News Sharing User Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions

Giovanni Brena,[1] Marco Brambilla,[1] Stefano Ceri,[1]
Marco Di Giovanni,[1] Francesco Pierri,[1] Giorgia Ramponi[1]
[1]Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Via Giuseppe Ponzio, 34, I-20133 Milano, Italy
{firstname.lastname}@polimi.it

Abstract

Online social media are changing the news industry and revolutionizing the traditional role of journalists and newspapers. In this scenario, investigating the behaviour of users in relationship to news sharing is relevant, as it provides means for understanding the impact of online news, their propagation within social communities, their impact on the formation of opinions, and also for effectively detecting individual stances relative to specific news or topics.

Our contribution is two-fold. First, we build a robust pipeline for collecting datasets describing news sharing; the pipeline takes as input a list of news sources and generates a large collection of articles, of the accounts that provide them on the social media either directly or by retweeting, and of the social activities performed by these accounts. Second, we also provide a large-scale dataset, built using the aforementioned tool, that can be used to study the social behavior of Twitter users and their involvement in the dissemination of news items. Finally we show an application of our data collection in the context of political stance classification and we suggest other potential usages of the presented resources.

Introduction

In recent times social networks platforms have registered an increasing growth of online interactions, revolutionizing existing communication paradigms in the news-media industry. Statistics show that over a third of the world's population is nowadays connected to at least one social platform[1].

As a result, news consumption is massively shifting towards these social technologies, where users can easily ingest, share and discuss news with friends or other readers. The term "Citizen journalism" has been used (Bruns and Highfield 2012) to describe the tendency of users to actively participate in the process of producing, disseminating and consuming so-called "random acts of journalism" (Bruns and Highfield 2012) which unfold in a decentralized manner via online social media platforms such as Twitter. This affects traditional roles of journalists and news outlets, as traditional barriers for entering online media industry have dropped; producing content online is easier and faster than ever (Allcott and Gentzkow 2017).

Factors such as the existence of closed-knit communities known as "echo-chambers" (Sunstein 2001) and the so-called "algorithmic bias" (Morgan, Lampe, and Shafiq 2013) have been indicated as primary drivers of information diffusion. Some studies have instead highlighted the role of different agents (including bots and cyborgs) in the dissemination of news items on social media (Ferrara et al. 2016) and described the growing presence on these platforms of malicious kinds of information which raised global concern in recent times (Allcott and Gentzkow 2017). Other research has focused on analyzing the diffusion of news using epidemiological models (Jin et al. 2013) or developing network-based models for describing users news sharing behaviour (Raghavan, Anderson, and Kleinberg 2018).

In this landscape, the contribution of our work is two-fold:

1. **A data collection and enrichment pipeline** which allows to generate custom data collections that include features related both to news content and the social context starting from a pre-defined set of news sources.
2. **A dataset which includes news articles from US major news outlets and associated sharing activities on Twitter**, covering the content of the sharing tweets and details of the users. The dataset highlights users' involvement in the process of news dissemination as we believe that understanding news sharing behaviours can provide further insights on detecting users' opinions, stance and communities. In particular, we describe a practical usage of our dataset in the context of political stance classification.

The paper is organized as follows: we first describe in detail the data collection pipeline; then, we provide a quantitative and qualitative description of the dataset; then, we show a use case of this collection in the context of political stance classification; then, we present some other data repositories and data collection/ingestion tools related to our work; finally, we draw some conclusions on the potential applications of our work.

The full code implementation of the pipeline is available under *Apache License Version 2.0* online at: https://github.com/DataSciencePolimi/NewsAnalyzer.

The dataset is available at: https://doi.org/10.7910/DVN/5XRZLH.

[1] http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

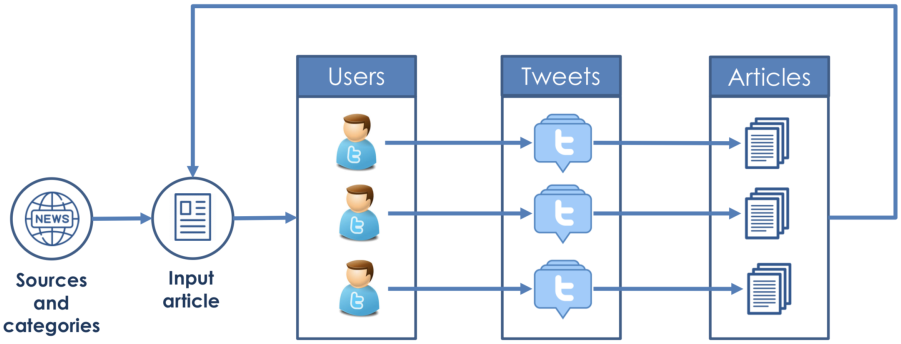


Figure 1: Overview of the data collection pipeline.

Data Collection Pipeline

In this section, we describe a tool which can be used to build custom collections of news articles and associated users and tweets. We also describe the settings for building the data collection presented in this work.

The entire framework is developed using `Python` (version 3.6.4): `Newspaper3k` library is used for news article crawling and extraction, `tweepy` is employed to interact with the Twitter Search API and `TextBlob` to compute polarity and subjectivity score. To build our enriched user model we use a few additional APIs, namely `Face++`, which offers a Face Detection tool, and `Yandex`, which allows to extract geographical information. In order to store the data we employ `MongoDB` which allows to store documents without a predefined structure (using JSON format).

The first module of the iterative pipeline takes as input a set of N articles URLs, gathered from a list of S predefined sources, which need to be manually specified by the user as to initialize the pipeline: they are crawled using the python library `Newspaper3k` and stored in the `MongoDB` database.

They are in turn used to query the Twitter Search API and extract the T most recent tweets which explicitly contain a link to some article, and save the accounts that tweeted them (which can be at most T distinct ones if each tweet is published by a distinct user). We further investigate the U most recent tweets from the Twitter history of these accounts, and select those tweets which contain URLs of other news of one of our sources, storing them along with their polarity and subjectivity score[2]. This last procedure allows to enrich the initial collection of URLs with new items which will be fedback to the initial module of the pipeline.

Parameters T and U control breadth and depth of the collection process: high T and low U entail a wide search over multiple users, collecting a small amount of tweets for each user; low T and high U result in a deep search on few users, collecting a large amount of tweets for each user.

In addition to these functionalities, we also provide support for category classification of articles, using Naive bayes, SVM, Logistic Regression and Random Forest classifiers, and bot score detection for users, using the `Botometer` API.

The entire process, which is detailed in Figure 1, is designed to run continuously within the limits imposed by the Twitter API. The aforementioned parameters N, S, T, U are manually specified when initializing the pipeline.

When building the data collection illustrated in the next section, we set $N = 2000$ (number of seed articles), $T = 100$ (maximum number of users retrieved for each article) and $U = 500$ (maximum number of recent tweets extracted from the timeline of each user). We uniformly sampled all the news sources as to obtain a set of 2000 articles which was used as initial input to the pipeline. This corresponds to news articles published in the last week of September 2018; by using the mechanisms previously described, we were able to extract articles which date back up to January 2018.

Data Description

The dataset that we provide is composed of 5 different entities: **news sources, news articles, news categories, tweets sharing the news, and users authoring the tweets**. Details on size and attributes of each entity are provided in Table 1.

Based on user features described in Table 1 we formalize an enriched user model which is built on four main dimensions:

1. a **social identity**, which includes demographic information (age, sex, ethnicity), the geo-localization and the Twitter profile;
2. features relative to **user-generated content**, which consist of statistics on the activity of the user (such as most frequent used words and hashtags, number of tweets/retweets, etc);
3. features derived from shared **news articles**, namely the distribution of user articles over sources, categories and topics and relative polarity;
4. the associated **Twitter network**, i.e. a list of followers and followees and the users engaged via mentions and

[2] The *polarity* score of a text is a value between -1 and +1 related to the negative or positive sentiment associated with the text. The *subjectivity* score of a text is a value between 0 and 1 related to how much the text transmits a personal and subjective thought.

Entity	Features	Size
Category	Name	8
Source	Name, URL	69
User	ID, Creation date, Description, Favourites count, Followers count, Language, Location, Geo-location Enabled (flag), Name, Profile Image Url, Screen Name, Statuses Count, Age, Ethnicity, Gender	37106
Article	ID, Author(s), Title, Text, Tags, Keywords, Publication Date, Category, URL, Article Source, Pipelined Flag	331769
Tweet	ID, Retweet Flag, Retweet Entities, Retweeted User ID, Coordinates, Creation Date, Entities, Favourite Count, Retweet Count, User ID, Screen Name, Language, Mentions, Article ID, Article URL, Article Source, Sentiment Polarity, Sentiment Subjectivity, Text	975788

Table 1: Description of the entities present in the dataset.

Sex	Male	55.2%
	Female	44.8%
Age	<20	1.6%
	20-30	20.4%
	30-40	20.5%
	>40	57.5%
Ethnicity	White	62.0%
	Black	21.5%
	Indian	9.3%
	Asian	7.1%

Table 2: Distribution of users by demographic information.

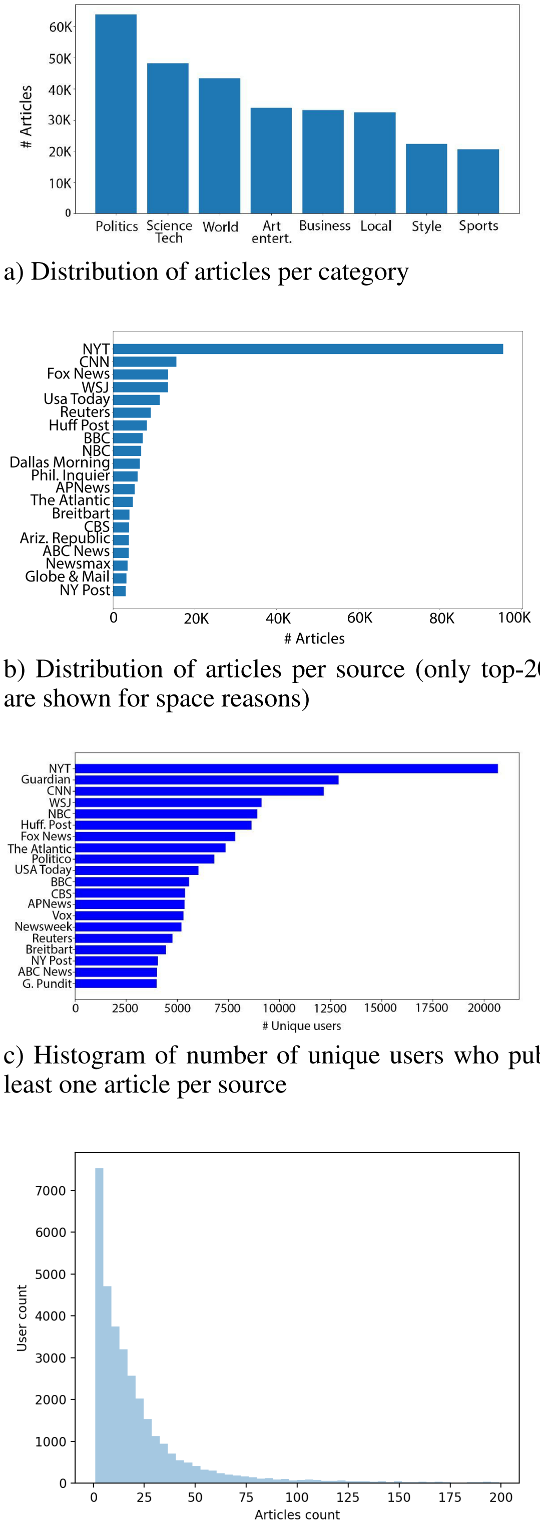
retweet.

To build the dataset, we manually selected a list of 69 popular U.S. online **news sources**, including 32 newspapers and 37 news agencies.

For 30 sources, we were able to extract the **category** of the extracted articles by looking at the URL structure; we inferred the others using a classifier built on several sets of features including full text, keywords and topics. Moreover, as different websites referred to the same categories using distinct labels, we manually aggregated them as to obtain a consolidated set of 8 categories: Politics, World, Business, Science/Technology/Health, Sports, National/Local News, Entertainment/Arts, Style/Food/Travel.

In Figure 2 we observe the distribution of **articles** by Category – with Politics being the most discussed –, by Source – it appears that ”The New York Times” is the most present mainstream outlet on Twitter – , by number of unique users that shared at least one article for each source and finally the distribution of users by number of articles shared – which follows a power law (with estimated coefficient 1.67), that is commonly peculiar of several social network statistics.

We provide in Table 2 the distribution of **users** by sex, age and ethnicity. Demographic enrichment based on profile



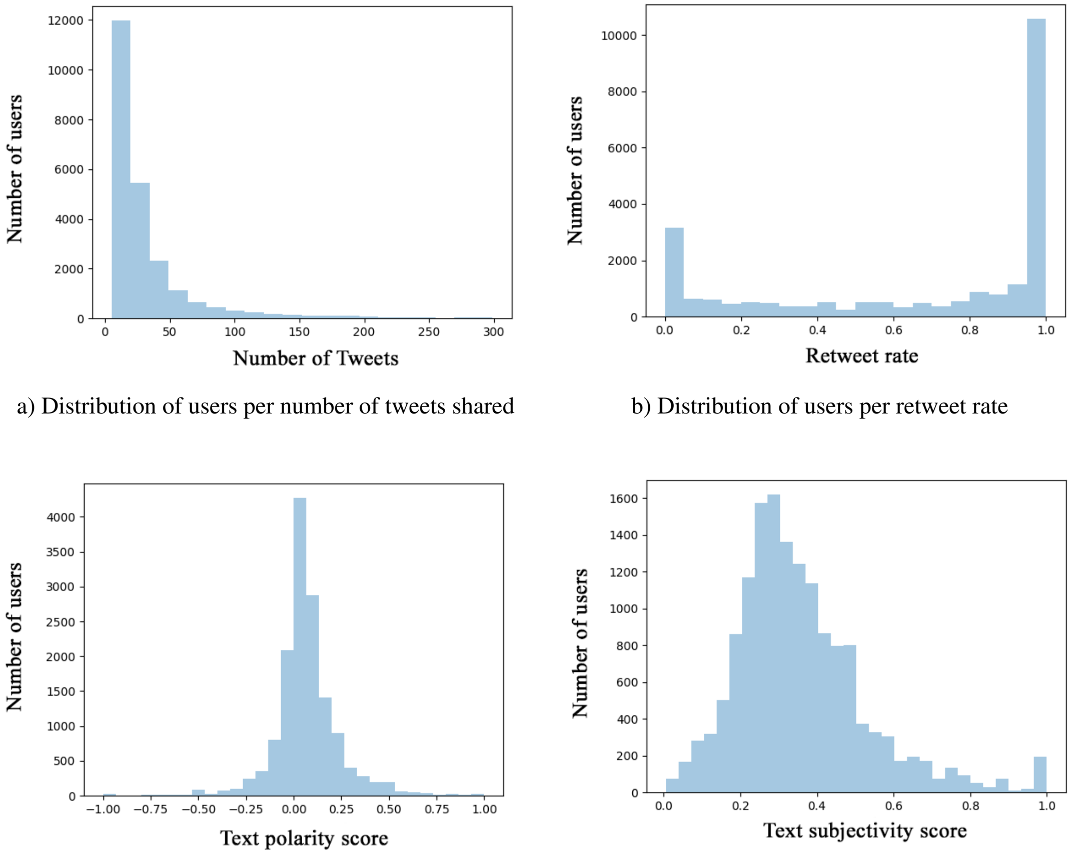
a) Distribution of articles per category

b) Distribution of articles per source (only top-20 sources are shown for space reasons)

c) Histogram of number of unique users who published at least one article per source

d) Histogram of number of users w.r.t the number of articles that they shared on Twitter

Figure 2: Statistics on articles, users and sources.



a) Distribution of users per number of tweets shared

b) Distribution of users per retweet rate

c) Distribution of users per average text polarity score

d) Distribution of users per average text subjectivity score

Figure 3: The distribution of users according number of tweets, retweet rate, average text polarity score and average text subjectivity score.

image was successful on 35% of the users while location extraction on 67%. We further considered US only users (approx. 15000) and checked whether our data truly reflected the actual population distribution in the US performing a Pearson correlation test with the population dataset provided by the U.S. Census Bureau in 2017, which held a coefficient value of 0.9 and p-value equal to $3.52e^{-20}$.

In Figure 3 we show the distribution of **tweets**, retweet rate, sentiment polarity and subjectivity by users. The first plot shows a power law distribution (with estimated parameter 1.64) with an average sample of ∼30 tweets per user which we believe is consistent to provide an estimate of user interests about news. The Retweet Rate instead highlights the existence of clearly separate groups with two evident peaks at 0 and 1, i.e. users who have a purely generative behavior (RT = 0) and those who spread only tweets from other users (RT = 1). The other two plots respectively show an overall neutral attitude (Sentiment Polarity) towards the generated content and a slight tendency to provide fairly objective texts within tweets w.r.t. express purely subjective thoughts.

Cluster name	Number of users
Republican Activists	1371
Democratic Activists	973
Republican Supporters	1032
Democratic Supporters	471
Targets	10545

Table 3: User groups and their cardinality.

Use Case: Political Stance Classification

In the following we describe a practical application of our data collection in the context of political stance classification. Given a set of users and the news articles that they shared on Twitter, the goal is to assess the proximity to the US Republican and Democratic parties. This use case shows that reaction to news can be used for interpreting important aspects of our society.

We started by manually collecting a list of 24 hashtags (12 for each party) which we assume are representatives of the two political factions, e.g. ”#bluewave” and ”#notmypres-

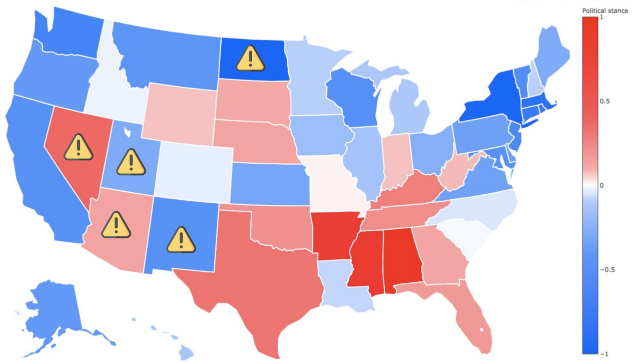


Figure 4: Prediction results for US 2018 Midterm elections with exclamation marks denoting wrong assignments.

ident” on the Democratic side and ”#maga” and ”#americafirst” for the Republican party.

Next, we used the following heuristic to label users in our dataset: we mark as ”Republican Activist” those who contain at least a Republican hashtag in their profile description and viceversa ”Democratic Activist” those who inserted at least a Democratic hashtag in the profile description; next we label those users who do not belong to the previous sets as ”Republican Supporters” if the Republican hashtags count in their tweets exceeds the Democratic hashtags count and ”Democratic Supporters” if the opposite holds; finally we label as ”Targets” those users who do not belong to any of the aforementioned categories. The cardinalities of each user group is described in Table 3.

We modeled the political stance as a continuous variable in the interval $[-1, 1]$, where the two extremes respectively indicate the maximum confidence towards Democrats and the Republicans. We evaluated several classifiers such as SVM, Random Forests and Logistic Regression and computed the political stance value for each user as the difference between the probabilities assigned by the model to Democratic class and Republican class. We used Supporters and Activists respectively as training and test sets and trained aforementioned models on different features, i.e. source, topics and categories. We observed the best performances when using all of them, achieving 90% accuracy – on the test set – with a Logistic Regression classifier.

For illustration purposes, we also used our predicted populations to anticipate the results of 2018 Midterm elections: we assigned Target users with known residence to states and then assigned to each state a label according to the average political stance of their residents (see Figure 4); only five states were incorrectly assigned in this way. Of course, this visualization is just for illustration purposes, as predicting electoral results goes beyond the objectives of this data collection.

Related Work

In the literature we can find a few recent contributions which are related to our work.

Hoaxy (Hui et al. 2018) is a running platform which has been conceived to track the diffusion on Twitter of news articles from a curated list of disinformation and fact-checking websites. Since the first introduction in 2016, this tool has collected millions of retweeted messages with links to thousands of articles from these domains.

NELA2017 (Horne et al. 2018) is a large political news data set which contains thousands of news articles collected from mainstream, satire, misinformation and hyper-partisan sources. They also compute a large set of content-based and social media engagement features which are meant to provide insights for different potential applications, such as news characterization, news attribution and content copying.

BuzzFace (Santia and Williams 2018) is a data collection which is composed of news stories posted on Facebook during September 2016. These articles were manually annotated by BuzzFeed journalists as to provide evidence in the context of news veracity assessment and social bot detection. Yielding over a million of text items, the collection provides different features including body text, images, links, Facebook and Disqus plugin comments.

FakeNewsNet (Shu et al. 2019) is a data repository, composed of hundreds of articles and thousands of social responses, which addresses the problem of fake news detection on Twitter. It includes a pipeline which automatically searches news articles based on the fact-checking activity of different organizations. For each item it provides several features relative both to news content and social engagement.

Similarly to Hoaxy and FakeNewsNet we provide a pipeline which automatically extracts information relative to news articles and Twitter interactions. However, they specifically focus on misinformation and fact-checking websites whereas our pipeline can be adapted to any news outlet. Moreover, they collect data in a real-time only fashion (using Twitter Streaming API) whereas our pipeline gathers data from users' timelines and it is not limited to present data.

NELA2017 and BuzzFace concern instead fixed sets of articles – whereas our collection can be dynamically updated – and are solely focused respectively on political and false news, two scenarios which can be easily reproduced with our pipeline.

Conclusions

In this paper we have presented a two-fold contribution: 1) a data collection pipeline which easily allows to build comprehensive collections of news articles and associated users and tweets from Twitter environment, starting from a pre-defined set of sources; 2) a comprehensive dataset (and relative descriptive statistics) which is conceived to describe the behaviours of social media users who are involved in the process of consuming and disseminating news items. We also showed a practical usage of this collection in the context of political stance classification.

We believe that our contribution may advantage several interesting applications in the future, from advanced users' profiling techniques which aim to characterize social media users based on news consumption aspects – such as clustering based on topics of interest – to large-scale studies of misinformation characterization and resolution – which take into account the news sharing behaviour of social media users.

References

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

Bruns, A., and Highfield, T. 2012. Blogs, twitter, and breaking news: the produsage of citizen journalism. In *Produsing Theory in a Digital World: The Intersection of Audiences and Production in Contemporary Theory*, volume 80. Peter Lang Publishing Inc. 15–32.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *International AAAI Conference on Web and Social Media*.

Hui, P.-M.; Shao, C.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018. The hoaxy misinformation and fact-checking diffusion network.

Jin, F.; Dougherty, E.; Saraf, P.; Cao, Y.; and Ramakrishnan, N. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 8. ACM.

Morgan, J. S.; Lampe, C.; and Shafiq, M. Z. 2013. Is news sharing on twitter ideologically biased? In *Proceedings of the 2013 conference on Computer supported cooperative work*, 887–896. ACM.

Raghavan, M.; Anderson, A.; and Kleinberg, J. 2018. Mapping the invocation structure of online political interaction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 629–638. International World Wide Web Conferences Steering Committee.

Santia, G., and Williams, J. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. *International AAAI Conference on Web and Social Media*.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2019. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Sunstein, C. R. 2001. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press.