath Iran nation force Israel resistant	7.13%
TA08	the Baghdad one expand day logic hour children poison ten three big kill car	7.39%
TA09	country politic operation united state no the government force meeting new other rights	7.44%
TA10	work year media company information office studying network university service program newspapers	6.94%

Table 2: Translations of the most frequent terms in the Arabic posts by topic. (TA = Topic in Arabic)

Blogging: a temporal mirror of war events

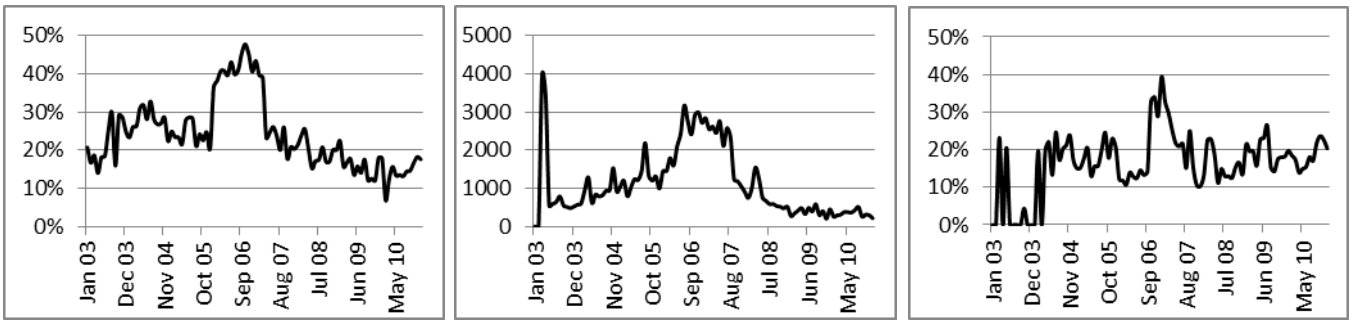
We first identified topics concerning war. Referring to Table 1, three independent coders interpreted topics TE01 and TE04 as topics concerning war. For example, included in TE01 were *American*, *war*, *Saddam*, *kill*, and *terrorist*. In TE04 were the words *kill*, *force*, *war*, *attack*, *military*, *bomb*, and *army*. These gave strong indication that the topics were about war. Figure 1a shows how the frequency of these war topics changed over time. War topics were considered *in proportion* to the overall topics discussed.

Next, we compared the blogged topics with external events related to the war to test whether the war topics corresponded with what was occurring in the world. We used the body counts index (number of civilian deaths) as a proxy for the level of magnitude or severity of the war. Data were collected from http://www.iraqbodycount.org, an independent nonprofit group that tracks the number of war-related civilian deaths in Iraq. According to the website, deaths are determined by cross-checking different news sources and reports from hospitals, morgues, and NGOs.

We correlated the body count with the proportion of war topics blogged (Figure 1a). We used one month as a time unit for conducting the correlation, a large enough unit to use to account for a lag in time from a war event (e.g. a bombing) to when the theme of war would surface in blog posts. Though we collected all posts from 2003 onwards, we only considered blogs and body counts starting from the year 2004, because very few blogs appeared prior. Keeping in mind that the Internet was not widely accessible in Iraq until after 2003, it took about a year for blogs to be written by enough people to enable study of thematic content.

Visually, the English blogs classified as containing war topics seems to mirror the magnitude of body count over the years of the war (Figure 1a and 1b) and indeed, we found a highly significant correlation of proportion of war topics in the English blogs, to body count: $R=.79$, $p<.0001$, $N=84$. This positive correlation reflects how, as the body count in Iraq rises, the proportion of war topics written in blogs about war also rises, and how as the body count drops, the number of war topics blogged also drops. Other topics rise as body count drops, such as topics concerning daily life. We consider the significant correlation we found to be an external validation of the topics identified as well as the use of the topic modeling technique.

A comparison of the English and Arabic blog war topics suggests that there is lag in Arabic posts (Figure 1a and 1c). Though we have no tangible explanation for this lag, we did note (during the manual analysis of posts) that Arabic posts tended to be more impersonal in nature and often relied on official reports of war events rather than personal accounts. Relying on official reports of events may mean that bloggers needed more time to find and synthesize appropriate reports. Other factors may have contributed to this lag. English posts may include those by Iraqi bloggers



a. Timeline of war topics in English **b. Body count** **c. Timeline of war topics in Arabic**

Figure 1: Timeline of Iraqi blog posts discussing war topics and civilian deaths (body count)

living abroad who have better access to the Internet (i.e. they are not constrained by power outages as those living in Iraq are). Thus, they may be able to post responses to war events more quickly than those within Iraq. More research is needed to ascertain the reasons for the difference.

War vs. daily life topics over time

Our first research question asked how blog topics change over the course of the war. With topic modeling, particular themes can be selected, tracked and analyzed over time. Other studies of blogs [cf 22, 36], examined in non-war environments, revealed that most blogs concern accounts about daily life. (We refer to daily life topics as those that concern aspects of people's daily life not related to war.) We examined the topics to see which addressed daily life. In Table 2, topic TE06 contains the following string of words: "god women Muslim love life man Islam men heart person religion woman Christian face eye." One interpretation of this topic might be concerning religion, or relationships—or in short, themes about daily life. Three independent coders determined the following topics as those concerning daily life (see Table 2): TE3, TE6, TE8 and TE10. These topics all contained words that concerned daily life, such as *friend*, *family*, *music*, etc.

Figure 2 shows the relation of daily life topics in comparison to the war topics over time (shown in Figure 1a). Initially, as the Internet and blogs first started being adopted in 2003, the discussion of daily life topics were higher in proportion to the war topics. However, we see that

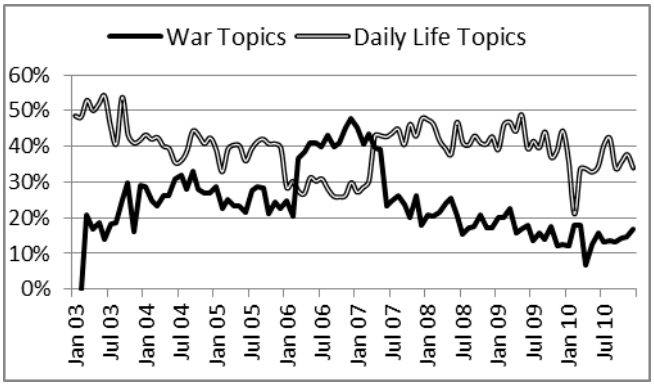


Figure 2: Percentage of topics concerning war vs. daily life over time

as the war progressed and as number of blogs increased, the discussion of daily life topics declined, especially as war topics increased as a proportion of all topics discussed in the blogs. The crossover point occurred just after December 2005, when the body count took a precipitous rise (see Fig. 1b), i.e. when violence of the war increased sharply. It was in December 2005 that the Iraq elections took place, with corresponding heightened violence.

A second crossover point occurred around April 2007. As the violence declined (see Fig. 1b), blogging of war topics decreased. However, what is especially interesting in Fig. 2 is the other perspective: blogging about daily life increased as the violence waned. This could reflect how when the crisis abates, people return to a sense of "normalcy," discussing topics about daily life.

Collective or personal expression of war?

Our second research question was to examine how identity is expressed in blog discussions about war. Following the idea that people form collective identity during trauma [2] combined with the idea that the use of first person plural pronouns ('we' and 'us') reflects the expression of a collective identity [34], we hypothesized that a collective identification with others about the war should be characterized by a high presence of first person plural pronouns in blogs.

We conducted an analysis of pronoun use in the Iraqi blogs as follows. We built a two-dimensional classifier of the posts. In the first dimension, a post can belong to one of three possible classes; (a) a war post if the sum of the distributions of war topics (TE01 and TE04) for that post exceeds 50%, (b) a daily life post if the sum of distributions of daily life topics (TE03, TE06, TE08 and TE10) for that post exceeds 50%, and (c) "other" if none of the previous two conditions is satisfied. In the second dimension, we defined an algorithm to distinguish 'I/me' from 'we/us' posts. A post was classified as an I/me post if the sum of the occurrences of the pronouns "I" and "me" in that post exceeds the sum of the occurrences of the pronouns 'we' and 'us', and vice-versa. We aggregated the posts by periods of equal lengths of one month.

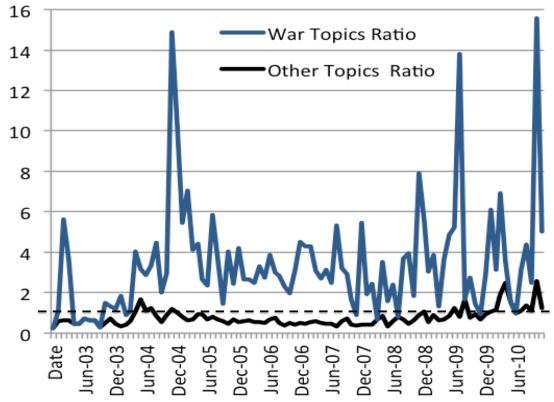


Fig. 3. Plot of ratios of 'We/Us' to 'I/Me' pronouns for War and all Other topics. Dotted line shows where ratio is equal.

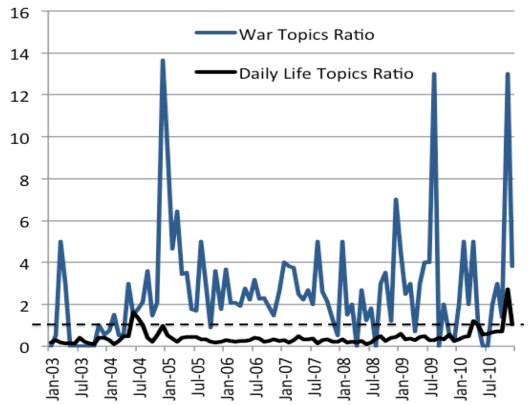


Fig. 4. Plot of ratios of 'We/Us' to 'I/Me' pronouns for War and Daily Life topics. Dotted line shows where ratio is equal.

We first examined first person plural pronoun use in the war topics compared to all other topics in our sample. We considered all the other topics as a "baseline" from which to compare the war topics. The dotted line in Figure 3 indicates where the ratio would be equal, i.e. where the first person plural and first-person singular pronouns would be used equally in the blogs. Figure 3 shows that the ratio of 'we/us' to 'I/me' pronouns exceeds a value of one for nearly all the blogs classified as war posts over the course of the war. This is consistent with the notion of reflecting a group or collective identity. What is interesting is that in the early war years, the ratio is less than one, which could indicate an individualistic reaction to a crisis, as Hoffman [24] found with victims in the Oakland firestorm.

However, it could be that blogging about any topic might reflect a collective identity. Figure 3 also shows that the war topics had far higher ratios of 'we/us' to 'I/me' pronoun use in relation to the rest of the blog sample, over nearly all the years of the war. A one-sample t-test of the difference of the ratios of war and all other topics over time (compared to a hypothesized mean of zero) shows that this difference is significantly greater than zero ($t(86)=8.23$, $p<.0001$). Thus, war blogs have a significantly higher use of first person plural pronouns compared to the rest of the blogs in the sample.

We then further contrasted the use of first person plural pronoun use for the war blogs compared to blogs about daily life. We cannot assume whether daily life topics would use a particular personal or collective voice, as the topics are so varied. Daily life topics may be written in a personal voice, as when one explains about their personal relationship or beliefs. On the other hand, daily life topics can also be about family, friends, or group activities, in which case they may use a collective, or group voice.

Figure 4 shows a plot of the ratios of 'we/us' to 'I/me' pronouns for the war blogs in relation to the daily life blogs. For daily life posts, the ratio is less than a value of one, over nearly all the years of the Iraq war. This ratio seems to trend towards a value higher than one in recent years, though research is needed to understand more fully why. A one-sample t-test of the difference of the war and daily life ratios over time shows that this difference is significantly greater than zero ($t(87)=9.7$, $p<.0001$). Thus, the use of first person plural pronouns is higher in war blog posts compared to the daily life posts, i.e. the type of posts consistent with those examined in previous studies of blogs in non-war environments.

DISCUSSION

Our goal was to examine a large corpus of archived blog data during a protracted crisis to try to gain a long term societal view of how people experience a war. We examined how blog topics changed over eight years of the war. We found an interesting relationship: as the war progressed over time, and as violent incidents increased, bloggers turned their attention to discussing the war. Even though there need not necessarily be an inverse relationship between blog topics of war and daily life, the proportion of blogging about daily life lowered as the war progressed. This suggests that when war is in the forefront of people's minds, especially when violent events occur, topics concerning daily life move to the background. During the height of crisis, bloggers discussed far more topics concerning the war, perhaps to make sense of the violence and uncertainty in the environment.

Theoretical models have been developed to describe the societal level processes that occur during one-time impact hazards; however, societal effects of long term crises have not been modeled. Powell [14] identifies eight sociotemporal stages ranging from predisaster to recovery. Closer to our work is that of Hoffman [24] who describes that people transition into recovery and adapt to the effects of a crisis in three stages: an individual stage where victims feel isolated, a collaborative stage where victims bond, and a third stage of return to normal life. Though Hoffman looked at a single disaster (the Oakland firestorm), our data could reveal stages of adaptation over a protracted crisis

like war. We interpret the finding of the interaction of daily life and war topics, in relation to external war events, as follows. As violence erupts, people try to maintain daily life routines, as evidenced by a higher level of daily life topics in blogs. As a war progresses over time with increased violence, there is a gradual slipping away from normalcy, as measured by a decrease in discussion of daily life topics and increase in war topics blogged. War moves to the forefront of the blogosphere as the society copes with the violence; a collective identity in blogging about war is evident. As violence subsides, or perhaps as a society becomes adapted to the war environment, people try to recover behaviors experienced before the war. Topics concerning daily life begin to shift back into the blog discourse. Thus, we believe that these results have potential to be used to monitor the reaction to a protracted crisis such as war. This model could be tested and refined empirically, e.g. through interviews.

Why might blogging about daily life be interpreted as a return to normalcy for bloggers? Giddens [18] can provide one interpretation, as he has discussed how humans continually strive for normalcy:

Routine is integral both to the continuity of the personality of the agent, as he or she moves along the path of daily activities, and to the institutions of society, which are such only through their continued reproduction (pg. 60).

Thus, a means to regain normalcy is through the performance of routines; returning to a focus on daily life is a sign of reconstruction of the routine. As such, our analysis suggests that blog discussions of daily life in a crisis might be used as a gauge for a society's return to normalcy. This result suggests a future research direction to look more closely at the relationship of daily life blog topics to actual attitudes in the broader population.

The results also show how bloggers seem to reflect different types of identity for distinct topics through their use of language. They appear to reflect more of a collective identity when they discuss war, as evidenced by the higher use of first person plural pronouns. We did not find the same incidence of first person plural pronouns in the rest of the blog sample nor in the blog posts specific to topics of daily life. It is possible however that any topic about a public event might employ a collective voice. We were not able to test this as the topic modeling did not produce other topics that we could classify as "public" topics. However, even if other public topics did produce a higher incidence of first person plural pronouns, this does not discount our results concerning war. Finding a higher use of first person plural pronouns with war topics is consistent with the claim that trauma is associated with a collective identity [2, 34].

Collective identity is established through interactions [12]; thus, to establish a collective identity people must have a means for interaction. In war environments it can be dangerous to meet others in public. Blogging is done in an online public forum and affords bloggers the ability to share their perspectives with others who are also experiencing the war. Online interaction may provide the channel for a collective identity concerning a war or crisis to emerge. But whether this identity formation is directly bolstered by interaction with other bloggers in a public forum who feel the same or is rather a result of bloggers presuming that what they feel is a shared experience, is not possible to say with our data; nevertheless, there does appear to be a collective perspective regarding the blog discussion of common war events to a great degree.

War differs from other crises like an earthquake or flood that have a contained emergency period. War involves a pattern of continual emergencies and disruption. No sooner are people trying to recover from one crisis, e.g. a bombing, when they may be faced with renewed violence. By concentrating on smaller-scale individual events, such as with qualitative analysis, it might be difficult to identify recovery of a society from the effects of a war. However, by taking a macro and longitudinal perspective, we found that archived blog data could suggest evidence of a generalized recovery in terms of the pattern of topics blogged. This suggests further research to examine what insights archived data can yield about crisis recovery.

Methods for large-scale analysis

As the volume of data in blogs—and on the Internet in general—swells, we need to look toward methods to examine large-scale data. Analytically we found reassurance that topic modeling can sort content into meaningful topics; this was validated by a significant positive correlation of war topic blog entries with real-world war events. This suggests that blogs might be a good representation for monitoring the effects of external events in a society. Though hypotheses about the connections between current events and blogs seem plausible, what remained unclear about the topic model analysis alone was evidence of the degree of the connection and detection of shifts over time.

Topic models can provide several key types of information for researchers. First, they can provide an overview of what topics are "hot" in the blogging community at particular time points. In the case of crisis events, this can help researchers understand what topics are perceived as important from the perspective of the blogger (as opposed to the traditional news media). They can benefit researchers in a variety of fields. For example, policy experts can use the information to understand reactions to the implementation of a particular policy. Distilled topics may even have a predictive value of events that will transpire in a society, e.g. if bloggers discuss dissatisfaction with the current government. Evidence of this can be found in the 2011 uprising in Egypt [30].

Topic modeling also has the advantage of revealing rapid insights, in contrast to qualitative analysis of data. They enable us to deal with a volume of data that cannot be analyzed manually. Moreover, that volume allows us to

discover low frequency topics that would certainly be missed in sampling. This is advantageous in the dynamic context of a conflict situation. Last, topic models can be used to understand the trajectory of different blogged topics over time and how they relate to external events. More research needs to be done to understand better how topic models can reflect broad views of a society over time.

Ideally, a researcher performing an ethnographic study of a conflict would want to travel to the location of the event. However, it is not always possible to travel to or stay in conflict zones because of the risk involved, the remoteness of the location, and/or the duration of the event. We believe that blog data provide an opportunity for researchers to understand from a distance the "mood" of a region that is experiencing a conflict over an extended or indefinite period of time.

Limitations

A limitation to the use of topic modeling is that the results (the topics) are subject to the interpretation of the analyst. For this reason we recommend having at least two independent coders to interpret the results and reconcile discrepancies in interpretation. We found it useful to have a researcher validate a sample of the topic results by hand.

A caveat in using blogs as a basis for generalizing to a broader society is that bloggers are not necessarily representative of the general population. As with users of any social media, they may have higher education and income. They are a self-selected group who may be more technically savvy than average. Therefore, our results must be used with caution in generalizing beyond the blogosphere. Also, we did sample some blogs by the same authors and this could have had resulted in a disproportionate representation by those authors. Yet most of the blogs in our sample had unique authors so this should not be such an uneven representation. Nevertheless, bloggers do provide a good sample of opinions about a region and event. That the topics correlated with an external societal event (body count) suggests that bloggers can represent a national discourse about a crisis to some degree.

CONCLUSION

War diaries have a long legacy. The translation of this practice to blogs by a large set of citizen writers affected by war opens up new opportunities to understand war, communication, and interaction. Challenges in analyzing such a large corpus of data require analysts to consider how to parse it in meaningful ways. Topic modeling offers promise, and the analysis offered here demonstrates that it maps to external indices that suggest validity. Furthermore, as would be expected, when wartime events are on the rise, written accounts of daily life move to the background. Finally, with the information topically parsed, linguistic analysis of pronoun use reveals an interesting finding that references to the collective 'we' also rise during wartime strife, suggesting fortification of a collective identity. These results stage the possibility for other analyses about the content of the writings of a society during life-changing periods in history.

ACKNOWLEDGMENTS

This project was supported by NSF grants #0910640 and #1128008.

REFERENCES

1. Al-Ani, B., Mark, G. and Semaan, B. (2010). Blogging in a region of conflict: Supporting transition to recovery. *Proc. of CHI'10*, ACM Press, 1069- 1078.
2. Alexander, J. C. (2004). Toward a theory of cultural trauma. In *Cultural trauma and collective identity*, in J. C. Alexander, R. Eyerman, B. Giesen, N. J. Smelser, and P. Sztompka. Berkeley: University of California Press.
3. Anderson, K. M. and Schram, A. (2011). Design and implementation of a data analytics infrastructure in support of crisis informatics research. Proc. *of the 33rd Intern'l Conf on Software Engineering (ICSE'11)*, 4 pp.
4. Baron, D. (2006). *Persistent media bias*. *Jour of Public Economics*, vol. 90, Issues 1-2, January 2006, 1-36.
5. Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet allocation. *J of Machine Learning Res*, 3:993–1022, 2003.
6. Blogpulse (2011). http://www.blogpulse.com
7. Bollen, J., Mao, H., and Zeng X.-J. (2010). Twitter mood predicts the stock market. arXiv:1010.3003v1.
8. Brewer, M. and Gardner, W. (1996). Who is this "We"? Levels of collective identity and self representations. *J of Personality and Social Psychology,* 71(1), 83-93.
9. Brown, R. and Gilman, A. (1960). The pronouns of power and solidarity. In Sebeok, T. A. (ed.) *Style in Language* Cambridge: MIT press. 253-76.
10. Buckwalter T. (2002). Buckwalter Arabic morphological analyzer Version 1.0. Linguistic Data Consortium, U of Pennsylvania. LDC Catalog No.: LDC2002L49.
11. Chesnut, M. (1981). *Mary Chesnut's Civil War.* C. Vann Woodward (ed.), New Haven: Yale Univ. Press.
12. Coles, Roberta L. (2002). War and the contest over national identity. *The Sociological Review* 28:586–609.
13. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis. *J of Amer Soc for Inf Sci*, 41: 391–407.
14. Dynes, R. (1970). *Organized Behavior in Disaster.* Heath Lexington, Lexington, MA.
15. Erickson, K. (1976). *Everything in Its Path*. New York: Simon and Schuster.
16. Foot, K. A, and Schneider, S. M (2004). Online structure for civic engagement in the September 11 web sphere, *Electronic J of Communication*, 14:3-4, 1-10
17. Gardner, W., Gabriel, S. and Lee, A. (1999). "I" value freedom, but "We" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science,* 10(4), 321-326.

18.Giddens, A. (1984): *The Constitution of Society*, University of California Press, Berkeley, CA.

19.Green P,Ward T (2009). The transformation of violence in Iraq. *Brit J of Criminology,* 49:628–645.

20.Griffiths, T., and Steyvers, M. (2004). Finding scientific topics. *Proc of the Nat'l Acad of Sciences*, 5228-5235.

21.Hall D., Jurafsky D., and Manning, C. (2008). Studying the history of ideas using topic models. *Proc of the Conf on Empirical Methods in Natural Language Processing,* Assoc for Computational Linguistics, 363-371.

22.Herring, S. C, Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information, Technology, & People*, 18(22), 142–171.

23.Hoffman, B. (2004), *Insurgency and Counterinsurgency in Iraq*, OP-127-IPC/CMEPP, Santa Monica: RAND.

24.Hoffman, S. (1999). The worst of times, the best of times: Toward a model of cultural response to disaster. In Smith and Hoffman (eds.), *The Angry Earth*, New York: Routledge.

25.Hoffman, T, (1999). Probabilistic Latent Semantic Indexing. *Proc of SIGIR'99,* 50-57.

26.Kamvar, S. and Harrison, J. (2009). *We Feel Fine: An Almanac of Human Emotion*. NY: Scribner.

27.Kireyev, Kirill, Palen, L. and Anderson, K.M. (2009). Applications of topics models to analysis of disaster-related Twitter data. *Neural Information Processing Systems Foundation Workshop*, Seattle, WA.

28.Knights, D., Mozer, M. and Nicolov, N. (2009). Detecting topic drift with compound topic models. *Proc of the 3rd Int'l Conf on Weblogs and Social Media*. AAAI Press.

29.Koppel, M., Argamon, S., and Shimoni, A. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Comp.,* 17(4), 401-412.

30.Lister T. and Smith E., 2011. http://articles.cnn.com/2011-01-27/world/egypt.protests.social.media_1_social-media-twitter-entry-muslim-brotherhood?_s=PM:WORLD.

31.Liu, S. B. (2011). *Grassroots Heritage: How Social Media Sustain the Living Heritage of Historic Crises*. Ph.D. Dissertation, University of Colorado at Boulder. Available at: http://sophiabliu.com/sophiabliu-dissertation.pdf.

32.Mardini, R. A., (2008). *Socializing Realism's Balance of Power: Collective Identity as Alliance Formation in Iraq*. Ph.D. Dissertation, Ohio State University.

33.Mark, G., Al-Ani, B., Semaan, B. (2009). Resilience through technology adoption: Merging the old and the new in Iraq, *Proc. of CHI'09*, ACM Press, 689-698.

34.Matthiesen, K. (2003). On collective identity. *Protosociology* 18, p. 66-88..

35.Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pp. 880-889.

36.Nardi, B., Schiano, D., and Gumbrecht, M. (2006). Blogging as Social Activity, or, Would You Let 900 Million People Read Your Diary? *Proceedings of CSCW'04*, 222-231.

37.Oliver-Smith, A. (1996). Anthropological research on hazards and disaster. *Annual Review of Anthropology*, 1996, 25:303-28.

38.Palen, L & Liu, S. B. (2007). Citizen communications in crisis: Anticipating a future of ICT-supported participation, *Proc. of CHI* 2007, 727-736

39.Palen, L., Vieweg, S., Liu, S., Hughes, A. (2009). Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007 Virginia Tech Event. *Social Science Computing Review*, Sage, 467-480.

40.Pax, S. (2003). *Salam Pax: The Baghdad Blog*, Grove Press.

41.Pennebaker, M. Mehl, R., and Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annu. Rev. Psychol*. 2003. 54:547–77.

42.Polletta, F., and James, J. (2001). Collective identity and social movements. *Ann Rev of Sociology*. 27:283–305.

43.Porter M F (2001). Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/introduction.html

44.Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011). Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. *Proc of CSCW '11*, 25-34.

45.Ramage, D., Dumais, S., and Liebling, D. (2010). Characterizing microblogs with topic models. *Proc. of ICWSM'10*, 130-137.

46.Starbird, K. and Palen, L. (2011). "Voluntweeters:" Self-organizing by digital volunteers in times of crisis. *Proc of CHI 2011*, Vancouver, BC.

47.Torrey, C., Burke, M., Lee, M.L., Dey, A.K., Fussell, S.R., and Kiesler, S.B. (2007). Connected giving: Ordinary people coordinating disaster relief on the Internet. *Proceedings of HICSS'2007*.

48.Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet*. New York, NY: Simon & Schuster.

49.Yano, T., Cohen, W., Smith, N. 2009. Predicting response to political blog posts with topic models. *Proc of the 2009 Annual Conf of the North American Chapter of the Assoc for Computational Linguistics,* 477–485.