Time-Based Ensembles for Prediction of Rare Events In News Streams

***Abstract*—Thousands of news are published everyday reporting worldwide events. Most of these news obtain a low level of popularity and only a small set of events become highly popular in social media platforms. Predicting rare cases of highly popular news is not a trivial task due to shortcomings of standard learning approaches and evaluation metrics. So far, the standard task of predicting the popularity of news items has been tackled by either of two distinct strategies related to the publication time of news. The first strategy, *a priori*, is focused on predicting the popularity of news upon their publication when related social feedback is unavailable. The second strategy, *a posteriori*, is focused on predicting the popularity of news using related social feedback. However, both strategies present shortcomings related to data availability and time of prediction. To overcome such shortcomings, we propose a hybrid strategy of time-based ensembles using models from both strategies. Using news data from Google News and popularity data from Twitter, we show that the proposed ensembles significantly improve the early and accurate prediction of rare cases of highly popular news.**

Keywords*—*Social Media, Regression, Rare Cases, Time-based Ensemble

I. Introduction

Thousands of news stories are read and shared through social media platforms every day. Most of these news remain relatively unpopular [1], and only a few may be considered as highly popular news. In news streams, once an item obtains a certain level of social feedback according to a given social media platform, its popularity becomes obvious. However, predicting that a news item will be very popular as soon as it is published or in the first moments after (when related social feedback is unavailable), is very challenging. Solving this task is crucial for entities that aim at identifying the most significant events in a fast and accurate manner.

The standard task of predicting news popularity has been addressed in previous work with two distinct strategies differentiating as to the time of prediction: i) *a priori*, where prediction occurs before or at the moment of the contents' publication when no social feedback is available and ii) *a posteriori*, where the popularity is predicted using data generated by users after the contents' publication.

We observe several shortcomings raised by the use of either strategies. Although *a priori* models are useful at predicting news popularity when there is no social feedback, they do not correct or update these predictions once user-generated data become available. As for *a posteriori* models, relying only on the limited data available shortly after publication might not be enough for an accurate prediction of highly popular news.

In this paper we aim at bridging the gap between these two approaches by introducing a hybrid strategy for the prediction of news items popularity, focusing on the rare cases of highly popular news. We propose time-based ensembles of both strategies (*a priori* and *a posteriori*) in order to tackle the aforementioned issues. We show, through experimental evaluation, that by applying the proposed hybrid strategy it is possible to significantly increase the accuracy of tasks attempting the early prediction of highly popular news when compared to *a priori* or *a posteriori* strategies.

The main contributions of this paper are the following:

- A new hybrid strategy for quick and accurate prediction of highly popular events in news feeds;
- An extensive experimental evaluation to verify that our proposed strategy significantly improves the results of either existing strategies.

The remainder of the paper is organized as follows: related work is described in Section II. The problem definition is presented in Section III and prediction models of both strategies (*a priori* and *a posteriori*) described in Section IV. Our time-based ensemble proposal is presented in Section V. Materials and methods are described in Section VI and the experimental evaluation is presented in Section VII. Finally, conclusions are presented in Section VIII.

II. Related Work

The study of content popularity using social networks' data has proliferated in recent years. Figueiredo et al. [2] studied the popularity of Youtube videos as to its evolution in time and extraction of trends. Lerman and Ghosh [3] studied how the structure of a network affects the dynamics of information spread by analyzing data from Digg and Twitter. Tatar et al. [1] provides a survey on prediction of web content popularity. Concerning news popularity and the Twitter platform, Petrovic et al. [4] conclude that Twitter covers almost all news-wire events but the opposite is not true. Osborne and Dredze [5] claim that Twitter is the preferred medium for breaking news, almost consistently leading Facebook or Google Plus.

Concerning the task of news popularity prediction, as previously mentioned, proposals are mainly focused in two distinct scenarios: *i) a priori* and *ii) a posteriori* prediction. The majority of related work attempts to tackle this task with approaches that are based on the average behaviour of news popularity (*i.e.* mostly news with low popularity), which is not our objective. Our goal is to quickly and accurately predict rare cases of highly popular news. In this section we focus on work that has been done in this context.

In the *a priori* scenario, Tsagkias et al. [6] approach this task using diverse sets of features for a two-step procedure: first, to predict if a news will receive comments and secondly the volume (low or high) of comments. Results show a solid

performance on the first step but a degraded performance in the second. Bandari et al. [7] propose classification and regression approaches using four predictors: source, category, subjectivity in the language and named entities mentioned. To the best of our knowledge, the work of Moniz et al. [8] is the only that proposes approaches focusing on the prediction of highly popular news. In that work, the authors propose using resampling strategies with standard regression models using a bag-of-words and sentiment scores of title and headline as predictors.

Regarding the *a posteriori* scenario, the challenges raised by this task of predicting rare cases of highly popular news are depicted in the work of Kim et al. [9]. Nonetheless, important work has been presented focusing on the general prediction of the number of tweets a news item will obtain within a given time-window. According to Tatar et al. [1], this scenario has been explored by three main approaches: *i)* cumulative growth, by studying the amount of attention items receive from being published until the prediction moment (*e.g.* [10], [11]); *ii)* temporal analysis, studying the evolution of content popularity over time until the prediction moment (*e.g.* [12], [13]); and *iii)* popularity evolution trends, using clustering methods to find similar items (*e.g.* [14], [15]).

Focusing on the most popular approach, cumulative growth, early work by Kaltenbrunner et al. [16] proposes a constant growth approach for predicting the popularity of Slashdot stories, depending on the publication hour of the news stories. Szabo et al. [17] propose two methods for predicting the popularity of YouTube videos and Digg stories: the *linear log* and *constant scaling* methods. Tsagkias et al. [18] and Tatar et al. [19] used the former methods in news popularity prediction tasks, showing their reliability.

To overcome the previously described shortcomings raised by proposals from both these strategies, we propose a hybrid strategy by combining models of both existing strategies in time-based ensembles, to improve predictive accuracy concerning rare cases of highly popular news.

III. Problem Definition

In this work we address the problem of early and accurately predicting rare cases of highly popular news. The main task consists of predicting the number of tweets news will receive[1], focusing on achieving high prediction accuracy for the highly popular news stories.

This task can be modelled as a non-standard (due to its focus on the rare cases of extreme high values) regression problem, where the target variable is the number of tweets each incoming news item is expected to have. Our assumption is that it is possible to map the number of tweets concerning news items at a given time into the total number of tweets they will receive after a predefined period. In this work we set this period to two days, based on the work of Yang and Leskovec [20][2].

[1]Note that this number includes re-tweets of the story.

[2]Although the work of the authors suggests that the popularity of online content may change until four days past its publication, our data suggests that two days is sufficient to determine the degree of popularity attributed to a given news item.

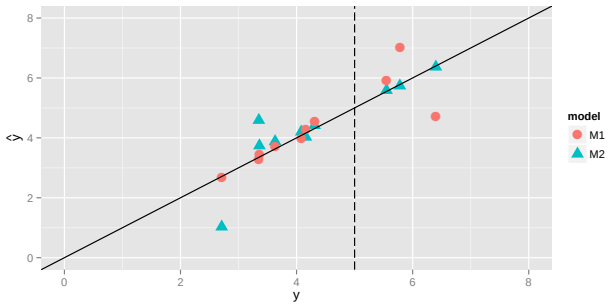

Fig. 1. Misleading Scenario for Standard Error Metrics with Artificial Data.

The unknown function we want to approximate is defined as $\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i)$, where $\hat{y}$ is the final number of tweets and $\mathbf{x}$ is a vector of predictors. Vector $\mathbf{x}$ is different when predicting with *a priori* or *a posteriori* models. The former uses news metadata such as sentiment scores or bags of words, the latter is solely based on social feedback, (*e.g.* the number of tweets a news accumulated until the moment of prediction).

A. Handling Imbalanced Distributions

While many existing algorithms solve standard regression tasks, they are designed to find the model which optimises standard error metrics (e.g. mean squared error). As mentioned, this may be problematic when the objective is predictive accuracy on an interval of the target variable which is underrepresented in the data. We exemplify these potential problems in the scenario depicted in Figure 1 using synthetically generated data. In this scenario, two models (M_1 and M_2) provide their respective sets of artificial predictions.

Observing Figure 1 we find that model M_1 clearly shows a superior predictive accuracy at low values of the data and that model M_2 is far more accurate at the highest values. However, if we calculate popular metrics such as Mean Squared Error and Mean Absolute Deviation (*MSE* and *MAD*, respectively) we will find no difference between these two models. Both models obtain a score of 0.461 for *MSE* and a score for *MAD* of 0.397. In order to overcome this issue of evaluating models in regression tasks focused on rare values, we resort to the utility-based framework proposed by Ribeiro [21] and the evaluation metrics described in Section VI-B.

This framework distinguishes rare and normal cases based on the distribution of the target variable and the concept of relevance as proposed by Ribeiro [21]. In this paper, the concept of relevance relates to the rareness of news items popularity. Since the distribution of news items' popularity resembles a power-law distribution [1], the higher the popularity, the more relevant the case. When expert knowledge is not available, as in our case, Ribeiro proposes an approach to automatically generate relevance functions based on box plot statistics, mapping values of the target variable into a scale of relevance $[0, 1]$. By combining this mapping and a relevance threshold provided by the user, it is possible to determine the cases in our data considered most relevant, the ones we want to apply some bias in the prediction models.

In Figure 2 an example with a sample of data from the topic *economy* (described in Section VI-A) illustrates this process

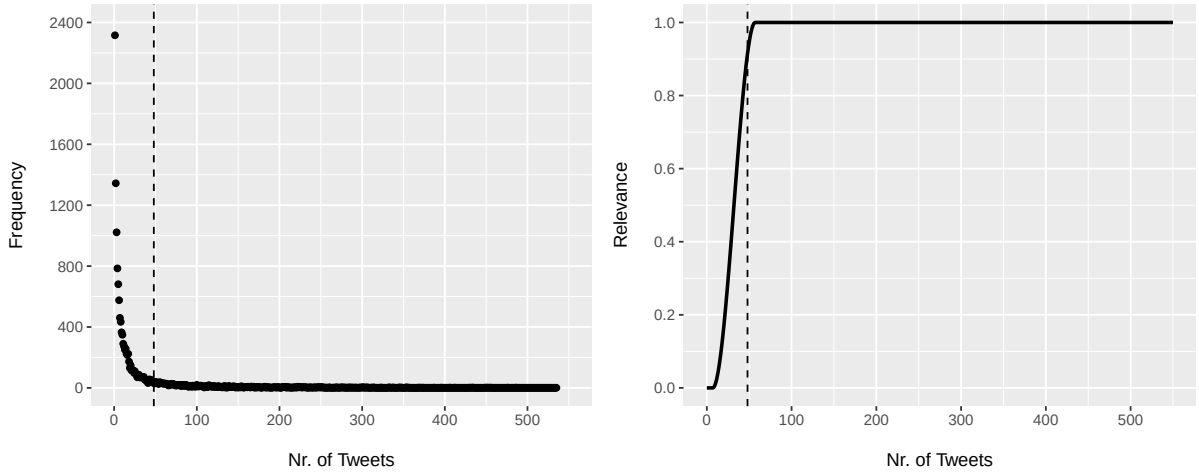

Fig. 2. Frequency of news items per number of tweets (left) and the automatic generation of a relevance function by the approach of Ribeiro [21] for a sample of news from the topic *economy* with a relevance threshold of 0.9.

using a relevance threshold of 0.9. In addition to confirming that the distribution of the popularity of news items resembles a power-law distribution [1], we observe that according to the data in this sample, the generated relevance function and the 0.9 relevance threshold, news are considered to be most relevant when they obtain 48 or more tweets, representing a small proportion of the data (15%).

IV. Prediction Models

In this section we describe models used in this paper for the prediction of news popularity for both the *a priori* and *a posteriori* strategies. These models are the basis of time-based ensembles and the hybrid strategy proposed in this paper.

A. A Priori Models

As previously stated, *a priori* models do not use data on social feedback of news items as predictors. They are solely based on descriptors of news items, and the prediction is independent of the alive-time of the items (*i.e.* constant).

We will use the approach proposed by Moniz et al. [8] which combines resampling strategies with standard regression tools. This decision is based on it being, to the best of our knowledge, the only *a priori* approach focused on accurately predicting highly popular news. The authors used four learning algorithms, namely random forests, multiple linear regression, support vector machines and multivariate adaptive regression splines, using as input the headline of the news items as a bag-of-words and sentiment scores of the title and headline.

Resampling strategies work by changing the distribution of the available training data in order to meet the preference bias of the users (*i.e.* highly popular news). Their main advantage is that they do not require any special algorithms to obtain the models - they work as a pre-processing method that creates a "new" training set upon which one can apply any learning algorithm. In Moniz et al. [8], the authors used two resampling strategies: (i) under-sampling, and (ii) SMOTEr. These methods were originally developed for classification tasks [22], [23] where the target variable is nominal, and recently extended for regression tasks [24], [25]. The basic idea of under-sampling is to decrease the number of observations with the most common target variable values (*i.e.* news with low popularity) with the goal of balancing these observations and the ones with the interesting target values that are less frequent. SMOTEr works by combining under-sampling of the frequent classes with over-sampling of the minority class (*i.e.* highly popular news). In SMOTEr, new cases of the minority class are artificially generated by interpolating between existing cases.

B. A Posteriori Models

The models proposed for the *a posteriori* strategy of predicting news popularity are based on the evolution of the number of tweets. In this paper we focus on two *a posteriori* prediction models proposed by Szabo et al. [17]: the linear regression on a logarithmic scale (*LinearLog*) and the constant scaling (*ConstScale*) prediction models. In *a posteriori* models, the time of prediction is important. We will refer to this as time slices $t, t \in (1, 2, \ldots, 144)$ which represent periods of 20 minutes after their publication (*i.e.* the first time slice represents the alive-time of 20 minutes; the third time slice represents the alive-time period between 40 and 60 minutes).

As previously noted, we did not find work concerning *a posteriori* models that were successful in the early and accurate prediction of rare cases of highly popular news. Nonetheless, we will use these two models in order to obtain ground for comparison.

The *LinearLog* model is described as

$$\hat{y}\left(n_j, t\right) = exp\left(ln\left(p_j^t\right) + \beta_0\left(t, t_f\right) + \frac{\sigma_0^2\left(t, t_f\right)}{2}\right), \quad (1)$$

where p_j^t is the number of tweets a given news n_j received until the alive-time t and t_f represents a two-days period; the parameter β_0 is computed using maximum likelihood parameter estimation given the regression function $ln\left(p_j^{t_f}\right) = \beta_0\left(t, t_f\right) + ln\left(p_j^t\right)$, on the training set; and the estimate of the variance of residuals on a logarithmic scale is given by σ_0^2.

The *ConstScale* model is expressed as $\hat{y}(n_j, t) = \alpha_2(t, t_f) \times p_j^t$, where

$$\alpha_2(t, tf) = \frac{\sum_a \frac{p_j^t}{p_j^{t_f}}}{\sum_a \left[\frac{p_j^t}{p_j^{t_f}}\right]^2}. \quad (2)$$

The parameters used to express this model are the same as those used to describe the *Linearlog* models: t is a given alive-time, t_f the final alive-time for the news item (two-days period) and p_j^t represents the number of tweets accumulated by a news item n_j until a given time t.

V. Time-Based Ensemble

Our proposal of time-based ensembles is focused on combining *a priori* and *a posteriori* models in order to overcome their individual shortcomings and provide a more early and accurate prediction of the rare cases of highly popular news. Ensemble methods train several learners to tackle the same problem and combine their outcome [26]. The most common combination methods are averaging and voting. We will focus on averaging methods which are more appropriate for regression tasks. A simple version of these methods consists of averaging the output of the learners directly. One may also use the weighted averaging method, where the combined output is obtained by averaging the outputs of each learner with different weights. This approach is common when we want to attribute different levels of importance for each learner, as in our work. Therefore, we resort to weighted averaging methods in our time-based ensemble proposals.

The scarcity of social feedback is related to the recency of the events. Therefore, the alive-time t of news items is the main factor to consider when combining models of both strategies. In Figure 3 we show the evolution of the mean proportion of available social feedback data from a sample of data of news from the topic "Obama" used in our experiments (described in Section VI-A). We remind the reader that we decided on a two-days period to observe the evolution of news popularity. The dashed line shows the evolution of the mean proportion of available data for the rare cases of highly popular news (with a relevance score equal to or above 0.9).

As depicted in Figure 3 we observe that most of the news will obtain half of their final number of tweets in the first moments succeeding its publication. As expected, for the cases of highly popular news this evaluation is slower. Given this analysis, we draw three assumptions that are the basis of our proposed hybrid strategy:

1) When social feedback is unavailable, only *a priori* models are able to predict news popularity;
2) When news items are recent, the available social feedback may be insufficient to confirm *a priori* predictions or to accurately predict popularity using *a posteriori* models;
3) As time passes since the publication of news, the available social feedback increases the accuracy of *a posteriori* predictions.

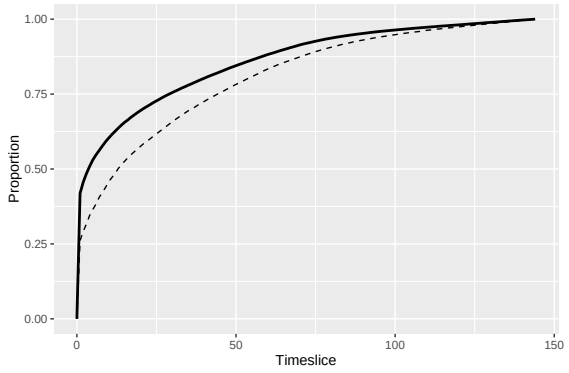

Fig. 3. Example of the evolution in mean proportion of available data in a data set of news from topic "obama". The dashed line represents the evolution of the cases considered to be highly popular

Given that the ability of *a posteriori* models to accurately predict our target cases is related to the available data on the news items, we propose to relate the weights of each learner in an ensemble with the evolution of the mean proportion of the available data at a given time t, using train data to learn its evolution. This is the reason for the time-based property of our proposed ensembles: the weights attributed to the models of each strategy are dependent of the time t in which the prediction occurs.

We propose two approaches to the combination of models in time-based ensemble resorting to the weighted averaging method: *i)* by combining the numeric predictions of the models (ENS_t), and *ii)* by combining the relevance of the models' predictions (ENS_ϕ). We recall that the concept of relevance, as proposed by Ribeiro [21] and described in Section III-A, is related to the rareness of the number of tweets in this paper: higher the popularity, higher the relevance of the news item.

When a news is published there is no related social feedback. Therefore, the predicted popularity of news items when $t = 0$ is solely based on the *a priori* models' predictions. As such, the weight of *a priori* models in the combination method of the ensemble is $w_{pr}^0 = 1$, and the weight of the *a posteriori* models is $w_{po}^0 = 0$. This is true for both our proposed approaches to time-based ensembles.

When $t \in (1, 2, \ldots, 144)$, where 144 is the final time slice, we associate the weight of *a posteriori* models to the mean proportion of available data, k_t at a given time slice t, learned with data from the training set Tr,

$$k_t = \frac{\sum_i^j \frac{p_i^t}{p_i^{t_f}}}{j}, n_i \in Tr, \quad (3)$$

where p_i^t is the number of tweets of news $n_i, i \in 1, \ldots, j$, and t_f is the final time slice. As such, the weight of the *a posteriori* models is given by $w_{po}^t = k_t$. Conversely, the weight of *a priori* models is given by $w_{pr}^t = 1 - k_t$.

The first proposed approach (ENS_t) applies weighted averaging to the numeric predictions of *a priori* models, $\hat{y}_{pr}$, and of *a posteriori* models, $\hat{y}_{po}$, where weights are associated

as previously described. Therefore, the formalization of this time-based ensemble proposal is:

$$\hat{y} = w_{po}^{t} \times \hat{y}_{po} + w_{pr}^{t} \times \hat{y}_{pr}. \quad (4)$$

We should note that *a priori* models introduce a bias in the training data due to the use of resampling strategies. This causes the average value of predictions to grow. As such, for news with very low popularity these models could cause an over-estimation of popularity. We reiterate that the prime objective of these *a priori* models is to predict rare cases of highly tweeted news, and predicting the exact number of tweets a news will receive is important, but not the main objective.

To tackle these potential issues, we propose a second approach (ENS_{ϕ}) where the weights used in the time-based ensemble $w_{po} = k_t$ and $w_{pr} = 1 - w_{po}$ are applied to the relevance of the predicted values of each model ($\phi(\hat{y}_{po})$ and $\phi(\hat{y}_{pr})$ respectively) instead of being applied to the predicted numeric target. Using this combined relevance value, we use the inverse (ϕ^{-1}) of the previously mentioned relevance function [21] to derive a predicted value of the number of tweets (*i.e.* popularity). Formally, the second proposed approach for time-based ensembles is described as such:

$$\hat{y} = \phi^{-1}(w_{po}^{t} \times \phi(\hat{y}_{po}) + w_{pr}^{t} \times \phi(\hat{y}_{pr})). \quad (5)$$

Also, we should note that one rule is introduced in both our proposals: when the number of tweets p_j^t a given news item n_j received until the time of prediction t obtains maximum relevance ($\phi(p_j^t) = 1$), we attribute the entire weight of the ensemble to the *a posteriori* models. This decision is based on the third assumption previously set forth.

VI. Materials and Methods

A. Data Used

In order to train and evaluate all the described models, it is necessary to have data on news items and their respective evolution in terms of popularity. As such, our input data consists of two datasets. The first dataset includes news items retrieved from Google News. The second describes the popularity evolution of each news item according to Twitter (*i.e.* number of tweets).

We selected four topics for our experiments: *economy*, *microsoft*, *obama* and *palestine*. The topics were selected according to the following criteria: *i)* being a permanently active topic, and *ii)* representing different types of entities (*e.g.* sector of society, company, a person and a country).

The first dataset N consists of data on news items appearing in Google News over a timespan of three and a half months (from April 1st to July 15th 2015). During the collection period, a query for each of the referred topics was posed on the Google News service every 20 minutes (*i.e.* for each time slice), and the top-100 news were retrieved. For each news item $n_i \in N$ retrieved, we collected its title, headline and publication date. Figure 4 illustrates a smoothed approximation of the amount of news per day for each topic (left), and the amount of news per topic used in our experiments (right), with an overall total of 48,527 items retrieved by 27,662 queries.

The second dataset P contains the evolution in number of tweets of news items. To obtain this data, the Twitter API[3] was used, also in 20-minute intervals. Data was collected during the interval spanning from the first moment when the news was recommended by Google News until two days past its original publication date. For 337 (0.7%) news items no information was retrieved, since they were more than two days old when recommended in the top-100 by Google News. These cases were excluded from our input data. Also, 6,550 news (13.45%) were never tweeted (*i.e.* the items obtained zero tweets).

B. Prediction Evaluation Metrics

The focus of our evaluation is on a small amount of cases, those which are rare due to their high number of tweets. As previously demonstrated, applying standard evaluation metrics may be problematic due to their focus on the average behaviour of the models. For this reason, we employ the utility-based regression metrics proposed by Torgo and Ribeiro ([21], [27]).

The authors propose that the user should be able to specify the sub-range of the target variable values which are considered to be the most relevant, relating to the concept of relevance previously described in Section III-A and depicted in Figure 2. In our experimental evaluation, the relevance threshold was set at 0.9 translating to approximately 15% of the most tweeted news being tagged as most relevant in all topics.

In our evaluation process we mainly rely on the utility-based regression metric F1, denoted as $F1_{\phi}$. It integrates the values of precision and recall according to the adaptation for the previously mentioned utility-based regression framework. An important reason for this choice is that the $F1_{\phi}$ metric considers false positive cases, where the models forecast a high number of tweets for a given case but it has in fact a low number of tweets. As such, if a model is biased to predict every case as being rare, it will obtain a low $F1_{\phi}$ score.

VII. Experimental Evaluation

We evaluate the prediction models in terms of their predictive accuracy focusing on the most popular news items. Our task is a numeric prediction task where we evaluate and compare the proposals for time-based ensembles described in Section V and baseline models of the *a priori* and *a posteriori* strategies described in Section IV, using the metric $F1_{\phi}$. The objective of our experiments is to assess if by employing time-based ensembles it is possible to achieve a significant improvement in the predictive accuracy of highly popular news in comparison to the baseline models.

The characteristics of our data require that the original temporal order of the news to be maintained. Monte Carlo simulation is used as the experimental methodology in order to obtain reliable estimates of the proposed approaches. Monte Carlo simulation randomly selects points within the available data and, for each of these points in time, it selects a window before the point for training data (Tr) and a subsequent window for test data (Ts), forming a train+test set for each of the points. To guarantee a reliable pairwise comparison, all approaches are compared using the same train+test sets. We

[3]Twitter API: https://dev.twitter.com/docs/api. The *count* method was deprecated on 20^{th} of November, 2015.

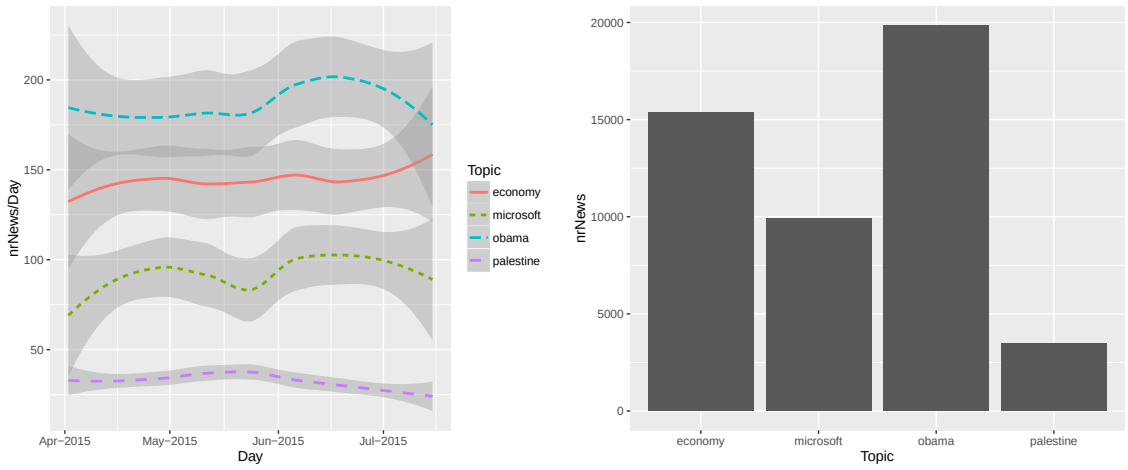

Fig. 4. Smoothed approximation of the amount of news per day (left) and Number of news per topic (right) for each topic.

should note that news that were published 2 or more days ago are not included in the test set since their observed number of tweets is already know.

Concerning baselines, we resort to the proposals for *a priori* models in the work of Moniz et al. [8] and the *a posteriori* models *LinearLog* and *ConstScale* proposed by Szabo et al. [17].

Our results were obtained through 50 repetitions of the Monte Carlo estimation process with 50% of the data as train set and the following 25% of the data as test set[4]. This process is carried out using the **R** package **performanceEstimation** [28]. The ground-truth values in these experiments are given by the total number of tweets obtained by each news in the time-span of two days since their publication.

Results include all four topics aforementioned (*economy*, *microsoft*, *obama* and *palestine*), using the metric $F1_\phi$. Our objective is to assess the predictive accuracy of the proposed and baseline approaches concerning highly popular news items. However, our focus is also on the first moments after its publication, given that after a certain period of time the prediction becomes obvious. Therefore, Table I presents the summary of the estimated metric scores for all the proposed setups in the first three time slices, denoted as $F1^1_\phi$, $F1^2_\phi$ and $F1^3_\phi$. The learning algorithms used are denoted as **rf** for random forests and **svm** for support vector machines. The use of resampling strategies in *a priori* models are denoted as: *i)* **U** when using the under-sampling strategy, and **SM** when applying the SMOTEr resampling strategy. The models using the time-based ensembles include an $\mathbf{ENS}_\phi$ tag when applied to the relevance of the models' predictions and $\mathbf{ENS}_t$ when applied directly to the numeric predictions of the models. The *a posteriori* approaches are denoted as **CS** for the *ConstScale* proposal and as **LL** for the *LinearLog* proposal. The highest scores for each topic are denoted in bold.

Results show that our hybrid strategy increases predictive performance of the rare cases of highly popular news, in the first moments after a news item is published. Comparing the baseline *a priori* models (*rf.U*, *rf.SM*, *svm.U*, *svm.SM*) and approaches from our hybrid strategy, we observe that the best results come from the latter. In most of these comparisons, the best results are obtained by time-based ensemble models combining the relevance of models' predictions (ENS_ϕ), using the respective *a priori* model and the *a posteriori* model *ConstScale*. In the topic *palestine* the best approach is given by the time-based ensembles combining the numeric predictions of models (ENS_t), the respective *a priori* models and the *a posteriori* model *LinearLog*.

Concerning the comparison of our proposed time-based ensembles and *a posteriori* models, we observe that our proposals show a better performance as to the predictive accuracy of rare cases of highly popular news. We observe that the analysis previously presented concerning the comparison between our proposed time-based ensembles and the *a priori* models also apply in the case of *a posteriori* models: time-based ensembles combining the relevance of models' predictions (ENS_ϕ) and using the *a posteriori* model *ConstScale* present the best results in most cases, with the exception of the topic *palestine*, where the time-based ensembles based combining the numeric models' predictions (ENS_t) and the *a posteriori* model *LinearLog* obtain the best results in comparison to the baseline version of the *a posteriori* models.

In order to evaluate whether these results are statistically significant (with $p-\text{value} < 0.01$), we resort to Wilcoxon signed rank tests. The results of these tests comparing the hybrid strategy to the baseline *a priori* models (as proposed by Moniz et al. [8]) and the baseline *a posteriori* models (as proposed by Szabo et al. [17]) are included in Tables II and III respectively. Results show supporting evidence of our claim that the proposed models provide a significant improvement of predictive accuracy towards the rare cases of highly popular news, when comparing to both baselines.

Concerning *a priori* models, we observe that our proposals of time-based ensembles obtain significant improvements in comparison to each of the baseline models proposed by Moniz et al. [8], when using the *a posteriori* model *ConstScale*. Nonetheless, concerning the baseline *a priori* models using the regression algorithm **svm**, the time-based ensemble proposal

[4]In order to prevent overlap, all news items in the rankings with a publication date in the last two days of the training set were also removed.

TABLE I. PREDICTION MODELS EVALUATION USING THE UTILITY-BASED EVALUATION METRIC $F1_\phi$ FOR ALL TOPICS. RESULTS REPORT METRIC SCORES IN THE FIRST THREE TIME SLICES t (PERIODS OF 20 MINUTES) OF NEWS ITEMS, DENOTED AS $F1^t_\phi$. BEST RESULTS DENOTED IN BOLD.

	economy			microsoft			obama			palestine		
Model	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$
rf.U	0.429	0.429	0.429	0.465	0.465	0.465	0.454	0.454	0.454	0.408	0.408	0.408
rf.U.ENS$_\phi$.CS	**0.584**	**0.595**	**0.600**	**0.678**	**0.696**	*0.708*	**0.563**	**0.572**	**0.582**	0.428	*0.438*	*0.448*
rf.U.ENS$_\phi$.LL	0.301	0.299	0.297	0.215	0.221	0.214	0.351	0.349	0.352	0.410	0.418	0.403
rf.U.ENS$_t$.CS	0.457	0.464	0.468	0.561	0.590	0.613	0.459	0.461	0.463	0.423	0.425	0.427
rf.U.ENS$_t$.LL	0.411	0.403	0.400	0.367	0.350	0.338	0.456	0.458	0.458	*0.434*	0.437	0.439
rf.SM	0.429	0.429	0.429	0.465	0.465	0.465	0.454	0.454	0.454	0.409	0.409	0.409
rf.SM.ENS$_\phi$.CS	**0.584**	**0.595**	**0.600**	**0.678**	**0.696**	**0.709**	**0.563**	**0.572**	*0.582*	0.428	*0.437*	*0.448*
rf.SM.ENS$_\phi$.LL	0.301	0.299	0.297	0.215	0.221	0.214	0.351	0.349	0.352	0.410	0.418	0.403
rf.SM.ENS$_t$.CS	0.457	0.464	0.468	0.562	0.590	0.613	0.460	0.461	0.464	0.423	0.425	0.427
rf.SM.ENS$_t$.LL	0.411	0.403	0.400	0.366	0.349	0.337	0.455	0.457	0.457	*0.433*	0.436	0.438
svm.U	0.444	0.444	0.444	0.461	0.461	0.461	0.056	0.056	0.056	0.450	0.450	0.450
svm.U.ENS$_\phi$.CS	*0.583*	*0.594*	*0.599*	**0.678**	*0.695*	*0.708*	*0.562*	*0.568*	*0.579*	0.404	0.438	0.441
svm.U.ENS$_\phi$.LL	0.301	0.298	0.296	0.215	0.221	0.214	0.350	0.348	0.350	0.410	0.417	0.403
svm.U.ENS$_t$.CS	0.478	0.487	0.493	0.610	0.634	0.650	0.480	0.481	0.485	0.509	0.492	0.488
svm.U.ENS$_t$.LL	0.432	0.423	0.420	0.392	0.370	0.354	0.547	0.550	0.546	**0.538**	**0.549**	**0.555**
svm.SM	0.441	0.441	0.441	0.471	0.471	0.471	0.047	0.047	0.047	0.446	0.446	0.446
svm.SM.ENS$_\phi$.CS	*0.583*	**0.595**	*0.599*	*0.677*	*0.695*	*0.708*	*0.562*	*0.568*	*0.579*	0.424	0.438	0.447
svm.SM.ENS$_\phi$.LL	0.301	0.298	0.297	0.215	0.221	0.214	0.350	0.348	0.350	0.410	0.418	0.403
svm.SM.ENS$_t$.CS	0.478	0.487	0.493	0.608	0.635	0.652	0.491	0.491	0.502	*0.521*	0.526	0.523
svm.SM.ENS$_t$.LL	0.430	0.421	0.418	0.390	0.369	0.352	0.546	0.549	0.546	0.516	*0.536*	*0.545*
ConstScale	*0.539*	*0.552*	*0.560*	*0.656*	*0.675*	*0.689*	*0.503*	*0.513*	*0.525*	0.395	*0.413*	*0.416*
LinearLog	0.275	0.274	0.272	0.197	0.204	0.197	0.321	0.318	0.322	*0.399*	0.406	0.393

TABLE II. RESULTS (p-VALUES) FROM WILCOXON SIGNED RANK TESTS DESIGNED TO TEST THE HYPOTHESIS THAT THE PROPOSED PREDICTION MODELS ARE SIGNIFICANTLY GREATER THAN THE BASELINE *A PRIORI* MODELS, ACCORDING TO THE EVALUATION METRIC $F1_\phi$.

	rfUNDER			rfSMOTE			svmUNDER			svmSMOTE		
Model	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$
ENS$_\phi$.CS	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
ENS$_\phi$.LL	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
ENS$_t$.CS	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
ENS$_t$.LL	1.00	1.00	1.00	1.00	1.00	1.00	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**

TABLE III. RESULTS (p-VALUES) FROM WILCOXON SIGNED RANK TESTS DESIGNED TO TEST THE HYPOTHESIS THAT THE PROPOSED PREDICTION MODELS ARE SIGNIFICANTLY GREATER THAN THE BASELINE *A POSTERIORI* MODELS, ACCORDING TO THE EVALUATION METRIC $F1_\phi$.

	ConstScale			LinearLog		
Model	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$	$F1^1_\phi$	$F1^2_\phi$	$F1^3_\phi$
rfU.ENS$_\phi$	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
rfU.ENS$_t$	1.00	1.00	1.00	**<0.01**	**<0.01**	**<0.01**
rfSM.ENS$_\phi$	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
rfSM.ENS$_t$	1.00	1.00	1.00	**<0.01**	**<0.01**	**<0.01**
svmU.ENS$_\phi$	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
svmU.ENS$_t$	1.00	1.00	1.00	**<0.01**	**<0.01**	**<0.01**
svmSM.ENS$_\phi$	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
svmSM.ENS$_t$	1.00	1.00	1.00	**<0.01**	**<0.01**	**<0.01**

combining models' predictions (ENS_t) using the *a posteriori* model *LinearLog* also obtains a significant increase in predictive accuracy.

Concerning *a posteriori* models, our proposal of time-based ensembles combining the relevance of the models' predictions (ENS_ϕ) is the only proposal capable of providing significant improvements in predictive accuracy concerning both the *a posteriori* models *ConstScale* and *LinearLog*, despite the baseline *a priori* model used in the ensemble. Also, results show that both proposals of time-based ensembles (ENS_t and ENS_ϕ) provide a significant increase in predictive accuracy concerning the baseline *a posteriori* model *LinearLog*.

Overall, in comparison to the baseline models, our proposed models are capable of improving the predictive ability of this task for which the goal is to accurately predict highly popular news (*i.e.* the most tweeted news). Also, we observe that the most robust model is the time-based ensembles combining the relevance of models' predictions (ENS_ϕ) since it is shown that it is the proposal that is capable of providing significant increases in predictive accuracy over both *a priori* and *a posteriori* models used in our experiments.

VIII. CONCLUSIONS

In this paper we introduce a new strategy for the early and accurate prediction of rare events in news streams. As illustrated in this paper, these rare events report to highly popular news items. Previous work has discussed the issues raised by the use of standard learning tools and evaluation metrics in tasks where predictive accuracy is focused on ranges of the target variable which are underrepresented. Since these tools and metrics are focused on predicting or evaluating the average behaviour of the data, their use is not sufficient to withdraw the conclusions as to our task of predicting the rare cases of highly popular news.

To overcome issues raised by the use of previously proposed strategies, we propose two approaches for a hybrid strategy that combines models from both *a priori* and *a posteriori* strategies, using time-based ensembles. The first proposed approach (ENS_t) combines the numeric predictions of the models of each strategy using weighted averaging. The second approach proposed (ENS_ϕ) applies weighted averaging to the relevance of the models' predictions, and the final prediction is obtained by an inverse function of relevance.

Results show that our proposed approaches are capable of significantly improving the timeliness and accuracy when predicting rare cases of highly popular news in comparison to models of both *a priori* and *a posteriori* approaches. The first proposal of time-based ensembles shows the ability to

significantly improve predictive accuracy when using the *a posteriori* model *ConstScale* but only in comparison to the baseline *a priori* models. Generally, the use of the *a posteriori* model *LinearLog* has shown a low predictive accuracy towards highly popular news in comparison to the *ConstScale* models. Results concerning the second proposed strategy (based on time and relevance) when using the *a posteriori* model *ConstScale* obtained the best overall results, showing ability to significantly improve models of both strategies in every comparison to baseline strategies.

For reproducibility, all code (written in **R**) and data necessary to replicate the results are available in the Web page http://tinyurl.com/zdt2qxb.

REFERENCES

[1] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *JISA*, vol. 5, no. 1, 2014.

[2] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, "On the dynamics of social media popularity: A youtube case study," *TOIT*, vol. 14, no. 4, pp. 24:1–24:23, Dec. 2014.

[3] K. Lerman and R. Ghosh, "Information contagion: an empirical study of the spread of news on digg and twitter social networks," in *Proc. of 4th ICWSM*, 2010.

[4] S. Petrovic, M. Osborne, R. McCreadle, C. Macdonald, I. Ounis, and L. Shrimpton, "Can twitter replace newswire for breaking news?" in *Proc. of the 26th AAAI*, 2013.

[5] M. Osborne and M. Dredze, "Facebook, twitter and google plus for breaking news: Is there a winner?" in *Proc. of the 8th ICWSM*, 2014.

[6] M. Tsagkias, W. Weerkamp, and M. de Rijke, "Predicting the volume of comments on online news stories," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 1765–1768. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646225

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

[8] N. Moniz, L. Torgo, and F. Rodrigues, "Resampling approaches to improve news importance prediction." in *IDA*, 2014, pp. 215–226.

[9] S. Kim, S. Kim, and H. Cho, "Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity," in *Proc. of the 2011 IEEE 11th CIT*. IEEE Computer Society, 2011, pp. 449–454.

[10] A. Tatar, P. Antoniadis, M. D. d. Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–12, 2014. [Online]. Available: http://dx.doi.org/10.1007/s13278-014-0174-8

[11] C. S. Lee and L. Ma, "News sharing in social media: The effect of gratifications and prior experience," *Comput. Hum. Behav.*, vol. 28, no. 2, pp. 331–339, 2012.

[12] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 365–374. [Online]. Available: http://doi.acm.org/10.1145/2433396.2433443

[13] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao, "Adapting boosting for information retrieval measures," *Inf. Retr.*, vol. 13, no. 3, pp. 254–270, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10791-009-9112-1

[14] G. Gürsun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *INFOCOM, 2011 Proceedings IEEE*, April 2011, pp. 16–20.

[15] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: Predicting the evolution of popularity in user generated content," in *Proc. of 6th ACM WSDM*. New York, NY, USA: ACM, 2013, pp. 607–616.

[16] A. Kaltenbrunner, V. Gomez, and V. Lopez, "Description and prediction of slashdot activity," in *Proc. of the LA-WEB*. IEEE, 2007, pp. 57–66.

[17] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[18] M. Tsagkias, W. Weerkamp, and M. Rijke, "News comments: Exploring, modeling, and online prediction," in *Proc. of the 32nd ECIR*. Springer Berlin Heidelberg, 2010, pp. 191–203.

[19] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proc. of the 1st WIMS*. ACM, 2011, pp. 67:1–67:8.

[20] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. of the 4th ACM WSDM*. NY, USA: ACM, 2011, pp. 177–186.

[21] R. Ribeiro, "Utility-based regression," Ph.D. dissertation, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011.

[22] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. of the 14th Int. Conf. on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *JAIR*, vol. 16, pp. 321–357, 2002.

[24] L. Torgo, P. Branco, R. Ribeiro, and B. Pfahringer, "Re-sampling strategies for regression," *Expert Systems*, vol. (to appear), 2014.

[25] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "Smote for regression," in *EPIA*, ser. Lecture Notes in Computer Science, L. Correia, L. P. Reis, and J. Cascalho, Eds. Springer, 2013, pp. 378–389.

[26] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed. Chapman & Hall/CRC, 2012.

[27] L. Torgo and R. Ribeiro, "Utility-based regression," in *Proc. of 11th European Conf. PKDD*. Springer, 2007, pp. 597–604.

[28] L. Torgo, "An infra-structure for performance estimation and experimental comparison of predictive models in r," *CoRR*, vol. abs/1412.0436, 2014.