Spreading One's Tweets: How Can Journalists Gain Attention for their Tweeted News?

Claudia Orellana-Rodriguez, Derek Greene and Mark T. Keane

Insight Centre for Data Analytics, School of Computer Science, University College Dublin, Ireland

ABSTRACT

Traditional news media face many serious concerns as their distribution channels are gradually being taken over by third parties (e.g., bloggers, citizen journalists, and news aggregators). If traditional media is to remain competitive, it needs to develop innovative strategies around these channels, to maximize audience engagement with the news it provides. In this paper, we focus on the issue of developing one such strategy for spreading news on Twitter. Using tweet corpora from two national news ecosystems – 1.7M tweets from 200 journalists in Ireland and 1.2M tweets from 364 journalists in the UK – and audience responses to these tweets, we develop predictive models to identify the features of 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). Using these findings, we suggest a set of guidelines for journalists, designed to help them maximize engagement with the news they tweet. Finally, we discuss how such analyses can inform innovative dissemination strategies in digital media.

Keywords: Computational journalism; Social media; Audience engagement; News events; Twitter; Attention to News; Journalism

ISSN 2332-4031; DOI 10.1561/106.00000009

1 Introduction

Traditional news media now competes, on a daily basis, with bloggers, citizen journalists and news aggregators to gain audience attention for their news. In this ultra-competitive environment, these news providers no longer control the distribution of their news, as articles are shared and forwarded on social media platforms (e.g., Facebook and Twitter). Indeed, the greatest threat to this industry may come from its loss of control of the distribution channels for its products. Facing these challenges, it has become critical for professional journalists to develop tailored strategies for different distribution channels, to maximize engagement with and attention to their news. In this paper, we address this problem in Twitter; an influential social media distribution channel for news. We perform extensive analyses of Twitter corpora from two national news ecosystems (i.e., Ireland and the UK) to identify the features of journalists and their tweets that predict audience engagement with tweeted news. We then use these analyses to suggest guidelines for journalists to help them maximize consequential audience attention to their tweeted news.

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 from their own company (what we will call *news tweets*). Twitter has also become a platform that *is the news*; as politicians tweet their views, celebrities tweet breakups, and citizen journalists report events they have witnessed (e.g., Hudson *et al.*, 2014; Sakaki *et al.*, 2010). 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 *Reuters Institute Digital News Report* 2015).

However, ultimately, Twitter is a distribution channel for news that is not controlled by the news providers. Journalists tweet links to their news articles, but it is the response of the Twitter community that determines whether that news is widely distributed (Park *et al.*, 2013). Therefore, a critical problem for journalists and news organizations is to determine the best strategy to maximize audience engagement with their news in this third-party distribution channel or, to put it more simply, *"What is the best way to spread 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 attention. Indeed, many news outlets are at a point where they have yet to identify 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 by identifying the features of both journalists and their tweets that predict audience engagement. Previous research on Twitter has shown that many tweets tend to be about news (Kwak *et al.*, 2010), that news can first break on Twitter (Osborne and Dredze, 2014), and has identified some of the factors that influence the dissemination of a tweet (Romero *et al.*, 2011b). However, this prior work has seldom specifically focused on journalistic tweeters or, indeed, on news tweets in determining audience engagement (see Section 2).

The present work makes two novel advances. First, it analyzes *news tweets* from two distinct corpora based on journalistic and corporate Twitter accounts in Ireland and the UK; 1.77M tweets involving 200 Irish journalistic accounts and 1.22M tweets involving 364 British journalistic accounts (see Orellana-Rodriguez *et al.*, 2016 for earlier, previous analyses of the Irish corpus). We specifically focus on the *news categories* of these accounts (i.e., tweeted news from the lifestyle, science and technology, politics, sports, breaking news, or business categories), as we hypothesize that the news category would significantly impact engagement. Using the two separate corpora we develop regression models to predict engagement involving these journalistic accounts and tweets. These models give us insights into the features of journalists and their tweets that garner attention on Twitter. Second, from these analyses, a set of guidelines are proposed for journalists when they are tweeting their news; guidelines that are designed to increase audience engagement, and, by extension, attention to their news.

In the next section, we 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, due to 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 using Twitter in their work (Heravi *et al.*, 2014). As such, millions of people consider this social media platform as a primary source of news, and actively seek relevant content to be informed of the latest developments. Twitter alone generates between 12% and 13% of weekly referrals to online newspapers in Ireland and the UK (*Reuters Institute Digital News Report* 2016), and its users differ from those of other social media sites because they are actively seeking news within the platform, instead of just encountering it, as would be the case for Facebook users (*Reuters Institute Digital News Report* 2015). Having said this, all social media interaction, whether it be Twitter or Facebook, is important to news providers. For instance, an analysis of 337 daily newspapers, has shown that online traffic for a newspaper's website is proportional to the size of that newspapers' social media network (Hong, 2012).

Twitter has also emerged as a key source of breaking news (Hudson *et al.*, 2014) and has become a 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 (*Reuters Institute Digital News Report* 2015).

On the negative side, Twitter is nowadays a communication space in which journalists face challenges that can 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 (Lasorsa *et al.*, 2011). Indeed, Lee (Lee, 2015) has shown that self-disclosure and social media interactions by journalists can negatively influence an audience's perception of their professionalism. Furthermore, with the rise of Twitter as a news channel there is concern about the blurring of the boundaries between what the public shares in social media and what the news media publishes online (Olteanu *et al.*, 2015). Arguably, this situation is exacerbated by the citizen journalism aspect of Twitter, which raises new issues about validating Twitter information sources and establishing their veracity, to separate genuine from fake eyewitnesses (Diakopoulos *et al.*, 2012) and genuine from fake news (Waters *et al.*, 2016).

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, though this research has tended not to specifically single out journalist tweeters and news tweets (Kwak *et al.*, 2010; Zhao *et al.*, 2015)

Many of the seminal papers on Twitter address the question of which 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. (Suh *et al.*, 2010) 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.

However, further studies have shown that factors like *popularity* and *use of hashtags* are more complex than first appreciated. With respect to the role of *popularity* (i.e., essentially, numbers of followers), 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 (Romero *et al.*, 2011a). Also, the *use of hashtags* is less straight forward; several studies have found significant variations in the way that hashtags spread across topics and over time (Romero *et al.*, 2011b).

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 has defined efficiency on Twitter as the ratio between the activity employed by users and the emergent collective response as a result to that activity (Morales *et al.*, 2014). This study shows that the effective dissemination of a tweet depends not only on its content but also on the user who posts it. The structure of the social network is also important for the dissemination of news; this structure affects the dynamics of the information spread differently depending on the platform, for example, although in Twitter stories spread slower than in other social media sites (i.e., Digg), they spread farther, depending on the total number of votes they receive, e.g., likes, retweets (Lerman and Ghosh, 2010). Engaging users to propagate news is not a simple task, though feature-based models, that exploit the content of people's tweets and social interactions, have been used to profile users' willingness to propagate information by retweeting a given tweet (Lee *et al.*, 2015).

While many of these studies report key results, they only really hint at possible author 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.

2.3 News Sharing & Popularity in Social Media

The present work specifically addresses the author features of the journalists (e.g., account type -individual, corporate-, gender, and workplace) and the content features of news tweets (e.g., time of creation, if it is original or a retweet, and if it contains mentions) that influence attention to their tweeted news within different news categories. A number of previous studies have addressed aspects of this problem in considering the factors that affect news sharing and popularity.

The sharing of news can be influenced by features of the authors of that news. In (Matias and Wallach, 2015), the authors analyzed a sample of 156K news articles, fitting Poisson regression models to predict impressions (i.e., counts of likes, shares, and reshares). They showed that online news audiences discriminate between articles written by men and women. Articles written by women receive fewer likes, shares, and reshares than those authored by men, and the magnitude of this difference changes depending on the newspaper section (e.g., whether the news category is sports or politics).

Indeed, previous research has also shown that different journalists have different styles of interaction on Twitter and that there are distinct classes of journalistic accounts. Bagdouri (Bagdouri, 2016) analyzed the usage patterns of 5,000 journalists and news organizations on Twitter interacting with 1 million news consumers. He found that Arab journalists' tweets tend to be less personal than English ones; specifically, that English journalists are more engaging and mention other users more often than Arab journalists, who tend to broadcast. He also found that corporate and individual journalist accounts exhibit different behaviors and audience responses; corporate accounts seem to avoid a personal style, while journalists use more personal pronouns (i.e., I, am, my, mine) and ask more questions. Bagdouri also compared Irish and British journalists, as two samples who speak the same language but belong to different countries, and found that the two groups are highly similar in many respects (e.g., retweets received, favorite counts, and number of followers). These findings are relevant to the present work because we also compare Irish and British journalists, while differentiating corporate from individual journalist Twitter accounts.

The sharing of news is also influenced by content features of the items. In (Diakopoulos and Zubiaga, 2014), the authors have shown that people are inclined to share news items more often when they reference socially deviant events. Using a corpus of 8,000 news stories, posted on Twitter by eight major U.S. news outlets, they found that stories involving robbery, homicide, or violence were retweeted more often than those not reporting violations of social or legal norms. Interestingly, these findings did not apply to all the newspapers in the study, perhaps reflecting audience differences, or different interests/focus.

Finally, as in the present study, some previous work has examined the combined influence of author and content features on popularity. In (Bandari *et al.*, 2012), the authors used four features extracted from the content of news articles to predict the popularity of these items on Twitter: (i) the source of the article, (ii) the news category of the article, (iii) the subjectivity in the language, and (iv) the named entities mentioned. They found that one of the most important predictors of popularity is the source of the article, and that top news sources on Twitter are not necessarily the conventionally popular news agencies. In the present work, we explore a much larger set of author and content features (40 in total) across a range of different news categories.

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 speaking, if the category of news matters when tweeting, then journalists may need different strategies, depending upon the topic of their article being discussed or promoted. A sports journalist may need to tweet about sports differently than the way a political journalist tweets about politics. In this section, we describe the collection of 2.9M tweets associated with 564 journalistic accounts and perform exploratory 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 and 364 British journalists' Twitter accounts. These accounts were selected to cover the major national and regional media outlets in the two countries, in addition to individual journalists writing for these outlets.

Irish and British news sources were chosen for two reasons. First, these journalists have been shown to be particularly active in social media, by global standards (Heravi *et al.*, 2014). Second, we wished to build a relatively complete profile of a news ecosystem in two given locales and analyze the extent to which similar patterns of attention to news emerge in different English speaking countries.

Using the Twitter Streaming API[1], we collected all tweets and retweets sent by each of the 200 Irish journalistic accounts for three periods in 2013, 2014 and 2015-16, for a duration of 71, 50 and 163 days, respectively. These periods cover a series of major international news events including the death of Nelson Mandela and the Charlie Hebdo shooting. For the UK dataset, we collected all tweets and retweets posted by 364 UK journalistic accounts for a 238 day period in 2015-16, covering important events such as the refugee crisis and the November 2015 Paris attacks. Besides collecting the tweets sent by the journalists, we also collected tweets reflecting interactions with them (e.g., tweets replying or mentioning any of the journalistic accounts). The datasets are summarized in Table 1.

[1] Every 30 minutes, the crawler collected the new tweets posted and received by the accounts of interest.

Country	Year	Period	Tweets
Ireland	2013	Sep 30 – Dec 09	378,893
Ireland	2014	Nov 20 – Jan 08	335,940
Ireland	2015 – 2016	Aug 10 – Jan 20	1,062,681
UK	2015 – 2016	Aug 10 – Apr 5	1,219,449
Total			2,996,993

Table 1: Data collection (Note that these numbers reflect exclusively the tweets sent by journalistic accounts).

Each account was manually labelled 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, @BBCNews) while individual accounts are those which can be directly associated with an individual journalist (e.g., @conor_pope, @NickyAACampbell).
- Organization: the newspaper or news outlet with which the account is associated. For example, @irishtimes is associated with The Irish Times, @RTEsoccer with RTE, @TimesNewsdesk with The Times, the journalist @conor_pope works for The Irish Times, or @NickyAACampbell works for the BBC.
- Gender: the gender of the journalist. We assign a value of zero to corporate accounts.

3.2 Finding the News Categories of Tweets

Our hypothesis is that the news category of a news tweet may importantly determine how people come to engage with that tweet (see e.g., Asur *et al.*, 2011; Romero *et al.*, 2011b). We consider the problem of identifying the news categories of tweets, as a precursor to using this variable in subsequent analyses of engagement.

Categories of News. Most news providers explicitly present and label their news articles in high-level, thematic *news categories*, including, sports, business, lifestyle, science and technology, politics, and breaking news.[2] 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 *"BBC business journalist covering banks, economy, EU, companies, UK & Ireland, consumers, government, markets etc. I sometimes do #r4Today"*. 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 mapping of journalistic Twitter accounts to particular news categories non-trivial. While this problem could be addressed by automated methods (e.g., LDA, clustering), to ensure quality, we used manual annotation by independent judges to identify the news category of journalistic accounts.

Aspect	Distribution
Ireland	
Account type	83 corporate and 117 individual accounts
Organization	79 different news outlets
Gender	31 female and 86 male journalists
UK	
Account type	58 corporate and 306 individual accounts
Organization	90 different news outlets
Gender	85 female and 221 male journalists

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

In this analysis, we associate each journalist to a single news category, which corresponds to the one that she/he specializes on and the main one covered by her/his tweets. The working assumption is that individual journalists tweet about a single news category from their account, that they do not tweet equally on multiple news categories, and that they do not use these accounts to tweet mainly on non-news issues. This assumption is based on the following observations: (i) during their career, journalists tend to become experts and specialize on one single section of the news. Their focus is reflected on their Twitter accounts, e.g., expert sports journalists commenting and analyzing rugby championships will rarely, if at all, dedicate the same coverage to other categories such as science and technology, business, or politics, (ii) our news category annotation of accounts is based on manually reviewing samples of tweets from a given journalist, and in all the cases, as we describe later in this section, the annotators were able to assign one main news category to each account. Individual accounts tweeting mainly on non-news subjects, as reported by our annotators, were excluded from the analysis. A more detailed study to account for the existence of particular cases on which journalists tweets span more than one news category would be interesting, but at present, it is out of the scope of our current work.

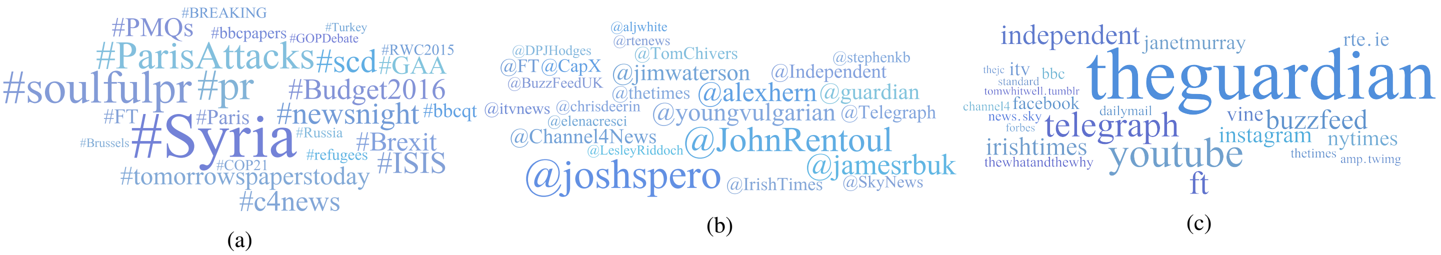
Separating Corporate from Individual Accounts. Before submitting the accounts to our judges, we divided them into corporate and individual journalist accounts. Out of the 564 Twitter accounts, 423 are individual and 141 are corporate. Corporate accounts are quite distinct from the accounts of individual journalists as they present different patterns of participation and content sharing (De Choudhury *et al.*, 2012). The 141 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 all of its news categories).

Judging the News Category of an Individual Account. Three judges manually annotated the news category to which each journalistic account belonged in the Irish and UK corpora. The news categories included business, lifestyle, breaking news, science and technology, politics, and sports. To judge the category of each account, the annotators were given (i) a random sample of 50 tweets sent by the individual journalist and (ii) a list of the top-100 terms used by the journalist in her/his tweets during the period of interest, ranked by TF-IDF score. Annotators were also asked to decide whether the tweets from an account were non-news tweets; any account that was found with such non-news tweets was removed from the analysis.

To make the annotation process clearer, we also provided definitions[3] of the news categories to our annotators. These definitions were

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

[3] The definitions were extracted from the Wikipedia's pages on the corresponding journalism branches.



(a) (b) (c)

Figure 1: Top-25 (a) hashtags, (b) mentions, and (c) domains used by individual journalists in their tweets (normalized by total number of hashtags, mentions, and domains, respectively).

as follows:

- **Business.** Branch of journalism that tracks, records, analyzes and interprets the economic changes that take place in a society. It could include anything from personal finance, to stock exchange, entrepreneurship, business at the local market, and shopping malls, to the performance of well-known and not-so-well-known companies.

- **Lifestyle.** News on relationships, real people, families, health, travel, fitness, fashion, interiors.

- **Breaking News.** Current issues that broadcasters feel warrant the interruption of scheduled programming and/or current news in order to report their details. Its use is also assigned to the most significant story of the moment or a story that is being covered live.

- **Politics.** Includes coverage of all aspects of politics and political science, although the term usually refers specifically to coverage of civil governments and political power. Political journalism is a frequent subject of opinion journalism, as current political events are analyzed, interpreted, and discussed by news media pundits and editorialists.

- **Sci and Tech.** News on the techniques, methods or processes used in the production of goods or services or in the accomplishment of objectives, such as scientific investigation, or any other consumer demands. Gadgets, technology of any kind, explanations and predictions about nature and the universe.

- **Sports.** Reports on sporting topics and competitions.

When annotators had assigned all the accounts to news categories the inter-rater agreement was computed for their judgments. For 59 out of the 117 Irish accounts (50%) judged in this way, the three annotators agreed on the news category. Of the remaining 58 accounts, at least two annotators agreed on the judgment for 52 cases (a further 44%). Majority voting was used to assign the final news category label to a given account. In the case of the remaining six accounts where there was less agreement, the first author assigned a label after further analysis of the tweets and top-scoring terms. The Fleiss' Kappa inter-rater agreement for Irish accounts is $\kappa = 0.51$. For 164 out of the 306 British accounts (54%) the three annotators agreed on the news category. Of the remaining 142 accounts, two annotators agreed on 129 cases (42%), and for the remaining 13 accounts where there was less agreement among the annotators, the final label was assigned by the first author. For British accounts, the Fleiss' Kappa inter-rater agreement was similar to the one found for Irish data, $\kappa = 0.58$. The distribution of accounts across the six news categories is shown in Table 3.

News Category	**Ireland**	**UK**
Business	13 (11%)	26 (8%)
Lifestyle	15 (13%)	80 (26%)
Breaking News	30 (26%)	70 (23%)
Politics	25 (21%)	91 (30%)
Science and Technology	6 (5%)	28 (9%)
Sports	28 (24%)	11 (4%)
Total	117	306

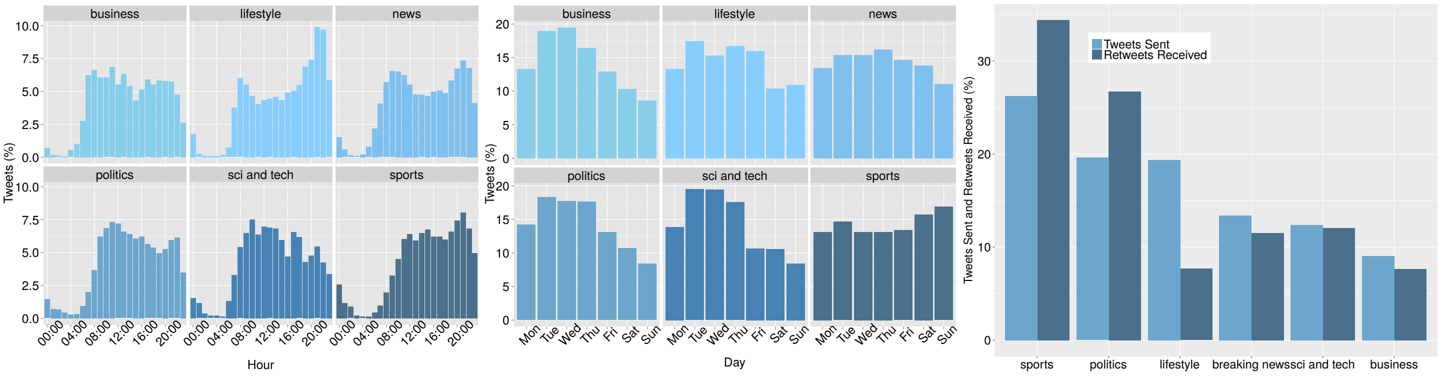
Table 3: News categories and corresponding number of individual journalists' accounts.

The news oriented nature of these accounts can be gleaned from high-level descriptions of them. First, of the 423 individual journalists' accounts categorized by our annotators, 209 (49.6% of the Irish accounts and 49.3% of the British accounts) are verified[4] by Twitter. Second, the top-25 hashtags, mentions and domains used in these accounts uniformly address news topics and current affairs (see Figure 1). The most used hashtags refer to topics, such as *#Syria, #ParisAttacks,* and *#Brexit*, and to journalists/broadcasters, such as *#bbcpapers, #FT,* and *#c4news*. The top mentions include journalists, such as *@joshspero*, deputy editor at Financial Times Special Reports, *@johnRentoul*, chief political commentator at The Independent, and *@jamesrbuk*, special correspondent at BuzzFeed UK, as well as news outlets, such as *@guardian, @Independent,* and *@IrishTimes*. Third, the domains referenced foiund to be news providers, such as *the Guardian, the Independent, the Telegraph,* and *the Irish Times.* Fourth, we verified that the owners of these accounts were using them predominantly for tweeting news and not for other non-news communications. We selected a random sample of 1,000 tweets from these accounts and had two annotators manually label them as news-related or non-news related; this experiment revealed that 92% of this tweet-sample were news-related, with a Cohen's Kappa inter-rater agreement of $\kappa = 0.86$.

3.3 Exploring News Categories

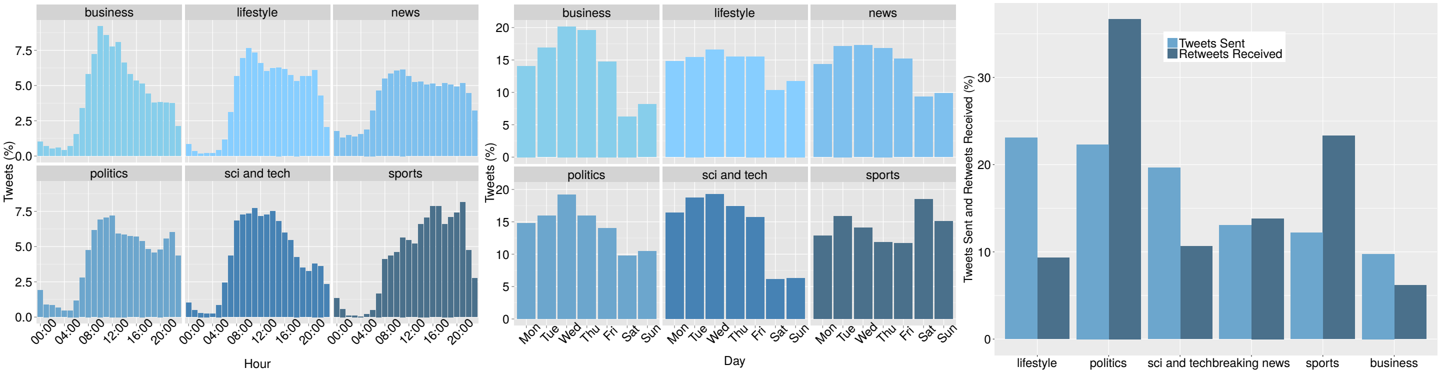
Having made the division between corporate and individual accounts and having labeled the news category of individual accounts, we explore whether there appear to be any systematic differences between

[4] A verification badge is given by Twitter to accounts of public interest to confirm their authenticity: https://support.twitter.com/articles/119135



(a) (b) (c)

Figure 2: Tweets posted by news categories in Ireland (a) per hour, (b) per day (normalized by the total number of tweets per category), and (c) proportion of tweets sent and retweets received by news category (normalized by number of individual accounts in each category in Ireland).



(a) (b) (c)

Figure 3: Tweets posted by news categories in the UK (a) per hour, (b) per day (normalized by the total number of tweets per category), and (c) proportion of tweets sent and retweets received by news category (normalized by number of individual accounts in each category in the UK).

these subsets of tweets on other dimensions (e.g., time of day). In this exploration we use the 2.9M tweets – 1.7M for Ireland and 1.2M for the UK.

Individual Accounts: Activity Levels Across Countries

Figures 2 and 3 illustrate the tweeting activity for the different news categories in Ireland and the UK, respectively. In terms of the time-of-day, in both countries (see Figures 2a and 3a), tweeting about sports and lifestyle news starts later in the day relative to other categories and tweeting levels are skewed towards the end of the day. Many of these sports/lifestyle tweets are posted at approximately 10:00, with tweeting activity peaking between 19:00 and 21:00 (though lifestyle tweets peak a bit earlier in the day in the UK). Journalists tweeting on business and politics start with a burst of tweets early in the morning and then maintain a relatively constant level of tweeting throughout the day, with a notable decrease close to midnight. Interestingly, in the UK as opposed to Ireland, a notable peak in activity in these two categories is reached between 08:00 and 12:00. Breaking news' journalists also post tweets throughout 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, perhaps corresponding to the morning and evening news broadcasts. The evening peak, although present in both countries, is more marked in Ireland than in the UK; also, tweeting activity around breaking news is higher in the UK from midnight to 06:00. Tweeting about science and technology news occurs at a fairly constant level, though the most popular time for posting tweets is approximately 10:00 in both countries.

In most categories, the mid-week days (i.e., Tuesdays, Wednesdays and Thursdays) show the highest levels of tweeting, though this pattern is more pronounced in business, science and technology, politics, and lifestyle news (see Figures 2b and 3b). Approximately 50% of the total weekly tweets are sent during these mid-week days. However, tweeting around sports runs counter to this pattern showing activity peaks in the weekend days. In Ireland, this sports-news tweeting is particularly active on Sundays; whereas, in the UK, Saturdays are the most active weekend day. Interestingly, Tuesdays present a higher activity for sports tweets than any other week day for both countries. Breaking news tweets are more evenly spread throughout the week, with a decrease on the weekends that is more notable in the UK.

When we break out the proportions of tweets sent and retweets received by these accounts (see Figures 2c and 3c), there are notable differences between news categories. In Ireland, sports and politics account for the lionshare of tweeting and also have the highest levels of engagement, with approximately a 60% proportion of the total retweets being received by these two categories (see Figure 2c). The other four categories account for less of the overall tweeting activity;

notably, in the lifestyle category the response (proportion of retweets received) is substantially lower. In the UK, lifestyle and politics account for the lionshare of tweeting, closely followed by science and technology though again we can see that proportionally, retweets received by politics and sports are much higher than those received by the rest of the accounts (see Figure 3c). In the other news categories, breaking news shows a relatively good balance between the tweets sent and retweets received, while business contributes substantially less activity.

It is important to note that these activity levels for different news categories are not a simple function of the number of journalistic accounts in the category; for instance, in Ireland the order of categories based on their activity (i.e., proportion of tweets sent) is sports, politics, lifestyle, breaking news, science and technology, and business, but their order based on account numbers is breaking news, sports, politics, lifestyle, business, and science and technology (see Table 3).

Corporate Accounts: Activity Levels Across Countries

Figures 4 and 5 illustrate the activity levels of the six news organizations that generated approximately 50% of the tweets for all corporate accounts in Ireland and the UK, respectively.

In Ireland, the *Irish Independent*, the *Irish Times*, the *Irish Examiner, Newstalkfm, The Journal*, and *Irish Independent Sport (IndoSport)*, are responsible for a high proportion of all news tweets sent from corporate accounts in the country (see Figure 4). These accounts correspond to the most important news outlets nationwide. The *Irish Independent* and the *Irish Times* have high levels of tweeting activity between 06:00 and 08:00 (see Figure 4a). *The Journal* begins tweeting slightly later than the rest of the news outlets and maintains a fairly constant activity throughout the day. For the *Irish Examiner, Newstalkfm* and *IndoSport*, the activity peaks around noon. Most of the tweeting activity by news organizations in Ireland takes place towards the middle of the week (see Figure 4b), with the exception of *IndoSport* that, as in the case of the sports category (see Figure 2b), has the most active days on weekends.

Figure 4c 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 number of tweets of the *Irish Independent* and receiving a comparable proportion of retweets. These patterns show us that different news organizations have diverse tweeting policies, policies that elicit very different levels of engagement.

In the UK, we observe a similar phenomenon to that found in Ireland. Forty-one percent of all news tweets are sent by the six accounts shown in Figure 5. These highly active Twitter accounts are *Sky News, Huffington Post UK, Financial Times, FT, BBC News*, and *The Economist*. *Sky News* and the *Huffington Post* are most active around noon (see Figure 5a), while *BBC News, Financial Times*, and *FT* present a more distributed activity with discrete peaks early in the morning and in the evening. Of all of these accounts *The Economist* differs in that it has a high, constant activity that spans midnight and early morning hours without a marked decrease, perhaps reflecting a policy to reach across time-zones to a more international readership. Note, that even though *The Economist* differs in being a weekly magazine published each Friday, its Twitter account is active from day to day. As illustrated in Figure 5b, weekdays are highly active for British news organizations. In particular, Tuesdays, Wednesdays, and Thursdays. *The Economist*, however, presents a quasi-constant activity that, in contrast to all the other accounts, increases towards the weekend.

Figure 5c shows the proportion of tweets sent and retweets received by each British news outlet, revealing two important phenomena: (i) 32% of the total retweets are received by *The Economist* and *Sky News*, and (ii) while the *Financial Times, FT, BBC News*, and *The Economist* have roughly equal levels of tweeting, they each engender very different patterns of engagement (some attract low levels of retweeting, others like *The Economist* attract massive engagement). This evidence suggests that different news organizations have different tweeting policies, that have markedly different outcomes in terms of engagement (as measured by retweets received). Indeed, these diverging patterns can even be seen witin the same news outlet. The account *@FinancialTimes* posts news stories, features and updates from the *Financial Times* whereas *@FT* (also from the *Financial Times*) posts only the headlines corresponding to this news. Both accounts post tweets at approximately the same times and days but *@FT* receives close to double the number of retweets than the *@FinancialTimes* (see Figure 5c), possibly indicating that the *Financial Times*' audience is more prone to read and share the headlines than longer news tweets.

Discussion

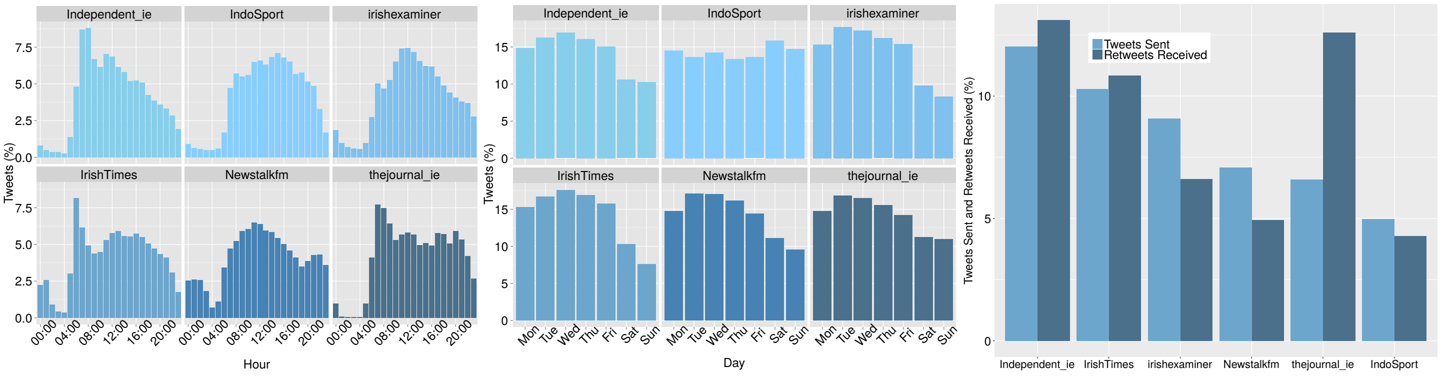
This initial exploration of journalistic tweets shows interesting differences in tweeting and retweeting activity across news categories. This diversity might be due to the audience demand that journalists need to satisfy, or simply to the production of news on each category throughout the day or week. For both countries, Ireland and the UK, we observe similarities between corresponding categories. For example, business and politics are most prolific in the mornings and midweek, sports is more active over weekends, and breaking news seems to follow the morning and evening news broadcasts. Notably, this data indicates that news category is clearly an important variable in determining levels of engagement when tweeting the news.

A further conclusion warranted by this data is that, for diverse reasons, news outlets are differentially successful in eliciting engagement from their readers. For example, *The Irish Independent* (Ireland) and *Sky News* (UK) carry out high volume tweeting that appears to elicit correspondingly high levels of retweeting. However, other outlets (e.g., *The Journal* in Ireland and *The Economist* in the UK) tweet much less but elicit very high levels of retweeting engagement.

In the remainder of this paper we analyze the drivers for these very different behaviors, to understand what it is about journalists and their tweets that leads to high levels of audience engagement.

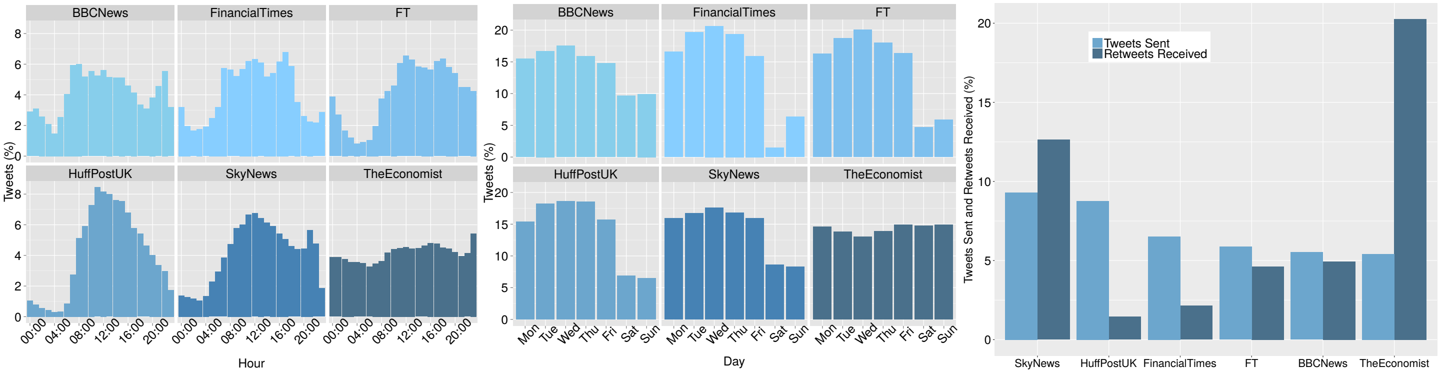
4 Predicting Engagement

To determine the features of journalists and tweets that impact engagement, we performed an analysis of our two Twitter corpora from Ireland (1.06M tweets from 200 Irish journalistic accounts) and the UK (1.22M tweets from 364 British journalistic accounts). Note that, for the feature analyses, we only use the tweets collected in the years 2015-2016 for both countries, in order to have two data samples from



(a) (b) (c)

Figure 4: Tweets posted by corporate accounts in Ireland (a) per hour, (b) per day (normalized by the total number of tweets per news outlet), and (c) tweets sent and retweets received per news outlet (normalized by total number of tweets sent & retweets received by these accounts in Ireland).



(a) (b) (c)

Figure 5: Tweets posted by corporate accounts in the UK (a) per hour, (b) per day (normalized by the total number of tweets per news outlet), and (c) tweets sent and retweets received per news outlet (normalized by total number of tweets sent & retweets received by these accounts in the UK).

roughly the same periods of time and thus avoid introducing possible noise due to external factors, such as platform changes. Retweet counts can be misleading in these corpora (e.g., for the Irish corpus each tweet has M = 1.58 retweets, SD = 6.38), as the overall distribution is exponential with a long tail in which many journalists' tweets received no retweets (see Figure 6a). Accordingly, we used the natural logs of these retweet counts[5] in our analyses (see Figure 6b).

Each of the country tweet corpora was divided into tweets from corporate accounts (e.g., *@IrishTimes*, *@BBCNews*) and tweets from individual accounts (e.g., the sports' journalist *@MiguelDelaney*); intuitively, engagement with the former appears to be quite different to that with the latter. The tweets from individual journalist accounts were further subdivided into the six, main news categories (i.e., lifestyle, sports, politics, breaking news, science and technology, and business). Note that corporate accounts cannot be separated by news category because they often tweet across all of them.

Taking these datasets, a set of user features and tweet/content features was extracted and each tweet was represented as a feature vector to be used in predicting audience engagement, which was operationalized as *retweets received* (a commonly used measure of engagement; see e.g., Said *et al.*, 2014). For the large majority of tweets, the lifespan[6] is merely hours, almost 100% of the tweets are rarely retweeted after 72 hours since being posted (Kong *et al.*, 2012). To take into consideration the lifecycle of the tweets, the retweet counts were computed only after the data collection process was completed.

Several different regression methods were explored to find the key predictive features for audience engagement and assess the relative importance of these features in different news categories. As we shall see, Gradient Boosting Trees were found to give the best results and, therefore, formed the basis for our subsequent analyses.

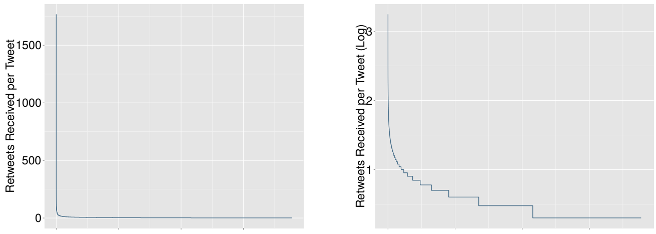
4.1 Method & Procedure

Feature Extraction. We represent all the tweets in the corpora as two-part vectors consisting of user features (e.g., individual or corporate account, 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 Features:** relating to time and day of creation of the tweets, e.g., *tweets per day segment*, *tweets per day of the week*.

[5] We computed the natural logs of (retweet count + 1), to account for those tweets with a retweet count of zero.

[6] Period of time where the tweet is receiving retweets



(a) Retweets per tweet.

(b) Retweets per tweet (Log).

Figure 6: Distribution of (a) retweets and (b) natural log of retweets received per tweet in Ireland.

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

Previous work on Twitter has shown that certain *social network features* are important to engagement, features such as the number of followers/followees or number of times the account has been listed by other users (Cha *et al.*, 2010; Suh *et al.*, 2010). Although, initially, these features appeared to be important, research shows that "the correlation between popularity and influence is quite weak, with the most influential users are not necessarily those with the highest popularity" (Romero *et al.*, 2011a). Hence, we concentrated more on other features that appeared to be more important in the journalistic context. Having said this, we do address socially-related features, as we are examining the users that actively engaged with the journalists' tweets, rather than those acting as passive consumers of information.

Task & Regression Methods. For our audience engagement prediction task, a regression analysis was used 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 those that are most important in explaining the response variable (Yan and Su, 2009).

In the analyses, we used three different methods for regression: 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 (Seal, 1967). 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 trees involved (Breiman, 2001). 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 (Friedman, 2002).

Journalist/News outlet Features	
Temporal Features	
Feature	**Description**
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
Hashtags, Mentions, URLs Features	
Feature	**Description**
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
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
User and Popularity Features	
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
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 mentions	Unique users mentioned by this user
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 Features	
Temporal Features	
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
Hashtags, Mentions, URLs Features	
Feature	**Description**
Contains hashtags	If the tweet contains hashtags
Hashtags simhash [7]	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
Is retweet	If the tweet is original or retweet

Table 4: List of journalist/news outlet features and tweet features (grouped into conceptual categories).

Corpora & Data Splits. In these experiments, the Ireland and UK corpora were treated separately and, hence, results are reported by country. The datasets were 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 subsets, we created time-wise training, validation, and test splits. For example, the dataset from the UK spans from Aug 10, 2015 until April 05, 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 sampled the latter

[7] SimHash is a similarity hash function which stores a set of hash keys and auxiliary data per file, to be used in determining file similarity (Sadowski and Levin, 2011).

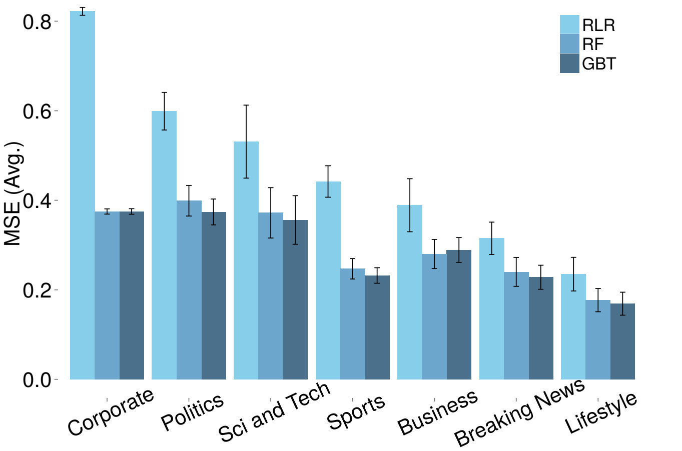
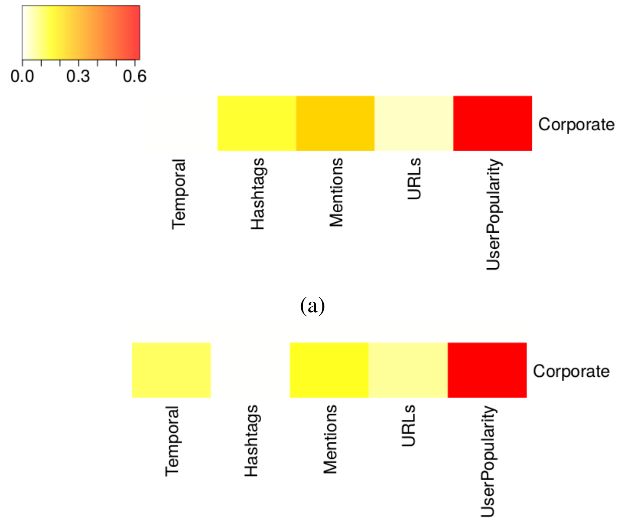


Figure 7: Average MSE values for the different models in predicting engagement based on the Ireland Twitter corpus, showing 95% confidence intervals (as these are error values, the lower the value the better the method). A similar pattern was observed for the UK Twitter corpus.

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., tweets sent in the first 80% of the days). To account for variability, the results reported are averaged over 10 rounds of experiments considering 95% confidence intervals. For the Ireland corpus we use the same procedure.

Parameters Settings for Methods. For the Ireland corpus 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 we explored different numbers of estimators. For GBT the number of estimators that performed best on 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, sports, and corporate tweets. In the UK corpus, good results were obtained for RLR by setting the regularization constant to 0.01 and the learning rate of 0.001. The only exception was for the science and technology category, where the best parameter values were 0.1 and 0.001, for the regularization constant and the learning rate, respectively. For GBT, the models of business, breaking news, science and technology, sports, and corporate tweets, reached good results with 150 estimators; while for lifestyle and politics, 500 estimators performed better. For RF, 100 estimators for the model of the lifestyle category, 150 for politics, and 500 for business, breaking news, science and technology, sports, and corporate tweets performed better on the validation splits.

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 (e.g., Metrics, 2015). We computed the MSE for each tweet in the test set and then took the average value across all these tweets.



(a)

(b)

Figure 8: Feature importance of five feature groups (columns) in predicting engagement for corporate accounts in (a) Ireland and (b) the UK. Intensity of color indicates higher importance.

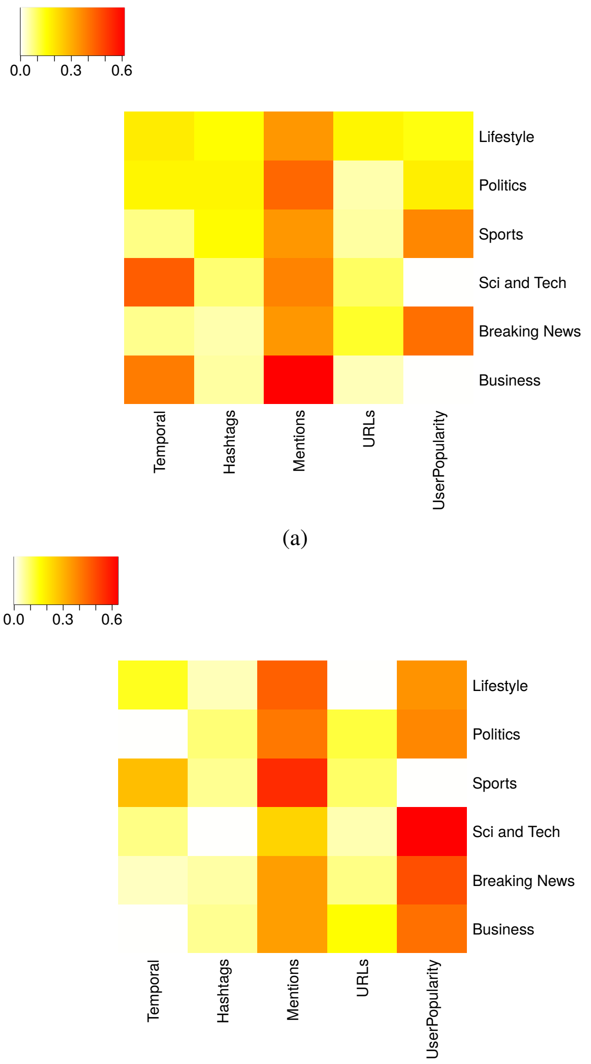
4.2 Results

Figure 7 shows the prediction performance for the three regression methods applied to the Irish Twitter corpus. 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 models for the tweets associated with corporate accounts show a slightly higher average error; perhaps, due to variety of content in these tweets (i.e., they cover many different news categories) and their diverse tweeting strategies. This result was found in the corpora for both countries. On the basis of these results, we chose to use GBT for the regression task. GBT have been shown to outperform other models in classification and regression tasks and have previously been used to predict audience engagement (e.g., Diaz-Aviles *et al.*, 2014).

As one of our goals is to develop specific guidelines for journalists to optimize their tweeting strategy, we need to understand the differential importance of features in predicting engagement. Hence, from each GBT model, we extracted the top-10 features that contributed most to the predictions; that is, the features that the models find more important for predicting how many retweets a tweet will receive. The heatmaps shown in Figures 8 and 9 summarize the relative importance scores of features, for the Ireland and UK corpora, in both corporate and individual accounts, with the individual accounts further broken out by news category. Overall, these analyses indicate that corporate accounts differ from individual accounts in the relative importance of features and that in individual accounts there are differential patterns of feature importance in different news categories (i.e., tweeting about sports demands a different strategy to tweeting about business). Interestingly, also, there are notable differences between countries.

Corporate Accounts: Importance of Features

In corporate accounts for both countries, the relative importance of different features is broadly the same. In both countries, the most



(a)

(b)

Figure 9: Feature importance for five feature groups (columns) per news category (rows) for individual accounts in (a) Ireland and (b) the UK. Intensity of color indicates higher importance.

important feature for audience engagement is user popularity (see Figures 8a and 8b); presumably reflecting some sense of brand loyalty. That is, engagement is largely predicated on the reading audience valuing of these accounts, possibly as an authoritative source. The second most important feature group, in both countries, is the mentions group, which could be indicating that the audience of these accounts does respond to news tweets that have a personal-like touch added to them. This feature group is followed by URLs, perhaps reflecting the tendency of corporate accounts to cite their articles in their tweets. Lesser roles are played by hashtags and temporal features. As we shall see, this pattern of relative importance of features turns out to be quite distinct from what we find in the different individual accounts.

Individual Accounts: Importance of Features

The individual accounts, for both countries, were divided by news category to determine if the category of the tweeted news impacts engagement. Notably, we find that groups of features are differentially important in different news categories. We also see that though there are broad similarities across Ireland and the UK, the specific relative importance of features is not identical (see Figures 9a and 9b). The corporate versus individual account differences are also interesting. In the individual accounts (see Figure 9) user-popularity features are less critical than in corporate accounts (see Figure 8), in favor of a greater reliance on the mentions feature group.

In comparing countries, for individual accounts, there is a general tendency for mentions to be the most important feature in Ireland and the UK. However, the countries diverge in the next most important feature groups. In Ireland, the temporal aspects of the tweet become critical with the use of hashtags and, to some extent, user popularity playing more of a role. In the UK, user popularity is the next most important feature, after mentions, with the use of URLs coming third. In this respect, engagement in the UK appears to be more personality-driven, based on who is doing the tweeting than it is in Ireland. Overall, looking at the Figures 9a and 9b, perhaps the most notable difference between Ireland and the UK is the interplay of features. In Ireland, all feature groups play a nuanced role across different news categories, whereas in the UK there seems to be clearly dominant feature groups within news categories. This may reflect a smaller-audience effect (in Ireland) where more diverse preferences emerge than in a larger audience where wisdom-of-the-crowd effects reduce diversity (n.b., these effects cannot be attributed to differential corpora size).

In comparing news categories, for individual accounts, the patterns of relative feature importance are quite different (i.e., comparing rows). This result underscores the importance of breaking out the news category in future analyses of engagement. To consider each news category in turn:

- *Lifestyle*: audience engagement depends mostly on the use of mentions, followed, in Ireland, by temporal issues (i.e., day and time tweets were sent) and in the UK by the popularity of the journalist; for both countries, the inclusion or absence of hashtags has an impact on engagement, although this effect is more notable in Ireland (see first row in Figures 9a and 9b).

- *Politics*: is strongly influenced by content features, especially by mentions, but *who* posts the tweet is also important, particularly in the UK. In Ireland, the day/time of posting the tweets is similar in importance as is the journalist's popularity.

- *Sports*: stands out as being strongly influenced by mentions, with URLs and hashtags playing a relatively important role in engaging the audience. Temporal features are important for both countries, possibly reflecting an audience engagement that depends on the coverage of relevant sports events. The popularity of the journalists is of higher importance for Ireland than for the UK, in this category.

- *Science and Technology*: in contrast to sports is more driven by user popularity in the UK and by temporal aspects in Ireland. The timeliness of such stories is, however, of importance for the two countries, with content features such as mentions being the most relevant.

- *Breaking News*: shows a primacy for *who* is doing the tweeting (user popularity) and mentions. Adding or omitting URLs is more important for engagement than the use of hashtags.

This category shows a highly similar behavior in the two countries. The breaking news category is similar to the corporate accounts in that it covers news across several topics; interestingly, this similarity is also observed in the resulting feature importance (see Figure 8).

- *Business*: engagement with the tweets depends highly on the use of mentions, as well as on temporal aspects for Ireland and on user popularity for the UK. Other important features in this category are the inclusion of URLs and hashtags.

For all the news categories we also observed the features ranked by our models as the least important in predicting engagement, and find that the *organization* the journalist works for and the *gender* of the journalist show little to no impact in the predictions.

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: Helping Journalists Gain Attention

The previous analyses have revealed the features of importance in predicting audience engagement for news related posts on Twitter for two English-speaking countries, namely, Ireland and the UK. In this section, we consider the more practical goal of converting these analyses into actionable guidelines for journalists. These guidelines should be specific enough to enable news providers to design innovative strategies for improving audience engagement.

The predictive analyses reveal the key features that influence audience engagement. They show 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 results.

To develop guidelines from these analyses, we need to further interpret the features, to understand what they specifically mean and, in some cases, to determine their direction of influence. For example, Figures 9a and 9b show that the use of mentions in tweets affects engagement, but the direction of influence for this feature is uncertain, as it is not clear whether tweets receive engagement by virtue of having greater or fewer mentions. Hence, to develop guidelines, we perform a separate set of analyses using individual decision trees that, together with the results presented in Section 4, allow us to interpret the direction of influence of the different features. Note that 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 different features on the predictions.

In these new analyses, the tweet corpora (1.06M tweets from Ireland and 1.2M from the UK) are separated into individual and corporate accounts, as clearly the guidelines should differ for each type. Then, as before, we split the individual accounts 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. Finally, each resulting tree is traversed 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 feature importance results discussed in Section 4.

Using this methodology, separate guidelines were 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 as opposed to 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 factors that are important 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 Corporate Accounts

The guidelines for corporate accounts are quite general, in part, because they tend to tweet on many different news categories. Indeed, in the case of these accounts, it is possible that the tweeting activity is already being regulated by the use of scheduling products or algorithms to optimize tweeting times for different audiences; however, despite these efforts, it is clear that using corporate accounts does not present a particularly successful or focused way to distribute one's news; largely, because these accounts fail to have the personal aspect that is a key feature of impact on Twitter. In one sense, these are non-social accounts trying to exploit aspects of a fundamentally social enterprise. The guidelines for these accounts hinge on making them more social, tapping their brand-loyalty aspects:

- 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.

5.2 Guidelines for Individual Accounts

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

- *Getting Personal*. Mentions that reflect direct interactions with other tweeters are well received by the news audiences in both Ireland and the UK; this confirms the long-standing advice that there is a personal aspect to Twitter posts, 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, in Ireland 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. In the UK, we see a different scenario in which URLs are slightly more important than hashtags.

The current analyses found that news category matters when tweeting, and that the features impacting audience engagement vary for different categories of news. We also observed that overall, the patterns of importance for different feature groups are quite similar in Ireland and the UK. These findings prompt us to propose specific guidelines for journalists working in different content areas:

Lifestyle

- Wednesdays and Tuesdays 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.
- Including mentions in the tweets is important in this category, however, to maximize engagement, journalists should attract an audience that also mentions them, as this results in more retweets.

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. This proposal applies especially to journalists in Ireland.
- 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.). In the UK, we observe that political journalists who are highly active in this specific time frame (e.g., working hours) also get retweets for occasional tweets sent at irregular times, for example, late in the evening. Which might suggest that when the British audience is aware of a journalist's activity, they are more prone to react to posts sent outside regular times.
- Having a wide audience of unique users that retweet one's news and mention one in their posts, 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, and who has built up a significant following for this reason where they are 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

- Weekends are the key days to engage the audience; presumably, as this is when major sports events typically occur and when people follow their sporting interests in their spare time.
- On weekdays, the best time to gain retweets for sport posts is Tuesday evening. Possibly suggesting a working audience interested in updates on events that took place in the past weekend or on schedules for the upcoming days.
- In Ireland, being active on a daily basis is important for sports journalists; those who post more than 2 tweets a day have a better response from their readers.
- The popularity of the journalist attracts retweets in this category; aspects such as the high number of retweets received by the journalist in the past are important in both countries, but in Ireland, high numbers of unique retweeters also 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 more responses from the audience.
- Weekdays are better than weekends to gain retweets in this category. Particularly, Thursdays and Fridays.
- In the UK, the audience's response also depends on user popularity aspects.

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. In Ireland, no one weekday shows significantly more importance than others. In the UK, tweets posted on Tuesdays and Wednesdays obtain more retweets.

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 in Ireland, 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. In the UK, readers also show certain engagement with business news on Saturdays.

- Before 5:00 p.m. is the time period in which tweets receive more retweets in this news category. Particularly in Ireland.

6 Conclusion: Caveats, Criticisms & 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 remain competitive. To address these challenges, we collected a corpus of news focused tweets from 564 news provider accounts in Ireland and the UK and analyzed them to develop a set of guidelines for journalists, designed to improve audience engagement. As such, this work has two main contributions: (i) 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, and (ii) 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.

Obviously, as with any piece of research, there are a number of caveats and criticisms that need to be considered, ones that may usefully define future work in the area. In this paper, we have explored a wide set of author and content features to assess their impact on engagement, in the context of different news categories. However, it is clearly the case that, even using 40 features, we have not exhausted the full set of potentially important ones. There may well be other features inherent to news, that also affect engagement (e.g., newsworthiness, importance, temporal aspects of sharing). For instance, traditionally news providers consider the notion of *newsworthiness* as a key feature that attracts reader attention; we saw earlier that research has shown a preference for deviant behavior stories (see Diakopoulos and Zubiaga, 2014). If it is possible to find an operational definition for such features, it is clear that they deserve to be explored in future work.

Furthermore, a working assumption in our analysis was that the tweets sent by journalists were news oriented and tended to focus on a single news category. To this end, we excluded any accounts that were found to be tweeting non-news items. However, we did not subdivide these tweets based on whether or not the journalist involved expressed personal opinions or tended to be more discussive in their interactions (Lee, 2015). Clearly, work could be done to automate the classification of tweets in order to reveal the more subtle features of news commentary (see e.g., Diakopoulos and Zubiaga, 2014). A detailed analysis of tweet content could provide a more definitive indication of the extent to which journalists stay "on topic" in relation to their area of expertise, and whether this behavior is consistent across different geographic regions and media outlets.

With respect to the guidelines, one could argue that they are "obvious" or "already known" to journalists. However, there is little evidence to suggest that this is the case; as the consistent use of particular strategies is not evident. 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 concern might be expressed over the generality of the results found given our focus on Ireland and the UK. Journalists worldwide are increasingly active in social media; it has been shown 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 (Heravi *et al.*, 2014). Hence, the countries we have targeted 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. 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.

Notably, the Irish journalistic cohort may well be 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 Reuters Institute (*Reuters Institute Digital News Report* 2015) 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 most of their news on mobile platforms using smartphones. Furthermore, this report showed that Irish and British news providers compete with other English-speaking news sources in a way that did not occur in non-English speaking jurisdictions (e.g., The Netherlands). This report also found that these outlets competed relatively successfully with much larger, international news sources. 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 may well be 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 its 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. There are future directions to our research, including an experimental evaluation of the proposed guidelines to measure their impact in day-to-day journalistic usage, and an analysis of the important predictors for audience engagement, such as *mentions*, to explore how different aspects of such predictors (e.g. *who* is being mentioned) impact on the number of retweets received. The analyses and results presented in this paper, surface the ways in which news audiences interact across different news categories. These findings can help inform journalists strategies for spreading news on Twitter.

Acknowledgments

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

References

Asur, S., 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. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2815.

Bagdouri, M. “Journalists and Twitter: A Multidimensional Quantitative Description of Usage Patterns.” In: *ICWSM*. AAAI Press. 22–31. ISBN: 978-1-57735-758-2. URL: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2016.html#Bagdouri16.

Bandari, R., S. Asur, and B. A. Huberman. “The Pulse of News in Social Media: Forecasting Popularity”. *CoRR*. abs/1202.0332. URL: http://dblp.uni-trier.de/db/journals/corr/corr1202.html#abs-1202-0332.

Breiman, L. “Random Forests”. English. *Machine Learning*. 45(1): 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A%3A1010933404324.

Cha, M., 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*.

De Choudhury, M., 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*. Seattle, Washington, USA: ACM. 241–244. ISBN: 978-1-4503-1086-4. DOI: 10.1145/2145204.2145242. URL: http://doi.acm.org/10.1145/2145204.2145242.

Diakopoulos, N., 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*. ACM. 2451–2460. DOI: 10.1145/2207676.2208409. URL: http://doi.acm.org/10.1145/2207676.2208409.

Diakopoulos, N. and A. Zubiaga. “Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance.” In: *ICWSM*. The AAAI Press. ISBN: 978-1-57735-659-2. URL: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#DiakopoulosZ14.

Diaz-Aviles, E., 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*. Foster City, CA, USA: ACM. 41:41–41:46. ISBN: 978-1-4503-3188-3. DOI: 10.1145/2668067.2668072. URL: http://doi.acm.org/10.1145/2668067.2668072.

Friedman, J. H. “Stochastic Gradient Boosting”. *Comput. Stat. Data Anal.* 38(4): 367–378. ISSN: 0167-9473. DOI: 10.1016/S0167-9473(01)00065-2. URL: http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

Heravi, B., N. Harrower, and M. Boran. “Social Journalism Survey: First National Survey on Irish Journalists’ use of Social Media”. Ed. by G. HuJo Insight Centre for Data Analytics National University of Ireland. URL: http://www.irishtimes.com/business/media-and-marketing/irish-journalists-among-world-s-heaviest-social-media-users-study-finds-1.2101471/.

Hong, S. “Online news on Twitter: Newspapers’ social media adoption and their online readership.” *Information Economics and Policy*. 24(1): 69–74. URL: http://dblp.uni-trier.de/db/journals/iepol/iepol24.html#Hong12.

Hudson, L., A. Iskandar, and M. Kirk. *Media Evolution on the Eve of the Arab Spring*. *The Palgrave Macmillan Series in International Political Communication*. Palgrave Macmillan. ISBN: 9781137403162.

Kong, S., L. Feng, G. Sun, and K. Luo. “Predicting Lifespans of Popular Tweets in Microblog”. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. *SIGIR ’12*. Portland, Oregon, USA: ACM. 1129–1130. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348503. URL: http://doi.acm.org/10.1145/2348283.2348503.

Kwak, H., 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*. ACM. 591–600. DOI: 10.1145/1772690.1772751. URL: http://doi.acm.org/10.1145/1772690.1772751.

Lasorsa, D. L., S. C. Lewis, and A. E. Holton. “Journalism Studies”. In: chap. Normalizing Twitter. URL: http://dx.doi.org/10.1080/1461670X.2011.571825.

Lee, J. “The Double-Edged Sword: The Effects of Journalists’ Social Media Activities on Audience Perceptions of Journalists and Their News Products.” *J. Computer-Mediated Communication*. 20(3): 312–329. URL: http://dblp.uni-trier.de/db/journals/jcmc/jcmc20.html#Lee15.

Lee, K., 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. DOI: 10.1145/2700466. URL: http://doi.acm.org/10.1145/2700466.

Lerman, K. and R. Ghosh. “Information contagion: An empirical study of the spread of news on Digg and Twitter social networks”. In: *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.

Matias, J. N. and H. Wallach. “Working Paper: Modeling Gender Discrimination by Online News Audiences.” In: *Computation + Journalism Symposium, Columbia University*.

Metrics, K. Ed. by kaggle.com. URL: https://www.kaggle.com/wiki/Metrics.

Morales, A. J., J. Borondo, J. C. Losada, and R. M. Benito. “Efficiency of human activity on information spreading on Twitter”. *Social Networks*. 39: 1–11. DOI: 10.1016/j.socnet.2014.03.007. URL: http://dx.doi.org/10.1016/j.socnet.2014.03.007.

Olteanu, A., 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*. AAAI Press. 288–297. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10583.

Orellana-Rodriguez, C., D. Greene, and M. T. Keane. “Spreading the News: How Can Journalists Gain More Engagement for Their Tweets?” In: *Proceedings of the 8th ACM Conference on Web Science*. *WebSci ’16*. Hannover, Germany: ACM. 107–116. ISBN: 978-1-4503-4208-7. DOI: 10.1145/2908131.2908154. URL: http://doi.acm.org/10.1145/2908131.2908154.

Osborne, M. 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. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8072.

Park, S., 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. DOI: 10.1145/2464464.2464523. URL: http://doi.acm.org/10.1145/2464464.2464523.

“Reuters Institute Digital News Report”. *Tech. rep.* URL: http://www.digitalnewsreport.org/.

“Reuters Institute Digital News Report”. *Tech. rep.* URL: http://www.digitalnewsreport.org/.

Romero, D. M., 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)*. ACM. 113–114. DOI: 10.1145/1963192.1963250. URL: http://doi.acm.org/10.1145/1963192.1963250.

Romero, D. M., 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*. ACM. 695–704. DOI: 10.1145/1963405.1963503. URL: http://doi.acm.org/10.1145/1963405.1963503.

Sadowski, C. and G. Levin. "SimiHash: Hash-based Similarity Detection".

Said, A., S. Dooms, B. Loni, and D. Tikk. "Recommender Systems Challenge 2014". In: *Proceedings of the 8th ACM Conference on Recommender Systems*. *RecSys '14*. Foster City, Silicon Valley, California, USA: ACM. 387–388. ISBN: 978-1-4503-2668-1. DOI: 10.1145/2645710.2645779. URL: http://doi.acm.org/10.1145/2645710.2645779.

Sakaki, T., 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*. Raleigh, North Carolina, USA: ACM. 851–860. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772777. URL: http://doi.acm.org/10.1145/1772690.1772777.

Seal, H. L. "Studies in the History of Probability and Statistics. XV: The Historical Development of the Gauss Linear Model". English. *Biometrika*. 54(1/2): 1–24. ISSN: 00063444. URL: http://www.jstor.org/stable/2333849.

Suh, B., 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*. Washington, DC, USA: IEEE Computer Society. 177–184. ISBN: 978-0-7695-4211-9. DOI: 10.1109/SocialCom.2010.33. URL: http://dx.doi.org/10.1109/SocialCom.2010.33.

Waters, R., M. Garrahan, and T. Bradshaw. "Harsh truths about fake news for Facebook, Google and Twitter". *Financial Times*. Nov. [Online].

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

Zhao, Q., 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*. Sydney, NSW, Australia: ACM. 1513–1522. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783401. URL: http://doi.acm.org/10.1145/2783258.2783401.