Kimberly Glasgow, Clay Fink, and **Jordan Boyd-Graber**. **Our grief is unspeakable: Measuring the community impact of a tragedy**. *The International AAAI Conference on Weblogs and Social Media*, 2014, 9 pages.


```
@inproceedings{Glasgow:Fink:Boyd-Graber-2014,
Title = {Our grief is unspeakable: Measuring the community impact of a tragedy},
Author = {Kimberly Glasgow and Clay Fink and Jordan Boyd-Graber},
Booktitle = {The International AAAI Conference on Weblogs and Social Media},
Year = {2014},
Location = {Ann Arbor},
Url = {http://umiacs.umd.edu/~jbg//docs/2014_icwsm_grief.pdf},
}
```


Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

"Our Grief is Unspeakable": Automatically Measuring the Community Impact of a Tragedy

Kimberly Glasgow and **Clayton Fink**
Johns Hopkins University Applied Physics Laboratory
kimberly.glasgow@jhuapl.edu clayton.fink@jhuapl.edu

Jordan Boyd-Graber
University of Maryland iSchool and UMIACS
jbg@umiacs.umd.edu

Abstract

Social media offer a real-time, unfiltered view of how disasters affect communities. Crisis response, disaster mental health, and—more broadly—public health can benefit from automated analysis of the public's mental state as exhibited on social media. Our focus is on Twitter data from a community that lost members in a mass shooting and another community—geographically removed from the shooting—that was indirectly exposed. We show that a common approach for understanding emotional response in text: Linguistic Inquiry and Word Count (LIWC) can be substantially improved using machine learning. Starting with tweets flagged by LIWC as containing content related to the issue of death, we devise a categorization scheme for death-related tweets to induce automatic text classification of such content. This improved methodology reveals striking differences in the magnitude and duration of increases in death-related talk between these communities. It also detects subtle shifts in the nature of death-related talk. Our results offer lessons for gauging public response and for developing interventions in the wake of a tragedy.

1 Introduction

On December 14, 2012, twenty school children and six faculty members were shot at Sandy Hook Elementary School in Newtown, Connecticut. It was the deadliest primary school shooting in US history. Most victims were six years old. This tragic event affected the entire United States. It received extensive media coverage and was widely discussed on social media.

It also elicited powerful, personal expressions of pain and loss in the aftermath:

@gil Gil, I wish I could give you better news. Our friend, T.'s dear god-brother, is no longer with us. Our grief is unspeakable.

Successfully identifying such social media discussions could assist in assessing the depth of traumatic loss on a community and effect across a geographic region. It could help focus disaster mental health efforts, assess risk in indirectly affected communities, and help monitor community resilience and recovery. We describe the importance of assessing health and life impacts of disasters on communities, and how social media can serve as a proxy in Section 2.

Section 3 documents the broad adoption of Linguistic Inquiry and Word Count (Pennebaker, Francis, and Booth 2001, LIWC) for text analysis of a wide range of personality characteristics and emotional traits. While this approach is standard, elegant, and well-understood, it does not always capture relevant psychological properties in text. For example, the author of a tweet such as *My throat is killing me.* is experiencing discomfort, but almost certainly will not die. Section 4 describes how death is discussed in social media.

We propose an alternative method to LIWC, using data-driven active learning (Section 5). By annotating a handful of messages and building a supervised classifier, we quickly and accurately classify hundreds of thousands of social media messages (tweets) to determine which are truly about death, and which are "false positives", unrelated to the event.

In Section 6, we show that our classifier allows us to compare communities that directly or indirectly experienced a traumatic mass casualty event. Because we have accurately discriminated social media messages truly relating to death, we can observe stark differences between communities after the tragedy. Those who directly experienced the event spoke far more frequently about it in social media, and sustained elevated rates for a substantial period of time. We also observe more subtle shifts in the directly affected community, perhaps indicative of how they are sensitized to the losses and suffering being experienced by their fellow citizens. These differences are less apparent or even undetectable through LIWC. Thus, our method more effectively measures the effects of traumatic events and the later resilience and recovery using social media.

2 Trauma and Disaster in a Community

We focus our work on a mass violence event that directly affected tens of thousands and indirectly affected millions. We examine how affected communities differ on two critical factors: geographic proximity and directness of threat to loved ones (Bonanno et al. 2010). Disasters are traumatic events that are collectively experienced. Mass violence events affect communities more deeply than other forms of disasters, like natural disasters or technological disasters, such as oil spills (Norris et al. 2002). A traumatic event typically involves exposure to "death, threatened death, actual or threatened serious injury, or actual or threatened sexual violence" (American Psychiatric Associa-

tion 2013). Unexpected and intense, it can trigger feelings of helplessness, fear or horror, and have both short and long term psychological and physiological consequences (Stith Butler, Panzer, and Goldfrank 2003; Ursano 1995).

Prior research on disasters has employed standard instruments to assess the impact of these tragedies on individuals. The Impact of Event Scale-Revised (IES-R) and the Symptom Checklist 90 Revised (SCL-90-R) have been the most frequently used instruments in the field (Norris et al. 2002). The IES-R is a well-established psychological instrument to measure subjective distress caused by traumatic events (Weiss and Marmar 2004). The SCL-90-R is a screening measure for general psychiatric symptomatology.

However, these instruments have limitations. They are subjective and are completed by individual respondents retrospectively, at a single point in time after the event. No pre-event baseline rate of symptomatology is available for the population. Respondent sample sizes have commonly numbered in the hundreds or fewer in studies of disaster (Norris et al. 2002). Social media could complement such instruments, providing streaming, natural observations from a vast number of individuals.

Individual suffering is only one component of disaster. Whole communities can be damaged and unable to provide necessary services. Community resilience is the ability of a community or social unit to withstand external shocks to its infrastructure (Norris et al. 2007) such as those caused by a mass traumatic event. Community resilience emerges from the ability to adapt to stress and return to healthy functioning. The speed with which a community can mobilize and use resources during and following a disaster event is strongly dependent on its various capacities to adapt to change and is related to the strength of its social networks (Sammantha L. Magsino 2009). Measuring resilience is complex, and social media provides a unique channel for assessing adaptation and return to health.

2.1 Social Media as a Proxy for Assessing Health and Life Impacts of Traumatic Events

There is a clear need for improvements to existing screening and surveillance tools and procedures to support disaster behavioral health response and for research in the areas of risk, resilience, and other factors relating to recovery (Watson, Brymer, and Bonanno 2011). Such methods must consider measures aimed at the broader community level and not focus solely on the individual. Social media provide options in this regard.

Merely assessing the presence or absence of psychopathology in a population in the aftermath of disaster is not enough to understand how communities react to disaster. More important is measuring functioning in work/school settings, observing healthy patterns of behavior, and overall quality of life. These are better signals for determining when life returns to "normal" or a "new normal" is found (Norris et al. 2007).

Moving beyond the results of psychological instruments, we can examine patterns in language use for information about individual and group health and well-being. Differences in language use have been associated with emotions and psychological states (Tausczik and Pennebaker 2009) and psychological conditions such as depression (Rude, Gortner, and Pennebaker 2004). There is evidence for subtle, perhaps subconscious shifts in language use in certain contexts, such as social bonding and courtship (McFarland, Jurafsky, and Rawlings 2013), and power differentials (Danescu-Niculescu-Mizil et al. 2012).

Patterns in language use have also been related to the study of sentiment, opinion, and subjectivity (Pang and Lee 2008). While we do not distinguish between objective and subjective posts about death, the relationships between word sense, context, and interpretation (Wiebe and Mihalcea 2006; Turney et al. 2011) are highly relevant to our classification of death-related tweets.

A mass traumatic event has far-ranging public health implications. From a public health perspective, social media is a real-time source of information about the thoughts, feelings, behaviors, symptoms, perceptions, and responses to events for a population. It can both augment existing public health capabilities, such as surveillance, and create new capabilities that use the spontaneous expressiveness of the population (Dredze 2012).

For example, Twitter's public sharing helps us understand the health impacts of major life events (De Choudhury, Counts, and Horvitz 2013). Twitter provides a timely source of immediate reactions to events from a large number of individuals. Twitter data have been successfully applied to track responses to unfolding natural disasters (Starbird and Palen 2011), and to human-caused disasters like the London riots (Glasgow and Fink 2013).

Social media are also transforming how we express and experience grief and mourning. They provide new mechanisms and affordances to move through the tasks and stages associated with loss, such as accepting the reality of the loss, working through the pain, readjusting to the environment, and reinvesting in life and forming a continuing bond to the deceased (Falconer et al. 2011).

For these reasons, we examine social media produced before and after the Newtown school shootings. Before describing the data in Section 5, we first discuss existing baselines for discovering psychologically-relevant features from text.

3 LIWC: The Standard for Psychological Text Measurement

Linguistic Inquiry and Word Count (LIWC), a text analysis program, has been used for assessing text for a range of social and psychological phenomena (Pennebaker, Francis, and Booth 2001). LIWC's central premise is that words people use reveal their mental, social, or emotional state. A quantitative approach to text analysis, LIWC was developed in 2001 to support measurement of language from a psychologically-informed perspective.

LIWC is the standard tool for text-based studies of personality, mood, emotion, and self-esteem, mental health and psychopathology (including depression and suicide), social processes, personal and shared distress and upheaval, changes in psychological health, and effectiveness of ther-

apy (Pennebaker, Mehl, and Niederhoffer 2003). Tausczik and Pennebaker (Tausczik and Pennebaker 2009) catalog 121 publications in the psychological literature based on LIWC word analysis, and dozens more are published each year. There is no other automated text analysis method used as broadly or frequently. Research examining social media from a social or psychological perspective has also adopted LIWC as a tool (Golbeck, Robles, and Turner 2011; Golder and Macy 2011; Brubaker et al. 2012).

Functionally LIWC is simple. LIWC is organized into a set of dozens of categories which contain word stems. For example, the *positive emotion* category contains "happy" and "happi*". For a document, LIWC provides the percentage of the terms associated with each category.

3.1 Relating LIWC to Death

LIWC has been applied to traditional media to examine psychological response to a mass death event. In this case, the collapse of a traditional campus bonfire killed 12 college students. Gortner et al. (2003) applied LIWC to articles from the campus newspapers of the afflicted university and a nearby one. The campus where the students died used fewer words from the *death* category than the comparison university.

The LIWC *death* category concerns death and dying. Its dictionary contains twenty-nine stems of content words including dead, burial, and coffin. The application of any dictionary to natural language will of course have limitations. Measurement of discussion of death using the LIWC *death* dictionary will have errors of both precision and recall.

Errors of precision arise due to polysemy, reference to proper names, or other creative or figurative use of language. For example, the word "dead" has 21 WORDNET (Miller 1990) senses, only two of which refer to death in a literal sense (no longer living or having life). Usage of dead in any of the other senses ("came to a dead stop", "a dead battery") would not reflect discussion of death or dying. Errors of recall occur because people will use terms outside the 29 enumerated words/stems to talk about death.

Moreover, LIWC's inventory of death-related terms is shallow: it does not reflect important distinctions of how death is discussed in social media. In the next section, we develop a more comprehensive taxonomy of death-related discussion on social media.

4 Before and After: Discussions of Death

We examined Twitter data from two communities: Fairfield County, Connecticut and Montgomery County, Maryland from December 2012 until mid-January 2013. Fairfield was directly affected by the tragedy, while Montgomery county was only affected indirectly. The communities differed with respect to geographic proximity to the event, as well as actual or perceived risk to self or loved ones.

We gathered user content from Twitter for Fairfield County, Connecticut using Twitter's Search API. We collected the tweets of users with any tweet from a location in Fairfield county, producing 360,000 tweets over the period. We repeat the process for an area centered on southern Montgomery County in Maryland to provide a comparison population. This dataset contains approximately 460,000 tweets. Both areas are well-off: high median income, high educational attainment, and low poverty (U.S. Census Bureau 2013). Both counties serve as bedroom communities for big cities (New York City and Washington, DC).

Not every mention of death, dying or killing refers to real-world loss of life. The "death" metaphor can express amusement or exasperation or be used to talk about mundane events and experiences. We split these usages into three categories: literal, figurative, and proper name usages. We describe these categories in more detail and provide examples in the following sections. Wherever appropriate throughout this paper, we have anonymized references to person's full names or Twitter handles.

4.1 LIWC as a Baseline for Death-related Tweets

LIWC provides us with a first pass of how often death is discussed in these two areas. One simple heuristic is to consider any tweet mentioning a term from LIWC's *death* category as being about death.

Using this metric, death is an infrequent topic, both before and after the tragedy. In Montgomery County, any word from the LIWC *death* list appears in 1.76% of the total tweets preceding the day of the school shootings, and in slightly higher (1.82%) afterwards. For Fairfield, the county where the tragedy occurred, death is discussed in 2.08% of the tweets before the shootings, and in 2.25% after.

We used standard z-scores to compare how both communities differed from their normal patterns of tweeting:

$$z_{n,i} = \frac{x_{n,i} - \mu_i}{\sigma_i}, \tag{1}$$

where μ_i is the average value and σ_i is the standard deviation. We compute the mean and standard deviation for the days preceding the shooting on December 14$^{\text{th}}$. Figure 1 illustrates the results, standardizing for the entire period. Death-related tweets spike on the day of the shooting for both communities. Fairfield jumps seventeen standard deviations, and Montgomery increases by nearly five. This increase dropped off within the week to rates close to baseline for both communities.

4.2 A Taxonomy of Death-Related Tweets

However, as previously noted, reliance on LIWC for measuring death-related discussion will fail on both precision and recall. To explore this and more accurately identify tweets in these data that explicitly discuss actual death, such as the murders of children in Newtown, we create and test a classifier that can distinguish between literal references to death, figurative references, and references to proper names that incorporate a death-related word. This classifier is based on the following taxonomy.

Literal Death Tweets in this category are about 'real' death. They relate to the end of life of a human or some other living organism. They may describe accoutrements of death such as caskets or funerals, or mention specific kinds or manners of dying or killing (homicide, suicide).

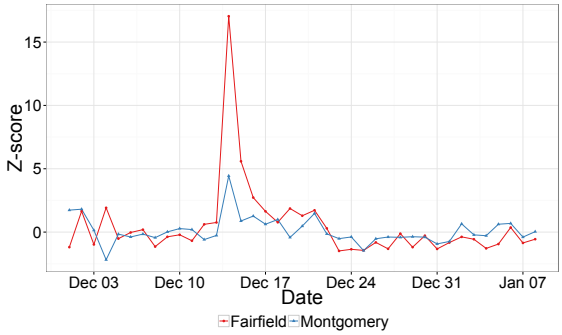

Figure 1: LIWC-based measurement of death tweeting in directly (Fairfield) and indirectly (Montgomery) affected communities shows increased activity on the day of the school shooting, and a return to baseline within days.

- *Dave Brubeck died yesterday. A truly significant figure in music.*
- *Reports are saying a teacher has been shot in #Newtown #CT. Hoping for no fatalities.*
- *RT @L: make it stop I can't take it. #rip #newtownfamily*
- *rest in peace to one of gods newest angels, daniel. just keep swimming buddy. lessons this summer wont be the same without you.*
- *Yo Westboro Baptist Church, come up to picket these children's funerals and watch me throw a bible in your face.*

Figurative Death These tweets reference death or killing in a metaphoric or figurative fashion. They do not involve actual loss of life of a biological organism. They may discuss cessation of function or performance of devices or machines, or the end of existence or relevance for some event, activity or organization. They may hyperbolically describe an emotional response, or otherwise intensify a description. Alternately, they may be jokes, or references to fictional events or characters.

- *Kill me now! (@ Macy's)*
- *The last thing I ever wanted was to wake up sick today and yet here I am, nice and stuffy with a killer sore throat*
- *Well some good came out of my car dying..at least I didn't have to go to my doctor's appointment!*
- *#BigBangTheory is killing me! #hilarious*
- *I love her to death and she knows it*
- *patrick on the swim team just killed it at pfeiffer idol. just got goose pumps no joke!*
- *Penalty has been killed! Full strength with 1:00 to play*

Proper Names Tweets that mention named entities with death or killing-related names fall in this category. Commonly these entities are musical groups, or books, television shows, or movie titles. They may be named individuals or organizations.

- *I want to see Texas Chainsaw Massacre 3D on January 4!*
- *HEY GUYS. GUESS WHAT!? I AM SEEING THE KILLERS TOMORROW. AND YOU'RE NOT. MUAHAHAHAHAHA MY LIFE IS AWESOME*

5 Beyond LIWC: Data-Driven Classification

In this section, we use the tools of computational linguistics to go beyond LIWC's simplistic determination of whether a particular piece of text discusses death or not. Since death is generally a rare topic on Twitter, we create a dataset that includes sufficient death-related tweets to train a classifier. We quickly develop a supervised classifier through an active learning paradigm. This classifier uses text features and tweet-level annotations to produce classification probabilities for each of the death categories. Coupling the output of that classifier with the LIWC *death* binary feature provides a simple and accurate way to determine if a tweet is truly about death, has figurative references to death, incorporates a mention of death in a proper name, or is entirely unrelated to the topic.

5.1 Creating a Balanced Dataset

Over 95% of tweets from Fairfield and Montgomery counties are not death-related, based on the values predicted by LIWC. Thus, we face a common challenge for real-world classification tasks. The classes we are interested in are only a small percentage of the actual data. A randomly selected dataset would most likely contain too few instances of the three classes of death-related tweets to effectively train and test a classifier. Further, Twitter's patterns change after tragic events like Sandy Hook. To address this, we consider tweets preceding the date of the shootings (December 14th) and those occurring on or after the 14th separately.

We use the LIWC *death* feature as a proxy to ensure sufficient death-related tweets are initially in the dataset. We randomly oversample tweets with this feature to address class imbalance (Van Hulse, Khoshgoftaar, and Napolitano 2007). To compensate for LIWC's low recall, we use pointwise mutual information (PMI) to identify terms that frequently co-occur in tweets with LIWC death dictionary terms, such as RIP (rest in peace), *RIPangels*, and *gunman*, and include additional tweets with these terms. To avoid training the model with terms specific to this event, we do not include high-PMI terms that mention a place name or person associated with the shootings. We randomly select tweets based upon these criteria.

After we performed the annotation described in the rest of this section, we were able to estimate proportions of death-related tweets in this collection. The final balanced dataset contains 5014 Fairfield tweets annotated for ground truth. This dataset is used for training and testing the model. Literal death tweets comprise 21%, figurative death 20%, proper names 2%, with the remainder unrelated to death. Inter-annotator agreement on coding is high at 0.94 (Cohen's $\kappa = 0.91$) for a randomly selected set of 200 tweets from this dataset.

5.2 Training the Classifier

First, we used DUALIST (Settles 2011), an interactive, active learning framework for both document-level annota-

tions and feature labeling for text classification. In addition to unigrams and bigrams, it incorporates Twitter-specific features such as emoticons, usernames, URLs, and hashtags. DUALIST uses a multinomial naïve Bayes model. We use four classes. Three distinguish between documents that refer to literal death, figurative death, and proper names that incorporate death-related terms. The fourth class contains all other, non-death related documents. Each document contains the words from a single tweet. DUALIST has performed successfully at a variety of classification tasks, including word sense disambiguation. The training data contain over 50,000 features. From this, DUALIST generated four scores (probabilities), one for each class, for each document. The highest-scoring class can be considered the label generated by the model for that document.

This model substantially outperforms LIWC at labeling literal death tweets (F_1 of .78 and .61 respectively, on test data described below). We observe, however, that the model's accuracy diminishes as the probability assigned to the most probable class decreases, which suggests that reranking the confidence may improve performance, especially given the biased nature of the full datasets.

Combining generative models' generalizability and with discriminative models' precision improves task performance (Shen et al. 2006; Fujino, Ueda, and Saito 2005; Mullen and Collier 2004); we adopt this approach. In our case, we employ a support vector machine to correct the false confidence of a naïve Bayes model trained on a skewed training set. We used the classes and scores of our initial model, as well as the LIWC *death* feature, to train the SVM using a 70-15-15 split, and a radial basis kernel function (Joachims 1999). This new model performs significantly better than chance at classifying tweets (0.90 accuracy across all classes, versus the 0.57 attainable by automatically assigning all tweets to the most frequent class) and outperforms the previous DUALIST multinomial naïve Bayes model. It achieved 0.86 precision and 0.82 recall for the literal death tweets, the class that contains tweets that are truly about death.

	Literal	**Fig.**	**Proper**	**Other**
Precision	0.86	0.85	0.50	0.95
Recall	0.82	0.74	1.00	1.00
F1	0.83	0.79	0.67	0.97

Table 1: Performance of SVM model at classifying tweets related to death (based on test data)

Since the LIWC *death* feature was used in selecting tweets for training, it is unsurprising that LIWC achieves high recall for literal death (0.95) on test data. However LIWC precision is poor for literal death (0.45), and F_1 is only 0.61. We do not present results for figurative or proper name death, as LIWC cannot distinguish these usages. And we expect that if we had included a larger proportion of tweets containing high-PMI terms in the training and test data, LIWC's performance would have been negatively affected.

6 More, and Different, Death Talk in Directly Affected Communities

From the full datasets, the final classifier identified 11,444 tweets from Fairfield County as literal death, and only 3721 from Montgomery. Both values are substantially different from the counts produced by running LIWC on these datasets (8695 and 8397 tweets). Furthermore, many of the LIWC-identified death tweets are likely to be false positives, indicative of figurative usage, based on the performance seen against test data.

As before, we compute a baseline percentage and standard deviation for the literal death tweets for the days preceding the shooting. We use these values to determine z-scores. These are used to show the magnitude of variation from the baseline in rates of tweeting for the literal death class. Figure 2 compares these results with the LIWC death results seen in Figure 1. Particularly for Fairfield, the differences are dramatic. On December 14th, the community that directly experienced the shootings spiked at a staggering 130 standard deviations above the prior baseline rate of literal death tweeting. The increase in Montgomery County was 27 s.d. Both increases dwarf the LIWC estimates (17 and 4.5, respectively). In Fairfield, this elevation drops slowly, and rates of literal death tweeting remain above baseline into January. The LIWC results mask the magnitude and duration of the increase in actual death talk, and are least accurate for the more deeply affected community and during the early aftermath of the disaster.

It is reassuring that the LIWC death variable is indeed able to identify a spike in death-related tweets. It clearly detected a change in communicative behavior after a traumatic mass shooting. This helps further validate its earlier role in selecting tweets for use in building our model, and its inclusion as a feature for the model. However, it has less applicability for more complex or granular tasks. Its inability to distinguish the non-literal usage of terms in its death vocabulary poses an additional limitation.

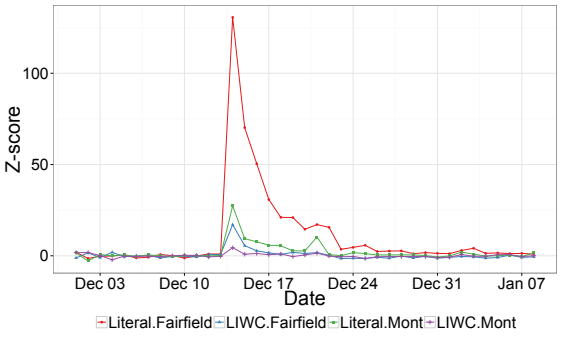

Figure 2: Tweeting about death increases after the shooting, particularly in Fairfield County. The SVM classifier detects a 130 standard deviation increase on the day of the shooting, while LIWC predicts only 17 s.d. above baseline. Montgomery is less strongly affected. (Full datasets for both counties.)

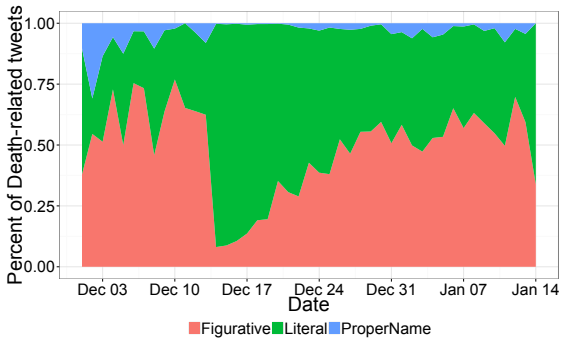

Figure 3: Fairfield County has a substantial drop in the proportion of figurative death usage after the tragedy, perhaps showing sensitivity in the aftermath of actual killings.

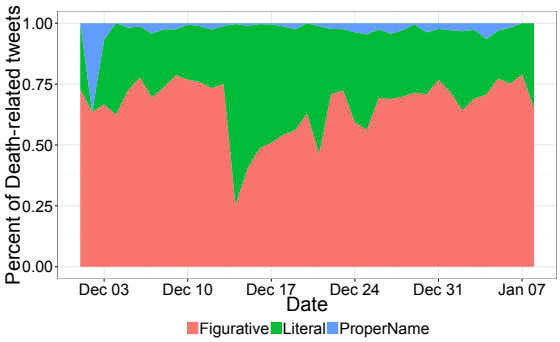

Figure 4: A smaller, shorter decrease in relative proportion of figurative death is seen in Montgomery County, the county that did not sustain losses in the shootings.

6.1 Shifts in Proportion of Figurative Usage of Death

We expect the amount of tweeting of death proper names to be relatively unaffected, with fluctuations that relate to events, such the broadcast of a TV show episode, a movie release, or a concert. Prior to the shootings, figurative usage was the most common. In Fairfield, it accounts for nearly two-thirds of the tweets identified by the SVM classifier as belonging to one of the three death classes. However, if people perceive that figurative references to death (*kill me now*, *hahahaha dying*) are insensitive or inappropriate while they and their fellow citizens are coping with the traumatic loss of dozens of their friends, family and neighbors, we should see a dip in this type of usage. Indeed, the proportion of tweets using death figuratively drops considerably in Fairfield immediately after the shootings, and only gradually climbs back. It trends below its baseline mean as a percentage of total tweets for the remainder of the time frame (an average of $0.34z$ below the baseline on 81% of subsequent days (Poisson $p < .01$). This is shown in Figure 3.

For Montgomery, the results from the classifier are similar, though more subdued. An immediate drop in figurative usage was visible, but the drop was neither as deep nor as sustained in this community that was only indirectly affected by the tragedy. Just ten days after the shootings, figurative usage again comprises 59% of Montgomery death talk, though only 39% in Fairfield.

In comparison to its baseline mean, the impact in Montgomery on the figurative death tweet rate as smaller. It fell an average of $0.17z$ below the baseline on 68% of subsequent days. Crucially, these shifts cannot be detected through LIWC at all, as its classification cannot distinguish literal death from other usages.

Prior work has identified expressions of humor after natural disasters as a factor in resilience for disaster survivors (Garrison and Sasser 2009), and the role of gallows humor in medical personnel and first responders exposed to violent death has been documented, though such humor tends to be expressed outside public view (McCarroll et al. 1993). While figurative usages of death in tweets were commonly made for humorous effect, we observe no occurrences of humorous reference to the school shootings whatsoever in the public world of social media as generated in Fairfield County.

6.2 Error Analysis

This classification task is not trivial. Human annotators themselves find some tweets ambiguous or difficult to classify. While the SVM performs well overall, it does make errors. In some cases, these errors reflect intrinsic challenges relating to the task, while in others, additional training or features might improve performance. These errors were observed in test data.

Tweets that fit multiple classes The system and the human annotator assign each tweet to a single class. Inspection of tweet content reveals a more complex reality. Human authors may employ both figurative and literal senses of death-related words in a single tweet.

- *RT @A: States slowly killing capital punishment...*
- *RT @B: It kills me to hear of the tragedy in CT. My prayers are with all of you*

In these cases, the human annotator labeled the tweet with one of the two classes that actually appeared, and the classifier labeled it with the alternate class. This is a basic challenge for any classifier that assigns a single label to an item.

Lack of training data In a few cases, an unambiguous reference to death like an obituary notice was misclassified as a figurative usage. In other cases, a specific manner of killing (e.g., "decapitation") was missed. A lack of data during training is a likely cause, and adding a simple feature (e.g., "obituary") during training would remedy this error.

- *Obituary: Rose Marie M. S., Age 83*
- *RT @C: An adult giraffe has a kick so powerful, they can decapitate a lion.*

Death as a source of humor or sarcasm Jokes and ironic or snarky comments are common on Twitter. These should be classified as figurative usages, since no true death of a

living creature is involved. Automatically recognizing humor, irony, and related phenomena in natural language is a complex problem that poses a challenge to automated approaches (Reyes, Rosso, and Buscaldi 2012).

- *RT @D: Kyle W. is like an appendix. Completely useless and you only know he's there when he's killing you. #Nyjets*

Figurative features appearing in literal contexts Many figurative usages of death include various indicators of excitement or emotion, such as *hahaha* and LMAO (laughing my a** off). OMG (Oh my god) is another such feature. When these features appear in a literal context, the classifier may mislabel the tweet.

- *Oh my god the numbers of deaths are still rising*
- *@F: You can live for weeks without eating, but will most likely die after 11 days without sleep. @G OMG, GET SLEEP!!*

7 Discussion

The Sandy Hook school shootings took the lives of 20 school children and 6 faculty. The toll on the community was brutal, as one community member who lost a loved one poignantly expressed: "Our friend is no longer with us. Our grief is unspeakable."

We consider one aspect of this tragic mass shooting - the explicit discussion of death in social media. We develop a classifier that produces a more complete, accurate, and nuanced understanding of this phenomenon. We are able to distinguish between literal discussion of death and usages that are not truly related to loss of life.

This provides a window into the community perception of and response to loss, with excellent temporal granularity. A community stricken by disaster has large, sustained rates of discussion of death compared to its pre-event baseline. It also shows a subtle shift away from figurative usages of death terms that might seem to trivialize profound, tragic loss. Fairfield never quite returns to "normal" over the timeframe we study.

These effects could not be detected by commonly used psychological instruments. For the Impact of Event Scale-Revised (IES-R), respondents rate how frequently they experienced symptoms such as intrusive thoughts or avoidant behavior in the previous week (Weiss and Marmar 2004). They score the frequency of each item from 0 (not at all) to 4 (extremely frequent). For the SCL-90-R, individuals rate their experience with each of 90 symptoms over the past week on a five-point scale. At best, week-over-week change can be assessed for an individual if the instrument is repeatedly administered. Twitter provides natural observations from the population time stamped to the second.

Consistent with theory on community response to disaster (Bonanno et al. 2010), a community that was geographically distant and not directly at risk showed smaller shifts in rates of discussion of death in social media. We hypothesize that our approach and these findings will generalize across other traumatic mass events.

This work could be extended to contribute to our understanding of sense-making in the aftermath of violent loss, and the development of complicated grief (Currier, Holland, and Neimeyer 2006). In our time frame of only a month following the mass shooting, we observe numerous expressions of the senselessness of the tragedy. We also find discussion of the difficulty of constructing an understanding of the experience of violent loss:

> *It touched everyone in a different way&no one will ever be able to understand why little innocent lives were taken & that's the hardest part*

We find that while LIWC clearly adds value to our understanding of texts, its *death* variable has major limitations when applied to Twitter, and perhaps any genre where texts are short, and language use is informal and creative. These are typical features of social media.

We focus our examination of the impact of a traumatic mass event specifically on posts relating to death. Of course, the actual social media response to the tragedy was richer and broader. While outside the scope of this paper, we observed striking differences in the use of Twitter hashtags between the direct and indirectly affected community.

In many cases these differences in Fairfield appear to reflect the use of social media to amplify and sustain community solidarity and cohesion. In the weeks following the shooting, death-related tweets discussed and publicized local memorial and tribute events. They also pushed back against outsider attempts to frame them:

> *Sandyhook(Newtown) is a community not an incident. Please refer to the tragedy as the "12-14 elementary school shooting"*

Analysis of social media may thus augment survey-based efforts to assess community solidarity after a mass tragedy (Hawdon and Ryan 2011).

Topic models hold promise for illuminating other aspects of community behavior. For example, after running a topic model (Blei, Ng, and Jordan 2003) using MALLET (McCallum 2002) on Fairfield death-related tweets, initial analysis identified one topic that seemed to capture wearing Sandy Hook's school colors as a response to the killings. This is a symbolic way to show cohesion within the community in the face of adversity.

Applying machine learning to social media augments what can be learned by administering established instruments for assessing psychopathology in the aftermath of disaster. It allows us to extend our knowledge and can contribute to the development of new theory relating to loss and resilience, and may suggest ways to help communities heal.

Acknowledgments We thank the anonymous reviewers and Philip Resnik for their insightful comments. Boyd-Graber is supported by NSF Grants CCF-1018625 and IIS-1320538. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 5th edition.

Blei, D. M.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Bonanno, G. A.; Brewin, C. R.; Kaniasty, K.; and Greca, A. M. L. 2010. Weighing the costs of disaster: Consequences, risks, and resilience in individuals, families, and communities. *Psychological Science in the Public Interest* 11(1):1–49.

Brubaker, J. R.; Kivran-Swaine, F.; Taber, L.; and Hayes, G. R. 2012. Grief-stricken in a crowd: The language of bereavement and distress in social media. In *ICWSM*.

Currier, J. M.; Holland, J. M.; and Neimeyer, R. A. 2006. Sense-making, grief, and the experience of violent loss: Toward a mediational model. *Death studies* 30(5):403–428.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, 699708.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, 32673276.

Dredze, M. 2012. How social media will change public health. *Intelligent Systems, IEEE* 27(4):8184.

Falconer, K.; Gibson, K.; Norman, H.; and Sachsenweger, M. 2011. Grieving in the internet age.

Fujino, A.; Ueda, N.; and Saito, K. 2005. A hybrid generative/discriminative approach to semi-supervised classifier design. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INLIGENCE*, volume 20, 764.

Garrison, M. B., and Sasser, D. D. 2009. Families and disasters: Making meaning out of adversity. *Lifespan Perspectives on Natural Disasters: Coping with Katrina, Rita, and Other Storms* 113.

Glasgow, K., and Fink, C. 2013. Hashtag lifespan and social networks during the london riots. In Greenberg, A. M.; Kennedy, W. G.; and Bos, N. D., eds., *Social Computing, Behavioral-Cultural Modeling and Prediction*, number 7812 in Lecture Notes in Computer Science. Springer Berlin Heidelberg. 311–320.

Golbeck, J.; Robles, C.; and Turner, K. 2011. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, 253262. New York, NY, USA: ACM.

Golder, S. A., and Macy, M. W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881. PMID: 21960633.

Gortner, E.-M., and Pennebaker, J. W. 2003. The archival anatomy of a disaster: Media coverage and community-wide health effects of the texas A&M bonfire tragedy. *Journal of Social and Clinical Psychology* 22(5):580603.

Hawdon, J., and Ryan, J. 2011. Social relations that generate and sustain solidarity after a mass tragedy. *Social Forces* 89(4):1363–1384.

Joachims, T. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. chapter 11.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit.

McCarroll, J. E.; Ursano, R. J.; Wright, K. M.; and Fullerton, C. S. 1993. Handling bodies after violent death: Strategies for coping. *American Journal of Orthopsychiatry* 63(2):209–214.

McFarland, D. A.; Jurafsky, D.; and Rawlings, C. 2013. Making the connection: Social bonding in courtship situations [1]. *American Journal of Sociology* 118(6):1596–1649.

Miller, G. A. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography* 3(4):245–264.

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, 412418.

Norris, F. H.; Friedman, M. J.; Watson, P. J.; Byrne, C. M.; Diaz, E.; and Kaniasty, K. 2002. 60,000 disaster victims speak: Part i. an empirical review of the empirical literature, 19812001. *Psychiatry: Interpersonal and Biological Processes* 65(3):207239.

Norris, F. H.; Stevens, S. P.; Pfefferbaum, B.; Wyche, K. F.; and Pfefferbaum, R. L. 2007. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American Journal of Community Psychology* 41(1-2):127–150.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1135.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology* 54(1):547–577. PMID: 12185209.

Reyes, A.; Rosso, P.; and Buscaldi, D. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering* 74:1–12.

Rude, S.; Gortner, E.-M.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18(8):1121–1133.

Sammantha L. Magsino, R. N. R. C. 2009. *Applications of Social Network Analysis for Building Community Disaster Resilience:Workshop Summary*. The National Academies Press.

Settles, B. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 14671478.

Shen, J.; Li, L.; Dietterich, T. G.; and Herlocker, J. L. 2006. A hybrid learning system for recognizing user tasks from desktop activities and email messages. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI '06, 8692. New York, NY, USA: ACM.

Starbird, K., and Palen, L. 2011. Voluntweeters: self-organizing by digital volunteers in times of crisis. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, 10711080.

Stith Butler, A.; Panzer, A. M.; and Goldfrank, L. R. 2003. *Preparing for the Psychological Consequences of Terrorism: A Public Health Strategy*. Washington, D.C.: The National Academies Press.

Tausczik, Y. R., and Pennebaker, J. W. 2009. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

Turney, P. D.; Neuman, Y.; Assaf, D.; and Cohen, Y. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, 680690.

Ursano, R. J. 1995. *Individual and community responses to trauma and disaster: the structure of human chaos*.

U.S. Census Bureau. 2013. State & county quickfacts.

Van Hulse, J.; Khoshgoftaar, T. M.; and Napolitano, A. 2007. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, 935942.

Watson, P. J.; Brymer, M. J.; and Bonanno, G. A. 2011. Postdisaster psychological intervention since 9/11. *American Psychologist* 66(6):482–494.

Weiss, D. S., and Marmar, C. R. 2004. The impact of event scale-revised. *Assessing psychological trauma and PTSD* 2:168189.

Wiebe, J., and Mihalcea, R. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 10651072. Stroudsburg, PA, USA: Association for Computational Linguistics.