To Post or Not to Post: Using Online Trends to Predict Popularity of Offline Content

Sofiane Abbar
Qatar Computing Research Institute
Doha, Qatar
sabbar@qf.org.qa

Carlos Castillo
Universitat Pompeu Fabra
Barcelona, Catalunya, Spain
chato@acm.org

Antonio Sanfilippo
Qatar Environment & Energy Research Institute, HBKU
Doha, Qatar
asanfilippo@hbku.edu.qa

ABSTRACT

Predicting the popularity of online content has attracted much attention in the past few years. In news rooms, journalists and editors are keen to know, as soon as possible, the articles that will bring the most traffic into their website. In this paper, we propose a new approach for predicting the popularity of news articles before they go online. Our approach complements existing content-based methods, and is based on a number of observations regarding article similarity and topicality. We use time series forecasting to predict the number of visits an article will receive. Our experiments on real data collections demonstrate the effectiveness of the proposed method.

ACM Reference Format:
Sofiane Abbar, Carlos Castillo, and Antonio Sanfilippo. 2018. To Post or Not to Post: Using Online Trends to Predict Popularity of Offline Content. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3209542.3209575

1 INTRODUCTION

Monitoring the performance of news articles is a core task within any news media organization. The highly crowded news market, and the fast growth of online news platforms and applications in recent years, have pushed editors into a fierce competition for the attention of news readers. Currently, editors focus on *popularity* in terms of number of visits and visitors to news websites as the most important performance metric for news articles online. While social media has changed the way people consume news, their contribution in terms of visits to websites is relatively small. Andrew Miller, Guardian News and Media CEO, estimated all social media combined are around 10% of their site's traffic [20]

Measuring popularity, however, is not sufficient. The ability to *anticipate* online news popularity enables editorial teams to take strategic decisions to maximize the impact of their online content, such as promoting or demoting articles in their web pages. Given the high velocity of news, editors need to have forecasts for news articles as early as possible after publishing the article – and ideally, before publication.

The research community has addressed the problem of predicting the popularity of news articles in several recent papers including [3, 5, 11], typically addressing this as a regression problem (i.e. predict the popularity of an item), and sometimes as a classification problem (discretization of popularity into different levels/classes). Approaches that can dispense with early popularity measurements have been explored through the development of predictive models that use features such as the words in the title of the article, e.g. [8]. Our approach is complementary to such methods, and provides a novel extension where topic popularity forecasts are used to improve news article popularity predictions.

We explore two forecasting algorithms that exploit these observations, and test them on a large collection of news articles published by *Aljazeera Network*[1] – an international news organization – over 18 months in 2013 and 2014. The ensuing results yield a mean average percentage error (MAPE) as low as 11%, demonstrating the efficacy of the approach in predicting news article popularity.

2 RELATED WORK

We review in the following three categories of work related to our paper. *(i.) Methods Based on Early Measurements.* The success of the auto-correlation approach pioneered by [17] has encouraged many researchers to use early popularity measurements as predictors of future popularity. Examples include [7] for votes in Digg; [10] for comments to articles; [12] for visits to articles; [1] for views in YouTube and Vimeo, and for votes in Digg. *(ii.) Methods Based on Topics.* [3] used information about the category of a news article (e.g. sports, politics) together with information about the source, subjectivity, and named entities to predict the popularity of news articles in social media, prior to their publication. [18] predicted the number of comments to articles on a large news website. *(iii.) Methods Based on Keywords.* [19] studied the prediction of comments on news articles, using features such as publication date, number of articles posted at the same

HT '18, July 9–12, 2018, Baltimore, MD, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery.
ACM ISBN 978-1-4503-5427-1/18/07…$15.00
https://doi.org/10.1145/3209542.3209575

[1] http://www.aljazeera.com

time, and named entities. [8] measured the impact of titles on the popularity of image *re-posts* for different communities in *Reddit*. Our approach differs from and is complementary to the approaches reviewed in this section, in that it relies on article similarity and topic modeling. Our approach can also use recent historical data at any level of granularity (e.g. days, hours) to predict online content popularity.

3 DATASET

We use data provided by *Al Jazeera*, a large international news network operating multiple television channels and websites. We harvested articles from the English version of this website, which has millions of visits per month. The data covers a time span from September 2012 through April 2014. Our collection comprises two types of articles: *News* (8,065) and *Opinion* (4,357). The first category refers to breaking news. The latter refers to opinions and features contributed by named writers sharing their analysis of a topic of public interest. For each article, we also retrieved a time series of the number of visits the article gets after its publication.

4 PREDICTING TOPIC VOLUME

The first task we describe is the prediction of the total volume of visits to a topic u, i.e. the sum of the visits of all articles that have the topic u as the main topic. We use the Latent Dirichlet Allocation algorithm (LDA) to uncover the topics in our collection of articles [4].

4.1 Determining the Number of Topics

As many topic modeling methods such as non-negative matrix factorization [9] and Probabilistic Latent Semantic Analysis [6], LDA assumes the number of topics k given. Empirically, a small k leads to broad topics such as "politics" and "sports" whereas a large k leads to specialized topics such as US elections. We use supervised classification to find the "appropriate" number of topics k^*. The intuition is that k^* topics should yield a partition of the documents in the dataset that can be accurately recognized by a classifier trained on k^* classes of documents. First, we run LDA with different number of topics ($k \in \{10, \dots 100\}$). Let $\mathcal{T}^{(k)}$ be the topic set produced by LDA for each value of k. For each set of topics, we label every article $a \in \mathcal{D}$ with its primary topic $u_a^{(k)}$ such that $u_a^{(k)} = \text{argmax}_{u \in \mathcal{T}^{(k)}} \text{rel}(a, u)$. Next, we split articles into train (80%) and test (20%) sets, and train a Multinomial Naive Bayes classifier (MNB) [14] on the tf $\cdot$ idf scores of stems within each article. Results are reported in Figure 1. While the precision of the classifier is almost the same at $k = 10$ and $k = 20$, the recall and F_1 scores are maximized for $k = 20$. Hence, we set $k^* = 20$.

4.2 Topic Volume Prediction Results

We use Pearson's correlation (r^2) to measure auto-correlations and cross-correlations between topics, and Mean Absolute Percentage Error (MAPE) to evaluate forecasting results.

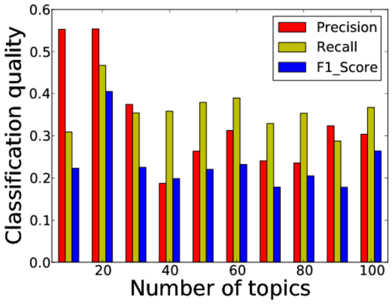


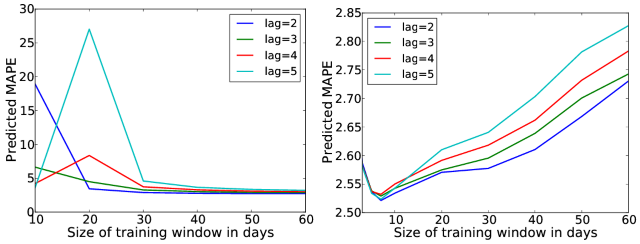
Figure 1: Quality of MNB for different numbers of topics obtained using LDA. F_1 maximizes at about 20 topics.

Determining the size of the training window. We now address the selection of the appropriate size of the time-window for training. A larger training window means more data is used for training, but *if the underlying process changes over time*, then incorporating training data that is too old may actually be counterproductive. The number of time lags δ to use is another important parameter. A larger δ means more variables are used for the prediction, which may lead to over-fitting. We train our prediction models with training windows of different sizes and different time lag values. We vary the sliding training window size to take values in $\{3, 5, 7, 10, 20, 30, 40, 50, 60\}$, and the lag $\delta \in \{2, 3, 4, 5\}$, both values expressed in days.

Figure 2 reports the average MAPE scores for different values of time lags and sizes of training sets. Each reported MAPE value is the average of scores achieved at predicting different steps-ahead (2, 3, 7, 15, and 30). Linear regression (LR) results are shown in Figure 2(a). A high variation of MAPE scores is observed for small sizes of the training set (≤ 30) before the scores stabilizes starting from training sets of size 50. SVR results are shown in Figure 2(b). MAPE scores are lower compared to those of LR, for all the values of the training set size we consider. It also shows that the ideal size of the training window is 7 days. Finally, adding more lags (larger δ) also increases the error rate. To summarize, the best prediction model is SVR with feature selection, a training window size of 7 days, and $\delta = 2$ or $\delta = 3$ as time lags.

5 ARTICLE PREDICTIONS

We now address the problem of predicting the number of visits to an *article*. Our objective is to assess to which extent topicality and article similarity can help predict the number of visits an article will receive. First we compute the popularity of an article as a function of the popularity of similar previously posted articles (Section 5.1). Then, we include topic popularity into the model (Section 5.2). Next, we integrate *predicted* topic popularity into the overall forecasting model (Section 5.3). Finally, we complement our prediction with early traffic observations to improve over both methods (Section 5.4).



(a) LR (b) SVR

Figure 2: MAPE of LR and SVR for # training sizes & lags.

5.1 Prediction Based on Article Similarity Using Nearest Neighbors (NN)

We hypothesize that similar articles posted within a relatively small time window receive a similar number of visits. The rationale behind this hypothesis is that people who visited an article about a developing story yesterday (or a few days ago), are likely to visit similar articles published today or at a later day. Sets of follow-up articles can be understood as playing the role of ephemeral pseudo-topics.

We measure article similarity by representing articles $\mathcal{D}$ using tf $\cdot$ idf vectors over the concatenation of their content and title. The similarity between each pair of articles is measured using cosine similarity $\mathrm{sim}_{\cos}(\cdot,\cdot) \in [0,1]$.

To predict article visits, we use these similarities as input to a nearest-neighbors estimation method (NN). This method consists on estimating the value of a function at given point, as an aggregate of the value of that function for a set of points near it [2, 15]. We use a variant of the kNN method applied to popularity prediction by [13], where the number of views of an item is the weighted sum of the number of views of similar items in the past few days.

Given an article a posted on day t_a, and a similarity threshold θ, we define $N^t_\theta(a)$ as the set of articles published on day t whose similarity with a is greater than or equal to θ:

$$N^t_\theta(a) = \{b \in \mathcal{D}, \mathrm{sim}_{\cos}(a,b) \geq \theta \wedge t_b = t\}\ . \qquad (1)$$

We next define a function which gives the weighted average of the number of visits to articles in $N^t_\theta(a)$ (for $t < t_a$) up to date t_a:

$$X_a(t) = \sum_{b \in N^t_\theta(a)} \frac{\mathrm{sim}_{\cos}(a,b) \cdot V_b(t_a)}{\sum_{b \in N^t_\theta(a)} V_b(t_a)} \qquad (2)$$

where $V_b(t_a)$ is the cumulative number of visits received by article b from its publication up to and including the publication date of a, t_a. Finally, our estimator is based on linear regression:

$$\widehat{V}_a(t_a + h) = \alpha_i + \sum_{i \in \delta(t_a)} \beta_i X_a(i) + \varepsilon \qquad (3)$$

where as before $\delta(t_a) = \{t_a - \delta, t_a - \delta + 1, \ldots, t_a - 1\}$ is the set of time lags under consideration, α and β_i are the linear regression coefficients, and ε is the residual term.

Results are shown on Figure 3(a). The model is trained on 80% of the articles, and tested on the remaining 20%. We vary δ from 1 to 7 days and set θ to values in $\{0.05, 0.1, 0.2, 0.3\}$. We observe that adding more days does not improve significantly the results. Values of θ close to 0.1 and 0.2 yield in general better results than 0.05 (which may cover too many articles distantly related to the one for which the prediction is being done) or 0.3 (which may be too strict as a criterion and include too few neighbors). We experimented with SVR and found the results to be no better than those obtained with linear regression (LR); in the remainder we report only the results with LR which is a simpler model.

5.2 Prediction Based on Topic Volume (NN+T)

Let us now consider a predictor of visits to article a based on the topic volume of its main topic u_a. This predictor is simply:

$$\widehat{V}_a(t_a + h) = \alpha_i + \sum_{i \in \delta(t_a)} \beta_i Y_{u_a}(i) + \varepsilon \qquad (4)$$

where $Y_{u_a}(i)$ is the number of visits to topic u_a at time i. The result is the dashed line in Figure 3(a). We observe its MAPE value is 1.33 percentage points lower than the one obtained with the method based on NN. Given that this method is complementary to the one using nearest neighbors, we can combine them using:

$$\widehat{V}_a(t_a+h) = \alpha_i + \sum_{i \in \delta(t_a)} \beta_i X_a(i, t_a) + \sum_{i \in \delta(t_a)} \gamma_i Y_{u_a}(i) + \varepsilon \quad (5)$$

where $X_a(i, t_a)$ is the aggregate of visits to nearest neighbors defined in Equation 2. Results are shown on Figure 3(b). We observe that the combined method is better than the method based only on topic volume for $\delta > 1$, and that in general the MAPE for $\delta = 3$ or $\delta = 4$ is lower than for $\delta = 1$.

5.3 Prediction Based on Predicted Topic Volume (NN+T+PT)

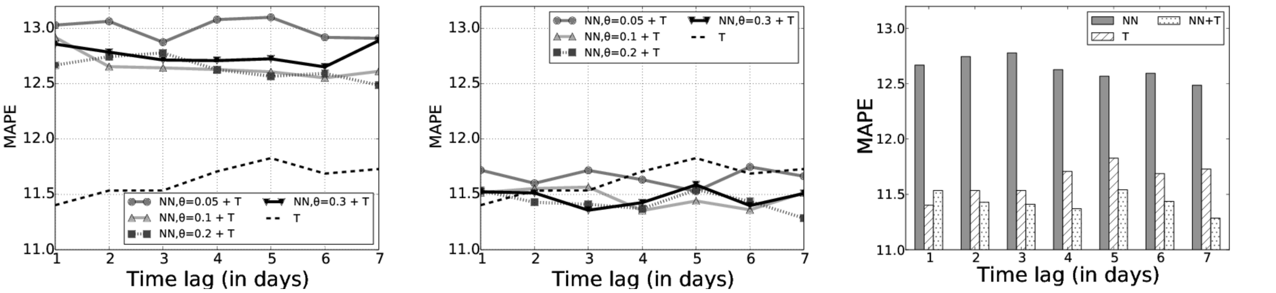
We further improve the results by creating an *ensemble* forecasting that operates in two steps. First, we predict the future popularity of a's topic u_a at time $t_a + h$, $\widehat{Y}_{u_a}(t_a + h)$ using the best estimator from Section 4.2. Next, we incorporate this as an input variable for the regression:

$$\begin{aligned} \widehat{V}_a(t_a + h) \quad = \quad & \alpha_i + \sum_{i \in \delta(t_a)} \beta_i X_a(i, t_a) + \sum_{i \in \delta(t_a)} \gamma_i Y_{u_a}(i) \\ & + \eta \widehat{Y}_{u_a}(t_a + h) + \varepsilon \end{aligned}$$

Results are shown on Figure 4. We observe a small but consistent improvement when incorporating this variable to our best predictor so far. Again, best results are observed using $\delta = 3$ or $\delta = 4$.

5.4 Incorporating Early Observations

Finally, we compare our method to the standard auto-regressive models based on early measurements (e.g. [13, 16, 17]). Recall that early-measurements methods rely on the existence



(a) Performance of nearest-neighbors method NN (continuous line) vs. topic-based method T (dashed line)

(b) Performance of combined NN+T method (continuous line) vs. topic-based method alone (dashed line)

(c) Comparison of NN, T, and NN+T with a fixed $\theta = 0.2$

Figure 3: Article popularity prediction using the nearest-neighbors method (NN), the topic-based method (T), and a combined method (NN+T). The first two plots vary $\theta \in \{0.05, 0.1, 0.2, 0.3\}$. The last plot fixes $\theta = 0.2$.

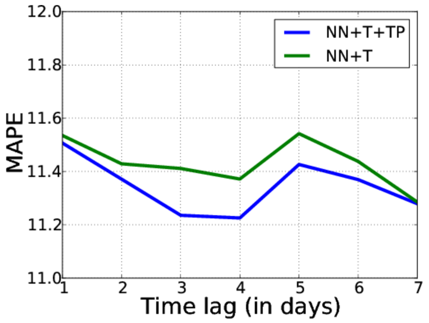


Figure 4: Predicting visits using nearest neighbors, observed topic volume, and predicted topic volume. θ is set to 0.1.

of a strong correlation between early and later times observed popularities, logarithmically transformed [17]. The forecasting formula of such a model is given below:

$$\widehat{V}_a(t_a + h_2) = \alpha + \beta V_a(t_a + h_1) + \varepsilon(h_2, h_1) \qquad (6)$$

Where t_a is the publication time of article a, $\widehat{V}_a(t_a + h_2)$ is the predicted future popularity of article a at time $t_a + h_2$, $V_a(t_a + h_1)$ is the popularity observed at time $t_a + h_1$, and $\varepsilon(h_2, h_1)$ is the noise term. Results of predicting the popularity of articles at three days ($h_2 = 3days$) at different early-measurements periods ($h_1 \in \{5min, 1h, 6h\}$) are shown on Figure 5. We observe that our method yields an error rate on the same scale as methods that use early observations. There is a smooth transition between the error rate resulting from our method (which can be used before publishing the article), and the error rate resulting from methods that use 5 minutes, 1 hour, or 6 hours of early observations. On average, our method yields a MAPE of 11.47%, while early predictions after 5 minutes, 1 hour and six hours obtain error rates of 9.59%, 6.83%, and 4.75% respectively. In the news domain, it is not realistic that an editor would publish a news article just to verify if it will have a large impact or not. Once a news is published, it can not be withdrawn without compromising reputation. Hence, our method provides a unique competitive advantage over the early-measurements-based methods.

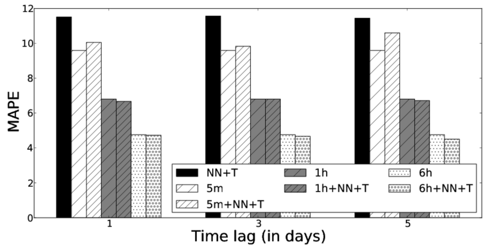


Figure 5: Visits prediction using our method, compared to methods using early measurements. The threshold θ is set to 0.1.

6 CONCLUSIONS

Predicting the popularity of an article before its date of publication requires combining content-based methods, which capture the article's communicative frame, with time series methods, which capture the evolution of people's attention around different issues. Our approach successfully combines two dimensions in the forecasting of visits for an article: the popularity of similar articles of recent issue, and the popularity of the topics that the article treats. More specifically, we have shown that an integration of these two dimensions rivals the performance of each dimension on its own. In future work, we will integrate information about sources, potential reach, as well as possible sources of competition for attention (e.g. similar articles on the same day), as a way of increasing the accuracy and robustness of the approach we have presented.

ACKNOWLEDGMENTS

C. Castillo is partially funded by La Caixa project LCF/PR/PR16/11110009.

REFERENCES

[1] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content. In *Proc. of WSDM*. ACM, Rome, Italy, 607–616. DOI:http://dx.doi.org/10.1145/2433396.2433473

[2] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. 1997. Locally Weighted Learning. *Artificial Intelligence Review* 11, 1-5 (1997), 11–73. http://citeseer.ist.psu.edu/atkeson96locally.html

[3] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity. In *Proc. of ICWSM*.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[5] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the Life Cycle of Online News Stories Using Social Media Reactions. In *Proc. of CSCW*. ACM, Baltimore, Maryland, USA, 211–223. DOI:http://dx.doi.org/10.1145/2531602.2531623

[6] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proc. of SIGIR*. 50–57.

[7] Salman Jamali and Huzefa Rangwala. 2009. Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. In *Proc. of WISM*. IEEE Computer Society, Washington, DC, USA, 32–38. DOI:http://dx.doi.org/10.1109/WISM.2009.15

[8] Himabindu Lakkaraju, Julian J. McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Proc. of ICWSM*. AAAI Press.

[9] Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for Nonnegative Matrix Factorization. In *In NIPS*. MIT Press, 556–562.

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

[11] Kristina Lerman and Tad Hogg. 2010. Using a Model of Social Dynamics to Predict Popularity of News. In *Proc. of WWW*. ACM, Raleigh, North Carolina, USA, 621–630. DOI:http://dx.doi.org/10.1145/1772690.1772754

[12] Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *Proc. of WWW*. ACM, Raleigh, North Carolina, USA, 621–630. DOI:http://dx.doi.org/10.1145/1772690.1772754

[13] Haitao Li, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, and Ke Xu. 2013. On popularity prediction of videos shared in online social networks. In *Proc. of CIKM*. ACM, 169–178.

[14] Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*. AAAI Press, 41–48.

[15] A. Navot, L. Shpigelman, N. Tishby, and Vaadia. 2006. Nearest neighbor based feature selection for regression and its application to neural activity.. In *Proc. of NIPS*.

[16] Matthew Rowe. 2011. Forecasting audience increase on YouTube. In *Workshop on User Profile Data on the Social Semantic Web*. Heraklion, Greece.

[17] Gabor Szabo and Bernardo A. Huberman. 2010. Predicting the Popularity of Online Content. *Commun. ACM* 53, 8 (Aug. 2010), 80–88. DOI:http://dx.doi.org/10.1145/1787234.1787254

[18] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo D. de Amorim, and Serge Fdida. 2011. Predicting the popularity of online articles based on user comments. In *Proc. of WIMS*. ACM, Sogndal, Norway. DOI:http://dx.doi.org/10.1145/1988688.1988766

[19] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2009. Predicting the volume of comments on online news stories. In *Proc. of CIKM*. ACM, 1765–1768.

[20] Twitter Blog. 2013. The Guardian Social Media Traffic. The Twitter blog. (2013). https://blog.twitter.com/2013/guardian-says-twitter-surpassing-other-social-media-for-breaking-news-traffic