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

Nuno Moniz[1], Luís Torgo[1], Magdalini Eirinaki[2]

[1]LIAAD – INESC Tec; Sciences College, University of Porto
[2]Department of Computer Engineering, San Jose State University

3nd International Workshop on Data Science for Social Media and Risk
(SOMERIS @ ICDM'2016)

Introduction

- Thousands of news stories are read and shared through social media platforms every day.

- In news streams, once an item obtains a certain level of social feedback according to a given social media platform, its popularity becomes obvious.

- Most news remain relatively unpopular, and only a few may be considered as highly popular news.

- Predicting that a news item will be very popular as soon as it is published or in the first moments after, is very challenging.

Who benefits from this?

- In general, any entity that requires the identification of the most significant events in a fast and accurate manner in order to act.

- Examples include media advertising, content caching, movie revenue estimation, traffic management, macro-economic trends forecasting and early detection of catastrophes.

Previous Work

- The standard task of predicting the overall popularity of web content has been tackled by two types of approaches:

- Meta-data Based Models (a priori)

- Social Feedback Based Models (a posteriori)

Strengths/Weaknesses

- Meta-data Based Models (a priori)

✔ Does not depend on the popularity evolution of items;
✔ Allows for prediction when social feedback is unavailable or scarce;
✗ When social feedback becomes available, it does not update its predictions.

- Social Feedback Based Models (a posteriori)

✔ They are simple and fast (computationally);
✔ Does not depend on the description of news items;
✗ Requires social feedback to be available.

Problem Definition

- The main task consists of predicting the number of tweets news will receive, being highly accurate for highly popular news stories.

- This task can be modelled as a non-standard regression problem, where the target variable is the number of tweets each incoming news item is expected to have after a certain period of time.

- In this work we set this period to two days.

Prediction of Rare Values

- A common issue is the imbalanced distribution of the target variable

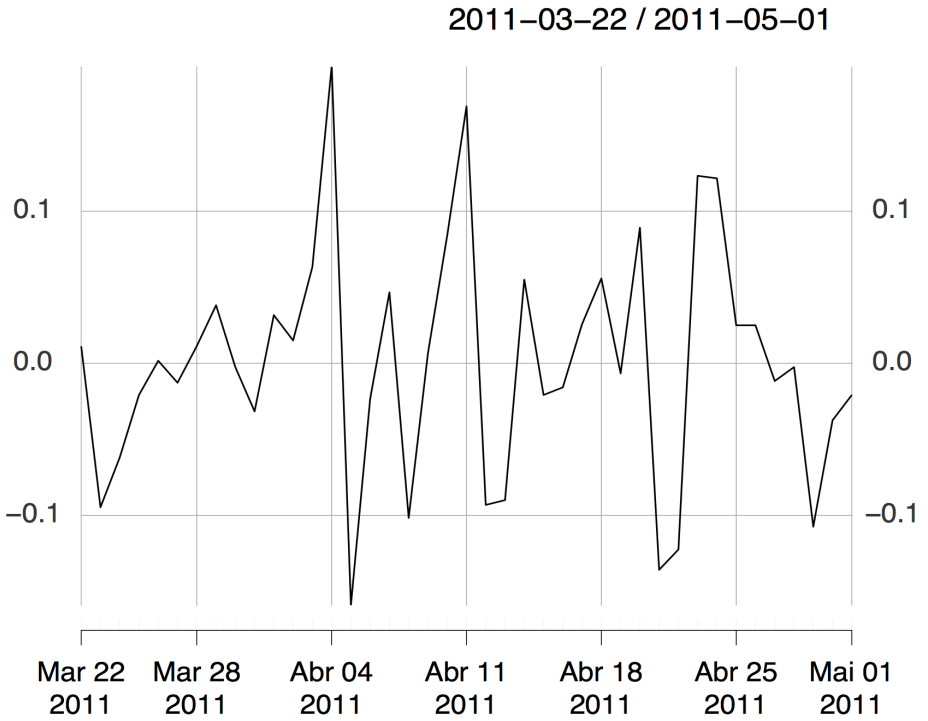
- Standard learning algorithms bias the models towards the most frequent situations, away from user preference biases.

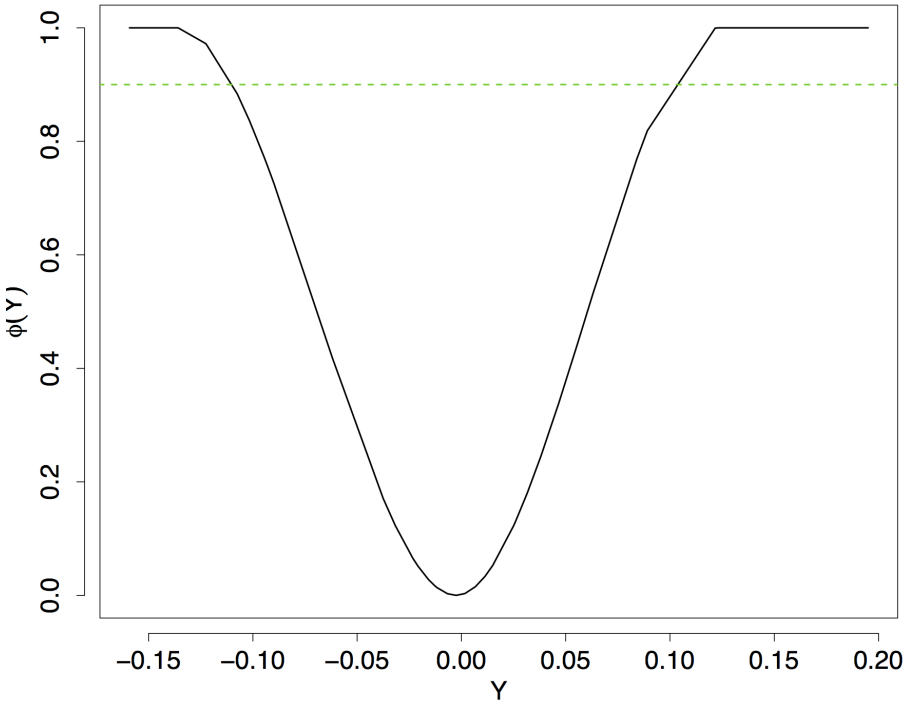
- Therefore, it is not clear if previous approaches to popularity prediction tasks are capable of predicting highly relevant news.

- The distribution of news items' popularity resembles a power-law distribution: the higher the popularity, the more relevant the case.

Relevance?

- Ribeiro [Ribeiro, 2011] proposes an approach to obtain a relevance function that maps the domain of continuous variables into a [0,1] scale of relevance: $\phi(Y)$: $Y \rightarrow [0,1]$





Our Proposal

- Time-Based Ensembles
- Combining a priori and a posteriori models in order to overcome their individual shortcomings and provide a more early and accurate prediction of the rare cases of highly popular news
- Models are combined using weighted averaging

Time-Based Ensembles

- Three assumptions that are the basis of our proposed hybrid strategy:

- 1) When social feedback is unavailable, only a priori models are able to predict news popularity;

- 2) When news items are recent, the available social feedback may be insufficient to confirm a priori predictions or to accurately predict popularity using a posteriori models;

- 3) As time passes since the publication of news, the available social feedback increases the accuracy of a posteriori predictions.

Models Ensemble

• The scarcity of social feedback is related to the recency of the events. Therefore, the alive-time t of news items is the main factor to consider when combining models of both strategies.

• A posteriori models' ability to accurately predict our target cases is related to the available data on the news items.

• We propose to relate the weights of each learner in an ensemble with the evolution of the mean proportion of the available data at a given time t, using train data to learn its evolution.

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

Proposed Approaches

- The first proposed approach applies weighted averaging to the numeric predictions of a priori models, and of a posteriori models, where weights are associated as previously described.

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

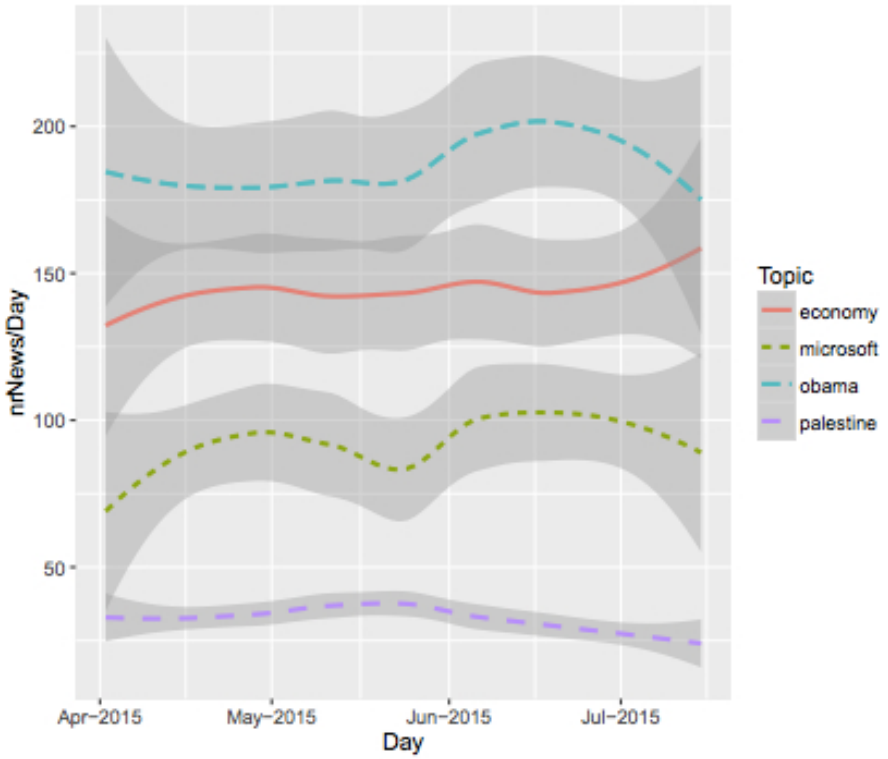
- The second proposed approach applies weighted averaging to the relevance of the numeric predictions of a priori models and a posteriori models. To obtain a numeric prediction of popularity, we use the inverse function of relevance.

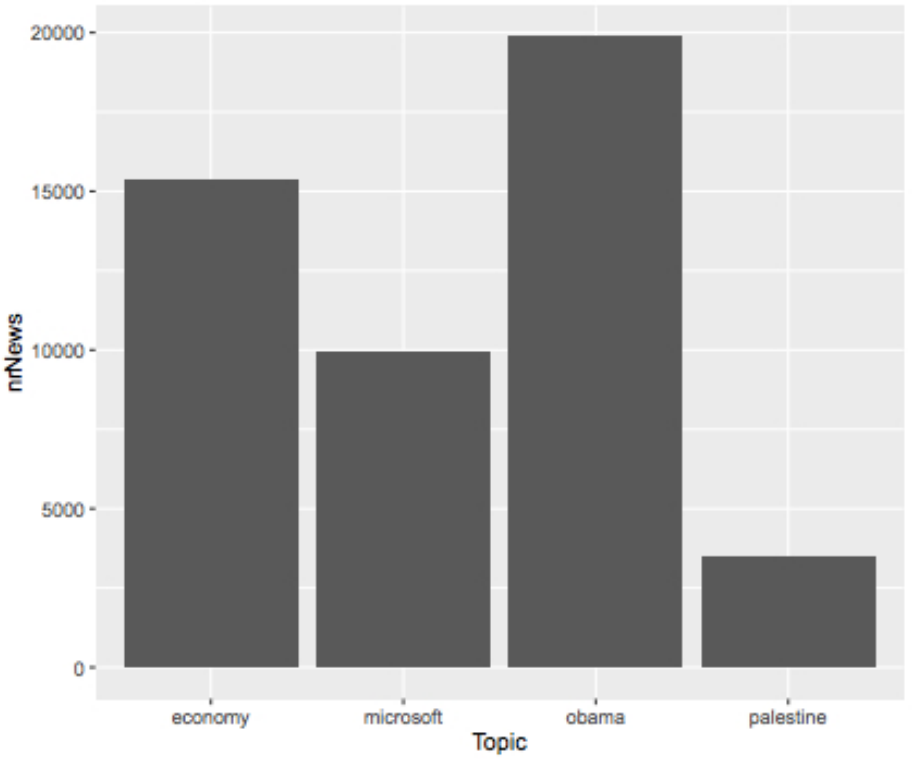
$$\hat{y} = \phi^{-1}(w_{po}^{t} \times \phi(\hat{y}_{po}) + w_{pr}^{t} \times \phi(\hat{y}_{pr})).$$

Experimental Evaluation

- 2 data sets: Google News and Twitter (~3 months and half)
- Monte Carlo Estimates (50% for training, 25% for testing)
- Significance tests: paired comparison using Wilcoxon signed rank tests ($p<0.01$)

Data


200
150
100
50
nrNews/Day
Apr−2015
May−2015
Jun−2015
Jul−2015
Day
Topic
economy
microsoft
obama
palestine


20000
15000
10000
5000
0
nrNews
economy
microsoft
obama
palestine
Topic

Methods

- A priori models

We use the work of Moniz et al.[1] which combines regression algorithms (Random Forest and SVM) with resampling strategies (under-sampling and SMOTEr).

- A posteriori models

•We use the work of Szabo et al. [2] which proposes a constant scale and a linear log approaches.

[1] N. Moniz, L. Torgo, and F. Rodrigues, “Resampling approaches to improve news importance prediction.” in IDA, 2014, pp. 215–226.
[2] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” Commun. ACM, vol. 53, no. 8, pp. 80–88, Aug. 2010.

Evaluation Metrics

- Standard evaluation metrics are not suitable to evaluate rare case prediction tasks (e.g. mse, mae, …)
- We resort to the utility-based framework proposed by Ribeiro [Ribeiro, 2011]
- This framework proposes several evaluation metrics designed for rare case prediction tasks
- We use F-Score, based on the definition of precision and recall for regression proposed by Ribeiro [Ribeiro, 2011]

$$precision = \frac{\sum_{\phi(\hat{y}_i)>t_R}(1+u(\hat{y}_i,y_i))}{\sum_{\phi(\hat{y}_i)>t_R}(1+\phi(\hat{y}_i))} \qquad recall = \frac{\sum_{\phi(y_i)>t_R}(1+u(\hat{y}_i,y_i))}{\sum_{\phi(y_i)>t_R}(1+\phi(y_i))}$$

Results

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

Improvement?

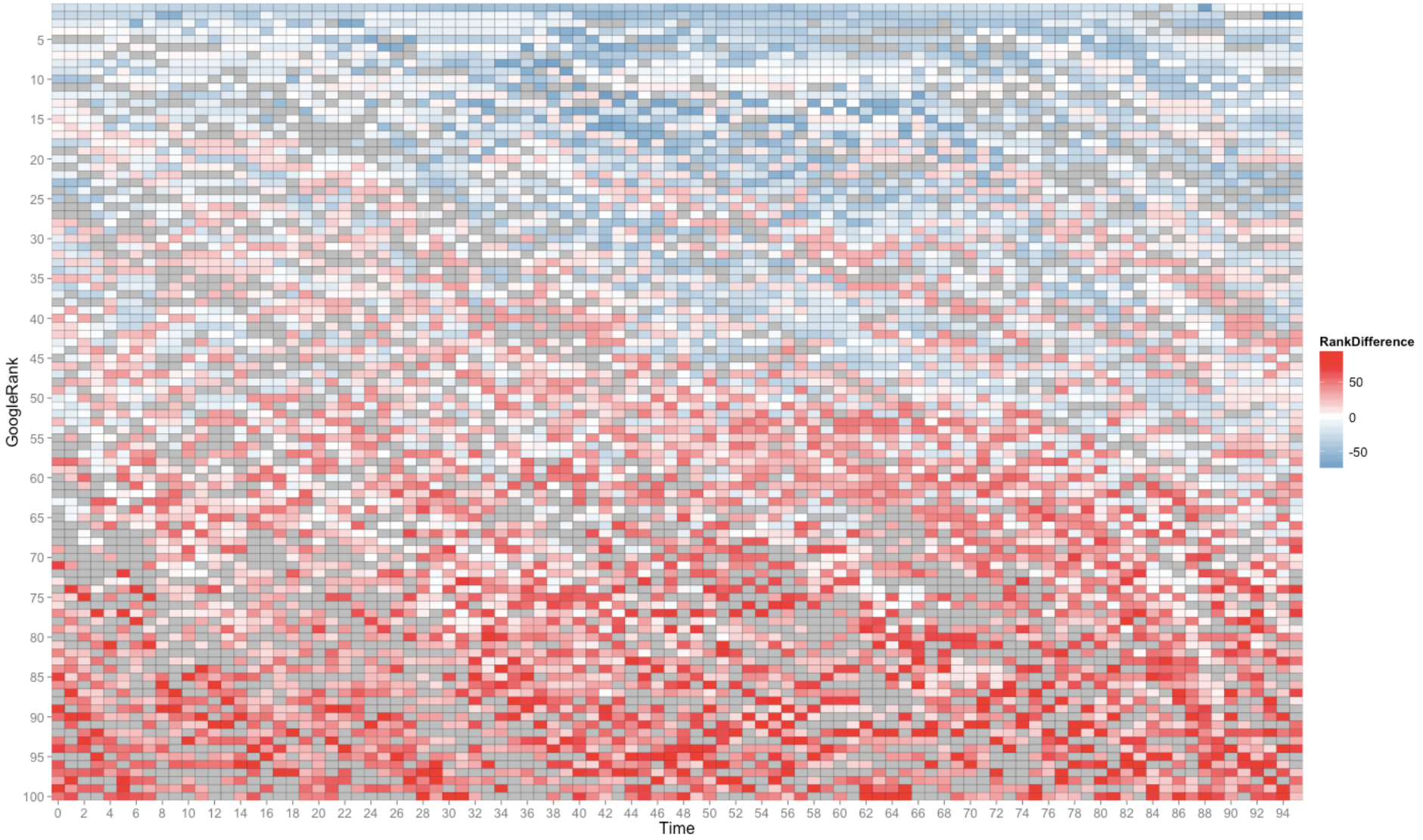
vs. a posteriori

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

vs. a priori

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

What about news aggregators?



Conclusions

- A new strategy for the early and accurate prediction of rare events in news streams is introduced;
- This strategy provides significant improvements over approach from both existing strategies;
- Results concerning the second proposed approach for time-based ensembles (time and relevance), when using the a posteriori model constant scaling, obtained the best overall results in all topics.

Thank you.

Nuno Moniz
nmmoniz@inescporto.pt

Luís Torgo
ltorgo@dcc.fc.up.pt

Magdalini Eirinaki
magdalini.eirinaki@sjsu.edu

Code + Presentation @ github:nunompmoniz