Soc. Netw. Anal. Min. (2014) 4:206
DOI 10.1007/s13278-014-0206-4

ORIGINAL ARTICLE

Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack

Pete Burnap · Matthew L. Williams · Luke Sloan · Omer Rana · William Housley · Adam Edwards · Vincent Knight · Rob Procter · Alex Voss

Received: 16 February 2014 / Revised: 22 April 2014 / Accepted: 23 May 2014

Abstract Little is currently known about the factors that promote the propagation of information in online social networks following terrorist events. In this paper we took the case of the terrorist event in Woolwich, London in 2013 and built models to predict information flow *size* and *survival* using data derived from the popular social networking site Twitter. We define information flows as the propagation over time of information posted to Twitter via the action of retweeting. Following a comparison with different predictive methods, and due to the distribution exhibited by our dependent *size* measure, we used the zero-truncated negative binomial (ZTNB) regression method. To model *survival*, the Cox regression technique was used because it estimates proportional hazard rates for independent measures. Following a principal component analysis to reduce the dimensionality of the data, social, temporal and content factors of the tweet were used as predictors in both models. Given the likely emotive reaction caused by the event, we emphasize the influence of emotive content on propagation in the discussion section. From a sample of Twitter data collected following the event ($N = 427{,}330$) we report novel findings that identify that the sentiment expressed in the tweet is statistically significantly predictive of both size and survival of information flows of this nature. Furthermore, the number of offline press reports relating to the event published on the day the tweet was posted was a significant predictor of size, as was the tension expressed in a tweet in relation to survival. Furthermore, time lags between retweets and the co-occurrence of URLS and hashtags also emerged as significant.

Keywords Social network analysis · Twitter · Information flows · Information propagation · Information spreading · Social media · Sentiment analysis · Opinion mining · Predictive models

P. Burnap (✉) · O. Rana
School of Computer Science and Informatics, Cardiff University, Cardiff, UK
e-mail: p.burnap@cs.cardiff.ac.uk

O. Rana
e-mail: o.f.rana@cs.cardiff.ac.uk

M. L. Williams · L. Sloan · W. Housley · A. Edwards
School of Social Sciences, Cardiff University, Cardiff, UK
e-mail: williamsm7@cardiff.ac.uk

L. Sloan
e-mail: sloanls@cardiff.ac.uk

W. Housley
e-mail: housleyw@cardiff.ac.uk

A. Edwards
e-mail: edwardsa2@cardiff.a.uk

V. Knight
School of Mathematics, Cardiff University, Cardiff, UK
e-mail: knightva@cardiff.ac.uk

R. Procter
Department of Computer Science, Warwick University, Coventry, England
e-mail: rob.procter@warwick.ac.uk

A. Voss
School of Computer Science, St Andrews University, Fife, Scotland
e-mail: alex.voss@st-andrews.ac.uk

Published online: 13 June 2014

1 Introduction

Open and widely accessible social micro-blogging technologies, such as Twitter, are increasingly being used by citizens on a global scale to publish content in reaction to real-world events. The diffusion of this information following events can manifest itself in a number of ways, ranging from supporting social resilience through calls for assistance or help, and spreading of advice and information (Procter et al. 2013a); to the socially disruptive, by providing a platform for the distribution of misinformation, rumour (Procter et al. 2013b) and antagonist commentary (Burnap et al. 2013).

The recent terrorist events in Boston, USA and in Woolwich, London, UK sparked widespread reaction and news reporting via social media. In both cases, information pertaining to the event had both positive and negative impacts in the immediate aftermath. In relation to positive impacts, Twitter was used by law enforcement officials and journalists to request information and relay assurances to the public. On the negative side, in Boston the vast amount of information posted by the public on Twitter led to law enforcement becoming overwhelmed with multiple lines of enquiry; in the UK there were a number of arrests made following the event due to alleged religiously offensive comments being posted on Twitter. Given such significant impacts following these kind of events, it is important for those with a remit to ensure community safety to understand the predictive factors for the propagation of information flows (Lotan et al. 2011) as a first step towards being able to mitigate their impact. We define information flow propagation as the process of information spreading to a greater number of people over time via Twitter through the action of *retweeting*. That is, when a Twitter user reads a tweet, they perform an action to 'forward' that tweet to all other Twitter users who *follow* them, incrementally widening the potential readership of the original tweet. This information can contain textual content, hashtags and URLs, from which a variety of temporal, content and social metrics can be derived for modelling purposes.

This modelling is important to understand how long a piece of antagonistic text might continue to be propagated in the *Twittersphere* as it may pose a risk to social cohesion. Likewise, it could be important to understand the factors that are likely to lead to information from official sources, such as law enforcement, reaching a large number of people in a short space of time (Procter et al. 2013a). In this paper we are particularly interested in information flow propagation in the aftermath of a terrorist event. This can be considered as an aspect of the broader study of information diffusion (Guille and Hacid 2012). Understanding the diffusion of information in social networks has received significant attention from a topological and social influence perspective, but as a recent survey identified, further research is required to better understand the significance of opinion and social factors (Guille et al. 2013). Therefore, our key research question is—what are the opinion and social factors that predict large (the size) and long lasting (the survival) information flows following a terrorist event?

To model information diffusion we derived two features of an information flow: *size* and *survival*. We took the frequency of retweets surrounding the event as our *size* dependent measure, and the duration between the first and last retweet as our *survival* dependent measure. These two features are independent as they measure two distinct properties of information flow propagation. In terms of size, the number of retweets is a measure of the level of public interest and endorsement of the information, while survival is a measure of persistence of interest over time. The two features are visualized in Sect. 4 (see Figs. 1, 2) where we discuss the features of the study dataset. This is consistent with previous work on modelling information diffusion in social networks (e.g. Zaman et al. 2010, Lin et al. 2013).

This paper contributes to the existing literature by studying quantitative measures of opinion and emotion (*sentiment* and *tension*) expressed in tweets for the purposes of predicting information flow *size* and *survival* in the case of a terrorist event. Guille et al. (2013) identified the requirement for additional work examining such features in the study of information diffusion. In addition to this, we also examine other relatively unexplored factors, including the influence of daily newspaper coverage of the event; the co-occurrence of content linking factors (hashtags and URLs) within the tweet; and build on previous work investigating the significance of rapidity of early user interaction intervals in predicting "thread" size (Backstrom et al. 2013). We also employ Kaplan–Meier estimation, which produces a survivor function plot of the declining propagation of information flows over time, allowing us to visually interpret the influence of independent variables on survival.

2 Related work

Suh et al. (2011) aimed to better understand the features that are important indicators of whether an individual tweet would be retweeted. They identified three factors relating to (1) author profile—number of followers, followees, and tweets; (2) content features—URLs and hashtags; and (3) retweets and followers—separating those who have been retweeted a lot and have a large number of followers, from those who tweet a lot and favourite a lot. They built a Generalized Linear Model (GLM) that suggested URL, hashtag and age of account to be most useful for retweet

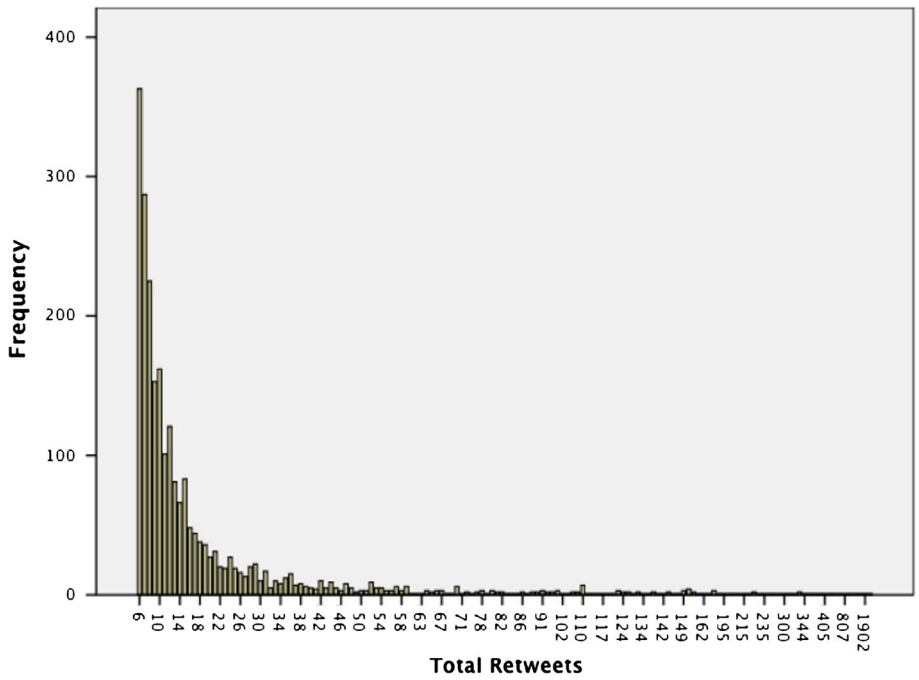

Fig. 1 Distribution of size

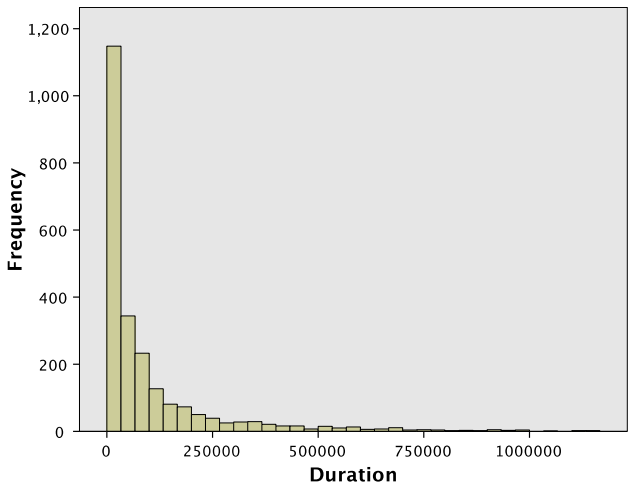

Fig. 2 Distribution of survival

prediction, with follower/followee status less significant, but still important. Similarly, Zaman et al. (2010) used the Matchbox algorithm to predict retweet probability for individual tweets, finding that attributes of the tweeter and the retweeter (similar to author profile of Suh et al.), were most accurate for prediction. However, they identified that using the content of the tweet was detrimental to prediction performance, even when normalized. Tsur and Rappoport (2012) also investigated Twitter content, specifically hashtags, in the context of the spread of ideas and memes. They identified that the emotive aspects of hashtags were not predictive of the spread of information. This is interesting as Berger and Milkman (2012) studied the likelihood of people propagating news stories based on the emotion invoked by the content and found that, contrary to common perception, people were more likely to share stories that invoke positive emotion as opposed to negative. Bandari et al. (2012) undertook the task of predicting the popularity of news stories on Twitter prior to their release. Using classification and regression techniques with features relating to content subjectivity, source and topic, they were able to achieve a reasonable accuracy in predicting a range of propagation likelihood scores, but were less efficient in predicting information flow size.

Guille and Hacid (2012) developed a model to predict the diffusion of information in online social networks, focusing on social, temporal and content factors. They identified a Bayesian logistic regression as a favoured predictive model. While the model performed well for diffusion, it was less effective at predicting size, indicating that predictive features of information diffusion and information flow size are independent. Zaman et al. (2013) also used Bayesian models for predicting the number of retweets using a time-series model, predicting at certain points in time as opposed to projecting the final size in the early stages of tweet lifetime. Neither of these models included latent subjectivity and emotion/opinion within the tweet as a feature.

Backstrom et al. (2013) identified that temporal factors including the rapidity of comments posted in response to a Facebook status update were predictive of overall thread length. Twitter interaction is slightly different to that of Facebook as retweets are propagated within and between networks as opposed to being visualized in structured "conversations" between "friends" or groups, but it could

be possible to use temporal retweet factors in a similar predictive manner, i.e. to predict total number of retweets using the rapidity of occurrence of initial retweets. Macskassy and Michelson (2011) built information propagation behaviour models for Twitter using temporal features such as the time lapsed since the original tweet was published and the recency of communication between a tweeter and other users. Both papers suggest that time is an important factor in modelling information propagation and that more investigation is required to use rapidity of retweeting to predict size and survival.

Yang and Counts (2010) constructed a topic-based diffusion model based on user mentions, where a mention constitutes the propagation of information from one user to another. They aimed to predict speed (time taken to reach first mention), scale (the number of first-order mentions of the user), and reach (the number of hops the mention produced). They employed the Cox proportional hazards regression model to quantify the degree to which features of the tweet or the original tweeting user were useful in predicting speed, scale and reach. They found that predictive features varied across topics for speed and scale with the amount a user was mentioned in the past emerging as most predictive.

Aside from the work of Lin et al. (2013) on the growth, survival and context of hashtags, our study of the existing literature at the time of writing failed to identify any work in the field related to information flow content and its relationship to survival over time, beyond hashtags alone.

3 Research questions

Based on an examination of the existing literature on information diffusion and propagation we developed five hypotheses to be tested in relation to identifying significant covariates of *size* and *survival*. As per conventional scientific methodology, *null* hypotheses (H_0) were generated from the five *alternative* hypotheses below, indicating the absence of statistically significant associations in each case. The models reported later provide evidence to reject the null hypotheses, lending support for our alternative hypotheses (Tabachnick and Fidell 2013).

H_1: Due to the nature of the study, and the focus on antagonistic content following a terrorist event, we hypothesized that tweets containing negative sentiment and high levels of tension wouldbe statistically positively associated with large and long-living information flows;

H_2: Based on previous research that identifies an association between offline media content and online reaction to real-world events (Burnap et al. 2013; Williams et al. 2013), we hypothesized that the number of newspaper stories relating to the event would be statistically positively associated with both the size and survival of information flows surrounding the event;

H_3: Based on previous research on retweet prediction, which identified linking content features of the tweet as important predictors (Suh et al. 2011), we hypothesized that the presence, frequency and co-occurrence of hashtags and URLs within the tweet would be statistically positively associated with both the size and survival of information flows surrounding the event;

H_4: Based on previous research on structured thread lengths, which identified the rapidity of responses to an original post on Facebook was a predictor of thread length (Backstrom et al. 2013; Macskassy and Michelson 2011) and tweet time lapses, we hypothesized that temporal factors relating to time lags between the time a tweet was posted and the initial n retweets would be statistically positively associated with both the size and survival of information flows surrounding the event;

H_5: Based on previous research on retweet prediction, which identified features of the tweet author profile as important predictors (Suh et al. 2011), we hypothesized that social factors such as number of followers, total number of tweets and potential reach of the tweet would be statistically positively associated with both the size and survival of information flows surrounding the event.

4 Data collection, analysis methods and predictive measures

The data collection period spanned a month following the terrorist event in Woolwich, London that took place on May 23rd 2013. Data were derived from the Twitter streaming API based on a manual inspection of the highest trending keyword following the event (i.e. "Woolwich"). An examination of Web search trends using the "Woolwich" keyword to query the Google Trends service indicated that the issue attention cycle (Downs 1972) around this event (the duration within which public attention to this event rises and falls away) spanned 14 days. This became the analysis sampling time frame for our study, during which we collected $N = 427{,}330$ tweets. The sample was subjected to data pre-processing and recoding prior to modelling, as outlined below. Table 1 provides a statistical description of the sample, including the range, mean and standard deviation of all dependent and independent variables.

4.1 Dependent measures

We took the frequency of retweets surrounding the event as our *size* dependent measure, and the duration between

Table 1 Descriptive statistics for the sample ($N = 2,336$)

Variables	Range	Mean	Std. Dev.
Dependent			
Size	6–4,079	26.30	112.64
Survival	58–1,135,479	98,699.55	161,346.70
Independent			
NumFollowers	0–8,820,174	172,432.30	666,498.50
TimelagRT1	0–146,016	782.42	6,824.65
TimelagRT2	1–494,753	1,860.08	15,417.01
TimelagRT3	1–515,929	3,238.54	21,294.03
TimelagRT4	3–835,865	5,121.39	31,198.04
TimelagRT5	5–835,881	8,028.82	38,406.12
ReachAtRT1	12–8,821,297	183,795.50	675,910.70
ReachAtRT2	140–8,821,642	187,787.40	683,277.20
ReachAtRT3	154–8,821,874	193,375.20	709,906.20
ReachAtRT4	316–1.13e+07	197,134.50	731,021.90
ReachAtRT5	448–1.13e+07	199,339.90	731,807.47
URLAndHASH	0–1	0.29	0.46
BinarySent	−1–1	−0.55	0.68
Tension	0–3	0.92	0.64
HashtagFreq	0–6	0.63	0.88
URLFreq	0–2	0.78	0.45

the first and last retweet as our *survival* dependent measure. We identified retweets via the presence of the 'RT' convention at the beginning of the tweet. This is an inherent feature of retweets. We extracted all tweets with this feature present using text pattern matching. We then derived a count measure to determine the frequency of unique tweets and their subsequent retweets by taking each original tweet *t* (i.e. not containing 'RT'), and using pattern matching on the text of all identified retweets to determine how many of them were retweets of *t*. We classified *size* as the number of retweets a unique tweet received. When aiming to predict Facebook thread length Backstrom et al. (2013) used a minimum length of five. Our data exhibited an extreme positive skew (see Figs. 1, 2) and tweets with a retweet count of 5 or lower were found in the long right tail of the distribution. Given our aim to model information propagation we use the cut-off point as identified by Backstrom et al. as a baseline convention to denote information that has propagated to a non-trivial level, and thus remove all tweets that were retweeted less than five times. Following this data processing exercise, the final sample size was $N = 2,336$. Given our sampling technique ensured the collection of all tweets containing the term 'Woolwich' for 14 days following the terrorist event, we are confident that this final sample size is representative of non-trivial information flows on Twitter surrounding this event.

4.2 Independent measures

4.2.1 Social features

Number of followers, *followees* and *total number of previous tweets* were extracted from the streaming API metadata as social features of the tweets. Given each tweet in the study dataset had already received at least five retweets, we created a variable to represent the *Reach* of the information flow after its fifth retweet by extracting and summing the number of followers of each subsequent retweeting user from the data. *Reach* was therefore the cumulative number of Twitter users that could possibly have read the tweet at that point. These factors were entered into the models outlined later in the paper as continuous predictors.

4.2.2 Temporal features

As with Reach, *Timelag* variables were derived by summing the number of seconds elapsed between the original tweet and each of the first five retweets. For example, if the first retweet occurred after 5 s, the second after 15 s, and the third after 75 s, *TimelagRT1 = 5, TimelagRT2 = 20 and TimelagRT3 = 95*.

4.2.3 Content features

Six content related features were derived from the 140 characters.

Due to our interest in the predictive influence of emotive content on information propagation following a terrorist event, we applied opinion-mining techniques to the textual content to derive independent variables. To understand the application of opinion-mining in this context it is useful to consider the definition of opinion expressed within text as a quintuple (*e, a, s, h*, and *t*), as defined in (Liu 2012), where *e* is a named entity representing the target of an opinion (e.g. a camera); *a* is an aspect of *e* (e.g. the picture quality of the camera); *s* is the sentiment or emotive feeling expressed towards *a*, where *s* can be positive, negative or neutral, and be expressed with different levels of strength [e.g. −5 (−ve) to +5 (+ve)]; *h* is the holder of the opinion; and *t* is the time at which the opinion was expressed. This framework for opinion mining is used in sentiment analysis applications that aim to measure the strength of feeling towards named entities and particular attributes. Sentiment analysis was performed on the content of tweets using the SentiStrength tool (Thelwall et al. 2010), which has been robustly evaluated for classifying emotive content as positive, negative or neutral in short, informal text such as tweets. It works by identifying sentiment-bearing words in the text and calculating a summative emotive value for the

overall text string. The output was recoded to provide negative, neutral and positive scores in the form of an ordinal variable (−1, 0, 1).

Tension analysis was also performed on the content of tweets using the COSMOS[1] tension engine. Defined by the UK Police as "any incident that would tend to show that the normal relationship between individuals or groups has seriously deteriorated", tension requires an additional focus on the interaction between communicating parties in order to measure the strength of relationships. Tension analysis follows the same quintuple formalism as opinion with the key difference that the sentiment aspect is modified to identify emotive content specific to antagonistic and accusatory content within particular contexts, including racial and religious hate related events. It has been found to provide an independent measure to sentiment in previous research (Burnap et al. 2013; Williams et al. 2013) and works in a similar way to sentiment analysis by identifying emotive words, but additionally enhances sentiment alone by considering terms that are directly related to the event, such as accusatory and attribution terms that carry an emotional weighting. We can say therefore that tension is quantifiable measure of emotion that links sentiment, directed emotion, an event, and any actors involved; which we suggest is particularly relevant to predicting the size and survival of information flows following a nationally publicized terror-related event. The outputs from this analysis were entered as ordinal values into the models (0 = low tension through 3 = high tension).

Additionally, we considered content published outside social media that related to the event. Issue attention cycle theory (Downs 1972) posits that an event is followed by a time period within which public attention to this event rises and falls away. For the purposes of understanding if peaks and troughs in the issue attention cycle were relevant predictors of the dependent variables we queried Lexis UK, a global database of publicly published news articles (e.g. national newspapers), to derive count frequencies of news stories that were published each day that included the terms "Woolwich" in the title. This variable was entered into the models as a continuous measure.

The frequency of URLs and hashtags were calculated for each tweet by applying pattern matching techniques to search for the patterns *http* and # in the tweet string. Frequencies were entered as discrete predictors in the models. We also derived a binary predictor variable based on the co-occurrence of a URL and a hashtag in a tweet.

[1] www.cosmosproject.net.

Table 2 Principal components analysis results for *size*

	Component 1	Component 2	Component 3
NumFollowers	0.984		
TimelagRT1		0.443	
TimelagRT2		0.464	
TimelagRT3		0.922	
TimelagRT4		0.898	
TimelagRT5		0.904	
ReachAtRT1	0.985		
ReachAtRT2	0.987		
ReachAtRT3	0.991		
ReachAtRT4	0.985		
ReachAtRT5	0.985		
URLAndHASH			0.714
BinarySent			0.439
Tension			−0.283
HashtagFreq			0.502
URLFreq			0.64

4.2.4 *Control variables*

Based on previous research we identified several control variables that have been shown to influence the propagation of information flows in social media (Zarrella 2009). Using the tweet metadata, time of day and day of week of when the tweets were posted were included in both models.

4.3 Principal components analysis

We conducted exploratory factor analysis using the principal component analysis (PCA) method to identify underlying features and their relative explanatory power in relation to our dependent variables of information flow *size* and *survival*. The rationale for conducting the PCA analysis was to identify correlations between sets of features ahead of their inclusion in the models reported later. This allowed us to generate sub-models based on the identified feature sets that were found to be highly correlated in the PCA, with the aim of determining which sub-model (i.e. which sets of features) explained the most variance in the dependent variables. This is useful in a practical sense in that it allows us to draw conclusions on which types of features are most important in creating large and long-lasting information flows. For example, are content features more or less important than social features?

The results reported in Tables 2 and 3 represent the optimum solution, based on an iterative analysis using a number of features. The Kaiser–Meyer–Olk in measure of sampling adequacy was 0.755 for size and 0.829 for survival, indicating small partial correlations between

Table 3 Principal components analysis results for *survival*

	Component 1	Component 2	Component 3
NumFollowers	0.985		
TimelagRT1		0.595	
TimelagRT2		0.844	
TimelagRT3		0.896	
TimelagRT4		0.879	
TimelagRT5		0.81	
ReachAtRT1	0.989		
ReachAtRT2	0.993		
ReachAtRT3	0.993		
ReachAtRT4	0.993		
ReachAtRT5	0.993		
URLAndHASH			0.899
BinarySent			0.287
Tension			−0.28
HashtagFreg			0.761
URLFreg			0.5

components. Bartlett's test for Sphericity was used to test the null hypothesis that the variables in the population correlation matrix were uncorrelated, and the result was significant for both dependents ($p < 0.000$), indicating the correlation matrices of features was factorizable. Based on an inspection of the screen plot, component and pattern matrices, and parallel analysis three components were extracted across both analyses. These three components are highly correlated with each other, as indicated by the presented rotated factor loading in Tables 2 and 3. The principle components explain a total of 71 and 69 % of the variance in the dependent measures—size and survival, respectively. The first component, which relates to social factors—namely number of followers and potential reach, explained 47 and 37 %, respectively; the second component, which relates to time lags between the first five retweets, explained 19 and 21 %; the third, relating to content, including URL and hashtag presence, number and co-occurrence, sentiment and tension explained 14 and 11 %. Based on the resulting amount of variance explained, the components were included in the models as experimental predictors for both dependent measures.

5 Methods

5.1 Size model selection

Having established the principal components we considered the application of various models for inferring the size of information propagation following the event under study. We identified models based on attributes exhibited by our size dependent measure (i.e. a Poisson skewed distribution) and previous research that had identified successful applications. First we selected a statistical regression technique. The zero inflated negative binomial (ZTNB) technique was selected for the size dependent measure based on the theoretical foundation that it was developed to handle continuous count data that is right skewed (see Fig. 1), over-dispersed (the conditional mean and variance were not equal), and can account for the absence of zero counts in the dependent measure. The mean of the *size* variable was 26.30 and the variance was 12,688.79, thus the data is certainly over-dispersed, ruling out the use of the more conventional Poisson technique.

To ensure we adopted the most appropriate and accurate modelling method, we compared the regression model with two machine-learning techniques that have been evidenced to produce favourable results in modelling information diffusion (Guille and Hacid 2012). The VGAM package was used to perform ZTNB model building using the R statistical software toolkit, and the Weka toolkit was used to build the machine-learning models. In order to fit our size dependent measure to these categorical techniques we employed data transformation to derive a binary classification: medium (M) (where retweet count $\leq$49) and large (L) (where retweet count $\geq$50) information propagation. By transforming the data in this fashion we acknowledge that a degree of complexity is lost in our dependent measure. To compare models we applied the standard classification measures of: *precision* (the fraction of retrieved documents that are relevant to the search—i.e. for each class how many of the retrieved texts were of that class); *recall* (fraction of documents that are relevant to the search that are successfully retrieved—i.e. for each class how many tweets coded as that class were retrieved); and *F-Measure*, a harmonized mean of precision and recall. The optimum result for each score is 1.

Table 4 summarizes the prediction accuracy for the three models. It shows BLR obtained the lowest F-Measure across the classes. This is likely due to BLR decisions being based on probability distributions, and the majority of tweets in the sample (and therefore the training dataset) being of the *medium* class. Because the training data is overwhelmingly biased to the *medium* class, and instances of the *large* class were rare, the probabilistic classifier appears to have labelled all new instances as *medium*. The decision tree obtained the highest F-Measure for the large class, however, its model is more specific because of the partitioning algorithm on which it is based (Guille and Hacid 2012). ZTNB performed best in relation to the medium class but obtained less favourable results in relation to the large class. Based on these results and other factors, including (1) the ZTNB is designed to fit the

Table 4 Classifier performances on a tenfold cross-validation

Classifier parameters	Precision		Recall		F-Measure	
Size dependent	*M*	*L*	*M*	*L*	*M*	*L*
Zero-truncated negative binomial (ZTNB)	1	1	0.99	0.10	0.99	0.18
Decision tree (C4.5)	0.95	0.69	0.98	0.44	0.97	0.54
Bayesian logistic regression (BLR)	0.92	0.00	0.98	0.00	0.95	0.00

original form of our size dependent measure relating to our hypotheses (continuous count data with a Poisson distribution) and (2) unlike the C4.5 decision tree model, ZTNB is multivariate, meaning it deals with attribute correlation, allowing the understanding of the influence of each attribute to the whole process while controlling for other attributes by holding them constant, we concluded the ZTNB model was most appropriate for estimating the size of information propagation in relation to the event under study.

5.2 Size model: zero-truncated negative binomial regression

The first dependent variable (size) exhibited a skewed distributed when plotted, therefore required a non-parametric model. We generated an estimate of the dispersion co-efficient to determine if the data were over-dispersed. An examination of the 95 % confidence interval indicated an estimate greater than zero, which suggests over-dispersion (conditional mean not equal to the conditional variance), ruling out the use of a Poisson model. Furthermore, because we are only interested in tweets that were retweeted more than five times, the data did not contain any zero counts for size. Based on these requirements we selected a ZTNB model to conduct the regression analysis. This regression technique is used to model count data for which the value zero cannot occur and when there is evidence of over-dispersion. The cumulative probability distribution function $F(X = x)$ for the discrete random variable x denoting the count, is formulated as follows, where i represents the number of retweets for each instance and $0 < p \leq 1$:

$$F(X = x) = \sum_{i=0}^{x} \binom{s+i-1}{i} p^s (1-p)^i$$

5.3 Survival model—Cox proportional hazards regression

The second dependent (survival) was a measure of the lifetime of an information flow. Our interest was to model the factors that affect the survival of information flows following the terrorist event. For example, we were interested in determining whether a factors such as the polarity of sentiment, had an effect on the hazard rate for information flow survival. Put another way, does expressing negative sentiment increase the lifetime of an information flow? This question can be posed as one of hazards to survival, thus we adopted Cox's proportional hazards model (Cox 1972) to estimate the explanatory factors. These models produce a survival function that predicts the probability that an event has occurred (in this case the last retweet present in the timeframe) at a given time t for the given values of the predictor variables X denoted by $\lambda(t|X)$. This can be formulated as follows:

$$\lambda(t|X) = \lambda_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n\right)$$

Based on this, the partial likelihood for X can be calculated using:

$$L(\beta) = \sum_{i:C_i=1} \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_i}$$

Where for a given tweet i, C_i is an indicator of the time corresponding to the tweet and $\theta_i = \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n\right)$.

As the study was bounded by a 14 days data analysis window, we were mindful that some information flows may survive the study period. Therefore, based on a study over 1 billion tweets that identified only 0.03 % of retweets occur beyond 48 h of the original tweet,[2] tweets that were posted within 48 h of the curtailment of data analysis were *right-censored* (i.e. the last retweet may not have occurred and we assume the information flow to still be active or "alive").

5.4 Kaplan–Meier estimation

We used Kaplan–Meier (KM) estimation to plot the survival functions of covariates that were most significant in the Cox proportional hazards model. These plots provide useful visualizations that demonstrate the relative affects on survivability based on the condition of the factors, e.g. the impact on survival rates when sentiment is positive or negative. A plot of the KM estimate of the survival function is a series of horizontal steps of declining magnitude. An important advantage of the KM curve is that the method can take into account right-censored data as are present in our analysis (as indicated by small vertical tick marks on the plot). The estimator is formulated as follows:

$$\widehat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

When there is no censoring, n_i is the number of survivors just prior to time t_i. With censoring (as in our case), n_i

[2] http://www.sysomos.com/insidetwitter/engagement/.

is the number of survivors minus the number of losses (censored cases). At time t_i only those surviving cases and are considered to be "at risk" of "dying" (receiving no more retweets).

6 Results

6.1 Information flow size

Results of the zero-truncated negative binomial model (Table 5) indicated several statistically significant associations between the dependent variable (size) and our predictive factors, and variability in the strengths of association as indicated by the incidence rate ratio (IRR). The IRR is derived by the exponentiation of the negative binomial regression coefficients, allowing for the interpretation of retweet incidence rates (as opposed to logs of expected retweet counts). We can therefore use the IRR to report the strength of causal associations between certain factors and the information flow size, enabling us to identify quantitatively which factors are more important than others.

Table 5 Size model results

Predictors	IRR	Std. Err.	Sig.	*Z* value
Social factors				
NumFollowers	1.0000010	0.0000	0.0000	4.7200
ReachAtRT5	1.0000000	0.0000	0.4640	−0.0900
Following	1.0000000	0.0000	0.4100	0.2300
TweetCount	0.9999979	0.0000	0.0000	−3.7300
Temporal factors				
TimelagRT1	0.9999988	0.0000	0.3885	−0.2800
TimelagRT2	1.0000000	0.0000	0.4670	0.0800
TimelagRT3	0.9999986	0.0000	0.2135	−0.8000
TimelagRT4	1.0000030	0.0000	0.0135	2.2100
TimelagRT5	0.9999948	0.0000	0.0000	−5.2900
CommuteMorn	1.1261560	0.1041	0.0995	1.2800
Work	1.2053270	0.0867	0.0045	2.6000
CommuteEve	1.2342040	0.1041	0.0065	2.4900
Evening	1.3008160	0.1000	0.0005	3.4200
Sunday	1.5336850	0.1790	0.0000	3.6700
Monday	1.1657050	0.1470	0.1120	1.2200
Tuesday	1.2523810	0.1612	0.0400	1.7500
Thursday	1.1473530	0.1361	0.1235	1.1600
Friday	0.9955436	0.1304	0.4865	−0.0300
Saturday	1.5184780	0.1763	0.0000	3.6000
Content factors				
URLFreq	0.6497726	0.0399	0.0000	−7.0300
HashtagFreq	0.8454206	0.0296	0.0000	−4.8000
URLAndHASH	1.7844330	0.1319	0.0000	7.8400
NewsStories	1.0004700	0.0003	0.0365	1.7900
BinarySent	1.1090050	0.0383	0.0015	2.9900
Tension	1.0360830	0.0448	0.2065	0.8200

6.2 Social factors

Holding all other factors constant, number of followers and number of previous tweets were statistically significant. In relation to the former, an increase in number of followers was positively associated with size of information flow (IRR = 1.00, $z = 4.72$, $p < 0.00$). Conversely, an increase in the number of previous tweets made by the tweeter was negatively associated with information flow size (IRR = 0.99, $z = -3.73$, $p < 0.00$).

6.3 Temporal factors

Of the variables of interest, the by-the-second time lag association with size varied between positive and negative. Only at time lag 5 did the associations begin to emerge as significant. By retweet 5 it was evident that the time lag was highly significant (IRR = 0.99, $z = -5.29$, $p < 0.00$) and negatively associated with size. Of the control variables, day of week and time of day emerged as statistically significant, confirming previous work (Zarrella 2009).

6.4 Content factors

All but tension emerged as statistically significant. Most significant were the URL and hashtag measures, which confirm the results of related work (Suh et al. 2011). We conducted several analyses using various measures of URL and hashtag presence, including binary (absent or present). We investigated both presence and frequency of URLs and hashtags and it was evident that the measure of frequency as opposed to presence explained more of the variance in the dependent and resulted in a better overall model fit. The presence of a URL was negative associated with size ($z = -4.92$, $p < 0.01$), whereas presence of a hashtag was positively associated ($z = 3.25$, $p < 0.01$). Both URL and hashtag frequency were negatively associated with size (URL frequency IRR = 0.65, $z = -7.03$, $p < 0.00$; hashtag frequency IRR = 0.85, $z = -4.90$, $p < 0.00$). Of novel importance here is the strong positive statistical association between the co-occurrence of URLs and hashtags in a tweet, and the information flow size (IRR = 1.78, $z = 7.84$, $p < 0.00$). Out of all the predictors in the model, this variable explains the most variance in the dependent.

Of particular interest to this study, the emotional aspect of sentiment also emerged as statistically significant, with tweets containing positive sentiment more likely to become large information flows (IRR = 1.11, $z = 2.99$, $p < 0.01$).

Table 6 Size model sub-factor analysis

	Full model	Social factors	Temporal factors	Content factors
N	2,334	2,334	2,334	2,334
−2 LL	−9,434.86	−9,513.774	−9,769.379	−9,732.987
BIC	19,079.14	19,151.65	19,670.61	19,605.59

Table 7 Survival model results

	Estimate β (B)	Std. Error	Wald	Sig.
Social factors				
NumFollowers	3.30E−08	1.24E−07	0.073	0.3935
ReachAtRT5	−1.60E−07	1.14E−07	1.979	0.08
Following	−0.000002	0.000002	0.619	0.2155
TweetCount	0.000003	5.29E−07	33.888	0.000
Temporal factors				
TimelagRT1	−0.000007	0.000004	3.58	0.029
TimelagRT2	6.31E−07	0.000002	0.064	0.400
TimelagRT3	−9.47E−07	0.000002	0.193	0.3305
TimelagRT4	0.000003	0.000002	2.083	0.0745
TimelagRT5	−0.000005	0.000001	15.181	0.000
CommuteMorn	0.27	0.09	8.879	0.0015
Work	0.116	0.071	2.715	0.0495
CommuteEve	0.253	0.083	9.398	0.001
Evening	0.155	0.075	4.292	0.019
Sunday	−0.008	0.126	0.004	0.4745
Monday	0.05	0.136	0.138	0.3555
Tuesday	−0.131	0.138	0.89	0.1725
Thursday	0.105	0.129	0.665	0.2075
Friday	0.024	0.142	0.029	0.432
Saturday	−0.18	0.127	1.999	0.0785
Content factors				
URLFreq	−0.222	0.062	12.756	0.000
HashtagFreq	−0.018	0.035	0.246	0.31
NewsStories	0	0	1.228	0.134
URLAndHASH	−0.152	0.072	4.455	0.0175
BinarySent	−0.082	0.033	6.113	0.0065
Tension	0.095	0.041	5.437	0.01

The issue attention cycle also plays a role as the number of offline press reports relating to the event published on the day the tweet was made was positively associated with size of information flow (IRR = 1.00, $z = 1.79$, $p < 0.05$).

The diagnostics for the full model indicated a good fit to the data (−2 Log-likelihood = −9,434.86, BIC = 19,079.14). Based on the components derived from the PCA we specified sub-models to determine which set of factors explained most of the variance in the dependent measure of size (Table 6). Social factors explained the largest amount of variance (−2 Log-likelihood = −9,513.774, BIC = 19,151.65), followed by content factors (−2 Log-likelihood = −9,732.987, BIC = 19,605.59) and temporal factors (2 Log-likelihood = −9,769.379, BIC 19,670.61).

6.5 Information flow survival

Results of the Cox proportional hazards (Table 7) model also indicated several statistically significant associations between the dependent variable (survival) and our predictive factors. Because the model is focused on explaining proportional hazards, a positive estimate is interpreted as increasing hazards to survival, and therefore reduces the survival of the information flow. We can therefore use the estimate β to report the strength of causal associations between certain factors and the information flow survival, enabling us to identify quantitatively which factors are more important than others.

6.6 Social factors

Holding all other factors constant, only the number of tweets previously posted by the author of a new tweet were statistically significant in predicting hazards to survival. Tweet count was found to be positively associated with hazards to survival (β=0.00, Wald = 33.89, $p < 0.01$), indicating more previous tweets reduces the survival of the information flow.

6.7 Temporal factors

Of the variables of interest, the by-the-second time lag association with survival follows an inverse pattern to that exhibited in the size model. Though the earlier stage time lags are far less significant (and in most cases not significant) than time lag 5, which emerges as highly significant and negatively associated with hazards ($\beta = -0.00$, Wald = 15.181, $p < 0.01$). Thus, the results suggest information flows following this event will survive longer where the number of seconds between the original tweet and the 5th retweet is higher, which is intuitive in a time-series model. Of the control variables, time of day emerged as statistically significant for predicting survival; the model suggests there is a higher likelihood of hazards to survival at times of the day when tweet traffic is known to be highest (Zarrella 2009).

6.8 Content factors

As with the size model, we conducted analysis with presence and frequency of URLs and hashtags, the latter being reported here because it explained more variance in the

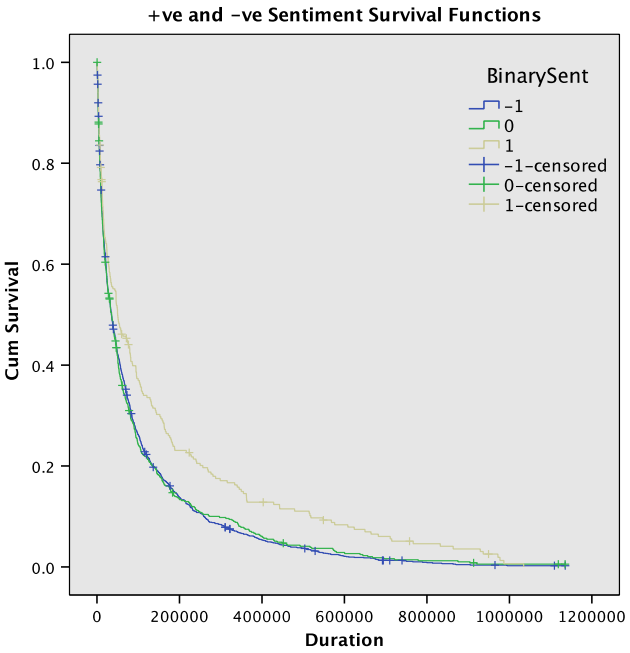

Fig. 3 Survival rates for sentiment

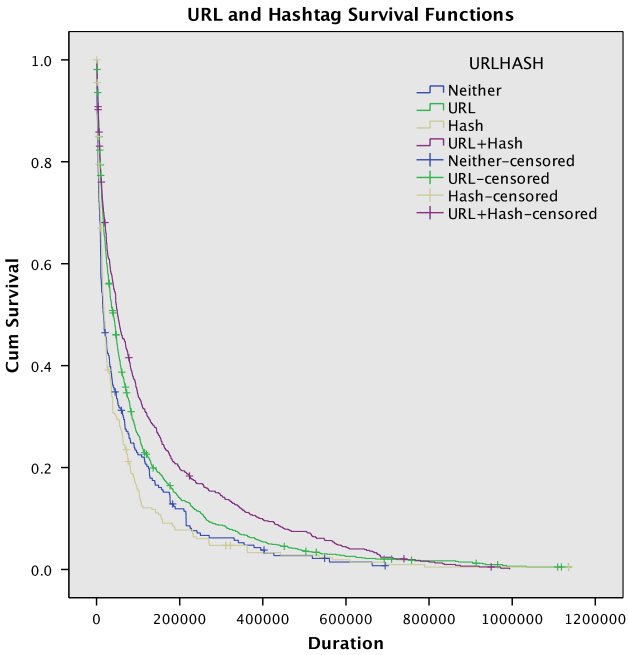

Fig. 5 Survival rates for URL + Hashtag

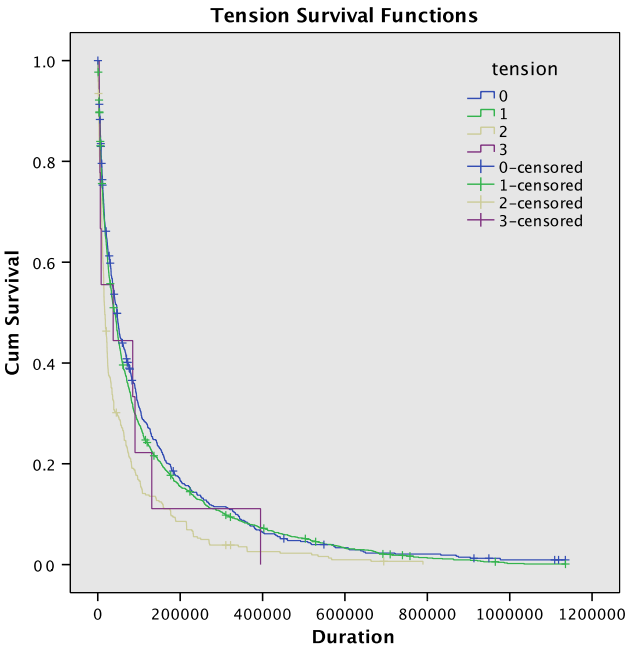

Fig. 4 Survival rates for tension

dependent. Both measures were negatively associated with hazards. URL frequency emerged as negatively associated with hazards and highly significant ($\beta = -0.22$, Wald $= 12.76$, $p < 0.01$), suggesting more URLs included in a tweet leads to longer survival. The co-occurrence of a URL and a hashtag in a tweet was also negatively associated with hazards and statistically significant ($\beta = -0.15$, Wald $= 4.46$, $p < 0.05$).

Both emotive content features, sentiment and tension, were highly statistically significant. Positive sentiment within a tweet was negatively associated with hazards ($\beta = -0.08$, Wald $= 6.11$, $p < 0.01$) and high levels of tension were positively associated ($\beta = -0.10$, Wald $= 5.44$, $p < 0.01$). This suggests tweets containing positive sentiment and a lack of tension in relation to the event produce longer living information flows. Unlike size, hashtag frequency and offline press reports were not statistically significant. Using Kaplan–Meier estimation, Figs. 3, 4, and 5 illustrate the comparative survival rates of tweets for all significant content measures. Figure 3 shows that tweets containing negative and neutral sentiment have low survival rates compared to those containing positive sentiment. Similarly Fig. 4 shows that tweets containing high levels of racist tension also die out more quickly than tweets with low levels of tension. Finally, Fig. 5 shows that the co-occurrence of URLs with hashtags (i.e. features that link the text to other text) also extends the life of tweets following such events.

The diagnostics for the full model indicated a good fit to the data (LRT $= 225.4$, $R^2 = 0.92$). We repeated the sub-model analysis (Table 8) and determined that content factors explained the largest amount of variance (LRT $= 122.3$, $R^2 = 0.051$), followed by temporal factors (LRT $= 103.9$, $R^2 = 0.044$) and social factors (LRT $= 86.36$, $R^2 = 0.036$).

Table 8 Survival model sub-factor analysis

	Full model	Social factors	Temporal factors	Content factors
N	2,334	2,334	2,334	2,334
LRT	225.4 ($p < 0.01$)	86.36 ($p < 0.01$)	103.9 ($p < 0.01$)	122.3 ($p < 0.01$)
Wald	198 ($p < 0.01$)	80.02 ($p < 0.01$)	81.43 ($p < 0.01$)	123.2 ($p > 0.01$)
R^2	0.092	0.036	0.044	0.051

7 Discussion

The statistical analyses revealed several significant associations between the independent variables and the dependent variables of size and survival of information flows surrounding a terrorist event. In relation to the first hypothesis that stipulated tweets containing negative sentiment and high levels of tension would produce large and long-living information flows, the results indicated negative sentiment was predictive of smaller size ($p < 0.01$) and shorter survival ($p < 0.01$), providing support to the null hypothesis. Based on an inspection of the incidence rate ratio (IRR) positive retweets were expected to have a rate of 1.11 (11 %) higher compared to negative tweets, while holding all other factors constant. Similarly, positive tweets were also more likely to propagate for longer during this event. In relation to high tension, the evidence was mixed. Results suggest there was no significant association between tension and size of information flow, however significant associations did emerge in relation to survival ($p < 0.01$). Contrary to the hypothesis, tweets containing high tension had shorter survival rates than those with lower level of tension. Mirroring the sentiment finding, this suggests Twitter users were less likely to prolong the information flow where the content of a tweet was antagonistic. In order to make sense of these associations, we are in the process of conducting further qualitative examination of the content of positive, negative and tense tweets.

Previous research identified an association between offline media content and online reaction to real-world events (Williams et al. 2013). We hypothesized that the number of newspaper stories relating to the event would influence both the size and survival of information flows surrounding the event. The models revealed mixed support for this, as there was no significant association with survival. However, for size the number of related newspaper stories published on the day a tweet was posted had a significant positive association ($p < 0.05$). Holding all other factors constant, an increase in 100 news headlines about the event increased the rate of retweets by a factor of 1.05 (5 %). This provides evidence to suggest that people react to stories published in offline media through the mechanism of online social media in relation to this kind of event.

Based on previous research on retweet prediction, which identified linking content features of the tweet as important predictors (Suh et al. 2011), we hypothesized that the presence, frequency and co-occurrence of hashtags and URLs within the tweet would influence both the size and survival of information flows surrounding the event. Hashtag and URL frequency were negatively associated with size, with an increase of hashtag and URL instances decreasing the rate of retweets by 15 and 35 %, respectively. This could be because including a number of hashtags and URLs reduces the space available in the 140-character post for the inclusion of information that may be of interest and subsequently retweeted. However, of novel importance here is the strong positive statistical association between the co-occurrence of URLs and hashtags in a tweet, and the information flow size and survival. The presence of a URL and a hashtag increased the rate of retweets by a factor of 1.78 (78 %). This may suggest that tweets that are identified as relating to the event via the use of a hashtag, and provide additional sources of information via a URL, could become large and long-lasting information flows due to their discoverability and enriched content.

Based on previous research on structured thread lengths, which identified the rapidity of responses to an original post on Facebook was a predictor of thread length (Backstrom et al. 2013), we hypothesized that temporal factors would influence both the size and survival of information flows surrounding the event. The results suggested that by the fifth time a tweet is retweeted, the time lag is highly significant ($z = -4.30$, $p < 0.01$) and negatively associated with size of the potential information flow. For every additional second time lapse between the original tweet and the fifth retweet, the retweet rate reduces by 1 %. This suggests the quicker a tweet is retweeted the larger it will become. Furthermore, time elapsed between the original tweet the fifth retweet was highly significant (Wald = 15.18, $p < 0.01$) and positively associated with survival rate, indicating the slower a tweet is retweeted, the longer it will survive. Or put another way, tweets that receive their first five retweets faster also die out quicker following a burst of sudden interest. The association between rapid propagation of information and size could be

explained through the theory of the issue attention cycle (Downs 1972), where the prominence of interest (i.e. large number of retweets) is a factor of enthusiasm in a current issue. This rapid burst of interest in a particular tweet could explain the size. Similarly, using the same theory, following a burst of enthusiasm, comes a rapid decline in interest, and a rapid decay in information propagation. Therefore, tweets that receive a large amount of interest early in their lifetime tend to survive for less time as interest wanes.

Also based on previous research on retweet prediction, which identified features of the tweet author profile as important predictors (Zaman et al. 2010), we hypothesized that social factors such as number of followers, potential reach of the tweet, and number of previous tweets will influence both the size and survival of information flows surrounding the event. We found that, for size, the number of followers of the tweeter was positively associated and highly significant ($p < 0.01$). An increment of 100,000 followers increased the rate of retweets by a factor of 10 %. Interestingly, the reach after the fifth retweet was not significantly correlated with size, while number of followers was, suggesting that the number of followers of the original tweeter is more significant as a predictor than the number of people it might have reached upon propagation. Neither were significant for survival. In relation to number of previous tweets, there was a significant ($p < 0.01$) negative association with both size and survival. This is contrary to the findings of Suh et al. (2011), who found no significant association with retweet likelihood. This new finding possibly suggests that by increasing the total number of tweets the user increases the freshness of their tweet-stream, decreasing the relevance of older information and possibly curtailing its propagation and survival in favour of the new content. Another possible explanation is that during these kinds of events, new information becomes available very frequently and given the user's propensity to tweet a lot, it is likely they could be relaying new information as it emerges.

8 Conclusion

In this paper we built models that predicted information flow size and survival on Twitter following a terrorist event. We defined information flows as the propagation over time of information posted to Twitter via the action of retweeting. We used zero-truncated negative binomial and Cox proportional hazards regression techniques to investigate the predictive value of social, temporal and content factors of the tweet. The novel findings were that time lags between retweets, the co-occurrence of URLS and hashtags, and the sentiment expressed in the tweet, were statistically significantly predictive of both size and survival of information flows of this nature. Furthermore, the number of offline press reports relating to the event published on the day the tweet was made were significant predictors of size, and the tension expressed in a tweet was predictive of survival.

The sub-model analysis for size suggested that social factors explained the largest amount of variance, followed by content factors and temporal factors. Interestingly, despite all but one of the content features being highly significant, it appeared that the number of followers and previous tweets alone accounted for more of the variance in the dependent of size for an event of this nature. Thus, to create a large information flow following a terrorist event, the most important factor is the social features of the tweeter. In contrast, for survival, content factors explained the largest amount of variance, followed by temporal factors and social factors. This suggests that to create a long-lasting information flow following such an event, the tweeter should focus on content features, such as including hashtags, URLs, positive sentiment and low tension.

This study has provided evidence that the opinion/emotional factors of tweets are statistically significant predictors of both information flow size and survival following a terrorist event. Contrary to popular perception, while negative and tense content did emerge on Twitter following the event, it failed to propagate via retweeting. In short, negative and tense content remained small in number and lasted for short periods, dying out quickly. This indicates that on the whole, members of the public who used Twitter in this case chose to propagate positive and supporting content. Following on from this work we intend to test the predictive efficacy of our models on other cases that exhibit similar characteristics (e.g. the Boston terrorist event and the 2011 riots in England). The outcome of this ongoing research will be a general predictive model for Twitter information flow size and survival following major socially disruptive events.

References

Backstrom L, Kleinberg J, Lee L, Danescu-Niculescu-Mizil C (2013) Characterizing and curating conversation threads: expansion, focus, volume, re-entry. Paper presented at the proceedings of the sixth ACM international conference on web search and data mining, Rome, Italy

Bandari R, Asur S, Huberman BA (2012) The pulse of news in social media: forecasting popularity. CoRR. http://arxiv.org/abs/1202.0332

Berger J, Milkman K (2012) What makes online content viral? J Mark Res 49(2):192–205

Burnap P, Rana O, Avis N, Williams M, Housley W, Edwards A, Morgan J, S L (2013) Detecting tension in online communities with computational Twitter analysis. Technol Forecast Social Change. doi:10.1016/j.techfore.2013.04.013

Cox D (1972) Regression models and life tables. J Roy Statist Soc B 34:187–220

Downs A (1972) Up and down with ecology—the 'issue-attention cycle'. Public Interest 28:28–50

Guille A, Hacid H (2012) A predictive model for the temporal dynamics of information diffusion in online social networks. Paper presented at the 21st international conference companion on World Wide Web, Lyon, France

Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: a survey. SIGMOD Rec 42(1):17–28. doi:10.1145/2503792.2503797

Lin Y, Margolin D, Keegan B, Baronchelli A, Lazer D (2013) #Bigbirds never die: understanding social dynamics of emergent hashtags. In: Proceedings of the seventh international AAAI conference on weblogs and social media, Boston, MA

Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5(Special Issue):1375–1405

Macskassy S, Michelson M (2011) Why do people retweet? antihomophily wins the day. In: International conference on weblogs and social media (ICWSM)

Procter R, Crump J, Karstedt S, Voss A, Cantijoch M (2013a) Reading the riots: what were the police doing on Twitter? Polic Soc 23(4):1–24. doi:10.1080/10439463.2013.780223

Procter R, Vis F, Voss A (2013b) Reading the riots on Twitter: methodological innovation for the analysis of big data. Int J Soc Res Methodol 16(3):197–214. doi:10.1080/13645579.2013.774172

Suh B, Hong L, Pirolli P, Chi E (2011) Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: SocialCom

Tabachnick B, Fidell L (2013) Using multivariate statistics, 6th edn. Allyn and Bacon, Boston

Thelwall M, Buckley K, Paltogou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inform Sci Technol 61(12):25442558

Tsur O, Rappoport A (2012) What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. Paper presented at the proceedings of the fifth ACM international conference on web search and data mining, Seattle, Washington, USA

Williams ML, Edwards A, Housley W, Burnap P, Rana O, Avis N, Morgan J, Sloan L (2013) Policing cyber-neighbourhoods: tension monitoring and social media networks. Polic Soc 23(4):1–21. doi:10.1080/10439463.2013.780225

Yang J, Counts S (2010) Predicting the speed, scale, and rangeof information diffusion in Twitter. In: International conference on weblogs and social media (ICWSM)

Zaman T, Herbrich R, Van Gael J, Stern D (2010) Predicting information spreading in Twitter. In: Workshop on computational social science and the wisdom of crowds (NIPS). http://arxiv.org/abs/1304.6777

Zaman T, Fox E, Bradlow E (2013) A Bayesian approach for predicting the popularity of tweets. CoRR

Zarrella D (2009) The science of retweets