Spreading the News: How Can Journalists Gain More Engagement for their Tweets?

Claudia Orellana-Rodriguez, Derek Greene, Mark T. Keane
Insight Centre for Data Analytics
University College Dublin, Ireland
{claudia.orellana, derek.greene, mark.keane}@insight-centre.org

ABSTRACT

News media face many serious concerns as their distribution channels are gradually being taken over by third parties (e.g., people sharing news on Twitter and Facebook, and GoogleNews acting as a news aggregator). If traditional media is to survive at all, it needs to develop innovative strategies around these channels, to maximize audience engagement with the news they provide. In this paper, we focus on the issue of developing one such strategy for spreading news on Twitter. Using a corpus of 1M tweets from 200 journalist Twitter accounts and audience responses to these tweets, we develop predictive models to identify the features of both journalists and news tweets that impact audience attention. These analyses reveal that different combinations of features influence audience engagement differentially from one news category to the next (e.g., sport versus business). From these findings, we propose a set of guidelines for journalists, designed to maximize engagement with the news they tweet. Finally, we discuss how such analyses can inform innovative dissemination strategies in digital media.

CCS Concepts

•Human-centered computing → Web-based interaction;

Keywords

Computational journalism; social media; audience engagement; news events

1. INTRODUCTION

News media are concerned by the emergence of new distribution channels and by third parties (e.g. citizen journalists and news aggregators) gradually taking over them. Professional journalists need to understand the audience of these channels and develop novel strategies to maximize the audience attention to the news they provide. In this paper, we look at developing one such strategy for one of the major social media distribution channels, that of Twitter.

In recent years, Twitter has emerged as *the* social media platform for news. It is the preferred tool for both consumers actively searching for news and for journalists trying to reach as wide an audience as possible; journalists typically tweet links to their online articles or retweet news items form their own company (what we will call *news tweets*). Twitter has also become a platform that *is the news*; as politicians tweet their views, as celebrities tweet breakups and citizen journalists report events they have witnessed (e.g., [10, 22]). Although Facebook may now account for more referrals to news websites, Twitter still retains a special status, as it seems to reach an influential (albeit smaller) audience for news *per se* (See 2015 Reuters Institute Digital News Report [25]).

However, ultimately, Twitter is a distribution channel for news that is not controlled by the news media. Journalists tweet links to their news articles, but it is the distributed response of the Twitter community that determines whether that news spreads [18]. Therefore, a key problem for journalists and news organizations is to determine the best strategy for maximizing audience engagement with their news in this third party distribution channel or, to put it more simply *what is the best way to spread for one's news?*

Unfortunately, at present, no clear answers to this question have been forthcoming. It is still unclear, from both research studies and journalistic practice, how to optimize audience engagement for news tweets. Many news agencies are struggling to determine whether one style of reporting news on Twitter is more successful than others, and to identify the variables that most influence audience engagement. Indeed, they are still to determine the best metrics to quantitatively assess the impact of their Twitter strategies.

In this paper, we attempt to find solutions to some of these problems. In it, we identify the features of both journalists and their tweets that help predict audience engagement with news tweets. Previous research on Twitter has shown that many tweets tend to be about news [11], that news can first break on Twitter [17], and identified some of the factors that influence the dissemination of a tweet [20]. However, this prior work has seldom specifically focused on journalistic tweeters, or indeed on news tweets in the assessment of audience engagement (see Section 2).

To fill the gap in the literature, we did two things. First, we performed a series of analyses of a corpus of tweets, from 200 journalist and news outlets accounts, across six news categories, namely lifestyle, science and technology, politics, sports, breaking news, and business, generating regression models for engagement prediction. These models give us insights into which features of journalists and their tweets are critical to garnering attention on Twitter. Second, from these analyses, we propose a set of guidelines for journalists in tweeting their news designed to increase audience engagement, when applied.

In summary, this work has two main contributions:

- We surface the main features that impact audience engagement for journalistic news tweets and reveal the ways in which they interact across different news categories.

WebSci '16, May 22-25, 2016, Hannover, Germany

DOI: **http://dx.doi.org/10.1145/2908131.2908154**

- We formulate a set of concrete guidelines for news producers to inform their strategy for spreading news on Twitter, whether that news provider is an individual journalist or a corporate body.

In the next section, we briefly review the related work in this area before presenting our analysis of journalistic tweeting (Section 3), developing predictive models (Section 4) and guidelines for journalistic practice (Section 5).

2. RELATED WORK

With millions of users and non-stop messages, it is increasingly harder for journalists to reach key audiences on Twitter, enabling their news to spread further; especially, when one considers that the majority of users are passive information consumers rather than actively responding (e.g., by tweeting or favoriting). Having said this, Twitter is still *the* social network for news dissemination and, as such, there is a considerable body of relevant research that addresses the problem of audience engagement. However, often this research has not specifically separated journalistic tweeters from other tweeters, or indeed news tweets from other tweets in the assessment of audience engagement.

2.1 A Good & Bad Journalistic Tool

On the positive side, Twitter has attained a special status as a tool for journalists given its capabilities to post and read real-time updates of events. In countries such as Ireland, the UK and France, more than 90% of the journalists report to use Twitter for work [9]. As such, millions of people consider this social media site as a source of news and actively seek relevant content to be informed of the latest developments. Twitter alone generates more than 10% of the visits to online newspapers, and its users differ from those of other social media sites because they are actively seeking news within the platform, instead of just coming across it, as is the case for Facebook users [25]. Twitter has also emerged as a key source of breaking news [10] and has become a key conversational channel for many journalists to build a following around their news articles, and a public channel for published news via the corporate twitter accounts of the major news organizations [25].

On the negative side, Twitter has also become a communication space in which journalists face challenges that could compromise their professional norms and practices. Content analyses have shown that journalists express themselves more freely on Twitter, in a more social media style, than they do in news articles, possibly conflicting with journalistic norms of objectivity [12]. Furthermore, with the rise of Twitter as a news channel there is some concern about the blurring of the gap between what the public shares in social media and what the news media publishes online [16]. Finally, while the citizen journalism aspect of Twitter is very important, it has raised new issues about validating Twitter information sources and establishing their veracity, to separate real from fake eyewitnesses [5].

2.2 Information Spread & Engagement

There is a substantial body of research on how information spreads in social networks like Twitter and on the ways that audiences engage with content [11, 27].

Many of the seminal papers on Twitter address the question of what tweets tend to be retweeted. Typically, they analyze crawls of all tweets (not just news tweets) for a selected calendar period, finding evidence for the impact of *user factors* (e.g., number of followers of a user, age of user's account, number and frequency of tweeting by a user) and *content factors* (e.g., presence of URLs, hashtags and mentions). For example, Suh et al. [24] analyzed a corpus of 10,000 tweets, using Principal Components Analysis (PCA), and found that the presence of URLs, use of hashtags, numbers of followers/followees and the age of the account were predictive of retweetability; a result they verified against a larger crawl of 74M tweets. Interestingly, they also showed that the particular URL used mattered; for example, if the URL was www.youtube.com or www.bbc.com then more retweets were likely.

These initial studies led to deeper explorations of the user factor of *popularity* in information spreading (i.e., essentially, numbers of followers). However, this variable turns out not to be a simple and constant influence: an analysis of 2.5 million Twitter users has shown that, even among active users, high popularity does not mean high influence in information spreading [19]. Furthermore, the content factor of hashtag use has also been found to be less straight forward; some studies have found significant variations in the way that hastags spread across topics and over time [20].

Effectively spreading news in Twitter is not only about finding influential readers but also about the tweeters themselves *becoming* an efficient source of information. Recent research define efficiency on Twitter as the ratio between the activity employed by users and the emergent collective response as a result to that activity [15]. The study shows that the effective dissemination of a tweet depends not only on its content but also on the user who posts it.

Engaging users to propagate news is not a simple task; however, feature-based models that exploit the content of people's tweets and social interactions can be used to profile the willingness of users to propagate information by retweeting a given tweet [13]. In [6] the authors show that users are inclined to re-share news items more often when they reference socially deviant events.

In conclusion, while many of these studies report key results, they only really hint at possible user and content factors, because they have not specifically addressed journalistic tweeting and audience engagements with news tweets. Therefore, this is the focus of the current work.

3. DO NEWS CATEGORIES DIFFER?

In our examination of audience engagement, we make two strategic choices. First, we focus on journalist accounts and the activity around them. Second, we adopt a content focus in our analyses, distinguishing between different categories of news. We believe the latter distinction to be critical. Different news categories may have different audiences (e.g., one person may only read about sports, while another mainly reads business articles) or the same reader may interact with different categories of news, differently (e.g., Alice may read the business pages during the work week and leave the lifestyle pages to the weekends). If this is, indeed, the case then any analysis of audience engagement must recognize this variable and then determine whether it impacts audience engagement.

Practically, if news categories matter in tweeting, then journalists may need different strategies according to their categories of news. A sports journalist may need to tweet about sports differently to a political journalist that tweets about politics. In the present section, we describe the collection of tweets associated with 200 journalistic accounts and perform initial analyses to determine what aspects of tweeting the news seem to matter; specifically, whether tweeting about one news category may differ from another.

3.1 Data Collection

To study journalistic tweeting, we manually curated a list of 200 Irish journalists' Twitter accounts. These accounts were selected to cover all of the major national and regional media outlets for Ireland, in addition to individual journalists writing for these outlets.

Year	Period	Tweets
2013	Sep 30 – Dec 09	378,893
2014	Nov 20 – Jan 08	335,940
2015	Aug 10 – Jan 20	1,062,681
Total		1,777,514

Table 1: Data collection.

The focus on Irish news sources was an intentional design choice for two reasons. First, these journalists have been shown to be particularly active in social media, by global standards [9]. Second, we wished to build a relatively complete profile of a news ecosystem, in a given locale (see Section 6 for more on this choice).

Using the Twitter streaming API we collected all tweets and retweets posted by each one of the 200 media sources and journalists accounts for three periods in 2013, 2014 and 2015, for a duration of 71, 50 and 163 days, respectively. These periods cover a series of news events including the death of Nelson Mandela, the Charlie Hebdo shooting and the Paris Attacks. We describe the dataset in Table 1.

Each account was manually labeled according to the following aspects (see Table 2 for descriptive statistics):

- Account type: we consider two types of accounts, corporate and individual. Corporate refers to those accounts which do not represent an individual but a corporation as a whole, e.g., @irishtimes, @thejournal_ie, while individual accounts are those which can be directly associated with an individual journalist, e.g., @conor_pope, @MaryFitzger.
- Organization: the newspaper or news outlet with which the account is associated. For example, @irishtimes is associated with The Irish Times, @RTEsoccer to RTE or the journalist @conor_pope works for The Irish Times.
- Gender: the gender of the journalist. We assign a value of zero to corporate accounts.

3.2 Judging the News Categories in Tweets

Our goal in this paper is to identify different characteristics of news tweets that impact upon audience engagement. Research has shown that audience attention is often topic-dependent [1, 20] and, as such, it seems reasonable to expect that the characteristics of tweets that trigger this engagement might also be topic-sensitive.

Categories of News. Most news providers explicitly present and label their news in high level, thematic *news categories* including, sports, business, lifestyle, science and technology, politics, and breaking news.[1] Some journalistic twitter accounts use the *description* field to identify the news category they belong to, for instance, we have seen descriptions such as *"Ireland's premier breaking news website providing up to the minute news and sports reports"* or *"Legal Affairs Correspondent for RTE News. Views my own"*. In many cases, this description alone provides a concise summary of the news category covered by the journalist or news outlet. However, this sort of information is not always present, making the identification of journalistic Twitter accounts with particular news categories non trivial. For this reason, we opted for having independent annotators judge each account to determine its appropriate news category.

Aspect	Distribution
Account type	83 corporate and 117 personal accounts
Organization	79 different news outlets
Gender	31 female and 86 male journalists

Table 2: Distribution of the 200 Twitter accounts, according to type, organization, and gender.

[1]Note that the same news categories can have different names depending on the news provider, we show here particularly representative ones.

Separating Corporate from Individual Accounts. Before we subjected the accounts to this classification judgment, we divided all the 200 accounts into corporate and individual journalist accounts. Out of the 200 Twitter accounts, 117 are individual and 83 are corporate. We consider these two types of accounts as separate in our study because corporate accounts present different patterns of participation and content sharing than individual ones [4]. The 83 corporate accounts are not included in the news categories judgment process, because they often promote news from a wide range of different news categories (e.g., the main *Irish Times* twitter account tweets right across its news categories).

Judging the News Category of an Individual Account. We had three judges manually annotate the news category to which each journalistic account belonged. The possible news categories are business, science and technology, lifestyle, breaking news, politics, and sports. For each of the 117 individual accounts we retrieved a sample of tweets from each year 2013, 2014, and 2015 and raters used this information to make their judgments. Each of the three annotators was assigned one whole year of tweeting for each of the 117 accounts; that is, the first annotator did 2013, the second did 2014 and the third 2015. To judge the category of each account, the annotators were given (i) all the text of the tweets sent by the individual journalist, (ii) a list of the top-100 terms from those tweets, ranked by TF-IDF score. When annotators had assigned all the accounts to news categories we computed the agreement for the categorizations for each account. The annotators were assigned to different years-of-tweets for the same accounts to explore whether a journalist transitioned from one category to another in the periods of study (as journalists are sometimes re-assigned to different news sections over time). [2]

For 59 out of the 117 accounts (50%) judged in this way, the three annotators agreed on the news category of the journalist's account. Of the remaining 58 accounts, at least two annotators agreed on the judgment for 52 cases (a further 44%) and we used majority voting to assign the final label; using Fleiss Kappa we found that the annotators showed an agreement of $\kappa = 0.51$. For the six accounts where there was no agreement, the first author assigned a label after further analyzing the tweets and top terms. The distribution of accounts across the six news categories is shown in Table 3.

News Category	Journalists (number of accounts)
Business	13
Lifestyle	15
Breaking News	30
Politics	25
Science and Technology	6
Sports	28
Total	117

Table 3: News categories and corresponding number of individual journalists' accounts.

[2]Indeed, there was no evidence of such transitions in the judgements made.

3.3 Exploring News Categories

Having made the division between corporate and individual accounts and the judgments of news category differences within individual accounts, we explored whether there appeared to be systematic differences among these subsets of tweets on other dimensions (e.g., time of day or week).

Individual Accounts

Figure 1 illustrates the tweeting activity for the different news categories. Regarding the time of the day (see Figure 1a), sports and lifestyle have a later start with respect to the other categories and the tweeting activity peaks towards the end of the day, particularly between 19:00 and 21:00. Journalists in the business and politics category start tweeting early in the morning and keep a fairly constant activity along the day with a notable decrease close to midnight. Breaking news' journalists post tweets along the day but have two main activity peaks, one in the morning between 08:00 and 11:00 and one in the evening between 19:00 and 22:00, which might correspond to the morning and evening news broadcasts. In science and technology, while being active along the day, the most popular time for posting tweets is 10:00.

In most categories, including business, politics, lifestyle and science and technology, the days with more tweeting activity are Tuesdays and Wednesdays (see Figure 1b). Approximately 40% of the total tweets for the week are sent in these two days in the business category, while for politics Thursdays are also highly active. Sports presents a different behavior, it is the only category for which the activity peaks on the weekend, particularly on Sundays. Breaking news tweets are more evenly spread along the week, with a slight peak on Thursdays.

Figure 1c shows the proportion of tweets sent and retweets received by each category. Sports is the most active and the most popular category, posting approximately 35% of the tweets and receiving more than 40% of the retweets. The second most active category is politics, receiving a greater proportion of retweets than the tweets sent. Breaking news and lifestyle are fairly active as well; however, tweets in the lifestyle category receive considerably fewer retweets. Business and science and technology contribute the least to the tweeting activity during the periods considered in this analysis. It is important to note that the activity shown by each category is not a simple function of the number of journalists' accounts; for instance, the three most popular and active categories are sports, politics and breaking news, but their order based on account numbers is breaking news, sports, and politics (see Table 3).

Corporate Accounts

Figure 2 illustrates the activity of the six news outlets that produce more than 50% of the tweets for the corporate accounts, namely the Irish Independent, the Irish Times, the Irish Examiner, Newstalkfm, The Journal and Irish Independent Sport (IndoSport). These accounts correspond to the most important news outlets in Ireland.

The Irish Independent and the Irish Times have a high tweeting activity between 06:00 and 08:00 (see Figure 2a). The Journal begins tweeting slightly later than the rest of the news outlets and maintains a fairly constant activity along the day. For the Irish Examiner, Newstalkfm and IndoSport the activity peaks around noon.

Most of the tweeting activity among news outlets takes place towards the middle of the week (see Figure 2b), with the exception of IndoSport that, as in the case of the sports category, has the most active days on weekends.

Figure 2c shows the proportion of tweets sent and retweets received by each news outlet. The Irish Independent is the most active and the account that receives the greater proportion of retweets. Posting approximately 13% of the tweets and receiving more than 15% of the retweets. The second most active news outlet is the Irish Times, followed by the Irish Examiner. Interestingly, the second most popular Twitter account among the top news outlets is The Journal, which gets approximately 15% of all the retweets received by corporate accounts. It is worth noting that The Irish Independent seems to opt for a brute-force strategy of tweeting news, being the news outlet with the highest proportion of tweets. The Journal, however, does not follow the same strategy posting half the tweets of The Irish Independent and receiving a comparable proportion of retweets.

Discussion

The initial exploration of journalistic tweets evidences the difference and uniqueness of the tweeting activity across news categories. This diversity might be due to the audience demand that journalists need to satisfy, or simply to the production of news on each category along the day or week.

There are important questions that arise of this exploration, for example, what effect do these differences in activity and volume of tweeting have on the audience engagement to news tweets? Are there characteristics of news delivery through Twitter that engage the readers more than others?

In the remaining of this paper we further explore the differences among news categories and also among news outlets and attempt to shed some light on the characteristics of news tweets that impact on audience engagement.

4. PREDICTING ENGAGEMENT

In this section, we analyze our Twitter corpus of 1M tweets from 200 journalist accounts to determine the features that most impact audience engagement with news tweets. Though, on average, each tweet receives one and a half retweets (M=1.58, SD=6.38), this statistic is somewhat misleading; overall the distribution is exponential with a long-tail in which many journalists' tweets receive no retweets (see Figure 3a; in our analyses we used the natural log of these counts, see Figure 3b).

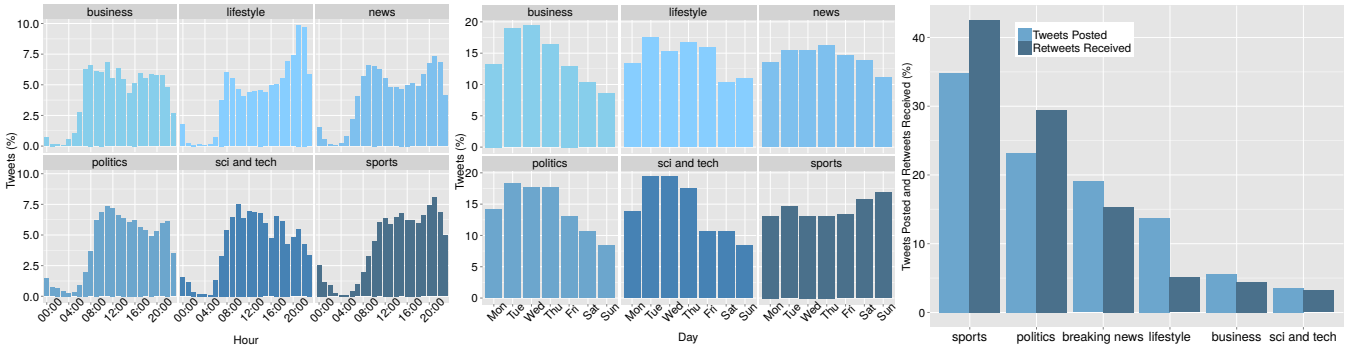
We split our tweet corpus into tweets from corporate accounts (e.g., *@IrishTimes*, *@rte*) and tweets from individual accounts (e.g., the sports' journalist *@MiguelDelaney*); intuitively, interactions with the former seem quite different to the latter. Also, we further subdivide tweets from individual journalist accounts into the six, main news categories (i.e., lifestyle, sports, politics, breaking news, science and technology, and business). As we saw in the previous section, the tweeting behavior of individual journalists differs from one news category to another.

We then extract a set of user features and tweet/content features and represent each tweet as a feature vector to be used in our prediction of audience engagement, which we operationalize as *retweets received* (a commonly used measure of engagement, see e.g., [21]). We explore several regression methods to find the key features that predict audience engagement and assess the relative importance of these features with respect to each of the news categories.

4.1 Method & Procedure

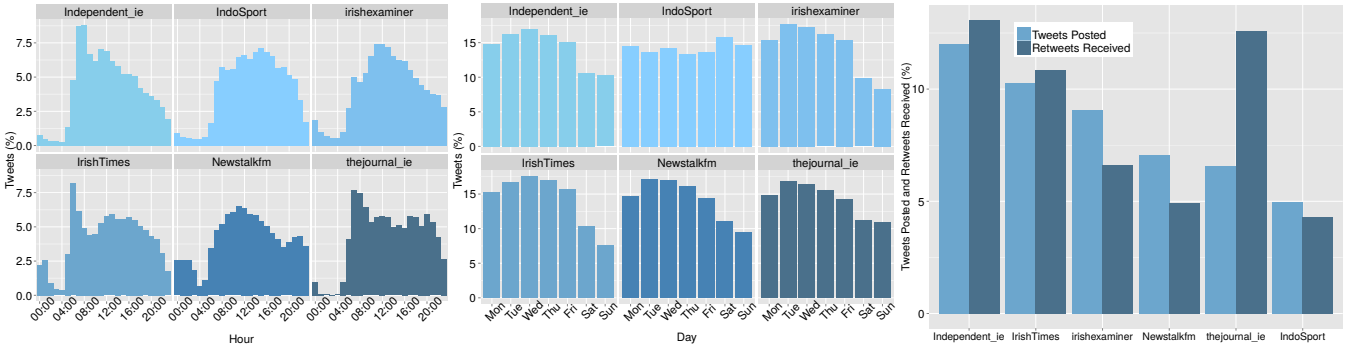
Feature Extraction. We represent all the tweets in the corpus as two-part vectors consisting of user features (e.g., individual or corporate, gender, organization) and content features (e.g., time of day, hashtags, mentions, etc). The complete list of features is presented in Table 4 and can be conceptually grouped into:

- **Temporal:** relating to time and day of creation of the tweets, e.g., *tweets per day segment*, *tweets per day of the week*.



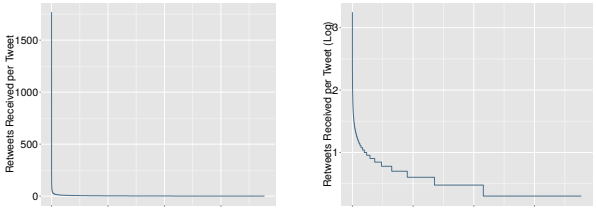
(a) (b) (c)

Figure 1: Distribution of Tweets per (a) hour and (b) day (normalized by the total number of tweets per category) and (c) tweets sent and retweets received per news category (normalized by the total number of tweets sent and retweets received by individual accounts).



(a) (b) (c)

Figure 2: Distribution of Tweets per (a) hour and (b) day (normalized by the total number of tweets per news outlet) and (c) tweets sent and retweets received per news outlet (normalized by the total number of tweets sent and retweets received by corporate accounts).



(a) Retweets per tweet. (b) Retweets per tweet (Log).

Figure 3: Distribution of (a) retweets and (b) natural log of retweets received per tweet.

- **Hashtags**, **Mentions**, and **URLs:** relating to the use and content of diffusion mechanisms, e.g., *contains hashtags*, *mentions per tweet*, or *URLs per retweet*.
- **User popularity:** related to interactions between the user and other users, e.g., *unique mentioners*, *unique retweeters*, or *mentioned by others*.

A considerable amount of Twitter literature has examined "social-network features" such as number of followers/followees or number of times the account has been listed by other users [3, 24]. Although, initially, these features appear to be important, research shows that "the correlation between popularity and influence is quite weak, with the most influential users not necessarily being the ones with the highest popularity" [19]. Hence, we concentrate on other features that may be more important in the journalistic context. Having said this, we do explicitly address socially-related features as we are examining the users that actively engaged with a given journalist's tweets, rather than those acting as passive consumers of information.

Task & Regression Methods. For our audience engagement prediction task, we use regression analysis to estimate the relationship between user and content features and the target variable of audience engagement (i.e., received retweets). We use regression analysis because it can (i) predict a target variable based on a set of values and (ii) screen variables to identify which ones are more important than others to explain the response variable [26].

We explore three different methods for regression, namely, Regularized Linear Regression (RLR), Random Forest (RF) and Gradient Boosting Trees (GBT). In regression the goal is to estimate the relation between one or more independent variables and a single dependent variable, a linear regression model estimates this relation by using a linear predictor function [23]. Random forest is a state-of-the-art meta estimator that fits a number of decision trees on different samples of the dataset, it improves the accuracy of the prediction by averaging the decisions of the involved trees [2]. Gradient boosting produces a prediction model as an ensemble of weak decision trees and it allows the optimization of an arbitrary loss function to avoid the problem of overfitting [8].

Corpus & Data Splits. For these experiments, we use over 1M tweets collected during the six-month period, from August 2015 until January 2016 (see Table 1). We limit our analyses to this set

Journalist/News outlet	
Feature	**Description**
Account type	Personal or corporate
Organization	Account owner/ journalist workplace
Gender	Female, Male or None (if corporate)
Tweets	Tweets posted by this user
Retweets	Retweets posted by this user
Retweets/tweets	Retweets received per each tweet sent
Avg. tweets per day	Avg. number of tweets sent per day
Tweets per day med.	Median of tweets per day
Avg. retweets per day	Avg. number of retweets sent per day
Retweets per day med.	Median of retweets per day
Tweets per day	Tweets sent per each day of the week
Retweets per day	Retweets sent per each day of the week
Tweets per day segment	00:00-08:59, 09:00-16:59, or 17:00-23:59
Retweets per day segment	00:00-08:59, 09:00-16:59, or 17:00-23:59
URLs	URLs included in this user's tweets and retweets
URLs per tweet	Avg. URLs in this user's tweets
URLs per retweet	Avg. URLs in this user's retweets
Hashtags	Hashtags included in this user's tweets and retweets
Hashtags per tweet	Avg. hashtags in this user's tweets
Hashtags per retweet	Avg. hashtags in this user's retweets
Mentions	Mentions included in this user's tweets and retweets
Mentions per tweet	Avg. mentions in this user's tweets
Mentions per retweet	Avg. mentions in this user's retweets
Unique mentions	Unique users mentioned by this user
Mentioned by others	Times this user was mentioned by others
Diff. in mentions	If this user is mentioned more than s/he mentions others
Unique mentioners	Unique users mentioning this user
Total retweets	Total retweets this user received
Unique retweeters	Unique retweeters of this user's posts
Retweets/retweeters	Retweets received per each unique retweeter
Tweet	
Feature	**Description**
Time of creation	00:00-08:59, 09:00-16:59, or 17:00-23:59
Is weekend	If the tweet was posted on a weekend or not
Day of week	The day of the week when the tweet was posted
Is retweet	If the tweet is original or retweet
Contains hashtags	If the tweet contains hashtags
Hashtags simhash	Simhash of the hashtags in the tweet
Contains mentions	If the tweet contains mentions
Mentions simhash	Simhash of the hashtags in the tweet
Contains URLs	If the tweet contains URLs
Domains simhash	Simhash of the domains in the tweet
Retweets received	Retweets received by this tweet

Table 4: List of journalistic account and tweet features.

of tweets because it is the largest set in our collection and because the manual annotation described in Section 3 suggests that journalists are consistent on the news categories they report on over time. Thus, making this tweet set representative for our analyses.

This dataset was split on the corporate/individual dimension, with the latter being further split into the six news categories (lifestyle, sports, politics, breaking news, science and technology, and business). For each one of the seven sub-sets, we created time-wise training, validation, and test splits. This data collection spans from Aug 10, 2015 until Jan 20, 2016, tweets sent within the last 20% of these days, chronologically ordered, are assigned to the test split. Then, from the remaining 80% of the days, we sample the latter 10% to form our validation split, which will help us to select the hyperparameters of our models, and use the former 70% for training. The idea behind the chronological splits is to build models that can learn from past tweet-audience interactions and predict future ones. After selecting the best hyperparameters, we retrained our models on the union of the training and validation splits (i.e., using the tweets sent within the first 80% of the days in our period of study). To account for variability, the results reported are the average of 10 rounds of experiments considering 95% confidence intervals.

Parameters Settings for Methods. Using the validation splits we found that for RLR a regularization constant of 0.1 and a learning rate of 0.0001 led to good results. In the case of GBT and RF

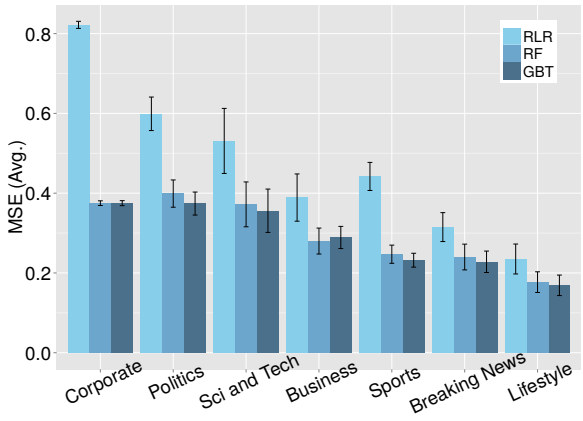


Figure 4: Average MSE values for the different models considering 95% confidence intervals (as these are error values, the lower the value the better the method).

we explored different number of estimators. For GBT the number of estimators that led to good results with the validation splits are 100 for the models lifestyle, breaking news and science and technology, 150 for sports, and 500 for business, politics and corporate tweets. For RF the estimators are 100 for business and breaking news, 150 for politics and 500 for lifestyle, science and technology and corporate tweets.

Metric Used. In order to measure the prediction quality of our models, we use the *Mean Squared Error (MSE)* measure. MSE is a risk function that measures how close a fitted line is to the data points and that is widely used in prediction competitions [14]. We computed the MSE for each tweet in the test set and then look at the average value.

4.2 Results

Figure 4 shows the prediction performance for the three regression methods. GBT and RF perform better than RLR in terms of MSE. The models generated using GBT have a lower error than those using RF, although the difference is not significant. Also, it appears to be harder to predict audience engagement for some news categories than for others. In particular, the model for the tweets associated with corporate accounts shows a high average error across all methods; perhaps, due to the mixed content in these tweets (i.e., they cover many different news categories) and variety of tweeting strategies used across them.

On the basis of this result, we choose to use GBT for our regression task. GBT have shown to outperform other models in classification and regression tasks and have been used successfully for audience engagement prediction [7].

Differential Importance of Features

Apart from being able to predict audience engagement, when one starts to consider guidelines for journalists, it is clearly necessary to understand the relative importance of different features. From each GBT model, we extracted the top-10 features that contribute the most to the predictions, i.e., the features that the models find more important for predicting how many retweets a tweet will receive. To get a broad sense for what features are relevant to each news category, we explored the relative importance of groups of features (according to the five feature-groups defined earlier in this section). The heatmaps in Figure 5 summarize the effects of different feature groups for each of the news categories of the individual accounts (Figure 5a) and for corporate accounts (Figure 5b).

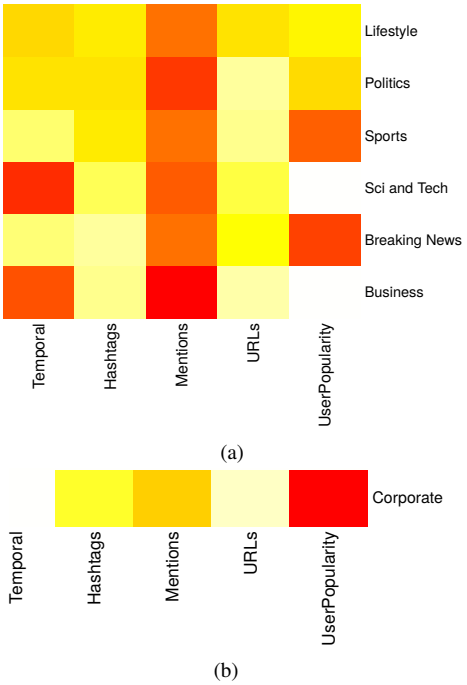


Figure 5: Feature importance for five feature groups (columns) for (a) individual accounts divided into news categories (rows), and (b) corporate accounts. A more intense color indicates this feature group is more important.

For individual journalist accounts, the patterns of importance for different news categories are not the same. Although, the mentions feature group is more important across all of these categories more so than for corporate accounts, the horizontal pattern of importance strengths for each row are notably not identical for any pair of news categories. These patterns of results underscore the importance of breaking out the news category dimension in any future analyses. To consider each news category in turn:

- *Lifestyle*: audience engagement here depends mostly on the use of mentions, followed by temporal issues (i.e., days and times tweets were sent); the inclusion or absence of URLs impacts engagement more than other content features such as hashtags, and user popularity has a role to play (see first row in Figure 5a).
- *Politics*: is strongly influenced by content features, especially the presence/absence of mentions, but *who* posts the tweet is also important, as is the day/time of posting.
- *Sports*: stands out as being strongly influenced by the journalists' popularity and with mentions and hashtags playing a relatively important role in engaging the audience (more than URLs or temporal features).
- *Science and Technology*: by contrast to sports is not driven by user popularity but rather by temporal aspects, perhaps the timeliness of such stories is a critical feature, with content features such as mentions being the most important followed by URLs and hashtags.
- *Breaking News*: is most similar to sports in its profile showing a primacy for who is doing the tweeting (user popularity) and mentions. For breaking news, however, adding or omitting URLs is more important for engagement than it is for news in the sports category, for which hashtags show a higher impact.
- *Business*: tweets attraction of different levels of attention depends highly on the use of mentions, as well as on temporal aspects. Other important features in this category are the inclusion of hashtags and URLs; it is most similar to the science and technology category.

For all the news categories we also observe the features ranked by our models as the least important in predicting engagement and find that *organization* and *gender* show none or little impact in the predictions.

For corporate accounts, the most important feature for audience engagement is user popularity (see Figure 5b); presumably reflecting some sense of brand loyalty. The corporate source of the tweets is valued by audiences of these accounts more than any other aspect. The second most important feature group is that related to mentions, followed by hashtags and URLs, whereas temporal features do not appear to play a significant role.

In the next section, we consider how these results on feature importance might be turned into concrete guidelines for news organizations and journalists tweeting in different areas of the news.

5. GUIDELINES TO HELP JOURNALISTS

Thus far, in this paper, we have uncovered the features that predict audience engagement for news posts in Twitter (see Section 4). In this section, we turn to our second, more practical, goal of turning these analyses into actionable guidelines for journalists; guidelines that are specific enough to enable news providers to design innovative strategies for improving audience engagement.

The predictive analyses revealed the key features that influence audience engagement. These analyses showed that not all features are equal, that some are more important than others and, significantly, that the relative importance of different features changes by news category (e.g., sports versus business). However, a set of guidelines cannot be developed by just "reading off" these features.

To develop guidelines from these analyses we need to further interpret these features, to understand what they specifically mean and, in some cases, to determine their direction of influence. For example, Figure 5a shows that the use of mentions in tweets affects engagement, but the direction of influence for this feature is unclear; that is, it is unclear whether tweets receive engagement by virtue of having greater or fewer mentions. Hence, to develop guidelines, we performed separate analyses using individual decision trees to interpret the different features. These decision trees are not expected to have the predictive power of the ensemble models but they do allow us to interpret the effect of features on the predictions.

First, we separate the tweet corpus into individual and corporate accounts in these analyses, as the guidelines should differ for each. Second, as before, we split the corpus by news category, with the full tweet-set in each category being used to train individual decision trees, casting the problem as an audience engagement prediction task. Then, we traverse each resulting tree to extract the decision rules that lead to the larger values of engagement in the leaves of each tree. The guidelines are then developed from inspecting these outputs and the features' importance discussed in Section 4.

We develop a set of different guidelines. As described above, separate guidelines are developed for individual as opposed to corporate accounts, as engagement with respect to each is quite different. Within the individual accounts analyses, the guidelines were

also divided into general versus specific ones. *General guidelines* deal with steps that can be taken to increase engagement, irrespective of the news category in which the journalist is working. *Specific guidelines* address the guidelines that are applicable within a particular news category (e.g., sports versus business). As we shall see, the latter guidelines are perhaps the most significant, as they suggest very specific interventions that individual journalists can take to promote their news.

5.1 Guidelines for Individual Accounts

Irrespective of the particular news category in which an individual journalist works, two main guidelines are suggested by our analyses:

- *Getting Personal*. Tweets that include mentions that reflect direct interactions with other tweeters are well received by the news audience; this confirms the long-standing advice that there is a personal aspect to Twitter interactions, that a journalist needs to build their audience by direct interaction with them.
- *Enriched Content*. Enriching tweets with hashtags, URLs and/or media content helps to increase engagement; interestingly, the inclusion of URLs has a lower impact than hashtags and, in and of themselves, URLs do not attract better engagement, running counter to the standard practice adopted by most news providers of tweeting links to their articles.

An important finding from the current analyses is that news-category matters, that the features impacting audience engagement change for different categories of news. This finding prompts very specific guidelines for individual journalists working in different content-areas of the news:

Lifestyle

- Weekdays before 5:00 p.m. and Sundays between 9:00 a.m. and 5:00 p.m. are the best times to elicit audience engagement; strongly suggesting a weekend supplement reading audience and perhaps a commuting one.
- Journalists with about 50 unique mentioners attract more retweets to their news, indicating that this category has a strong personal dimension; being known and getting involved in conversations with other users has a positive impact on audience engagement.

Politics

- Mondays, Tuesdays and Thursdays are the best days to attract retweeted responses; as people engage most in the earlier parts of the working week.
- Tweets sent early in the morning (before 9:00 a.m.) and during working hours engage the audience more than if sent during the evening (after 5:00 p.m.).
- Having a wide audience of unique users that retweet one's news, promotes expanding audience engagement; this looks like a *rich gets richer* effect in which a political journalist develops a reputation as *the* expert on a particular topic, who has a wide following for this reason and is promoted by this following, accordingly.
- Interaction with users through mentions is of general relevance for gaining retweets; however, for news in the politics category tweets with mentions are particularly valued.

Sports

- Friday and weekends are the key days to engage the audience; presumably, as this is when major sports events typically occur and people indulge sporting interests in their spare time.
- Being active on a daily basis is important; journalists who post more than 2 tweets a day have a better response from their readers.
- The popularity of the journalist is very important in attracting retweets in this category; aspects such as the high number of retweets received by the journalist in the past and high numbers of unique retweeters determine audience engagement.

Science and Technology

- The inclusion of URLs and particularly the content of these, defines the engagement to the tweets in this category.
- Active journalists who post more than 2 tweets a day receive a better response from the audience.
- Thursdays are the best days to gain retweets for science and technology news.

Breaking News

- Popularity matters in breaking news, as having a larger audience with unique retweeters increases engagement; this feature suggests that one will do better in the breaking news, if many active readers have eyes on your posts.
- Active journalists who retweet and mention others' posts engage more readers; again, perhaps, the personal aspect of being known for breaking stories.
- Temporal aspects, such as the day of the week when the tweet is posted, impact the readers' reactions, as weekdays seem to be better than weekends to gain retweets; however, no one day shows significantly more importance than others, perhaps reflecting the fact that breaking news can occur on any day.

Business

- As in the case of politics, the inclusion of mentions causes a particularly positive impact on engagement for tweets in this category.
- Weekdays are better than weekends to gain retweets and Mondays are the best days to elicit audience engagement; reflecting a mixture of weekend, leisure-time getting up to date with business news and starting the working-week in an engaged way.
- Before 5:00 p.m. is the time period in which tweets receive more retweets in this news category.

5.2 Guidelines for Corporate Accounts

Corporate accounts are analyzed separately from individual journalist accounts. The guidelines advanced here for corporate accounts are of a general nature, in part, because these accounts tend to tweet on many different news categories. Overall, what emerges from their analysis is that such accounts do not present a particularly successful or focused way to distribute one's news. There are many reasons why this might be the case, but perhaps the major one is that these accounts fail to have the personal aspect that is a

very important feature in Twitter use. In one sense, these are non-social accounts trying to exploit aspects of a fundamentally social enterprise. Indeed, one interpretation of the guidelines that emerge here, is that these accounts are really only successful by virtue of brand loyalty in the audience. The guidelines that emerge here are:

- Features concerning user popularity influence the audience engagement for corporate tweets more than any other group of features; in particular, the number of unique retweeters and mentioners is critical as the more people interacting with the account's posts, the more the tweets spread.
- Using mentions, hashtags and URLs leads to more retweets.
- There is no best time of the day to attract retweets in these accounts; however, on any day after 5:00 p.m. tweets can receive a slight increase in audience engagement.

6. CONCLUSION AND FUTURE WORK

This paper began with a discussion of the challenges faced by news media, with respect to third party control of their distribution channels via social media, and their need to develop innovative strategies to deal with such challenges. To answer this innovation challenge, we collected a corpus of news focused tweets from 200 news provider accounts and analyzed them to develop a set of guidelines for journalists who wish to spread their news. As such, this work has two main contributions:

- The main features that impact audience engagement for journalistic news tweets have been surfaced and the ways in which they interact across different news categories revealed.
- These findings have been used to analytically formulate a set of concrete guidelines for news producers to inform their strategy for spreading news on Twitter, whether that news-provider is an individual journalist or a corporate body.

It could be argued that these guidelines are "obvious" or "already known" to journalists. However, there is little evidence to suggest that this is the case; as we do not see any consistent usage of these. Even a cursory glance at the Twitter strategies of major news media, shows no clear agreement on the best way to tweet news. Indeed, some of the current strategies conflict with the guidelines proposed (e.g., the widespread use of corporate accounts).

A further issue might be raised, with respect to this work's emphasis on news providers on Twitter that are largely active on the island of Ireland. This focus was a conscious design choice in our study as we wanted to analyze a coherent news ecosystem in a particular locale. For example, if we had mixed culturally different news organizations (e.g., Irish and French news outlets) a less clear picture may have emerged. Heravi et al. (2014) found that journalists worldwide are increasingly active in social media, reporting that 92% of Irish journalists use Twitter for work, the same percentage as in the UK (92%), and very close to their Canadian (89%), Australian (85%), and American (79%) peers [9]. Hence, our selected sources appear to be representative of a fairly sophisticated, English-speaking news cohort, that should parallel news providers in countries such as the USA, UK, Australia and New Zealand. [3] We would be more cautious about generalizing to very different, non-English speaking cultural contexts (e.g., France, Germany, or Arabic States), where language differences can create very different competitive conditions for news consumption.

[3]At present, we are gathering data to substantiate this claim

Notably, the cohort we have analyzed is quite sophisticated in the use of social media based on recent surveys of the Irish news media ecosystem. In 2015, a country based report by the Reuter's Institute [25] showed that news consumers in Ireland are much more digitally oriented than many other European countries. Irish news readers are heavy consumers of digital news, rely more on social media distribution, and read more of their news on mobile platforms using smartphones. Furthermore, it showed that Irish news providers had to compete with other English-speaking news sources outside Ireland (e.g., BBC, Huffington Post) in a way that did not occur in non-English speaking jurisdictions (e.g., The Netherlands). This report also unveiled that they competed relatively successfully with these much larger, international news sources. Furthermore, another pan-European study has shown that Irish journalists are much heavier users of social media to promote their news, relative to journalists in other countries [9].

In short, the evidence suggests that the journalistic group and audience we have analyzed, appears to be representative of an advanced social media ecosystem for news that is close to best practice or in advance of current practice in other countries.

In the last few years, there has been a concerted move from considering Twitter in general to considering it in niche aspects of the Twitter population. An important part of this move has been a more focused analysis on how journalists and news providers are using Twitter and the consequences of the same. The present work sits within this broad research movement. Future directions of our research will be:

- To experimentally evaluate the proposed guidelines and measure the impact that their practical usage has on audience engagement.
- To further analyze important predictors for audience engagement, such as *mentions*, and explore how different aspects of such predictors (e.g. *who* is being mentioned) impact on the number of retweets received.
- To analyze the extent to which our findings relate to other online news distribution channels and represent different English-speaking news providers.

7. ACKNOWLEDGMENTS

The authors would like to thank *The Irish Times* for their funding and help on this project and Dr. David Coyle for his valuable feedback and advice. This work is supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

8. REFERENCES

[1] S. Asur, B. A. Huberman, G. Szabó, and C. Wang. Trends in social media: Persistence and decay. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*. The AAAI Press, 2011.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[4] M. De Choudhury, N. Diakopoulos, and M. Naaman. Unfolding the event landscape on twitter: Classification and exploration of user categories. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*,

CSCW '12, pages 241–244, New York, NY, USA, 2012. ACM.

[5] N. Diakopoulos, M. D. Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *CHI Conference on Human Factors in Computing Systems, CHI '12*, pages 2451–2460. ACM, 2012.

[6] N. Diakopoulos and A. Zubiaga. Newsworthiness and network gatekeeping on twitter: The role of social deviance. In *ICWSM*. The AAAI Press, 2014.

[7] E. Diaz-Aviles, H. T. Lam, F. Pinelli, S. Braghin, Y. Gkoufas, M. Berlingerio, and F. Calabrese. Predicting user engagement in twitter with collaborative ranking. In *Proceedings of the 2014 Recommender Systems Challenge*, RecSysChallenge '14, pages 41:41–41:46, New York, NY, USA, 2014. ACM.

[8] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002.

[9] B. Heravi, N. Harrower, and M. Boran. Social journalism survey: First National Survey on Irish Journalists' use of Social Media, 2014.

[10] L. Hudson, A. Iskandar, and M. Kirk. *Media Evolution on the Eve of the Arab Spring*. The Palgrave Macmillan Series in International Political Communication. Palgrave Macmillan, 2014.

[11] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 591–600. ACM, 2010.

[12] D. L. Lasorsa, S. C. Lewis, and A. E. Holton. *Journalism Studies*, chapter Normalizing Twitter. 2011.

[13] K. Lee, J. Mahmud, J. Chen, M. X. Zhou, and J. Nichols. Who will retweet this? detecting strangers from twitter to retweet information. *ACM TIST*, 6(3):31, 2015.

[14] K. Metrics, January 2015.

[15] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Efficiency of human activity on information spreading on twitter. *Social Networks*, 39:1–11, 2014.

[16] A. Olteanu, C. Castillo, N. Diakopoulos, and K. Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, pages 288–297. AAAI Press, 2015.

[17] M. Osborne and M. Dredze. Facebook, twitter and google plus for breaking news: Is there a winner? In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014*. The AAAI Press, 2014.

[18] S. Park, M. Ko, J. Lee, A. Choi, and J. Song. Challenges and opportunities of local journalism: a case study of the 2012 korean general election. In *Web Science 2013 (co-located with ECRC), WebSci '13*. ACM, 2013.

[19] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011 (Companion Volume)*, pages 113–114. ACM, 2011.

[20] D. M. Romero, B. Meeder, and J. M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pages 695–704. ACM, 2011.

[21] A. Said, S. Dooms, B. Loni, and D. Tikk. Recommender systems challenge 2014. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 387–388, New York, NY, USA, 2014. ACM.

[22] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[23] H. L. Seal. Studies in the history of probability and statistics. xv: The historical development of the gauss linear model. *Biometrika*, 54(1/2):pp. 1–24, 1967.

[24] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 177–184, Washington, DC, USA, 2010. IEEE Computer Society.

[25] University of Oxford. Reuters Institute for the Study of Journalism. Reuters institute digital news report 2015. Technical report, 2015.

[26] X. Yan and X. G. Su. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2009.

[27] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1513–1522, New York, NY, USA, 2015. ACM.