# The Utility Problem of Web Content Popularity Prediction

Nuno Moniz
LIAAD - INESC Tec
University of Porto
Porto, Portugal
nmmoniz@inescporto.pt

Luís Torgo
Faculty of Computer Science
Dalhousie University
Halifax, Canada
ltorgo@dal.ca

## ABSTRACT

The ability to generate and share content in social media platforms has changed the Internet. With a growing rate of content generation, efforts have been directed at making sense of such data. One of the most researched problem concerns predicting web content popularity. We argue that the evolution of state-of-the-art approaches has been optimized towards improving the predictability of average behaviour of data: items with low levels of popularity. We demonstrate this effect using a utility-based framework for evaluating numerical web content popularity prediction tasks, focusing on highly popular items. Additionally, it is demonstrated that gains in predictive and ranking ability of such type of cases can be obtained via naïve approaches, based on strategies to tackle imbalanced domains learning tasks.

## KEYWORDS

Machine learning, Social networking sites, Social recommendation

**ACM Reference format:**
Nuno Moniz and Luís Torgo. 2018. The Utility Problem of Web Content Popularity Prediction. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 9 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

One of the most challenging tasks involving web content concerns accurately forecasting the degree of interest shown by users, i.e. popularity. However, an important characteristic in social media data is that items' popularity is best described by heavy-tail distributions [34]. The uneven distribution of popularity would not be a problem in this context if the under-represented items (tail) were not relevant. However, intuitively, the objective of prediction tasks involving this domain is mostly concerned with accurately predicting such under-represented cases, i.e. those with high popularity levels, since these items are those that will probably be suggested to users, or automatically promoted to improve user-experience in online platforms.

This non-standard learning scenario of data imbalance and non-uniform domain preferences (i.e. towards highly popular items) is commonly described as imbalanced domain learning [6]. Solving such tasks is an interesting and open issue, considered to be one of the most important problems in machine learning and data mining [41]. However, the majority of previous work concerning prediction of web content popularity has addressed this problem as a standard learning task, where it is assumed that the distribution is balanced and/or that users have uniform domain preferences.

This raises several issues, given that standard learning algorithms commonly optimize models by attempting to reduce a given standard evaluation metric, which is focused on the average behaviour of the data [6, 20, 23]. This approach may lead to models specialized towards well-represented cases of the data (i.e. low popularity items), resulting in models with a sub-optimal performance towards under-represented cases (i.e. high popularity items).

This paper provides an analysis of numerical web content popularity prediction approaches and their ability to forecast and rank highly relevant web content, based on the formalization of the predictive modelling task as an utility-based regression task [27]. This study includes the analysis of 5 previously proposed prediction approaches, and 2 new proposals based on strategies to tackle imbalanced domain learning tasks. Experimental evaluation efforts are focused on the online news type of social media data. The contributions of this paper are summarized as follows:

(1) Previous work proposals to popularity prediction with the best ability in predicting the average behaviour of the data, are also those with the worst performance in predicting highly popular cases, in the first moments after publication;
(2) The proposals made in this paper, based on imbalanced domain learning strategies, show significant improvements in anticipating highly popular content shortly after publication;
(3) Rankings generated by prediction models confirm that the approaches proposed in this paper obtain the best results concerning timely suggestions of highly popular content.

The remainder of this paper is organized as follows. An overview of previous work is presented in Section 2. The problem definition is formalized in Section 3 and the concept of utility-based regression in Section 4. Section 5 describes new proposals for popularity prediction, and the experimental study is presented in Section 6. Section 7 concludes the paper and provides outlook on future work.

## 2 WEB CONTENT POPULARITY

Accounting for the potential of web content popularity prediction tasks, researchers have addressed this problem differently, regarding several dimensions, such as the time of prediction and the data mining task used to formalize the problem.

Concerning the time of prediction, proposals have focused on two separate approaches: *i) a priori*, and *ii) a posteriori* prediction.

*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The first concerns the task of predicting the popularity of web content items before or upon their publication, when no social feedback from users is available (e.g. [4]). The second is related to the problem of predicting the popularity of web content items after they are published, using social feedback on the items (e.g. [32]). Regarding the data mining tasks used to formalize this prediction problem, several have been used, such as classification (e.g. [30]), regression (e.g. [3]), time series forecasting (e.g. [26]) and clustering (e.g. [40]), learn to rank (e.g. [11]), and others (e.g. [21, 39]). Tatar et al. [34] provide a survey on web content popularity prediction.

The study proposed in this paper is based on the analysis of approaches to the problem of *a posteriori* and numerical prediction of web content popularity. This includes approaches formalized as regression or time series forecasting tasks. Such proposals range from statistical approaches to applying learning algorithms, such as the proposals by Szabo and Huberman [32] and Pinto et al. [26], often used as experimental baselines. These are well representative of the types of approaches proposed over the years.

Szabo and Huberman [32] propose two statistics-based approaches, the constant scaling and the log-linear approaches. The constant scaling approach is based on the calculation of a time-dependent factor similar to a growth factor, which is multiplied by the popularity of all the items chosen for prediction, at a given prediction time. The log-linear approach explores the linear relationship of popularity values at a given timeslice (sampling intervals) and their final value, using logarithmic transformations. Other authors have proposed modelling the dynamics of popularity according to a given type of distribution, such as the PCI [17] or the Poisson distributions [29].

Concerning the application of learning algorithms, Pinto et al. [32] propose two approaches focusing on the dynamics of items' popularity. The first proposal is a multivariate linear regression model, denoting each timeslice as a popularity *delta*, i.e. the difference in popularity between consecutive intervals. A second proposal extends the multivariate linear model by accounting for the similarity of cases, using features obtained by the application of Radial Basis Functions [7]. Tatar et al. [35] and Asur and Huberman [3] rely on linear models to learn the dynamics of popularity. The former proposes a direct approach, where the model represents the relation between the popularity of items in a training set w.r.t. a given timeslice, and their final popularity. The latter extends this direct approach by including features related to sentiment analysis.

In this paper, popularity is described as the general attention obtained by web content shared in social media platforms. Different definitions have been used, focusing on the popularity of a given user or groups of users (*e.g.* [16, 31, 42]). We assume a global view of publications, considering all users' interactions (e.g. shares) equally, and using them as input to the numeric prediction process.

## 2.1 Motivation

As previously stated, one of web content popularity's most distinct characteristics relates to its distribution. This issue has been previously discussed, showing evidence that it can be generally described with heavy-tail distributions [9, 17, 22]. This poses several issues for standard learning tools and most evaluation metrics, given their focus on the average behaviour of the data. Previous contributions confirm this problem, having noted the high difficulty in predicting highly popular web content items [2, 15, 18, 19, 29].

In classification tasks, this problem is observed when using the evaluation metric Accuracy (e.g. [4, 30]): the Accuracy Paradox [43] states that a prediction model with a given Accuracy score may have better predictive ability than a model with higher Accuracy scores. Tasks with numerical target variables are also prone to such learning and evaluation issues due to the use of metrics for model optimization and evaluation that are unable to distinguish where, in the domain of the target variable, the errors occur [6, 20, 23]. Consider the synthetic prediction scenario illustrated in Figure 1 where two models ($M_1$ and $M_2$) provide sets of predictions.

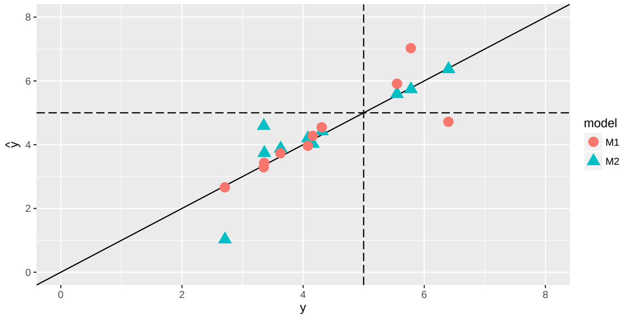

**Figure 1: Misleading scenario for standard error metrics in a regression task, using artificial data.**

Model $M_1$ obtains better predictive accuracy at low values and model $M_2$ is more accurate at the highest values. However, results from standard metrics Mean Squared Error (*MSE*) and Mean Absolute Error (*MAE*), show that both models obtain a score of 0.461 in *MSE* and a score of 0.397 for *MAE*. This shows that using standard metrics may lead to an over-estimation of prediction models' ability to anticipate the popularity of highly relevant cases.

# 3 PROBLEM DEFINITION

This work addresses two main tasks: *i)* identifying highly popular items by predicting its popularity, and *ii)* transforming this knowledge into rankings considering both recency and popularity.

The first task can be modelled as a non-standard regression problem (due to its focus on the rare cases of extreme high values) where the target variable is the popularity each incoming item is expected to have, after a predefined time period. This task is solely based on data collected in real-time, i.e. *a posteriori* prediction. Our assumption is that it is possible to map the evolution of popularity at a given point in time into the final popularity the item will obtain after a predefined period. In this work we decided on a period of two days based on the work of Yang and Leskovec [40].

The unknown function $\hat{y}()$ we want to approximate performs the mapping between the popularity $p_j^t$ of a given item $n_j$ for a given alive-time $t$, and its final popularity, $p_j^{t_f}$. The alive-time $t$ of the story, will be henceforth called a timeslice, and the final timeslice is represented by $t_f$ (*i.e.* two days). In this work, we set the duration of each timeslice to 20 minutes (*e.g.* timeslice 1 represents the period within 0-20 minutes of alive-time, timeslice 6 represents the period within

100-120 minutes, etc.). To obtain a model which approximates this unknown function we use a training set with instances of the function mapping (i.e. $D = \{\langle p_i^t, p_i^{t_f} \rangle\}, i \in (1, \cdots, n), t \in (1, \cdots, t_f)$.

Proposals have been presented that are capable of dealing with non-standard learning and evaluation scenarios as those described. Based on the concept of utility, Ribeiro [27] proposes the formalization of such learning tasks as utility-based regression tasks, and describes appropriate evaluation metrics to account for non-uniform users' preferences in imbalanced domains (Section 4.

The second task consists in generating rankings using the predictions of the approaches to the previous task, and evaluating them. The objective of this second task is to confirm that it is possible to generate timely suggestions of highly popular web content. The experimental study carried out in this paper uses a data set of online news items and their observations of popularity in several social media platforms (Section 6.1).

## 4 UTILITY-BASED REGRESSION

The problem of learning with imbalanced domains arises when two conditions are verified [6]: *i)* a subset of the target variable domain is attributed more relevance by the user w.r.t its predictive performance; and *ii)* the most relevant cases for the user in the training set are severely under-represented, causing a poor predictive performance in such cases. Unlike standard learning tasks, these conditions describe a scenario where users assign different levels of relevance to distinct types of cases.

The majority of efforts in solving tasks with such conditions have been focused on classification tasks [6, 20, 23], and according to Crone et al. [8] the majority of research concerning regression tasks does not consider uneven judgments of values' relevance. Although these properties may be valid for tasks where users' domain preferences are uniform, when regression tasks involve imbalanced domains, work by Ribeiro [27] shows that the outcome could be misleading. The author proposes the concept of utility-based regression, focusing on how to enable tasks of evaluation, comparison and model selection while accounting for uneven judgments of items relevance. This framework is based on two concepts: *i)* relevance functions, and *ii)* utility surfaces.

To define the importance of target values, Ribeiro [27] proposes the use of relevance functions allowing the user to assign relevance scores to each of the values in a given target variable, concerning a given domain: $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$. This is carried out via interpolation of relevance, given a set of user-defined pairs of relevance scores and target values, i.e. control points, using the Piecewise Cubic Hermite Interpolating Polynomials [13] (*pchip*) method. Additionally, Ribeiro proposes an automatic approach for defining relevance functions, based on generating control points via box plot statistics [37]: given a target variable $Y$, *a)* the median of $Y$ receives a relevance score of 0; *b)* the lower and upper whiskers a relevance score of 1; and *c)* all outliers a relevance score of 1. This is appropriate when users have no domain knowledge.

Figure 2 shows the interpolation of relevance ($\phi(Y)$) based on box plot statistics (top) of popularity values $Y$. In addition, users may establish relevance thresholds, $t_R$, representing a boundary for the user-definition of relevant values. It should be stressed that this boundary does not serve the purposes of discretization or the definition of irrelevant values. Its objective is to define the values that, according to the user, are the most relevant in a given domain. The data used for this illustration is described in Section 6.1.

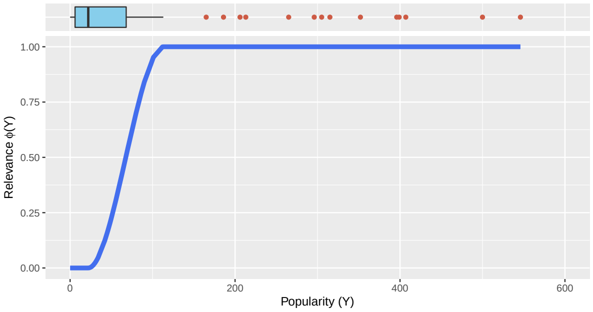

**Figure 2: Example of a relevance function with boxplot statistics (top) depiction.**

Based on the concept of relevance functions, Ribeiro [27] defines the principle of utility for imbalanced domain regression tasks. Unlike standard regression tasks, where utility is considered a function of the error of predictions, $L(\hat{y}, y)$, utility-based regression tasks consider utility as a function of such errors and the relevance of both predicted ($\hat{y}$) and true ($y$) values, bounded by $-1$ and 1.

Using this process, one is able to obtain utility functions for continuous domains, also denoted as *utility surfaces*. These can be interpreted as a continuous version of benefit matrices used in cost-sensitive learning, with classification tasks [14]. However, this proposal raises an important caveat as it allows for cases correctly predicted as highly relevant to have negative utility, when involving considerable numerical error [24].

To ensure that cases that are correctly predicted as highly relevant or non relevant have a utility bounded by $[0, 1]$, and that in analogous cases the utility is bounded by $[-1, 0]$, Moniz et al. [24] proposed a rule-based approach for generating utility surfaces. Given the domain of web content popularity, and assuming that correctly predicting highly relevant cases of popularity should not be penalized with negative utility, this approach will be employed in the experimental study. For demonstration purposes, Figure 3 exemplifies a utility surface as proposed by Moniz et al. [24], based on the relevance function depicted in Figure 2.

## 5 PREDICTION MODELS

In *a posteriori* prediction tasks it is commonly assumed that one is in possession of a set of cases $C$ describing the dynamics of their popularity evolution (training set), and that the objective is to predict the final popularity values of a second set of cases $P$, for which the evolution of popularity is known until a given timeslice $t$ (test set). In this paper, timeslices are defined as consecutive periods of 20 minutes, after the publication of a given web content item. The sets $C$ and $P$ are defined as follows, where $t_f$ corresponds to the final timeslice, i.e. prediction horizon of 2 days ($t_f = 144$), and $n$ and $m$ the size of $C$ and $P$, respectively.

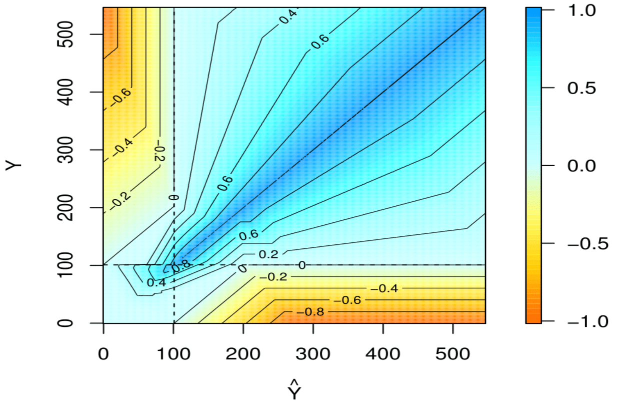

**Figure 3: Example of rule-based utility surface as proposed by Moniz et al. [24], with a relevance threshold of** $0.9$ **(dashed line).**

$$C = \begin{bmatrix} c_1^1 & \cdots & c_1^{t_f} \\ \cdots & \cdots & \cdots \\ c_n^1 & \cdots & c_n^{t_f} \end{bmatrix} \quad (1) \qquad P = \begin{bmatrix} p_1^1 & \cdots & p_1^t \\ \cdots & \cdots & \cdots \\ p_m^1 & \cdots & p_m^t \end{bmatrix}, t < t_f \quad (2)$$

A major difficulty in *a posteriori* prediction is that models must be able to forecast the final value of popularity having different levels of available data. However, in the first moments, it may be difficult to distinguish which observations relate to cases that will obtain a high level of popularity. Given the focus of standard learning algorithms in optimizing the predictions of models towards the average behaviour of the data, these will only be able to detect highly relevant cases that distinguish themselves by obtaining a high level of popularity in a short amount of time.

This problem can be tackled if the focus of learning algorithms is altered, in order to be more sensitive to highly relevant cases early on. As such, to tackle *a posteriori* prediction tasks, the use of algorithm-based methods is proposed. Such methods are a well-known strategy for tackling imbalanced domain learning [6], focused on altering learning algorithms in order to relax the influence of well-represented cases in the correct prediction of rare cases.

In this paper, we present new proposals concerning altered versions of kernel regression and $k$-nearest neighbour methods, attempting to improve predictive ability of under-represented cases in comparison to state-of-the-art approaches. The distinguishing characteristic of the proposed approaches concern the use of a biased case selection procedure. As such, instead of basing the prediction of cases on overall statistics (e.g. the average slope), the objective of the proposals is to implement this process locally.

## 5.1 Kernel-Based Approach

The first proposal is a kernel-based approach, based on the concept of kernel regression [25, 38]. The idea of this proposal is to capture the local popularity dynamics of items using locally weighted averages instead of coarse-grain statistics. This relates to the procedure of case selection, where distinct training cases should provide a differentiated contribution to the prediction of future values, depending on their distance w.r.t. the target case.

To achieve this outcome, a distance factor is introduced in the training case selection process. This factor enables the determination of an interval around the popularity value of a given web content item. This allows the selection of cases that have similar levels of popularity w.r.t. the target prediction case, at a given timeslice $t$. Concerning the definition of the interval, this proposal uses the interquartile range (IQR), considered to be a basic robust measure of dispersion and tolerant to the impact of outliers. It is defined as $IQR = Q_3 - Q_1$, where $Q_3$ and $Q_1$ report to the third and first quartile of a given continuous target variable.

The popularity of a given item may vary in consecutive timeslices, and as such the distribution of the target variable as well. Given this, the value of $IQR$ is calculated for each timeslice $t$ and is denoted as $IQR_t$, where $t \in (1, \cdots, t_f)$.

*Formalization.* Given a case for prediction $p_j$ in timeslice $t$, the kernel-based approach formulates the predictive problem as follow:

$$\hat{p}_j^{t_f} = f(p_j^t, C^t), \tag{3}$$

where $\hat{p}_j^{t_f}$ is the predicted value of popularity for the item $p_j$ in the final timeslice $t_f$, $p_j^t$ is the level of popularity at the reference time (time of prediction), and $C^t$ represent the popularity values of cases from a given training set $C$ in timeslice $t$.

By using the value of the target case at the reference time ($p_j^t$), a procedure is applied to obtain a set of cases that are within the interval of $IQR_t$, i.e. the maximum admissible distance for considering an example as similar. This is carried out by defining the lower and higher value thresholds of $p_j^t$ concerning $IQR_t$, and retrieving the index of items in $C^t$ with a value within the mentioned interval:

$$low_j^t = max(0, p_j^t - IQR_t) \quad (4) \qquad high_j^t = p_j^t + IQR_t \quad (5)$$

$$A_j^t = \{i : c_i^t \in [low_j^t, high_j^t], c \in C^t\} \tag{6}$$

Given the set of indexes $A_j^t$ representing cases similar to the target case $p_j^t$, it is necessary to calculate the weight that each train case has when predicting its final value, $p_j^{t_f}$.

The motivation for the kernel-based approach is that the number of cases that are considered as the basis for the prediction should be restricted. As such, the weight of cases is defined as the inverse distance between the popularity at the timeslice $t$ of each case $c_a^t$ where $a \in A_j^t$, and the popularity value of the target case $p_j^t$. These values are normalized into a $[0, 1]$ scale. The calculation of the weights is formalized in the following equation.

$$W_a^t = \{1 - \frac{|p_j^t - c_a^t|}{c_a^t}, \forall a \in A_j^t, c \in C^t\} \tag{7}$$

Using the train cases considered similar to the target case, and their calculated weights, the prediction of the popularity value at the final timeslice $t_f$ for a given case $p_j^t$ is carried out as follows:

$$\hat{p}_j^{t_f} = \frac{\sum_a w_a \times c_a^{t_f}}{\sum_a w_a}, a \in A, c \in C^t \tag{8}$$

## 5.2 kNN-Based Approach

The second proposal concerning algorithm-based methods is based on the $k$-nearest neighbour algorithm [1] ($k$NN). Typical settings of the method operate by deriving a subset of $k$ train cases presenting the smallest distance to the target case. Using this subset, the predictions are given by the average of their target values. In comparison to the original $k$NN algorithm, instead of providing a fixed number of neighbours $k$, in the proposed $k$NN-based approach this value is given by the amount of cases in a training set that, in timeslice $t$, have a similar popularity value w.r.t. the target case, i.e. within $IQR_t$ distance.

*Formalization.* The formalization of the $k$NN-based approach is very similar to the formalization of the kernel-based approach. The main difference between these two approaches is the non-use of weights (Equation 7). Disregarding the influence of such factors, the formalization of the $k$NN-based approach is given by the following, concerning a given target case $p_j$ and the prediction of its final popularity value w.r.t. to the reference timeslice $t$:

$$\hat{p}_j^{t_f} = \frac{\sum_a c_a^{t_f}}{|A_j^t|}, a \in A_j^t, c \in C^T \tag{9}$$

where $A_j^t$ (Equation 6) is an index set regarding cases in $C^t$ (set of values at timeslice $t$ from the train set) with a popularity value framed within an interval of $IQR_t$ (Equations 4 and 5), at the reference timeslice $t$.

# 6 EXPERIMENTAL STUDY

In this section an experimental study is presented, concerning the predictive and ranking ability of web content popularity forecasting approaches. The objectives of this study are: *i)* to assess the predictive ability of proposals on web content popularity prediction concerning highly popular items; *ii)* to compare the results with the common evaluation approach in previous work, where it is implied that users consider each item equally important; and *iii)* to draw conclusions on which formalization provides the best outcome in terms of the rankings generated by the prediction models, in a timely manner. Two new prediction approaches, based on a strategy to tackle imbalanced domain learning (algorithm-level methods) are also included in the experiments.

This study uses data concerning online news and its popularity in social media platforms. Online news is one of the most researched types of data in popularity prediction tasks. This interest is due to it being massively diffused over social media platforms, but having a short life-span, raising greater interest in the early and accurate prediction of highly popular items.

The following sections present the data, methods and evaluation methodology used, in addition to experimental results concerning both tasks formalized in Section 3: *i)* numerical *a posteriori* prediction of web content popularity focusing on highly popular items, and *ii)* the ability of these predictions in providing rankings of web content where such type of items are favoured in a timely manner.

## 6.1 Data

Experiments are based on news items concerning four specific topics: *economy, microsoft, obama,* and *palestine*. These topics were chosen due to two factors: their (worldwide) popularity and that they report to different types of entities. For each topic, a data set was generated with news suggested by Google News and Yahoo! News during a period spanning roughly eight months (November $10^{th}$, 2015 until July $7^{th}$ 2016), with the following procedure. Google News and Yahoo! News were queried simultaneously for each topic, in 20 minute intervals, and the top-100 recommended news items were collected in each query. Duplicated cases were handled by grouping items by title, headline and news outlet, and maintaining the oldest case. Figure 4 shows the global number of news per topic during the retrieval period (left) and a smoothed approximation of the amount of news per day for each topic (right). The total amount of news retrieved in both official media sources is 93,239.

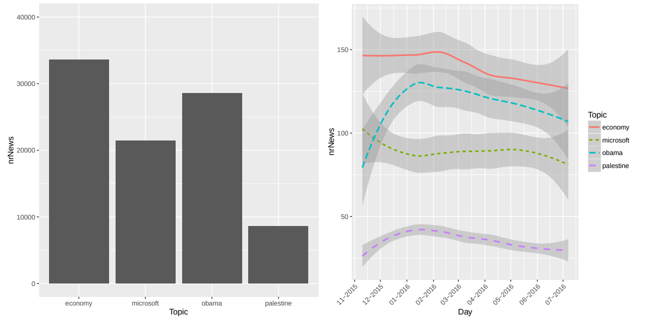

**Figure 4: Number of news per topic from Google News and Yahoo! News (left) and a smoothed approximation of the news per day ratio in each topic**

After each query to the official sources, the popularity of all known news items, with an alive-time below the defined period of two days, is obtained from the social media sources Facebook, Google+ and LinkedIn, simultaneously. The procedures for obtaining popularity levels from each social media source are as follows.

- For obtaining information from Facebook, the Facebook Graph API[1] is used, by querying for information concerning the URL of each news item. The data retrieved reports the number of shares concerning each unique URL, which is used as a popularity measure.
- The social media platform Google+ does not allow to obtain the number of shares of a given URL, but allows one to check the number of times users "liked" the URL's. Despite differences with other social media sources, it is a valid metric of received attention by news stories. This process is carried out by querying a public end-point[2] to obtain the amount of "+1" (similar to "like") a URL received.
- Finally, concerning the LinkedIn platform, the number of times each news story URL was shared is obtained by querying its public end-point[3], designed for such purposes.

Since a news item may appear in more than one position in official media sources rankings at different moments, additional

[1] The Graph API: https://developers.facebook.com/docs/graph-api

[2] Obtained by appending the respective URL to https://plusone.google.com/_/+1/fastbutton?url=. This approach is no longer available.

[3] Obtained by appending the respective URL to https://www.linkedin.com/countserv/count/share?format=json&url=.

data sets are built: for each topic and both Google News and Yahoo! News, a data set is constructed containing a news item identifier, the timestamp of the query and the respective position in the rank. This data is used in the ranking task described in Section 3.

## 6.2 Methods

To evaluate each of the tasks (prediction and ranking), two sets of baselines are used. The baseline methods used in the experimental evaluation of prediction models include the constant scaling (*ConstScale*) and log-linear (*Linear-log*) proposals made by Szabo and Huberman [32], the multivariate linear regression (*ML*) proposal and its extension using features by the application of Radial Basis Function (*MRBF*), by Pinto et al. [26], and the linear regression approach using sentiment analysis features (*LM*) proposed by Asur and Huberman [3]. These proposals are described in the related work, in Section 2. It should be noted that concerning the proposal *LM* by Asur and Huberman [3], the authors use a distribution parameter as a predictive feature. This proposal was originally focused on predicting box office revenues, using the number of theaters a given movie is presented in as a distribution parameter. In this paper, the distribution parameter is defined as the accumulated popularity of the web content item, until the moment of prediction.

Concerning the baselines used for evaluating the ranking task, the effectiveness of the proposed prediction models in generating timely news rankings is compared to three baseline strategies: *Time*, where news are ranked by time of publication with the most recent first [33]; *Live*, where news are ranked by the amount of popularity accumulated until the reference time (time of prediction) [33]; and *Source*, where news are ranked by the average final popularity of news items from their news outlet, available in the train set.

## 6.3 Evaluation Methodology

Concerning the predictive task, the focus of this experimental study is to assess the predictive approaches ability in anticipating highly popular items, but also to clarify the relation between standard and non-standard predictive approaches, such as those described in this paper. As such, the evaluation of the predictive ability of models in this experimental study is based in two metrics: *i)* the root mean squared error and *ii)* a utility-based F-Score [27].

The root mean squared error (*RMSE*) is a standard evaluation metric accounting for squared prediction errors and common in previous work concerning our scope. This metric assumes uniform domain preferences, i.e. each item is equally important. As for the utility-based F-Score, this proposal is based on the precision/recall evaluation framework [10] commonly used in classification tasks. Based on the previously detailed concepts of relevance and utility (Section 4), Ribeiro [27] presented a formulation of precision and recall for regression tasks with imbalanced domains, as in our case. The following is an alternate definition of the original proposal, for simplification purposes, based on the work of Branco [5]:

$$prec^{u}_{\phi} = \frac{\sum\limits_{\phi(\hat{y}_i)>t_R, \phi(y_i)>t_R} (1 + u(\hat{y}_i, y_i))}{\sum\limits_{\phi(\hat{y}_i)>t_R} (1 + \phi(\hat{y}_i))} \tag{10}$$

$$rec^{u}_{\phi} = \frac{\sum\limits_{\phi(\hat{y}_i)>t_R, \phi(y)>t_R} (1 + u(\hat{y}_i, y_i))}{\sum\limits_{\phi(y_i)>t_R} (1 + \phi(y_i))} \tag{11}$$

where $\phi(y_i)$ and $\phi(\hat{y}_i)$ is the relevance associated with the true value $y_i$ and predicted value $\hat{y}_i$, respectively; $t_R$ is a user-defined relevance threshold, above which cases are signalled as highly relevant for the user, and $u(\hat{y}_i, y_i)$ is the utility of making the prediction $\hat{y}_i$ for the true value $y_i$, normalized to $[-1, 1]$, using rule-based utility surfaces [24]. In this paper $t_R$ is set to 0.9, denoting approximately 10% of the items as highly relevant, in each topic. The utility-based F-Score metric $F^u_\beta$ is based on the traditional definition of the metric [28], combining both Precision ($prec^u_\phi$) and Recall ($rec^u_\phi$) with an harmonic mean, including a $\beta$ factor denoting the importance attributed to the components. In this paper it is set as 1, equally weighting precision and recall.

To evaluate the ranking task, the Normalized Discounted Cumulative Gain metric is used. It measures search result quality of ranking functions by assigning high weights to items in highly ranked positions and reducing the ones in lower ranks. Its definition is presented as follows, where $Rel_{i,q}$ is an ad-hoc relevance judgment of the $i^{th}$ ranked item for query $q$. The normalization to a value between 0 and 1 is done by dividing the *DCG* value for the ideal ordering of the ranking (*idealDCG*).

$$DCG@k(q) = \sum_{i=1}^{k} \frac{2^{Rel_{i,q}} - 1}{log_2(1+i)} \tag{12}$$

$$NDCG@k = \frac{\sum_{q=1}^{Q} \frac{DCG@k(q)}{idealDCG@k(q)}}{Q} \tag{13}$$

Finally, regarding the methods employed in this experimental study for metric estimation, given the scope of our problem and the context of the data (i.e. temporal order), the Monte Carlo simulation [36] method is used. This methodology randomly selects a set of points in time within the available data, and then for each of these points selects a certain past window as training data and a subsequent window as test data, with the overall process repeated for each random point. All alternative models are compared using the same training+test sets to ensure fair pairwise comparisons of the obtained estimates.

## 6.4 Predictive Modelling Results

This section presents the results of the experimental evaluation concerning the first task formalized in Section 3, providing evidence to address two of the objectives of this study: *i)* to assess the predictive ability of previous work on web content popularity prediction approaches concerning highly popular web content, and *ii)* to compare such results with the common evaluation approach in previous work, where it is implied that users consider each item equally important. Given that the goal is also to enable accurate predictions as early as possible (final popularity becomes evident after some time), this study is focused on the predictive ability of approaches in the first 3 timeslices, i.e. first hour after publication.

Experiments are carried out using the previously detailed data set of online news. Estimations of prediction errors concerning the *RMSE* and $F^u_1$ evaluation metrics are obtained via Monte Carlo simulation method, with 20 repetitions using 50% of cases as training

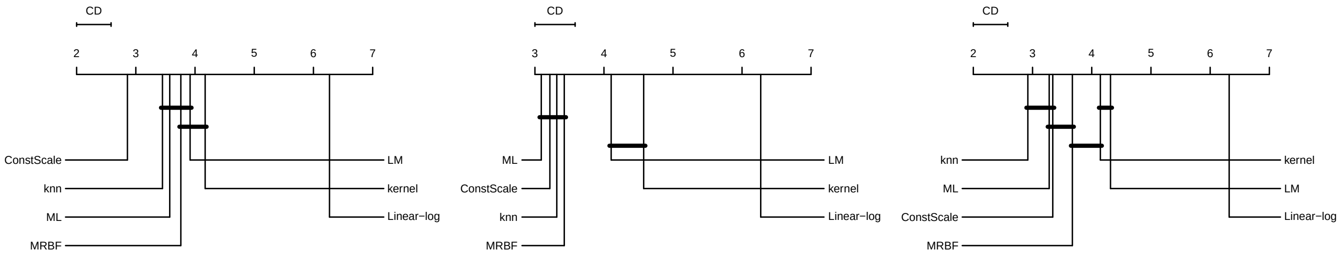

**Figure 5: Critical difference diagram concerning the results of the evaluation metric *RMSE* for models in *a posteriori* prediction in the first (left), second (center) and third (right) timeslices.**

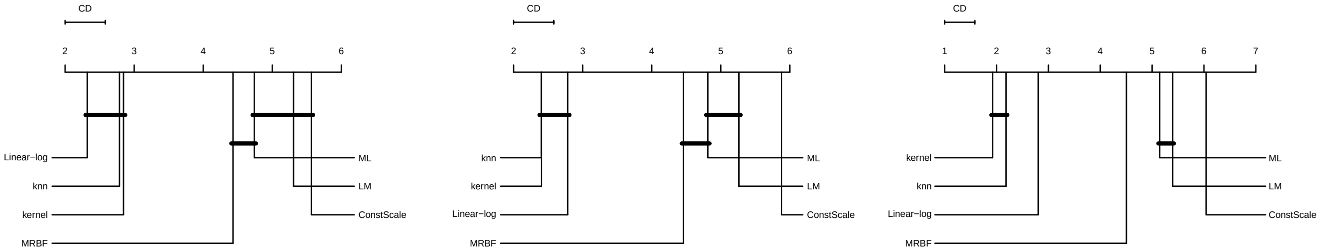

**Figure 6: Critical difference diagram concerning the results of the evaluation metric $F_1^u$ for models in *a posteriori* prediction in the first (left), second (center) and third (right) timeslices.**

set, and the subsequent 25% as test set. Results from different prediction models are compared according to the guidelines provided by Demsăr [12], and illustrated in Figures 5 and 6 concerning the *RMSE* and the $F_1^u$ metrics, respectively.

Results obtained with the *RMSE* metric show that the best predictive approaches for the first 3 timeslices are *ConstScale*, *ML* and *knn* models. The outcome is unexpected w.r.t. the *ConstScale* method, since most of the remaining models from previous works observed that their proposals provided an increased predictive ability concerning *ConstScale*. Regardless, it should be highlighted that the proposed algorithm-based method *knn* is capable of obtaining good results, presenting one of the best overall predictive approaches in the first 3 timeslices, i.e. first hour after publication.

Focusing on the utility-based evaluation metric $F_1^u$, results show an advantage of the methods proposed in this paper, the *kernel* and *knn* approaches w.r.t. previous work proposals. This outcome confirms the intuition motivating such proposed methods, concerning the possible issues of previously proposed approaches and the influence of the imbalanced distribution of web content popularity. Nonetheless, it should be noted that although *Linear-log* models show a poor predictive performance concerning the standard evaluation metric *RMSE*, they show a considerable advantage in comparison to other methods in predicting highly popular content. Conversely, the best approach according to the *RMSE* metric *ConstScale*, presents the worst outcome w.r.t. $F_1^u$ metric.

Given the outcome of the experiments and the objectives for this evaluation, results show that regarding the prediction of highly popular items shortly after publication: *i)* the proposed algorithm-based methods present the best overall results; and *ii)* the optimization of web content popularity prediction models using standard evaluation metrics can lead to an over-estimation of their ability in predicting highly popular web content.

## 6.5 Ranking Results

This section presents the results of the experimental evaluation concerning the second task formalized in Section 3, providing evidence to answer the third objective of this study: which formalization of the predictive task provides the best outcome in terms of the rankings generated by prediction models, in a timely manner.

Unlike the previous experiment, the train and test cases are rankings provided by the official media sources Google News and Yahoo! News. Prediction models are built using data from news items included in rankings of the train set, and predictions report to items from rankings in the test set. Items in the rankings of the test set are not used to build prediction models, to avoid overfit.

In addition, there is an issue concerning the short lifetime of the items. As time passes, popularity levels become evident and the usefulness in anticipating it becomes irrelevant. As such, to correctly evaluate the rankings, the ground-truth and predicted values of ranked items' final popularity are weighted by a linear factor of temporal decay $\frac{t_f - t}{t_f}$, where $t_f$ is the final timeslice (the prediction horizon is two days), and $t$ is the timeslice of prediction.

Concerning the definition of $k$ in the *NDCG@k* evaluation metric, given that the objective is to assess the quality of the ranking in its top positions, this value is defined as 10 (*NDCG@*10). This metric also requires the ad-hoc definition of relevance judgments concerning items' importance. Given that the rankings have 100 items, it is defined ad-hoc that the judgments of relevance for an

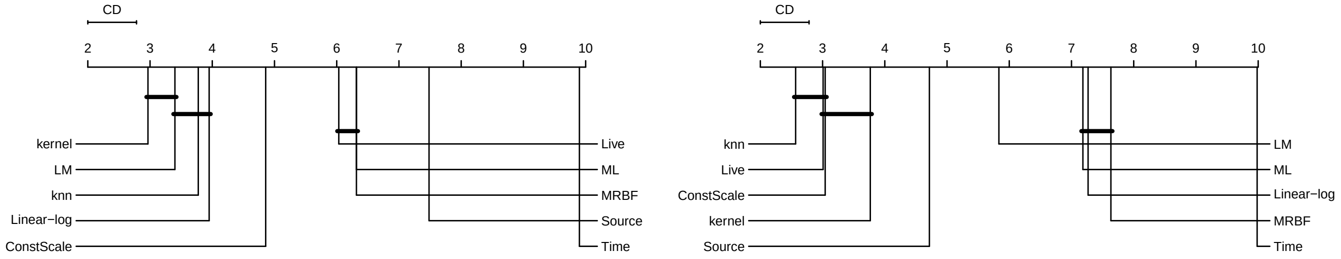

**Figure 7: Critical Difference diagram concerning rankings generated by all prediction approaches and baselines, using Google News (left) and Yahoo! News (right) rankings, according to the $NDCG@10$ evaluation metric.**

instance space of $\{0, 1, 2, 3\}$ are as such: items with the top-10 popularity values have a relevance of 3 and the remaining items in the top-25, top-50 and top-100 a relevance of 2, 1 and 0, respectively.

Estimations of the $NDCG@10$ evaluation metric are obtained via Monte Carlo simulation method, with 20 repetitions using 50% of cases (rankings from official media sources) as training set, and subsequent 25% as test set. Also, results are compared according to the guidelines by Demsăr [12], illustrated in Figure 7.

Concerning the objective of determining which formalization of the predictive task provides a better ability to rank highly popular items in a timely fashion, results show that the algorithm-level methods proposed in this paper are capable of obtaining an advantage over previous proposals to the problem of web content popularity prediction. Also, the results are coherent in terms of the models providing the worst ranking outcome. These include the *ML*, *MRBF* models, and the *Time* baseline (ranks items by recency).

However, there are discrepancies between the results when using data from Google News or Yahoo! News that should be studied further. For example, while the *Linear-log* approach is evaluated as one of the top-performers concerning the Google News rankings data, that is not true for Yahoo! News rankings, and the inverse situation is reported regarding the *LM* model. Also, results show that the baseline *Live* (using the actual observations of popularity), provides an advantage over all the approaches tested when using Yahoo! News data, with the exception of the proposed *knn* approach.

## 7 CONCLUSIONS AND OUTLOOK

In this paper an experimental study on the ability of web content popularity prediction approaches in forecasting and ranking highly popular items is presented. The motivation is two-fold. First, previous research agrees that popularity is best described by heavy-tail distributions [34]. As such, most web content obtains low levels of popularity, and a small set of items achieves very high levels. Second, the usefulness of solving this predictive task is to provide means to anticipate highly relevant items' popularity, resulting in their faster suggestion. This setting resembles the problem of imbalanced domain learning, which provides evidence on issues raised by standard learning and evaluation tools, and their over-estimation of models' ability in predicting the target cases, i.e. highly popular items. This has also been observed by previous contributions to the web content popularity prediction problem [2, 15, 18, 19, 29].

This study is based on the formalization of the predictive task as a utility-based regression problem (Section 4), accounting for the imbalanced data setting and uneven domain preferences of users. The objectives are *i)* to assess the ability of web content popularity prediction models in forecasting highly popular web content, *ii)* to compare such results with the standard evaluation approach used in previous work, and *iii)* to conclude which formalization provides the best outcome concerning the generation of timely rankings. We evaluate 5 predictive approaches from previous work and 2 new approaches based on algorithm-level methods [6, 20, 23]. The data used concerns online news obtained from Google News and Yahoo! News, and their respective popularity in Facebook, Google+ and LinkedIn, with a prediction horizon of 2 days.

Results show that previous proposals that obtain the best results in predicting the average behaviour of the data are also those with the worst performance in predicting the popularity of highly relevant cases, concerning the first 3 timeslices, i.e. first hour after publication. Also, it shows that the predictive approaches proposed in this paper provide the best predictive ability concerning the target cases. Rankings generated by the outcome of prediction models also confirm that the proposed approaches obtain the best results in terms of a timely suggestion of highly popular content.

In future work, several aspects of this experimental study should be extended, such as studying the sensitivity of the parametrization used, e.g. relevance threshold $t_R$; the use of other measures of dispersion instead of *IQR*, e.g. standard deviation; or the inclusion of a more diversified set of evaluation metrics, to better understand the potential and shortcomings of prediction approaches, with emphasis on ranking evaluation: results in our study show some disparity. Regardless, concerning the novel approaches proposed in this paper, results show that algorithm-based methods are capable of providing a considerable improvement in both prediction and ranking of highly popular web content. Given the simplicity of the approach, and the results obtained in this study, research concerning this topic should further explore the various strategies from imbalanced domain learning tasks, such as algorithm-level methods employed in this paper, but also data-level methods (i.e., data pre-processing) and hybrid methods (i.e. cost-sensitive learners and ensemble solutions).

Finally, despite the focus of this study on the online news type of web content, Shulman et al. [30] have shown the ability of prediction models in this domain to generalize well, with comparable performances in data sets of different social media sources. Regardless of such evidence, for thoroughness, this study should be extended to other types of web content in future work, to provide further evidence of such conclusions.

For the sake of reproducible science, code for prediction models and data necessary to replicate the results shown in this paper are available in the Web page https://tinyurl.com/y6wejrbm. All code is written in the free and open source R software environment.

## ACKNOWLEDGMENTS

This work is financed by the ERDF – European Regional Development Fund through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

## REFERENCES

[1] N. S. Altman. 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* 46, 3 (1992), 175–185.

[2] I. Arapakis, B. B. Cambazoglu, and M. Lalmas. 2014. On the Feasibility of Predicting News Popularity at Cold Start. In *Proc. of 6th Int. Conf. SocInfo (SocInfo 2014)*, L. M. Aiello and D. McFarland (Eds.). Springer, Barcelona, Catalonia, 290–299.

[3] S. Asur and B. A. Huberman. 2010. Predicting the Future with Social Media. In *Proc. of 2010 Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT '10)*. IEEE Computer Society, Washington, DC, USA, 492–499.

[4] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity.. In *Proc. 7th International Conference on Weblogs and Social Media*. The AAAI Press, Dublin, Ireland, 26–33.

[5] Paula Branco. 2014. *Re-sampling Approaches for Regression Tasks under Imbalanced Domains*. Ph.D. Dissertation. Universidade do Porto.

[6] P. Branco, L. Torgo, and R. P. Ribeiro. 2016. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.* 49, 2, Article 31 (Aug. 2016).

[7] D.S. Broomhead and D. Lowe. 1988. Multivariable Functional Interpolation and Adaptive Networks. *Complex Systems* 2 (1988), 321–355.

[8] Sven F. Crone, Stefan Lessmann, and Robert Stahlbock. 2005. Utility Based Data Mining for Time Series Analysis: Cost-sensitive Learning for Neural Network Predictors. In *Proceedings of the 1st International Workshop on Utility-based Data Mining (UBDM '05)*. ACM, New York, NY, USA, 59–68.

[9] Easley David and Kleinberg Jon. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge Press, New York, NY, USA.

[10] Jesse Davis and Mark Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. In *Proc. of 23rd ICML*. ACM, New York, NY, USA, 233–240.

[11] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From Chatter to Headlines: Harnessing the Real-time Web for Personalized News Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, Seattle, Washington, USA, 153–162. https://doi.org/10.1145/2124295.2124315

[12] J. Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *JMLR* 7, Jan (2006), 1–30.

[13] R. L. Dougherty, A. Edelman, and J. M. Hyman. 1989. Nonnegativity-, Monotonicity-, or Convexity-Preserving Cubic and Quintic Hermite Interpolation. *Math. Comp.* 52, 186 (1989), 471–494. https://doi.org/10.2307/2008477

[14] Charles Elkan. 2001. The Foundations of Cost-sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978. http://dl.acm.org/citation.cfm?id=1642194.1642224

[15] Shuai Gao, Jun Ma, and Zhumin Chen. 2015. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 107–116.

[16] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 57–58.

[17] Andreas Kaltenbrunner, Vicenc Gomez, and Vicente Lopez. 2007. Description and Prediction of Slashdot Activity. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*. IEEE Computer Society, Washington, DC, USA, 57–66. https://doi.org/10.1109/la-web.2007.59

[18] Y. Keneshloo, S. Wang, E. H. S. Han, and N. Ramakrishnan. 2016. Predicting the shape and peak time of news article views. In *2016 IEEE International Conference on Big Data*. IEEE, Bethesda, MD, USA, 2400–2409.

[19] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. 2011. Predicting the Virtual Temperature of Web-Blog Articles as a Measurement Tool for Online Popularity. In *IEEE 11th International Conference on Computer and Information Technology*. IEEE, Paphos, Cyprus, 449–454.

[20] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5, 4 (2016), 221–232.

[21] Jong G. Lee, Sue Moon, and Kave Salamatian. 2010. An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors. In *IEEE Conference on Web Intelligence*. IEEE, Toronto, Canada, 623–630.

[22] E Limpert, WA Stahel, and M Abby. 2001. Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience* 51, 5 (2001), 341–352.

[23] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 250 (2013), 113–141.

[24] N. Moniz, L. Torgo, M. Eirinaki, and P. Branco. 2017. A Framework for Recommendation of Highly Popular News Lacking Social Feedback. *New Generation Computing* 35, 4 (2017), 417–450.

[25] E. A. Nadaraya. 1964. On Estimating Regression. *Theory of Probability & Its Applications* 9, 1 (1964), 141–142.

[26] H. Pinto, J. M. Almeida, and M. A. Gonçalves. 2013. Using Early View Patterns to Predict the Popularity of Youtube Videos. In *Proc. of 6th ACM Int. Conf. WSDM (WSDM '13)*. ACM, New York, NY, USA, 365–374.

[27] R. Ribeiro. 2011. *Utility-based Regression*. Ph.D. Dissertation. Dep. Computer Science, Faculty of Sciences - University of Porto.

[28] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann, Newton, MA, USA.

[29] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI, Québec City, Québec, Canada., 291–297.

[30] B. Shulman, A. Sharma, and D. Cosley. 2016. Predictability of Popularity: Gaps between Prediction and Understanding. In *Proc. of 10th ICWSM*. AAAI, Cologne, Germany, 348–357.

[31] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. 2010. Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proc. of the 2nd IEEE SOCIALCOM*. IEEE, DC, USA, 177–184.

[32] G. Szabo and B. A. Huberman. 2010. Predicting the Popularity of Online Content. *Commun. ACM* 53, 8 (Aug. 2010), 80–88.

[33] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. 2012. Ranking News Articles Based on Popularity Prediction. In *Proc. of International Conference ASONAM 2012 (ASONAM '12)*. IEEE Computer Society, Washington, DC, USA, 106–110.

[34] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis. 2014. A survey on predicting the popularity of web content. *JIAS* 5, 1 (2014), 1–20.

[35] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida. 2011. Predicting the Popularity of Online Articles Based on User Comments. In *Proc. of 2011 WIMS (WIMS '11)*. ACM, New York, NY, USA, Article 67.

[36] L. Torgo. 2014. An Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models in R. *CoRR* abs/1412.0436 (2014), 1–40.

[37] J. W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley, Princeton, NJ.

[38] Geoffrey S. Watson. 1964. Smooth Regression Analysis. *The Indian Journal of Statistics, Series A* 26, 4 (1964), 359–372.

[39] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *Proc. of 13th AAAI Conference (AAAI'16)*. AAAI Press, Phoenix, Arizona, 272–278.

[40] Jaewon Yang and Jure Leskovec. 2011. Patterns of Temporal Variation in Online Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 177–186. https://doi.org/10.1145/1935826.1935863

[41] Q. Yang and X. Wu. 2006. 10 challenging problems in data mining research. *Int. J. of Inf. Tech. & Dec. Mak.* 05, 04 (2006), 597–604.

[42] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. 2014. A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics* 8, 3 (2014), 1583–1611.

[43] X. Zhu and I. Davidson. 2007. *Knowledge discovery and data mining: challenges and realities*. Information Science Reference.