Data Analytics Methods for the Dissemination and Consumption of News

by

Claudia Orellana Rodriguez
Student Number: 14207084

This thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy

in

School of Computer Science and Informatics

Head of School: Professor Pádraig Cunningham

Principal Supervisor: Professor Mark Keane

Doctoral Studies Panel Membership:

Professor Pádraig Cunningham
Assoc. Prof. Neil Hurley

May 2019

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

While much is known about how people tweet and interact on Twitter, surprisingly little is known about how the news items tweeted by journalists – *news-tweets* – act as a distribution channel for the news that spreads on social media. Traditional news providers have been assaulted by one disruption after another from people sharing news on social media, to people becoming *citizen journalists*, to news aggregation sites re-packaging their products. As always, however, such disruptions present as many challenges as they do opportunities; opportunities that depend on understanding the dynamics of news in this brave new world of social media.

Most research in news production, dissemination, and consumption in social media, aims at making sense of the large amount of user generated content and interactions available in these platforms. The focus of most works has been determined by tasks such as automatic news stories summarization, commentary filtering, fake news detection, or personalized news recommendations. However, a journalist's success in using social media platforms depends more on a deep understanding of current journalistic practices and news consumers behavior on these platforms.

In this thesis, we study current practices of professional news providers and news consumers on Twitter, and then use these analyses to address important tasks that will help journalists develop new strategies relevant to news production, dissemination, and consequent consumption. Specifically, we present analyses on three domains of the news ecosystem: 1) journalistic practices on Twitter, 2) news providers, and 3) news audiences.

To lay the basis of our work, we start by presenting an extensive analysis of *current journalistic practices on Twitter,* accompanied by a survey in which English-speaking journalists report on their professional usage of the platform. The outcome of this first part allows us to identify important problems and challenges faced by professional journalists who use Twitter as a tool for news distribution.

The second part of the thesis takes as input the knowledge gathered from the analysis of journalistic practices and focuses on *news providers*, their strategies for news production and dissemination on Twitter and how these differ depending on the news category. Using tweet corpora from two national news ecosystems – Ireland and 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.

In the third part of this thesis, we demonstrate how behavioral patterns of news consumption can be used to automatically profile *news audiences*. Here we present the Twitter News Model (TNM), a computational data-driven approach to elucidate the dynamics of news consumption on Twitter. We apply the TNM to a dataset of interactions between users and journalists/newspapers to uncover *what* drives users to consume news on Twitter, and predictively relate users' news beliefs, motivations, and attitudes to their consumption of news. Our findings reveal that news motivations, followed by news attitudes and news beliefs, impact users' behavior of news consumption on Twitter.

In the fourth and final part of this thesis, we present some practical tools that can benefit journalists who use Twitter as a news dissemination platform. We discuss how our analyses can inform innovative dissemination strategies in digital media, and present a software tool to help newspaper editors compose effective headlines for online publication. The system identifies the most salient keywords in a news article and ranks them based on both their overall popularity and their direct relevance to the article, it also uses a supervised regression model to identify headlines that are likely to be widely shared on social media. The user interface is designed to simplify and speed the editor's decision process on the composition of the headline. As such, the tool provides an efficient way to combine the benefits of automated predictors of engagement and search-engine optimization (SEO) with human judgments of overall headline quality. As the final step in the headline design process, recommendations given by our tool are assessed by a news editor, who makes the ultimate selection on what keywords to use in the headline based on the character and culture of the news organization.

With the ongoing proliferation of social media platforms, the need for the development of novel strategies relevant to news production, dissemination, and consequent consumption will continue apace. Our work contributes to advance our understanding of digital news delivery and consumption on Twitter.

Statement of Authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Signature ______________________

Para Aurora.

Que en tu corazón siempre haya lugar para nuevas aventuras. Que tu curiosidad y deseos de aprender no conozcan límites. Que nada ni nadie te detenga.

ACKNOWLEDGEMENTS

This journey would not have been possible without the support of many people. First and foremost, I would like to thank my supervisor, Mark Keane. Before starting my PhD, I used to think how much I would like to share this journey with a supervisor who deeply and sincerely cared about my research and also about providing guidance and support through good and bad times. Having Mark as my supervisor has been exactly that. He supported me and encouraged me on each and every project, he always made time to read my drafts, discuss ideas and even write with me. The excitement with which he approaches research is contagious.

During my Ph.D., I have been supported by the amazing team at the Insight Centre for Data Analytics. I would like to express my sincere gratitude to the PIs, lecturers, postdocs, administrative staff, and fellow Ph.D. students who were always there when I needed them. They went to great extents to ease concerns, give support, provide guidance and feedback, and in more than one occasion, to lend an ear. Special thanks go to Dr. Georgiana Ifrim, Dr. Derek Greene, Dr. David Coyle, and fellow students Hannah, Andy, Terry, Maryanne, Yuriy, Erika, An, Gevorg, and Igor. This journey has been much more enjoyable because of them.

I would also like to thank my family. To pursue this path has meant to be far away from them for several years, and despite the time and distance, I have always felt their love and support. Special thanks go to my mom and dad, Isolina and Ismael, whose lives are an example of respect, diligence and determination. To my sister, Flor, whose hard work and eagerness to succeed are inspiring. And to my niece and nephew, Juliette and Christian, whose joy, laughs, and curiosity are hearth-warming. This achievement has their mark on it.

Finally, I would like to express my deepest gratitude and appreciation to Ernesto. It has been amazing to share this journey with him. His endless encouragement, support, and love have been a source of strength to me. Thank you for sharing my dreams.

LIST OF PUBLICATIONS

- C Orellana-Rodriguez and MT Keane. **Attention to News and its Dissemination on Twitter: A Survey**. Computer Science Review, Elsevier, 2018.

  COR gathered and reviewed the literature. MTK provided supervisory guidance for the design of the review. COR wrote the manuscript in consultation with MTK.

- C Orellana-Rodriguez and MT Keane. **Modeling and Predicting News Consumption on Twitter**. The 26th Conference on User Modeling, Adaptation and Personalization (UMAP), Singapore, Singapore, July 2018.

  COR gathered the data and performed experiments. COR and MTK contributed to the design and implementation of the research. COR wrote the manuscript in consultation with MTK.

- C Orellana-Rodriguez, D Greene and MT Keane. **Spreading One's Tweets: How Can Journalists Gain Attention for their Tweeted News?**. The Journal of Web Science, Now Publishers, 2017.

  COR and DG gathered the data. COR performed experiments. COR and MTK contributed to the design and implementation of the research. COR wrote the manuscript in consultation with MTK and DG.

- G Ifrim, D Greene, MT Keane, C Orellana-Rodriguez, B Shi and G Poghosyan. **On Supporting Digital Journalism: Case Studies in Co-Designing Journalistic Tools**. Computation + Journalism Symposium, Evanston, IL, USA, October 2017.

  COR, DG, BS and GP gathered data. COR, BS and GP designed and performed experiments. GI, DG and MTK provided supervisory guidance on the design of the experiments. MTK consolidated the research into the manuscript.

- C Orellana-Rodriguez, D Greene and MT Keane. **Spreading the News: How Can Journalists Gain More Engagement for their Tweets?**. The 8th ACM Conference on Web Science (WebSci), Hannover, Germany, May 2016.

  DG gathered the data. COR performed experiments. COR and MTK contributed to the design and implementation of the research. COR wrote the manuscript in consultation with MTK and DG.

- T Szymanski, C Orellana-Rodriguez and MT Keane. **Helping News Editors Write Better Headlines: A Recommender to Improve the Keyword Contents & Shareability of News Headlines**. Natural Language Processing meets Journalism. IJCAI. New York, NY, USA, July 2016.

  COR and TS gathered the data and performed experiments. COR, TS and MTK contributed to the design and implementation of the research. COR and TS wrote the manuscript in consultation with MTK.

These publications have emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

CHAPTER

ONE

Introduction

While much is known about how people tweet and interact on Twitter, surprisingly little is known about how the news items tweeted by journalists – *news-tweets* – act as a distribution channel for the news that is spread by social media reading and sharing.

Social media has disrupted news dissemination and the limits that once separated journalists from news consumers are now blurred. Sites such as Twitter, Facebook, and YouTube enable users to create, read and share news content in a social way. Twitter, in particular, has attained a special status as *the* social media platform for news; it has become a venue where news consumers and journalists converge to report, read, discuss and share the news. Unlike other social media platforms (e.g., Facebook, Instagram, or Youtube) where news is mainly encountered as a side-effect of other interactions, many Twitter users specifically turn to the site to track news developments [134, 1, 133]. Accordingly, this micro-blogging platform is favored by journalists and news organizations to live-tweet news events that are happening in real-time or to share incremental reporting threads and other content [7].

Apart from enabling the spreading of news and updates, Twitter presents great potential for learning valuable insights from journalists and audience usage of the platform. As a result, research in areas such as news sourcing and distribution [41, 64], development of journalistic tools for generation, recommendation or filtering of news articles [135], and the analysis of news audiences' behavior and needs [25], has increased. In order for news providers to take advantage of the many possibilities that social media have to offer, a deeper understanding of journalists and news audiences' behavior in such platforms is necessary. The work presented in this thesis tries to span analyses of both, news providers and readers, and provide valuable insights into the world of social media for news.

1.1 Motivation

Traditional news providers have been assaulted by one disruption after another from people sharing news on social media, to everyone becoming a potential "citizen journalist", to a tweeting President that refuses to give traditional press briefings. Importantly, each of these disruptions are progressively undermining the business model of these news organizations, damaging their traditional income streams and depleting their paying audience. As always, however, such disruptions present as many challenges as they do opportunities; opportunities that depend on understanding the dynamics of news in this brave new world of social media. If traditional news outlets are to survive then, critically, they need to understand the dynamics of news production, consumption and dissemination in social media.

1.1.1 News production and dissemination

On a daily basis, conventional news media now competes with bloggers, citizen journalists and news aggregators to gain audience attention for their news. In this ultra-competitive environment, news providers no longer control the production and distribution of their news, as articles are posted, 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 thesis, we address these problems in Twitter.

Twitter 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 – what we will call *news-tweets* – from their own company. 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 [1]). Furthermore, 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., [72, 152]).

Just over a decade after its creation, nearly all major news providers use Twitter to both advertise their news articles and promote interaction with their reader audience. And, while social media has clearly changed journalistic practice in many positive ways, it also appears to have changed news publication in negative ways. Many worry that the *social* aspects of Twitter undermine journalistic integrity; for instance, while a highly interactive journalist benefits from more engagement with their audience they may do so at the expense of appearing less objective by virtue of these interactions [36, 93].

The key issue in all of this is that, Twitter is a distribution channel that is not controlled by 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 [129]. Therefore, a critical problem for journalists and news organizations is to determine the best strategy to maximize the dissemination of 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 the dissemination of 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.

1.1.2 News consumption

Daily newspaper reading and the viewing of national TV news, have been traditionally correlated with a civic obligation to stay informed about current events [106]. For older people, the daily habit of following the news and this civic sense seems to have persisted as news has moved online and onto social media. For younger people, news consumption has, perhaps, become a more personalized activity to gather information about events that directly affect them or, more a matter of social interaction, as they use news items in online conversations with friends and colleagues [1].

For news providers, it is critically important to understand the dynamics of news consumption in the social media context, or *what* drives news consumption in social media users; this would allow them to better serve their audiences and motivate future news consumption [110]. Indeed, users who read the news on social media often play an important role in modern journalism, not only as direct consumers, but also as gatekeepers, with almost half the social media users sharing and reposting news stories, images, or videos, and discussing news issues or events online [133]. In addition, journalists who post/share their news and interact with social media audiences, can use their knowledge of such factors to influence people's satisfaction [82], which may in turn, have a positive impact in news consumption.

Social media is a particular environment in which news consumption is not a passive activity, but rather one in which users interact with news providers, express their interests, ask questions, or request more information, via mechanisms such as shares, likes/dislikes, retweets, or mentions. In this thesis, beyond addressing popular tasks such as the prediction of news consumption, we attempt to present valuable insights obtained in the process. These insights are important because if journalists learn what makes users consume news, they can apply this knowledge to future interactions and thus, keep their audience and/or possibly attract more readers.

1.2 Research Objectives

Most research in news production, dissemination, and consumption in social media, aims at making sense of the large amount of user-generated content and interactions available on these platforms. The focus of most works has been determined by tasks such as automatic news stories summarization, commentary filtering, fake news detection, or personalized news recommendations. However, journalists' success in using social media platforms depends more on a deep understanding of current journalistic practices and news consumers behavior on these platforms. In this thesis, we study current practices of professional news providers and news consumers in Twitter, and then use this analysis to address important tasks that will help journalists develop novel strategies relevant to news production, dissemination, and consequent consumption.

Specifically, the thesis has the following research objectives:

1. **To explore journalistic strategies for news distribution and how such strategies impact audience attention to news.** To achieve this objective we take a two-step approach, we (i) survey English-speaking journalists on several aspects of their professional usage of Twitter, and (ii) analyze two Twitter corpora which contain real journalists interactions with the platform and the corresponding audience responses. These explorations are presented in Chapters 3 and 4.

2. **To design practical tools that can help journalists get more attention for their news.** In an effort to bridge the gap between technical findings and journalistic practice, this thesis presents a method to interpret results obtained from a prediction task so that they can be directly used by journalists in their daily practices.

   In addition, we explore the construction of news headlines for online newspapers and design a recommender system that suggests modifications of proposed headlines, based on their keyword content and predicted *shareability* in social media platforms. These tools are presented in Chapter 6.

3. **To understand what drives people to consume news on Twitter.** Beyond predicting news consumption, in this thesis we explore *what* drives people to consume news. The findings from this exploration could inform strategies that journalists might use to promote their news. A detailed description of this analysis is presented in Chapter 5.

1.3 Contributions

While a vast amount of work aims to understand the usage and benefits of social media for news, their focus has seldom been specific to journalistic practices and news providers' behavior in such platforms. In fact, many of these works do not use news-specific datasets nor have tried to understand news distribution, dissemination and consumption within a journalists-news consumer ecosystem.

In this thesis, we concentrate on news-specific environments within the vast world of social media, and make five main contributions, the work:

1. Establishes how recent and current literature concretely informs the news industry's response to their current challenges.
2. Uncovers the features of *news-tweets* that impact audience readership, attention, and engagement.
3. Addresses the important issue of closing the gap between computer science research and journalistic practice by considering practical guidelines for optimizing journalistic use of social media platforms when publishing the news.
4. Reviews the key theoretical descriptors (i.e., news beliefs, motivations, and attitudes) that impact news consumption in *conventional* news media; then proposes a Twitter-specific instantiation of these descriptors, and demonstrates the usefulness of the proposed approach by applying it to a news consumption prediction task in Twitter. Importantly, this thesis does not only aim at predicting users' news consumption, but at providing interpretable insights on *what* drives users to such behavior.
5. Presents a recommender system that facilitates the decision-making process facing a news editor in composing a headline.

1.4 Approach

In the first part of the thesis, we present an extensive analysis of *current journalistic practices in Twitter*, accompanied by a survey where English-speaking journalists report on their professional usage of the platform. The outcome of this first part allows us to identify important problems and challenges faced by professional journalists who use Twitter as a tool for news distribution.

The second part of the thesis takes as input the knowledge gathered from the analysis of journalistic practices and focuses on *news providers*, their strategies for news production and dissemination on Twitter and how these differ based on the news category. Furthermore, the analyses presented in this part, unveil how different strategies for news distribution impact audience attention to news.

In the third part of this thesis, we concentrate on *news audiences*, particularly, we model people's news consumption and use this information to predict news consumption in users with a similar behavior. In addition, we demonstrate how behavioral patterns of news consumption can be used to automatically profile news audiences.

In the fourth and final part of this thesis, we present practical tools that can benefit journalists who use Twitter as a news dissemination platform.

1.5 Structure of the Thesis

Following this introductory chapter, in Chapter 2 we present background and preliminaries that will serve as a guide to contextualize the tasks addressed in this thesis. In addition, Chapter 2 presents and discusses work related to our research.

The next four chapters are organized as follows: in Chapter 3, we describe our exploration on current professional journalistic practices on Twitter. Chapter 4 presents the analyses on attention to news and its dissemination on Twitter. In Chapter 5, we model and predict news consumption for Twitter audiences. And in Chapter 6, we present practical tools derived from our analyses, oriented to assist journalists in their use of social media for news reporting.

In Chapter 7, we conclude the thesis and discuss future directions.

CHAPTER

TWO

Related Work

This chapter reviews the literature surrounding the *professional reporting of news on Twitter*. Hence, we study how journalists and newspapers use Twitter as a platform to disseminate news and the factors that affect readers' attention and engagement with that news on Twitter. We also consider whether any practical advice for journalists and news professionals can be gleaned from this research, to optimize the use of Twitter for news.

2.1 Background

2.1.1 Online Developments: News Moves Online and Social

Throughout the 1990s, most major news providers moved online establishing significant digital presences. For example, by 1996, most major American newspapers – including *U.S.A. Today*[1], *The New York Times*[2] and *The Washington Post*[3] – had established websites and were publishing digital versions of their papers. Inevitably, this led to digital-only news providers; in 2000, the *Southport Reporter*[4] was launched as the first online-only UK newspaper. Such offerings continued apace with the emergence of digital-first news; for instance, by 2006, *The Guardian* was offering a "web first"[5] service where news was first published online before it appeared in the physical newspaper. Nowadays, the online publication of news typically precedes newspaper publication

[1] http://static.usatoday.com/about/timeline/

[2] http://www.niemanlab.org/2016/01/20-years-ago-today-nytimes-com-debuted-on-line-on-the-web/

[3] https://www.washingtonpost.com/apps/g/page/national/washington-post-co-timeline/374/

[4] http://www.southportreporter.com/

[5] https://www.theguardian.com/media/2006/jun/07/theguardian.pressandpublishing

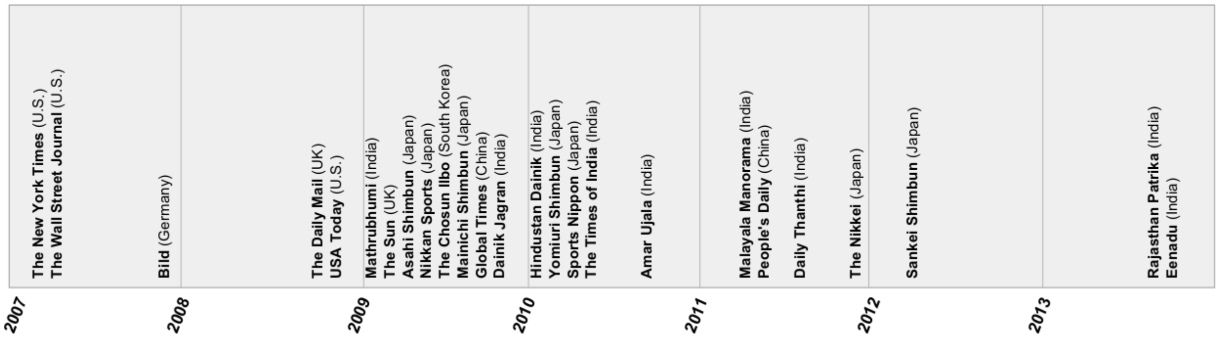


Figure 2.1: Dates on which the top-25 newspapers (by circulation) created official Twitter accounts.

thus becoming the first and, in some cases, primary source of most of the news that people read.

In the 2000s, with the emergence of social media sites – Facebook (2004), YouTube (2005), and Twitter (2006) – there was yet another turning-point in online news delivery and consumption. These sites enabled the real-time sharing of online news and gave the audience the opportunity to interact with news providers, directly. Furthermore, as these social media platforms were "always-on", they fast became key fora for collating updates, gathering comments and reporting breaking news. Indeed, in many cases, they became the news *themselves* [134].

Of all social media sites, Twitter has emerged as *the* platform for news [62]. Unlike other social media platforms (e.g., Facebook, Instagram, Youtube) where news is mainly encountered as a side-effect of other interactions, many Twitter users specifically use the platform to track news developments [134, 1, 133]. Accordingly, this microblogging platform is favored by journalists and news organizations to live-tweet news events that are being broadcast in real time or to share incremental reporting threads and other content [7].

Figure 2.1 shows the top-25 newspapers, by worldwide circulation, ordered by the dates in which they created their official Twitter accounts. *The New York Times* joined Twitter on March 2^{nd} 2007, being the first of the five most popular American newspapers to use the service, swiftly followed by *The Wall Street Journal* (March 31^{st} 2007), and the German tabloid *Bild*[6] (October 2^{nd} 2007). Most of the other newspapers in this list, created their official accounts between 2008 and 2010, though two other outlets – *Rajasthan Patrika* and *Eenadu* from India – did not officially join Twitter until late 2013.

By the late 2000s, news providers were not only reporting news on Twitter, but, with the increasing popularity of the platform, were also using it as a tool to post immedi-

[6]http://www.bild.de/

Style	Actors	Platforms
Citizen Journalism	public	web, social media
Traditional Journalism	professional journalists	web, newspaper
Computational Journalism	professional journalists, computer scientists	web, social media, newspaper
Data Journalism	professional journalists, computer scientists	web, social media, newspaper
Ambient Journalism	public, professional journalists, data scientists	web, social media, newspaper

Table 2.1: New and old types of journalism, the actors performing them and the platforms on which their news is typically delivered.

ate updates from citizens and eyewitnesses. Journalists had also started *sourcing* news from tweets, some of which were posted from anonymous or unverified Twitter accounts; this "publish first, ask questions later" strategy marked the coverage of certain events, such as the popular protests in Iran in June 2009 [160]. Of course, in time, a cottage tech-industry has grown up around verifying and curating such Twitter sources (e.g., [40]). Now, different news outlets (e.g., BBC, NYT) and different news platforms (e.g., TV, online, in print) compete using a wide variety of strategies and journalistic styles played out on social media [14].

Just over a decade after its creation, nearly all major news providers use Twitter to both advertise their news articles and promote interaction with their audience. And, while social media platforms have clearly changed journalistic practice in many positive ways, they also appear to have changed news publication in negative ways. Many worry that the *social* aspects of Twitter undermine journalistic integrity; for instance, while a highly interactive journalist benefits from more engagement with their audience they may do so at the expense of appearing less objective by virtue of these interactions [36, 93].

2.1.2 Emerging Developments: From Computational to Citizen Journalism

These major changes in the news ecosystem have created new conceptions of journalistic practice; changes that have re-defined the skill-sets of those who call themselves journalists (as in *Computational Journalism*) and, perhaps, widened the definition of "journalist" to include ordinary citizens (as in *Citizen Journalism*).

Computational Journalism was coined in 2006 as a descriptor for the then newly emerging form of journalism[7] that draws on innovations in computer science and informatics. It changes "how news stories are discovered, presented, aggregated, monetized, and archived" largely through the use of new computational technologies [30]. Figure 2.2 shows how different journalistic and news-publication processes can be enabled and augmented by information technology "all while upholding the core values of journalism such as accuracy and verifiability" [39]. For example, for the process of *information gathering*, information technology has facilitated the finding and characterization of source experts, the cross-referencing of eyewitness reports and the determination of originating source of information. As for the *communication/presentation* process, journalists are now able to make interactive models or data that inform more than a static story.

The Guardian's Datablog was introduced in 2009, as a new service that specifically gleaned news stories from data analysis.[8] Such offerings popularized the concept of *Data Journalism* as the intersection between journalism, design, computer science, and statistics [77]. Other news outlets, including *The Irish Times*, joined the movement a few years later (2015) by introducing *Irish Times Data*[9], an effort to put data journalism to the foreground.

The importance of social media, and particularly of Twitter, for news production and distribution became evident in 2010 with the emergence of *Citizen Journalism* and *Ambient Journalism*. *Citizen Journalism* refers to the power of the audience to create and disseminate news, often as part of their direct participation in target, news events.[10,11] Though the concept already existed by 2010, it reached a peak of popularity during The Arab Spring (December 17^{th} 2010 – December 2012). The coverage of the Arab Spring saw ordinary citizens shaping and spreading the news based on millions of tweets from "young, urban, relatively well-educated individuals, many of which were women" [68]. This Citizen Journalism concept has inspired the idea of *Ambient Journalism*, where Twitter is viewed as an awareness system that offers a "means to collect, communicate, share and display news and information serving diverse purposes" [62]. Table 2.1 presents these emerging types of journalistic practice comparing them in terms of the main actors involved and the platforms on which their news is typically delivered.

[7] http://www.gatech.edu/hg/item/182791
[8] https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism
[9] https://www.irishtimes.com/news/ireland/irish-news/welcome-to-irish-times-data-1.2089414
[10] http://www.hypergene.net/wemedia/weblog.php
[11] http://archive.pressthink.org/2008/07/14/a_most_useful_d.html

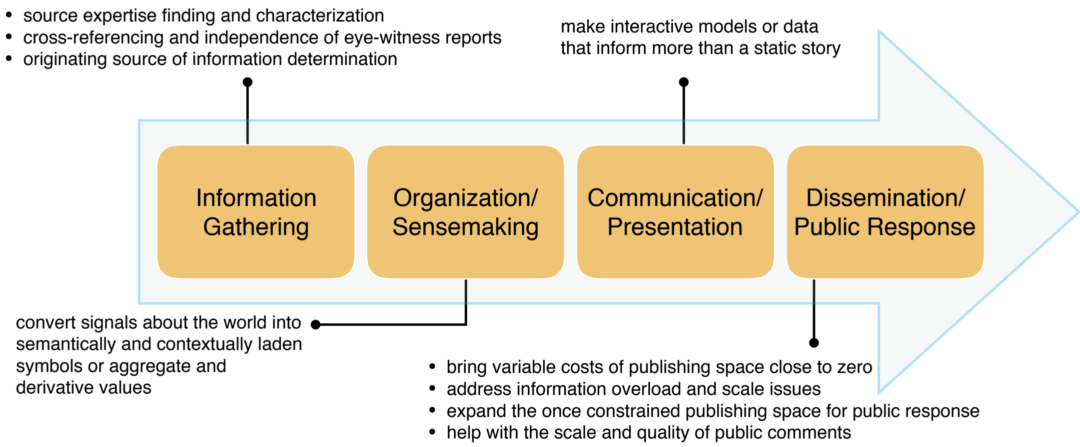


Figure 2.2: Journalistic and publishing processes impacted by computing (adapted from [39]).

2.1.3 The Good and Bad of News on Twitter

Most digital innovations impact our daily lives in good and bad ways and the use of Twitter for news is no exception. Many positive effects have arisen from Twitter's arrival in the news ecosystem; for instance, journalists can now track breaking news more efficiently and directly [142]. However, many negative effects have also arisen; for instance, arguably, Twitter has undermined some of the key tenets of traditional journalism. It is important to appreciate these wider aspects of the news ecosystem, to contextualize the technological innovations arising from the use of Twitter before moving to considering the literature in detail and what it has found.

The Good: How Twitter Enhances Journalism

The use of Twitter for news has transformed many aspects of journalistic practice in positive ways. Twitter has become an important tool for journalists by (i) supporting real-time information gathering and validation, (ii) creating a conduit to websites for news articles, (iii) providing a self-selected, target audience, (iv) engaging audiences in news commentary, and (v) enabling self-promotion and branding.

Real-time information gathering & validation. During rapidly-changing news events, Twitter can serve as a real-time monitor for journalists and first responders (e.g., during natural disasters, disease outbreaks, elections or civil disturbances). For example, maps of weather events can be built from crowdsourced information, by asking people to report how much snow they have in their backyards during a snowstorm [167]. Twitter has also been used as an early warning system for disease outbreaks, based on people reporting their symptoms (e.g., the 2011 e-coli outbreak in Germany [43]). It has also been used as a direct source of news in elections, where politicians' tweets themselves become the subject of news reports [21]. Notably, journalists also use Twitter to perform real-time news collection by questioning eyewitnesses using tweets. For example, during the Arab Spring, journalists made extensive use of Twitter to satisfy a variety of information needs: to check facts, to find information sources, and to solicit and clarify opinions [59]. For breaking news, Twitter can be seen as a distributed sensing platform, through which eyewitnesses – people *in situ* during the event – share their experiences, verify details of the event and provide personal opinions [40].

A conduit to news websites. A large number of tweets posted by news outlets and journalists provide links to a news organization's website or to other news sites.[12] Twitter can be used to direct traffic to a news-provider's home-page while making it easier for users to share one's news with followers, leading to wider dissemination (e.g., by retweeting) [128]. It has been shown that having a Twitter presence increases traffic to a news-provider's website [66]. However, this behavior is changing in several ways. Increasingly, readers are moving away from accessing news via the homepages of news providers in favor of accessing news via social media and news-aggregator sites (e.g., Facebook, Twitter, and GoogleNews) [2]; between 2013 and 2015, the number of Americans accessing news via Twitter increased from 52% to 63% [11]. People are also accessing news on Twitter by following news categories; while the *breaking news* category has been used in this way for some time, people are now following other categories such as *national government and politics*, *international affairs*, *business* and *sports* [11]. This trend has led to the category-based promotion of news articles by some news providers. Indeed, it is now common practice for news organizations to use several Twitter accounts to post tweets reflecting different news categories. As of 2011, *The Washington Post*, *The New York Times*, and *The Wall Street Journal* each had more than 70 different Twitter accounts, covering a wide range of news categories.[13] Furthermore, recent work has shown that providing links to external sites (e.g., other news outlets) to support "story-focused reading" – where users read the same story from multiple sources – results in higher user engagement with a news-provider's home site [94].

Provision of a self-selected, target audience. As a social media platform, Twitter delivers a self-selected, target audience for journalists and news organizations; an audience that is known to be interested in either them or their news or both. Recent research has shown that this audience is a younger demographic that routinely uses social sharing. The 2016 Reuters Digital News Report [2] found that younger people (ages 18-

[12] http://www.journalism.org/2011/11/14/how-mainstream-media-outlets-use-twitter/
[13] See footnote 11

34) in developed economies (i.e., the U.S., Germany, France, Japan, Ireland, UK, and Sweden) use their smartphones as the main device for accessing news. Indeed, this demographic segment tends to rely on social media sites as their main news source, checking these sites repeatedly throughout the day (as often as five times a day). Also, the "concise delivery" of news in Twitter may be an attractive feature for this audience. With limited characters per tweet, journalists provide this audience with a preferred, stripped-down summary of the news (typically, a headline or link text). Analyzing the retweets received on two *Financial Times'* corporate accounts (i.e., *@FinancialTimes* and *@FT*),[14,15] shows that the account that tweets the news with headlines alone, attracts higher levels of sharing than the one tweeting news with the headline plus commentary (see Chapter 4 for further discussion).

Audience engagement in news commentary. News distribution is no longer a one-way street where journalists and news providers push news out to passive *readers*. Twitter has made news dissemination a three-way interaction, transforming it into a participatory activity where users interact, engage, ask questions, express viewpoints, have conversations, not only with news providers but, also, with other readers [167] (i.e., active attention, see Section 3.2). Twitter now represents a medium through which journalists and audiences can *bond*. For journalists, these interactions have created a new channel for feedback on their work. For audiences, these interactions have created a new sense of being part of a dialogue about the news, possibly increasing their motivation to consume it [36, 5, 91].

Journalistic self-promotion. Journalists can also use their professional Twitter accounts to promote themselves and, even, to create a personal brand. The top-6 most active Twitter accounts, for news in Ireland, have engagement levels for individual journalists that sometimes outstrip the corporate accounts of their employers (see further discussion in Chapter 4). It has been shown that the extent to which journalists promote themselves depends on several factors, including the organization that employs them and their gender. Elite journalists (working for major news organizations) are less prone to use Twitter to create a personal image than non-elite journalists (working for "lesser" organizations); indeed, the latter were found to be more "open" to posting tweets that mixed news with opinions and personal life stories [19]. Similar behaviors have been observed in female journalists, who in contrast to their male counterparts, tend to be significantly more transparent about their activities and themselves, leading to increases in relatedness and audience engagement [87].

[14] https://www.ft.com/

[15] Both accounts tweet at approximately the same days and times

The Bad: How Twitter Undermines Journalism

The use of Twitter for news can also negatively impact the provision of news and journalism. Twitter can have negative impacts on news publication by: (i) making journalists act in more subjective ways, (ii) increasing the perception of bias in the news, (iii) facilitating the emergence of filter bubbles, and (iv) supporting the spread of fake news.

Making journalists more subjective. Traditionally, journalists are meant to be "committed observers" which means that though they are part of the community, the need to maintain a detachment from that community to maintain a perspective on events [37]. Traditional journalism attempts to draw a clear line between *objective* descriptions of events and *subjective* evaluations of these events; a distinction that was enshrined in the separation of news and op-ed sections in traditional newspapers. However, Twitter encourages journalists to act in more subjective ways undermining traditional journalistic norms of objectivity [88, 164, 58]. People use Twitter to engage with a journalist or news provider, directly; in the same way that they follow *famous people* with whom they would normally have no direct contact [146]. In this medium, social interaction and personal disclosure are the norm and one-to-one interactions are not only encouraged but expected [93, 92]. On Twitter, arguably, the journalist is actively pressured to be more subjective; they are expected to converse with their readers and express themselves more freely. Indeed, Twitter users tend to pay more attention to tweets reflecting such interactions (e.g., mentions: *@user*). Readers also prefer to follow and interact with accounts belonging to individual journalists (e.g., *@AnnaHolmes*) rather than those that are official corporate accounts (e.g., *@nytimes*), showing the personalized nature of this type of news consumption [124, 146].

The perception of bias. Twitter's subjectivization of the journalistic role has a number of unfortunate side-effects, most notably an increase in the perception of bias. With Twitter, the context for news publication has changed radically from the closely-edited, objective delivery of news to a loosely-edited, real-time personalized interaction [63]. So, just like non-journalistic tweeters, a high proportion of journalists' tweets deal with personal life stories, humor, and opinions and not the news *per se* [87, 90, 113]. Hence, they come to be perceived as "just" another member of a news-sharing community. This more informal delivery of the news, often without source accountability or editorial inspection [113], strongly suggests that "the traditional conception of the *objective* journalist does not apply on Twitter" [36]. Indeed, in [46], the authors have recently proposed quantified measures of a media outlet's political and socio-economic bias using Twitter data.

The filter bubble. Furthermore, arguably, Twitter also enables the conditions for yet another negative, social media effect on news; namely, the filter bubbles that emerge from, amongst other things, following and sharing behaviors. *Filter bubbles* (aka "echo chambers") refer to the narrowing of people's perspective on current events – created by social media features, personalization and recommendation – as they literally *only* see the news that confirms their prior beliefs [90, 118, 55, 38]. So, the technology actively restricts their exposure to a diversity of views, thus confirming their existing beliefs about current events. Recent research of Danish Facebook users has found that *sociality*, namely number of page likes, group memberships and friends, is the best significant predictor for being in a filter bubble [12]. On Twitter, as in other social media platforms, filter bubbles can emerge when users (including journalists) follow, receive and share news from a few selected sources; sources that are already aligned with their own viewpoint and interests, thus remaining unaware of information from other sources, commentaries and events, that could provide a wider and more informed perspective [38, 161]. For example, in the 2016 US Presidential election, an analysis of 1B journalistic tweets showed that the majority of journalists followed Hilary Clinton rather than Donald Trump and based their opinions on official polls, leading to a serious underestimate of support for Trump in the electorate (see The Electome Project[16]). Indeed, it has been shown that many of the news audiences in the US election stayed within their filter bubbles for much of the campaign (see MIT analysis[17]), which aligns with the idea that regular readers of partisan articles are almost exclusively exposed to only one side of the political spectrum [49].

Fake news. Twitter also, perhaps inadvertently, enables the emergence of "fake news". *Fake News* refers to fabricated articles presented as news, that convey deliberate misinformation or hoaxes, designed to either attract attention or mislead an audience [4, 157, 170]. Many major cases of fake news have emerged in recent elections with a view to influencing voters, leading to regulatory steps being taken by some governments[18]: including, the 2016 U.S. presidential election [4], the 2016 Italian Constitutional Referendum [38], the 2017 French presidential election [67] and the 2017 German federal elections [116]. While the phenomenon of fake news is not new (e.g., the "Yellow Journalism" of the Hearst era) it has gained a new life with the advent of social media platforms. Twitter has contributed to the fake news phenomenon in two main ways. First, by encouraging more personalized interaction with news providers, Twitter has increased the subjectivization of journalism, leading to the perception that many news sources are biased, blurring the lines between "fake" and "real" news. Second,

[16]http://www.electome.org/

[17]https://news.vice.com/story/journalists-and-trump-voters-live-in-separate-online-bubbles-mit-analysis-shows

[18]http://www.politico.eu/article/germany-election-campaign-fake-news-angela-merkel-trump-digital-misinformation/

the social media functionality of Twitter (e.g., following, retweeting, liking) allows the rapid spread of fake news articles before they can be debunked; as the constant repetition of the fake news item lends it a spurious truth. For example, Shao et al. [154] found evidence that social bots (software-controlled profiles or pages)[19] support the spread of fake news as they are early spreaders of claims and tend to target influential users, though they may not spread fake news any faster than human spreaders [170]. However, the ways in which rumors come to be propagated or debunked on Twitter are quite complex [170, 26, 159, 182, 181]. Early research found that Twitter does well in debunking inaccurate information and scotching rumors [26], though other findings question this conclusion [170, 159]. Recently, more temporally-focused work has found that while the overall tendency is for users to support unverified rumors in the early stages, there is a shift toward supporting true rumors and debunking false rumors as time goes on [182, 181]. Indeed, journalists appear to play a key role as power users in this debunking process [6, 158] Having said this, a recent study has found that "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information" [170]. So, while Twitter is not solely responsible for fake news (indeed, Facebook may play a larger role in the phenomenon), it has certainly contributed to the conditions that facilitate its occurrence and promotion.

2.2 The Anatomy of News and Attention to Them

In the previous section, we outlined the broad positive and negative impacts that Twitter has had on journalistic practice and people's consumption of the news. We now turn to the core literature dealing with the dynamics of professional news publication on Twitter, beginning with definitions of the *news-tweet* and the *attention* it attracts.

In this thesis, we focus on *news-tweets* (see Section 2.2.1); those tweets that are newsworthy or specifically designed to convey news items. These tweets are quite different from the millions of other tweets that discuss personal events, topics and happenings, that have nothing to do with the news (e.g., spam, birthdays, adverts and jokes). We also address how news providers use such news-tweets to gain attention for their news and on what features of these tweets impact audience attention. Broadly speaking, we use the term *attention* to cover all forms of engagement with a news-tweet whether that is simply looking at/reading the tweet (i.e., impression) or actively engaging with

[19]Notably, this emergence of social bots has made the "author profiling task" an important challenge for the community, see e.g. [138, 139]

Figure 2.3: Examples of three types of news-tweet: where a journalist or news organization refers to (a) their own news article, (b) another journalist's or news organization's article, and (c) include journalistic commentary on events/comments/opinions tweeted by relevant figures (e.g., by eyewitnesses, celebrities, or politicians) and that journalists or news organizations consider newsworthy.

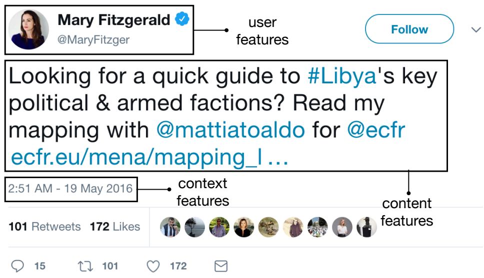


Figure 2.4: Types of tweet features: content, user and context.

it (i.e., by liking, quoting, or retweeting it). However, we also advance more precise definitions for several different types of attention (see Section 2.2.2).

2.2.1 The News-Tweet: Definition and Features

News-tweets are tweets by journalists and news organizations that, typically, (a) link to articles they have written, (b) link to articles from other sources, or (c) include journalistic commentary on events/comments/opinions tweeted by relevant figures (e.g., by eyewitnesses, celebrities, or politicians) and that journalists or news organizations consider newsworthy. Figure 2.3 shows actual examples of these three main types of news-tweet.

All of these news-tweets can be divided into three categories of features; content, user, and context features (see Figure 2.4). *Content features* refer to what is written in the tweet itself; for instance, the topic of the tweet (e.g., sports, politics, or education) and/or its inclusion of hashtags or URLs or other media (e.g., photos or videos). *User features* are those related to the user who posts the tweet: the gender, number of followers/followees, the volume and frequency of his/her tweeting, the status of the user (e.g., verified or unverified, personal or corporate account), or the news topics he/she discusses in the tweets. *Context features* relate to the conditions under which the tweet is posted, for example, time and day of publication, location of the author, or geographical focus of the tweet. These features hold for both original tweets and retweets, even though the content-part of the latter may itself contain an embedded tweet. We have found this tripartite division of the features of news-tweet to be a good organizer for the literature in the field (see Section 2.3).

2.2.2 Attention to News-Tweets: Definition and Features

Every news-tweet competes for the limited attention of users depending on where it falls in their respective timelines [25, 177]. When a news-tweet is posted, assuming it is seen, it may be attended to in several different ways. The attention to it can be a *passive attention*, where it is simply scanned or read by a user, or an *active attention*, where it is actively engaged with by users, by liking, retweeting or quoting. Indeed, very few tweets attract active attention; attention to news-tweets, as measured by number of retweets has been found to follow an exponential distribution, with a few tweets receiving a lot of attention and most tweets receiving no attention in a long tail. Furthermore, these two classes of attention can be further sub-divided into *primary attention* and *secondary attention*. Consider each type of attention, in turn.

Passive attention can be operationalized in terms of *impressions*, based on a simple count of the times a news-tweet is shown to a user[20]. Impressions indicate that the tweet has appeared in a user's timeline or in search results without necessarily attracting further active engagement (e.g., likes or retweets). In analyzing impressions, the amount of attention that a reader has given to a tweet is unclear; that is, it is not known whether it was just visually scanned as a user scrolls through their tweets, or read quickly or carefully considered. As such, passive attention does not tell us much about a user's response to a news-tweet. Notably, passive attention is the *modus operandi* for most Twitter users.[21] Although reading a news-tweet is indeed a signal of attention, for wider dissemination, active attention is much more desirable, as it broadcasts the news to other Twitter users [143].

Active attention refers to the explicit engagement of a user with a news-tweet (by liking, retweeting or quoting); it is much more informative about the user and important to dissemination. This active attention also exposes the tweet to those who follow that user; thus, increasing the size of the audience for the news item and, possibly, attracting further attention to the post, either in the form of more retweets and likes, or even in new followers for the tweet's author [179]. Users may also actively react to news-tweets by engaging in *conversations* with journalists by either *mentioning* journalists in their tweets or *replying* to journalists' posts. In many cases, journalists themselves start these conversations, leading to more active attention, especially from users who favor *participatory journalism* (i.e., journalistic environments as "social systems" in which audiences participate actively [18]). All of these engagement actions potentially promote the popularity of a given news-tweet and, hence, its visibility to other users,

[20] https://support.twitter.com/articles/20171990
[21] http://guardianlv.com/2014/04/twitter-users-are-not-tweeting/

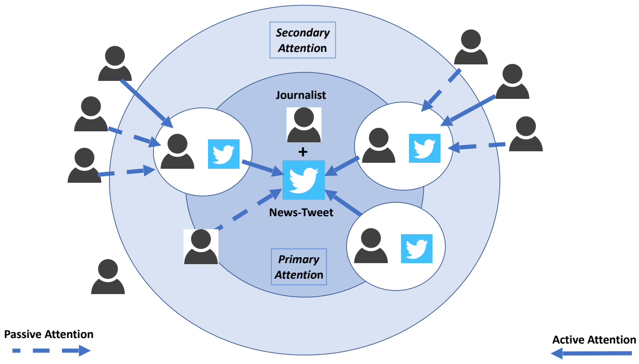


Figure 2.5: The original news-tweet produced by a journalist receives primary attention from several users, some interacting passively (e.g., reading) others actively (e.g., re-tweeting, shown as white circle with new tweet). These active attention interactions then enter the sphere of secondary attention where users may, in turn, passively or actively attend to them.

increasing the likelihood of receiving more attention from a wider audience.

Primary attention refers to the passive or active attention of a user to an original posting of a news-tweet. It concerns the user's direct evaluation and potentially active response to the news-tweet in its original form as posted by the journalist (see Figure 2.5).

Secondary attention refers to the passive or active attention of a user to a tweet referring to an original news-tweet; that is, the reading or engagement with a tweet that has been liked, retweeted or quoted by any user, other than the original tweet's author. Secondary attention, is bound in some ways, to be different from primary attention as the original news-tweet is now being presented in a new context (see Figure 2.5). For example, if *@nytimes* posts a news-tweet about an article on the US Presidential Election and Hilary Clinton retweets it with some comment, known in Twitter as a *quote*, then the original tweet is framed in a new context, a context that would be very different to one in which Donald Trump quotes it with a different comment. Understanding the dynamics of primary attention to news-tweets is likely to be significantly easier than understanding the dynamics of secondary attention; though, it should be noted that, an understanding of the latter is critical to explaining dissemination and diffusion.

2.3 Features of News-Tweets: Content, User and Context

In the previous section, we clarified the definition of a news-tweet and described how it can be cast as a collection of content, user and context features; we also considered some key definitional distinctions in the notion of attention (e.g., passive versus active, primary versus secondary). In the current section, we begin to consider how the literature might be organized using these feature-categories of the news-tweet.

There is now a substantial literature on the dynamics of professional news publication on Twitter. However, it is hard to get a coherent picture of what has been discovered, because this literature addresses a babel of different tasks: from predicting popularity [177, 71], to tracking the news life-cycle [25] and topic coverage [10], to identifying gender imbalances in journalistic environments [104, 103], and the role of journalists and news audiences in social media [41]. Hence, there is a pressing need to put some structure on the literature to understand its findings and to identify clear recommendations for journalistic practice. Our approach to putting some shape on the findings from this work, rests on two steps: (i) focusing on "attention" to news-tweets as a fundamental phenomenon of importance (acknowledging the different sub-categories of attention, see previous section), and (ii) partitioning the literature based on the different feature-categories of news-tweets that impact attention in different ways (namely, content, user and context features).

In an environment where new posts and updates are constantly emerging, news providers have a narrow time-window in which to attract attention to their news-tweets. The attraction of attention to news on Twitter is a complex process in which a news-tweet's content, a reader's knowledge of the tweet's author and the overall context act together to influence people's behavior. For any given reader, the deployment of attention will depend on filtering out some stimuli, balancing other stimuli and, perhaps, assessing their emotional significance [140]. However, ultimately, attention to a news-tweet must depend on various facets of the tweet itself. These facets can be subdivided into three categories of features: *content features* based on the news-tweet's content, *user features* based on the news-tweet's author and *context features* that capture the context surrounding the posting of that news-tweet (see Figure 2.4). These classes of features have the potential to impact (i) primary or secondary attention (though, the literature primarily addresses the former) and (ii) passive or active attention (though, the literature has primarily addressed the latter because it provides more visible output measures).

Table 2.2 summarizes these three main categories of features, lists the particular features within each category and references the literature that addresses different aspects of the feature in question. It is probably the case that this list of features does not exhaust all the relevant factors affecting attention to news-tweets, but it does present those that have been identified, to date.

In the following sections, we survey the specific studies that have been carried out on how these features impact attention to news-tweets, before considering the literature that addresses interactions between these features.

2.4 Content Features of News-Tweets

As we have seen, there are there main categories of features that can be explored in understanding attention to news-tweets – content, user and context features – into which we can map the existing literature (see Section 4). Content features refer to what is written in the news-tweet itself; for instance, the topic of the news-tweet (e.g., sports, politics, or education), or, indeed, its use of hashtags or URLs, or embedded media (e.g., photos or videos). There is an extensive body of research examining a wide range of specific content features (see Table 2.2). We consider each of these features, in turn, summarizing their main findings for attention to news-tweets.

2.4.1 Headlines and What They Convey

In the age of social media, headlines have become very important, as they may be the only visible artifact that gives readers access to a news provider's articles; that is, increasingly readers *only* encounter news items in social media (indeed, in many cases it may be the only aspect of the article read [51]). As news readers on Twitter are exposed to a continuous stream of attention-demanding headlines, from which they try to find the most relevant ones [65], these headlines are increasingly *tailored* to capture users' attention [28, 81]. Although news providers use different strategies when posting news-tweets, the *story headline + link* format remains the most commonly-used strategy [93, 88]. This way of presenting headlines is, presumably, designed to invite the reader to click on the link, to read the article and/or to like or to share it. In many respects, using headlines in this way in news-tweets leverages tactics that have a long history in newspapers (e.g., using headlines with particular styles and content, such as social deviance). However, news providers are increasingly adopting more sophisticated strategies for headline use; for instance, *The Washington Post* now uses several different headlines for the same story, which are A-B tested online to be personalized for different segments of their readership[22]. Furthermore, several research efforts concentrate on developing tools and models to suggest "good headlines" for journalists [81, 166].

[22]https://agency.reuters.com/en/insights/articles/articles-archive/how-washington-post-data-driven-product-development-engages-audiences.html

Type	Feature	Related research
Content	Headlines	Social deviance and sensationalism [41, 80], news worthiness, values and style [135, 136]
	News Category & Topic	Sports journalism [47], breaking news [23, 169, 127], political journalism [130], financial news [44], all news categories [124, 125], popularity and overlap [10] eventful vs. non-eventful topics [78], real world events [71], climate change [120], topic coverage [150], disaster events in rural contexts [33], riots [169], personal, informal or formal content [24], serendipity [105], soft vs. hard news coverage [22], broadcast vs. targeted news [9]
	Tweet Elements	Hashtags [71, 156, 124, 10], mentions and replies [125, 124, 10, 22], URLs and propagated web-links [143, 95, 51]
	Textual Elements	Standout phrases [34], commentaries with personal interactions [24], quotes & memes for news dissemination [96, 162, 22], quotes with named entities and people [10, 22]
User	Account Owner	Corporate vs. individual journalist account [23, 9, 70, 100, 124, 125], trust and tweet lifespan [13], elite vs. non-elite news organizations [9], broadcast vs. targeted communication [9, 23], event type [35], mentioning patterns [70]
	Branding & Reputation	Meformer vs. informer [71], personal brand creation [113, 113], gatekeeping role [171, 89], impartiality and nonpartisanship [89], emotions and personal voice [10], perception of journalists [93]
	Followers & followees	Social media adoption, online readership and the size of newspapers' social media network [66], social network structure [71], influence and number of followers [95]
	Activity volume	Daily tweets [95, 9]
	Journalistic Process	Accountability [113, 89], elite vs. non-elite sources [9], source review [40], perception of journalists [93]
	Gender	Gender, sharing and impressions [104], gender and journalistic transparency [87], follower bias [103]
	Nationality & Ethnicity	Arab and British journalists [9], Australian journalists [23], Irish and UK Journalists [124, 125], white males [103]
Context	Temporal Aspects	Time & date of publication, news crowds [95], visitation patterns [25], breaking news on Twitter [70]
	Locational Aspects	Geographical aspects of users and real world events [71], news recommendation [69], reporting with a crowd [33], events of local-interest [114]
	Visibility	Visibility and divided attention [65]

Table 2.2: Features that impact the tweeting of news studied in the literature.

Social deviance. Many of the effects of headlines on attention in Twitter rely on well-established practice in writing headlines from traditional news, such as emphasizing social deviance. Banner headlines on the front pages of newspapers were an innovation of the late 1800s; before then, front pages were filled with adverts, often, in very small type [175]. Traditionally, news headlines were fashioned to draw readers in, encouraging them to either buy the newspaper and/or to read the target article. So, there is a time-honored tradition of reporting social deviance in news headlines (e.g., the infamous tabloid headline "Freddy Starr Ate My Hamster" published by The Sun in 1986[23]). Socially deviant news involves items about violations to social or legal norms (e.g., robbery or homicide[24]). In Twitter, the effects of socially-deviant headlines on audience attention have been analyzed with a view to predicting shareability. Diakopoulos and Zubiaga [41] analyzed 107,066 tweets, using the *<story headline>* + *<link>* format, posted by the eight top U.S. newspapers (by circulation) with more than 100K Twitter followers between November 2011 and October 2012. Using Amazon Mechanical Turk (AMT)[25] the tweets were labeled as "deviant" or "non-deviant" (with deviant meaning that the headline in the tweet refers to an event that breaks social norms). This study found that newspapers have a preference for tweeting deviant headlines, especially tabloid newspapers (e.g., the *NYPost* and *NY Daily News*). Furthermore, it was also found that news audiences on Twitter pay more attention to deviant headlines than to non-deviant ones, and are more likely to retweet them. More generally, the use of social deviance is another manifestation of the use of sensationalism, which is widely used across all news categories (i.e., including "hard news" categories such as government affairs and science) to attract online attention, even though readers do not necessarily respond to such sensational treatments [80, 52].

Newsworthiness versus style. Traditional news media has always made judgments about the newsworthiness of events with a view to publishing or prioritizing stories. Accordingly, headlines have been fashioned to convey this newsworthiness, reflecting so-called *news values* (i.e., prominence, sentiment, magnitude, proximity, surprise, and uniqueness). Indeed, Trilling et al. [168] have argued that newsworthiness is a good predictor of "shareworthiness". Headlines have also been written with a distinct *style* represented by brevity, simplicity, and unambiguity. These features of headlines have been analyzed to determine their effects on attention [135, 136, 51, 15]. In [135], Piotrkowicz et al. used headlines alone to predict the popularity of news articles. They used a collection of news headlines from *The Guardian* and *The New York Times* and

[23] https://www.thesun.co.uk/archives/tv/895099/freddie-starr-ate-my-hamper/

[24] http://www.aejmc.org/home/wp-content/uploads/2012/09/Journalism-Quarterly-1991-Shoemaker-781-951.pdf

[25] https://www.mturk.com

their corresponding popularity scores in Twitter (tweets and retweets) and Facebook (likes and shares). The authors partitioned the features of the headlines into news-value and style features and using regression models found that style features on their own, achieve better performance than news values, suggesting that the style of the headline, independently of the article content, plays a greater role in capturing the attention of social media users (see also [136]). However, this does not mean that the content of the article is irrelevant; as we shall see in the next section, it appears that the content features of a tweeted article also significantly affect the extent to which it garners attention.

2.4.2 News Categories and Topics: Reader and Journalistic Aspects

Different readers have different interests, preferring some news stories over others; indeed, such interests may manifest themselves in broad preferences for one category of news over another (e.g., sports over politics). These preferences seen in traditional news consumption, carry forward into the domain of tweeted news and are reflected in new behaviors appropriate to this ecosystem. Similarly, journalists working in diverse areas of news may manifest different norms in their professional use of social media; that is, sports journalists may tweet their news and interact with their readership in different ways than, say, business journalists. Overall, these behaviors act to make news content and news categories important factors impacting attention to news.

Readers' engagement with news categories & topics. In Twitter, readers' preferences for a given news category are reflected in their engagement with tweets related to topics covering similar events. By looking at the content of a user's past tweets and at those of their followees, it is possible to predict future engagement on sports, politics, or business topics. For instance, a user who normally tweets about a geographical area, is also likely to engage with news covering sports teams in that area [71]. This influence of news category has been shown for several different types of events, including real world events [71], climate change [150], disaster events [33], and riots [169].

These news category effects are, perhaps, most clearly manifested in the "breaking news" category; where tweets about sudden or developing events receive considerable attention.[26] In situations where real-time updates on events are critically important (e.g., earthquakes, floods, or sport events) news-tweets that accurately convey breaking news tend to gain traction on Twitter before other social media (e.g., Facebook or

[26]https://www.forbes.com/sites/quora/2017/01/10/why-twitter-is-still-the-best-place-for-breaking-news-despite-its-many-challenges

Google Plus [127]). Indeed, Keane et al. [78] have shown, by means of daily LDA topic modeling of of 350K Reuters news articles and 2 billion tweets, that certain types of "eventy' topics (such as Robin Williams' death) receive consistently higher levels of attention in Twitter. Notably, it has been shown that, compared to news wires, Twitter exhibits a larger number of hyper-local or real-life events news-tweets that attract the attention from audiences in search of increased event coverage [150, 132]. Furthermore, these news audiences tend to engage more with topics presented informally and that tend to be more personal, reflecting a more reciprocal relation between readers and journalists [24].

In other work, using data collected from Feedzilla[27] covering 31 news categories (e.g., world news, business, sports and art), Bandari et al. [10] found that technology-related and health-related news stood out. Technology-related news receives a higher average number of tweets than other news categories, while health-related news appears to have a loyal, niche audience that is ready to tweet and share its links, even if the number of these links is lower than in other categories.

News agencies and their readers may often differ in what they consider relevant or interesting [180]; there may be a gap between what newspapers publish online and what people share in social media. Through their analysis of news and tweets on climate change, Olteanu et al. [120] showed that ordinary events receive more coverage in Twitter than in mainstream media, and that individual actions from non-elite people "neither rich, nor powerful, nor famous" generate peaks of attention comparable to those involving important organizations or even governments.

Finally, it should be said that sometimes the content of a news-tweet can capture attention, serendipitously, just because it is surprising in some way (and not necessarily in a socially-deviant way). As readers search through news-tweets apparently non-relevant, surprising content can attract users' interest and capture their attention [105].

Journalistic practice within news categories. Just as readers tend to respond differentially to different categories of news, so too do journalists; specifically, journalistic practice in the use of Twitter varies over content categories.

In *breaking news*, Bruns [23] found that individual journalists covering a story on Twitter, adopted a more collaborative dynamic in which several parties worked together to find and share the latest updates of the developing events.

Along the same lines, Vis [169] studied two journalists – Paul Lewis (*The Guardian*) and Ravi Somaiya (*The New York Times*) who actively covered breaking news about the UK

[27] www.feedzilla.com

summer riots of 2011 – finding that a type of tweeting activity that is similar to live blogging, gained them significant national and international attention on Twitter.

In *sports news*, using a combination of in-depth interviews and content analysis of print and online articles, English [47] examined when and why sports journalists adopt Twitter. This work found that promoting articles and engaging readers are reasons why some sports journalists start their accounts in the first place. Indeed, several journalists reported that having a large number of followers led to more reader interaction that was helpful in sourcing ideas and opinions before deciding if a piece is suitable for publication.

In a similar way, *financial news* on Twitter has been increasingly used as a direct line of communication with other journalists, governments, banks, politicians, and investors sharing interests, analyses and priorities; thus, increasing attention from readers, who consider Twitter a newswire for financial news [44]. Although, it also seems that different news providers act somewhat differently in certain news categories. In countries such as the Netherlands and the United Kingdom, elite newspapers turn to Twitter to source "hard news" (e.g., politics, business and economy, and health care), as well as "soft news" (e.g., human-interest and sports topics), to a lesser degree. For non-elite counterparts, the opposite occurs, as they pay more attention to soft news [22]. In contrast, the tweeting of "political news" has been changed less as both journalists' awareness and organizational norms bound the changes so they do not cause a major shift in traditional journalistic practices [130]. Interestingly, in general, it has been shown in a recent qualitative analysis of Norwegian news outlets, that soft news items tend to be shared more than hard news items [76].

2.4.3 Tweet Elements: Hashtags, Mentions, Replies & URLs

We have seen that the content of the tweet clearly impacts attention, just as in traditional newspapers. However, there are also specific elements in tweets that impact attention to news-tweets, promoting diffusion; namely, hashtags, mentions, and replies, as well as the URLs that users can click to get more details on news articles of interest. These elements are regularly used by many journalists to gain greater attention for their news-tweets.

Hashtags. User-defined tags, so-called hashtags – such as *#CharlieHebdo* – are added to news-tweets to make them easier for readers to find. In theory, if a hashtag clearly discriminates one news-item from others, then readers are more likely to encounter a news-tweet containing it, creating the opportunity for more engagement and subsequent diffusion. In general, the use of hashtags in all tweets is relatively low (typically, $\sim 20\%$ of tweets). However, in breaking news-events, tweets use hashtags more often ($\sim 42\%$ of tweets) [71]. As such, it is clear that, under certain conditions, hashtags are used by readers and journalists to promote engagement with stories. While, in general, there is evidence that hashtags promote attention and engagement, we have found few tests of their specific impact on news-tweets. However, Shi et al. [156] have shown that tweets with no hashtags or with the hashtag *#news*, attract equivalent numbers of impressions, engagement and URL clicks; while tweets with tailored hashtags generate more impressions, engagement and clicks, and a higher engagement rate (0.86% vs. 0.47%). There is also a noted tendency for individual journalistic accounts to use hashtags more in news-tweets than corporate accounts; often leading to increased attention and engagement (see Chapter 4 for further discussion).

Mentions & replies. Like hashtags, mentions and replies are other content elements of a tweet that are used to make news-tweets easier for "selected" readers to find. Mentions/replies direct a tweet to specific people and/or create an awareness of their involvement in an event (e.g., using a journalist's user-name *@MaryFitzger*). Mentions and replies look very similar but they differ in their physical location within the tweet and, possibly, in the intent behind them.

Mentions can be inserted anywhere in the tweet and do not necessarily reflect that a particular person is being referenced, but rather that the tweet refers to a topic that involves this person in some way.

Replies are always in the beginning of the tweet and reflect an intent to engage the referenced individual in conversation. Mentions and replies are similar in that they are both explicit signals by a tweeter to engage journalists (or other readers).

Intuitively, tweets that include one or more mentions of known people, places, or organizations, attract more attention from news readers [22]. However, the relationship between the presence of known entities and tweet popularity is not easily evidenced; researchers claim that for engagement-prediction tasks it is better to use this feature (i.e., mentions of entities) in combination with other features such as article source, news category, and subjectivity of language [10]. This multivariate aspect of popularity and sharing has been a hallmark of more recent predictive models (see [79, 98]).

URLs. The final element used in tweets that impacts attention is the use of URLs; as clickable entities they provide a method for sharing access to the actual news article published by a news organization or journalist. Click data associated with these URLs can be used to understand user interests, to measure influence of the journalist/news-outlet involved, or to predict the number of clicks similar URLs (aka similar articles) will obtain in the future [143]. Furthermore, studying audience attention to URLs on Twitter, reveals that different news events attract groups of people with different characteristics that can be exploited to identify related news events [95].

Click-Per-Follower (CPF; i.e. proportion of followers that click on a URL/link in a tweet sent by a user) has been used as a measure to distinguish differences in engagement for different Twitter accounts. CPF shows that URLs widely shared by official Twitter accounts, although receiving considerable attention in terms of mentions and retweets, are less appealing to users as far as clicks are concerned – i.e., people do not necessarily click the URL to read the article on the website, despite having shared or mentioned the tweet [51].

2.4.4 Textual Elements: Stand-out Phrases, Commentaries, Quotes & Memes

Apart from Twitter-specific elements, there are other identifiable textual-elements in news-tweets that can be separately assessed for their impact on attention: namely, stand-out phrases, commentaries, quotes and memes.

Standout phrases. Some phrases can act to set one news-tweet apart from others, affecting their *memorability*. Danescu-Niculescu-Mizil et al. [34] found that memorable phrases are (i) lexically distinctive (i.e., they include words that are less common but use a common part-of-speech structure), and (ii) general (i.e., they can be easily applied in other contexts). These phrasal effects are akin to ones found in the psychological studies of text processing, where quite superficial characteristics of a text have been shown to influence its understanding (e.g., [32]).

Commentaries. As a social-media platform, Twitter bridges the communication gap between journalists and news readers, and a growing number of tweets contain commentaries and interactions [93, 88]. However, the use of commentaries is not uniform across all journalistic accounts. Focusing on two daily regional newspapers in the UK, Canter [24] analyzed journalist's tweets reflecting commentaries or interactions with the readers, finding that journalists gradually built relationships by engaging in conversations with their audience. However, some news organizations do not promote these types of exchange, and differences exist between individual and corporate Twitter accounts; in general, corporate accounts manifest less interaction than individual journalists accounts.

Quotes & memes. For news to spread, its content should prove accessible, timely and relatable to news readers. Quotes may be excerpts from political speeches, declarations, or interviews, that can be inserted into tweets to give users a snippet of news. When parts of those quotes are modified they may become memes, spreading further by virtue of their shareability. In some cases, professional news providers use memes (or mutations of memes) to attract attention by including them in their news-tweets [22, 96]. Quotes or memes might also be specifically used in newspapers when reporting on news events that break on Twitter (e.g., declarations from celebrities, athletes, the public and politicians; see [162]).

2.5 User Features of News-Tweets

As outlined earlier, there are three main categories of features that can be explored in understanding attention to news-tweets – content, user and context features – into which we can map the existing literature (see Section 4). Beyond the features that make up a news-tweet's content, there are also user features that impact attention and engagement, those that depend on *who* posts the tweet; so-called *user features*. User features' center on the owner of the Twitter account (i.e., whether it is owned by a corporate entity or individual journalist), the self-brand of the individual journalist, the tweeting style of the account and a variety of demographic factors (e.g., gender, nationality, ethnicity).

In considering user features, a fundamental division arises between corporate and individual journalistic accounts; indeed, the news organization behind a specific corporate account is important too (see Section 2.5.1). With respect to individual journalistic accounts, there are a host of user features to do with how journalists project themselves to readers and how they create a "brand" that attracts attention to their news-tweets (see Section 2.5.2). Similarly, the tweeting style adopted in an account may also impact attention; for instance, in interacting with followers/followees, in the volume of news-tweets generated and in the way tweets are posted (see Section 2.5.3). Finally, and more generally, some user features reflect demographic characteristics – such as gender, nationality and ethnicity – that have been shown to impact attention (see Section 2.5.4).

2.5.1 Account Ownership: Corporate Versus Individual Accounts

Corporate Twitter accounts have distinct user-related features that distinguish them from individual journalist accounts, leading to different patterns of attention and engagement that may sometimes depend on the organization behind the account.

Corporate accounts present the corporate-brand of the news organization as a whole (e.g., *@IrishTimes*, or *@nytimes*), whereas *individual journalist accounts* present the personal-brand of a specific journalist (e.g., *@jstallman*,*@maryfitz*). On average, corporate accounts have 39% more followers than individual accounts and also tend to be added to more lists[28]; presumably, reflecting a belief amongst users that following a corporate account will alert them to all the news published by that source [23, 9]. In contrast, individual accounts tend to be used to reach out to a specific journalist, to make personal contact with them, to converse with them and to participate directly in the news cycle [23, 9].

Indeed, the way corporate and individual accounts attract attention can change depending on where a story is in the news cycle [35]. In breaking news, before a given story is confirmed by the majority of media outlets, people attend more to individual journalists than to corporate accounts (e.g., especially when the journalist may have been first to tweet this news item); but as soon as a news item is officially verified, people engage more with corporate accounts, possibly seeking to add credibility to their own tweets, when spreading the news to their network of followers [70].

News organizations with a strong Twitter presence appear to have an authority, in themselves, that people rely on [100]. It has been shown that the more social recommendations people get for a particular news organization, the more likely they are to trust and develop an interest in its news [91]. For certain news organizations – such as the *BBC*, *Mashable*, and *The New York Times* – the URLs they tweet have a longer life, reflecting a greater engagement with their news-tweets; indeed, these tweets can spread for several hours (or even days) after the URLs first appear on Twitter. The tweeted URLs of other organizations – such as *Bloomberg*, *Wired*, and *Forbes* – often have shorter lifespans, attracting lower levels of engagement [13].

Furthermore, some journalists leverage people's recognition of the reputation of news organizations by referring to them in their tweets [169], perhaps adding greater credibility while exposing their posts to the larger audiences attracted to these organizations. Indeed, the tweeting behavior of journalists is, to some extent, shaped by the organization that employs them; journalists working for *elite* news organizations tend to be more careful and adhere more to journalistic norms than their colleagues working for *non-elite* organizations [9]. In fact, elite news organizations, such as *The New York Times*, periodically update policies for their journalists, to regulate the subjectivity of their Twitter interactions.[29]

[28] A list is a curated group of Twitter accounts.
[29] https://www.nytimes.com/2017/10/13/reader-center/social-media-guidelines.html

2.5.2 Branding & Reputational Features

We have already seen that Twitter, arguably, undermines journalism by making it appear more subjective (see Section 2.1.3). This growing subjectivization of the news has emerged as journalists present themselves in ways that are known to attract more attention and engagement. This behavior can be viewed as a form of *self-branding*, where journalists are projecting an image of themselves to their readers, one that will attract greater attention to their news-tweets. For example, journalists that interact with readers and that make self-disclosures attract more attention and engagement, especially from younger audiences [93]. Having said this, some of the traditional virtues of journalism still matter. A journalist's expertise and interest in the topic area is a good predictor of readers' positive perceptions of them and of engagement with their news-tweets, especially if the reader and journalist share common beliefs and interests [93].

Three specific tactics have been shown to be important in self-branding; self-reference, increased subjectivity and individualized curation. *Self-reference* in a tweet is one method used by journalists to promote themselves. One distinct category of such tweets has been noted: *meformer tweets* [71]. Meformer tweets are those that include users' comments, opinions, and thoughts that are highly personal [113]. Obviously, journalists using meformer tweets are taking a much more *subjective* approach to the news. However, the attentional and engagement benefits for a journalist of being highly subjective remain to be proven. Indeed, the measured subjectivity of individual journalists has not been verified as a good predictor of engagement [10]. Finally, the *gatekeeping* actions of the individual journalist may also help to shape their brand. Traditionally, news organizations have acted as gatekeepers, performing a curation process for readers, as editors select stories and promote certain views of events; though, digital gatekeeping is now performed by individuals (journalists and non-journalists) and, even, platforms [171]. In Twitter, individual journalists perform a similar gatekeeping, curation role that, again, may attract greater attention to their news-tweets [103, 89]. For example, readers may follow a given journalist because they value their *expert* commentary on events (e.g., following a Nobel prize economist's commentary on the U.S. economy). Finally, journalists may take a more objective approach to the news, using so-called *informer tweets*, that include actual information about events of public interest [71].

2.5.3 User Features Conveyed By Tweeting Style

Apart from directly projecting their Twitter personality, journalists may also project themselves by the way they tweet, their *tweeting style*. This set of user features is manifested in (i) how journalists manage their followers and followees, (ii) the frequency and volume of their tweeting behavior, and in (iii) how much they reveal the journalistic process in their tweets.

Followers and followees. Many journalistic accounts build up significant follower and followee lists, presumably based on the simple assumption that the larger these groups, the higher the level of attention and engagement; that there will be a *rich get richer* effect as more followers and followees will see a news-tweet and respond to it [95]. However, the dynamics of following are not this simple; that is, higher levels of engagement do not necessarily occur from larger groups of followers/followees, one must have followers that become *active* spreaders of one's tweets [143]. Indeed, there is evidence to suggest that one can sometimes have too many followers, leading to a decrease of awareness, engagement, and participation due to several factors, including cognitive overload and higher noise (e.g., followers feel that interactions become less personal) which can demotivate the audience from spreading a news-tweet [71]. There are also significant biases around gender in the following behaviours that male and female journalists attract [103].

Volume of tweeting. The volume of tweets delivered from a journalistic account (e.g., number of tweets published in a given time period) clearly has some impact on the levels of attention and engagement that arise, though the causal relationship is not straightforward. Corporate accounts tweet at very high rates by comparison to individual accounts [9]; but, as we have seen, these accounts play a special role in spreading news. It is important for journalists to find a balanced volume of tweeting. Indeed, tweeting too often may decrease the likelihood of a news-tweet being seen, as these will remain at the top of followers' timelines for less time [65]. On the other hand, as we will discuss in Chapter 4, being an active tweeter raises awareness among a journalist's audience, which may lead to more engagement.

There also appears to be culture differences in tweeting styles by journalists in different countries, presumably, based on some assessment of what works. For example, Arab journalists tend to *broadcast* their tweets (i.e., high tweeting volume and frequency, use of hashtags and links), and this behavior seems to attract a favorable response from their audience. In contrast, British journalists are more likely to *target* their tweets, directly engaging with their readers, to attract significant audience attention [9].

Revealing the journalistic process. Journalists that adopt a tweeting style that reveals aspects of the journalistic process impact attention to their news-tweets; this is often achieved by including eyewitness evidence in news-tweets (e.g., pictures and videos) or by revealing discussions with other journalists. This style of tweeting has been more apparent in news-tweets covering fast-moving breaking news; for example, in the Tunisian and Egyptian uprisings, journalists included user-generated multimedia or links to images in their tweets, to help readers follow developing events [64]. The posting of tweets that are viewed as transparent and informative (e.g., with verified information) can improve a journalist's standing and attract more attention to their work. This transparency is often conveyed by including "job talk" in tweets (i.e., professional, job-related jargon), by discussions between journalists/public-figures, and by providing links to relevant external resources [113]. In certain contexts – such as events involving medical emergencies, disasters, fatalities, or floods – people are more prone to engage with tweets when these are easier to find and the information comes from a trusted source, which does not necessarily have to be the author of the original tweet [121]. Indeed, increasingly, journalistic tools are being developed to support the verification of sources based on a tweets' geographical location, time of creation, and a user's social network [40].

2.5.4 Features Reflecting Demographics: Gender, Nationality & Ethnicity

Apart from the individual features of a journalist that impact attention and engagement, research has shown that certain demographic characteristics of groups of journalists are also important; notably, gender, nationality and ethnicity.

Gender. News articles by female and male journalists posted on Twitter receive different responses from audiences, responses that can vary across news category. With respect to this feature, it is important to appreciate that gender representation in different categories of news, may not be balanced from the outset; that is, many major news organizations, to begin with, have more male than female writers (e.g., from 143K opinion articles published by *The Guardian* between July 2011 and June 2012, only 21% were written by women). But, even when female journalists get to tweet their news, they often receive lower levels of attention in social media (in Facebook and Google Plus, as well as Twitter). For instance, in the *Daily Mail*, an article written by a female journalist in the sports section receives 33% of the impressions received by an article written by a male journalist [104]. Furthermore, even though female journalists are more open and transparent in their news-tweets than their male counterparts [87], characteristics that are valued by audiences, these audiences still engage more with male-authored tweets. Female journalists also suffer from follower bias, though systems to remediate this have been proposed [103]. Indeed, in general, news items tend to be shared more by males (that are white) [141].

Nationality & ethnicity. Other demographic factors have also been identified as being important in news tweeting and re-tweeting, notably nationality and ethnicity. There are differences in journalistic-tweeting across geographical boundaries, perhaps reflecting different cultural norms in different parts of the world. We have already seen that Arab journalists tend to broadcast tweets and their audience seems to respond positively to that style of tweeting; whereas English journalists tend to target tweets, which gets positive attention from their audience [9]. In Australia, journalists tend to be more gregarious in their tweeting, being approachable, willing to interact, and to create ties to other people of public interest [23]. In Ireland, journalists tend to interact with users as well, although this behavior varies depending on the news category, as is the public engagement with their tweets (see Chapter 4 for further discussion). Overall, in promoting their news, journalists in different countries differ in their use of sensationalism on Twitter [80, 52]. In addition, in general, it has recently been shown that white, male users tend to be more active in terms of sharing news, suggesting that there may be a propagation bias to news dealing with the interests of this demographic group [141].

2.6 Context Features of News-Tweets

We have argued that the three main categories of features that can be explored in understanding attention to news-tweets are content, user and context features (see Section 2.3). We have seen how the contents of a news-tweet and the features of the journalistic user posting that tweet can impact attention and engagement. Beyond these types of features, the final category captures the conditions or *context* in which that tweet is posted. Context features deal with the factors surrounding the posted news-tweet; namely, the temporal aspects of the post, locational aspects of it, and the visibility of the post in the tweet stream. The same piece of tweeted news sent by a given journalist may be responded to very differently depending on whether it was read in the morning or evening by an audience; for example, in the morning it could be a "breaking-news first", whereas by the afternoon it could be "old news".

2.6.1 Temporal Aspects of News Tweeting

The way an audience responds to different categories of news can change at different times of the day or, indeed, at different times of the week. For instance, breaking news needs to be timely; posting a news-tweet a few minutes later than competitors may result in the loss of an audience for that news item [25]. In the case of Osama Bin Laden's death, three journalists (i.e., *@jacksonjk*, *@keithurban* and *@brianstelter*), received significant attention to their tweets when they reported the *possible* death of Osama; indeed, their breaking tweet convinced the audience of this event, before any official announcement was aired [70]. In contrast, posting a news-tweet with a link to a long-read, in-depth article should probably be targeted at times of the day when readers are known to have the time to read such articles (e.g., during commuting times or at the weekends).

Audience reactions to news-tweets at different times can reveal how interest levels wax and wane for different news events in a population of readers. Such data can be used to fashion similar articles with a view to eliciting similar engagement patterns. Using Twitter data around news stories published by the BBC and Al Jazeera, Lehmann et al. [95] discovered that after a news-tweet is posted, people's subsequent response (i.e., tweets or retweets) are correlated during a brief period of time, before dispersing again, and that similar news articles attracted similar crowds.

2.6.2 Locational Aspects of News Tweeting

When a journalist is *in situ* covering a news event and tweeting about it, more attention and engagement with those news-tweets will occur. The classic case of this phenomenon occurs when a tweeting journalist is an eyewitness at the event and is playing a key informational role for first responders and/or community volunteers. In [33], the authors report on such a case during Hurricane Irene, when professional journalists in a rural community successfully led an online volunteer community effort. On foot of such analysis, it is clear that the geographical sources of tweets provide important information to news providers about reader interests in different geographical locations [69, 114]. Indeed, recent work has focused on analyzing the potential of tools to enrich news stories with network insights of virtual crowds during emerging events [3]. Having said this, it should also be noted that, users' interests can go beyond their location, as they often interact with tweets involving news occurring in places far away from where the users are located [71] (e.g., where people become involved in fund-raising for natural disasters in other countries).

2.6.3 The Visibility of News-Tweets

All news-tweets appear somewhere in the Twitter timeline over the 24-hours of any day on which they are posted. In theory, all these tweets are visible. However, readers are operating under attentional-limitations as they search for posts of interest [65]. This means that all the posts that are physically displayed may never become psychologically visible depending on the time, effort and attention of readers. Once a journalist has posted a news-tweet, its visibility is affected by two opposing forces. On the one hand, if a user (e.g., journalist's follower) has many followees (i.e., more than 250) then she/he is increasingly exposed to stimuli and thus new tweets remain visible for a shorter period of time and are harder to find. On the other hand, if many of the user's followees tweet or retweet the same post, the probability of that tweet being seen, read and subsequently shared, increases [65].

2.7 Studying the Interactions Between Features

In the preceding sections, we have teased out the effects of three distinct categories of features – content, user and context features – on attention to and engagement with news-tweets. This partitioning of features was adopted to develop a simple and clear picture of these effects. However, in everyday news distribution on Twitter, all of these features will be interacting together sometimes working in an additive way, sometimes in a subtractive way and other times in a neutral fashion. We have seen that several studies have been referenced under different feature-categories, when features interact with each other; however, there are significant issues in surfacing such interactions.

There are two major issues that arise in attempting to understand such feature interactions. First, there are very few controlled studies that have specifically examined the interactions between these variables. Second, a very real problem that arises in considering feature interactions, is the diversity of tasks examined in the literature. Part of the fragmentation in the literature arises from the very different questions that researchers have attempted to answer; for instance, some have tried to predict the popularity of news on Twitter, others try to identify bias in online sources, others attempt to understand the factors that affect people's engagement with real-world events, and to find the key groups in advancing news stories on Twitter.

Indeed, to complicate matters further, these studies often use different data sources (e.g., news articles, metadata, online traffic logs, distributions of social media interactions), different output measures (e.g., attention, types of engagement, diffusion), and

Title	Task & *Findings*	Data	Features (following our typology)
The pulse of news in social media: forecasting popularity [10]	Predict popularity of news on Twitter: *Attention can be predicted from news source, news category, language subjectivity, entities*	news articles, metadata and number of link shares on Twitter;	*content:* news category, named entities; *user:* news source, subjectivity
Transient news crowds in social media [95]	Help journalists rapidly detect follow-up stories to their articles: *Users re-tweeting an article can reveal "transient crowds", sources of follow-on news events*	news articles plus tweets including those articles URLs	*content:* URLs, hashtags, topics; *user:* ratio followers/followees, number of followers/followees tweets per day, favorited tweets
Social media news communities: gatekeeping, coverage and statement bias [150]	Identify bias in online news sources and social media communities: *Selection, coverage & statement biases found in news sources, more geographical than political*	RSS feeds and corresponding corporate Twitter accounts of prominent news sources	*content:* mentions, URLs; *user:* partisanship, number of tweets; *context:* geographical region
Newsworthiness and network gatekeeping on Twitter: the role of social deviance [41]	Study the role of social deviance and the resharing of news headlines by network gatekeepers on Twitter: *Twitter users show same preference for social deviant news items as journalists*	tweets with links to news articles	*content:* deviant/non-deviant tweets; *user:* news organization
Predicting user engagement on Twitter with real-world events: [71]	Understand the factors that affect people's engagement with real world events; *Users' prior Twitter activity topical interests, geolocation, social network structures correlate with engagement*	tweets and users related to a list of real world, events plus all tweets of those users in the last six-months	*content:* directed and broadcast tweets, hashtags, topics; *user:* total tweets, tweets/hour, followers/followees, meformer/informer tweets; *context:* geographical proximity
What journalists retweet: opinion, humor and brand development on Twitter [113]	Explore what journalists retweet: *Tweeting journalists often violate norms of objectivity and independence*	tweets and retweets from 8 American journalists	*content:* topics; *user:* number of retweets, name, news organization, description
Journalists and Twitter: a multidimensional quantitative description of usage patterns [9]	Compare Twitter usage by journalists, news organizations and news consumers: *Arab, English & Irish journalists have different tweeting strategies for news*	tweets from journalists and news orgs.	*content:* URLs, hashtags, mentions, replies; *user:* number of followers; *context:* tweet source (mobile, desktop, app)
Sharing news articles using 140 characters: a diffusion analysis on Twitter; [13]	Examine how news articles diffuse on Twitter: *Network & temporal analyses show propagation patterns determining life-spans of news*	tweets with news articles' URLs	*content:* mentions, replies, URLs; *user:* number of followers, followers participation
Breaking news on Twitter[70]	Identify the key groups in advancing the story of Osama Bin Laden's death: *Twitter broke death story then spread by a few "opinion formers"*	tweets covering the first rumors of the news and president Obama's speech	*user:* mentions, URLs, news category
Characterizing the life cycle of online news stories using social media reactions [25]	Study the lifecycle of news articles posted online: *Social media reactions can predict traffic of article within first 10 mins*	news articles with over 3.6M visits and at least 235K social followers/followees (Facebook + Twitter)	*content:* tweets' entropy; *user:* tweets and shares/min, corporate accounts
Modeling gender discrimination by online news audiences [104]	Analyze differences in for online liking, sharing, re-sharing of articles by male & female journalists: *Differences in engagements for articles by men vs. women, women 1/3 rate of men's rate*	news articles plus number of likes, authors' gender; for each article in Facebook, Twitter and Google Plus	*content:* news category; *context:* day of week

Table 2.3: Sample of research papers showing the tasks addressed, findings made (in *italics*), data analyzed, and features explored (following our typology).

examine different platforms in their work (e.g., Twitter, Facebook, Reddit). To understand this diversity, we show how some representative studies can be cast in terms of the features discussed here (see Table 2.3).

2.7.1 Structuring the Diversity of Research

Much of the fragmentation in the literature, arises from different researchers analyzing different tasks and collections of features. In this section, we try to show how this diversity can be structured by dividing studies into their focal task, the datasets they use and the category-features they examine in their work.

A number of studies in the literature were sampled to show how they can be parsed into these different dimensions, allowing us to align them in terms of their content, user and context features (see Table 2.3). In sampling these papers, we have selected ones that represent the diversity that exists in the literature. Hence, some works are relatively highly-cited [10, 25] while others are not [9, 104], some papers are relatively recent [9] whereas others are older [10], and some examine very specific, narrow phenomena (e.g., social deviance [41]) whereas others examine very broad phenomena (e.g., life-cycle of news articles, predicting news popularity on Twitter). Overall, while the sample is small, we consider it represents some of the diversity that exists in the literature providing a good test of whether the proposed framework is useful.

A differentiator between these studies is the task adopted; this plays a key role in selecting the dataset to be used and, perhaps, the features that are deemed to be important in the analyses carried out. Several research groups explore attention to news-tweets but with a different focus and, hence, use different features. In [13], for instance, attention is operationalized as secondary attention with diffusion and social-network features being examined (i.e., how tweets diffuse from one user to the next in Twitter's followers structure), while for others, attention might be operationalized as primary attention measured by retweeting, so social-network aspects play less of a role.

One interesting observation that arises from this organization of studies, is how researchers *customize* the features they use in examining a focal task. For example, in order to identify how journalists share news on Twitter [41], the authors concentrate on the news organization that posts the tweets, and on tweet-content classified as deviant or non-deviant.

For other tasks, namely, predicting user engagement with real-world events [71], describing the usage patterns of journalists [9], identifying transient news crowds [95], and modeling how news articles diffuse on Twitter [13], a more extensive set of features is used (e.g., max. number of tweets per hour or minute, number of followers/followees, total mentions in retweets, number of hashtags, whether the tweet is a meformer or an informer, or the entropy of the tweet vocabulary).

2.8 Summary

The past decade has seen radical disruptions in the news business, in how news is produced, distributed and handled in journalistic practices. One of the main drivers for these changes has been the emergence of social media and, in particular, the use of Twitter as a distribution platform for news.

The focus of this chapter has been on reviewing and discussing the distribution of news on Twitter, particularly, our interest lies in understanding *what* makes news readers actively interact with news-tweets and pay attention to them. To this end, we presented different types of news posts found in Twitter, and discussed how computer science researchers *dissect* such tweets in order to extract features that are later used in a variety of tasks.

In the last part of this chapter, we took a detailed view of three types of features of news-tweets, namely content, user, and context features, and examined how these are used together to address research tasks oriented at understanding both journalists and news readers' behavior and attention to news on Twitter. Finally, we discussed efforts by both news agencies and computer scientists, directed at helping journalists and news organizations to take advantage of the possibilities that Twitter offers as a news distribution platform.

In the next chapter, we present and discuss a survey designed to gain insights into the use of Twitter for news dissemination, as reported by professional journalists.

CHAPTER

THREE

Journalistic Practices on Twitter

Journalists are a key group of Twitter users, they not only are active and attract large audiences but also contribute with high quality content. As such, their practices, the challenges they face, and their purposes when using the platform, should be studied. However, their actual professional practices and needs are rarely analyzed in isolation from those of other groups of users (e.g., artists, athletes, or politicians).

In this chapter we present a survey designed to gain information and understanding on current journalistic practices on Twitter, as expressed by professional journalists themselves. The survey's results open a door into journalists' view of Twitter as a tool for disseminating their news, highlight open problems, and provide important input for identifying the role of different characteristics of tweets and journalists in the quest for audience attention. Furthermore, together with current literature (see Chapter 2), the findings of this survey help scope the research presented in the subsequent chapters of this thesis.

The rest of this chapter is organized as follows: in Section 3.1 we describe the survey's design and participants, in Section 3.2 we present the responses and results, and in Section 3.3 we discuss the findings and conclude.

3.1 Survey Design

The survey design is based on actual advice given by Twitter to journalists and news organizations who wish to make the most out of their tweets [107, 108] (see Appendix A for full survey). The survey contains a total of eighteen questions, with a combination of multiple choice and free-form responses, and targets five main areas: *Twitter usage* (four questions), *joining the conversation* (three questions), *increasing followers and improving engagement* (three questions), *getting to know your audience* (four questions), and *measuring engagement on Twitter* (three questions). The remaining question asks journalists about the news category they report. We created the survey and collected responses using Qualtrics[1].

The questions on *Twitter usage* are concerned with understanding the journalist's basic use of the platform. We ask journalists how often they use Twitter, their searching and retweeting behavior, and the situations in which they reply to and mention others in their tweets. The questions about *joining the conversation* are designed to understand the mechanisms journalists use for sharing news and information with a broader audience: the inclusion of hashtags, media or quotes in their tweets, whether they post news as soon as it occurs, and how often they interact with their audience either by replying or answering questions.

Questions on *increasing followers and improving engagement* relate to the techniques journalists use to grow their audience and increase engagement with readers. Here we ask journalists about the regularity with which they tweet and if they follow up and update their audience on past stories, whether they actively seek to interact with other prominent users, and about the actions they take to increase engagement with their readers. To gain understanding about the information that journalists find valuable and important about the audience, we ask each participant if and what tools they use to track their Twitter audience, the features that they find important or would like to know about their readers, and their preferences regarding audience size (in the section *getting to know your audience*). In the final questions: *measuring engagement in Twitter*, we ask our participants if and how they measure audience engagement to their tweeted news and to indicate characteristics of both journalists and tweets that they consider influential in the audience reaction.

Terminology. A Twitter user connects with others by following them, which enables them to see, read and retweet the tweets that they send. A retweet is a re-post of another user's original tweet. Users can informally categorize or tag the content of their tweets by using tokens called hashtags, for example: #health, or #fashion. Other functionality to enable interaction between users are replies and mentions. Replies are direct responses to another person's tweet and start with that person's @username. Mentions occur when a user includes one or more @usernames anywhere in the tweet. Users can also include URLs, photos, or videos within their tweets.

[1] https://www.qualtrics.com

3.1.1 Participants

Twenty professional journalists who are regular Twitter users participated in the survey. We used Twitter and email messages to contact them and waited for their consent before sending the link to the survey. No monetary compensation was offered.

Among our participants we have 8 female and 12 male journalists working for well-established news outlets including *The Irish Times* and the *Irish Independent*, as well as other publications such as *Global Voices*[2], or *Migrante21*[3]. As such, they know the challenges of reporting both for traditional media, as well as using Twitter to reach a broader, potentially more diverse audience with their tweeted news.

The most popular news categories reported by the participants in our survey are politics (25%), science and technology (15%), sports (10%), business (5%), and breaking news (5%). The rest of the participants report on other categories such as consumer affairs, lifestyle, human rights, social movements, underrepresented or under-reported topics, freedom of expression, digital rights, and education.

A Note on Participants Selection. A total of 100 journalists were contacted by email and/or Twitter. The decision on what journalists to contact was based on the Twitter accounts under study (i.e., we contacted popular and active journalists whose Twitter data was available to us for further analysis). We also contacted English-speaking journalists from smaller, more niche news organizations outside Ireland, for example Migrante21 in the United States. The idea of including such participants, was to observe differences in Twitter strategies based on culture, geographical location, or news organization.

Besides the gender, we did not collect any demographics or other data from our participants. The reason behind this decision, is that several journalists did agree to participate in our survey but did not agree to be identified nor to give more details about themselves, which would leave us with only partial information, not usable for posterior analysis.

3.2 Findings

We describe and summarize the results based on the five main areas that structure our survey, namely *Twitter usage*, *joining the conversation*, *increasing followers and improving engagement*, *getting to know your audience*, and *measuring engagement on Twitter*.

[2]https://globalvoices.org/about/

[3]http://migrante21.com/en/about-us/

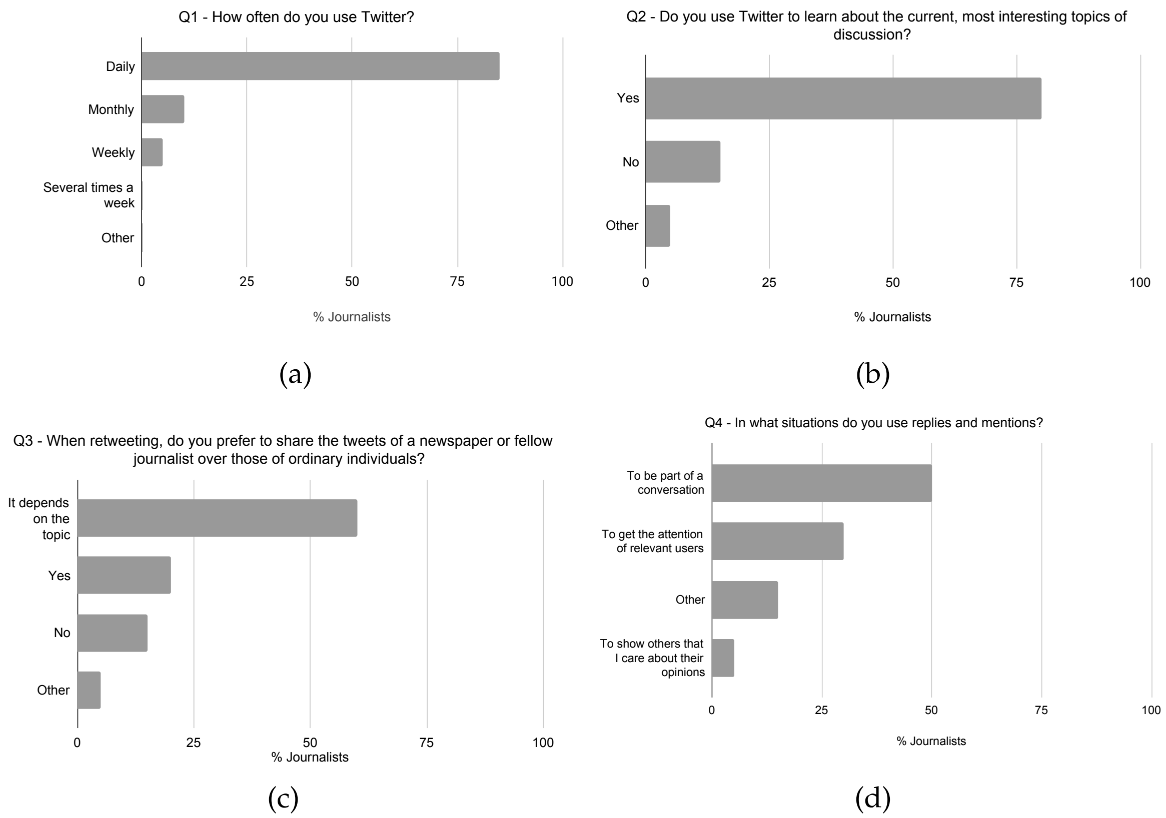


Figure 3.1: Responses (in %) to questions on *Twitter Usage*.

Twitter Usage. Most of our participants (85%) indicated that they use Twitter on a daily basis (see Figure 3.1a) and that they use it to learn about interesting topics of discussion (80%). One of them reported: *"[I use Twitter for] escapism and second stop for emerging news (news organisation for digest and context then Twitter for viewpoints and comments before looking for wider views elsewhere)"* (see Figure 3.1b). Sixty percent of the journalists said that, depending on the topic, they sometimes prefer to share the tweets posted by a known source (e.g., news outlet or fellow journalist) rather than those of an ordinary individual. Twenty percent said that – despite the topic –they prefer to retweet journalistic sources, and 15% have no preference. One journalist mentioned that the decision on whose tweet to retweet depends not only on the topic, but also on the opinion being shared and the trustworthiness of the sharer (see Figure 3.1c).

Regarding the situations when they reply to or mention other users in their tweets, the journalists said to do it mainly for two reasons: to be part of the conversation (50%) and to get the attention of relevant users (30%). One participant; however, said to use replies and mentions to show others that she/he cares about their opinions, and the remaining three journalists mentioned that their use of replies or mentions is based on a combination of these reasons and beyond: *"sometimes just to be polite if someone has mentioned me"* (see Figure 3.1d).

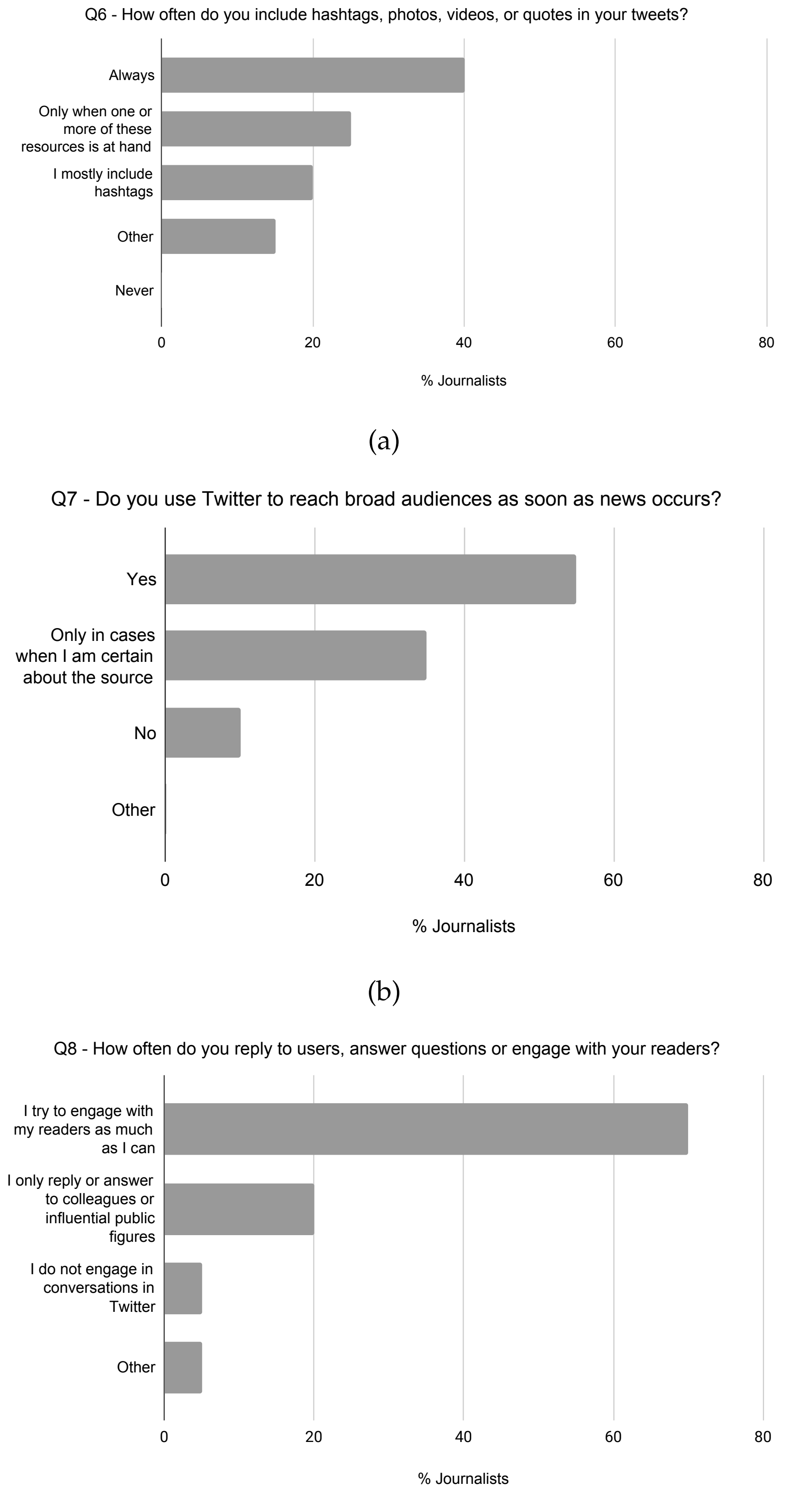
Joining the Conversation. Nearly half of our participants (40%) always include hashtags, media or quotes in their tweets. A further 25% reports using these resources only if they have them to hand at the time of posting the tweets, and 20% use mostly hashtags. The rest of the journalists stated they use hashtags, media or quotes only in cases *"when it is relevant and illustrative to do so"*. One journalist said: *"I use hashtags sparingly, I think they can be annoying. If I'm tweeting a newspaper article I will usually extract a quote from the piece and tweet that for context. I use photos quite frequently, rarely use video"* (see Figure 3.2a).

Most of the participants (55%) said they tweet news as soon as it occurs, 10% said they do not, and the remaining 35% does so only when the source is trustworthy and the topic is trending (see Figure 3.2b). Regarding how often they reply to users and answer audience's questions (see Figure 3.2c), 70% said that they try to engage with their readers as much as they can, while an important 20% restrict their replies only to colleagues and influential figures. The remaining 10% either do not engage in conversations on Twitter or do so in their own terms; for example, one participant stated: *"I try to engage with my readers as much as I can but avoid hate talk and all forms of trolling"*.

Increasing Followers and Improving Engagement. Seventy five percent of the journalists indicated they post updates and follow ups to their stories and include media whenever possible (see Figure 3.3a), the rest of the participants do not follow up on stories, or if they do, they do not include any new media.

Twitter offers journalists a platform to spread their news, and to establish connections, or conversations with other users. Seventy percent of the participants in our survey, admit seeking to engage with prominent users either as often as possible or only when in need of spreading a relevant post. Of the remaining 30%, 20% do not actively seek to connect with prominent users on Twitter, and 10% said that rather than trying to connect, they seek to alert people they mention in their news, or that they tend to have conversations with other journalists covering similar stories (see Figure 3.3b).

All of our participants take actions to increase engagement with their readers (see Figure 3.3c). Hashtags are the single most used artifact to this end (35%), followed by retweets (15%), mentions (15%), and replies (10%). Twenty five percent of the journalists said they use a combination of hashtags and mentions, the *like* button to acknowledge they have read a tweet, reply to tweets where they are being mentioned, retweet relevant tweets, and mention users if they are quoted on a news piece. One participant points out: *"I include hashtags and mention users when appropriate or to direct the audience towards them and/or offer verification and leap point for the audience"*.

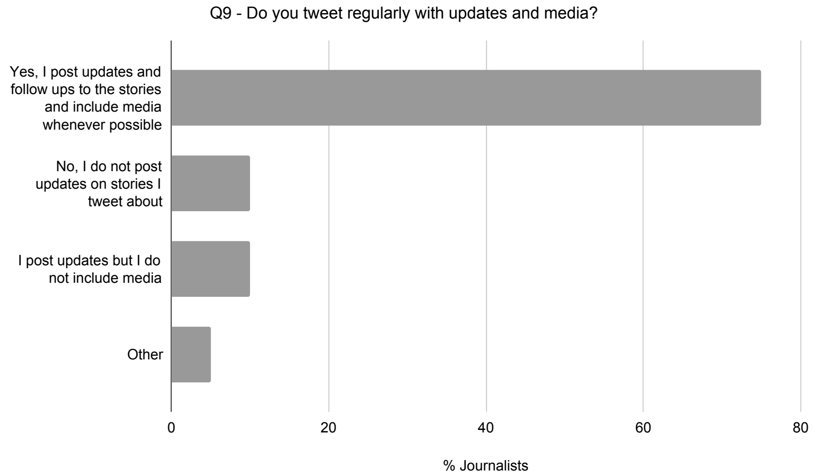


(a)

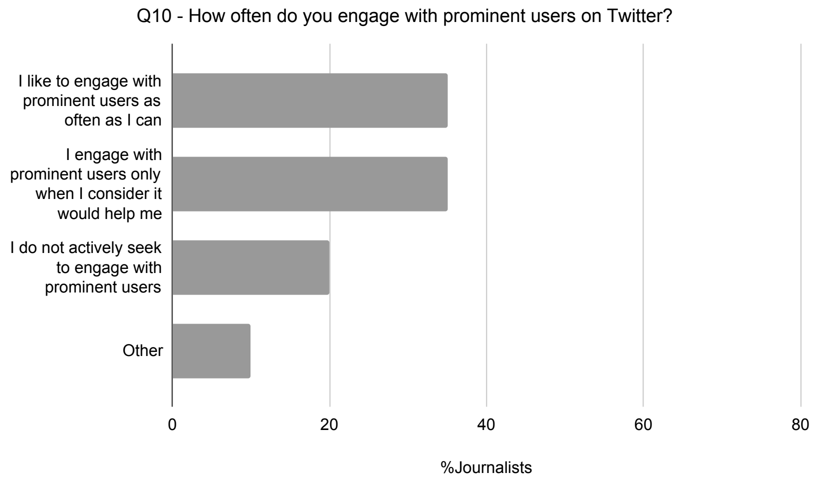
(b)

(c)

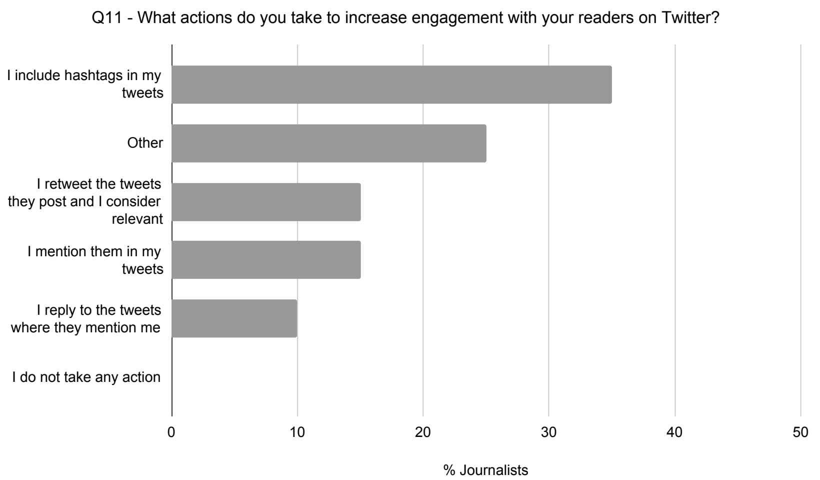
Figure 3.2: Responses (in %) to questions on *Joining the Conversation*.



(a)



(b)



(c)

Figure 3.3: Responses (in %) to questions on *Increasing Followers and Improving Engagement*.

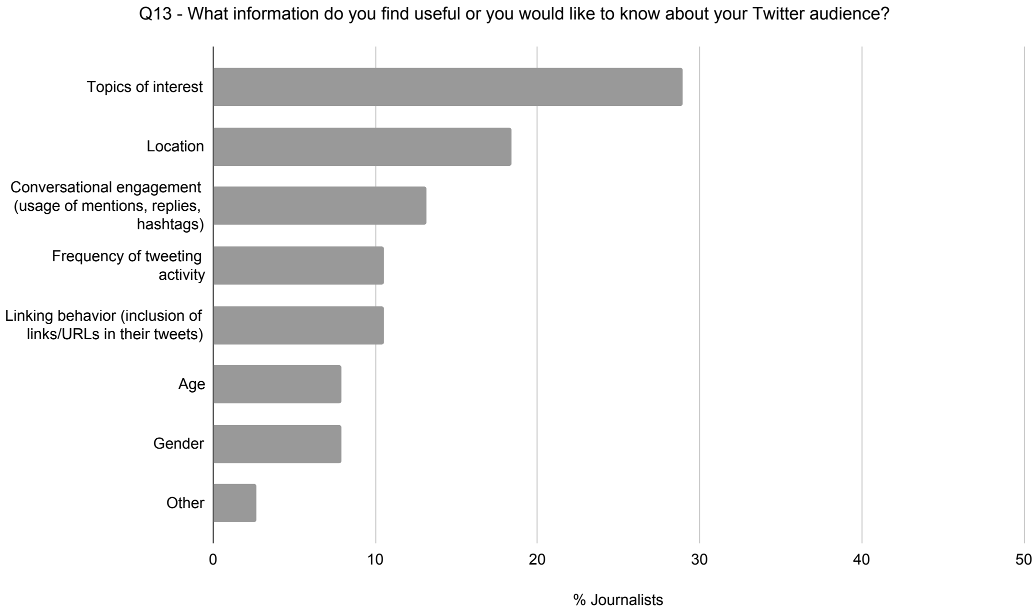
Getting to Know your Audience. To keep track of their Twitter audience, our participants use Twitter analytics[4], Tweepsmap[5], Buffer[6], or Muck Rack[7]. When asked about the audience information they would consider most useful to know, in 29% of the responses journalists said they would like to know the topics of interest for their Twitter audience, followed by location (19%), conversational engagement (13%), and linking behavior (11%). One participant mentioned: *"I would love to be able to track users who share our pieces but don't mention us directly"* (see Figure 3.4a).

In terms of audience size, we asked journalists whether they preferred to have a small but loyal and active group of readers or a larger, more diverse group even if they are less active. Approximately 60% of the participants indicated that they would like to have a larger audience; however, more so due to the diversity this brings than merely for the number of readers. The remaining 40% prefer to have a small audience, but one that is loyal and, in some cases local.

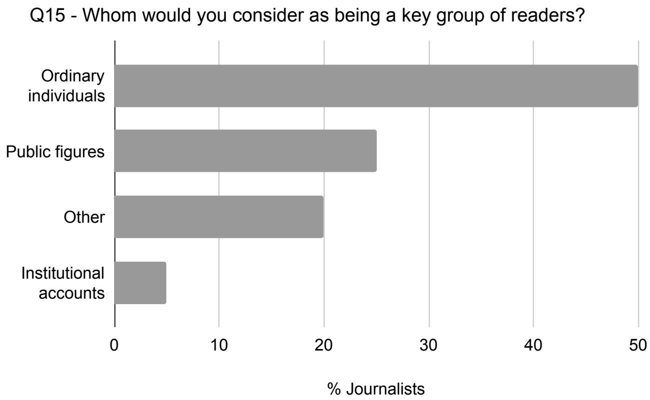
Half of the journalists consider ordinary individuals as their key group of readers in Twitter. A further 45% are more inclined toward public figures or trusted Twitter users in general (i.e., they can be public figures, ordinary individuals and/or institutional accounts). One participant said: *"Other people in media following my account is important. Twitter is a very good networking tool and tweets that I post often lead to TV or radio appearances"* (see Figure 3.4b).

Measuring Engagement. Our participants indicated that they measure the audience reaction to each one of their tweets mostly by the number of retweets the post receives (30%), the number of different users who react (i.e., retweet, favorite) to the tweet (30%), and by how fast the tweet starts getting reactions after being posted (15%). Another metric mentioned is the traffic from Twitter that goes to the articles on their websites (see Figure 3.5a).

[4]https://analytics.twitter.com/
[5]https://tweepsmap.com/
[6]https://buffer.com/
[7]http://muckrack.com/whoshared/

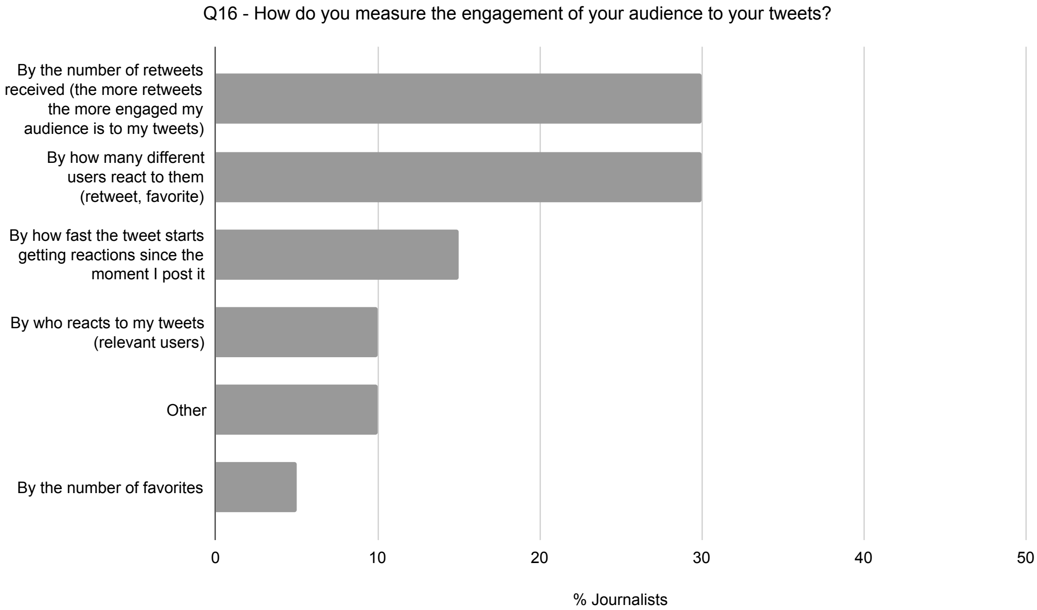


(a)

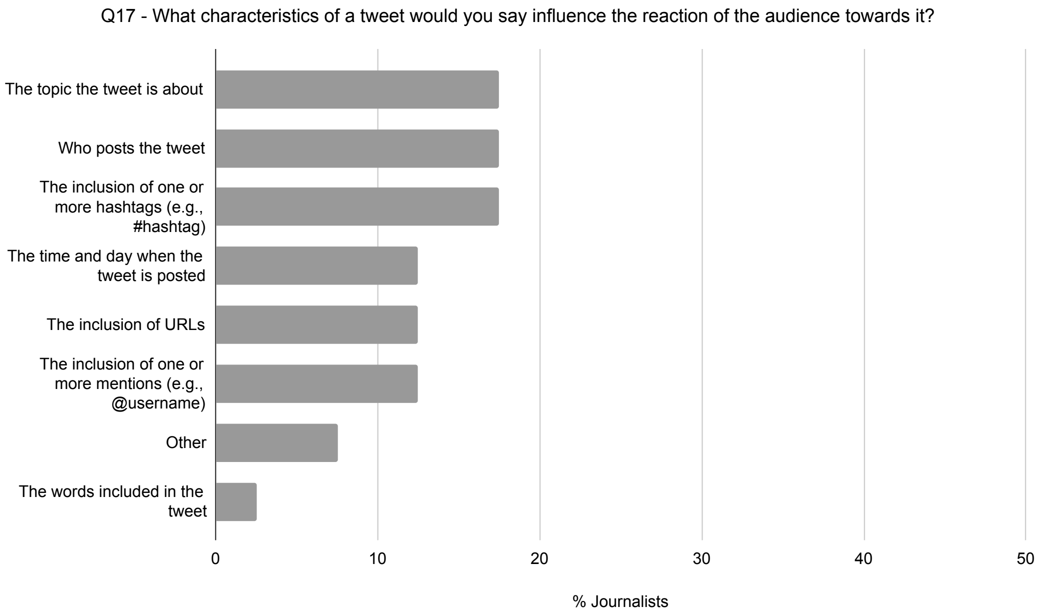


(b)

Figure 3.4: Responses (in %) to questions on *Getting to Know your Audience*.

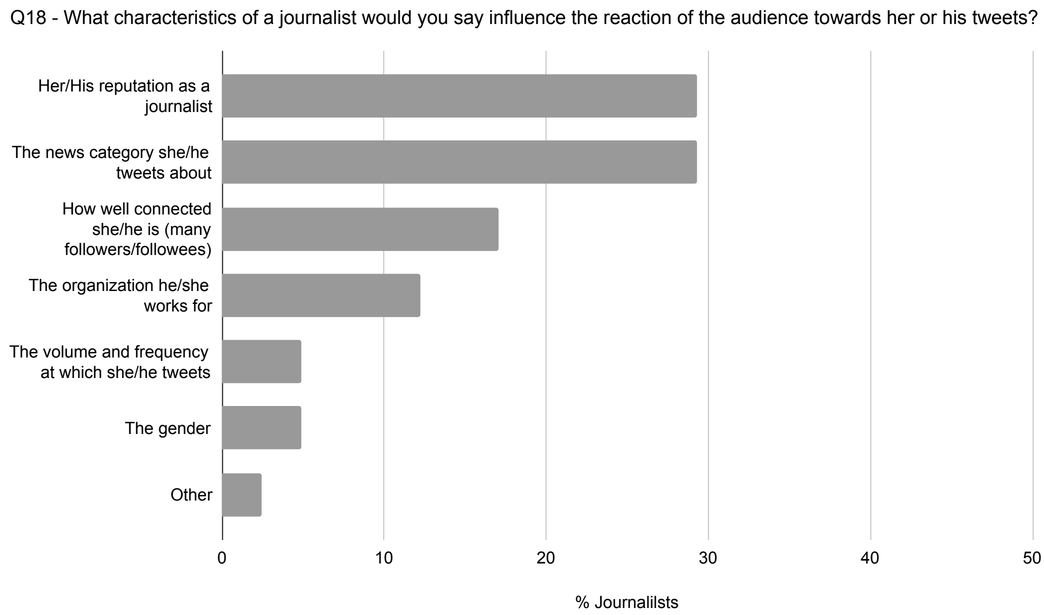


(a)



(b)

Figure 3.5: Responses (in %) to questions on *Measuring Engagement*.



(c)

Figure 3.5: Cont. responses (in %) to questions on *Measuring Engagement*.

The characteristics of the tweets considered to have the most influence on audience reaction are the topic of the tweet (18%), its author (18%), and if it contains one or more hashtags (18%). The time when the tweet is posted, and if it contains URLs and mentions, are also considered to influence audience reactions, more so than the actual words of the tweet itself. Beyond the above mentioned characteristics, journalists reported: *"People respond most to humor and surprise"*, *"I think the way a tweet is phrased is very important - if you are funny or neatly encapsulate an opinion or statement, it tends to get traction"* (see Figure 3.5b).

Audience reactions to news-tweets might be influenced by more than just the content of such tweets (see Figure 3.5c). 30% of our participants considers that the reputation of a journalist impacts whether and how readers react to her/his tweets, and a further 30% highlighted the news category that a journalist reports as being one of the drivers of readers' attention. The number of followers/followees of a journalist and the organization they work for, were also considered important. Among other characteristics, one journalist mentioned that audiences are influenced by the level of engagement a journalist shows to their readers, and *"if she or he is verified"*[8].

[8] A verified account – known by a blue badge next to the username – is Twitter's means to let people know that an account of public interest is authentic.

3.3 Discussion

The survey presented in this chapter, provides a glimpse into the world of professional journalism on Twitter. From the responses we learned that most journalists use Twitter on a daily basis, that they are aware and make use of the different interaction mechanisms available on the platform (e.g. hashtags, mentions, or replies), that they employ diverse tools to gain information about their audience, and that they consider concrete values such as number of retweets, or unique retweeters as good proxies to measure engagement to the news they post.

We also learned that, unlike in traditional news rooms, journalists who use Twitter as a platform for news reporting have little or no guidance on matters such as code of conduct, strategies to use Twitter effectively, audience interests and expectations. Indeed, it is evident from the responses that journalists approach news reporting on Twitter in an experimental manner; that is, many use hashtags, contact prominent users, or engage with the public in an effort to spread their news or to gain access to useful information.

Based on the findings of this survey and the research presented in Chapter 2, in the following chapters we conduct a series of data analyses, model audiences' behaviors, and design tools that aim at assisting journalists in their reporting of news-tweets.

CHAPTER

FOUR

Attention to News on Twitter

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, 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, perhaps 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 chapter, 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.

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 articles they have written, links to articles from other sources, or include journalistic commentary on events/comments/opinions tweeted by relevant figures and that they consider newsworthy (what we 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., [72, 152]). 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 [1]).

However, ultimately, Twitter is a distribution channel for news that is not controlled by 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 [129]. 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?"*

Prior to this thesis, no clear answers to this question had been forthcoming. It was still largely unclear, from both research studies and journalistic practice (see Chapters 2 and 3), how to optimize audience engagement for news-tweets. 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. To add to this, it has been shown that beyond knowledge gain, the key role of social media in news content is the ability to engage users who may be passively receiving news on these sites [119].

In this chapter, 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 [86], that news can first break on Twitter [127], and has identified some of the factors that influence the dissemination of a tweet [145]. However, this prior work has seldom specifically focused on journalistic tweeters or, indeed, on news-tweets in determining audience engagement.

The work in this chapter has been informed by findings presented in Chapters 2 and 3, specifically:

- In Chapter 2, we learned about tweet's characteristics found in existing literature. Such characteristics capture different aspects of tweets, can be used to extract features and address tasks such as audience predictions, and can be grouped into three categories: user, content, and context. In this chapter, and based on our analysis of the literature, we use 40 features of news-tweets that fit in the above mentioned categories, to form feature vectors that are later used in the prediction of attention to news-tweets.
- In Chapter 3, we presented a survey where 20 journalists reported on their usage of Twitter as a news dissemination platform. The findings of this survey reveal that one of the preferred measures of audience engagement among these journalists are the number of retweets, and that characteristics such as the topic of a tweet, who posts it, the usage of hashtags and urls, the reputation of the journalists, the organization they work for, and the volume and frequency at which they tweet, are considered to impact the attention a news-tweet receives. Consistent

with our findings in Chapter 2, it confirms our selection of news-tweets features and the decision of using "number of retweets" as our proxy for audience engagement.

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 UK journalistic accounts. 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, these analyses help formulate a set of guidelines for journalists when they are tweeting their news (these guidelines are presented in Chapter 6).

In the next sections, we present our analysis of journalistic tweeting (Section 4.1), and develop predictive models for news consumption for six news categories (Section 4.2) before concluding in Section 4.3.

4.1 Do News Categories Differ?

In the 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 its impact.

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 tweeting about politics. In the present section, we describe the collection of tweets associated with 564 journalistic accounts and perform analyses to determine what aspects of *tweeting the news* seem to matter; specifically, whether tweeting about one news category may differ from another.

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			2,996,963

Table 4.1: Data collection (Note that these numbers reflect exclusively the tweets sent by journalistic accounts).

4.1.1 Data Collection

To study journalistic tweeting, we manually curated a list of 200 Irish and 364 UK 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 UK news sources were chosen for two reasons. First, these journalists have been shown to be particularly active in social media, by global standards [61]. 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 these accounts (e.g., tweets replying or mentioning any of the journalistic accounts). The datasets are summarized in Table 4.1.

Each account was manually labelled according to the following aspects (see Table 4.2 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

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

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.

4.1.2 Judging the News Categories in 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., [8, 145]). 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 al-

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

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

ways 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.

In this analysis, we associate each journalist with a single news category, which corresponds to the one that she/he specializes in 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 in one single section of the news. Their focus is reflected in 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 outside of the scope of this 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 [35]. 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 *The Irish Times* Twitter account tweets right across all of its news categories).

Judging the News Category of an Individual Account. Three judges, all PhD students and regular Twitter users, manually annotated the news category for each journalistic account in the Irish and UK corpora.

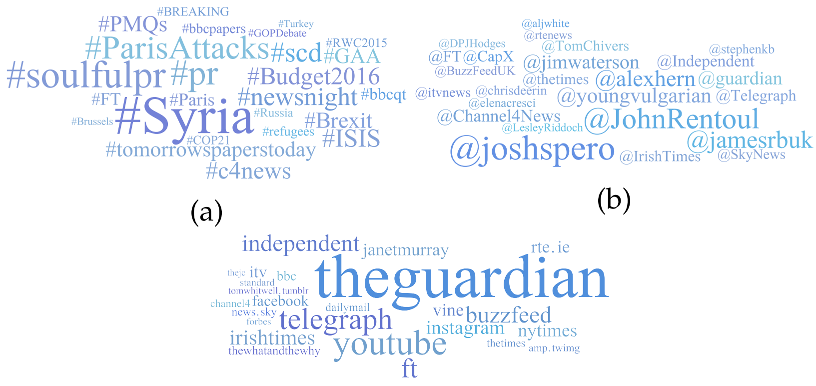
The news categories included business, lifestyle, breaking news, science and technology, politics, and sports. To judge the category of each account, the annotators were given (i) a random sample of 50 tweets sent by the individual journalist and (ii) a list of the top-100 terms used by the journalist 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 as follows:

- **Business.** News that tracks, records, analyzes and interprets the economic changes that take place in a society. It could include anything from personal finance, to the 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. It includes 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 editors.
- **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 PhD candidate assigned a label after further analysis of the tweets and top-scoring

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



(c)

Figure 4.1: 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).

terms. The Fleiss' Kappa inter-rater agreement for Irish accounts is $\kappa = 0.51$. For 164 out of the 306 UK 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 PhD candidate. For UK 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 4.3.

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 4.3: 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 UK accounts) are verified by Twitter. Second, the top-25 hashtags, mentions and domains used in these accounts uniformly address news topics and current affairs (see Figure 4.1). 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 *The 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 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. A random sample of 1,000 tweets from these accounts[4] were selected and two annotators manually labelled 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 $\kappa = 0.86$.

4.1.3 Exploring News Categories

Having made the division between corporate and individual accounts and labeled the news category of individual accounts, we explore whether there appear to be any systematic differences between these subsets of tweets on other dimensions (e.g., time of day). In this exploration we use 2.9M tweets – 1.7M for Ireland and 1.2M for the UK.

Individual Accounts: Activity Levels Across Countries
Figures 4.2 and 4.3 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 4.2a and 4.3a), 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

[4]Our existing dataset of tweets by Irish and UK individual journalists was combined into one single file which lines were then shuffled and a random non-stratified sample of 1000 tweets were extracted. Tweets in this sample belong to our collection spanning the time periods shown in Table 4.1.

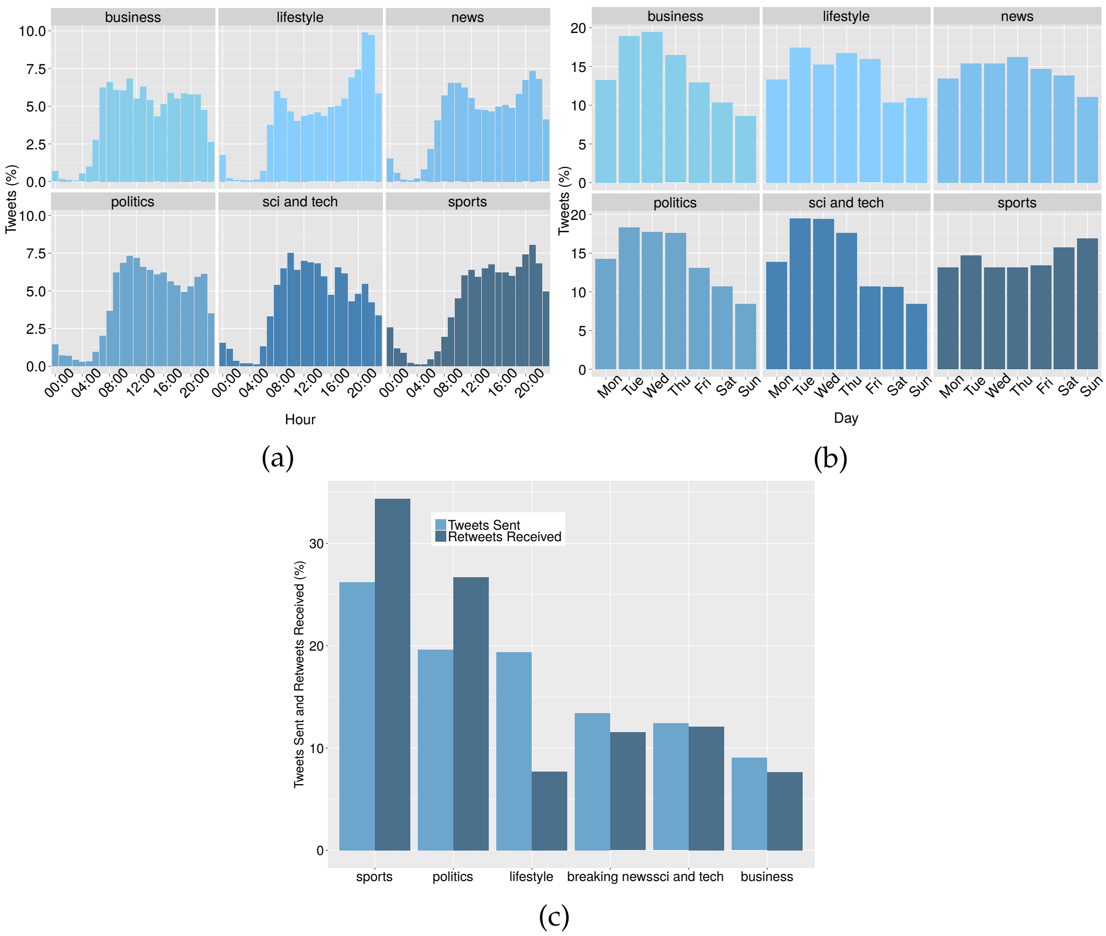
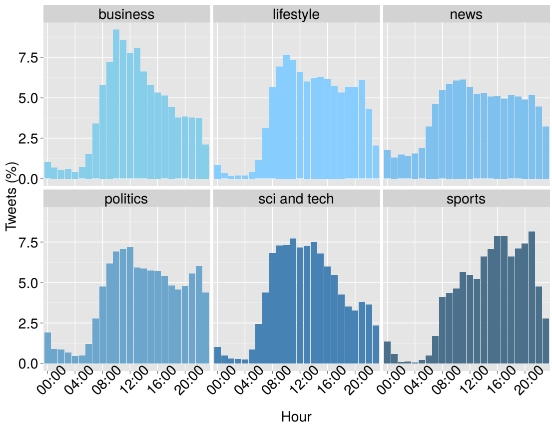


Figure 4.2: Tweets posted by news categories in Ireland (a) per hour, (b) per day (normalized by the total number of tweets per category), and (c) proportion of tweets sent and retweets received by news category (normalized by number of individual accounts in each category in Ireland).

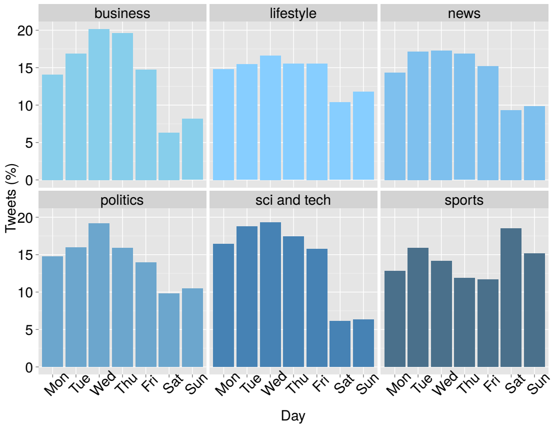
news occurs at a fairly constant level, though the most popular time for posting tweets is approximately 10:00 in both countries.

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 4.2b and 4.3b). 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.

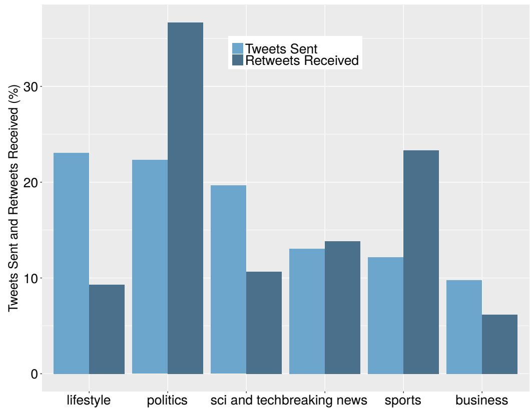
When we break out the proportions of tweets sent and retweets received by these ac-



(a)



(b)



(c)

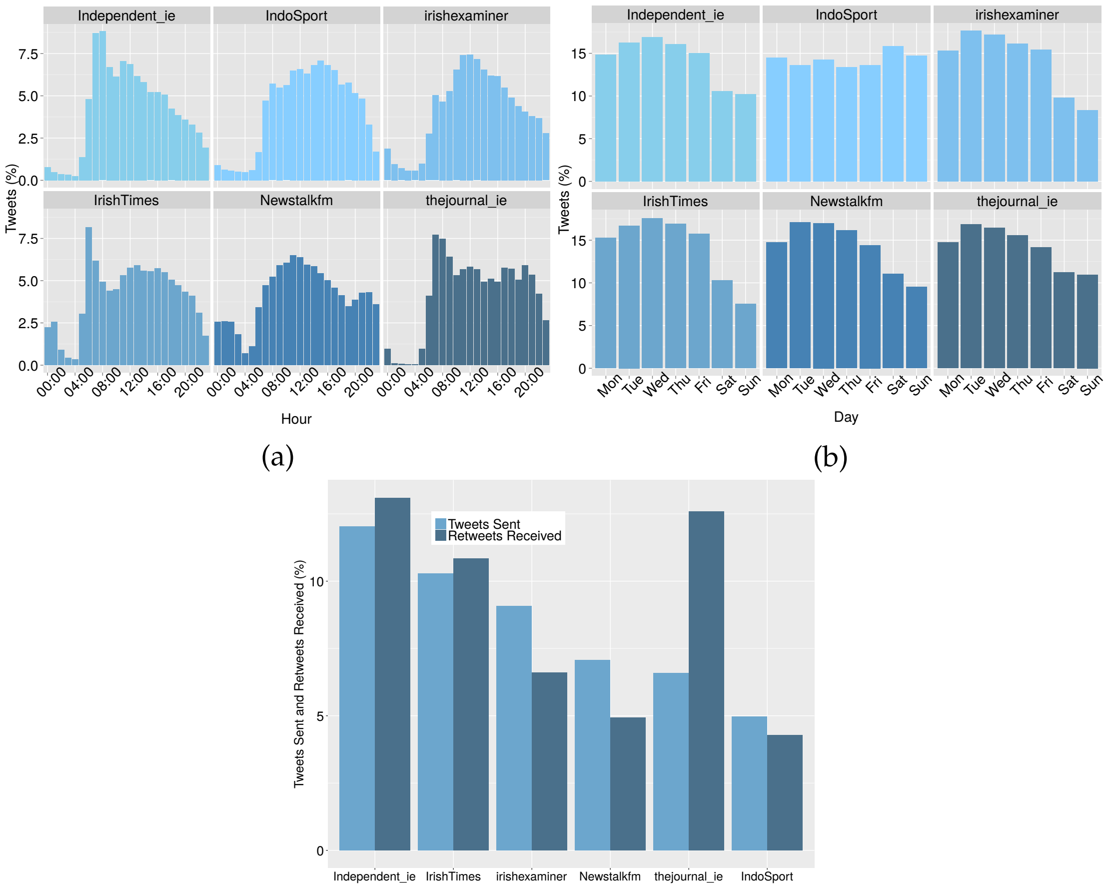
Figure 4.3: Tweets posted by news categories in the UK (a) per hour, (b) per day (normalized by the total number of tweets per category), and (c) proportion of tweets sent and retweets received by news category (normalized by number of individual accounts in each category in the UK).

counts (see Figures 4.2c and 4.3c), 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 60% of the total retweets being received by these two categories (see Figure 4.2c). 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 4.3c). 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 (see Figure 4.2c), but their order based on account numbers is breaking news, sports, politics, lifestyle, business, and science and technology (see Table 4.3).

Corporate Accounts: Activity Levels Across Countries

Figures 4.4 and 4.5 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.

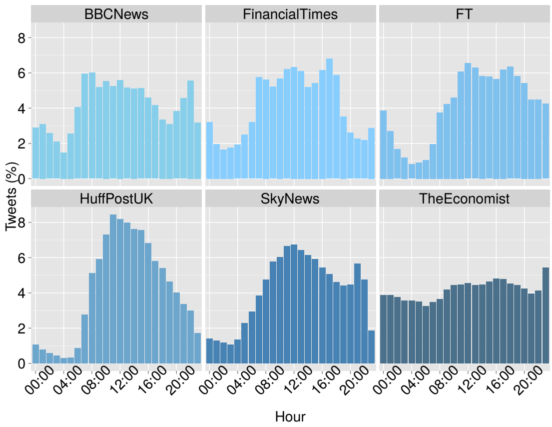


(a) (b)

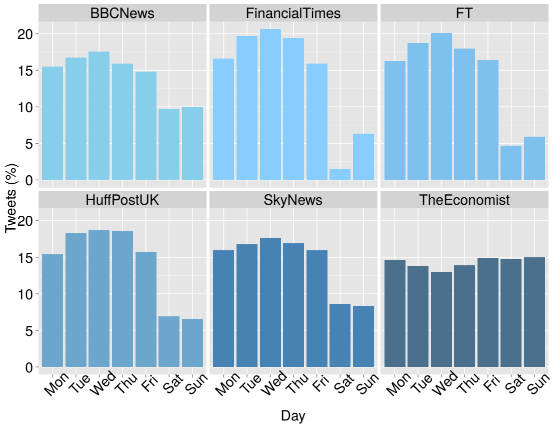
(c)

Figure 4.4: 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).

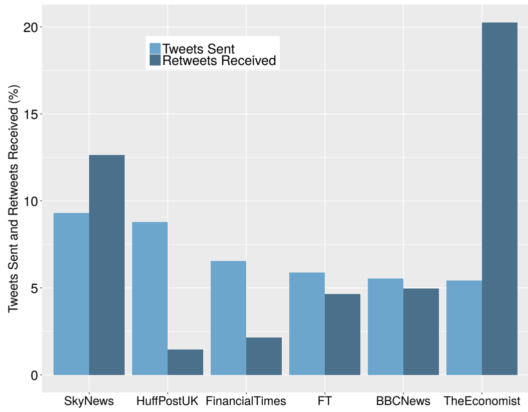
In Ireland, *The Irish Independent*, *The Irish Times*, *The Irish Examiner*, *Newstalkfm*, *The Journal*, and *Irish Independent Sport (IndoSport)*, are responsible for a high proportion of all news-tweets sent from corporate accounts in the country (see Figure 4.4). These accounts correspond to the most important news outlets nationwide.



(a)



(b)



(c)

Figure 4.5: 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).

The Irish Independent and *The Irish Times* have high levels of tweeting activity between 06:00 and 08:00 (see Figure 4.4a). *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 4.4b), with the exception of *IndoSport* which, as in the case of the sports category (see Figure 4.2b), has the most active days on weekends.

Figure 4.4c 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 4.5. 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 4.5a), 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 4.5b, weekdays are highly active for UK 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 4.5c shows the proportion of tweets sent and retweets received by each UK 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 within 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 4.5c), 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.

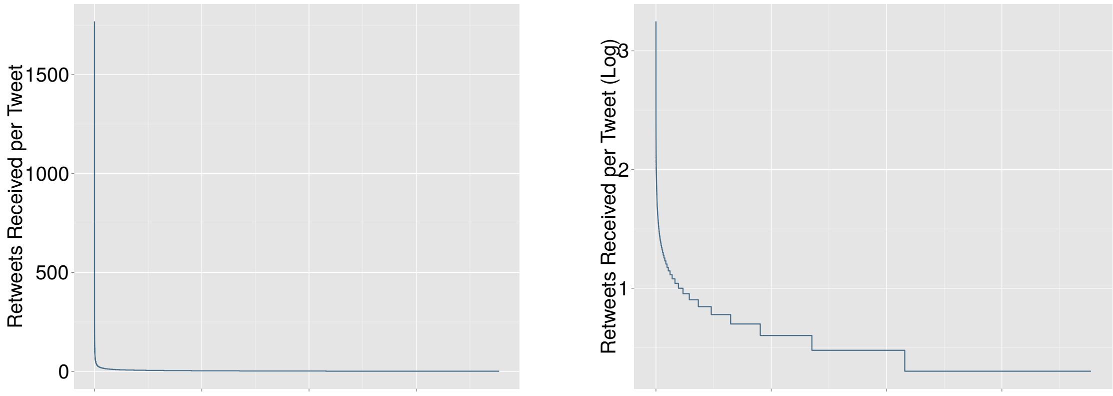
A further conclusion warranted by this data is that, for diverse reasons, news outlets are differentially successful in eliciting engagement from their readers. For example, *The Irish Independent* (Ireland) and *Sky News* (UK) carry out high volume tweeting that appears to elicit correspondingly high 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 chapter, we analyze the drivers for these very different behaviors, to understand what is it about news providers and their tweets that leads to high levels of audience engagement.

4.2 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 UK 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 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 4.6a). Accordingly, we used the natural logs of these retweet counts[5] in our analyses (see Figure 4.6b).

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



(a) Retweets per tweet.

(b) Retweets per tweet (Log).

Figure 4.6: Distribution of (a) retweets and (b) natural log of retweets received per tweet in Ireland. The X-axis shows 1M individual tweets and the Y-axis shows the corresponding number of retweets received by each of these tweets.

Each of the countries tweet corpora were 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 categories.

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., [151]). 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 [84]. 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.

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

4.2.1 Method and 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 4.4 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*.
- **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 [27, 163]. 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" [144]. 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.

Most of the news-tweets' features used in this thesis are based on those found in existing literature; however, we represent them so as to capture the activity and style of our target population, journalists and news organizations. For example, common features used in the literature such as *number of tweets*, are here divided by day segment (morning, afternoon, evening), and day of the week as well as weekday or weekend. This approach allows us to capture granular temporal aspects of the tweeting activity that are more adapted toward journalistic usage of the platform (e.g., tweeting activity corresponding to news broadcasts). Other features, such as *organization* and *account type*, were contributed by this thesis and are specific to the population under study.

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

As observed in Table 2.2, a common approach to working with tweets is to focus on very specific features, for example, textual elements such as standout phrases, hashtags usage, gender, nationality and/or ethnicity of the author, and temporal aspects such as the time and date of a tweet's publication. In this work, however, our selection of features ranges over three areas, namely content, user and context. Such combination of features, allows us to analyze attention to news-tweets from a different, more complete, perspective.

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 [176].

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 [153]. RF 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 [20]. GBT 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 [50].

Feature normalization. All the features that conform the news-tweets' vectors are important and one or a few of them should not dominate the final results, merely by having a value that is too different that the others. As a first step in our data processing, we discarded Twitter accounts whose levels of tweeting were too high in comparison with the rest of our subjects (i.e., avoid possible bots or bot-like behavior). Secondly, we performed min-max scaling for numerical features and assigned values of 0 and 1 to categorical features such as *is weekend*, *day of week*, *contains hashtags*, and *is retweet*. Finally, algorithms used in our experiments such as RF and GBT, are invariant to monotonic transformation of the features and per feature scaling will not impact the results [60].

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

[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 [149].

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 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., [109]). We computed the MSE for each tweet in the test set and then took the average value across all these tweets.

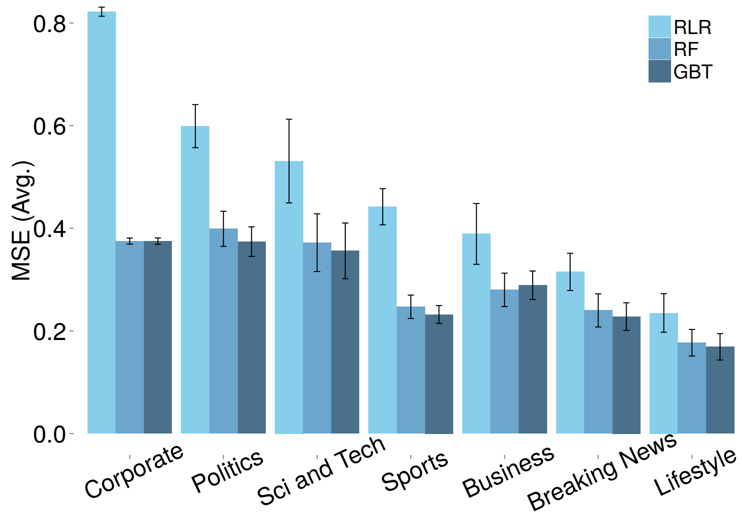


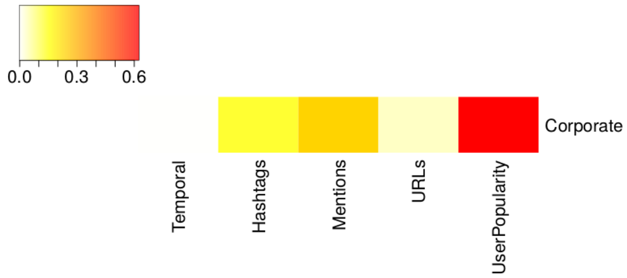
Figure 4.7: Average MSE values for the different models in predicting engagement based on the Ireland Twitter corpus, showing 95% confidence intervals (as these are error values, the lower the value the better the method).

4.2.2 Results

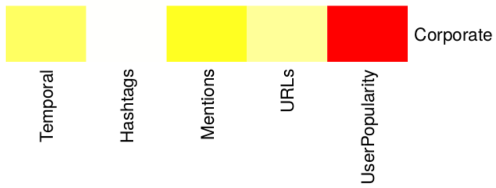
Figure 4.7 shows the prediction performance for the three regression methods applied to the Irish Twitter corpus[8]. GBT and RF perform better than RLR in terms of MSE. Except for the case of the business category, models generated using GBT have a lower error than those using RF, although the difference 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., [42]).

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 4.8 and 4.9 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 categor-

[8]A similar pattern was observed for the UK Twitter corpus



(a)



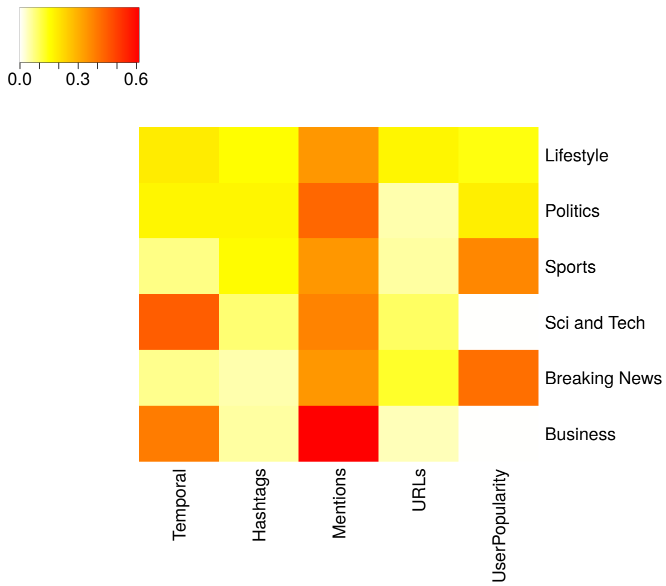
(b)

Figure 4.8: Feature importance for five feature groups (columns) in predicting engagement for corporate accounts in (a) Ireland and (b) the UK. Values of feature importance i per country add up to 1 and are represented by the intensity of color, i.e., white means that a particular feature group is not important for the prediction of engagement $i = 0$, pale colors represent values ranging $0 < i < 0.4$, and high intensity oranges and reds mean that the corresponding feature groups are highly important to the model's predictions $i >= 0.4$.

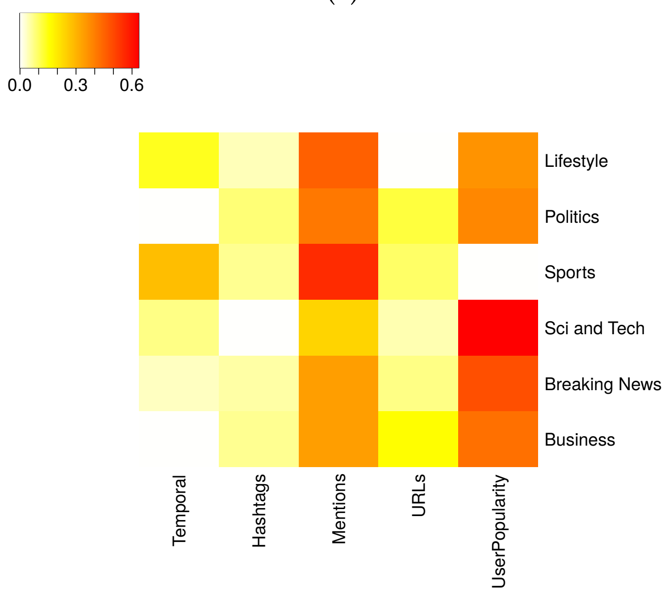
ies (i.e., tweeting about sports demands a different strategy to tweeting about business). Interestingly, also, there are notable differences between countries.

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 important feature for audience engagement is user popularity (see Figures 4.8a and 4.8b); 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 indicate that the audience for these accounts responds 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 news 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.



(a)



(b)

Figure 4.9: Feature importance for five feature groups (columns) per news category (rows) for individual accounts in (a) Ireland and (b) the UK. Values of importance i of the five feature groups (columns) per news category (row) add up to 1 and are represented by the intensity of cell color, i.e., white means that a particular feature group is not important for the prediction of engagement $i = 0$, pale colors represent values ranging $0 < i < 0.4$, and high intensity oranges and reds mean that the corresponding feature groups are highly important to the model's predictions $i >= 0.4$.

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 4.9a and 4.9b). The corporate versus individual account differences are also interesting. In the individual accounts (see Figure 4.9) user-popularity features are less critical than in corporate accounts (see Figure 4.8), 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 4.9a and 4.9b, 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.

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 4.9a and 4.9b).
- *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 4.8).
- *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.

4.3 Summary

This chapter 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 (ii) the ways in which they interact across different news categories revealed.

There are, however, a number of caveats that need to be considered, ones that may usefully define future work in the area. In this chapter, 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 [41]). 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 sub-divide these tweets based on whether or not the journalist involved expressed personal opinions or tended to be more discussive in their interactions [93]. 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., [41]). 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.

One particular aspect of this work is the use of prediction models to study the impact that different features have in the attention to news-tweets. While this is a common practice in the literature, our work differs in that its focus is not only on news-tweets' features and their importance for engagement, but also in understanding how such importance can be interpreted to offer more practical advice to journalists. We used three regression models namely RLR, RF, and GBT to predict attention to news-tweets, and a common measure of error, MSE, to decide which model to use further. The MSE, although widely used, can be difficult to interpret and the decision on whether an error is good or not, is highly dependent on the particular conditions of the problem.

In our case, MSE was used merely to discriminate among models based on its average value over 10 rounds. The lowest errors were shown by GBT – except for the business category – and as such, was the selected model. After using GBT for engagement prediction, we extracted the top-10 news-tweets' features based on their importance to the predictions, and these results were further interpreted to offer practical advice for journalists (see Section 4.2.2). This process resulted in a list of guidelines that apply to six different news categories and corporate Twitter accounts. The *goodness* of such guidelines has been validated by journalists and news readers who consider them coherent with actual journalistic practices and news consumption tendencies in Twitter.

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 [61]. 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, 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 [1] 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 UK 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.

In the next chapter we will present an analysis that focuses, rather than on the journalists, on Twitter news audiences: we will analyze what drives users to read news-tweets, and present our approach to predicting news consumption.

CHAPTER

FIVE

Modeling and Predicting News Consumption on Twitter

Daily newspaper reading and the viewing of national TV news, have been traditionally correlated with a civic obligation to stay informed about current events [106]. For older people, the daily habit of following the news and this civic sense seems to have persisted as news has moved online and onto social media. For younger people, news consumption has, perhaps, become a more personalized activity to gather information about events that directly affect them or, more a matter of social interaction, as they use news items in online conversations with friends and colleagues [1, 16]. Indeed, young people seem to be more socially engaged with news media on a national level, than with local/regional media [45]. In general, the shift toward online news consumption has fundamentally impacted consumption patterns themselves [112] and it has been found that social media use positively predicts total news consumption time [56].

For news providers, it is critically important to understand the dynamics of news consumption in this new social media context. They need to understand the beliefs, motivations and attitudes that drive news consumption in social media users. Such insights should allow news providers to better serve their audiences and motivate future news consumption [110]. Indeed, users who read the news on social media often play an important role in modern journalism, not only as direct consumers, but also as gatekeepers, with almost half the social media users sharing and reposting news stories, images, or videos, and discussing news issues or events online [133]. Also, journalists who post/share their news and interact with social media audiences, can use their knowledge of such factors to influence people's satisfaction [82], which may in turn, have a positive impact in news consumption.

Social media is a particular environment in which news consumption is not a passive activity, but rather one in which users interact with news providers, form communities, express their interests, ask questions, or request more information, via mechanisms such as shares, likes/dislikes, retweets, or mentions.

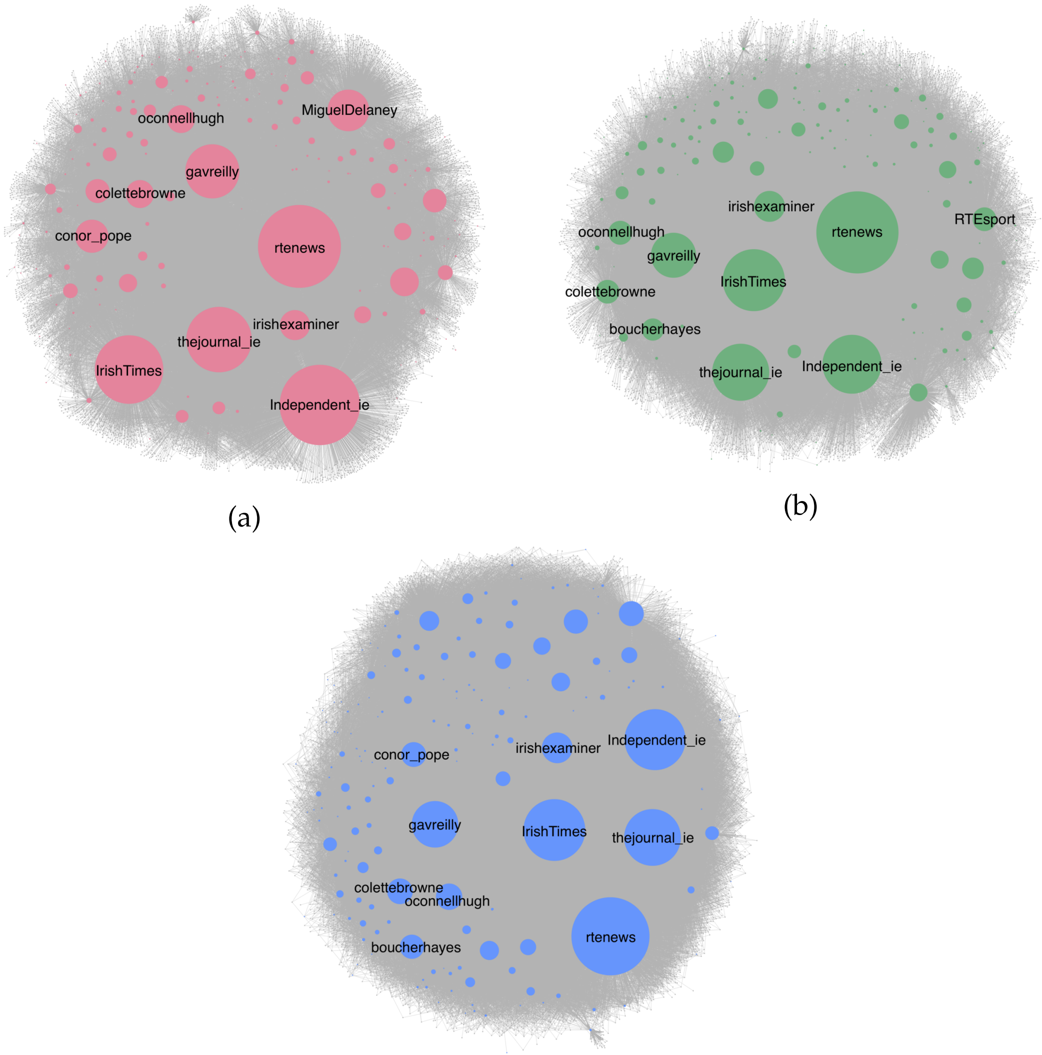
Figure 5.1 illustrates some of these dynamics of news consumption, as seen in the interactions of an Irish Twitter audience with 200 journalistic accounts[1] for a description of the dataset). Each graph represents news audiences (nodes in gray), journalists (nodes in color), and their interactions (edges from news audiences to journalists nodes). The larger the node, the more interactions it has received. We have labeled the top-10 journalists/news outlets in each graph, according to the number of interactions received from the audience. In Figure 5.1a, the edges represent the news audience mentioning journalists in their tweets. Corporate accounts such as *@Independent_ie*, *@rtenews*, *@IrishTimes*, and *@thejournal_ie* receive most of the audience mentions; however, individual accounts including political journalist *@gavreilly* and sports journalist *@MiguelDelaney*, receive considerable attention as well. In terms of retweets (see Figure 5.1b), corporate accounts *@rtenews* and *@IrishTimes* receive the bulk of them, with political journalists *@gavreilly*, *@Oconnellhugh*, and *@colettebrowne* also being important targets of audience retweets.

Overall, Irish news audiences tend to interact (i.e., mentions+retweets) more with corporate and political journalists (see Figure 5.1c).

In this chapter, we propose the *Twitter News Model (TNM)*, a computational data-driven approach to predict and explain people's consumption of news by analyzing their interactions with journalists and news-tweets. Our *Twitter News Model (TNM)* is based on the Motivational Consumption Model (MCM) [91], developed to better understand the consumption of "conventional" news. We apply the TNM to an empirical study of news consumption on Twitter, designed to (i) reveal *what* drives users to consume news, and (ii) predictively relate users' news beliefs, motivations, and attitudes to their consumption of news.

In the remainder of this chapter, we present the theoretical background on which our work is based (Section 5.1), instantiate our Twitter-specific model for news consumption (Section 5.2), and predict news consumption on Twitter (Section 5.3). We conclude in Section 5.4.

[1]These Irish journalistic accounts are the same accounts described in Section 4.1.1; however, the data used in this chapter spans a different time period



(a) (b) (c)

Figure 5.1: Dynamics of news audiences' interactions with journalists on Twitter. Gray nodes represent the audience and colored nodes represent journalists. Each edge represents a member of the audience (a) mentioning a journalist, (b) retweeting a journalist's tweet, and (c) overall interacting (mentioning + retweeting) with a journalist. The nodes' sizes are proportional to the number of (a) mentions, (b) retweets, and (c) overall interactions they received. On each figure, we have labeled the top-ten most popular Twitter accounts according to the corresponding metric.

5.1 Modeling News Consumption

Our aim in this chapter is to model news consumption on Twitter. Our starting point is the prior work that has modeled motivational aspects of human intention and specific models of news consumption in "conventional" news media. Hence, we explore a psychological theory (i.e., the Reasoned Action Model) that has been applied to the consumption of printed and digital news (i.e., Motivational Consumption Model). In

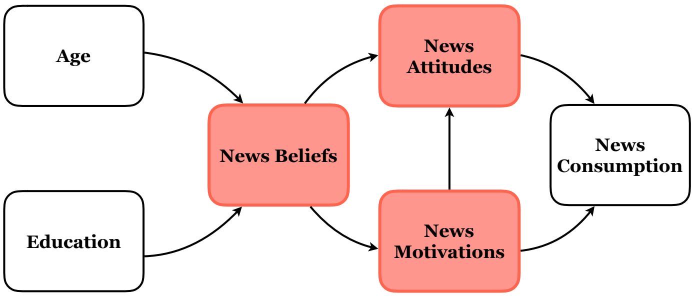


Figure 5.2: Motivational Consumption Model (MCM). Various demographic factors are seen as influencing news consumption via the mediators of news beliefs, news motivations, and news attitudes (shown in red). Source: Adapted from [91].

this section, we review these previous models before discussing how they can be operationalized to the consumption of news on Twitter (see section 5.2).

5.1.1 Reasoned Action Model

Fishbein and Ajzen's Reasoned Action Model (RAM) [48] maintains that there are three main predictors of intention to engage in a behavior, namely attitude, social norms, and self-efficacy. *Attitude* deals with a person's orientation towards performing the behavior. *Social Norms* consider the normative pressure perceived by the person to perform a given behavior in context. *Self Efficacy* relates to the behavioral control the person perceives themselves to have over the target behavior. RAM has been used to understand specific situations where people have manifested intentions to act. For instance, in political science, RAM has been used to model voters' attitudes towards candidates and political parties, and how these attitudes impact polls and subsequent voting behavior; in public health research, RAM has been used to understand the key beliefs that influence individuals' health care utilization [74].

5.1.2 Motivational Consumption Model

Lee and Chyi have adapted the RAM framework to understand news consumption in their Motivational Consumption Model (MCM) [91]. In the MCM, a variety of demographic factors are seen as influencing news consumption via the mediators of news beliefs, news motivations, and news attitudes (see Figure 5.2).

This adapted model backgrounds the effects of RAM's social norms and self-efficacy. Social norms play less of a role in news consumption because of the broadly shared belief that it is one's responsibility to follow the news, to stay informed. Self-efficacy also plays a lesser role as people can easily control their own news consumption given the ubiquitous access they have to news content via a variety of to-hand digital devices [91].

In their model, Lee and Chyi go on to describe and operationalize the three key mediators as follows:

– *News beliefs:* refer to the *value* that news has for people, that news can be conceived to be a means to empower and mobilize the public. These news beliefs were operationalized using 7-point Likert scale ratings of survey questions that asked participants (i) how important the news is to you, and (ii) whether being informed was empowering.

– *News motivations:* refer to the *reasons* driving people's consumption of news. For instance, some people consume news because it helps them keep up with current events that are topics of conversations in their social circle or it allows them to make informed decisions in their daily life; whereas others consume news as a source of entertainment. These news motivations were operationalized by combining ratings from five survey questions that asked participants whether they consumed news (i) to find out what is going on in the world, (ii) to keep up with the way your government functions, (iii) to make yourself an informed citizen, (iv) to gain important new information, and (v) to fulfill your need to know.

– *News attitudes:* reflect positive views of the news, that see news consumption as an enjoyable and advantageous behavior. These news attitudes were operationalized by survey questions that asked participants if they agreed with the following statements: (i) getting the news is enjoyable to you, and (ii) getting the news is advantageous to you.

In their survey, Lee and Chyi gathered data from a US-nationwide sample of 1.1K American adults that varied in several demographic features (e.g., gender, age, income, and race). The survey asked people about their news consumption – reading, watching, or listening – on several news sources, including *New York Times*, *CNN*, and *Google News*, in both traditional and digital formats. The responses were analyzed using a regression method. Based on the MCM, demographics such as age and education were found to positively influence news beliefs, and more positive news beliefs were found to lead to better news motivations and news attitudes, which in turn resulted in more frequent news consumption.

In the present chapter, we adapt the Motivational Consumption Model to apply it to news consumption on Twitter, by developing new operationalizations of its mediators that were appropriate to this social media context.

Mediator	Measure	Description
News Beliefs B	$B(u) = \begin{cases} 1, & \text{if } N_u \geq R_u \\ 0, & \text{otherwise} \end{cases}$	How valuable are news for user u?
News Motivations M	$\forall c \in C, M(u,c) = \frac{i_{uc}}{I_u}$	What motivates user u to consume news?
News Attitudes A	$A(u) = \frac{1}{N_u} \sum_{m_u} Polarity(m_u)$	What is user u's approach toward news?
News Consumption C	$C(u) = I_u$	How much does user u consume news in Twitter?

Table 5.1: New operationalizations for Twitter. Per each mediator in the MCM, we propose a measure in TNM within the context of Twitter. News beliefs B is operationalized as the proportion of retweets and mentions resulting from the interaction user-journalist. If the user mentions more than they retweet $B = 1$, $B = 0$ otherwise. News motivations M is measured by the proportion of interaction a user has with each news category (i.e., does a user look for entertainment, information for decision-making, or to have conversation topics). News attitudes A is represented by the overall polarity of the tweets or interactions between a user and a journalist. News consumption C is given by the number of interactions (mentions + retweets) of a user, in which a journalist or news organization is involved.

5.2 News Consumption on Twitter

Lee and Chyi applied their Motivational Consumption Model to the consumption of news found in newspapers (i.e., either printed or online), on news aggregators (e.g., *Google News*, *Yahoo News*), and on network Sunday talk shows [91].

In the present section, we aim to model news consumption on Twitter, which constitutes quite a different context. On Twitter, the dynamics of news consumption differ in that the users do not need to actively buy/access the news but rather browse their *timelines* at any time/place, and find news posts distributed (i.e., tweeted/retweeted) by Twitter accounts from followed journalists or newspapers, or that appear in the timeline because one or more of the other user's followees (e.g., family, friends, or public figures) has tweeted or retweeted them. In this social media context, our conception of news beliefs, news motivations and news attitudes needs to be adapted with Twitter-specific instantiations of these notions. We will call our adapted model the Twitter News Model (TNM) [126].

5.2.1 News Beliefs

News beliefs refer to the value that the news has for people (i.e., whether people find news to be important). In Twitter, users express the value that a news-tweet has for them by retweeting it, liking it, or mentioning the author of that tweet.

To quantify the value of news (B) to a Twitter user u, we use the proportion of retweets and mentions resulting from the interaction between the user and news-tweets, journalists, or newspapers. If a user u mentions journalists in her/his tweets more or as many times as she/he retweets news-tweets, then we can assume they value news on Twitter to a greater extent (see Equation 5.1).

$$B(u) = \begin{cases} 1, & \text{if } N_u \geq R_u \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

where N_u is the total number of tweets in which user u has mentioned a journalist or a newspaper, and R_u is the total number of news-tweets that user u has retweeted, in the same period of time.

Mentions, in comparison to retweets, require more effort and time. Mentions help users create awareness of a tweet's author and even spread the author's tweets to new audiences, including non-followers, thus increasing diffusion. By mentioning a journalist or a newspaper, user u might help other users find more tweets from such accounts on a timely manner [172].

5.2.2 News Motivations

News motivations refer to the reasons that drive people to consume news (e.g., to become an informed citizen). In Twitter, these motivations can be operationalized by tracking how a user interacts with different news categories (e.g., a user who only interacts with tweets about sports may be mainly motivated by entertainment).

To quantify news motivations (M) for news consumption on Twitter, the observed interaction behavior of each user u with tweets/journalists for different news categories is computed as follows:

$$\forall c \in C, M(u, c) = \frac{i_{uc}}{I_u} \tag{5.2}$$

where $C = \{$*sci & tech, sports, politics, business, breaking news, lifestyle, corporate*$\}$ represents the main content categories found for news[2], i_{uc} is the total number of interactions (i.e., retweeting of a news-tweet and/or mentioning a journalist or newspaper) by user u in a particular news category c, and I_u is the total number of interactions (i.e., retweets and mentions) by user u across all news categories. $M(u, c)$ falls in the range of $[0, 1]$

[2]A more detailed explanation on how we classify tweets into these categories can be found in Section 4.1.2.

where 0 means user u did not interact with category c and 1 means that all interactions of user u were with category c.

It should be noted that using news categories to identify motivations is quite a coarse-grained approach, as precise motivations within a category may vary [91]. A user only interacting with lifestyle news might be motivated by reasons of entertainment or social interaction. A user interacting mostly with business or political news, may be looking to make specific business decisions or just to track broad societal changes. So, while using news categories broadly partitions motivations, a more fine-grained analysis of these motives remains to be discovered. Finally, it should also be noted that the way users interact with the news, is known to vary by news category. For example, Irish Twitter users consuming sports news engage more with news-tweets that contain mentions, while those who engage with science and technology news are influenced more by the temporal arrival of the tweets (see Section 4.2.2).

5.2.3 News Attitudes

News attitudes refer to people's overall orientation to the consumption of the news (e.g., whether it is viewed as an enjoyable behavior). In Twitter, these attitudes can be operationalized by tracking the polarity of users' tweets. To quantify the news attitudes (A) of a user u to news consumption on Twitter, the following is computed:

$$A(u) = \frac{1}{N_u} \sum_{m_u} Polarity(m_u) \tag{5.3}$$

where $Polarity(m_u)$ is the polarity score (i.e., positive or negative) of each tweet in which user u mentions a journalist or newspaper, and N_u is the total number of tweets with mentions – of a journalist or a newspaper – posted by user u.

In our operationalization of news attitudes, we do not consider retweets, only the users' original tweets in which they mention journalists or newspapers. We make this distinction because news attitudes represent a person's tendency to react favorably towards the news [91]. Retweeting a news-tweet may signal users' interests; however, news-tweets are not tailored by the users themselves but rather by the source of the tweet (i.e., a journalist or a newspaper). Through mentions, a person can express themselves, in their own words, thus giving a better indication of their news attitudes.

5.2.4 News Consumption

News consumption is generally measured by the frequency with which people read/watch/listen to news. In Twitter, we measure a user u's news consumption (C) by the number of interactions between the user and news-tweets, journalists and newspapers, which are defined as follows:

$$C(u) = I_u \tag{5.4}$$

$$I_u = N_u + R_u \tag{5.5}$$

where I_u is the number of interactions of user u, N_u is the total number of tweets in which user u mentions a journalist or newspaper, and R_u is the total number of news-tweets that user u retweeted.

We consider that both retweets and mentions give us an indication of the users' overall tendency to consume news. For example, by retweeting a news-tweet, a person may indicate interest in a topic or a specific news article, while by mentioning a journalist or a newspaper, a person may express an interest in the work of the journalist or the coverage of the newspaper. Therefore, we use these interactions as a proxy for news consumption.

In conclusion, Table 5.1 summarizes the components of the Twitter News Model (TNM) with its news beliefs, motivations and attitudes as they are characterized in the context of Twitter, along with a Twitter-specific definition of news consumption. In the next section, we describe how the TNM can be leveraged to predict and explain news consumption.

5.3 Realizing the Twitter News Model

In this section, we apply the proposed Twitter News Model (TNM) to a dataset of interactions between users and journalists/newspapers to elucidate the dynamics of news consumption on Twitter. Specifically, this empirical study is designed to (i) reveal *what* drives users' consumption of news on Twitter, and (ii) predictively relate users' news beliefs, motivations, and attitudes to their consumption of news on Twitter. We characterize the various components of the model – news beliefs, motivations and attitudes – and relate them to the computation of consumption.

# News media accounts	200 (117 individual and 83 corporate)
# Interactions	352,500 (tweets with mentions and retweets)
# Users	16,683 unique users
Time period	from August 10th to December 10th, 2017

Table 5.2: Dataset statistics.

5.3.1 Dataset

For this task, we focus on the 200 Irish journalistic Twitter accounts described in Section 4.1.1. However, the dataset used in this chapter spans a different time period: four months from August 10th to December 10th, 2017. This period covers a series of news events including Hurricane Harvey, the North Korea's launch of missiles, and the Las Vegas shooting. In total, 16.6K users retweeted and/or mentioned one or more of the 200 journalists and newspapers under study, producing a total of 352,500 (μ = 22, σ = 48) interactions.

As our aim is to understand what makes audiences interact with news-tweets, journalists, and newspapers in Twitter, we focus on the 352.5K interactions (retweets and tweets with mentions) posted by the 16.6K users. We do not analyze the tweets posted, nor the interactions started by our 200 journalistic accounts.[3] Table 5.2 summarizes the main statistics of our Twitter corpus.

5.3.2 News Beliefs

News beliefs are defined in terms of user mentions of journalists and newspapers, more than merely retweeting their news-tweets (see Equation 5.1). In our corpus, we found a total of 52% ($N = 8.7K$) of users who post at least as many tweets with mentions as they do retweets (i.e., $B(u) = 1$). Of these, 4.3% ($N = 713$) of users post retweets and mentions in a 50/50 ratio.

Users mention journalists or newspapers for different reasons, including discussing topics of interest (e.g., *@IrishTimes Each year the housing prices rise. How do we rebuild at a rate of such increase? We can't. Stop housing...*), expressing opinions and concerns (e.g., *@gavreilly 2017 results at home terrible. Two draws and defeat from Wales, Austria and Serbia cost great chance to top the group*), or simply to establish conversations (e.g., *@MiguelDelaney Really great piece Miguel, fair play. I think the FAI need to import some bright minds from Germany...*). Users who mention news providers more often, spend

[3]Note that our data collection only includes the interactions between users and journalistic accounts, during the time-period under study; these users' interactions with other users (i.e., non-journalists), were not collected.

time and put effort into tailoring their tweets, which could indicate that news (and its discussion), is valuable to them. The remaining 48% ($N = 7945$) of users in our corpus, retweet more than they mention (i.e., $B(u) = 0$); and 54% ($N = 4330$) of them do so with a retweets-mentions ratio of at least 80/20.

These findings indicate that for the majority of users in our dataset, news is important to the extent in which they feel the urge to not only spread tweets from journalistic sources, but to express their own opinions, post their own tweets, and become active participants in the discussion of news.

5.3.3 News Motivations

News motivations are what drive people to consume news. In order to better understand the reasons behind the news consumption for each user in our corpus, we separated her/his interactions (i.e., retweets + mentions) by news category.

Our model proposes that the proportion of interactions a user has with a given news category, indicates their motivation for consuming news. For example, if a user only retweets sports news-tweets and/or mentions sports journalists, then this indicates they are largely consuming news for entertainment purposes, in contrast to the motivations of a user that only interacts with news-tweets/journalists in the business category.

To identify the news category of the tweets posted by the users in our corpus, we followed the process described in Section 4.1.2. We first separated the 200 Twitter accounts into individual and corporate. Corporate accounts (e.g., *@Independent_ie*) post tweets that span all news categories, while individual accounts (e.g., *@conor_pope*) tend to focus on a specific area, such as sports or business. Three annotators judged the news category of each individual account based on (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. The news categories considered are business, lifestyle, science and technology, breaking news, politics, and sports. We follow this human annotation approach, so we could obtain high quality labels that can be used as ground truth in related future tasks. The distribution of accounts across the six news categories is shown in Table 4.3. We group the remaining 83 Twitter accounts under a category that we will call *corporate*, as they cannot be classified under any specific news category. The Fleiss' Kappa inter-rater agreement over all news categories is $\kappa = 0.51$.

The interactions between a user and a news category were calculated as follows: if user

u mentions a journalist or retweets a news-tweet from a journalist who belongs to news category c (e.g., sports), then we add 1 to user u's interactions with c. We follow the same procedure for all user u's tweets. We then normalize the total number of user u's interactions per category (i_{uc}) by the total number of user u's interactions over all news categories (I_u). We repeat this procedure for all users (U) in our corpus.

Using this analysis a number of interesting regularities can be found in the way users interact with news categories. Figure 5.3 shows the number of users in our corpus who interact with one or more news categories. Out of the 16.6K users, $N = 4887$ (29.3%) interact (i.e., retweet and mention journalists) solely with one news category. The three most popular categories among these users are corporate (3383 users only interact with corporate accounts), sports ($N = 1134$ users), and politics ($N = 220$ users). A further third of this audience ($N = 5350$ or 32%) interact exclusively with two news categories. The most popular pairs of news categories are corporate-sports ($N = 1882$ users only interact with corporate and sports accounts, with $N = 717$ users interacting more with sports than with corporate accounts, and $N = 1165$ doing the opposite), corporate-politics ($N = 1847$ users), and corporate-breaking news ($N = 843$).

Of the remainder, $N = 3358$ (20.1%) of users interact with three different news categories. The most popular triplet for these users is politics-corporate-breaking news ($N = 1074$ users only interact with these three news categories). From four to seven news categories, we observe a sharp decrease in the number of users. Only, $N = 1796$ (10.8%) users interact with four news categories, $N = 877$ (5.3%) with five, $N = 338$ (2%) with six, and $N = 79$ (0.5%) with all the seven news categories.

These findings show that, for this user cohort, the majority of people who follow news on Twitter tend to interact with 1-3 news categories. Corporate accounts (e.g., *@Independent_ie*) seem to be a major focus of user attention and interactions, independently of the other specific news categories that a user follows. One possible reason for this behavior, may be that audiences in Twitter are consuming news as a source of general knowledge, and/or to keep up with events of public interest. Sport and politics are the two other news categories that receive considerable attention. This result evidences more niche interests, in which users have a news category of choice, that presumably fulfills more fine-grained motives such as entertainment (e.g., sports news), or keeping an up-to-date knowledge on government-related affairs (e.g., political news).

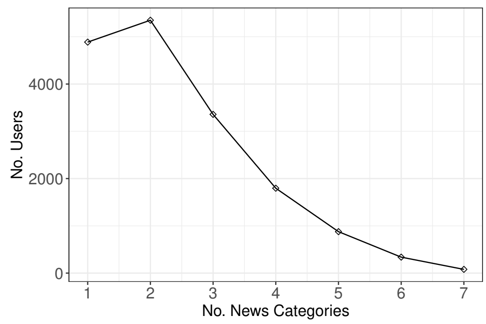


Figure 5.3: Audience and number of news categories they interact with.

5.3.4 News Attitudes

News attitudes reflect users overall orientation to the consumption of news and, in our model is measured by extracting the polarity of the users' mentions (i.e., tweets in which users mention journalists or newspapers). This means that for a given user, we first need to determine the polarity of her individual tweets and then average the polarities of all her tweets. These aims were achieved in three steps: (i) all the tweets of concern were transformed into vector representations using a word-embedding technique, (ii) a polarity classifier was trained on a corpus of tweets with known polarities, (iii) this classifier was used to find the polarity score for each tweet of a given user and then these scores were averaged over all the user's tweets. Each of these three steps is detailed further as follows.

(i) Transforming tweets into word vectors. Several word-embedding techniques have been proposed to transform words into vector representations (e.g., *Word2Vec* [111], and *GloVe* [131]). The tweets of concern here were processed using GloVe, an unsupervised learning algorithm that provides a collection of word vectors pre-trained on 2B tweets/27B tokens [131]. These vectors are available in different dimensionalities: 25, 50, 100, and 200 dimensions.[4] Using the 200-dimension vectors, we found the GloVe for each word in a tweet,[5] and averaged all these word vectors to obtain the overall tweet vector.

(ii) Training a polarity classifier. To train a tweet polarity classifier[6], we used the word vectors (i.e. GloVe) representation of the SemEval-2017[7] dataset [147] consisting of 19.9K tweets. These tweets were posted between 2013 and 2016 and had identified

[4]https://nlp.stanford.edu/projects/glove/

[5]If the word does not have a GloVe representation, we set to 0 all the 200 dimensions of the corresponding vector

[6]scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7]SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems, organized under the umbrella of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics.

polarities: 8157 positive, 8101 negative, and 3079 neutral tweets. In training the classifier, only positive and negative tweets were used; we used upsampling on the negative tweets to ensure a balanced dataset (i.e., 8157 positive and 8157 negative tweets). The classifier was trained to predict the probability that a given tweet belonged to the positive class (i.e., the closer to 1 the more probable the tweet was positive, the closer to 0, otherwise). Using 5-fold cross validation, the final classifier had an average F1 measure of $F1 = 0.78$. The SemEval datasets are widely used for sentiment analysis tasks [148, 29, 97]. An interesting extension to polarity classification, would be to analyze the presence of other emotions known to be expressed in tweets, such as anger, joy, surprise, or sadness, as done in [122, 123].

(iii) Aggregating the polarity of a user's tweets. For each user u in our corpus, we extract the polarity for each individual tweet in which user u mentioned a journalist or a newspaper. As a final step, we average the polarity scores of all user u's tweets (i.e., tweets with mentions) and assign that average as the overall user's polarity.

Out of the 16.6K users in our corpus, $N = 13,097$ (78.5%) show a positive attitude ($A(u) \geq 0.5$) toward journalists and newspapers; of these users, $N = 190$ had a strongly positive polarity average ($A(u) > 0.8$). The remaining $N = 3,578$ (21.5%) users show an average attitude that is more inclined toward a negative polarity ($A(u) < 0.5$); of these users, $N = 62$ had a strongly negative polarity average ($A(u) < 0.2$). Although we find some news consumers to be more negative than positive, the majority of people in our corpus show a positive attitude toward news and news providers.

5.3.5 News Consumption

News consumption is captured by the total number of interactions (i.e., retweets + mentions) that user u has with news-tweets, journalists, and newspapers. Here, we found that on average, each of our 16.6K users interacts 21.3 times ($\sigma = 48$) with news-tweets/news providers over a four-month period. Of these users, $N = 508$ (3%) interact 100 times or more, $N = 9,303$ (56%) users interact 10 times or less, and the remaining $N = 6,874$ (41%) of users interact between 11 and 99 times in the four-month period spanned by our data collection. These results show that 41% of the Twitter users in our dataset, turn to news on a regular basis. Some of these users tend to do so approximately once a day. A small group (3%), interacts with news-tweets and/or news providers on average 4 times a day, with some users retweeting and mentioning as much as 13 times a day, on average. For more than half of our news consumers (56%), interacting with news-tweets and news providers seems to be more of an sporadic activity.

5.3.6 Predicting and Explaining News Consumption

Having captured the behavior of our news consumers using the Twitter News Model, we turn our attention to building a predictive model to better understand the drivers of news consumption. We cast the news consumption prediction as a regression task in which our goal is to predict, based on users' features, their corresponding news consumption. Our aim is to train a machine learning regressor to predict the news consumption, which in our model corresponds to the *number of interactions (i.e., retweets + mentions)* between each user and news-tweets/news providers (cf. Section 5.2). We use the z-score of the users' interactions as our target variable.

User representation. Each user in our data collection is characterized using a feature vector. Users' features are extracted using three methods described as follows:

- *Twitter News Model (TNM).* We construct users' vectors using the TNM mediators; these are 9-dimensional vectors with one dimension for news beliefs (i.e., whether the user tends to mention more than retweet), seven for news motivations (i.e., one per each news category and one for corporate accounts), and one for news attitudes (i.e., the polarity score for the user).

- *Word-Embeddings (WE).* Using the same procedure described in Section 5.3.4, we construct feature vectors based on the content of the users' interactions. For each user, we averaged the GloVe word-embeddings [131] of her/his retweets and mentions to obtain a 200-dimensional vector that is fully based on the content of the user's tweets.[8]

- *Twitter News Model+Word-Embeddings (TNM+WE).* Vectors combining the Twitter News Model and the word-embeddings representations of the tweets. This results in vectors of 209 dimensions: 9 from the TNM and 200 from the word-embedding representations.

Experimental setup. We split our data into training and test sets using an 80%/20% ratio. We explore three different regression models, namely Random Forest (RF) [60], Gradient Boosting Trees (GBT) [60], and Extremely Randomized Trees (ET) [54] and gridsearch their best hyperparameters using 10% of the training set for validation. We conduct 25 rounds of experiments and measure the prediction quality of our models using the *Mean Squared Error (MSE)* metric. The results reported are the average MSE on the test set over the 25 rounds. Based on this exploration, the model that exhibits the best predictive performance for all three user representations is Extremely Randomized Trees (ET),[9] and therefore, the one chosen for our task. ET fits a number of randomized

[8] Unlike in the calculation of News Attitudes (Section 5.3.4), these vectors also take into account users' retweets, not just their original tweets with mentions.

[9] scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

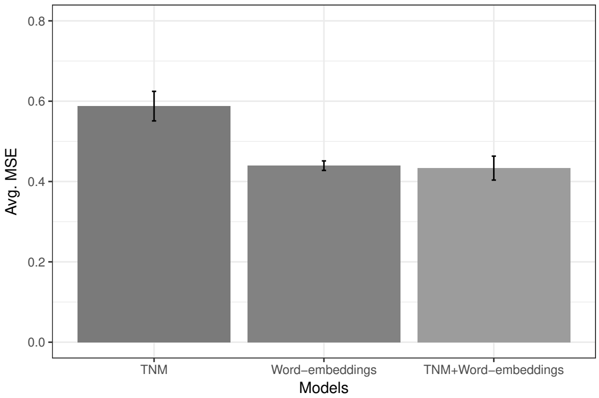


Figure 5.4: Average MSE values for the different models in predicting news consumption, showing 95% confidence intervals (as these are error values, the lower the better). MSE values are normalized to the [0,1] interval.

decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The best hyperparameters found for our task are 1000 estimators (n_estimators=1000) and a minimum of 3 samples required to be at a leaf node (min_samples_leaf=3).

We train three separate ET regression models, one per each set of features extracted, namely TNM, WE, and TNM+WE. Figure 5.4 shows each model's average MSE, whose values correspond to 0.587 for TNM, 0.439 for WE, and 0.433 for TNM+WE.

Interpretability vs. predictive power. As we observe from Figure 5.4, the model trained using the TNM features has the highest error. In terms of predictive power, the Word-embeddings and TNM+Word-embeddings models are better. However, the model based on the Twitter News Model has one overriding benefit; namely, that it is fully interpretable, in that we know the semantics of the corresponding TNM feature. Notably, the hybrid TNM+WE model, combines WE's predictive power with TNM's interpretability, allowing us to produce predictions of comparable quality (to those produced by the WE model alone), while identifying the meaningful features that explain these predictions. We cannot obtain the same interpretability from the WE model on its own given its nature, i.e., lack of semantics of the word embeddings.

From our hybrid model (TNM+WE), we extracted the features that were most important (to the ET regressors) in the prediction of news consumption. Interestingly, the top-2 features are TNM's news motivations, specifically, users' interaction with the news categories business and breaking news. In rank 3, we find a content feature (i.e., from the word-embeddings representation), followed by two other features of news motivations in ranks 4 and 5, the interaction with corporate accounts and the politics news category. The remaining features in the top-20 are content features, with the exception of the interaction with lifestyle news in position 13. For our hybrid model, the TNM's features considered less important for prediction were news attitudes (i.e., the polarity of users' interactions), users' interactions with the news categories of science and technology and sports (remaining features of news motivations), and news beliefs (i.e., whether the users tend to mention more than to retweet).

5.3.7 Discussion

By applying the Twitter News Model to our dataset, we were able to extract interesting insights into Twitter users' news consumption. We found that the most telling features in the prediction of users' news consumption are *news motivations*, or the reasons that drive them to consume news. This finding is consistent with the findings of the Motivational Consumption Model, in which the authors explain that news motivations influence both news attitudes and news consumption. While we are now able to indicate that users' interactions with certain news categories (e.g., business and politics) are useful in predicting their news consumption, further analyses are necessary in order to explain whether this means they consume more or less news than if they interacted with other news categories.

Although most of the observed effects can be attributed to news motivations, *news attitudes* also play a role. We found that the cohort in our sample tends to approach journalists with a positive attitude, as reflected in the polarity of the tweets in which they mention journalistic accounts.

The features reflecting *news beliefs* appeared to have the least importance in the prediction of news consumption. This finding is in line with the MCM, in which *news beliefs* are seen as having a direct effect on news motivations and news attitudes, and only through these, showing an effect on news consumption itself. This result might suggest that users who are active participants in the news conversation in Twitter, implicitly convey the value that news has for them, by virtue of such interactions (which explicitly reflect their news motivations and attitudes).

Using the TNM features alone to learn and predict users' news consumption, did not result in the best model; as the predictions had the highest error. Other approaches, including those using content-based features, for example word-embeddings, have higher predictive power. However, the TNM models help make results more interpretable. This characteristic is essential, particularly in environments where beyond predicting news consumption, is the knowledge obtained in the process that matters, for example, if journalists learn what makes Twitter users consume news, they can apply the knowledge to future interactions and thus, keep their audience and/or possibly attract more readers. For tasks such as news recommendation and personalization, journalists can use these results to segment their audience by creating *personas* based on the news consumption behavior. News could then be tailored and targeted to each persona, highlighting the aspects that will make it more likely to be consumed.

5.4 Summary

We have presented our Twitter News Model (TNM), a computational approach to modeling the news consumption behavior of Twitter users. Our approach is based on the Motivational Consumption Model (MCM), developed to model the consumption of news in conventional media such as newspapers, news aggregators, and TV (see Figure 5.2). To the best of our knowledge, this is the first time that MCM has been adapted to this social media context. We started by discussing the theoretical background on which the MCM was based, described how we used the theory to instantiate our Twitter-specific model for news consumption (TNM), showed the usefulness of our model by applying it to a real-world scenario in which we learn the dynamics of news consumption in Twitter, and used this knowledge in a news consumption prediction task. Our findings reveal that news motivations, followed by news attitudes, and news beliefs, impact users' behavior of news consumption in Twitter. In the future, these findings could be applied to tasks such as user segmentation, and to inform strategies that journalists might use to promote their news on Twitter, so that they can better understand the dynamics of news consumption in social media.

In the next chapter, we present two practical approaches aiming at bridging the gap between technical findings and journalistic practices.

CHAPTER

SIX

Designing Tools for Journalists

Thus far, we have discussed the dynamics of news on Twitter (see Chapter 3), uncovered the features that predict audience engagement for news-tweets (see Chapter 4) and identified the drivers of audience news consumption (see Chapter 5). In this chapter, we address the more practical goals of turning these analyses into actionable tools that aim to assist journalists in different aspects of their work.

We start by presenting a set of guidelines informed by the analyses described in Chapter 4. These guidelines are specific enough to enable news providers to design innovative strategies for improving audience engagement on Twitter (see Section 6.1). In Section 6.2, we present a recommendation system designed to facilitate the decision-making process facing a news editor in composing a headline. This system is informed, in part, by the analysis presented in Section 4.2.

6.1 Guidelines: Helping Journalists Gain Attention

The analyses described in Chapter 4 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 of Chapter 4 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).

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 4.9a and 4.9b, and the discussion presented in Section 4.2.2, 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 4.2.2, 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.

To formulate guidelines informed by our analyses in Chapter 4 and the decision trees that we describe later in this chapter, we do the following: for each group of features and news category in turn, we use as a guide the corresponding feature importance illustrated in Figures 4.9a and 4.9b, then we look for features belonging to that group within the decision tree produced for that corresponding news category, and extract the values of the features that lead to larger values of engagement in the leaves of the tree. Using this procedure, we produce guidelines that, beyond suggesting the interaction of features, provide more detail on their usage so as to assist journalists in their practical usage of the platform.

Experimental setup. In these new analyses, the tweet corpora (1.06M tweets from Ireland and 1.2M from the UK)[1] are separated into individual and corporate accounts, as clearly the guidelines should differ for each type. Then, as described in Section 4.1.2, we split the individual accounts by news category, and due to our focus on interpretability rather than on predictive power, use the full tweet set in each category to train individual decision trees (as opposed to having training and test sets as was the case in Section 4.2.1). The problem is cast 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 developed from manually inspecting these outputs and the feature importance results discussed in Section 4.2.2.

Results. 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

[1] Note that this is the same dataset described in Table 4.1 (years 2015-2016) . This is required because the guidelines shown in this chapter are partly based on the results obtained in Section 4.2

working. *Specific guidelines* address factors that are important within a particular news category (e.g., sports versus business). As we will see in the remaining of this section, the latter guidelines are perhaps the most significant, as they suggest very specific interventions that individual journalists can take to promote their news.

6.1.1 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 for these accounts; however, on any day after 5:00 p.m. tweets can receive a slight increase in audience engagement.

6.1.2 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 attracted by 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 the popularity of the journalist.

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.

Discussion. In this section, we presented a set of guidelines aimed at assisting journalists in their day-to-day use of Twitter. Specifically, we described how different guidelines apply according to the Twitter account type (i.e., corporate or individual) and the category of news (e.g., sports vs. politics). Current journalistic practices on Twitter (see Chapters 2 and 3) show little evidence to suggest that journalists have an informed strategy, or formal instruction provided by their news organizations on how to approach Twitter as a tool for news dissemination. A cursory glance at Twitter strategies of journalists working for major news media (see Chapters 3 and 4), 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). While we do not propose the guidelines to replace any newsrooms' rules for Twitter usage, we believe that data-based strategies, such as ours, could be highly beneficial for journalists in their quest for an efficient use of social media platforms for news distribution.

Interpretation of results. To create the guidelines, we fit one decision tree to the full set of tweets per each news category. We did not split the data into training and test sets, but instead used the full set of news-tweets to fit decision trees whose leaf nodes correspond to different values of retweets. We did not use any error measure for the results given by these decision trees, as our objective was not to compare such results to those obtained with other methods (i.e., the ensemble methods used in Chapter 4), but rather to show that prediction models can provide researchers with data that can be of help to design practical tools.

Once the decision trees were fit to the full set of tweets for each news category (i.e., six decision trees in total), we manually traversed each tree to extract the features and corresponding values that lead to leaf nodes with the highest number of retweets (i.e., our proxy for engagement).

The advantage of such approach, is that by fitting the model to the data we can learn specifics about their behavior and feature interaction, and produce detailed guidelines that can be actionable and put in practice by journalists. One limitation of our approach is that these results are not generalizable, i.e., each decision tree is fit to the specific input data. Also, given that we did not use these trees for prediction *per se* but for interpretation, we cannot argue that the obtained predictions are better than the ones obtained with ensemble methods.

We present these guidelines as, to the best of our knowledge, one of the first attempts to bridge the gap between technical findings and practical advice for journalists on Twitter.

Strengths and Limitations of the Proposed Guidelines. The guidelines described in this chapter were distributed to journalists in Irish news organizations and presented in workshops including "Analysing Social Media Data" at The Institute for Future Media and Journalism in Dublin City University. Overall, the feedback obtained was positive, with journalists showing interest in using them to structure their daily tweeting activity and to understand attention to news categories outside their expertise. Besides

Irish journalists, we received positive feedback from people at international news organizations in countries such as Bangladesh and the United States.

Our knowledge on the usefulness of these guidelines, beyond word-of-mouth, is however limited. A follow-up study on the application, usage and benefits of following the guidelines is necessary in order to have concrete numbers on the impact they have on news-tweets attention. It is also important to point out that the guidelines are based on two news ecosystems, namely Ireland and the UK, and that a generalization of their application to other geographies, languages and news cultures, cannot be made without proper experimentation and analysis. The scope of our work was to go one step further than presenting results that do not call for action nor present practical tools for journalists to use. A detailed analysis on the usage and practical impact of such tools, is an interesting extension for future work.

6.2 Designing Headlines for Attention

In this section, we exemplify how analyses of attention to news in social media (see Chapter 4) can be used as key components in real-world systems applications. Particularly, we describe the role that our analyses play in a keyword recommendation system for headlines.

Preliminaries. The headline is an extremely important component of every news article that performs multiple functions: summarizing the story, attracting attention, and signaling the voice and style of the newspaper [31]. In the online realm, headlines are expected to meet several new functions; for instance, to convey the article's contents in different online contexts, or to optimize the article for search engine queries (i.e, SEO). Indeed, arguably, the headline is now more important than ever, as it may be the only visible part of the article in microblog posts (e.g., Twitter), social media feeds and listings on news-aggregation sites. These multiple requirements on the news headline have complicated the composition task facing news editors, as they attempt to ensure that each headline is crafted as perfectly as possible.

Prior NLP work in the area of news headlines has mostly focused on the task of automatic headline generation, cast as "very short summary generation" in the DUC tasks of the early 2000s; tasks that produced much of the research on the topic. The best-performing system in the 2004 DUC task worked by parsing the first sentence of the article and pruning it to the desired length [178], an approach that works by leveraging human intelligence: journalists generally compose news articles in the "inverted pyramid" style, which places the most important information in the lead para-graph [31]. Other headline generation systems generally work by first using some metric to identify terms within the document that are likely to appear in the headline, and then constructing a headline containing these terms [115].

This latter approach has much in common with the task of keyword selection for SEO, which first caught the attention of major newspapers at least ten years ago [99], and continues to be a much-discussed issue today [165]. While even long-established, traditional news publications have begun to move away from classical forms of headlines towards more direct, keyword-laden ones, many copy editors would still prefer to write clever, witty headlines [174, 53], and readers of the news seem to value creativity in headlines over clarity or informativeness [73]. Therefore, one of the key considerations in the design of our system is to balance the mechanical act of filling a headline with informative, relevant keywords, against the creative act of writing headlines that appeal to human interests and emotions.

We expect that the most interesting and emotional stories are likely to be more popular with readers than the "average" story. Analysis of reader behavior has shown that there is no correlation between how much an article is shared on social media and how much of the article is read by an average user [57, 137]; a fact that could be taken as evidence supporting the widely-held view that people share articles online that they have not fully read themselves [101]. In this case, the headline –which people presumably read even if they do not read the full text– could be an important factor in determining the "shareability" of a news article; an idea that is another key motivation behind the design of our system.

The tool presented here is designed to *facilitate the decision-making process* facing a news editor in composing a headline. The software employs NLP and machine learning techniques to make its recommendations, but it is not designed to automatically generate headlines or to make decisions about a headline's goodness on its own.

6.2.1 Design and Behavior

From a user-interface perspective, the tool has two modes of operation: input mode and analysis mode. The input mode (illustrated in Figure 6.1) facilitates the entry of a news article and its corresponding headline and sub-headline, which may either be entered manually or selected from a feed of recent articles. In practice, this feed would be integrated into the newspaper's workflow so that an editor could review all new articles with the tool prior to publication.

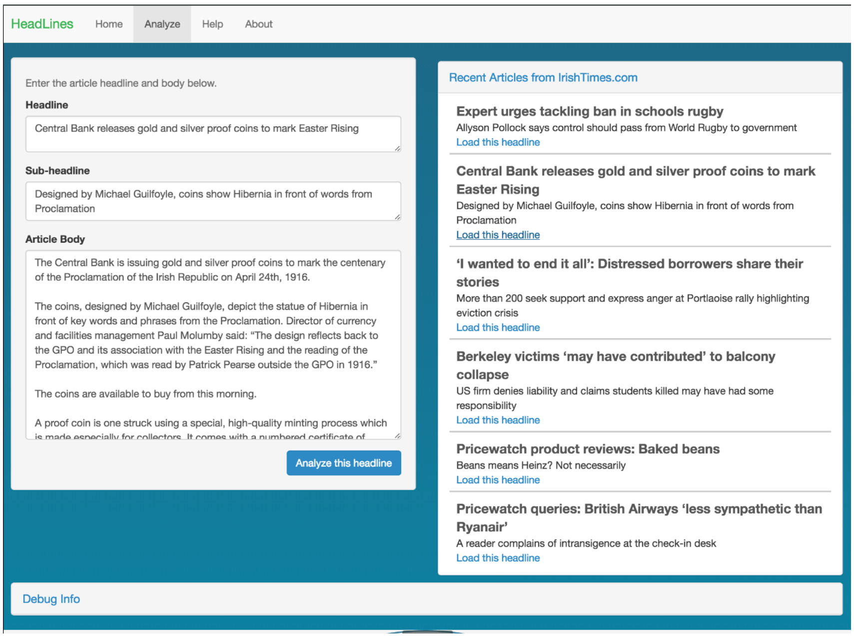


Figure 6.1: Screenshot of the tool in input mode, with input text areas on the left and a live feed of articles on the right.

After the editor-user selects an article, the system switches to the analysis mode, showing the results of the automated analysis (illustrated in Figure 6.2). This mode is designed to allow the user to quickly assess the strengths and weaknesses of the headline and decide whether any changes should be made to improve it. The five most highly-ranked keywords from the article are listed on the right side of the screen sorted by *weight*, a metric combining the keyword's *frequency* in the article and its *SEO score*, which respectively capture the keyword's local relevance to the article itself as well as its global prominence among news stories in general (see section 6.2.3 for details on these measures).

The keywords are color-coded to distinguish keywords which already appear in the headline (green) from those which do not appear in the headline (red), and size-coded according to their weight. Thus, any large, red keywords are those which an editor should consider adding to the headline. In the example in Figure 6.2, the top three recommended keywords are already present in the headline; the two remaining recommendations, "Irish Republic" and "GPO", are both sensible suggestions for the article.

In addition to the keyword recommendations, the system scores each headline for its "shareability" on two social media platforms: Twitter and Facebook; if the shareability score on either platform exceeds a threshold value, then an alert is displayed to the user. In the example in Figure 6.2, the article has exceeded the Facebook threshold but not the Twitter threshold, so only one of the two alerts is displayed. The newspaper's editor in

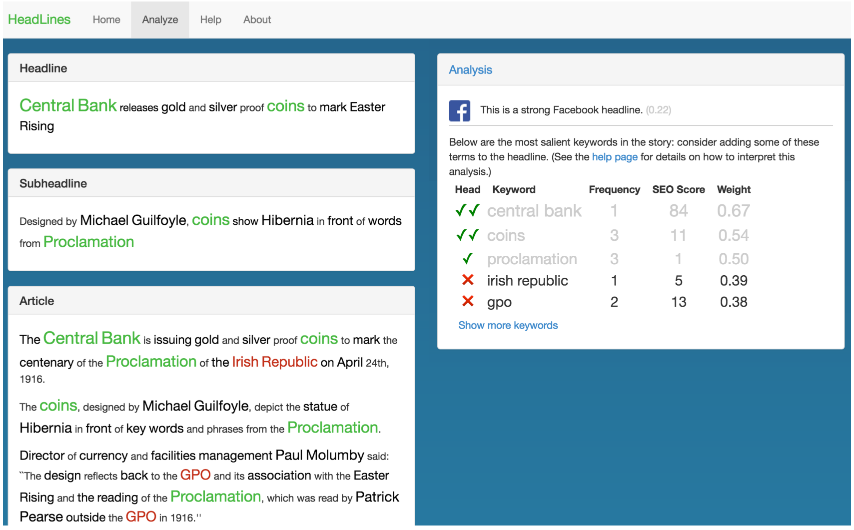


Figure 6.2: Screenshot of the tool in analysis mode. In this example, three of the top five keywords (highlighted in green) are already in the headline; the remaining two (in red) are recommendations that the user may consider adding to the headline.

charge of social media can use this information when deciding which stories should be posted and promoted on social media sites. The threshold is set to a relatively conservative value, so that most articles will not produce alerts, and only the most promising headlines will come to the editor's attention.

Ultimately, it is up to the editor to decide what action, if any, to take based on the information presented by the software. The editor has the leeway to add keywords in the headline in creative ways that fit the style of the story and the news organization, and she can also flexibly deal with any errors that may be produced by the keyword recommendation system, rather than blindly following its advice.

6.2.2 Implementation

The system consists of three major components: a user-interface front-end, a text analysis back-end, and a web server that mediates communication between the two.

The user interface is implemented with HTML and Javascript and accessed via a web browser; its behavior is described and illustrated in the previous section.

The web server is implemented in Python (based on the *Flask*[2] framework), which allows easy integration with the text analytics back-end, which is also mainly implemented in Python. We use the *sklearn* module for regression and Stanford's *CoreNLP* Java suite for NLP [102].

[2]http://flask.pocoo.org/

The entire system is deployed on a web server and accessed by the client's web browser.

The back-end consists of two components—keyword analysis and shareability analysis—which operate independently of one another and are discussed in detail below.

6.2.3 Keyword Analysis

The role of keyword analysis is to identify terms in the article body that are good candidates for inclusion in the headline. We believe that headlines containing informative and popular keywords can be both more appealing to readers and more prominent in users' search results and on news aggregator websites.

Processing of an input article begins with tokenization and named-entity recognition using *CoreNLP*, which identifies all entities (e.g. people, organizations, locations) in the article. Next, any known keywords appearing in the text are identified, by using a database of 90k keywords and their frequencies from Irish news articles in recent years, which we populated with data provided to us by two other Irish news-related projects [155, 17]. This process results in a list of entities, which may be unique to the given article, and a list of keywords, which are known to have been encountered in previous news articles. These keywords and named entities are linked using a simple, rule-based approach that resolves pairs like "Enda Kenny" and "(Mr.) Kenny", yielding a single list of resolved keywords, along with a list of all positions in the text where each keyword appears.

Our keyword ranking system aims to capture the intuition that salient keywords should ideally be both *locally* prominent (i.e. appearing frequently in the given news article) and *globally* popular (i.e. appearing frequently in articles other then the current one). Thus, we calculate the weight w of each keyword k in the document d as the weighted sum of its local weight w_{local} and its global weight w_{global}:

$$w(k, d) = \lambda \, w_{local}(k, d) + (1 - \lambda) \, w_{global}(k) \tag{6.1}$$

The local weight is calculated as the normalized within-document frequency of the keyword, so that the most frequent keyword in the document gets a w_{local} of 1. The global weight is calculated in a similar way, using the across-document frequencies from the keyword database and applying a nonlinear (log) transformation to compensate for the highly skewed distribution of these frequencies (note that it is possible for a keyword to have a zero global weight if it does not appear in our database; this is common for named entities in the article which have not been mentioned in the news before). The relative contributions of the local and global weights are balanced with the λ parameter.

This formula was chosen as the simplest method (a linear combination) of combining the two factors. It is similar to a *tf-idf* score in that it combines both term frequency and document frequency, but it is critically different in that it rewards, rather than penalizes, terms that occur in many documents. This is a good thing because we believe that terms which may be very common (e.g. the names of well-known politicians or celebrities) can be good headline terms, and also because our method of selecting terms (via a closed set of keywords and automatic named entity detection) generally avoids selecting words which may be high-frequency but low-quality (like stopwords).

We manually set the value of λ to achieve rankings which we subjectively deemed to be suitable.[3] This manual parameter setting allowed us to deploy our system quickly with acceptable performance, but a better option would be to learn these parameters automatically. To do so would require a dataset containing news articles, their headlines, and either some measure of the quality of the headline or an assurance that the headlines in the data set are "good", in order to guarantee that the parameters are set based on "good" headline examples. This type of data was not available to us when the system was under development.

This method ultimately assigns a weight to each keyword between 0 and 1.0, which determines its ranking in the analysis output (Figure 6.2).

In the user interface, the weight is displayed in a table alongside the keyword's "frequency" and "SEO Score", which we consider to be more user-friendly than w_{local} and w_{global} themselves (the frequency is exactly the number of times the term appears in the article, and the SEO score is just w_{global} scaled to the familiar scale of 0 to 100).

6.2.3.1 Shareability Analysis

The role of *shareability* analysis is to identify headlines that are likely to be shared on social media. With the rise of social media as dissemination channels for the news, headlines now need to be both informative and "shareable"; that is, the headline some-

[3] We found that a value of 0.6 (i.e. slightly favoring local frequency over global frequency) worked well for our data, but this value changed depending on which keyword list we used. Ultimately, we combined both keyword lists, which introduced a large number of noisy terms. To suppress these noisy terms, we added an additional term to boost the score of keywords which were identified as named entities in the article (up to 0.2 of the overall weight).

how needs to attract people to post, share, and engage with the article on social media, in order to reach a large online audience.

According to the Reuters Institute Digital News Report [117], Facebook and Twitter generate 54% of the visits to online news sites, suggesting that direct visits to the home pages of news providers are being supplanted by social media mediated access. However, Facebook and Twitter are known to have quite different audiences and engage users in different ways [83]. Users on Twitter generally actively search news and their consumption varies across news categories (see Chapter 4), whereas on Facebook, news tends to be just encountered incidentally or just by sharing amongst friends [75, 85, 16]. Therefore, in our system, we model the two social networks separately.

Fot this analysis, we used the 700K tweets and retweets posted by 200 Irish media outlets and journalists between 2013 and 2014 (see Table 4.1 – Ireland, 2013-2014). From these tweets, we extracted all the URLs and used the Facebook and Twitter APIs to collect the number of times each URL was shared on Facebook or posted on Twitter. Because these posts were made by journalists, the links in the tweets are mainly to news articles (see discussion in Section 4.1.2), from which we extracted headlines. This step yielded a data set of 55K headlines with corresponding counts of social shares for each one.

We used regression analysis to estimate the relationship between features of the headlines and the target variable of number of shares. Each headline in our collection is represented as a vector consisting of eight features covering three main aspects of the headline's content: the sentiment polarity (as computed by the *TextBlob* Python package), the presence of named entities, and the length in words. The complete list of features is presented in Table 6.1.

We train two different Gradient Boosting Trees (GBT) regression models, one for Facebook and one for Twitter. From the results, we observe that the two models behave differently: comparing the values of the MSE, predictions for shareable headlines on Facebook present an MSE of 41.8, while for Twitter the error is slightly smaller, 37.6.

Once the GBT models are trained, we store them and incorporate them into the system's pipeline. Every inputted headline receives two shareability scores, one for each social media site; however, in order to avoid triggering too many notifications to the journalist or news editor, the system only shows a result if the score is equal or larger than a manually-defined threshold of 3.7 and 1.7 for Facebook and Twitter, respectively, which correspond to the median number of shares (on each platform) received by the headlines in our collection.

Feature	Description
neutral	# of neutral sentiment words
positive	# of positive sentiment words
negative	# of negative sentiment words
organizations	# of ORGANIZATION entities
persons	# of PERSON entities
places	# of LOCATION entities
day	(T/F) headline contains the name of a day
length	total # of words in the headline

Table 6.1: Headline features used for regression. The first six features are normalized by the length of the headline.

6.2.4 Evaluation

The tool was developed in collaboration with *The Irish Times*, and several professional editors have tested its usability. The feedback from these sessions has been positive and has informed several design features. In particular, the color-coding and font-size features of the interface have been noted for their usability.

Testings of the system have identified some potential areas for improvement. The keyword system commonly fails to recognize when pairs of equivalent but non-identical keywords have the same referent; for example *Taioseach* and *Enda Kenny*, or *GPO* and *General Post Office*. While editors easily recognize this duplication, this error affects frequency counts, which in turn affect keyword rankings. This type of co-reference resolution is an open question in NLP research, with typical solutions relying on a rule-based or gazette-based approach to fix commonly-occurring cases.

The system could also be improved by moving from a static keyword database to a dynamic, real-time database. We were fortunate to be able to bootstrap our system with the keyword sources discussed in section 6.2.3; however neither of these sources were created with this specific use-case in mind, and the static nature of these lists means that the keyword database will become outdated over time. Updating the keyword frequency counts on a rolling basis is an easy first step; but a more sophisticated approach is probably required, where new entities are added to the database over time, and more recent articles are given a greater weight than older articles. Because our system already identifies named entities in news articles, these entities could be added as new keywords in our database as they are encountered.

An ongoing evaluation is also the impact of using the tool on SEO, based on determining whether it improves article rankings in news aggregators and search engines. While the lack of click-through from Google News has led some to question its effectiveness at driving traffic to news sites [173], for *The Irish Times'* website, Google News is a major source of referrals. By performing an analysis of 30k *The Irish Times* articles (from October 2015 to March 2016) we learned that articles listed on the Google News (Irish edition) front pages received significantly ($p < 0.01$) more page views than unlisted articles; with Google News listed articles receiving almost twice as many views ($n = 11,125$, $\mu = 1665.5$ views per article) as unlisted articles ($n = 19,339$, $\mu = 892.4$ views per article). Google News' ranking algorithm is not publicly known, so the exact factors leading to this correlation are opaque; however, for practical purposes, we could argue that if our keyword recommender leads to greater visibility on Google News, then it should increase readership.

Finally, the quality of our keyword recommendations can, in part, be assessed by noting whether the system's top-recommended keywords are already present in the original headline written for the article, as it shows that the system corresponds to human judgments (n.b., the keyword analysis only uses the article body, not the headline, as input). We processed a sample of roughly 3K *The Irish Times* headlines with our system, and found that a majority of these (64%) contained two or more of the top five keywords recommended by our system (in either the headline or the sub-headline), and a large majority (88%) contained at least one of the recommended keywords. We take this as evidence that the keywords recommended by our system generally correspond with the types of keywords that a human editor would normally include in the headline.

6.3 Summary

In this chapter, we presented a set of guidelines derived from analyzing news tweets distribution and consumption as reflected by the interactions between journalistic posts and news audiences on Twitter. We also presented a system for recommending keywords for inclusion in newspaper headlines and for identifying headlines with high potential shareability on social media.

Regarding the guidelines, their aim is to assist journalists on approaching Twitter as a tool for news dissemination. These guidelines, informed by our previous analyses on attention to news tweets, are specific enough to enable news providers to design innovative strategies for improving audience engagement for different news categories and types of Twitter accounts.

The headlines' recommender system identifies plausible keywords that are both relevant to the given news article and popular overall in past news articles headlines, in an effort to maximize both the reader interest and the SEO aspect of the headline. In addition, the system identifies headlines that are likely to receive above-average engagement on social media, allowing editors to effectively target their social media strategy. It is important to note, however, that this tool does not recommend headlines *per se*, but it rather suggests keywords so that the news editor can decide how to craft the headline by either using or discarding such keywords, based on the style of the newspaper or news organization, i.e., it is the editor's decision to avoid or to craft high-quality or click-bait headlines. We believe that this tool can be a helpful component in modern, online-oriented newsrooms.

In the next chapter, we conclude our work and discuss opportunities for future research.

CHAPTER

SEVEN

Conclusion

Technological advances and the popularization of social media and mobile devices, have greatly disrupted the dynamics of news distribution and consumption. This disruption has been the subject of significant research aimed at understanding the factors underlying audience attention and news dissemination on social media, and to provide new tools for journalists to better disseminate their news via these platforms. Nowadays more than ever, interdisciplinary collaboration between computer scientists and journalists is necessary to advance our understanding of news distribution and consumption in online channels.

7.1 Main Contributions

In this thesis, we explored the important aspects of the news cycle in Twitter. We did so by analyzing the behavior of journalists and audiences, as well as the mutual interactions between them throughout a series of topics ranging from the dynamics of news on Twitter, to the attention to and consumption of news, to the designing of tools that can be used by journalists who use Twitter for their work.

We found that there is no clear agreement among journalists on what is a good strategy for distributing their news on Twitter, that readers across news categories look for and value different aspects of the news, and that little is known about what drives users to consume news on Twitter.

We addressed these and other issues by conducting analyses that looked specifically at new tweets, journalists and news readers, and tackled each problem with two motivations: to automatically process, treat and model news on Twitter, and to present the results in a way that they give journalists a better understanding of not only the existing problems but of how these findings may assist them in making a better use of Twitter for news distribution.

We present a summary of our main contributions in the remaining of this section.

7.1.1 The Dynamics of News on Twitter

One of our major contributions, presented in Chapters 2 and 3 are the detailed studies of current research and dynamics of news distribution on Twitter.

First, we provided an integrative review of the literature on the professional reporting of news on Twitter; focusing on how journalists and news outlets use Twitter as a platform to disseminate news, and on the factors that impact readers' attention and engagement with that news on Twitter. Using the precise definition of a *news-tweet*, we produced one of the first research surveys that structures the literature to reveal the main findings on features affecting audience attention to news and its dissemination on Twitter.

We also presented and discussed a survey on the use of Twitter by journalists whose working language is English. The survey provided an overview of actual practices and needs as expressed by the journalists themselves.

These analyses shed light on and structured the main challenges that needed to be addressed in the rest of our work. Before this ground was established, little was known about the latent needs of the niche area: journalism and Twitter.

7.1.2 Attention to News-Tweets: Strategies and Tools

Another contribution of this thesis is the development of strategies for spreading news on Twitter (see Chapter 4) and, using these strategies as input, build tools that assist journalists in their use of this platform (see Chapter 6).

To develop such strategies, we started by analyzing two Twitter corpora which contain real journalists interactions and the corresponding audience responses. We then developed predictive models to identify the features of journalists and news-tweets that impact audience attention. These analyses revealed that different combinations of features influence audience engagement differentially from one news category to the next (e.g., sport versus business).

In addition, in an effort to bridge the gap between technical findings and journalistic practice, we presented methods to interpret results obtained from our prediction tasks so that journalists could find them applicable to their daily practices.

In Chapter 6, we suggested a set of guidelines for journalists, designed to help them maximize engagement with the news they tweet (see Section 6.1). These guidelines have been subject of many interesting discussions with journalists and communication professionals both inside and outside of Ireland.

We also explored the construction of news headlines for online newspapers and described a recommender system that suggests modifications of proposed headlines, based on their keyword content and predicted *shareability* in social media platforms (see Section 6.2). The headlines' recommender system was demoed at *IJCAI'16* where it generated immediate attention from media companies. This solution has been licensed by *The Irish Times*, which further confirms its applicability and relevance.

7.1.3 Drivers of News Consumption on Twitter

While much is known about how people tweet and interact with Twitter, surprisingly little is known about how the news items tweeted by journalists – *news-tweets* – act as a distribution channel for the news that is spread by social media reading and sharing.

We aimed to fill this gap by revealing what drives users to consume news, and by developing a news consumption prediction model. As our third contribution, we presented the Twitter News Model (TNM), a computational data-driven approach to elucidate the dynamics of news consumption on Twitter (see Chapter 5). We applied the TNM to a dataset of interactions between users and journalists/newspapers to reveal *what* drives users' consumption of news on Twitter, and predictively relate users' news beliefs, motivations, and attitudes to their consumption of news. Our findings revealed that news motivations, followed by news attitudes and news beliefs, impact users' behavior of news consumption on Twitter.

7.2 Discussion

In this thesis, we concentrated on news-specific environments and data within the vast world of social media. We approached the research problems by looking at the different participants of the online news distribution cycle: journalists and audience, and focused on advancing our understanding of each group's behavior and goals.

Furthermore, within each scenario, we considered the particularities of different news categories, i.e., sport news do not have the same audience nor the same qualities as political news and, therefore, should be analyzed separately. In the case of the headlines' recommendation system, we took into consideration that audiences behave differently depending on the social media platform they are using, and so the relevance of a headline may vary accordingly.

We believe that research at the intersection of journalism and computer science is of high relevance, and that there is a huge opportunity to open more channels for discussion and collaboration that will allow us to widen our understanding of such a dynamic and important area.

Overall, the most significant impact of this thesis is the usage and adaptation of data analytics methods for the advancement of journalistic strategies in the crafting and distribution of news in Twitter. Furthermore, by having prioritized the interpretation of models' outputs and produced practical tools, this thesis has proposed that an active, ongoing, and symbiotic collaboration between computer scientists and other disciplines, e.g., journalism, is necessary and possible.

7.3 Limitations

This thesis has focused on addressing important challenges that journalists and news organizations face in their distribution of news in social media. However, despite the work being as carefully designed and executed as possible, certain limitations remain.

Since the analyses here presented largely focus on social factors (e.g. times, days, and audience interests to news-tweets), more exhaustive content analysis (e.g. linguistic analysis) is out of the scope of the thesis.

The thesis performs large-scale analysis of news-tweets dissemination and consumption, as opposed to previous works that may have conducted finer-grained, small-scale analyses of tweets in other niche areas. Inevitably, this means having a limited amount of noise in the large-scale computational analysis (e.g., a marked disparity on the activity and attention some journalistic Twitter accounts perceive), however this noise has been minimized to the extent possible, for example by identifying and excluding from the analyses accounts that displayed a bot-like volume of tweeting, and/or whose tweets were not news content.

Certain aspects, such as the effect of re-posting news-tweets in different times of the day or with different texts, and the volume of news articles and how that may or may not match trends observed in tweet volumes, were not considered in our analyses and are left for future work.

With respect to the dynamics of news distribution in Twitter (see Chapter 3), while this thesis provides a glimpse into the usage of the platform as self-reported by journalists, the collection of demographic data was limited due to restrictions on the participants' side. In the future, and if a similar survey were to be run again, it would be valuable to collect data about the journalists that goes beyond their usage of the platform but can impact their usage of social media channels. For example their nationality, age, and news organization they work for.

7.4 Outlook

With the ongoing proliferation of social media platforms, the need for the development of novel strategies relevant to news production, dissemination, and consequent consumption will continue apace.

To deepen our understanding of the dissemination of news in digital social environments, we need to further investigate the inter-dependency between features of tweets (or other social media posts) and journalists. For example, whether journalists' styles when crafting online news posts have a direct effect on the relevant features of the post itself and their impact on audience attention, or whether the popularity of the journalist can affect the features' importance.

As for analyzing audiences behaviors when consuming news in social media, there is a need to develop native constructs that consider the particularities of the interactions between users and news in digital environments. Several current modelings, including our TNM, use aggregate behavioral measures, e.g., we define news beliefs as the ratio between retweets and mentions, to operationalize psychological constructs; however, the two can be theoretically and empirically different.

Technological approaches that assist journalists on tasks such as news posts crafting and headlines composition and that help audiences by, for example, enriching news with contextual information, will continue to emerge. We consider that this thesis has contributed to advance our understanding of digital news delivery and consumption on Twitter, and that the future of news will continue to pose new and interesting challenges for interdisciplinary research.

Appendix A

A.1 Survey: Twitter for Journalism

Currently, the news media face many serious challenges as their distribution channels are gradually being taken over by third parties (e.g., people sharing news on Twitter and Facebook, and Google News acting as a news aggregator). For journalists to remain competitive in social media and other emerging platforms, they need to develop innovative strategies to maximize audience engagement for the news they produce.

Survey Design

To understand current journalistic practices on Twitter, we designed a survey directed to professional journalists and people who identify themselves as journalists that assesses the extent to which they follow a strategy when tweeting news.

The survey's questions require a combination of multiple-choice and free-form responses, and covers the following areas:

Twitter usage. Deals with understanding the basic use of the platform; with questions asking each journalist how often they use Twitter, about their searching and retweeting behavior, and the situations in which they reply to and mention others in their tweets.

Joining the conversation. Concerned with understanding the mechanisms used by journalists for sharing news and information with a broader audience; addressing whether they included hashtags, media or quotes in their tweets, whether they post news as soon as it occurs, and how often they interact with their audience either by replying or answering questions.

Increasing followers and improving engagement. Questions the techniques that journalists use to grow their audience and increase engagement with readers; by asking them about the regularity of their tweeting and if they follow up and update their audience on past stories, whether they actively seek to interact with other prominent users, and about the actions they take to increase engagement with their readers.

Getting to know your audience. Probes on the steps that journalists take to improve their knowledge of their audience, by asking if and what tools they use to track their Twitter audience, the features that they find important or would like to know about their readers, and if they prefer to have a small but loyal and active audience or a large and more diverse group even if they are less active.

Measuring engagement on Twitter. Asks journalists about if and how they measure audience engagement to their tweeted news and about the characteristics of both journalists and tweets that they consider impact audience reaction.

As such, these five areas cover the main topics relevant to the proper and best use of the service by journalists.

Note: The survey is anonymous, we do not ask for personal information such as name, age, gender, or home organization.

Survey Questions

1. How often do you use Twitter?
   - Daily
   - Several times a week
   - Weekly
   - Monthly
   - Other (please specify)
2. Do you use Twitter to learn about the current, most interesting topics of discussion?
   - Yes
   - No
   - Other (please specify)
3. When retweeting, do you prefer to share the tweets of a newspaper or fellow journalist over those of ordinary individuals?
   - Yes
   - No

- It depends on the topic
- Other (please specify)

4. In what situations do you use replies and mentions?
    - To be part of a conversation
    - To show others that I care about their opinions
    - To get the attention of relevant users
    - Other (please specify)
5. What news category do you report?
    - Sports
    - Politics
    - Breaking News
    - Lifestyle
    - Business
    - Science and Technology
    - Other (please specify)
6. How often do you include hashtags, photos, videos, or quotes in your tweets?
    - Always
    - Only when one or more of these resources is at hand
    - Never
    - I mostly include hashtags
    - Other (please specify)
7. Twitter is live, public and conversational. Do you use Twitter to reach broad audiences as soon as news occurs?
    - Yes
    - No
    - Only in cases when I am certain about the source and the topic is trending
    - Other (please specify)

8. How often do you reply to users, answer questions or engage with your readers?
   - I try to engage with my readers as much as I can
   - I only reply or answer to colleagues or influential public figures
   - I do not engage in conversations in Twitter
   - Other (please specify)
9. Do you tweet regularly with updates and media?
   - Yes, I post updates and follow ups to the stories and include media whenever possible
   - No, I do not post updates on stories I tweet about
   - I post updates but I do not include media
   - Other (please specify)
10. How often do you engage with prominent users on Twitter?
    - I like to engage with prominent users as often as I can
    - I do not actively seek to engage with prominent users
    - I engage with prominent users only when I consider it would help me spread a relevant post
    - Other (please specify)
11. What actions do you take to increase engagement with your readers on Twitter?
    - I include hashtags in my tweets
    - I reply to the tweets where they mention me
    - I retweet the tweets they post and I consider relevant
    - I mention them in my tweets
    - I do not take any action
    - Other (please specify)
12. Do you use any tool to track your audience in Twitter? If yes, what tool do you use?
13. What information do you find useful or you would like to know about your Twitter audience? (you can select more than one option)

- Age
- Location
- Gender
- Frequency of tweeting activity
- Topics of interest
- Linking behavior (inclusion of links/URLs in their tweets)
- Conversational engagement (usage of mentions, replies, hashtags)
- Other (please specify)

14. Regarding the size of your audience, do you prefer to have a small but loyal and active group of readers or a larger, more diverse group even if less active?
15. Whom would you consider as being a key group of readers?
    - Institutional accounts
    - Public figures
    - Ordinary individuals
    - Other (please specify)
16. How do you measure the engagement of your audience to your tweets?
    - By the number of retweets received (the more retweets the more engaged my audience is to my tweets)
    - By the number of favorites
    - By how many different users react to them (retweet, favorite)
    - By how fast the tweet starts getting reactions since the moment I post it
    - By who reacts to my tweets (relevant users)
    - Other (please specify)
17. What characteristics of a tweet would you say influence the reaction of the audience towards it? (you can select more than one option)
    - The inclusion of one or more hashtags (e.g., #hashtag)
    - The inclusion of one or more mentions (e.g., @username)
    - The words included in the tweet

- The inclusion of URLs
- Who posts the tweet
- The topic the tweet is about
- The time and day when the tweet is posted
- Other (please specify)

18. What characteristics of a journalist would you say influence the reaction of the audience towards her or his tweets?
    - The news category she/he tweets about
    - The gender
    - The volume and frequency at which she/he tweets
    - The organization he/she works for
    - How well connected she/he is (many followers/followees)
    - Her/His reputation as a journalist
    - Other (please specify)

BIBLIOGRAPHY

[1] Reuters institute digital news report. Technical report, 2015.

[2] Reuters institute digital news report. Technical report, 2016.

[3] W. Ahmed and S. Lugovic. Social media analytics: analysis and visualisation of news diffusion using nodexl. *Online Information Review*, 10 2018.

[4] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May 2017.

[5] P. C. Álvarez. Journalism and social media: How spanish journalists are using twitter/periodismo y social media: cómo están usando twitter los periodistas españoles. *Estudios sobre el mensaje periodístico*, 18(1):31–53, 2012.

[6] A. Arif, J. J. Robinson, S. A. Stanek, E. S. Fichet, P. Townsend, Z. Worku, and K. Starbird. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 155–168. ACM, 2017.

[7] Associated Press. Social media guidelines for ap employees, 2013. [Online].

[8] 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.

[9] M. Bagdouri. Journalists and twitter: A multidimensional quantitative description of usage patterns. In *ICWSM*, pages 22–31. AAAI Press, 2016.

[10] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. *CoRR*, abs/1202.0332. 2012.

[11] M. Barthel, E. Shearer, J. Gottfried, and A. Mitchell. The evolving role of news on twitter and facebook, 2015.

[12] A. Bechmann and K. L. Nielbo. Are we exposed to the same "news" in the news feed? *Digital Journalism*, 6(8):990–1002, 2018.

[13] D. Bhattacharya and S. Ram. Sharing news articles using 140 characters: A diffusion analysis on twitter. In *Proceedings of the 2012 IEEE/ACM International Con-*

ference on Advances in Social Networks Analysis and Mining, ASONAM 2012, pages 966–971, 2012.

[14] D. Bhattacharya and S. Ram. Understanding the competitive landscape of news providers on social media. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 719–724. International World Wide Web Conferences Steering Committee, 2016.

[15] J. N. Blom and K. R. Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015.

[16] P. J. Boczkowski, E. Mitchelstein, and M. Matassi. "news comes across when i'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New Media & Society*, 20(10):3523–3539, 2018.

[17] G. Bordea, P. Buitelaar, and T. Polajnar. Domain-independent term extraction through domain modelling. In *Proceedings of TIA*, 2013.

[18] M. Borger, A. van Hoof, and J. Sanders. Expecting reciprocity: Towards a model of the participants' perspective on participatory journalism. *New Media & Society*, 18(5):708–725, 2016.

[19] M. Brautović, I. Milanović-Litre, and R. John. Journalism and twitter: Between journalistic norms and new routines. *MediAnali : International scientific journal of media, journalism, mass communication, public relations, culture and society*, 7(13):19–36, 2013.

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

[21] M. Broersma and T. Graham. Social media as beat. *Journalism Practice*, 6(3):403–419, 2012.

[22] M. Broersma and T. Graham. Twitter as a news source. *Journalism Practice*, 7(4):446–464, 2013.

[23] A. Bruns. Journalists and twitter: How australian news organisations adapt to a new medium. *Media International Australia*, 144(1):97–107, 2012.

[24] L. Canter. The interactive spectrum. *Convergence*, 19(4):472–495, 2013.

[25] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 211–223, New York, NY, USA, 2014. ACM.

[26] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[27] 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.

[28] S. Chesney, M. Liakata, M. Poesio, and M. Purver. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61, 2017.

[29] M. Cliche. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*, 2017.

[30] S. Cohen, J. T. Hamilton, and F. Turner. Computational journalism, 10 2011.

[31] M. Conboy. *The Language of the News*. Routledge, 2007.

[32] L. Connell and M. T. Keane. What plausibly affects plausibility? concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory & Cognition*, 2004.

[33] D. Dailey and K. Starbird. Journalists as crowdsourcerers: Responding to crisis by reporting with a crowd. *Computer Supported Cooperative Work (CSCW)*, 23(4):445–481, 2014.

[34] C. Danescu-Niculescu-Mizil, J. Cheng, J. M. Kleinberg, and L. Lee. You had me at hello: How phrasing affects memorability. In *ACL (1)*, pages 892–901. The Association for Computer Linguistics, 2012.

[35] 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.

[36] H. G. de Zúñiga, T. Diehl, and A. Ardèvol-Abreu. When citizens and journalists interact on twitter. *Journalism Studies*, 19(2):227–246, 2018.

[37] W. Dean. The journalist as a 'committed observer', 2013.

[38] M. Del Vicario, S. Gaito, W. Quattrociocchi, M. Zignani, and F. Zollo. News consumption during the italian referendum: A cross-platform analysis on facebook and twitter. In *Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on*, pages 648–657. IEEE, 2017.

[39] N. Diakopoulos. A functional roadmap for innovation in computational journalism, 04 2011.

[40] 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.

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

[42] 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 RecSys Challenge*, 2014.

[43] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd. In J. G. Breslin, N. B. Ellison, J. G. Shanahan,

and Z. Tufekci, editors, *Proceedings of the Sixth International Conference on Weblogs and Social Media*. The AAAI Press, 2012.

[44] M. Dredze, P. Kambadur, G. Kazantsev, G. Mann, and M. Osborne. How twitter is changing the nature of financial news discovery. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, DSMM'16, pages 2:1–2:5, New York, NY, USA, 2016. ACM.

[45] N. Drok, L. Hermans, and K. Kats. Decoding youth dna: The relationship between social engagement and news interest, news media use and news preferences of dutch millennials. *Journalism*, 19(5):699–717, 2018.

[46] E. Elejalde, L. Ferres, and E. Herder. On the nature of real and perceived bias in the mainstream media. *PloS one*, 13(3):e0193765, 2018.

[47] P. English. Twitter's diffusion in sports journalism: Role models, laggards and followers of the social media innovation. *New Media & Society*, 18(3):484–501, 2016.

[48] M. Fishbein and I. Ajzen. *Predicting and Changing Behavior*. Psychology Press, 2010.

[49] S. R. Flaxman, S. Goel, and J. M. Rao. Ideological segregation and the effects of social media on news consumption. In *SSRN Scholarly Paper ID 2363701*. Social Science Research Network, Rochester, NY, 2013.

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

[51] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout. Social clicks: What and who gets read on twitter? *ACM SIGMETRICS Performance Evaluation Review*, 44(1):179–192, 2016.

[52] V. García-Perdomo, R. Salaverría, D. K. Kilgo, and S. Harlow. To share or not to share: The influence of news values and topics on popular social media content in the united states, brazil, and argentina. *Journalism Studies*, pages 1–22, 2017.

[53] L. Gatti, G. Özbal, M. Guerini, S. Oliviero, and C. Strapparava. Automatic creation of flexible catchy headlines. *IJCAI-16 Natural Language & Journalism*, 2016.

[54] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, Apr. 2006.

[55] C. Grevet, L. G. Terveen, and E. Gilbert. Managing political differences in social media. In S. R. Fussell, W. G. Lutters, M. R. Morris, and M. Reddy, editors, *CSCW*, pages 1400–1408. ACM, 2014.

[56] L. Ha, Y. Xu, C. Yang, F. Wang, L. Yang, M. Abuljadail, X. Hu, W. Jiang, and I. Gabay. Decline in news content engagement or news medium engagement? a longitudinal analysis of news engagement since the rise of social and mobile media 2009-2012. *Journalism*, 19(5):718–739, 2018.

[57] T. Haile. What you think you know about the web is wrong. *Time Magazine*, 2014.

[58] Harvard Kennedy School, Shorenstein Center. Nyt public editor sees social media as 'double-edged sword' that is changing objective journalism, 02 2013.

[59] M. Hasanain, M. Bagdouri, T. Elsayed, and D. W. Oard. What questions do journalists ask on twitter? In *Social Media in the Newsroom, Papers from the 2016 ICWSM Workshop, Cologne, Germany, May 17, 2016*, 2016.

[60] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

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

[62] A. Hermida. *Journalism Practice*, chapter Twittering the News: The Emergence of Ambient Journalism. 2010.

[63] A. Hermida. #JOURNALISM: Reconfiguring journalism research about Twitter, one tweet at a time. *Digital Journalism*, 1(3):295–313, Oct. 2013.

[64] A. Hermida, S. C. Lewis, and R. Zamith. Sourcing the arab spring: A case study of andy carvin's sources on twitter during the tunisian and egyptian revolutions. *Journal of Computer-Mediated Communication*, 19(3):479–499, 2014.

[65] N. O. Hodas and K. Lerman. How visibility and divided attention constrain social contagion. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 249–257, Washington, DC, USA, 2012. IEEE Computer Society.

[66] S. Hong. Online news on twitter: Newspapers' social media adoption and their online readership. *Information Economics and Policy*, 24(1):69 – 74, 2012. The Economics of Digital Media Markets.

[67] P. N. Howard, S. Bradshaw, B. Kollanyi, C. Desigaud, and G. Bolsover. Junk news and bots during the french presidential election: What are french voters sharing over twitter? *Data Memo 2017.3. Oxford, UK: Project on Computational Propaganda*, March 2017.

[68] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening closed regimes: What was the role of social media during the arab spring? *SSRN*. 2011.

[69] C.-K. Hsieh, L. Yang, H. Wei, M. Naaman, and D. Estrin. Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 51–62, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[70] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2751–2754, New York, NY, USA, 2012. ACM.

[71] Y. Hu, S. Farnham, and K. Talamadupula. Predicting user engagement on twitter with real-world events. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 168–178, 2015.

[72] 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.

[73] E. Ifantidou. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, 41(4):699–720, 2009.

[74] J. JACCARD. The reasoned action model: Directions for future research. *The Annals of the American Academy of Political and Social Science*, pages 58–80, 2012.

[75] J. Kaiser, T. R. Keller, and K. K. von Königslöw. Incidental news exposure on facebook as a social experience: The influence of recommender and media cues on news selection. *Communication Research*, 0(0):0093650218803529, 0.

[76] B. Kalsnes and A. O. Larsson. Understanding news sharing across social media: Detailing distribution on facebook and twitter. *Journalism Studies*, pages 1–20, 2017.

[77] N. Kayser-Bril, A. Valeeva, and I. Radchenko. Transformation of communication processes: Data journalism. *CoRR*. 2016.

[78] N. Keane, C. Yee, and L. Zhou. Using topic modeling and similarity thresholds to detect events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 34–42, 2015.

[79] Y. Keneshloo, S. Wang, E.-H. Han, and N. Ramakrishnan. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441–449. SIAM, 2016.

[80] D. K. Kilgo, S. Harlow, V. García-Perdomo, and R. Salaverría. A new sensation? an international exploration of sensationalism and social media recommendations in online news publications. *Journalism*, 2016.

[81] J. H. Kim, A. Mantrach, A. Jaimes, and A. Oh. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1645–1654. ACM, 2016.

[82] Y. Kim, Y. Kim, Y. Wang, and N. Y. Lee. Uses and gratifications, journalists' twitter use, and relational satisfaction with the public. *Journal of Broadcasting & Electronic Media*, 60(3):503–526, 2016.

[83] N. Kirk, J. Suiter, and P. McNamara. *Reuters Institute Digital News Report (Ireland)*. 2015.

[84] S. Kong, L. Feng, G. Sun, and K. Luo. Predicting lifespans of popular tweets in microblog. In *Proceedings of the 35th International ACM SIGIR Conference on Re-*

search and Development in Information Retrieval, SIGIR '12, pages 1129–1130, New York, NY, USA, 2012. ACM.

[85] A. S. Kümpel. The issue takes it all? *Digital Journalism*, 7(2):165–186, 2019.

[86] 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.

[87] D. L. Lasorsa. Transparency and other journalistic norms on twitter. *Journalism Studies*, 13(3):402–417, 2012.

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

[89] D. L. Lasorsa, S. C. Lewis, and A. E. Holton. Normalizing twitter. *Journalism Studies*, 13(1):19–36, 2012.

[90] R. G. Lawrence, L. Molyneux, M. Coddington, and A. Holton. Tweeting conventions. *Journalism Studies*, 15(6):789–806, 2014.

[91] A. M. Lee and H. I. Chyi. Motivational consumption model: Exploring the psychological structure of news use. *Journalism & Mass Communication Quarterly*, 91(4):706–724, 2014.

[92] E.-J. Lee and S. Y. Shin. Are they talking to me? cognitive and affective effects of interactivity in politicians' twitter communication. *Cyberpsy., Behavior, and Soc. Networking*, 15(10):515–520, 2012.

[93] J. Lee. 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, 2015.

[94] J. Lehmann, C. Castillo, M. Lalmas, and R. Baeza-Yates. Story-focused reading in online news and its potential for user engagement. *Journal of the Association for Information Science and Technology*, 68(4):869–883, 2017.

[95] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman. Transient news crowds in social media. In *ICWSM*, 2013.

[96] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[97] Q. Li, S. Shah, A. Nourbakhsh, R. Fang, and X. Liu. funsentiment at semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs using word vectors built from stocktwits and twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 852–856, 2017.

[98] C. Liu, W. Wang, Y. Zhang, Y. Dong, F. He, and C. Wu. Predicting the popularity of online news based on multivariate analysis. In *Computer and Information Technology (CIT), 2017 IEEE International Conference on*, pages 9–15. IEEE, 2017.

[99] S. Lohr. This boring headline is written for Google. *The New York Times*, April 9 2006.

[100] M. M. Malik and J. Pfeffer. A macroscopic analysis of news content in twitter. *Digital Journalism*, 4(8):955–979, 2016.

[101] F. Manjoo. You won't finish this article. *Slate Magazine*, June 6 2013.

[102] C. D. Manning, M. Surdeanu, et al. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, 2014.

[103] J. N. Matias, S. Szalavitz, and E. Zuckerman. Followbias: Supporting behavior change toward gender equality by networked gatekeepers on social media. In *CSCW*, pages 1082–1095, 2017.

[104] J. N. Matias and H. Wallach. Working paper: Modeling gender discrimination by online news audiences. In *Computation + Journalism Symposium, Columbia University*, 2015.

[105] L. McCay-Peet, M. Lalmas, and V. Navalpakkam. On saliency, affect and focused attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 541–550, New York, NY, USA, 2012. ACM.

[106] M. McCombs and P. Poindexter. The duty to keep informed: News exposure and civic obligation. *Journal of Communication*, 33(2):88–96, 1983.

[107] T. Media. The best ways for news organizations to engage with their audience, 2018. [Online].

[108] T. Media. How journalists can best engage with their audience, 2018. [Online].

[109] K. Metrics, January 2015.

[110] S. Michailina, A. Masouras, and C. Papademetriou. Uses and gratifications in online news: Comparing social media and news media use by users. 05 2015.

[111] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS*, pages 3111–3119, 2013.

[112] J. Möller, R. N. van de Velde, L. Merten, and C. Puschmann. Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, 0(0):0894439319828012, 0.

[113] L. Molyneux. What journalists retweet: Opinion, humor, and brand development on twitter. *Journalism*, 16(7):920–935, 2015.

[114] A. Monroy-Hernández, d. boyd, E. Kiciman, M. De Choudhury, and S. Counts. The new war correspondents: The rise of civic media curation in urban warfare. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1443–1452, New York, NY, USA, 2013. ACM.

[115] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3), 2011.

[116] L.-M. Neudert, B. Kollanyi, and P. N. Howard. Junk news and bots during the german federal presidency election: What were german voters sharing over twitter?, 02 2017. [Online; 2017].

[117] N. Newman, R. Fletcher, D. A. L. Levy, and R. K. Nielsen. *Reuters Institute Digital News Report*. 2016.

[118] D. Nikolov, D. F. M. Oliveira, A. Flammini, and F. Menczer. Measuring online social bubbles. *CoRR*, abs/1502.07162. 2015.

[119] A. Oeldorf-Hirsch. The role of engagement in learning from active and incidental news exposure on social media. *Mass Communication and Society*, 21(2):225–247, 2018.

[120] 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.

[121] A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 994–1009, New York, NY, USA, 2015. ACM.

[122] C. Orellana-Rodriguez, E. Diaz-Aviles, and W. Nejdl. Mining emotions in short films: user comments or crowdsourcing? In *Proceedings of the 22nd International Conference on World Wide Web*, pages 69–70. ACM, 2013.

[123] C. Orellana-Rodriguez, E. Diaz-Aviles, and W. Nejdl. Mining affective context in short films for emotion-aware recommendation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 185–194. ACM, 2015.

[124] C. Orellana-Rodriguez, 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, pages 107–116, New York, NY, USA, 2016. ACM.

[125] C. Orellana-Rodriguez, D. Greene, and M. T. Keane. Spreading one's tweets: How can journalists gain attention for their tweeted news? *The Journal of Web Science*, 4(1):1–16, 2017.

[126] C. Orellana-Rodriguez and M. T. Keane. Modeling and predicting news consumption on twitter. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, pages 321–329, New York, NY, USA, 2018. ACM.

[127] 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.

[128] S. Owens. How the atlantic uses twitter to drive traffic to content, 2015. [Online].

[129] 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.

[130] J. H. Parmelee. Political journalists and twitter: Influences on norms and practices. *Journal of Media Practice*, 14(4):291–305, 2013.

[131] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[132] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[133] Pew Research Center. In changing news landscape, even television is vulnerable, 09 2012.

[134] Pew Research Center. News use across social media platforms 2016, 05 2016.

[135] A. Piotrkowicz, V. Dimitrova, J. Otterbacher, and K. Markert. Headlines matter: Using headlines to predict the popularity of news articles on twitter and facebook. In *ICWSM*, pages 656–659. AAAI Press, 2017.

[136] A. Piotrkowicz, V. Dimitrova, J. Otterbacher, and K. Markert. The impact of news values and linguistic style on the popularity of headlines on twitter and facebook. In *Proceedings of the Second International Workshop on News and Public Opinion (ICWSM NECO 2017)*. Association for the Advancement of Artificial Intelligence, 2017.

[137] A. Ramachandran, L. Wang, and A. Chaintreau. Dynamics and prediction of clicks on news from twitter. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, pages 210–214, New York, NY, USA, 2018. ACM.

[138] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–30, 2014.

[139] F. Rangel, P. Rosso, M. Potthast, and B. Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*, 2017.

[140] J. J. Ratey and A. M. Galaburda. *A User's Guide to the Brain: Perception, Attention, and the Four Theaters of the Brain*. Vintage Series. Vintage Books, 2002.

[141] J. Reis, H. Kwak, J. An, J. Messias, and F. Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 195–204. ACM, 2017.

[142] B. Ritholtz. Twitter is becoming the first and quickest source of investment news, 04 2013.

[143] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 113–114, New York, NY, USA, 2011. ACM.

[144] 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.

[145] 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.

[146] T. Rosenstiel, J. Sonderman, K. Loker, M. Ivancin, and N. Kjarval. *Twitter and the news: How people use the social network to learn about the world*. American Press Institute, 2015.

[147] S. Rosenthal, N. Farra, and P. Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[148] A. Rozental and D. Fleischer. Amobee at semeval-2017 task 4: Deep learning system for sentiment detection on twitter. *CoRR*, abs/1705.01306, 2017.

[149] C. Sadowski and G. Levin. Simihash: Hash-based similarity detection, 2011.

[150] D. Saez-Trumper, C. Castillo, and M. Lalmas. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1679–1684, New York, NY, USA, 2013. ACM.

[151] 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.

[152] 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.

[153] 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.

[154] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of fake news by social bots. *CoRR*, abs/1707.07592. 2017.

[155] B. Shi, G. Ifrim, and N. Hurley. Be in the know: Connecting news articles to relevant Twitter conversations. In *Proceedings of ECML*, 2014.

[156] B. Shi, G. Ifrim, and N. J. Hurley. Learning-to-rank for real-time high-precision hashtag recommendation for streaming news. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *WWW*, pages 1191–1202. ACM, 2016.

[157] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[158] K. Starbird, D. Dailey, O. Mohamed, G. Lee, and E. S. Spiro. Engage early, correct more: How journalists participate in false rumors online during crisis events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 105. ACM, 2018.

[159] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference 2014 Proceedings*, 2014.

[160] B. Stelter. Journalism rules are bent in news coverage from iran, 06 2009.

[161] J. Stray. Are we stuck in filter bubbles? here are five potential paths out, 2012.

[162] C. Suen, S. Huang, C. Eksombatchai, R. Sosic, and J. Leskovec. Nifty: A system for large scale information flow tracking and clustering. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 1237–1248, New York, NY, USA, 2013. ACM.

[163] 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.

[164] M. Sullivan. How social media affects journalistic objectivity, 02 2013.

[165] M. Sullivan. Hey, Google! Check out this column on headlines. *The New York Times*, April 18 2015.

[166] T. Szymanski, C. Orellana-Rodriguez, and M. T. Keane. Helping news editors write better headlines: A recommender to improve the keyword contents & shareability of news headlines. *IJCAI-16 Natural Language & Journalism; also arXiv preprint arXiv:1705.09656*, 2016.

[167] E. C. Tandoc and T. P. Vos. The journalist is marketing the news. *Journalism Practice*, 10(8):950–966, 2016.

[168] D. Trilling, P. Tolochko, and B. Burscher. From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly*, 94(1):38–60, 2017.

[169] F. Vis. Twitter as a reporting tool for breaking news. *Digital Journalism*, 1(1):27–47, 2013.

[170] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[171] J. Wallace. Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digital Journalism*, 6(3):274–293, 2018.

[172] B. Wang, C. Wang, J. Bu, C. Chen, W. V. Zhang, D. Cai, and X. He. Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1331–1340, New York, NY, USA, 2013. ACM.

[173] R. Wauters. Report: 44% of Google News visitors scan headlines, don't click through. *TechCrunch*, January 19 2010.

[174] D. R. Wheeler. 'Google doesn't laugh': Saving witty headlines in the age of SEO. *The Atlantic*, May 11 2011.

[175] K. Williams. *Read All About It!: A History of the British Newspaper*. Routledge, New York, NY, USA, 2010.

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

[177] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[178] D. Zajic, B. J. Dorr, and R. Schwartz. BBN/UMD at DUC-2004: Topiary. In *Proceedings of DUC*, 2004.

[179] L. Zhu and K. Lerman. Attention inequality in social media. *CoRR*, abs/1601.07200. 2016.

[180] A. Zubiaga. Newspaper editors vs the crowd: On the appropriateness of front page news selection. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 879–880, New York, NY, USA, 2013. ACM.

[181] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32, 2018.

[182] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

sfi
science foundation ireland
fondúireacht eolaíochta éireann

UCD DUBLIN

Insight