New Generation Computing manuscript No.
(will be inserted by the editor)

A Framework for Recommendation of Highly Popular News Lacking Social Feedback

Nuno Moniz · Luís Torgo · Magdalini Eirinaki · Paula Branco

Received: date / Accepted: date

Abstract Social media is rapidly becoming the main source of news consumption for users, raising significant challenges to news aggregation and recommendation tasks. One of these challenges concerns the recommendation of very recent news. To tackle this problem, approaches to the prediction of news popularity have been proposed. In this paper we study the task of predicting news popularity upon their publication, when social feedback is unavailable or scarce, and to use such predictions to produce news rankings. Unlike previous work, we focus on accurately predicting highly popular news. Such cases are rare, causing known issues for standard prediction models and evaluation metrics. To overcome such issues we propose the use of resampling strategies in order to bias learners towards these rare cases of highly popular news, and a utility-based framework for evaluating their performance. An experimental evaluation is performed using real-world data to test our proposal in distinct scenarios. Results show that our proposed approaches improve the ability of predicting and recommending highly popular news upon publication, in comparison to previous work.

Keywords Social media · Recommender systems · Predictive models · Resampling strategies

N. Moniz
DCC - Faculdade de Ciências, Universidade do Porto
LIAAD-INESC TEC
E-mail: nmmoniz@inescporto.pt

L. Torgo
DCC - Faculdade de Ciências, Universidade do Porto
LIAAD-INESC TEC
E-mail: ltorgo@dcc.fc.up.pt

M. Eirinaki
Computer Engineering Department
San Jose State University
E-mail: magdalini.eirinaki@sjsu.edu

P. Branco
DCC - Faculdade de Ciências, Universidade do Porto
LIAAD-INESC TEC
E-mail: paobranco@gmail.com

1 Introduction

The creation and proliferation of platforms which enable users to post information onto the web in real-time and interact with others has brought new challenges into light. A recent international study [33] on news consumption has shown that half (51%) of the Internet users use social media as a source of news each week. As real-time micro-blogging and other social network applications become the main channels of quick information dissemination, traditional information retrieval (IR) and web page ranking approaches demonstrated several shortcomings in identifying the most relevant information for users [10].

Data dissemination services such as news aggregators provide up-to-date, real-time information to users. One of their main challenges is addressing recency [56]: to quickly identify how popular a news will be in order to rank it appropriately. Most research concerning this issue uses social feedback to predict the popularity of such content. By social feedback we refer to the reaction of users in social media platforms, such as the number of times the content is shared or commented. However, these approaches introduce a waiting period, as they need social feedback to be available in a substantial amount, in order to make accurate predictions. In this work, a study is provided concerning the task of predicting news items' popularity upon their publication, without requiring any updates on related social feedback, and using the resulting prediction to produce news ranking.

An important characteristic of news popularity is that most news obtain a very low level of popularity [44]. Therefore, only a small set of cases may be considered highly popular. Unlike previous work, our proposal is focused on accurately predicting the popularity of these rare cases as soon as possible, in order to position them in the top of the recommendations. However, rare case prediction tasks raise issues with standard learning tools and evaluation metrics as they are focused on predicting and evaluating the average behaviour of the data (*i.e.* low popularity news). To overcome this issue, we leverage our previous work on resampling strategies [31] applied to the problem of predicting rare cases of highly popular news. Resampling strategies operate by removing or/and adding cases from a data set in order to correct its skewness (*i.e.* data imbalance). The hypothesis tested is that by correcting the data imbalance in news popularity data sets, it is possible to significantly improve predictive accuracy on highly popular news, and thus provide better rankings.

As to the evaluation of such tasks, we resort to previous work by Ribeiro [37] on a utility-based regression framework which enables the evaluation of predictive accuracy on specific intervals of target variables (the most popular news items). The combination of these biased learning and evaluation approaches are formalised as an integrated framework for tackling the task of accurately predicting and ranking highly popular news.

Using a real-world data set, collected over a period of 8 months, we build predictive models using resampling strategies and use the resulting prediction to produce concrete news rankings. We focus on achieving high prediction accuracy for recent and highly popular news, which is the key distinguishing aspect of our work. Our proposed framework can be deployed as a stand-alone solution, or integrated in a news recommendations' framework, using data from two types of sources: (i) official media; and (ii) the recommendations of Internet users as they emerge from their social network activity. In the framework instantiation presented in this article, we use Google News as the official media source and Twitter as a social media source,

where social feedback (*i.e.* popularity) is defined as the number of tweets where a given news item is shared.

The main contributions of this work are:

- an extension of previous work concerning resampling strategies in the context of news popularity prediction, by introducing novel strategies;
- using such extension in an integrated framework for the prediction of news popularity and recommendation of such items without requiring related social feedback, focusing on highly popular news;
- an extensive evaluation of the proposed framework accuracy when used as a standalone solution or when integrated in a news recommendation framework, using a data set of news items from Google News and social feedback from Twitter collected over a period of 8 months.

The rest of this paper is structured as follows. In Section 2 previous work is discussed. In Section 3 the tasks addressed in our work are described. Resampling strategies are presented in Section 4. Section 5 describes the used data, regression methods and evaluation procedures. Results are presented in Section 6 and discussed in Section 7. Conclusions are presented in Section 8.

2 Previous Work

Given the shortcomings raised by traditional IR and web page ranking approaches [10], the study of content popularity and proposals for prediction approaches using social media data have proliferated in recent years. Figueiredo et al. [16] studied the popularity of Youtube videos. Lerman and Ghosh [26] studied how the structure of a network affects the dynamics of information spread by analyzing data from Digg and Twitter. Concerning news popularity and the Twitter platform, Petrovic et al. [35] conclude that Twitter covers almost all news-wire events but the opposite is not true. Osborne and Dredze [34] find that Twitter is the preferred medium for breaking news, almost consistently leading Facebook or Google Plus. Tatar et al. [44] provide a survey on web content popularity prediction.

News popularity prediction tasks can be framed in two scenarios: *i) a priori*, when related social feedback is unavailable; and *ii) a posteriori* prediction, when related social feedback is available. A considerable distinction is that *a priori* proposals use meta-data concerning the target items and *a posteriori* proposals use related social feedback, as it become available.

Most previous work has been focused on the *a posteriori* scenario. According to Tatar et al. [44], 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; *ii)* temporal analysis, studying the evolution of content popularity over time until the prediction moment; and *iii)* popularity evolution trends, using clustering methods to find similar items. The evaluation of many of these proposals use data from different social media platforms, but are nonetheless important to consider given that recent research shows that they generalise well [38].

Regarding cumulative growth approaches, early work by Kaltenbrunner et al. [24] proposes a constant growth approach for predicting the popularity of Slashdot stories, depending on the publication hour of the news stories. Szabo et al. [41] propose two methods for predicting the popularity of YouTube videos and Digg stories: the *linear log* and *constant scaling* methods. Tsagkias et al. [51] and Tatar et al. [42] used the former methods in news popularity prediction tasks. Tatar et al. [45] extend their

work and show that the prediction methods outperform several state-of-art learning to rank approaches [18, 17, 53, 52]. Also, Lee et al. [25] propose a survival analysis approach with a Cox proportional hazard regression model to tackle this task.

As to the temporal analysis approach, Pinto et al. [36] propose to frame the popularity prediction task as a multivariate linear regression task; Wu et al. [52] use a reservoir computing approach and Gürsun et al. [20] propose a time series approach focusing on frequently-accessed items. As to the popularity evolution trends approach, Pinto et al. [36] propose to combine multivariate linear regression with radial basis functions and Gürsun et al. [20] propose an hierarchical clustering approach to predict rarely-accessed items. Also, Ahmed et al. [1] propose a transition model based on identified classes of behaviour.

The main requirement concerning *a posteriori* approaches is availability of social feedback related to the target items. However, when an item is published no social feedback is available, and some alive-time is required to have enough data for an accurate prediction. As previously mentioned, the focus of our work is on accurately predicting the popularity of news upon publication, when social feedback is nonexistent or scarce, making *a posteriori* approaches not suitable. To tackle this issue, *a priori* approaches have been proposed.

In the *a priori* scenario, Tsagkias et al. [50] use 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. [3] propose classification and regression approaches using four predictors: source, category, subjectivity in the language and named entities mentioned.

Recently, a third prediction strategy has been proposed: a hybrid strategy using time-based ensembles [32] to combine approaches from the two known strategies. The cerebration of this proposal lies in the shortcomings of both *a priori* and *a posteriori* strategies. The former allows for the prediction of news popularity when social feedback is unavailable, but does not allow for an update of such predictions when social feedback becomes available. On the other hand, the latter relies solely on social feedback. As such, is incapable of accurately predicting news popularity until there is a sufficient amount of feedback.

According to Tatar et al. [44] the popularity of online content is described with a power-law distribution. As such, most of the content obtains a low level of popularity and only a small amount of rare cases achieve very high popularity levels. Proposals in previous work tackle this task of predicting the popularity of online content mainly using approaches based on standard prediction models that by definition optimize performance for the average behaviour of the data. Although these approaches are capable of achieving good results, given the power-law distribution of online content popularity, most of the correctly predicted items have low popularity. The main distinction of our work and previous work is that our proposal is focused on accurately predicting rare cases of highly popular news in order to position such items at the top of the recommendations as soon as they are published.

Finally, in this paper a stochastic view of the popularity concept is assumed (*i.e.* aggregate behaviour), considering all tweets and retweets from every user equally, which are used as input in the numeric prediction task. Different tasks have been discussed, such as those focused on determining the number of "retweets" (i.e. Twitter functionality to re-publish a tweet) a given tweet will obtain ([40, 55, 21]) or those using data concerning the social network of individual users to predict the popularity of content generated by the user ([19, 22]). The task tackled in this paper is not

focused on the popularity of content generated by a single or a given group of users, but on the general popularity of content in social media platforms.

The current paper extends our previous work [31] by significantly increasing the amount of data analysed, by evaluating new resampling strategies, but mainly by using such strategies to propose an integrated framework for producing news rankings using the results of the previously proposed prediction methods and the evaluation of such rankings in different deployment scenarios.

3 Problem Description

The objective of our work is twofold: a) to accurately predict the popularity of news stories upon their publication, focusing on the news that will be highly popular, and b) to transform this outcome into news rankings.

The task of predicting news popularity has been previously formalised as a classification (*e.g.* [50, 19]) or ranking task (*e.g.* [11, 28]). However, one of the main challenges in our work is, as previously described, the ability to predict popularity as a numeric target variable. As such, the first task is a numeric prediction task (*i.e.* regression) where we are trying to forecast the number of tweets a news item will receive based on some description of the news. However, this task has one particularity: we are only interested in achieving high prediction accuracy on a small sub-set of the news - the ones that will be tweeted the most (*i.e.* the most popular). These are the news that the public deems as highly relevant for a given topic and these are the ones we will subsequently place at the top of our news recommendation.

The second task consists on the generation and evaluation of news rankings using the results of the first task. This second task should confirm the possibility of reducing the latency period (*i.e.* time between publication and an appropriate position in the recommendation) related to the recency of news.

In this paper we describe and test several approaches aimed at improving the predictive performance on the difficult task of predicting the popularity of a news item. We then apply these prediction models in a framework and evaluate them in two scenarios: *i)* as a standalone system where a news pool of recent news is provided, and *ii)* its application to ranks provided by Google News. The objective of the evaluation is to ascertain the ability of our framework in accurately predicting the most popular news at any given time.

3.1 Data Mining Tasks

The first task, where our goal is to forecast the number of tweets of a given news, is a numeric prediction task, usually known as a regression problem. This means that we assume that there is an unknown function that maps some characteristics of the news into the number of times this news is tweeted, *i.e.* $Y = f(X_1, X_2, \cdots, X_p)$, where Y is the number of tweets in our case, $X_1, X_2, \cdots, X_p$ are features describing the news and $f()$ is the unknown function we want to approximate. In order to obtain an approximation (a model) of this unknown function we use a data set with examples of the function mapping (known as a training set), *i.e.* $D = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$.

The standard regression task we have formalised can be solved using many existing algorithms, and most of them try to find the model that optimises a standard error criterion like the mean squared error. What sets our specific task apart is the fact that we are solely interested in models that are accurate at forecasting the rare and high values of the target variable Y, *i.e.* the news that are highly tweeted. Only this

small subset of news is relevant for our overall task of providing a ranking of the most popular news for a given topic.

Regarding the second task, rankings are produced using the outcome of the first task, the predicted number of tweets for each given news item in a set of news. This means that, given these predicted numbers of tweets, this second task executes the trivial process of ranking the news by decreasing predicted number of tweets.

3.2 Imbalanced Distribution of Popularity

Previous work [37, 48] has shown that standard regression tools fail in tasks where the goal is accuracy at the rare extreme values of the target variable (*i.e.* the most tweeted news), as in the problem we address with this work. We exemplify these potential problems in the scenario described in Table 1 and depicted in Figure 1 using synthetically generated data. In this scenario, two models (M_1 and M_2) provide their respective sets of predictions, presenting a scenario where resampling strategies are useful.

Table 1 Predictions made by two artificial models M_1 and M_2 with their respective error and the ground-truth values.

Predictions of Two Artificial Models										
True	2.71	3.35	3.36	3.63	4.08	4.16	4.31	5.55	5.78	6.40
M_1	2.67	3.29	3.43	3.73	3.97	4.28	4.54	5.91	7.03	4.72
Loss M_1	0.04	0.06	0.03	0.04	0.11	0.12	0.23	0.64	1.45	1.68
M_2	1.03	4.59	3.74	3.88	4.20	4.03	4.42	5.59	5.74	6.37
Loss M_2	1.68	1.24	0.72	0.45	0.88	0.77	0.89	0.04	0.04	0.03

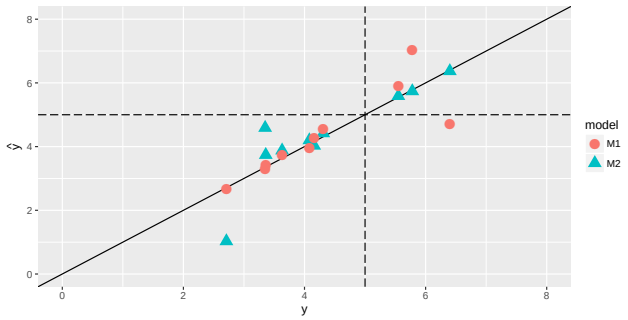

Fig. 1 Misleading Scenario for Standard Error Metrics with Artificial Data.

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 standard 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. This occurs because these metrics are unable to distinguish where (in the domain of the target variable) the errors occur. In order to overcome this issue of evaluating models in regression tasks focused on rare values, we resort to a utility-based framework proposed by Ribeiro [37] and the evaluation metrics described in Section 5.3.1.

This framework distinguishes rare and normal cases based on the distribution of the target variable and the concept of relevance as proposed by Ribeiro [37]. The

relevance expresses the bias in the domain and is defined as a continuous function $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$ mapping the target variable domain $\mathcal{Y}$ into a $[0, 1]$ scale of relevance. Being a domain-based function, this framework requires an input by the user concerning the relevance function but also the relevance threshold, from which we define the set of relevant cases.

In this paper, the concept of relevance relates to the rareness of news items popularity. This is due to the observation that the popularity of news items resemble a power-law distribution, as noticed by several authors [43, 39, 24]. One of the most common argument for this behaviour is related to the rich-get-richer effect [8], where items that become popular have a higher probability of becoming even more popular because they are promoted. Therefore, the large majority of published news items gathers a low level of popularity and a small set of rare news items obtain very high popularity levels. As such, we consider that 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 $\phi()$ based on box plot statistics. This automatic approach uses a piecewise cubic Hermite interpolation polynomials [12] (*pchip*) algorithm to interpolate a set of points describing the distribution of the target variable (*i.e.* popularity). The choice of interpolation method was based on its simplicity and effectiveness in interpolating discrete data. Moreover, by restricting the first derivative values at the control points they are capable of preserving local positivity, monotonicity and convexity of the data. These are most convenient properties in the context of our target scenarios. These points are given by box plot statistics. The outlier values (either extreme high or low) according to the box plot statistics of the target variable (*i.e.* popularity) are given a maximum relevance of 1 and the median value of the distribution is given a relevance of 0. The relevance of the remaining values is then interpolated using the *pchip* algorithm.

By combining this mapping and the relevance threshold t_R provided by the user, it is possible to determine the cases in our data considered to be "rare", the ones we want to apply some bias in the prediction models. This procedure is also crucial for the evaluation of such prediction models, as further discussed in Section 5.3.1.

In Figure 2, we show an example of the automatically generated relevance function for a sample of data of the topic *economy* described in Section 5.1, with a relevance threshold of 0.9. The value of the relevance threshold was defined in order to obtain a small set of cases considered to be highly popular. In this example, a news item is considered rare if it obtains a popularity of 47 or more, corresponding to 10% of the sample cases.

4 Resampling Strategies

Several methodologies have been proposed for addressing imbalanced domains, mainly in classification tasks. Resampling methods are among the simplest and most effective. Resampling strategies work by changing the distribution of the available training data in order to meet the preference bias of the users. 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 this paper we will experiment with four resampling strategies for regression tasks as implemented by Branco et al. in the **R** package **UBL** [6]: i) random under-sampling, ii) random over-sampling, iii) SMOTEr, and iv) importance sampling. Al-

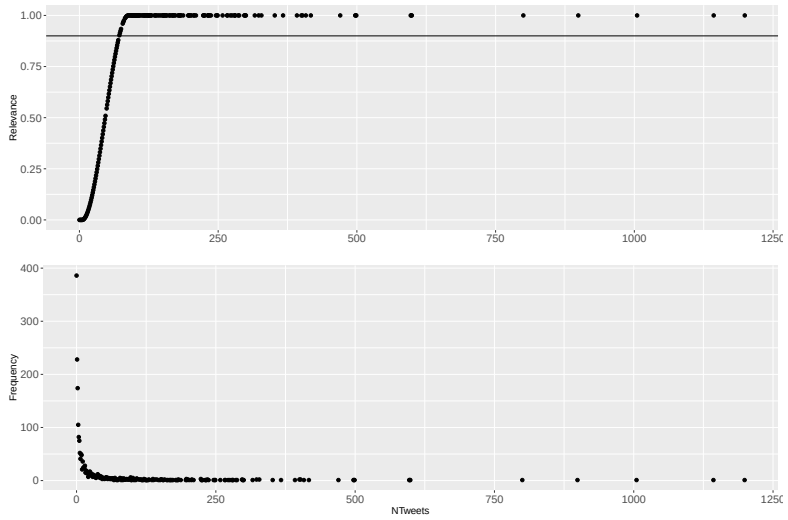

Fig. 2 Distribution of news popularity and the automatically generated relevance function for a sample of data from the topic *economy* described in Section 5.1.

though resampling strategies were originally developed for classification tasks where the target variable is nominal, recent work [48, 49] has adapted these methods for regression tasks where the target variable is numeric. The distinction between the "normal" and "rare" items is carried out by resorting to the relevance concept as proposed by Ribeiro [37] and previously described in Section 3.2. As such, in our particular case, the "normal" cases are those which have a relevance below the relevance threshold t_R which is set by the user, and the "rare" cases are those which have a relevance equal or higher than the threshold.

The objective of the first resampling strategy, random under-sampling, is to decrease the number of observations with the most common target variable values, to better balance the ratio between these observations and the ones with the interesting target values (which are less frequent). Oppositely, the objective of the second resampling strategy, random over-sampling, is to increase the number of observations with the less frequent target variable values by resorting to replication, with the same goal of improving the ratio between cases with "normal" and "rare" target values. In both these strategies, the selection of cases to remove or append is done randomly.

Concerning the SMOTEr strategy, it combines under-sampling of the frequent classes with over-sampling of the minority class, through artificial generation of new "rare" cases. The generation of synthetic cases is done by interpolating between pairs of existing cases. This process is not as trivial as in the original SMOTE algorithm [7] where the rare cases have the same target variable. In this case, although in a given pair of items both have high relevance, they may not have the exact same numeric value. To tackle this issue, the authors propose a weighted average of the target variable defining the weight as an inverse function of the distance between the generated case and the cases used.

The final resampling strategy, importance sampling, is a novel proposal that combines probabilistic under- and over-sampling strategies. This method does not require the setting of a relevance threshold, being solely based on the relevance function. The key idea of this method is to use the relevance function as a probability for both under-

and over-sampling the examples. Therefore, the higher the relevance of a case, the higher the probability of it being selected to be replicated. Oppositely, the lower the relevance of a case, the higher the probability of it being selected for removal.

In our experiments we apply and compare these strategies in order to determine the best in accurately forecasting the number of tweets of highly popular news items. We should note that in each of the presented strategies the percentage of cases to be removed and/or added by the methods is defined by the user.

In Figure 3 we show the impact of applying the presented resampling strategies. We depict the changes in density of the data[1], using the previously mentioned sample from data of the topic *economy*, described in Section 5.1. For the purposes of this demonstration we used a 0.05% percentage of under-sampling and a 1.1% percentage for over-sampling.

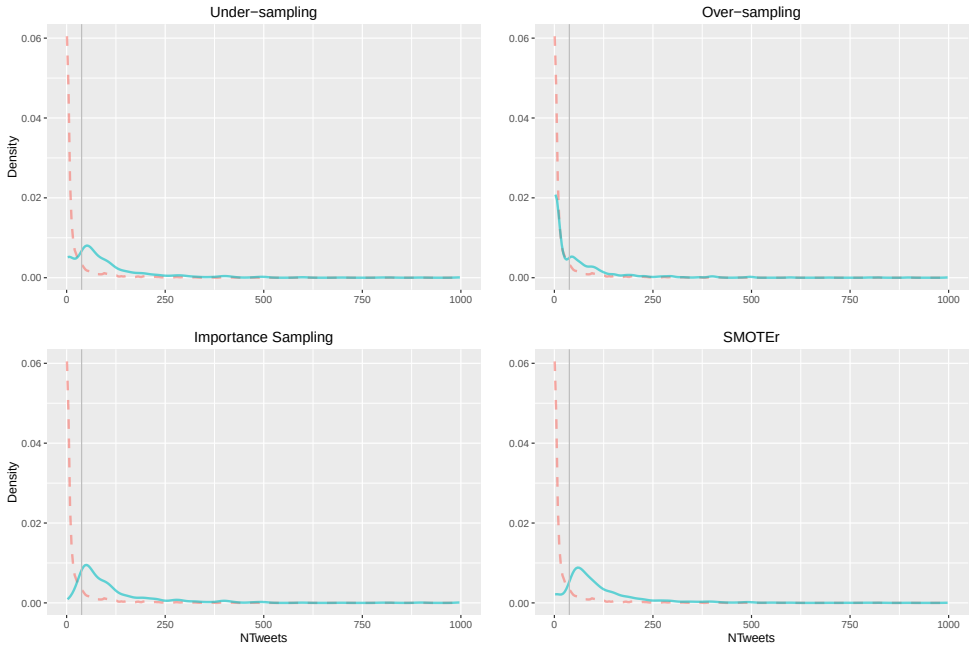

Fig. 3 Density of the data sample when resampling strategies are applied in comparison to the original data (in red/dashed). The grey line delimits the number of tweets as to "normal" or "rare", given a relevance threshold of 0.9.

5 Data and Methods

5.1 The Used Data

The experiments are based on news concerning four specific topics: *economy, microsoft, obama,* and *palestine*. These topics were chosen due to two factors: their (worldwide) popularity and the fact that they report to different types of entities (sector, company, person, and country, respectively). For each of the four topics we

[1] This process was carried out using the **density** function in **R**

have constructed a dataset with news suggested by Google News during a period spanning roughly eight months (2014-May-01 - 2015-Jan-01). The datasets were created by querying Google News every 30 minutes and collecting the top-100 recommended news. Figure 4 shows the total number of news per topic during this period (left) and a smoothed approximation of the amount of news per day for each topic (right).[2]

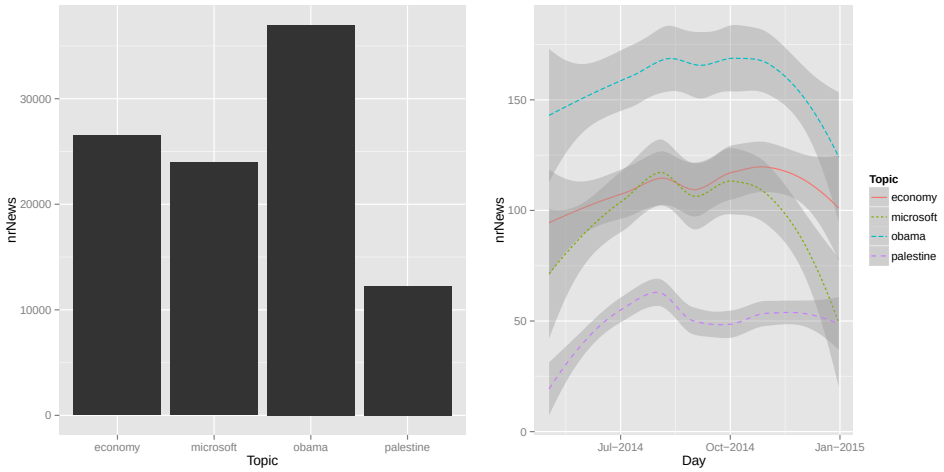

Fig. 4 Number of news per topic (left) and a smoothed approximation of the amount of news per day for each topic

For each news recommended by Google News the following information was collected: title, headline (the subtitle, also known as "lead"), publication date, the news source and its position in the ranking. For each of the four topics a dataset was built for solving the predictive task formalized in Section 3.1. These datasets were built using the following procedure. For obtaining the target variable value we have used the Twitter API[3] to check the number of times the news was tweeted in the two days following its publication. This two days' limit was decided based on the work of Yang and Leskovec [54] and our own observations, both of which suggest that after a few days the news stop being tweeted. Of the total number of news for all topics (106.456), in 6.411 cases (6%) it was not possible to obtain the number of tweets and in 19.719 cases (18, 5%) the news items were not tweeted.

Concerning predictor variables, previous work provides evidence concerning which variables have strong predictive power and those that do not. Bandari et al. [3] show that the media source of the news items is a strong predictor of popularity. Textual features are used by Tsagkias et al. [50], by extracting the top-100 most discriminative terms and observing that they show a strong predictive performance. Additionally, they use named entity identification, adding predictors concerning average behaviour (as to popularity) when certain entities are referenced in news. Subjectivity of language and sentiment analysis have also been used in this numeric prediction tasks using news popularity. Bandari et al. [3] used subjectivity of language, but report

[2] The total number of news for all topics is 106.456.

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

little predictive power by this predictor. However, in the case of sentiment analysis, Berger et al. [4] show that is it a good indicator of articles' virality.

Considering the indications provided by previous work, the predictor variables we used to describe each news are as follows. We have applied a standard bag-of-words approach to obtain a set of terms describing it, using the infrastructure provided by the R package **tm** [15]. This was applied to the headline of the news, given that initial experiments have shown that the headline provides better results than the title of the news item[4]. We have not considered the use of the full news text as this would require following the available link to the original news site and have a specific crawler to obtain this text. Given the wide diversity of news sites that are aggregated by Google News, this would be an infeasible task. Punctuation, numbers and stop words were removed, and sparse terms were also removed. Depending on the topic, we roughly used between 100 and 200 terms to describe the news items. To this set of predictors we have added two sentiment scores: one for the title and the other for the headline. These two scores were obtained using the sentiment dictionary described by Hu and Liu [23]. Finally, we use the average number of tweets for the respective news source, learnt using train data. Our decision on using such predictors is motivated by the attempt to understand the high-level popularity potential of news items in three dimensions: *i)* the presence of key terms; *ii)* the polarity and magnitude of sentiment shown in both the title and headline; and *iii)* the level of attention given to the respective media source of the news. Table 2 presents a summary of the information describing each news item in our data sets.

Table 2 The variables used in our predictive tasks

Variable	Description
$NrTweets$	The number of times the news was tweeted in the two days following its publication. This is the target variable Y.
$T_1, T_2, \cdots$	The term frequency of the terms selected through the bag of words approach when applied to all news headlines.
$SentTitle$	The sentiment score of the news title.
$SentHeadline$	The sentiment score of the news headline.
$SourceAvg$	The average number of tweets of the news source.

Since a news item may appear in more than one position in the Google News ranking at different timestamps, additional data sets were built for each topic with such information: for each topic and Google News query, a data set is constructed containing a news item identifier, the timestamp of the query and the respective position in the rank.

5.2 Regression Algorithms

In order to test our hypothesis that using resampling methods will improve the predictive accuracy of the models on the cases that matter to our application, we have selected a diverse set of regression tools. Our goal is to ensure that our conclusions are not biased by the choice of a particular regression tool.

Table 3 shows the regression methods and tools that were used in our experiments. To allow for an easy replication, we have used the implementations of these tools

[4] Preliminary results obtained with evaluation metric described in Section 5.3.1 show that the application of a bag of words approach to the headline of news produces better results than when applied to the title.

available in the free and open source R environment. Concerning the parameter settings, we implemented a method in order to discover the optimal parametrization (*i.e.* the setting that obtains the best possible results within a certain set of values of the parameters) for each of these regression methods. The search for optimal parameters was carried out for each combination regression method - data set, and the results are detailed in Appendix A.

Table 3 Regression algorithms and respective R packages

ID	Method	R package
RF	Random forests	randomForest [27]
SVM	Support vector machines	e1071 [29]
MARS	Multivariate adaptive regression splines	earth [30]

5.3 Evaluation Metrics

The metrics presented here are used to evaluate the performance of our approach in both their tasks: i) the prediction of rare cases of highly tweeted news and ii) the production of news rankings using those predictions.

5.3.1 Prediction Evaluation Metrics

Regarding the performance evaluation of prediction models, the focus of our problem is accurately predicting a small amount of cases, i.e. those which are rare due to their high number of tweets. Standard regression metrics (e.g. mean squared error) are unsuitable for these problems as previously shown in Section 3.2. In classification, the solution usually revolves around the use of the precision/recall evaluation framework [9]. Precision provides an indication on how accurate is the model in predicting the rare cases. Recall signifies how frequently the rare situations were identified as such by the model. To achieve a similar solution for regression tasks an extra degree of complexity must be considered: the notion of numeric precision.

Torgo and Ribeiro [46] and Ribeiro [37] presented a formulation of precision and recall for regression tasks with imbalanced distributions, as in our case. This framework for utility-based regression considers both the user preference biases and the issue of numeric accuracy.

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. In our paper, the concept of relevance relates to the rareness of news items popularity (*i.e.* higher the popularity, the more relevant the case). Ribeiro [37] describes an automatic method for obtaining the relevance function when the user goal is to be accurate at rare extreme values, as in our case. This method is detailed in Section 3.2 and is used to obtain the relevance functions of the data sets used in our experiences.

As to the performance evaluation of prediction models, we will rely on the regression variant of the F-measure, henceforth referred as $F1_\phi$, that depends on the precision ($prec_\phi$) and recall (rec_ϕ) measures proposed by Branco [5] and based on the mentioned framework of Torgo and Ribeiro [46] and Ribeiro [37]. In this context, precision and recall as defined as:

$$prec_\phi = \frac{\sum_{\phi(\hat{y}_i)>t_R}(1+u(\hat{y}_i,y_i))}{\sum_{\phi(\hat{y}_i)>t_R}(1+\phi(\hat{y}_i))} \quad (1) \qquad rec_\phi = \frac{\sum_{\phi(y_i)>t_R}(1+u(\hat{y}_i,y_i))}{\sum_{\phi(y_i)>t_R}(1+\phi(y_i))} \quad (2)$$

where $\phi(y_i)$ is the relevance associated with the true value y_i, $\phi(\hat{y}_i)$ is the relevance of the predicted value $\hat{y}_i$, t_R is a user-defined threshold signalling the cases that are

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 $[-1, 1]$.

Utility is commonly referred to as being a function combining positive benefits and costs (negative benefits). In the context of this paper, utility is a function of both the relevance of the true ($\phi(y)$) and predicted ($\phi(\hat{y})$) values, and the numeric difference between them (prediction error). In regression, the utility surface is a continuous version of the benefit matrix proposed by Elkan [13]. This surface allows the evaluation of regression models for tasks where the target values have a different relevance for users, such as the tasks tackled in this work.

The seminal work on utility-based regression by Ribeiro [37] proposes an approach to obtain utility surfaces. This proposal is focused on actionable forecasting tasks, motivated by the concept of activity monitoring [14]. Actionable forecasting entails the process of predicting the correct action, inferred by a function of a predicted numerical variable. The proposal for utility states two criteria for considering a prediction as beneficial (*i.e.* positive utility): *i)* the predicted value leads to the correct action, and *ii)* the deviation of the predicted value is within a maximum admissible loss (prediction error). The second criteria raises caveats regarding the tasks tackled in this paper, since it allows for highly relevant values which are predicted as such, to have a negative utility. Therefore, we propose a new approach for utility surfaces.

Our proposal builds on previous work [37], by also considering utility as a function of both: i) the relevance of true and predicted values and ii) a loss function. However, we establish an extra group of constraints regarding the calculated utility values, as follows:

1. If a case is correctly predicted as highly relevant or non relevant, its utility is bounded by $[0, 1]$;
2. If a case is incorrectly predicted as highly relevant or non relevant, its utility is bounded by $[-1, 0]$;

These restrictions, use a threshold on the relevance values to ensure that no cost (i.e. negative benefit) is assigned to points that are both either highly relevant or non relevant.

To interpolate the values of utility we resort to the R package **akima** [2] which provides a framework for the interpolation of irregularly and regularly spaced data. This interpolation is carried out by providing pairs of points (true and predicted values) where the utility value is known, described as such:

- When the predicted value is the same as the true value ($\hat{y} = y$), the utility is equal to the relevance of the true value, $\phi(y)$;
- Pairs of true and predicted values where one corresponds to its maximum value (*e.g.* $max(\hat{y})$) and the other to a value with the same relevance as the relevance threshold (*e.g.* $y : \phi(y) = t_R$), the utility value is equal to 0.
- Cases where the prediction error is maximized (*e.g.* $(max(\hat{y}), min(y))$), the utility value is equal to -1.

Figures 5 and 6 provide a visual comparison of the utility surface proposed by Ribeiro [37] and our proposal, using a data sample from the topic *economy*, and the regression algorithm MARS.

Confirming our previous considerations, the utility surface proposed by Ribeiro shows that it is not appropriate given the tasks dealt with in this paper, since it allows for the prediction of cases as highly relevant to obtain a negative utility. We will use

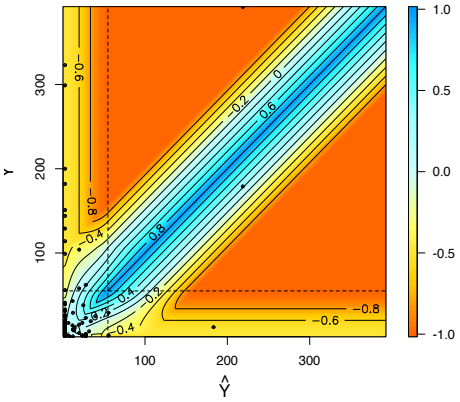

Fig. 5 Example of utility surfaces as proposed by Ribeiro [37].

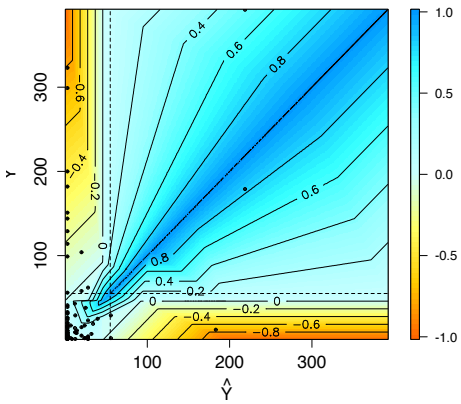

Fig. 6 An example of our proposal for utility surfaces.

the metrics $prec_\phi$, rec_ϕ and $F1_\phi$ using our proposed approach for utility surfaces, in order to compare and evaluate the type of tasks dealt in this paper. In our target application, the relevance threshold is set at 0.9 representing approximately 15% of the most popular news being tagged as rare.

We have shown that standard evaluation metrics are not appropriate in tasks where the focus of predictive accuracy is on a given numeric interval of the target variable. Nonetheless, given the widespread use of such metrics in previous work (*e.g.* [24, 41, 51]), it is important to provide a comparison of such metrics. One of the most used in previous work concerning our scope is the root-mean-square error ($rmse$), which is primarily focused on the numeric prediction error. In addition to this standard evaluation metric, it is important to analyse the impact of considering the concept of relevance in calculating the numeric prediction errors. As such, we propose the use of $rmse$ and a second metric ($rmse_\phi$), similar to the former, where items are weighted by their respective relevance (ϕ). These metrics are defined as follows:

$$rmse(\hat{y}, y) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \quad (3) \qquad rmse_\phi(\hat{y}, y) = \sqrt{\frac{\sum_{i=1}^{n}\phi(y_i) \times (\hat{y}_i - y_i)^2}{n}}, \quad (4)$$

where, the predicted and true values are respectively denoted as $\hat{y}$ and y; n is the number of items in the set; and $\phi(y)$ is the relevance of the items' true value.

By including these two metrics in the evaluation of prediction models, we are capable of providing interesting insights concerning two issues: *i)* does the application of resampling strategies provide worst results concerning standard evaluation metrics such as $rmse$, and *ii)* what is the impact of considering the relevance in such standard metrics and how do they relate.

5.3.2 Ranking Evaluation Metrics

Given the importance of ranking order to news recommendation, we used the following metrics: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and

Normalized Discounted Cumulative Gain (NDCG@k) metrics. The first metric is focused on the global outcome, and the remaining metrics consider the position of a given item in the ranking.

Average Precision computes the average precision for all values of k where k is the rank, n is the number of retrieved items and Rel_k is a binary function evaluating the relevance of the k^{th} ranked item, attributing 1 to the relevant items[5] at rank k and 0 otherwise. The Mean Average Precision (MAP) metric computes the fraction of relevant documents retrieved over a set of queries.

Taking into account ranking position, the Reciprocal Rank is defined as the inverse of the rank at which the first relevant document is retrieved. As in MAP, the Mean Reciprocal Rank is defined as the average of the reciprocal ranks over all queries.

The Normalized Discounted Cumulative Gain (Eq. 6) measures the search result quality of the ranking function by assigning high weights to documents in highly ranked positions and reducing the ones found in lower ranks. Its definition is presented as follows, where $Rel_{i,q} \epsilon \{0, 1, 2, 3\}$ is the ad-hoc relevance judgment of the i^{th} ranked item for query q. The normalization of Discounted Cumulative Gain (Eq. 5) to a value between 0 and 1 is done by dividing the DCG value for the ideal ordering of the ranking ($idealDCG$).

$$DCG@k(q) = \sum_{i=1}^{k} \frac{2^{Rel_{i,q}} - 1}{log_2(1+i)} \quad (5) \qquad NDCG@k = \frac{\sum_{q=1}^{Q} \frac{DCG@k(q)}{idealDCG@k(q)}}{Q} \quad (6)$$

6 Experimental Evaluation

This section presents results on three sets of experiments. The first set evaluates the ability of models accurately predicting the news items focusing on those with a high number of tweets. The second set of results relates to the application of our proposed framework as a stand-alone system and its evaluation using daily news rankings. Finally, the third set is a real-world usage evaluation where our framework is applied to news recommendations from Google News. In all of these experiments the ground-truth is derived from the number of tweets assuming that our objective is high accuracy in recommending those news which have the most number of tweets. In the following experimental evaluations we will refer to the resampling strategies as **UNDER** for random under-sampling, **OVER** for random over-sampling, **SMOTEr** for SMOTEr and **IS** for importance sampling. Also, we will refer to the regression algorithms using the nomenclature presented in Table 3: **SVM**, **MARS** and **RF**.

6.1 Prediction Models Evaluation

In this section we evaluate the predictive accuracy of the proposed models focusing on the rare cases of highly popular news. Our objective is to prove the main hypothesis of this paper, which states that the combination of resampling strategies proposed in Section 4 and standard learning algorithms is capable of improving predictive accuracy in comparison to previous work.

[5] We have established that the relevant items are those which belong to the top 10.

6.1.1 Baseline

As baselines for this experiment, in addition to the models produced by the learning algorithms without the application of resampling strategies, we consider the approach proposed by Bandari et al. [3].

The authors report that the best results are obtained by resorting to support vector machines (**SVM**) to model the data, using six features: source and category density score, subjectivity score, number of named entities, highest and average score among named entities. The source, category and named entities scores report to popularity (*i.e.* number of tweets). As for the subjectivity score, the authors examine if an article written in a more subjective voice can resonate stronger with the readers. Originally, the authors used a subjectivity classifier from LingPipe[6]. However, due to the issues concerning its implementation in **R**, we chose to use the sentiment polarity score for both the title and headline instead.

6.1.2 Results

Our data has a temporal order. In this context, one needs to be careful in terms of the process used to obtain reliable estimates of the selected evaluation metrics. This means that the experimental methodology should make sure that the original order of the news is kept so that models are trained on past data and tested on future data to avoid over-optimistic estimates of their scores. In this context, we have used Monte Carlo simulation as the experimental methodology to obtain reliable estimates of the selected evaluation metrics for each of the alternative approaches. 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 training+test process repeated for each point. All alternative approaches are compared using the same training and test sets to ensure fair pairwise comparisons of the obtained estimates.

Our results cover four topics: *economy, microsoft, obama,* and *palestine*. They are obtained through 50 repetitions of a Monte Carlo estimation process with 50% of the cases used as training set and the subsequent 25% as test set. This process is done using the infrastructure provided by R package **performanceEstimation** [47].

Table 4 presents a summary of the estimated metric scores for root-mean-square error ($rmse$), the relevance weighted root-mean-square error ($rmse_\phi$), precision ($prec_\phi$), recall (rec_ϕ) and F1-Score ($F1_\phi$). For each regression algorithm the best estimated scores are denoted in italics, whilst the best overall score is denoted in bold.

As expected, the best overall results concerning the standard evaluation metric $rmse$ are provided by prediction models where resampling is not applied. Nonetheless, we note that when we account the best results by all combinations of regression algorithms and topics, they show that in **SVM** and **RF** models the application of resampling strategies does in fact improve the $rmse$ metric. By comparing this metric and $rmse_\phi$, the best overall results show the impact of accounting for relevance in the numeric prediction error: the best results considering the $rmse_\phi$ metric are given by models which apply resampling strategies. We observe that concerning this metric, the best models are obtained by the combinations of the regression algorithm **MARS** and the random under-sampling strategy, and the **RF** algorithm and the importance sampling strategy. Concerning the baseline approach of Bandari et al. [3], results

[6] Lingpipe 4.1.0: `http://aliasi.com/lingpipe`

Table 4 Evaluation of prediction models with estimated scores for Precision, Recall and F1-Score, as given by the utility-based framework proposed by Ribeiro [37].

	economy					microsoft				
Model	$rmse$	$rmse_\phi$	$prec_\phi$	rec_ϕ	$F1_\phi$	$rmse$	$rmse_\phi$	$prec_\phi$	rec_ϕ	$F1_\phi$
bandari	128.694	127.070	0.614	0.550	0.580	154.274	151.384	0.629	0.583	0.605
svm	127.675	125.669	**0.623**	0.562	0.591	156.008	149.339	0.624	0.604	0.614
svm_UNDER	*125.517*	119.451	0.612	0.636	0.624	154.745	*143.243*	0.634	0.659	0.646
svm_OVER	126.527	119.954	0.601	0.625	0.613	*152.888*	145.840	0.640	0.630	0.635
svm_SMOTEr	126.623	119.126	0.608	0.649	*0.627*	156.925	143.922	*0.649*	*0.688*	*0.668*
svm_IS	133.422	*118.920*	0.572	*0.660*	0.613	165.494	143.611	0.596	0.669	0.630
mars	**123.910**	120.212	*0.622*	0.619	0.620	**149.032**	144.829	0.643	0.622	0.632
mars_UNDER	124.324	**115.687**	0.595	0.687	0.637	149.752	**138.388**	0.650	0.704	0.676
mars_OVER	124.050	116.079	0.602	0.680	0.639	150.182	140.118	0.654	0.695	0.674
mars_SMOTEr	142.079	120.001	0.574	**0.725**	**0.641**	170.697	138.460	**0.660**	**0.755**	**0.704**
mars_IS	133.788	120.535	0.597	0.687	0.639	171.027	138.754	0.652	0.718	0.683
rf	127.622	124.818	*0.619*	0.593	0.606	159.378	156.289	0.625	0.596	0.610
rf_UNDER	130.607	120.020	0.596	0.680	0.635	163.742	145.727	0.620	0.684	0.650
rf_OVER	*127.546*	119.576	0.600	0.665	0.631	*156.673*	146.119	0.627	0.651	0.639
rf_SMOTEr	132.099	119.151	0.588	0.700	*0.639*	160.578	147.712	*0.647*	0.695	*0.670*
rf_IS	142.825	*117.262*	0.556	*0.717*	0.626	185.088	*143.055*	0.569	*0.703*	0.629

	obama					palestine				
bandari	372.092	369.108	**0.633**	0.495	0.554	**211.697**	208.263	*0.583*	0.556	0.569
svm	370.619	366.630	0.597	0.497	0.541	222.275	207.260	0.562	0.586	0.573
svm_UNDER	377.734	338.396	0.581	0.620	*0.600*	227.749	*197.957*	0.544	0.649	*0.591*
svm_OVER	*364.515*	353.231	*0.600*	0.543	0.570	*216.006*	204.306	*0.575*	0.593	0.584
svm_SMOTEr	381.349	359.576	0.558	0.533	0.545	227.095	208.155	0.561	0.600	0.579
svm_IS	419.075	*334.940*	0.545	*0.661*	0.598	272.634	202.992	0.518	*0.678*	0.587
mars	**363.858**	357.793	*0.620*	0.518	0.564	*216.254*	*199.250*	*0.585*	0.623	0.603
mars_UNDER	379.900	334.351	0.582	0.669	0.623	237.835	206.259	0.567	0.683	0.619
mars_OVER	371.113	337.256	0.592	0.638	0.614	228.469	205.422	0.578	0.662	0.617
mars_SMOTEr	388.684	332.380	0.569	*0.698*	0.627	228.385	203.289	0.580	*0.691*	**0.631**
mars_IS	407.291	*331.973*	0.576	0.691	**0.628**	244.019	205.922	0.573	0.679	0.621
rf	*366.943*	361.819	*0.620*	0.503	0.554	*211.833*	204.468	**0.587**	0.593	0.590
rf_UNDER	392.007	336.097	0.570	0.660	0.612	234.310	193.899	0.544	0.689	0.608
rf_OVER	371.310	343.441	0.587	0.591	0.589	224.278	196.157	0.554	0.661	0.602
rf_SMOTEr	407.427	333.189	0.556	0.690	*0.616*	234.629	194.250	0.549	0.695	*0.613*
rf_IS	444.574	**330.798**	0.536	**0.715**	0.613	272.426	**193.080**	0.519	**0.709**	0.599

show that the models produced by regression algorithms without the application of resampling strategies provide better results than the baseline in most cases.

Concerning the utility-based evaluation metrics ($prec_\phi$, rec_ϕ and $F1_\phi$), we observe that in many cases, the best results of precision ($prec_\phi$) for each of the regression algorithms is given by models where resampling is not applied. However, these models also show a distinct decrease in recall (rec_ϕ). Considering these observations, they show that these models often incur in a wrongful prediction of rare cases. If we focus on the $F1_\phi$ metric, which combines results from both the utility-based precision and recall metrics, this intuition is confirmed. The best results concerning the utility-based metric $F1_\phi$ in all topics and all regression algorithms is given by models where resampling strategies are applied. Also we observe that in all topics, the best overall results are given by the regression algorithm **MARS**. These results are in most cases obtained by combining this regression algorithm and the resampling strategy **SMOTEr**. As such, we observe that the results obtained show an advantage for resampling strategies where both under- and over-sampling are applied. In relation to the baseline approach by Bandari et al. [3], the comparison of results between this approach and the other models, given by the utility-based metric $F1_\phi$, shows that in most cases the baseline approach provides the worst results.

Overall, this experimental evaluation describes a considerable advantage for the use of resampling strategies in combination with standard regression algorithms when the objective is to improve predictive accuracy concerning rare cases of highly popular news. However, it is still not clear if the outcome represents a statistically significant performance improvement. Therefore, we resort to Wilcoxon signed rank tests in order to test the hypothesis that the performance of our proposed prediction models provide significant accuracy improvements, according to the utility-based metric $F1_\phi$, and to infer the statistical significance (with $p-\text{value} < 0.05$) of the paired

differences of the approaches' outcome. The statistical tests concerning regression algorithms without and with the application of resampling strategies show that the latter provides a significant performance improvement in every comparison. However, concerning the comparison between models and the resampling strategies used in this paper, results are more diverse. These are shown in Table 5, by aggregating the approaches outcome by regression algorithm.

Table 5 Results from Wilcoxon signed rank tests designed to test the hypothesis that the application of given resampling strategies (rows) provide significant increases of predictive ability in comparison to other strategies (columns), according to the evaluation metric $F1_\phi$, aggregated by regression algorithm. Ticks represent statistical significance and crosses the lack of significance.

	svm				mars				rf			
Strategy	UNDER	OVER	SMOTEr	IS	UNDER	OVER	SMOTEr	IS	UNDER	OVER	SMOTEr	IS
UNDER	•	✓	✓	✓	•	✓	×	×	•	✓	×	✓
OVER	×	•	×	×	×	•	×	×	×	•	×	×
SMOTEr	×	✓	•	×	✓	✓	•	✓	✓	✓	•	✓
IS	×	✓	✓	•	✓	✓	×	•	×	✓	×	•

According to the significance tests carried out, results show that the resampling strategy that provides the most significant advantage, independently of the regression algorithm used, is **SMOTEr**. However, results also show that overall, resampling strategies are very effective in improving the predictive accuracy of different models for the specific task of forecasting the number of tweets of highly popular news. These methods are able to overcome the difficulty of these news being infrequent. With $F1_\phi$ scores between 0.6 and 0.7, although not optimal, we have more confidence in using the predictions of these models for ranking news items. This is particularly important within our goal, which requires the ability to accurately identify the news that are more relevant for the users, in order to improve the performance of news recommender systems.

6.2 Standalone Usage Evaluation

The goal of this evaluation is to check how effective are the news recommendation rankings produced using our proposed framework. Therefore, we will compare the rankings produced by our method against the ground truth. The ground truth rank is based on the observed number of tweets of each news item within the period of two days following its publication timestamp. Using the same process as in the first experiment, we predict the popularity of news items in the test set Ts and derive daily rankings using their publish date for aggregation. The ground truth rank is given by the sorting of news per day according to their final number of tweets, in a decreasing order.

6.2.1 Baselines

The effectiveness of the proposed prediction models in generating news rankings is compared to two baseline strategies, in addition to the ranks derived by the approach proposed by Bandari et al. [3]:

– *Time*: news are ranked by time of publication with the most recent first;

– *Source*: news are ranked by the mean number of tweets obtained by news of their respective news source.

The first baseline is a simple strategy usually used by news aggregators to promote popular content. As for the second, although we did not find any reference to it in previous work, it is an acceptable hypothesis that there are news sources that gather more attention than others. Therefore, we introduce this baseline in order to assess if knowing the source of news is enough to provide optimal news recommendations.

6.2.2 Results

In this experiment, the temporal order of the data is also crucial. As such, we again use Monte Carlo simulation as the experimental methodology, where each alternative approach is compared by the same pair of training (Tr) and test (Ts) sets. As such, our results are obtained through 50 repetitions of a Monte Carlo estimation process with 50% of the cases used as training set and the subsequent 25% used as test set. Based on these comparisons we will calculate the metrics described in Section 5.3.2, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain ($NDCG$@5 and $NDCG$@10), for the gathered data from each four topics (*economy, microsoft, obama,* and *palestine*).

Table 6 Evaluation of the rankings derived by the proposed prediction models and baselines with estimated scores for MAP@10, MRR, NDCG@5 and NDCG@10

	economy				microsoft			
Model	MAP@10	MRR	NDCG@5	NDCG@10	MAP@10	MRR	NDCG@5	NDCG@10
Source	0.493	0.739	0.618	0.544	0.700	0.927	0.813	0.711
Time	0.078	0.195	0.105	0.114	0.085	0.211	0.119	0.122
Bandari	0.493	0.733	0.614	0.546	0.695	0.919	0.799	0.705
svm_UNDER	0.482	0.726	0.594	0.527	0.664	0.890	0.769	0.671
svm_OVER	0.459	0.718	0.570	0.501	0.660	0.901	0.769	0.670
svm_SMOTEr	0.466	0.715	0.577	0.512	0.648	0.878	0.751	0.655
svm_IS	0.418	0.675	0.520	0.452	0.585	0.846	0.699	0.594
mars_UNDER	0.497	0.743	**0.622**	**0.549**	**0.708**	**0.930**	**0.815**	**0.714**
mars_OVER	0.496	0.743	0.620	0.547	0.704	0.923	0.812	0.713
mars_SMOTEr	**0.499**	**0.749**	0.620	0.545	0.705	0.926	0.814	0.713
mars_IS	0.491	0.737	0.615	0.542	0.698	0.918	0.807	0.707
rf_UNDER	0.460	0.706	0.571	0.513	0.651	0.905	0.763	0.663
rf_OVER	0.461	0.702	0.575	0.515	0.661	0.917	0.773	0.672
rf_SMOTEr	0.436	0.690	0.548	0.489	0.653	0.904	0.765	0.667
rf_IS	0.416	0.683	0.518	0.448	0.573	0.853	0.679	0.567
	obama				palestine			
Source	0.276	0.521	0.334	0.339	0.678	0.880	0.785	0.715
Time	0.048	0.158	0.074	0.067	0.185	0.418	0.277	0.268
bandari	0.273	0.515	0.330	0.335	0.665	0.878	0.776	0.713
svm_UNDER	0.217	0.441	0.276	0.278	0.526	0.823	0.655	0.578
svm_OVER	0.225	0.459	0.283	0.289	0.579	0.832	0.693	0.624
svm_SMOTEr	0.132	0.326	0.179	0.175	0.491	0.746	0.602	0.544
svm_IS	0.078	0.214	0.115	0.120	0.288	0.571	0.402	0.365
mars_UNDER	**0.279**	0.524	0.337	0.341	0.679	0.888	0.791	0.724
mars_OVER	**0.279**	**0.525**	**0.338**	**0.342**	0.678	0.887	0.790	0.724
mars_SMOTEr	0.276	0.522	0.336	0.339	**0.680**	**0.889**	**0.792**	**0.725**
mars_IS	0.273	0.517	0.334	0.338	0.672	0.888	0.786	0.717
rf_UNDER	0.134	0.305	0.189	0.196	0.587	0.865	0.713	0.643
rf_OVER	0.160	0.348	0.216	0.228	0.594	0.872	0.722	0.651
rf_SMOTEr	0.123	0.295	0.177	0.181	0.607	0.876	0.731	0.661
rf_IS	0.097	0.245	0.135	0.139	0.464	0.773	0.593	0.524

Overall, the results of this experimental evaluation clearly shows the difficulty of producing news rankings based on the prediction of news popularity. This is to be expected since, unlike the previous experimental evaluation, the success of this task requires not only an accurate prediction of highly popular news, but also to accurately predict their correct order. As such, in the topic *obama* for example, results show a non-optimal performance, which illustrates the margin for improvement.

Concerning the regression algorithms used, results show that the **MARS** algorithm provides the best results, which complies with the outcome of the previous experimental evaluation. However, results show that in most cases, the best rankings are given by models using random under- (**UNDER**) and over-sampling (**OVER**) strategies, which is an important difference in comparison to the previous evaluation.

Considering the best overall results, no clear advantage is observed concerning the use of a single resampling strategy. Nonetheless, results show a clear advantage of models using resampling strategies over those where such strategies are not applied. This observation is confirmed concerning each of the ranking evaluation metrics used. Amongst the results obtained by the baselines, we observe that the results of both the Source and the proposal of Bandari et al. [3] provide very similar results to the overall best. We should also note that although the results of these two baselines are very similar, the Source approach shows a slight improvement in the majority of cases. This goes to show the predictive ability provided by knowing the media source of news with regard to this ranking task.

Summarizing, results show that our approach is capable of producing daily news rankings which obtain better results than the baselines proposed, in all the four topics evaluated. This evaluation also shows that it is possible to improve the accurate ranking of recent news item sets, thus reducing the latency period related to their recency, for a considerable amount of cases.

6.3 Real World Usage Evaluation

In this section we present an evaluation of a real world case scenario where our predicted rank (PR_i) is derived from the news items proposed by Google News at a given time (GN_i) against the ground truth. Similarly to the previous experiments, the ground truth is given by the set of news items proposed by Google News, ordered by their popularity in a decreasing fashion. Our evaluation is focused on the top 10 news items as our objective is to check the ability to predict the users' reading preference on highly tweeted news items.

For this experiment, the starting point of our method is the top 100 news obtained from Google News for time Q_i, GN_i. This set can be decomposed into GN_i^{old}, the news with a time-alive greater than two days, and GN_i^{new}, the news with a time-alive lesser than two days. We build our predicted rank (PR_i) by obtaining the number of tweets for all news in GN_i. For those belonging to GN_i^{old} we use the known number of tweets, whilst for those in GN_i^{new} we use our models to predict this number of tweets. The predicted rank PR_i is then compared against the ground truth that is obtained using the observed number of tweets of all news in time step i. A concrete illustration of this process is shown in Table 7 where news items belonging to GN_i^{old} are shown in italic.

Table 7 Prepared evaluation data example (top 10 of 100) for a given query

GroundTruthRank	RealNTweets	Pred.Rank	Pred.NTweets
1	474	49	127.45
2	*398*	***1***	*398*
3	*299*	***2***	*299*
4	283	38	136.21
5	278	52	125.52
6	271	**3**	298.59
7	270	**9**	244.09
8	245	74	102.24
9	198	29	157.89
10	179	67	115.33

Considering the example in the table above, the prediction models made some new news appear in the top 10: one in position 3 and another in 9, when according to the ground truth ranking they should be in positions 6 and 7, respectively. Considering the difficulty of this task, the predictive ability shown in the example is our main goal in this experiment.

The comparisons between the predicted and ground truth rank are carried out for all time steps $i \in Ts$ and in the end we calculate the ranking evaluation metrics. This process is repeated 50 times with Monte Carlo estimations having 50% of the cases used as training set (Tr) and the following 25% cases used as test set (Ts). In this experiment, unlike the previous ones, the cases report to each of the rankings proposed by Google News and not the individual news items. The resulting metric scores are presented in Table 8 for all the proposed approaches and the baselines used in the previous evaluation.

Table 8 Evaluation of the proposed models and baselines according to metrics MAP@10, MRR and $NDCG$@10, when applied to data on Google News rankings

	economy			microsoft		
Model	MAP@10	MRR	NDCG@10	MAP@10	MRR	NDCG@10
Source	0.550	0.744	0.595	0.690	0.857	0.719
Time	0.033	0.124	0.059	0.043	0.137	0.070
bandari	0.698	0.849	0.722	0.796	0.924	0.816
svm_UNDER	0.758	0.874	0.779	0.847	0.932	0.862
svm_OVER	0.750	0.871	0.773	0.843	0.931	0.857
svm_SMOTEr	0.755	0.872	0.776	0.845	0.932	0.859
svm_IS	0.743	0.866	0.759	0.847	0.932	0.862
mars_UNDER	0.767	0.878	**0.793**	**0.858**	**0.938**	**0.875**
mars_OVER	0.766	0.878	0.792	0.855	0.937	0.872
mars_SMOTEr	**0.768**	**0.880**	0.792	0.857	**0.938**	0.874
mars_IS	0.764	0.877	0.791	0.856	**0.938**	0.874
rf_UNDER	0.744	0.873	0.778	0.845	0.933	0.859
rf_OVER	0.749	**0.880**	0.778	0.843	0.933	0.857
rf_SMOTEr	0.740	0.874	0.772	0.840	0.931	0.856
rf_IS	0.742	0.874	0.756	0.842	0.933	0.848
	obama			palestine		
Source	0.422	0.649	0.506	0.525	0.730	0.586
Time	0.073	0.209	0.113	0.052	0.172	0.080
bandari	0.436	0.667	0.511	0.827	0.903	0.870
svm_UNDER	0.506	0.707	0.578	0.880	0.944	0.887
svm_OVER	0.500	0.700	0.576	0.873	0.935	0.890
svm_SMOTEr	0.428	0.642	0.491	0.863	0.929	0.871
svm_IS	0.398	0.601	0.464	0.813	0.901	0.819
mars_UNDER	0.538	0.726	**0.618**	0.910	**0.964**	0.910
mars_OVER	0.536	0.724	0.616	0.909	0.962	**0.912**
mars_SMOTEr	**0.539**	**0.727**	0.617	**0.911**	0.963	**0.912**
mars_IS	0.535	0.724	0.616	0.909	**0.964**	0.909
rf_UNDER	0.448	0.646	0.536	0.880	0.940	0.890
rf_OVER	0.458	0.652	0.552	0.875	0.931	0.892
rf_SMOTEr	0.447	0.657	0.522	0.880	0.939	0.890
rf_IS	0.415	0.628	0.468	0.855	0.936	0.855

The results of this final experiment confirm the trend shown in previous experiments as to the regression algorithm that shows the best overall results: again, the **MARS** algorithm provides the best evaluation. Focusing on the results provided by the $NDCG$@10 metric, we find that, similarly to the previous evaluation, but contrary to the evaluation of prediction models, results show that in most cases the best overall evaluation is obtained by applying the random under-sampling (**UNDER**) strategy with the **MARS** algorithm.

Nonetheless, results from this experiment provide interesting results concerning the comparison between "simple" (**UNDER**, **OVER**) and "complex" (**SMOTEr**, **IS**) resampling strategies. If we observe the results of the metric MAP@10, which essentially accounts for the fraction of relevant items in the top-10 of the ranking, we observe that in most cases the best overall results are provided by applying the **SMOTEr** resampling strategy. This is also true for the metric MRR, denoting how

effective is a ranking in positioning a relevant item in its top positions. However, we should note that the differences in evaluation between the best results employing these two strategies (**UNDER** and **SMOTEr**) is residual.

Summarizing, results show that our proposal, when applied to Google News recommendations, is capable of producing news rankings using the prediction of news items' popularity with fairly good results. Finally, our results show that our best combinations are able to outperform all of the baseline strategies, and that, unlike the previous experimental evaluation, the approach proposed by Bandari et al. [3] outperformed the baseline Source.

7 Discussion

In the experimental evaluations carried out in this paper, we applied an optimal parametrization search method. This search included both regression algorithms' parameters and the percentages of under and over-sampling in the resampling strategies. Results show that regression algorithms are very sensitive to their parameter settings. In the first experimental evaluation, where predictive accuracy of models is tested, results show that the models where resampling strategies are not applied, achieve considerable results. However, several of the issues motivating our contribution are confirmed by analysing such results with the utility-based metrics in our evaluation. The results obtained by models where resampling strategies are not applied confirm that standard evaluation metrics are not appropriate to evaluate predictive accuracy on rare cases. While the $rmse$ metric shows that these models obtain the best results, the altered version of such metric $rmse_\phi$, which weights the items' relevance, describes a different outcome, where the application of resampling strategies provides the best results.

The optimal parametrization of regression algorithms without the application of resampling strategies also provides considerable results concerning the utility-based metric precision ($prec_\phi$). This is related to the optimal parametrization method used: although we observe an increased ability to detect rare cases, they also greatly increase the number of cases where the models mistakenly predict a case as highly popular, as reported by the utility-based metric recall (rec_ϕ). Focusing on the utility-based metric F1-Score ($F1_\phi$), results show the advantage of using resampling strategies.

Despite the regression algorithm **MARS** in combination with the resampling strategy **SMOTEr** having obtained the best results in the prediction models evaluation, it did not produce the most accurate rankings. Concerning the evaluation of our proposed framework in a stand-alone and real-world scenarios, the best results were obtained by the combination of regression algorithm **MARS** and in most cases the resampling strategy random under-sampling (**UNDER**).

This raises an issue concerning the use of "simple" resampling strategies, such as **UNDER** and **OVER**, and more "complex" resampling strategies (**SMOTEr** and **IS**). The ability of such strategies to accurately predict highly popular news and the ability to correctly rank such predictions entail different conclusions. Although the more "complex" resampling strategies show some advantage concerning the accurate prediction of rare cases, the resampling strategies **UNDER** and **OVER** have shown a greater advantage as to their outcome when evaluating the news rankings produced. A careful observation of the evaluation in all three experimental evaluations provides valuable insights to explain such results. According to the ranking evaluation metric MAP@10, the best model is the combination of the **MARS** algorithm and the **SMOTEr** resampling strategy, in most cases. However, the combination with the

UNDER resampling strategy often provides the best results as to the $NDCG@10$ evaluation metric. The difference between these two metrics is two-fold: *i)* unlike the $NDCG$ metric, MAP considers the position of the items in the ranking equally, and *ii)* the $NDCG$ metric uses multi-level relevance to evaluate the items in the ranking while the MAP metric does so in a binary manner. Considering these differences between evaluation metrics, we conclude the following. The combination of the **MARS** regression algorithm and the **SMOTEr** resampling strategy is more capable of accurately predicting top-10 items and positioning them accordingly, in most cases. However, if one considers that items in a ranking are not only described as "relevant" and "non-relevant", the combination of **MARS** algorithm and the **UNDER** resampling strategy provides the best news rankings. In other words, the combination with the **SMOTEr** resampling strategy includes more items of very low relevance in its rankings than the combination with the **UNDER** resampling strategy.

In general, results from the three experiments show that our proposal of combining resampling strategies with standard learning algorithms are capable of significantly improving the predictive accuracy of highly popular news. Nonetheless, they also show that despite this apparent advantage in predictive terms, the ranks produced by the predictions of such approaches are not equally superior. Thus showing that the best combination of learning algorithm and resampling strategy in predictive terms, is not necessarily the best approach when evaluating their generated news rankings.

8 Conclusions

We have presented an approach to news recommendation that aims at predicting the reading interests of users of these services. This approach uses the number of times a news is tweeted as a proxy for its relevance for users and, in this context, tries to predict this number focusing on very recent news items (*i.e.* those which have not received any social feedback). Given that these predictions will be used to rank news, the models should focus on improving the prediction accuracy towards highly tweeted news, which are rare cases. These are the news that we wish to rapidly position in the top of the recommendations. This fact leads to one of the main contributions of our work: the study and comparison of modelling techniques using resampling strategies, that are able to improve accuracy in forecasting highly popular news items. The importance of this prediction is to minimise the referred latency and enable the treatment and recommendation of recent news at any time. We propose a framework for the translation of these predictions into rankings, as well as their evaluation. This particular instantiation of the framework uses two data sources: Google News and Twitter. Results from the evaluation process of both prediction models and rankings recommended by the proposed framework demonstrated that it is possible to successfully approach and tackle the issue of latency related to the recency of news items, producing more robust solutions that are capable of taking into consideration the users' reading interests. We compared our results to previous work, such as the approach proposed by Bandari et al. [3] and two baseline strategies for the derivation of ranks. Concerning the proposal of rankings, we evaluated our framework in two different scenarios of deployment: *1)* as a stand-alone system containing a news pool of recently published news, and *2)* being applied to the news recommendations of Google News. In both cases, successful solutions where found, which obtained good results. We observe that the best overall approach was obtained by combining the learning algorithm MARS and random under-sampling strategy. However, concerning the prediction task, the best results were obtained by combining the regression

algorithm **MARS** and the **SMOTEr** strategy. Concerning future work, it is our intention to broaden the basis of analysis in order to include information from multiple official and social media sources, in addition to study the impact of using the full text of news items. Also, we plan to carefully analyse the impact of using resampling strategies on the individual predictors as this could provide important insight on the correlation to the evolution of news popularity.

For reproducibility purposes, all code (written in **R**) and data necessary to replicate the results are available in the Web page `http://tinyurl.com/jucjj44`.

Acknowledgements 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. The work of N. Moniz is supported by a PhD scholarship of FCT (SFRH/BD/90180/2012). The work of P. Branco is supported by a PhD scholarship of FCT (PD/BD/105788/2014). The authors would like to thank the anonymous reviewers for their insightful comments and suggestions for improving the quality of this paper.

Appendices

A Appendix 1

The following list describes the parameters tested in each of the regression algorithms in order to optimize the results concerning the experimental evaluation (Section 6) carried out in this paper. For SVM models, we tested the parameters *cost* (c) and *gamma* (g); for MARS models, the parameters nk, *degree* (d) and *thresh* (th); and for Random Forest models parameters *mtry* (mt) and *ntree* (nt) were tested. We remind that these parameters correspond to the mentioned implementation of such models in **R**. The optimal parametrization is presented in Table 9.

- **svm**: $c \in \{10, 150, 300\}$, $g \in \{0.01, 0.001\}$;
- **mars**: $nk \in \{10, 17\}$, $d \in \{1, 2\}$, $th \in \{0.01, 0.001\}$;
- **rf**: $mt \in \{5, 7\}$, $nt \in \{500, 750, 1500\}$;

References

1. Ahmed M, Spagna S, Huici F, Niccolini S (2013) A peek into the future: Predicting the evolution of popularity in user generated content. In: Proc. of 6th ACM WSDM, ACM, New York, NY, USA, pp 607–616
2. Akima H, Gebhardt A (2015) akima: Interpolation of Irregularly and Regularly Spaced Data. URL `https://CRAN.R-project.org/package=akima`, r package version 0.5-12
3. Bandari R, Asur S, Huberman BA (2012) The pulse of news in social media: Forecasting popularity. CoRR abs/1202.0332
4. Berger J, Milkman KL (2012) What makes online content viral? Journal of Marketing Research 49(2):192–205
5. Branco P (2014) Re-sampling approaches for regression tasks under imbalanced domains. PhD thesis, Universidade do Porto
6. Branco P, Ribeiro RP, Torgo L (2016) UBL: an r package for utility-based learning. CoRR abs/1604.08079

Table 9 Optimal parametrization for regression algorithms and resampling strategies, for all topics.

	Economy	**Microsoft**
svm	c=10, g=0.001	c=150, g=0.001
svm_UNDER	c=10, g=0.001, u=0.2	c=10, g=0.001, u=0.1
svm_OVER	c=10, g=0.001, o=5	c=10, g=0.001, o=5
svm_SMOTEr	c=10, g=0.001, u=0.2, o=2	c=10, g=0.001, u=0.1, o=2
svm_IS	c=10, g=0.001, u=0.2, o=3	c=10, g=0.001, u=0.1, o=2
mars	nk=10, d=1, th=0.01	nk=10, d=1, th=0.01
mars_UNDER	nk=10, d=1, th=0.01, u=0.2	nk=10, d=1, th=0.01, u=0.05
mars_OVER	nk=10, d=1, th=0.01, o=3	nk=10, d=2, th=0.001, o=10
mars_SMOTEr	nk=10, d=2, th=0.01, u=0.05, o=10	nk=10, d=2, th=0.001, u=0.05, o=10
mars_IS	nk=10, d=2, th=0.01, u=0.2, o=2	nk=10, d=2, th=0.01, u=0.05, o=3
rf	mt=7, nt=1500	mt=7, nt=750
rf_UNDER	mt=5, nt=500, u=0.2	mt=7, nt=1500, u=0.1
rf_OVER	mt=7, nt=500, o=5	mt=7, nt=750, o=10
rf_SMOTEr	mt=5, nt=1500, u=0.2, o=2	mt=7, nt=1500, u=0.2, o=2
rf_IS	mt=5, nt=1500, u=0.8, o=10	mt=7, nt=1500, u=0.2, o=5

	Obama	**Palestine**
svm	c=10, g=0.001	c=150, g=0.001
svm_UNDER	c=10, g=0.001, u=0.05	c=10, g=0.001, u=0.1
svm_OVER	c=10, g=0.001, o=3	c=10, g=0.001, o=2
svm_SMOTEr	c=300, g=0.01, u=0.6, o=2	c=10, g=0.001, u=0.8, o=10
svm_IS	c=10, g=0.001, u=0.05, o=5	c=10, g=0.01, u=0.2, o=3
mars	nk=10, d=2, th=0.01	nk=10, d=1, th=0.001
mars_UNDER	nk=10, d=1, th=0.01, u=0.05	nk=10, d=2, th=0.01, u=0.05
mars_OVER	nk=10, d=1, th=0.01, o=10	nk=10, d=2, th=0.001, o=5
mars_SMOTEr	nk=10, d=1, th=0.01, u=0.05, o=2	nk=10, d=2, th=0.01, u=0.8, o=10
mars_IS	nk=10, d=1, th=0.01, u=0.1, o=3	nk=10, d=2, th=0.001, u=0.1, o=2
rf	mt=5, nt=750	mt=7, nt=500
rf_UNDER	mt=5, nt=750, u=0.05	mt=7, nt=1500, u=0.1
rf_OVER	mt=5, nt=500, o=10	mt=7, nt=500, o=10
rf_SMOTEr	mt=5, nt=1500, u=0.05, o=10	mt=7, nt=1500, u=0.2, o=5
rf_IS	mt=5, nt=1500, u=0.05, o=10	mt=7, nt=750, u=0.1, o=5

7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. JAIR 16:321–357
8. David E, Jon K (2010) Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York, NY, USA
9. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proc. of 23rd ICML, New York, NY, USA, pp 233–240
10. De Choudhury M, Counts S, Czerwinski M (2011) Identifying relevant social media content: leveraging information diversity and user cognition. In: Proc. of the 22nd ACM HT, ACM, New York, NY, USA, pp 161–170
11. De Francisci Morales G, Gionis A, Lucchese C (2012) From chatter to headlines: harnessing the real-time web for personalized news recommendation. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, ACM, ACM, New York, NY, USA
12. Dougherty RL, Edelman A, Hyman JM (1989) Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. Mathematics of Computation 52(186):471–494
13. Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'01, pp 973–978
14. Fawcett T, Provost F (1999) Activity monitoring: Noticing interesting changes in behavior. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '99, pp 53–62
15. Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in r. Journal of Statistical Software 5(25):1–54
16. Figueiredo F, Almeida JM, Gonçalves MA, Benevenuto F (2014) On the dynamics of social media popularity: A youtube case study. TOIT 14(4):24:1–24:23

17. Freund Y, Iyer R, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. J Mach Learn Res 4:933–969
18. Friedman JH (2000) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29:1189–1232
19. Gupta M, Gao J, Zhai C, Han J (2012) Predicting future popularity trend of events in microblogging platforms. ASIS&T 75th Annual Meeting
20. Gürsun G, Crovella M, Matta I (2011) Describing and forecasting video access patterns. In: Proc. of 2011 IEEE INFOCOM, pp 16–20
21. Hong L, Dom B, Gurumurthy S, Tsioutsiouliklis K (2011) A time-dependent topic model for multiple text streams. In: Proc. of the 17th ACM SIGKDD, ACM, KDD '11, pp 832–840
22. Hsieh C, Moghbel C, Fang J, Cho J (2013) Experts vs. the crowd: examining popular news prediction performance on twitter. In: Proc. of WWW 2013
23. Hu M, Liu B (2004) Mining opinion features in customer reviews. In: Proc. of 19th AAAI, AAAI Press, pp 755–760
24. Kaltenbrunner A, Gomez V, Lopez V (2007) Description and prediction of slashdot activity. In: Proc. of the 2007 LA-WEB, IEEE, pp 57–66
25. Lee JG, Moon S, Salamatian K (2012) Modeling and predicting the popularity of online contents with cox proportional hazard regression model. Neurocomputing 76(1):134 – 145
26. Lerman K, Ghosh R (2010) Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: Proc. of 4th ICWSM
27. Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22, URL `http://CRAN.R-project.org/doc/Rnews/`
28. McCreadie RMC, Macdonald C, Ounis I (2010) News article ranking: Leveraging the wisdom of bloggers. In: Adaptivity, Personalization and Fusion of Heterogeneous Information, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, RIAO '10, pp 40–48
29. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2012) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. URL `http://CRAN.R-project.org/package=e1071`, r package version 1.6-1
30. Milborrow S (2013) earth: Multivariate Adaptive Regression Spline Models. URL `http://CRAN.R-project.org/package=earth`, r package version 3.2-6
31. Moniz N, Torgo L, Rodrigues F (2014) Resampling approaches to improve news importance prediction. In: Advances in Intelligent Data Analysis XIII, Springer International Publishing, vol 8819, pp 215–226
32. Moniz N, Torgo L, Eirinaki M, Branco P (2016) Time-based ensembles for prediction of rare events in news streams. In: 2016 IEEE International Conference on Data Mining Workshop, ICDMW, (accepted)
33. Newman N, Fletcher R, Levy DAL, Nielsen RK (2016) Reuters institute digital news report 2016. Tech. rep., Reuters Institute for the Study of Journal (University of Oxford)
34. Osborne M, Dredze M (2014) Facebook, twitter and google plus for breaking news: Is there a winner? In: Proc. of the 8th ICWSM 2014
35. Petrovic S, Osborne M, McCreadie R, Macdonald C, Ounis I, Shrimpton L (2013) Can twitter replace newswire for breaking news? In: Proc. of 7th ICWSM, The AAAI Press
36. Pinto H, Almeida JM, Gonçalves MA (2013) Using early view patterns to predict the popularity of youtube videos. In: Proc. of 6th ACM WSDM, New York, NY, USA, pp 365–374

37. Ribeiro R (2011) Utility-based regression. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto
38. Shulman B, Sharma A, Cosley D (2016) Predictability of popularity: Gaps between prediction and understanding. In: Proc 10th ICWSM, pp 348–357
39. Simkin MV, Roychowdhury VP (2015) Why does attention to web articles fall with time? Journal of the Association for Information Science & Technology 66(9):1847–1856
40. Suh B, Hong L, Pirolli P, Chi EH (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, pp 177–184
41. Szabo G, Huberman BA (2010) Predicting the popularity of online content. Commun ACM 53(8):80–88
42. Tatar A, Leguay J, Antoniadis P, Limbourg A, de Amorim MD, Fdida S (2011) Predicting the popularity of online articles based on user comments. In: Proc. of the 2011 WIMS, pp 67:1–67:8
43. Tatar A, Antoniadis P, de Amorim MD, Fdida S (2012) Ranking news articles based on popularity prediction. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Washington, DC, USA, ASONAM '12, pp 106–110
44. Tatar A, de Amorim M, Fdida S, Antoniadis P (2014) A survey on predicting the popularity of web content. JISA 5(1):8
45. Tatar A, Antoniadis P, Amorim MDd, Fdida S (2014) From popularity prediction to ranking online news. SNAM 4(1):1–12
46. Torgo L (2010) Data Mining with R, learning with case studies. Chapman and Hall/CRC
47. Torgo L (2014) An infra-structure for performance estimation and experimental comparison of predictive models in r. CoRR abs/1412.0436
48. Torgo L, Ribeiro RP, Pfahringer B, Branco P (2013) Smote for regression. In: Correia L, Reis LP, Cascalho J (eds) EPIA, Springer, Lecture Notes in Computer Science, vol 8154, pp 378–389
49. Torgo L, Branco P, Ribeiro RP, Pfahringer B (2015) Resampling strategies for regression. Expert Systems 32(3):465–476
50. Tsagkias M, Weerkamp W, de Rijke M (2009) Predicting the volume of comments on online news stories. In: Proc. of 18th ACM CIKM, New York, NY, USA, pp 1765–1768
51. Tsagkias M, Weerkamp W, Rijke M (2010) News comments: Exploring, modeling, and online prediction. In: Proc. of 32nd ECIR, pp 191–203
52. Wu Q, Burges CJ, Svore KM, Gao J (2010) Adapting boosting for information retrieval measures. Inf Retr 13(3):254–270
53. Xu J, Li H (2007) Adarank: A boosting algorithm for information retrieval. In: Proc. of 30th ACM SIGIR, New York, NY, USA, pp 391–398
54. Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proc. of 4th ACM WSDM, New York, NY, USA, pp 177–186
55. Zaman T, Fox EB, Bradlow ET (2013) A Bayesian Approach for Predicting the Popularity of Tweets. DOI arXiv:1304.6777
56. Özgöbek O, Gulla JA, Erdur RC (2014) A survey on challenges and methods in news recommendation. Proc of 10th WEBIST