Quality & Quantity
https://doi.org/10.1007/s11135-022-01448-z

Who pushes the discussion on wind energy? An analysis of self-reposting behaviour on Twitter

Loretta Mastroeni[1] · **Maurizio Naldi**[2] · **Pierluigi Vellucci**[1]

Accepted: 16 May 2022

Abstract
Discussions about wind energy and its environmental impact take place routinely over Twitter. Twitterers with a strong interest in the matter may also retweet their own tweets (aka self-reposting) as a means to increase their visibility and push their message across. Identifying the features that make self-reposted tweets different from tweets that are not retweeted (either by their originators or by other twitterers) is crucial to understand what drives self-reposting. In this paper, we examine several characteristics of self-reposted tweets, concerning when they occur, how frequently, their length, and the number of hashtags, hyperlinks, and exclamation points they contain. We conduct our analysis on a dataset comprising tweets about wind energy. We find out that: (a) twitterers repost their own tweets primarily on weekends (especially on Sundays) and in the afternoon; (b) self-reposted tweets tend to be longer and contain more hashtags; (c) self-reposting typically occurs when retweets by other twitterers become less frequent, probably driven by the need to refresh the message. Finally, we also observe that self-reposting is resorted to mostly by individual twitterers rather than companies.

Keywords Twitter · Retweeting · Social networks · Wind energy · Wind power

1 Introduction

Wind energy is one of the major renewable sources, but it has spurred a significant discussion about its environmental impact since its deployment, as investigated, e.g., in Saidur et al. (2011), Leung and Yang (2012), Dai et al. (2015). A large part of the debate has been channelled through social media, among which Twitter is probably the most popular,

✉ Pierluigi Vellucci
pierluigi.vellucci@uniroma3.it

Loretta Mastroeni
loretta.mastroeni@uniroma3.it

Maurizio Naldi
m.naldi@lumsa.it

[1] Department of Economics, Roma Tre University, Via Silvio D'Amico 77, Rome 00145, Italy

[2] Department of Law, Economics, Politics and Modern languages, LUMSA University, Via Marcantonio Colonna 19, Rome 00192, Italy

Published online: 10 June 2022

with 330 million monthly active users (see https://www.oberlo.com/blog/twitter-statistics), due to its short messages (aka tweets, shorter than 140 characters) (O'Reilly and Milstein 2011). Tweets may include *hashtags* (identified by the preceding # symbol), which allow associating a topic to the message: all messages related to a specific topic can be retrieved at once by searching for the associated hashtag. Users may decide to *follow* somebody and receive all his/her updates, i.e. by an opt-in mechanism.

Though anybody may send a message over Twitter, stakeholders in the wind energy debate may act more persistently to send their message through.

In fact, people do not constantly scroll Twitter all day and may miss tweets. Time zones further complicate the matter since a message sent in one's morning may be received by the recipient during his/her night, increasing the chance of missing it. Twitterers may therefore post their tweets again and again, on different days and at different times, increasing their chance of reaching their intended readers and hammering their message (like in the old Latin adage *repetita iuvant*). In the following, we refer to this practice as self-reposting to distinguish it from the general retweeting mechanism (at the basis of Twitter), where you share a tweet originally posted by others. Tools like Buffer (https:buffer.com, HootSuite (https:www.hootsuite.com), and Sprout Social (https://sproutsocial.com) make it easy easy to re-share the same tweet so as to publish it over and over. We guess this practice is heavily carried out by people having a strong interest in pushing the discussion about wind energy.

Twitter is aware of the practice and its potentially harmful consequences. The recent overhaul of its terms of service forbids users to schedule tweets with identical content across multiple accounts. This move was intended to get more original content on the platform and reduce spam and bots.

Though self-reposting is not necessarily spurred by obnoxious intentions, analysing its characteristics is helpful for several reasons. It helps build a deeper picture of twitterers' behaviour and the traffic they generate. It may also help Twitter's management staff tune their policies by identifying its features against the widely accepted retweeting mechanism. In our case, it helps us understand who wishes to take the lead in the discussion about wind energy over Twitter.

While some work has been conducted for mining textual content that users generate or analysing the social network structure in social perception on energy economics (Austmann and Vigne 2021; Reboredo and Ugolini 2018; Agarwal et al. 2020), no publication has studied the underlying mechanism of the retweeting behaviour yet. This lack is even more true in the context of renewable energies. This is not a secondary issue because retweeting became an established convention inside Twitter, and the literature sees it as a conversational practice in which "authorship, attribution, and communicative fidelity are negotiated in diverse ways" (Suh et al. 2010). For this reason, in our paper, we investigate the characteristics of self-reposting in the specific case of wind energy. We aim to understand where self-reposting is different from retweeting others' messages. Of course, we do not refer to their definition (i.e., retweeting one's own tweets rather than other people's ones). Instead, we wish to find out the meta-characteristics of messages (e.g. their length) and their occurrence (e.g., the day of the week when they are posted) that make them more likely to be self-reposted. In other words, we wish to identify the determinants of self-reposting.

In order to accomplish that task, we formulate the following three research questions (RQ):

RQ1 When do self-reports occur?

RQ2 Does retweeting frequency impact self-reposting behaviour?
RQ3 Are tweets meta-characteristics associated with self-reposting?

For our purpose, we employ both an exploratory analysis and a machine learning-based classification approach (through a decision tree), employing a dataset obtained by collecting tweets over the topic of wind energy.

After an analysis of the literature (Sect. 2), we report our original contribution:

- Self-reposting is mostly carried out by individual twitterers, in some cases having their account suspended due to some violation of the norms of conduct;
- International or governmental agencies, associations, and organisations resort to self-reposting less massively, while companies involved in building or managing wind power plants play quite a negligible role;
- The most relevant features to discriminate self-reposted messages vs non-self-reposted ones appear to be the day of the week, the time of the day, the length of the tweet, the number of hashtags, and the intertweet time interval;
- Weekends (especially Sundays) are the days of choice for self-reposting;
- Self-reposting takes place in afternoons mostly;
- Messages longer and containing more hashtags are more likely to be self-reposted;
- Twitterers tend to repost their tweets when their retweeting by others gets less frequent.

2 Literature review

We wish to analyse self-reposting behaviour as opposed to retweeting and consider wind energy as a debated topic where we observe self-reposting in action. In this section, we review the literature concerning retweeting behaviour and the social perception of wind energy separately. We also analyse the use of social media in connection to energy and environmental issues.

2.1 Retweeting

Social media have attracted a great deal of attention as non-conventional sources of data. They contain timely data capable of capturing public awareness and insights into human behaviours. Among them, Twitter is undoubtedly one of the most used to investigate a wide range of topics, such as investor sentiment (Reboredo and Ugolini 2018), social impact assessment (Sherren et al. 2017), NIMBY conflicts (Wang et al. 2019, 2021), social perception of energy (Li et al. 2019; Austmann and Vigne 2021) and social appreciations (Resce and Maynard 2018).

A way of expressing the support for an opinion expressed through a tweet is by retweeting it. The reasons for retweeting have been investigated, e.g., by Boyd et al. (2010), Webberley et al. (2011). Several factors have been identified in the literature to influence retweeting: number of followers (Kim et al. 2016; Li and Liu 2017), content features such as hashtags and URLs (Pang and Law 2017), linguistic features (Wang et al. 2012), topic and/or user virality (Hoang and Lim 2013; Li and Liu 2017; Park and Kaye 2019; Shi et al. 2017), users' and tweets' emotional status (Chen et al. 2017; Kim et al. 2016; Park and Kaye 2019), social behaviors performed by the sender and the receiver of the retweet (Chen and Deng 2020; Shi et al. 2017), different types of network structures (Kim et al.

2016). Though retweets often represent uncritical support of the original tweet, Gruber (2017) has analysed the commented variant, where the retweeter (critically) evaluates content (or author) of the retweeted text.

Several approaches have been proposed to model and predict the retweeting dynamics. Gao et al. (2015) propose an exponential reinforcement mechanism characterising the "richer-get-richer" phenomenon by capturing a power-law temporal relaxation function corresponding to the ageing in the ability of the message to attract new retweets. Wang et al. (2013) employ the classic Susceptible-Infectious-Susceptible (SIS) epidemic model. Other approaches are based on networks (Achananuparp et al. 2012), classification trees with recursive partitioning procedure (Nesi et al. 2018), social exchange theory and game theory (O'Leary 2016), factor graph model (Yang et al. 2010), probabilistic matrix factorization (Zhang et al. 2016). Other papers, like, e.g. Lin et al. (2016); Hu et al. (2018); Li et al. (2013); Xiong et al. (2012); Wang and Lee (2020); Kim et al. (2014); shin Lim and Lee-Won (2017); Cruz and Lee (2014) focus on the importance of retweeting, analysing in some cases the evolution of its propagation.

With the rise in popularity of social media platforms like Twitter, exerting great influence on such platforms allows influencing many decisions and shaping many opinions. (Gao et al. 2016) has shown that the popularity of tweets is distributed very unevenly, with a handful of key nodes dominating the scene and gaining a high number of retweets. A few efforts have been made in the literature to describe the phenomenon of *influencers*, who act through a constant production of data, attempting to make themselves algorithmically recognisable (Gillespie 2017) by social media algorithms and so respond to the coercive force of the *threat of invisibility* (O'Meara 2019; Cotter 2019). However, some bad practices have emerged, representing a threat to the credibility of these social networking platforms. For example, Arora et al. (2020) has investigated collusive retweeting activities, e.g., those services provided by paying for them to gain influence inorganically. The impact of spammers on Twitter networks has instead been addressed by Fronzetti Colladon and Gloor (2019), while Liu et al. (2019) has analysed the crowd-retweeting spam.

However, the literature does not appear to have investigated self reposting, so far, i.e., the practice by which influencers reshare their own same content multiple times (as if it were original) to increase their retweet count and their visibility.

2.2 Wind energy

Most of the literature on wind energy concerns its diffusion and the environmental policies adopted by governments.

The need to move towards sustainable, low-carbon and affordable energy systems is driven by the urgency to reduce greenhouse gas emissions and address global environmental problems. The required international efforts to combat climate change may include a shift in the national energy mix towards renewable energies and a new portfolio of electricity generation technologies (Hoffert et al. 2002; De Jesus et al. 2018; Dhakouani et al. 2019; Aleixandre-Tudó et al. 2019). An important contributor to that energy transition is the development and implementation of a wind energy infrastructure, as shown by Muñoz and Márquez (2018), Stephens et al. (2009), Zhao et al. (2016), Jethani (2016), both in its onshore and offshore implementations as described in Weinzettel et al. (2009), Sun et al. (2012). Countries are taking steps in that direction. We cite, e.g., the USA (where 29 states require minimum levels of wind generation through renewable portfolio standards Lamy et al. (2020)), Switzerland (where citizens approved a national energy strategy in

May 2017, and wind energy plays a fundamental role Vuichard et al. (2019)), and Denmark (where parliament has agreed to promote the establishment of a longer-term goal of satisfying 50% of Denmark's electricity needs through wind power by 2020 Borch et al. (2020)). From the year 2013 to the end of 2019, the total amount of wind energy capacity grew up at a rate of 12.6%, helped by steadily falling generation costs and an increase in the size of wind turbines as well as a low average construction time (Ali Sayigh 2020).

However, all renewable energy options have their own negative impacts, and wind power is no exception. Acceptance-related issues have hindered the deployment of wind energy. For example, people living near such projects may oppose their development (Horbaty et al. 2012; Vuichard et al. 2019). We are talking about the onshore wind farms, which are still the most popular type of wind farm in the world, but also the offshore wind farms seem to spur signs of conflict in the local communities (Lamy et al. 2020; van der Loos et al. 2020). In short, the social acceptance of wind plants continues to be the key challenge facing the governments and industry (Frantál 2015).

2.3 Discussion on social media

The use of social media in relation to energy and environmental issues has a relatively long history in the literature. Social media play various roles in that relationship, where social media users may play both active and passive roles.

An example of the former is the use of social media to express protests and one's social activism. Valenzuela (2013) have analysed that role in the framework of the massive demonstrations taking place in Chile to change the government's energy policy.

Social media also exert an influence on one's attitudes towards energy sources. Such influence may be driven by an intentional behaviour to influence other social media users or by the sheer pressure of the crowd of users expressing their opinion. The variety of ways whereby social media can develop their influence has been explored by Sivarajah et al. (2015) in the context of energy efficiency practices. More recently, Zobeidi (2020) has shown that trust in the information obtained from social media affects the attitudes towards conventional energy/fuels. Luedecke and Boykoff (2017) have shown that social media are largely overcoming traditional media in that role.

The overall sentiment about energy sources can now be measured on social media. Jain and Jain (2019) has employed Twitter as a data source to measure the sentiment towards renewable energy sources. Similarly, Boutakidis et al. (2014) has used questionnaires to elicit the opinions of social media users on the same issue.

Opinions expressed on social media can also be used as a proxy for physical measurements. Wang et al. (2017) have proposed to derive an environmental quality index from those opinions. For example, metrics could be obtained to measure water quality and air pollution.

3 Method

In this section, we describe the method we have applied to the analysis of tweets. We first provide a mathematical definition of self-reposting and its related variables. We then describe how we collect our data and the dataset we obtain.

3.1 Self-reposting of tweets

As hinted in the Introduction, we are interested in the phenomenon of self-reposting, i.e. in people posting again a message they have posted on Twitter in the (recent) past. Self-reposting is distinguished from the well-studied phenomenon of retweeting, where the same message is posted again by somebody other than the original twitterer. We can also refer to the same phenomenon as self-retweeting but in the following, we stick to using *self-reposting* to avoid confusion. In this section, we provide a mathematical description of the phenomenon.

We represent the set of all tweets by S. Tweets may be original or reposted. In the latter category, we include both tweets self-reposted by the author of the original message and tweets retweeted by somebody else. We group the original tweets in the set S_T and the reposted ones in the set S_R. Those two are subsets of S and represent a partition of S, so that $S = S_T \cup S_R$ and $S_T \cap S_R = \emptyset$.

Let N be the cardinality of S_T and M be the cardinality of S_R.

The generic element t_i of S_T, $i = 1, 2, \ldots, N$ is a 3-tuple $t_i := (t_i^{\text{text}}, t_i^{\text{time}}, t_i^{\text{user}})$, where t_i^{text} is the text enclosed in that tweet, t_i^{time} is the concatenation of date and UTC time when that tweet was created (e.g. *2020-03-18 11:00:29*), and t_i^{user} is the user who posted the tweet.

As to the set S_R, we may distinguish retweets based on the identities of the retwitterer and the original twitterer. If those identities coincide, we have self-reposting, where the twitterer of the original tweet posts it again. We group those self-reposted tweets in the S_{SR} subset. We collect those retweets that are instead retweeted by another twitterer in the S_{RR} subset. Those two subsets represent a partition of S_R, so that $S_R = S_{RR} \cup S_{SR}$ and $S_{RR} \cap S_{SR} = \emptyset$. We denote the cardinalities of the two subsets respectively as M_{RR} and M_{SR}, with $M = M_{RR} + M_{SR}$.

The generic element rr_j of S_{RR}, $j = 1, 2, \ldots, M_{RR}$, is a 4-tuple $rr_j := (rr_j^{\text{text}}, rr_j^{\text{time}}, rr_j^{\text{user}}, rr_j^{\text{orig}})$, where rr_j^{text} is the reposted text, rr_j^{time} is the day and time of reposting, rr_j^{user} is the user who reposted the tweet, and rr_j^{orig} is the user who posted the original tweet. Since all the elements in S_{RR} and S_{SR} are repostings of tweets in S_T, the following statement holds:

$$\forall j \in (1 : M_{RR}) \quad \exists! i \in (1 : N) / rr_j^{\text{text}} = t_i^{\text{text}}, rr_j^{\text{orig}} = t_i^{\text{user}}, rr_j^{\text{time}} > t_i^{\text{time}}. \tag{1}$$

Figure 1 illustrates the relationship between those sets.

We can now define the set S_{SR} of self-reposted tweets, i.e. those tweets that are reposted by the original twitterer itself. The generic element sr_j of S_{SR} is represented by the 3-tuple $sr_j := (sr_j^{\text{text}}, sr_j^{\text{time}}, sr_j^{\text{user}})$ with obvious meaning of the elements. We do not need a fourth element here since the original twitterer and the retwitterer coincide. The following statement holds similarly to what we have stated for S_{RR}

$$\forall j \in (1 : M_{SR}) \quad \exists! i \in (1 : N) / sr_j^{\text{text}} = t_i^{\text{text}}, sr_j^{\text{user}} = t_i^{\text{user}}, sr_j^{\text{time}} > t_i^{\text{time}}. \tag{2}$$

These are exactly the subsets we are interested in, i.e., the tweets making the self-reposting phenomenon. Since we are interested in the union of the two sets S_T and S_{SR}, we can conveniently define that union as $S_{TSR} = S_T \cup S_{SR}$. Each element $s_k \in S_{TSR}$ is defined by the 3-tuple $s_k := (s_k^{\text{text}}, s_k^{\text{time}}, s_k^{\text{user}})$. It is useful to order those elements by time, so that $s_{k+1}^{\text{time}} > s_k^{\text{time}}$. Each element in S_{TSR} is an original tweet or a self-reposted one. For each element $s_k \in S_{TSR}$ we can identify the subset $S_{RR}^{(k)}$ of its retweets by other twitterers, i.e. all the elements $rr_j \in S_{RR} : (rr_j^{\text{text}} = s_k^{\text{text}}) \wedge (s_k^{\text{time}} < rr_j^{\text{time}} < s_{f(k)}^{\text{time}})$, where $f(k)$ is the function pointing to the next element in S_{TSR} that is a self-repost of s_k, i.e. $f(k) = \min j : (s_j^{\text{text}} = s_k^{\text{text}}) \wedge (s_j^{\text{time}} > s_k^{\text{time}})$. If the tweet s_k is the last self-repost in its

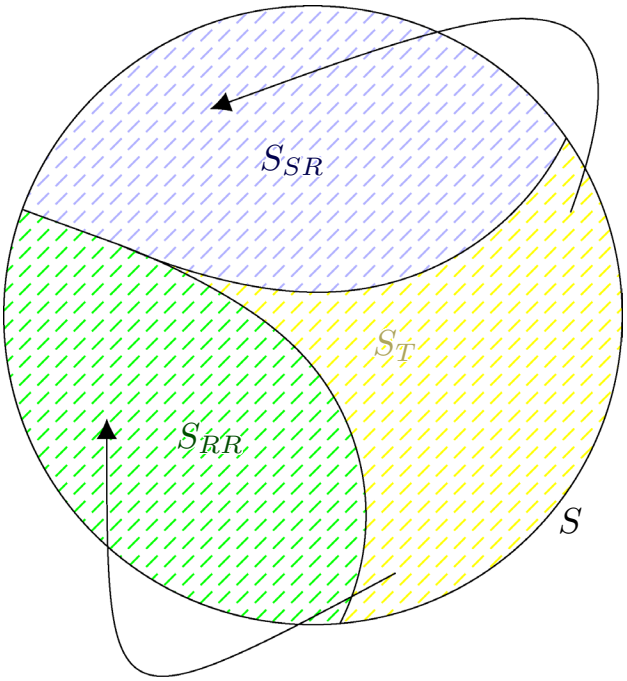

Fig. 1 The set of all tweets, S. S_T (yellow), $S_{SR} = \cup_{i=1}^{N} S_{SR}^{(i)}$ (blue) and $S_{RR} = \cup_{i=1}^{N} S_{RR}^{(i)}$ (green) form a partition of S. We can spot the presence of two different applications that map a tweet in S_T to a self-repost in S_{SR} or a retweet in S_{RR}

series, then the subset $S_{RR}^{(k)}$ is made of all the tweets $rr_j \in S_{RR} : (rr_j^{\text{text}} = s_k^{\text{text}}) \wedge (rr_j^{\text{time}} > s_k^{\text{time}})$. It is also expedient to define some other quantities that will be useful when we later define the features we employ. For each subset $S_{RR}^{(k)}$ we define the ordered set $U^{(k)}$ of all the indices of the elements of $S_{RR}^{(k)}$, i.e. the set of all j for which $rr_j \in S_{RR}^{(k)}$. The generic element of $U^{(k)}$ is $u_i^{(k)}$, $i = 1, 2, \ldots, |S_{RR}^{(k)}|$. We also introduce the set $D^{(k)}$ of all time differences between subsequent retweets of the same tweet, whose elements are $d_i^{(k)} = rr_{u_{i+1}^{(k)}}^{\text{time}} - rr_{u_i^{(k)}}^{\text{time}}$, $i = 1, 2, \ldots, |S_{RR}^{(k)}| - 1$. We can finally introduce the rank of each time difference, i.e., the set $RD^{(k)}$ of all $rd_j^{(k)} = \#d_i^k \quad : \quad d_i^{(k)} \geq d_j^{(k)}$.

3.2 Twitter scraping

We have scraped the tweets by Twitter's API (Application Programming Interface), accessible upon opening a Twitter developer account. Tweets were retrieved using the R package twitteR developed by Gentry (2015). The search interval has a 7-day limit, which means that only tweets posted in the latest seven days can be retrieved. Our inspection interval falls in the weeks from November 30, 2019, to March 23, 2020, for an overall collection period of 16 weeks.

We have focussed our investigation on tweets concerning wind power. For that reason, we have searched for all the tweets containing either of the following word combinations:

- *Wind* AND *power*;

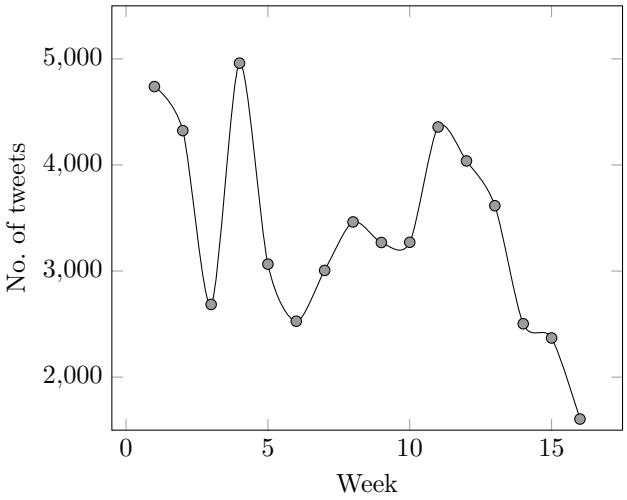

Fig. 2 Number of relevant tweets by week

- *Wind* AND *energy*.

Since many tweets contain both the above combinations, our basket after the retrieval phase may contain duplicates (i.e., tweets exhibiting the same 3- or 4- tuple identifying the tweet as described in Sect. 3.1). Before proceeding further, we have removed all duplicates.

We must, however, recognize the possibility of including tweets that, despite containing the words *wind* or *power* or *energy*, are not relevant to our actual theme, i.e., the use of wind to get electrical power.

The number of relevant tweets by week is shown in Fig. 2.

3.3 Dataset

As previously mentioned, our dataset has been built from tweets in English posted over roughly four months. This dataset contains 51,580 tweets ($|S| = 51,580$). The subsets of original and self-reposted tweets amount to 11,723 tweets. Precisely, we have $|S_T| = 10,230$ and $|S_{SR}| = 1,493$.

The target variable, which our supervised learning engine tries to predict, is the status of being self-reposted. Namely, the target variable *self_repost* is 1 if the tweet has been self-reposted and 0 otherwise. The definition of the target variable is then as follows: Fix $i = 1, 2, \ldots, N$ and consider $t_i \in S_T$. Let us denote by $S_{SR}^{(i)}$ the set of all sr_j that satisfy Eq. (2). Then, intuitively, the union of $S_{SR}^{(i)}$ and $\{t_i\}$ is the set of tweets with *self_repost* $= 1$. Let us now denote by $|S_{SR}^{(i)}|$ the cardinality of this set and define:

$$\begin{aligned} \mathcal{I}_0 &\triangleq \left\{ i \in \left(1 : \left(N + M_{SR}\right)\right) \middle| |S_{SR}^{(i)}| = 0 \right\} \\ \mathcal{I}_1 &\triangleq \left\{ i \in \left(1 : \left(N + M_{SR}\right)\right) \middle| |S_{SR}^{(i)}| \geq 1 \right\} \end{aligned} \tag{3}$$

Accordingly, the sets of tweets with *self_repost* $= 0$ and *self_repost* $= 1$ are, respectively $\left\{ t_i \middle| i \in \mathcal{I}_0 \right\}$ and $\left\{ t_i \cup S_{SR}^{(i)} \middle| i \in \mathcal{I}_1 \right\}$.

For each tweet in S_{TSR}, we have nine features, obtained either by simple retrieving from the scraped data or by processing those data. The features of the generic element $s_k \in S_{TSR}$ are:

- The retweet count $n_{\mathrm{RT}}(k)$, i.e. the number of times the tweet has been retweeted before being reposted, which is precisely defined as

$$n_{\mathrm{RT}}(k) = |S_{RR}^{(k)}|. \tag{4}$$

- The average time $\bar{t}(k)$ between two retweets, which can be computed as the ratio of the time between two subsequent self reposts to the number of retweets:

$$\bar{t}(k) = \frac{s_{f(k)}^{\mathrm{time}} - s_k^{\mathrm{time}}}{|S_{RR}^{(k)}|}. \tag{5}$$

- The weekday $wd(k)$, i.e., the day of the week when the tweet has been posted, extracted from s_k^{time};
- The time $dt(k)$ of the day when the tweet has been posted, extracted again from s_k^{time};
- The length $nc(k)$ of s_k^{text} in characters;
- The number $nh(k)$ of hashtags contained in s_k^{text};
- The number $nl(k)$ of hyperlinks (URLs) contained in s_k^{text};
- The number $ne(k)$ of exclamation points contained in s_k^{text}.
- The rank $o(k) = rd^{(k)}_{|S_{RR}^{(k)}|-1}$ of the last inter-retweet interval (i.e., after sorting all the inter-retweet intervals since the latest self-repost).

4 Determinants of self-reposting

In this section, we analyse the dataset to extract the most relevant features associated with self-reposting. We carry out a dual approach. We first employ an exploratory analysis, which may give us a clue about which features take different values in self-reposted vs non-self-reposted tweets. Then we apply a machine learning approach based on a decision tree to classify tweets by employing the nine features we have described in Sect. 3.3.

4.1 Exploratory analysis

Before applying a thorough machine learning approach, we conduct an exploratory analysis by seeing if any single feature actually takes different values for self-reposted and non-self-reposted tweets. For the feature to be really useful in discriminating between the two types of tweets, hence characterizing self-reposting, we expect it to behave differently.

We start by examining how frequent self-reposting is. We can draw two metrics from the composition of the dataset illustrated in Sect. 3.3. First, out of all original tweets, those that are self-reposted are $|S_{SR}|/|S_T| = 14.6\%$, which is a significant fraction. On the other hand, we may consider self-reposting as a special case of retweeting, i.e., where the twitterer tweeting the original tweet and that tweeting the retweet coincides. In that case, we can compute the fraction of retweets that are actually self-reposts as $|S_{SR}|/(|S_{SR}| + |S_{RR}|) = 3.6\%$.

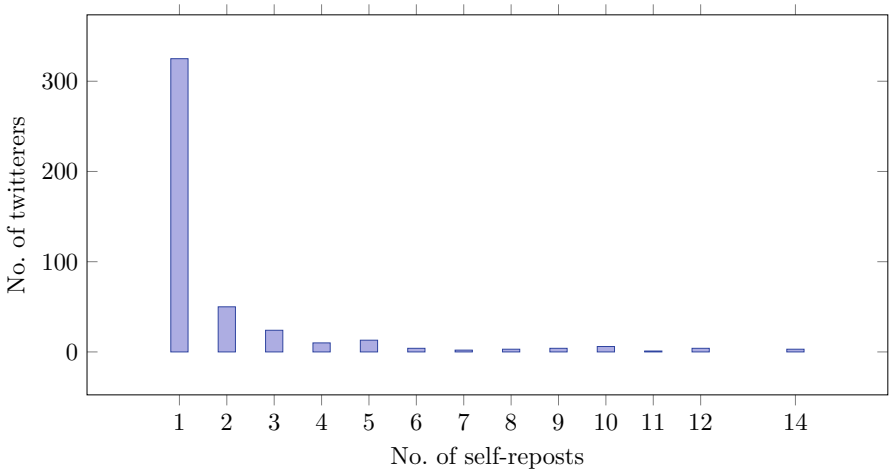

Fig. 3 Distribution of twitterers by the number of self-reposts carried out

We can also see how much self-reposting is resorted to by each twitterer. In Fig. 3, we see that the overwhelming majority of twitterers limit themselves to reposting their tweet twice, but there are (rare) cases where the twitterer applies self-reposting massively. Though we reported just the cases up to fourteen retweets, in a small percentage of cases (4.46% of the total), the authors have reposted their same post more than fifteen times. We mention the cases of single users who reposted the same tweet 35, 40, 50 or even 114 times. The last case concerned a tweet reposted by the same author nearly every day, roughly at the same time, but without ever earning a retweet.

If we look at the users who have applied self-reposting massively, we notice that in the top 20, there are: only 1 (small) company; 1 expert; 1 climate activist; 3 leading international or governmental agencies, associations, organizations; 3 users who seem not to be on Twitter any longer; 2 suspended accounts[1]; 9 generalist individual twitterers. The latter is a class of twitterers that promote websites on cryptocurrencies and appear to be related (they publish or retweet the same posts); their behaviour is reminiscent of social bots.

We can now examine the impact of each feature on self-reposting behaviour. In the following, we consider the two values of the target variables (i.e., being a self-repost or not) and report the statistical characteristics of each feature for the two values of the target variable.

We start with the number of retweets (by twitterers other than the original poster) taking place between two subsequent self-reposts. In the case of non-self-reposted tweets, we consider the overall number of retweets instead as a comparison. We report the two resulting probability density functions in Fig. 4. The comparison does not show significant differences.

If we take a look at how often retweets are posted in Fig. 5, we see that retweets after self-reposts take place more frequently than for tweets that are not self-reposted.

[1] Twitter suspends accounts that violate Twitter Rules (i.e. violence, terrorism, child sexual exploitation, abuse and harassment, hateful conduct, self-harm and suicide, sensitive content, illegal or regulated goods and services)

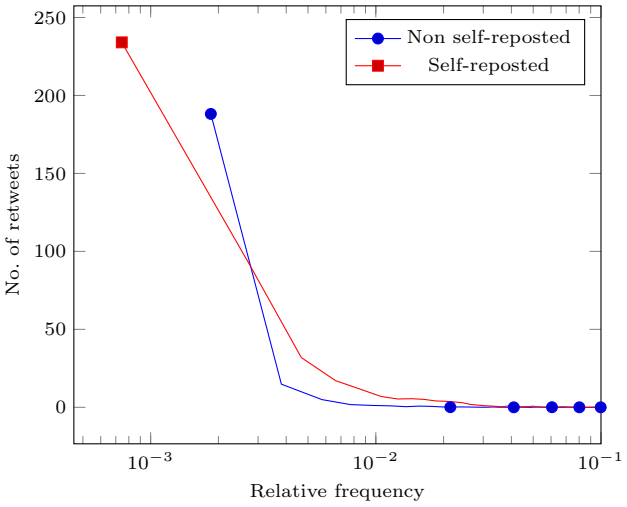

Fig. 4 Relative frequency of retweet count

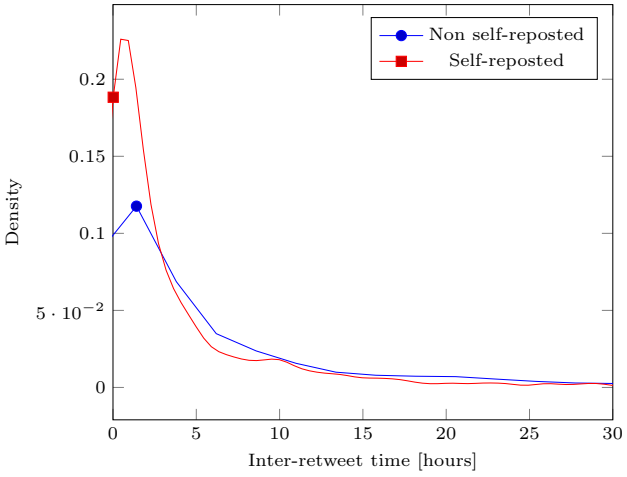

Fig. 5 Density plot of the average time between two successive retweets of the same tweet

However, the mode of the density (the most frequent inter-retweet time) is quite similar for both cases, being located at 0.47 hours (for self-reposted tweets) and 1.41 hours (for non-self-reposted tweets), respectively.

We can now examine *when* retweets are posted. We start with the day of the week. In Fig. 6. Here we see significant differences between the two subsets. Non-self-reposted tweets occur more frequently during weekdays, while self-reposts peak on Sunday. We can imagine a reviewing habit taking place for twitterers who review the status of their posts on Sundays and then decide whether to self-repost their tweets.

We can go deeper by examining the time of the day at which tweets are posted. In Fig. 7, we see that the distribution is far from uniform but different for the two cases.

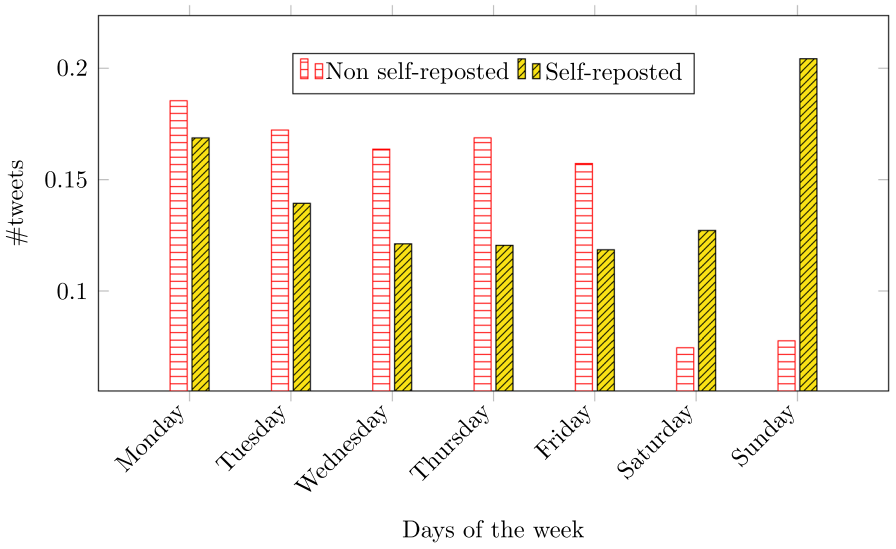

Fig. 6 Distribution of tweets by the week day in which they are posted

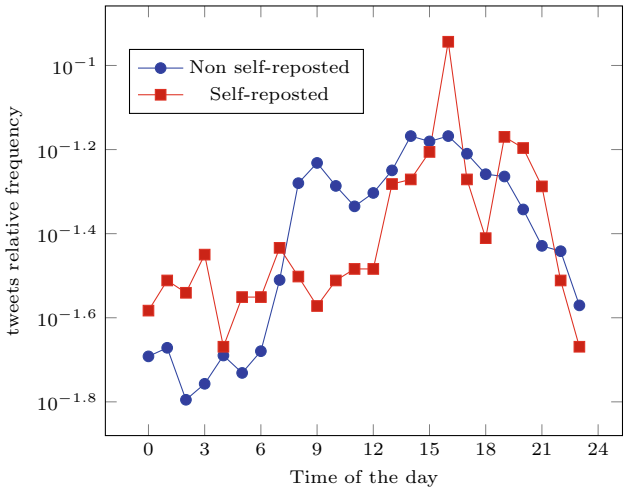

Fig. 7 Distribution of tweets by the hour of the day in which they are posted

Tweets that are not self-reposted take place relatively uniformly during the daytime. Self-reposts concentrate in the afternoon hours.

We can now move to the actual features of the tweet itself.

We first consider the length of the tweet, reported in Fig. 8. We see that there are no highly significant differences, except some occasional peaks around shorter lengths for tweets that are self-reposted.

Another important feature of tweets is the number of hashtags they contain. In Fig. 9, we see that self-reposted tweets have more embedded hashtags than non-self-reposted

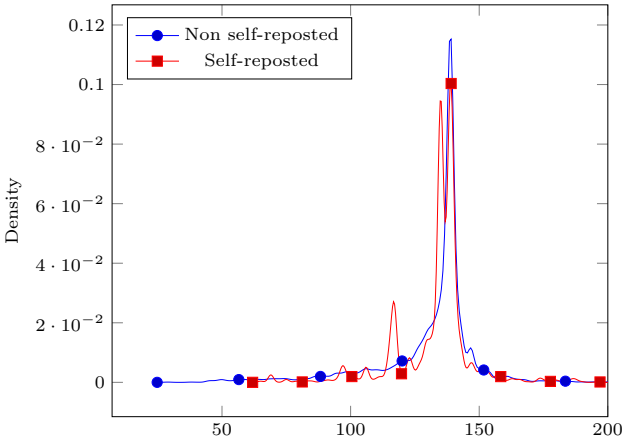

Fig. 8 Density plot of tweets length in characters

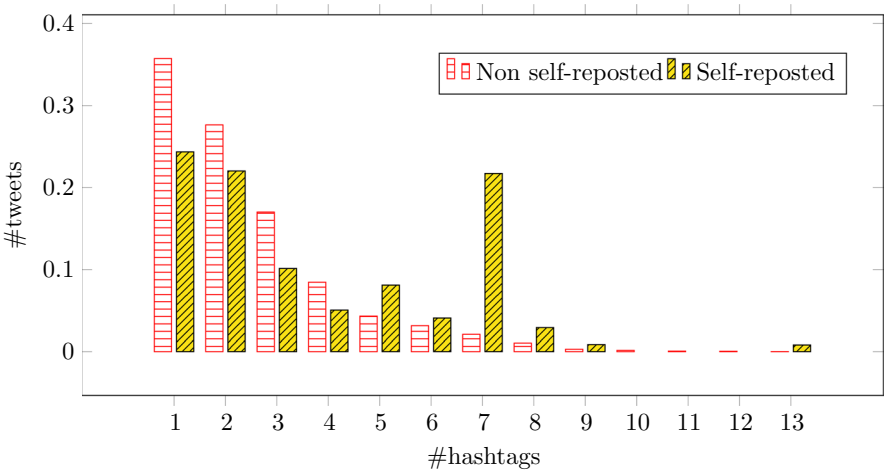

Fig. 9 Distribution of tweets by the number of hashtags contained in the tweet

ones. We can conclude that twitterers who repost their own tweets are more sensitive to catching readers' attention.

Instead, no significant differences appear if we turn to another embedded element such as hyperlinks. In Fig. 10, we see that the overwhelming majority of tweets contain just one hyperlink, with most of the remaining tweets containing just two. Though the presence of a single hyperlink is more frequent for self-reposts (and the opposite can be said for the case of two hyperlinks), the number of hyperlinks embedded in tweets does not seem a helpful feature to discriminate between self-reposts and non-self-reposted tweets.

We can reach the same conclusion for the number of exclamation points, which is shown in Fig. 11. Here, we have no exclamation points for 94% of the non-self-reposted tweets and 96% of self-reposts, quite a slight difference to be considered for discrimination (Fig. 12).

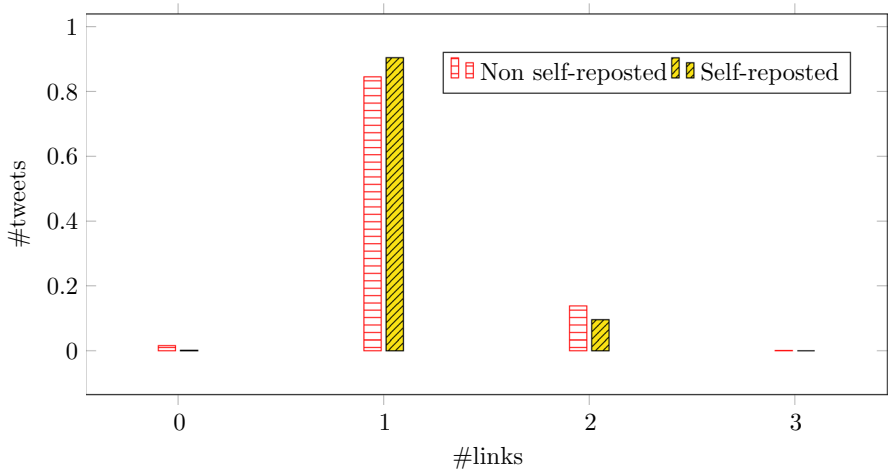

Fig. 10 Distribution of tweets by the number of links which they contain

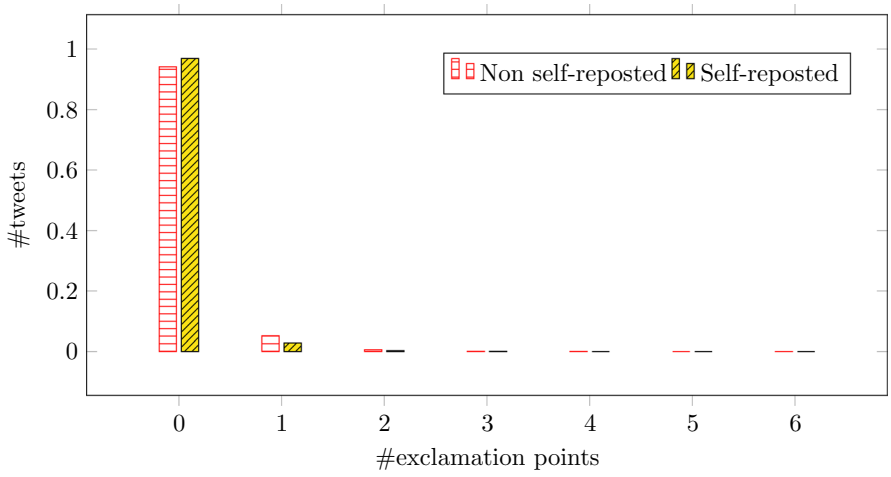

Fig. 11 Distribution of tweets by the number of exclamation points which they contain

4.2 Classification metrics

After performing the exploratory analysis, we now consider a machine learning approach to identify the determinants of self-reposting behaviour. The task of predicting whether a message is going to be self-reposted can be seen as a binary classification task, where we label self-reposted messages as positive instances. The resulting confusion matrix is reported in Table 1.

In order to evaluate the performance of our classifier, we employ the following metrics, which are well established in the literature:

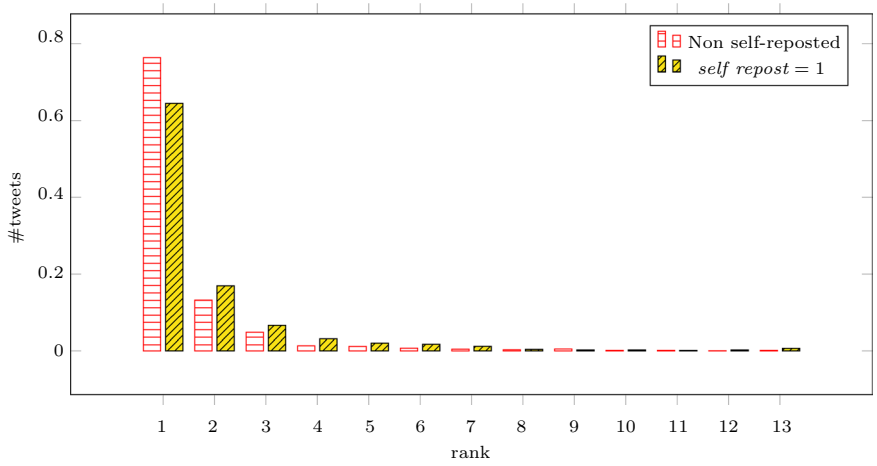

Fig. 12 Distribution of tweets by rank feature

Table 1 Confusion matrix

		Predicted	
		Self-reposted	Non self-reposted
Actual	Self-reposted	TP	FP
	Non self-reposted	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{9}$$

Since our dataset is imbalanced, accuracy may not be the best performance metric; hence we will mostly rely on the true positive and true negative rates (i.e., Sensitivity and Specificity).

4.3 Classification methodology

The task we are considering is a classification task, where we try to link the features identified in Sect. 3.3 to the target variable *self_repost*.

We employ the labelled dataset described in Sect. 3.3. We consider inputs $\mathbf{x}$ and output y, merged in the training set $\mathcal{D} = \left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^{N}$, where N is the number of training tweets. In our case, $\mathbf{x}$ contains the values of the 9 features introduced in Sect. 3.3 and y_i is the binary categorical variable *self_repost* $\in \{0, 1\}$. Formally, given the training set $\mathcal{D}$ of N tweets, the k-th tweet is represented by 9 features and a target variable:

$$\left(n_{\mathrm{RT}}(k),\, \bar{t}(k),\, wd(k),\, dt(k),\, nc(k),\, nh(k),\, nl(k),\, ne(k),\, o(k),\, self\_repost_k\right) \tag{10}$$

We opt for a classification tree as the classification model due to its simplicity and easy interpretability. The tree is built using the greedy procedure shown in Algorithm 1, which looks for the optimal partitioning of the data. This method is used by CART (Breiman et al. 1984; Murphy 2012), and its popular implementation (Therneau et al. 1997; Therneau and Atkinson 2019; Kuhn 2020, 2008). The split function chooses the best feature and the best value for that feature, as follows:

$$\begin{aligned}(j^*, t^*) = \arg \min_{j \in \{1,\ldots,D\}} \min_{t \in \mathcal{T}_j} \Big[& \mathrm{cost}\left(\left\{\mathbf{x}_i, y_i : x_{ij} \le t\right\}\right) + \\ & + \mathrm{cost}\left(\left\{\mathbf{x}_i, y_i : x_{ij} > t\right\}\right)\Big]\end{aligned} \tag{11}$$

where $\mathcal{T}_j$ is the set of possible thresholds for feature j; it can be obtained by sorting the unique values of x_{ij}. For example, if feature 1 has the values $\{4.5, -12, 72, -12\}$, then we set $\mathcal{T}_1 = \{-12, 4.5, 72\}$. For real-valued inputs, it makes sense to compare a feature x_{ij} to a numeric value t. In the case of categorical inputs we consider splits of the form $x_{ij} = c_k$ and $x_{ij} \neq c_k$, for each possible class label c_k. The training set $\mathcal{D}$ is then splitted in two subsets, the left subtree $\mathcal{D}_L$ and the right one $\mathcal{D}_R$, using a single feature j and a threshold t.

Algorithm 1 Recursive procedure to grow the classification tree

```
function fitTree(node, D, depth)
node.prediction = mean (y_i : i ∈ D)
(j*, t*, D_L, D_R) = split(D)
if not worthSplitting (depth, cost, D_L, D_R) then
    return node
else
    node.left = fitTree(node, D_L, depth+1)
    node.right = fitTree(node, D_R, depth+1)
    return node
end if
```

The function that checks if a node is worth splitting uses a stopping heuristic based on the reduction in cost Δ:

$$\Delta \triangleq \mathrm{cost}(\mathcal{D}) - \left(\frac{|\mathcal{D}_L|}{|\mathcal{D}|}\mathrm{cost}(\mathcal{D}_L) + \frac{|\mathcal{D}_R|}{|\mathcal{D}|}\mathrm{cost}(\mathcal{D}_R)\right) \tag{12}$$

To define the classification cost we resort to the so-called *impurity functions*. Let f be some impurity function and define the impurity of a node A as

$$I(A) = \sum_{i=1}^{C} f\left(p_{iA}\right) \tag{13}$$

where p_{iA} is the proportion of those in A that belong to class i (in our case we have $C = 2$).

Two candidates for f are the information entropy $f(p) = -p\log(p)$ and the Gini index $f(p) = p(1-p)$. The algorithm we built in this paper split the tree based on information entropy. Since we wish to define the *variable importance*, the reduction in the information entropy attributed to each variable at each split is tabulated, and the sum is returned.

The partitioning we employ is binary (it splits the data into two regions) and recursive (each splitting rule depends on the splits above it). However, we must decide how long we keep growing the tree. Though we could proceed till exploiting all the features, a very large tree could result in overfitting (of course, we may also face the opposite problem, since a small tree might not capture the important structure hidden in the data) (Friedman et al. 2001).

The adopted strategy here is to grow a large tree T_0 and then prune it back to find an optimal subtree.

Let us define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0 and denote, by $|T|$, the number of terminal nodes in T. Each node A_m represents a region with N_m observations; for example, the node A considered in (13) is just one of them.

Hence, we define a cost complexity criterion as follows (Friedman et al. 2001)

$$L_{c_p}(T) = \sum_{m=1}^{|T|} N_m I(A_m) + c_p|T| \tag{14}$$

where $I(A_m) = I(A_m, T)$ is defined by (13) and c_p is the complexity parameter. The idea is to find the optimal subtree to minimize $L_{c_p}(T)$. Large (small) values of c_p result in smaller (bigger) subtrees. Intuitively, with $c_p = 0$ the solution is the full tree T_0.

5 Experiments and results

We aim to identify the drivers of self-reposting. For that purpose, we have introduced a decision-tree model in Sect. 4.3. In this section, we report the results of applying that model. In addition, we examine the importance of each feature in the model. These results allow us to identify the most significant factors associated with self-reposting.

5.1 Testing and validation

As stated earlier, the subset of interest to us is S_{TSR}, which contains both the original tweets and the self-reposted ones. The overall size of S_{TSR} is 11,723. As reported in Sect. 3.3, the dataset is heavily imbalanced since the number of original tweets is much larger than the self-reposted ones. For the purpose of testing, we split that set into two parts, employed respectively for training the machine learning model and testing it. In order to avoid overfitting, the composition of both sets was not fixed, but tenfold cross-validation was employed Schaffer (1993).

Table 2 Performance of the decision tree model with full features

Accuracy	0.8972112
Sensitivity	0.9043062
Specificity	0.8346805
Balanced Accuracy	0.8694933

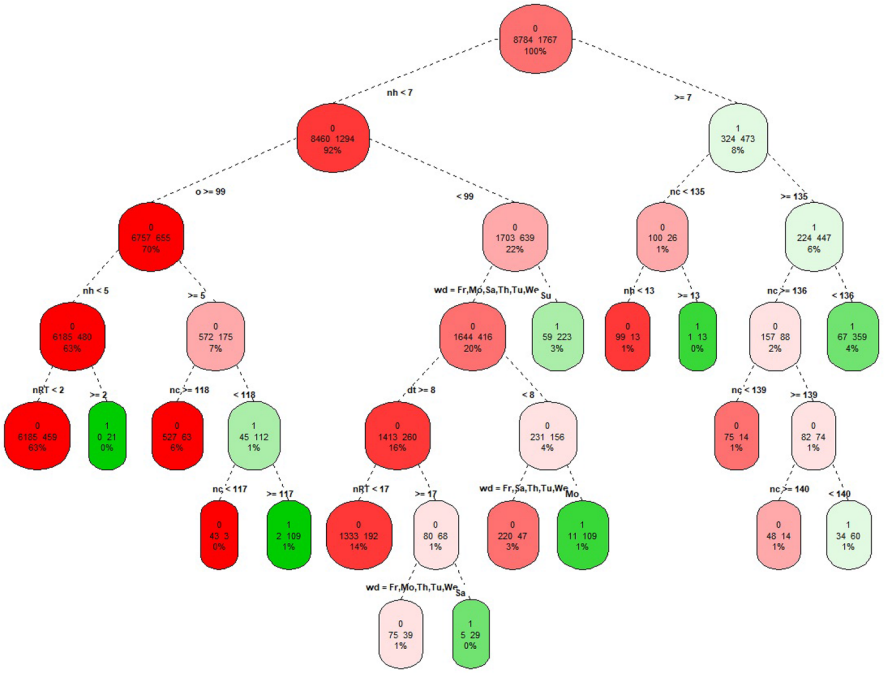

Fig. 13 Decision Tree for the full set of features (Gini index)

5.2 Full-feature decision tree classification

We first build a decision tree as described in Sect. 4.3 with the full set of features. The resulting performance metrics are reported in Table 2. As can be seen, the accuracy is close to 90%, which we can consider as a good value. We also see that the performance for the two classes, embodied by the sensitivity and specificity, are not far apart. In particular, the high sensitivity value, now a bit over 90%, shows that the full features model can capture the overwhelming majority of self reposts (here labelled as positive). Several non-self-reposted tweets are instead disguised as positive cases, as shown by the lower value of the specificity. However, though being dragged towards this lower value, the balanced accuracy is still above 85%.

We have built the tree using both impurity measures (Gini index and entropy). The resulting trees are shown respectively in Figs. 13 and 14. We can consider the variables employed in the levels closest to the tree root as the most significant ones. The first level split is carried out by employing the number $nh(k)$ of hashtags. On the second level, we use the rank $o(k)$ on the left subtree and the length $nc(k)$ of the tweet in characters on the right subtree. This partly agrees with the clue provided by the exploratory analysis in Sect. 4.1.

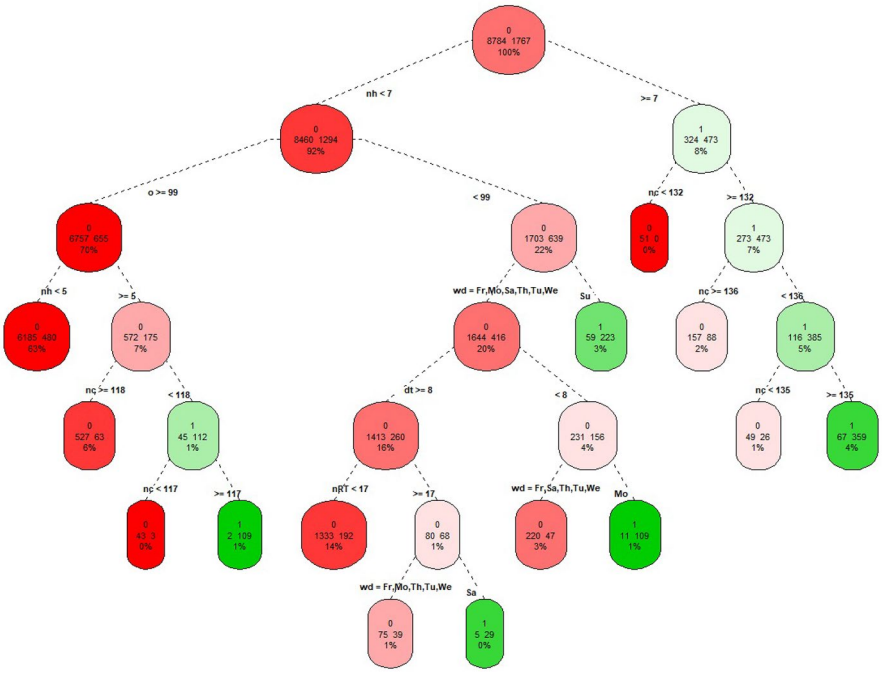

Fig. 14 Decision Tree for the full set of features (Information Entropy)

5.3 Feature selection

Though we have achieved good accuracy with the tree model built by considering the full set of features, we wish to extract those features that really matter in characterizing the self-reposting phenomenon.

We have already performed an exploratory analysis in Sect. 4.1, where the day of the week, the time of the day, the number of embedded hashtags, and the rank of the last inter-retweet interval appeared as the most discriminating features. The decision trees built for the full set of features identified the number of hashtags, the rank, and the length of tweets as the most relevant variables to split the dataset.

We now take a step forward by considering some metrics that allow us to support (or disprove) those findings under a more quantitative approach. In Duboue (2020) the following set of metrics is suggested, with our choices between parentheses:

- Chi-square test (disregarded);
- Pearson correlation coefficient (replaced by Cramer's V);
- Mutual information.

The chi-square test would be employed by setting the null hypothesis as that there is no difference between the distribution of the feature when the tweet is self-reposted or not. The features to be retained as significant are those for which the null hypothesis is rejected. However, it is well known that such significance tests are very sensitive to the sample size since even trivial differences may be considered as significant Bergh

Table 3 Feature ranking according to Cramer's V

Feature	V
Length of the tweet	0.51
Day of the week	0.32
Number of hashtags	0.24
Retweet count	0.13
Rank of the last inter-retweet interval	0.10
Time of the day	0.09
Average inter-retweet time	0.08
Number of hyperlinks	0.04
Number of exclamation points	0.02

(2015). In that case, the null hypothesis is rejected for most variables, concluding that all variables are significant. This is precisely what happened in our case: the chi-square test was not helpful, judging all the variables to be significant.

As to the second suggestion in the bulleted list above, the Pearson correlation coefficient is not very meaningful since most of our variables are categorical. We have opted then for Cramer's V association index, which is built for categorical variables Acock and Stavig (1979). We have discretized the intrinsically continuous variables by binning them as typically done in the computation of the chi-square statistics. In Table 3, we show the results, where the three features most strongly associated with the target variable are the length of the tweet, the day of the week, and the number of hashtags. These results confirm the suggestions provided by our exploratory analysis and the decision tree.

Finally, we have computed the mutual information between each feature and the target variable. The results are shown in Fig. 15, where now the day of the week, the length of the tweet, and the rank appear as the three top features.

Summing up the outcomes of the exploratory analysis and the three quantitative feature selection metrics, we can conclude that the most significant features to assess the propensity of a tweet to be self-reposted are

- The day of the week;
- Length of the tweet;
- Number of hashtags;
- Rank of the intertweet time interval;
- The time of the day.

The day of the week has received the broadest support, i.e., by the exploratory analysis, Cramer's V, and the mutual information. The tweet's length is supported by both Cramer's V and the mutual information. The exploratory analysis supports the number of hashtags and Cramer's V. The rank is supported again by the exploratory analysis and the mutual information. Finally, the time of the day is supported by the exploratory analysis only.

We can now build a new decision tree using just the features that have been deemed significant. We show them in Figs. 16 and 17 when the Gini index and the information entropy are employed respectively.

The performance achieved with the reduced set of features is shown in Table 4. If we compare these results with those obtained with the full set of features (shown in Table 2), we notice the performance is practically unaltered (actually, there is even a tiny improvement). The removed features add nothing to the prediction performance of the decision tree. We can

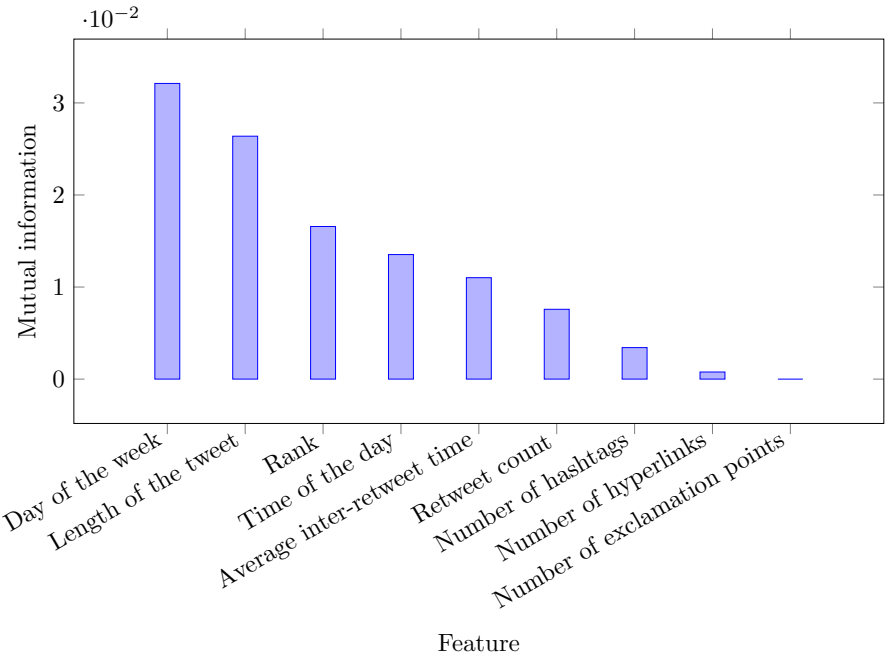

Fig. 15 Mutual Information between features and target variable

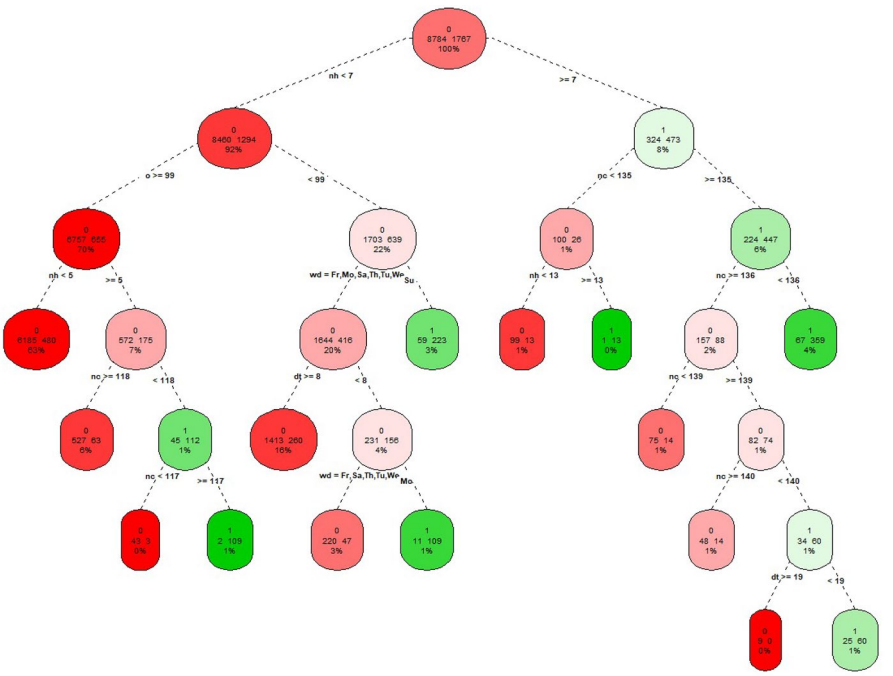

Fig. 16 Decision Tree for the reduced set of features (Gini index)

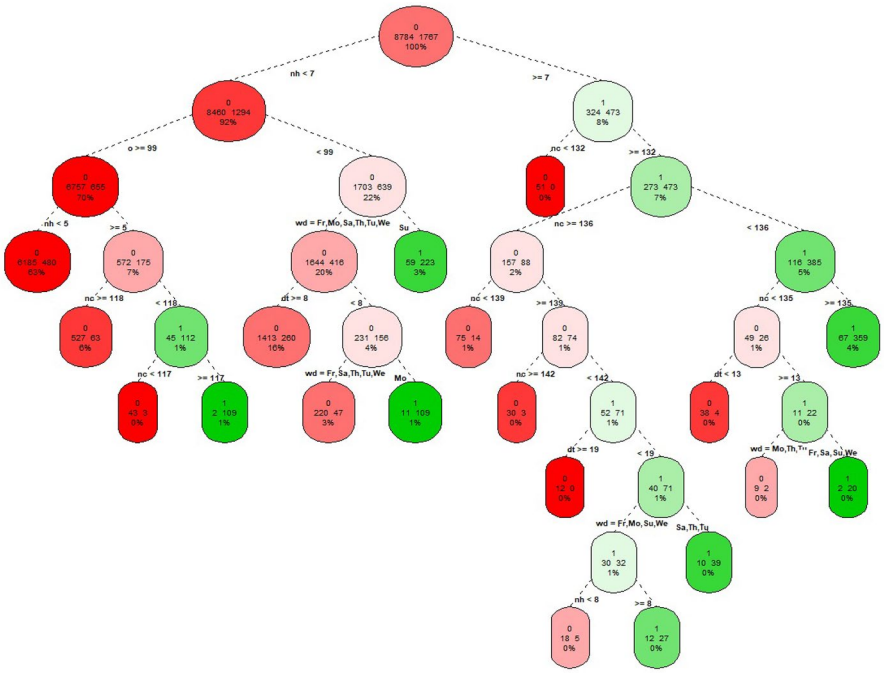

Fig. 17 Decision Tree for the reduced set of features (Information Entropy)

Table 4 Classification performance of the decision tree with the reduced set of features

Accuracy	0.8984053
Sensitivity	0.9052624
Specificity	0.836704
Balanced accuracy	0.8709832

then safely conclude that the features impacting the self-repost behaviour are the day of the week, the tweet's length, the number of hashtags, the rank of the intertweet time interval, and the time of the day. In particular, as can be seen from both decision trees in Figs. 16 and 17, self-reposts take place more often in the weekend and on times other than the central part of the day, for longer tweets containing a higher number of hashtags, and when a long time has elapsed since the last retweet.

Looking at Fig. 16 we can spot that the characterization of all the tweets reposted multiple times (for the reduced set of features) is the following

$$\begin{aligned}
&\Big\{\mathrm{nh} \geq 7\,,\ \mathrm{nc} \in [135, 136)\Big\} \ \vee\ \Big\{\mathrm{nh} \geq 7\,,\ \mathrm{nc} \in [139, 140)\,,\ \mathrm{dt} < 19\Big\} \vee \\
&\Big\{\mathrm{nh} < 7\,,\ \mathrm{o} < 99\,,\ \mathrm{dt} < 8\,,\ \mathrm{wd} \in \big\{\mathrm{Monday}\big\}\Big\} \vee \\
&\Big\{\mathrm{nh} < 7\,,\ \mathrm{o} < 99\,,\ \mathrm{wd} \in \big\{\mathrm{Sunday}\big\}\Big\} \vee \\
&\Big\{\mathrm{nh} \in [5, 7)\,,\ \mathrm{o} \geq 99\,,\ \mathrm{nc} \in [117, 118)\Big\}
\end{aligned} \tag{15}$$

6 Discussion and conclusions

In this paper, we have described an analysis framework to understand the phenomenon of self-reposting in Twitter discussions. For that purpose, we have examined a large set of tweets concerning wind energy. Though those tweets concern a specific topic, the framework and most conclusions can be considered generally valid.

The combined use of exploratory analysis and a machine learning approach (through decision trees) has shown that self-reposts typically take place on weekends and in the afternoon hours, have a length close to the limit, contain more hashtags, and occur when retweets from other twitterers become less frequent. Their typical placement over time suggests that they fall into a planning habit. The combination of day and time reveals that self-reposting is not a major activity to be conducted in the core times. Instead, it is part of a regular survey of social media activities. People review the status of their tweets every week and take some time off their non-working hours to examine their posts and resubmit some. A major criterion for self-reposting that emerges from our analysis is suggested by the prominent feature of the rank of the inter-tweet interval. It looks like twitterers decide to repost their own messages when retweets start to get rarer. We cannot go as far as identifying a threshold that acts as a trigger; we can imagine that, though implicitly, this can be related to the popularity of the tweet to be self-reposted. However, the intention to revive the interest in their messages is clear. It is not worthwhile to repost when the interest is still high, as embodied by a high number of retweets. As to the characteristics of the tweets that are self-reposted, they appear to be richer (more content and more hashtags) than non-self-reposted ones. It is not easy to say whether this is a driving feature for self-reposting or a natural characteristic of tweets posted by people who strongly wish to put their message across.

Regarding the specific topic of wind energy, we see that wind energy companies do not apply self-reposting. Indeed, among the users who have applied self-reposting massively, we noticed that the majority is formed by suspended accounts or accounts which seem to no longer be on Twitter and by generalist individual twitterers. It is to be seen whether the massive use of reposting by individuals, especially by temporary or banned users, is a feature of the specific topic or is a general result.

Aside from the characterization of the self-reposting phenomenon, the implications of this study concern the management of social platforms like Twitter. In fact, self-reposting may take an annoying or vexing flavour (or even overlap with a phenomenon like mail bombing), which may compel platform managers to take actions against massive self-reposters. Possible actions by Twitter could include limitations to twitterers reposting their own tweets. This could be done based on one or more indicators, which express related policies. For example, self-reposts could be blocked when they exceed a threshold set for a time interval (e.g., no more than three self-reposts per month). The rationale, in this case, would be to limit self-reposting straightforwardly, considering it a vexing activity anyway. A second possible measure would be to limit self-reposts when the original tweet has not received enough retweets (which triggers self-reposting in many cases). The rationale would be to avoid reposting tweets that have not caught the interest of twitterers. Finally, a similar measure by Twitter could block self-reposts when the tweet has not been retweeted for a long time. This third measure would avoid reposting tweets that do not longer catch the interest of twitterers though they did in the past.

Our analysis has shown that self-reposting is not a random behaviour but is triggered by some events (which we have identified as the timing of past retweets) and is linked to

some features of the tweet itself (e.g., its length and hashtag content). At the same time, it happens more frequently on weekends and in non-central times. Two branches appear to be most interesting after this initial study: (a) looking for additional features of self-reposted tweets; (b) extending the analysis to other social media platforms to examine if self-reposting is employed and exhibits characteristics different from Twitter.

Funding Open access funding provided by Università degli Studi Roma Tre within the CRUI-CARE Agreement. The authors have not disclosed any funding.

Declarations

Conflict of interest The authors declared that there is no conflict of interest.

References

Achananuparp, P., Lim, E.P., Jiang, J., et al.: Who is retweeting the tweeters? modeling, originating, and promoting behaviors in the twitter network. ACM Trans. Manage. Inf. Syst. **3**(3), (2012)

Acock, A.C., Stavig, G.R.: A measure of association for nonparametric statistics. Soc. Forces **57**(4), 1381–1386 (1979)

Agarwal, S., Foo Sing, T., Yang, Y.: The impact of transboundary haze pollution on household utilities consumption. Energy Econ. **85**(104), 591 (2020)

Aleixandre-Tudó, J.L., Castelló-Cogollos, L., Aleixandre, J.L., et al.: Renewable energies: worldwide trends in research, funding and international collaboration. Renew. Energy **139**, 268–278 (2019)

Ali Sayigh, D.M.: The age of wind energy: progress and future directions from a global perspective. Innovative Renewable Energy, 1st edn. Springer International Publishing, Berlin (2020)

Arora, U., Dutta, H.S., Joshi, B., et al.: Analyzing and detecting collusive users involved in blackmarket retweeting activities. ACM Trans. Intell. Syst. Technol. **11**(3), (2020)

Austmann, L.M., Vigne, S.A.: Does environmental awareness fuel the electric vehicle market? a twitter keyword analysis. Energy Econ. **101**(105), 337 (2021)

Bergh, D.: Chi-squared test of fit and sample size–a comparison between a random sample approach and a chi-square value adjustment method. J. Appl. Meas. **16**, 2 (2015)

Borch, K., Munk, A.K., Dahlgaard, V.: Mapping wind-power controversies on social media: Facebook as a powerful mobilizer of local resistance. Energy Policy **138**(111), 223 (2020)

Boutakidis, D., Aggelopoulos, S., Pavloudi, A., et al.: Attitudes and opinions of social media users on renewable energy. J. Environ. Prot. Ecol. **15**(4), 1727–1734 (2014)

Boyd, D., Golder, S., Lotan, G.: (2010) Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: 2010 43rd Hawaii International Conference on System Sciences, pp. 1–10

Breiman, L., Friedman, J., Olshen, R.A., et al.: Classification and regression trees. The Wadsworth statistics/probability series, CRC, Boca Raton (1984)

Chen, J., Liu, Y., Zou, M.: User emotion for modeling retweeting behaviors. Neural Netw. **96**, 11–21 (2017)

Chen, L., Deng, H.: Predicting user retweeting behavior in social networks with a novel ensemble learning approach. IEEE Access **8**, 148250–148263 (2020)

Cotter, K.: Playing the visibility game: how digital influencers and algorithms negotiate influence on instagram. New Med. Soc. **21**(4), 895–913 (2019)

Cruz, R.A.B., Lee, H.J.: The brand personality effect: communicating brand personality on twitter and its influence on online community engagement. J. Intell. Inf. Syst. **20**(1), 67–101 (2014)

Dai, K., Bergot, A., Liang, C., et al.: Environmental issues associated with wind energy-a review. Renew. Energy **75**, 911–921 (2015)
De Jesus, A., Antunes, P., Santos, R., et al.: Eco-innovation in the transition to a circular economy: An analytical literature review. J. Clean. Prod. **172**, 2999–3018 (2018)
Dhakouani, A., Znouda, E., Bouden, C.: Impacts of energy efficiency policies on the integration of renewable energy. Energy Policy **133**(110), 922 (2019)
Duboue, P.: The art of feature engineering: essentials for machine learning. Cambridge University Press, Cambridge (2020)
Frantál, B.: Have local government and public expectations of wind energy project benefits been met? implications for repowering schemes. J. Environ. Policy Plan. **17**(2), 217–236 (2015)
Friedman, J., Hastie, T., Tibshirani, R., et al.: The elements of statistical learning, vol 1. Springer series in statistics New York (2001)
Fronzetti Colladon, A., Gloor, P.A.: Measuring the impact of spammers on e-mail and twitter networks. Int. J. Inf. Manage. **48**, 254–262 (2019)
Gao, J., Shen, H., Liu, S., et al.: Modeling and predicting retweeting dynamics via a mixture process. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '16 Companion, pp. 33–34 (2016)
Gao, S., Ma, J., Chen, Z.: Modeling and predicting retweeting dynamics on microblogging platforms. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, pp. 107–116 (2015)
Gentry, J.: twitteR: R Based Twitter Client. https://CRAN.R-project.org/package=twitteR, r package version 1.1.9 (2015)
Gillespie, T.: Algorithmically recognizable: Santorum's google problem, and google's santorum problem. Inf. Commun. Soc. **20**(1), 63–80 (2017)
Gruber, H.: Quoting and retweeting as communicative practices in computer mediated discourse. Discourse Context Med. **20**, 1–9 (2017)
Hoang, T.A., Lim, EP.: Retweeting: an act of viral users, susceptible users, or viral topics? pp. 569–577 (2013)
Hoffert, M., Caldeira, K., Benford, G., et al.: Engineering: advanced technology paths to global climate stability: energy for a greenhouse planet. Science **298**(5595), 981–987 (2002)
Horbaty, R., Huber, S., Ellis, G.: Large-scale wind deployment, social acceptance. WIREs Energy Environ. **1**(2), 194–205 (2012)
Hu, J., Luo, Y., Yu, J.: An empirical study on selectiviey of retweeting behaviors under multiple exposures in social networks. J. Comput. Sci. **28**, 228–235 (2018)
Jain, A., Jain, V.: Renewable energy sources for clean environment: opinion mining. Asian J. Water Environ. Pollut. **16**(2), 9–14 (2019)
Jethani, J.: Wind power policy in India. World **6000**, 5358 (2016)
Kim, E., Sung, Y., Kang, H.: Brand followers' retweeting behavior on twitter: How brand relationships influence brand electronic word-of-mouth. Comput. Hum. Behav. **37**, 18–25 (2014)
Kim, E., Hou, J., Han, J.Y., et al.: Predicting retweeting behavior on breast cancer social networks: Network and content characteristics. J. Health Commun. **21**(4), 479–486 (2016)
Kuhn, M.: Building predictive models in r using the caret package. J. Stat. Softw. Artic. **28**(5), 1–26 (2008)
Kuhn, M.: caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret, r package version 6.0-86 (2020)
Lamy, J., de Bruin, W.B., Azevedo, I.M., et al.: Keep wind projects close? a case study of distance, culture, and cost in offshore and onshore wind energy siting. Energy Res. Soc. Sci. **63**(101), 377 (2020)
Leung, D.Y., Yang, Y.: Wind energy development and its environmental impact: A review. Renew. Sustain. Energy Rev. **16**(1), 1031–1039 (2012)
Li, Q., Liu, Y.: Exploring the diversity of retweeting behavior patterns in chinese microblogging platform. Inf. Process. Manage. **53**(4), 945–962 (2017)
Li, Q., Qu, H., Chen, L., et al.: Visual analysis of retweeting propagation network in a microblogging platform. In: Proceedings of the 6th International Symposium on Visual Information Communication and Interaction. Association for Computing Machinery, VINCI '13, pp. 44–53 (2013)
Li, R., Crowe, J., Leifer, D., et al.: Beyond big data: social media challenges and opportunities for understanding social perception of energy. Energy Res. Soc. Sci. **56**(101), 217 (2019)
Shin Lim, Y., Lee-Won, R.J.: When retweets persuade: the persuasive effects of dialogic retweeting and the role of social presence in organizations' twitter-based communication. Telemat. Inf. **34**(5), 422–433 (2017)

Lin, X., Lachlan, K.A., Spence, P.R.: Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on twitter and weibo. Comput. Hum. Behav. **65**, 576–581 (2016)
Liu, B., Ni, Z., Luo, J., et al.: Analysis of and defense against crowd-retweeting based spam in social networks. World Wide Web **22**(6), 2953–2975 (2019)
Luedecke, G., Boykoff, MT.: Environment and the media. Int. Encycl. Geogr. People Earth Environ. Technol. pp. 1–8 (2017)
Muñoz, C.Q.G., Márquez, F.P.G.: Wind energy power prospective. In: Renewable energies. Springer, p 83–95 (2018)
Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press, Cambridge (2012)
Nesi, P., Pantaleo, G., Paoli, I., et al.: Assessing the retweet proneness of tweets: predictive models for retweeting. Multimed. Tools Appl. **77**(20), 26371–26396 (2018)
O'Leary, D.E.: Modeling retweeting behavior as a game: comparison to empirical results. Int. J. Hum. Comput. Stud. **88**, 1–12 (2016)
O'Meara, V.: Weapons of the chic: Instagram influencer engagement pods as practices of resistance to instagram platform labor. Social Media+ Society 5(4):2056305119879671 (2019)
O'Reilly, T., Milstein, S.: The twitter book. O'Reilly Media, Inc. (2011)
Pang, N., Law, P.W.: Retweeting #worldenvironmentday: A study of content features and visual rhetoric in an environmental movement. Comput. Hum. Behav. **69**, 54–61 (2017)
Park, C.S., Kaye, B.K.: Expanding visibility on twitter: author and message characteristics and retweeting. Social Media+ Society 5(2):2056305119834595 (2019)
Reboredo, J.C., Ugolini, A.: The impact of twitter sentiment on renewable energy stocks. Energy Econ. **76**, 153–169 (2018)
Resce, G., Maynard, D.: What matters most to people around the world? retrieving better life index priorities on twitter. Technol. Forecast. Soc. Chang. **137**, 61–75 (2018)
Saidur, R., Rahim, N., Islam, M., et al.: Environmental impact of wind energy. Renew. Sustain. Energy Rev. **15**(5), 2423–2430 (2011)
Schaffer, C.: Selecting a classification method by cross-validation. Mach. Learn. **13**(1), 135–143 (1993)
Sherren, K., Parkins, J.R., Smit, M., et al.: Digital archives, big data and image-based culturomics for social impact assessment: Opportunities and challenges. Environ. Impact Assess. Rev. **67**, 23–30 (2017)
Shi, J., Lai, K.K., Hu, P., et al.: Understanding and predicting individual retweeting behavior: Receiver perspectives. Appl. Soft Comput. **60**, 844–857 (2017)
Sivarajah, U., Fragidis, G., Lombardi, M., et al.: The use of social media for improving energy consumption awareness and efficiency: An overview of existing tools. In: European, Mediterranean and Middle Eastern Conference on Information Systems (EMCIS) (2015)
Stephens, J.C., Rand, G.M., Melnick, L.L.: Wind energy in us media: a comparative state-level analysis of a critical climate change mitigation technology. Environ. Commun. **3**(2), 168–190 (2009)
Suh, B., Hong, L., Pirolli, P., et al.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE second international conference on social computing, IEEE, pp. 177–184 (2010)
Sun, X., Huang, D., Wu, G.: The current state of offshore wind energy technology development. Energy **41**(1), 298–312 (2012)
Therneau, T., Atkinson, B.: rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart, r package version 4.1-15 (2019)
Therneau, T.M., Atkinson, E.J., et al.: An introduction to recursive partitioning using the rpart routines. Tech. rep, Technical report Mayo Foundation (1997)
Valenzuela, S.: Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. Am. Behav. Sci. **57**(7), 920–942 (2013)
van der Loos, H.A., Negro, S.O., Hekkert, M.P.: Low-carbon lock-in? exploring transformative innovation policy and offshore wind energy pathways in the netherlands. Energy Res. Soc. Sci. **69**(101), 640 (2020)
Vuichard, P., Stauch, A., Dällenbach, N.: Individual or collective? community investment, local taxes, and the social acceptance of wind energy in Switzerland. Energy Res. Soc. Sci. **58**(101), 275 (2019)
Wang, A., Chen, T., Kan, M.Y.: Re-tweeting from a linguistic perspective. In: Proceedings of the Second Workshop on Language in Social Media, pp. 46–55 (2012)
Wang, H., Li, Y., Feng, Z., et al.: Retweeting analysis and prediction in microblogs: An epidemic inspired approach. China Commun. **10**(3), 13–24 (2013)
Wang, X., Lee, E.W.: Negative emotions shape the diffusion of cancer tweets: toward an integrated social network–text analytics approach. Internet Research (2020)

Wang, Y., Li, H., Zuo, J., et al.: Evolution of online public opinions on social impact induced by nimby facility. Environ. Impact Assess. Rev. **78**(106), 290 (2019)
Wang, Y., Zheng, L., Zuo, J.: Online rumor propagation of social media on nimby conflict: Temporal patterns, frameworks and rumor-mongers. Environ. Impact Assess. Rev. **91**(106), 647 (2021)
Wang, Z., Ke, L., Cui, X., et al.: Monitoring environmental quality by sniffing social media. Sustainability **9**(2), 85 (2017)
Webberley, W., Allen, S., Whitaker, R.: Retweeting: A study of message-forwarding in twitter. In: 2011 Workshop on Mobile and Online Social Networks, pp. 13–18 (2011)
Weinzettel, J., Reenaas, M., Solli, C., et al.: Life cycle assessment of a floating offshore wind turbine. Renew. Energy **34**(3), 742–747 (2009)
Xiong, F., Liu, Y., Zhang, Z., et al.: An information diffusion model based on retweeting mechanism for online social media. Phys. Lett. A **376**(30), 2103–2108 (2012)
Yang, Z., Guo, J., Cai, K., et al.: Understanding retweeting behaviors in social networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '10, pp. 1633–1636 (2010)
Zhang, K., Yun, X., Liang, J., et al.: Retweeting behavior prediction using probabilistic matrix factorization. In: 2016 IEEE Symposium on Computers and Communication (ISCC), pp. 1185–1192 (2016)
Zhao, X., Li, S., Zhang, S., et al.: The effectiveness of china's wind power policy: an empirical analysis. Energy Policy **95**, 269–279 (2016)
Zobeidi, T.: Impact of social media on perceptions and use of renewable energy sources. Young scientists summer program, International Institute for Applied systems Analysis (2020)