On the Feasibility of Predicting Popular News at Cold Start

Ioannis Arapakis[*1], B. Barla Cambazoglu[†1], and Mounia Lalmas[‡2]

[1]Yahoo Labs, Barcelona, Spain
[2]Yahoo Labs, London, UK

March 21, 2016

Abstract

Prominent news sites in the Web provide hundreds of news articles daily. The abundance of news content competing to attract online attention, coupled with the manual effort involved in article selection, necessitates the timely prediction of future popularity of these news articles. The future popularity of a news article can be estimated using signals indicating the article's penetration in social media (e.g., number of tweets) in addition to traditional web analytics (e.g., number of page views). In practice, it is important to make such estimations as early as possible, preferably before the article is made available on the news site (i.e., at cold start). In this paper, we perform a study on cold-start news popularity prediction using a collection of 13,319 news articles obtained from Yahoo News, a major news provider. We characterise the popularity of news articles through a set of online metrics and try to predict their values across time using machine learning techniques on a large collection of features obtained from various sources. Our findings indicate that predicting news popularity at cold start is a difficult task, contrary to the findings of a prior work on the same topic. Most articles' popularity may not be accurately anticipated solely on the basis of content features, without having the early-stage popularity values.

1 Introduction

In recent years, news consumption has transcended the boundaries of the printed press into the area of online content distribution. However, as more web content becomes available, user attention stretches even thinner across the online space, making it possible only for a small percentage to receive significant traffic. Consequently, out of the many news articles published daily, very few reach a large audience and eventually go viral. In the mean time, major news providers constantly seek ways to attract more visitors and increase the visibility of their content, given that advertising revenues form a large share of their total income [Manduchi and Picard, 2009].

[*]Email: arapakis@yahoo-inc.com. Tel: +34 93 183 8879 (Corresponding author)
[†]Email: barla@yahoo-inc.com. Tel: +34 93 183 8830
[‡]Email: mounia@acm.org. Tel: +44 (077) 95644805

Traditionally, the published content is selected by editorial teams from a large pool of articles to maximise the coverage of important news articles. Given the manual effort involved, timely prediction of an article's future popularity is considered a high-value task, but also a challenging one. For example, if the number of times an article will be tweeted or liked can be estimated before it becomes available online, this information can be used to adjust the life span of the article, moving it to headlines, or even not publishing it at all. Ultimately, this can lead to better monetisation.

So far, some research effort has been made to address the problem of news popularity prediction, relying on early-stage measurements and user-generated content associated with the articles. The cold-start prediction scenario has been investigated, for the most part, in the context of recommender systems. To our knowledge, the only exception is the recent, widely cited work of Bandari et al. [Bandari et al., 2012], who investigate the problem using exclusively content-based features available at cold start. The performance results reported by the authors suggest that cold-start popularity prediction may be feasible.

Our work challenges the positive interpretation of the performance results reported in [Bandari et al., 2012]. To this end, we first try to reproduce the results in [Bandari et al., 2012] by following their experimental setting and methodology. We then expand their methodology and integrate more appropriate performance metrics in a step-by-step fashion. Our work introduces a large number of new features (including those used in [Bandari et al., 2012]) which may further help in predicting future article popularity. As the popularity metric, in addition to tweet counts (the only metric used in [Bandari et al., 2012]), we also use Facebook likes, shares, and comments, as well as view counts of article pages.

Although we could mostly reproduce the findings of [Bandari et al., 2012] and obtain similar results, our final findings, which are obtained after a rigorous evaluation and interpretation, indicate that predicting the popularity of news articles at cold start is not really a viable task with the existing techniques. We point at the high skewness in the popularity distribution as the source of the problem (i.e., large number of unpopular articles and very few popular articles). We show that the techniques are biased to predict the large class of unpopular articles more accurately than the small class of popular articles (a common phenomenon in machine learning). Therefore, popular articles, which are more important to detect early, cannot be predicted and surfaced to a large extent.

This paper is organised as follows. A survey of the related work is given in Section 2. In Section 3, we provide details about the data used in our work and our experimental setup. Section 4 presents some analyses on the characteristics of the investigated popularity metrics. In Section 5, we provide a detailed study of the features extracted from the data, aiming to reveal the correlations between these features and the popularity metrics. The extracted features are then used in a machine learning task for cold-start article popularity prediction, presented in Section 6. We conclude the paper in Section 7 with a discussion on our findings and possible extensions.

2 Related Work

Trend analysis is the practice of collecting and analysing time-varying data to identify the trends or to predict the future outcome of a target variable. A typical problem, which has been investigated in different contexts, is the prediction of an item's popularity over time. The related work on this problem include popularity prediction of multimedia content [Pinto et al., 2013, Shamma et al., 2011],

social marketing and stock market prediction [Yu et al., 2011, Zhang et al., 2011], election prediction [Tumasjan et al., 2010], impact prediction of research articles [Brody et al., 2006], topic volume prediction [Lehmann et al., 2012, Ruan et al., 2012], popularity prediction of microreviews [Vasconcelos et al., 2014a, Vasconcelos et al., 2014b], or early detection of popular online content in social media [Kim et al., 2011, Mathioudakis et al., 2010].

There is also a fairly large number of studies on popularity prediction in the context of online news [Ahmed et al., 2013, Freyne et al., 2010, Garimella and Castillo, 2014, Gupta et al., 2012, Jamali and Rangwala, 2009, Lerman and Hogg, 2010, Marujo et al., 2011, Szabo and Huberman, 2010, Tatar et al., 2011, Tsagkias et al., 2010]. Most of these work focused on prediction, exploiting early-stage popularity metrics (i.e., not cold-start prediction). Often, early measurements of a popularity metric (e.g., publication hour, page views, number of comments) were used to train a prediction model that will identify classes of temporal patterns and, eventually, forecast the future popularity of news articles.

In [Jamali and Rangwala, 2009], Jamali and Rangwala estimate the future popularity of online news based on the user comments left in response to the news articles, as well as other information derived from social network features. Similarly, in [Tatar et al., 2011, Tsagkias et al., 2010], the authors observe the volume of comments over a short period after an article's publication and use it as an indication of its importance, as well as a metric for forecasting online popularity. Freyne et al. [Freyne et al., 2010] demonstrate the merits of analysing social network events for personalised recommendations in a news feed. By harnessing user interactions and identifying patterns, they manage to deliver more accurate predictors of relevance. Garimella and Castillo [Garimella and Castillo, 2014] perform real-time traffic predictions of online news sources by considering social media features, as well as the entropy of the vocabulary of messages posted in Twitter. Gupta et al. [Gupta et al., 2012] propose an approach to predict the future popularity trend of news events on microblogging platforms (Twitter feeds).

Agarwal et al. [Agarwal et al., 2012] study the online actions (e.g., printing, commenting, rating, and sharing) that users perform when reading a news article to inform a ranking algorithm according to the probability that a user would take a post-read action on an article. Similarly, Szabo and Huberman [Szabo and Huberman, 2010] predict long-term popularity of online content based on early measurements of user access (e.g., views, votes, and downloads). The authors can forecast online popularity even 30 days ahead by observing the access to news articles during the first two hours after their publication. Lerman and Ghosh [Lerman and Hogg, 2010] propose a prediction model that can forecast the online popularity of news based on users' initial reactions. More specifically, their model provides an estimate of inherent story quality by considering early voting behaviour and predicts how many votes the story will receive after a number of days. In [Marujo et al., 2011], the authors investigate various approaches for automatically predicting the number of clicks a news article will receive within the first hour after its publication. The prediction is computed using a combination of popularity metrics, as well as content- and time-related features. Finally, in [Ahmed et al., 2013], a clustering technique is proposed to predict the future popularity of web content using two simple features based on the past popularity of the content.

The above studies have addressed the popularity prediction problem in the context of online news, by accounting for the early-stage activity and the user-generated content associated with the news articles. Although related to ours, none of these work investigates the popularity prediction problem in a cold-start scenario. More importantly, the exploited features are limited to the standard web analytics (e.g., page views, clicks, downloads) and popularity measurements (e.g., number of comments, votes, tweets), which are suitable only for an early-stage prediction task.

The feature set we use in our work is more diverse and is also more suitable for the cold-start prediction task. For example, we extract contextual information like article genre, news source, date of publication, and certain linguistic features. We also perform sentiment analysis and compute the sentimentality and polarity of news articles and their titles. Finally, we compute statistics regarding the mention of entities in the article, in Twitter, web search queries, and Wikipedia.

So far, some research efforts have addressed the cold-start prediction problem in the context of recommender systems. In [Liu et al., 2011], the authors present an approach to identify representative users and items using representative-based matrix factorisation. In [Levi et al., 2012], Levi et al. discuss a hotel recommender system that employs context-based features. The authors overcome the cold-start problem by mining contextual information and analysing it for common traits per context group. Emphasis is put on hotel reviews written with the same intent or by reviewers with a similar background. In [Givon and Lavrenko, 2009], the authors demonstrate how cold-start book recommendations based on social-tags can be combined with traditional collaborative filtering methods to improve performance. Finally, Quercia et al. [Quercia et al., 2010] address the problem of cold-start social event recommendation, using the home location of the mobile phone users and the social events they have attended in the past. Their findings indicate that the availability of large datasets is key to identify events that are not only popular among the residents of an area, but also of interest in a city. These examples highlight the importance of selecting representative samples in attaining high coverage and better performance rates. In a similar manner, our approach to the cold-start problem involves the use of an extensive feature set, which we construct using information from both the article content and external sources. In addition, we perform an exhaustive correlation analysis to determine the features that are more strongly correlated with our target variables and identify the best predictors.

To the best of our knowledge, the only work that has tackled the cold-start popularity prediction problem in the context of online news is the work of Bandari et al. [Bandari et al., 2012]. In their work, the authors use a measure of popularity based on the number of times a news article is shared on Twitter. They devise a machine learning framework using some basic features including news source, genre, subjectivity of the language, and entities in the articles. The performance results reported by the authors suggest that popularity prediction is possible using only the limited information available before a news article is published. In this work, we reproduce the experimental results of [Bandari et al., 2012] and demonstrate, using a more uniform dataset and a larger set of features, that predicting news popularity at cold-start is not a viable task with the existing techniques. Contrary to the findings of Bandari et al., we show that an article's popularity cannot be accurately estimated, solely on the basis of content features without incorporating any early-stage popularity information.

This paper significantly extends a preliminary version of this work published as a short paper [Arapakis et al., 2014]. We introduce three additional metrics that quantify the popularity of a news article. More specifically, we use the number of shares, likes, and comments obtained from Facebook as the new metrics in addition to the two metrics used in the previous work (number of tweets and page views). We repeat all previous experiments using these metrics and report new results. Moreover, we discuss our features in greater detail and present correlation analyses, not previously reported. The new observations are in line with the previous findings in that cold-start popularity prediction remains as a challenging task with the newly introduced metrics as well.

Table 1: The popularity metrics used in the study

Facebook shares (`Shares`): the number of times users have posted or shared the URL of a news article on Facebook.
Facebook likes (`Likes`): the number of times Facebook users have liked a news article or any comments about the article.
Facebook comments (`Comments`): the number of comments Facebook users have made on a shared article.
Tweets (`Tweets`): the number of times a news article was posted or shared on Twitter.
Page views (`Pageviews`): the number of times a news article page has been viewed by the users.

3 Data and Setup

Our analysis was conducted on a dataset consisting of 13,319 news articles taken from Yahoo News. We opted for a single news portal to be able to extract features that are consistent across all articles. The dataset was constructed by crawling news articles over a period of two weeks. During the crawling period, we connected to the RSS feed API of the news portal every 15 minutes and fetched newly published articles. Each article was identified by its unique URI and stored in a database, along with meta-data like genre (e.g., politics, sports, crime), publication date, and article's HTML content at the time of publication.

To quantify the online popularity of news articles, we opted for five different metrics (four sociometrics and one web analytic metric): `Shares`, `Likes`, `Comments`, `Tweets`, and `Pageviews` (please refer to Table 1 for their description). The choice of sociometrics was informed by the fact that, nowadays, an increasing number of users are interacting with social media applications and exchanging content. These online communities serve as conduits for information flow and can thereby help assess better the virality of online content. We also include page visits as a metric since it is commonly used as a proxy for website engagement and online content popularity.

In our setup, every request to the RSS feed API was followed by a request to the Facebook and Twitter public APIs to collect data about the popularity metrics. For all articles stored in the database, the metric values were sampled every half an hour, over a period of one week after the articles' publication. This resulted in 337 observations per article and per metric, a total of 18,164,300 observations. In addition, we collected information about the page views, also every half an hour, from the access logs of the news site.

4 Characterisation of Metrics

Figure 1 shows the variation of `Shares`, `Likes`, `Comments`, and `Tweets` over time (the values are averages over all articles). We report both the original values (inner plot) and normalized values (outer plot). Regarding the accumulative behaviour of the popularity metrics, we can make the following observations: i) `Tweets` exhibit the highest rate of change and precede all other metrics, ii) `Shares` appear to precede `Likes` in the early stages; the intersection point indicates the point in time (+9 hours after publication) beyond which `Likes` exceed `Shares`, and iii) the popularity metric that reaches the highest value after saturation is `Likes`.

Figure 2 shows the distribution of `Shares`, `Likes`, `Comments`, and `Tweets` in decreasing order of

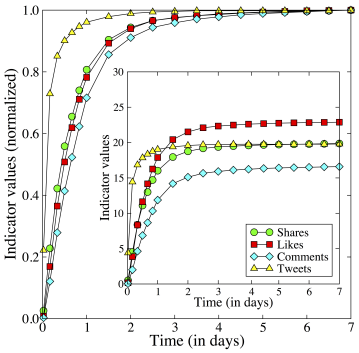


Figure 1: Variation of metrics over time.

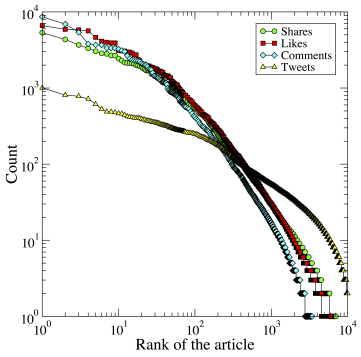


Figure 2: Distribution of metrics in decreasing rank order.

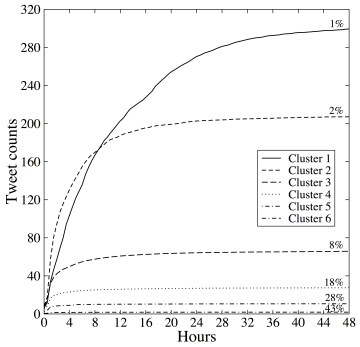


Figure 3: Tweet counts in different clusters of articles.

observed values. As expected, all distributions are heavily skewed, i.e., most articles attract very little attention while very few articles receive a lot of interest. Consequently, as we will see in Section 6, the problem of identifying soon-to-be-popular news articles becomes quite challenging.[1]

To characterise news articles based on the popularity metrics, we applied a clustering analysis. For each metric, we used the 337 available values and also computed the rate of change at each sample point. Therefore, each article is represented as a feature vector of $2 \times 4 \times 337 = 2,696$ dimensions. Given all feature vectors, we clustered the news articles using the SimpleKMeans algorithm with the Manhattan distance metric. As the number of clusters, we evaluated values ranging from $k = 1, \ldots, 10$. We applied the elbow method [Thorndike, 1953] to determine the optimal number of clusters, using the ratio and difference of the intra- and inter-cluster distances. In our case, the optimal choice was $k=6$ for both measures.

Figure 3 shows the average number of `Tweets` across the articles within each cluster.[2] Based on the volume of `Tweets`, the distribution of the articles across the six clusters is as follows: cluster 1: 72 (1%), cluster 2: 234 (2%), cluster 3: 1,128 (8%), cluster 4: 2,453 (18%), cluster 5: 3,789 (28%), and cluster 6: 5,799 (43%). For all metrics, clusters 1 and 2 consist of articles that were associated with very high counts starting from the first few hours following the article's publication and were regarded as highly popular. The news articles appearing in clusters 3, 4, 5, and 6 were the remaining majority that received little or no attention, further demonstrating the challenge in predicting news popularity beyond a binary decision.

To determine if and to what extent the popularity metrics are related, we run a correlation analysis on their hourly, daily, and weekly volumes. Given the non-normal distribution of the data, we opt for the Spearman's rho non-parametric test. Table 2 presents several statistically significant, positive correlations. When we examine the shared variation (r^2), we observe that it is likely to be lower in case of the hourly volumes (in the $[0.006\%, 0.371\%]$ range). However, it is considerably higher in case of the daily volumes (in the $[0.046\%, 0.549\%]$ range) and the weekly volumes (in the $[0.113\%, 0.596\%]$ range). In practice, because certain popularity metrics originate from the same social media domain (e.g., Facebook), one could potentially use only a representative subset of the metrics and omit the rest. For example, weekly `Shares` appear to be highly correlated with weekly

[1] The `Pageviews` metric is omitted in Figures 1–3 because of its confidential nature.

[2] The plots for the other metrics were omitted because they exhibited very similar patterns.

Table 2: Correlation matrix of popularity metrics

	Shares	Likes	Comments	Tweets	Pageviews
	Hourly volumes				
Shares	1.000	0.334*	0.273*	0.608*	0.297*
Likes		1.000	0.405*	0.259*	0.164*
Comments			1.000	0.145*	0.077*
Tweets				1.000	0.285*
Pageviews					1.000
	Daily volumes				
Shares	1.000	0.741*	0.682*	0.684*	0.376*
Likes		1.000	0.730*	0.533*	0.275*
Comments			1.000	0.430*	0.214*
Tweets				1.000	0.297*
Pageviews					1.000
	Weekly volumes				
Shares	1.000	0.772*	0.714*	0.690*	0.505*
Likes		1.000	0.759*	0.553*	0.419*
Comments			1.000	0.454*	0.337*
Tweets				1.000	0.415*
Pageviews					1.000

*Correlation is significant at the 0.01 level (2-tailed).

`Likes` ($r = 0.772, p < 0.01$) and account for 0.59% of the variation in `Likes`. To put this number into perspective, it leaves approximately 40% of the variation still to be accounted for by other factors. We nonetheless keep all metrics in our study.

Finally, the `Pageviews` metric appears to be the least correlated metric. We speculate that this is due to the different information domain (page access logs), which is largely disconnected from that of Facebook or Twitter. Another explanation is that it is not always the biggest news that penetrate the social media. More niche, topical, or negative news have been shown to have higher potential to become viral [Phelan et al., 2009, Thelwall, 2006, Wu et al., 2011]. In addition, `Pageviews` appears to be consistently more correlated with `Shares` and `Tweets` in hourly, daily, and weekly volumes, compared to the remaining popularity metrics. This suggests that a news article which had larger `Shares` than `Comments` is more likely to have a higher rate of page access or the other way round, given that correlation does not indicate causation.

5 Features

We use a larger number of features that we extract from the content of the news articles as well as external sources. Table 3 shows the list of features, which are categorised under ten main headings depending on how or where the feature is obtained from. In addition, we perform a correlation analysis to identify potentially significant correlations between our features and the popularity metrics to obtain insights for the predictive power of each feature.

Table 3 shows the correlation coefficients for different types of features (interval, ordinal, nominal, and binary) and popularity metrics, after one hour and one week following the articles' first publication time. For the interval and ordinal features we compute Spearman's rank correlation coefficient (r_s), given that our data violates the normality assumption. For the binary features, we

Table 3: The features used by the prediction model and their correlations with the predicted values, i.e., the volumes of the five popularity metrics one hour (`Hour`) and one week (`Week`) after the articles are published

Features			Popularity Metrics									
			Shares		Likes		Comments		Tweets		Pageviews	
Type	Name	Values	Hour	Week	Hour	Week	Hour	Week	Hour	Week	Hour	Week
Time	HourOfTheDay	Nominal	.047	.048	.044	.047	.038	.045	.117‡	.115‡	.075‡	.057‡
	AM/PM	Binary	.003	.009	.008	.004	−.001	.003	−.062‡	−.064‡	.035‡	.002
	DayOfTheWeek	Nominal	.024	.025	.020	.028	.018	.020	.081‡	.084‡	.023	.050‡
News source	NewsSource	Nominal	.107‡	.159‡	.069†	.135‡	.052	.108‡	.272‡	.207‡	.126‡	.166‡
Length	#OfCharacters	Ordinal	.064‡	.125‡	.030‡	.123‡	.007	.092‡	.060‡	.090‡	.022‡	.111‡
	#OfWords	Ordinal	.054‡	.113‡	.027‡	.116‡	.006	.088‡	.050‡	.081‡	.015	.105‡
	#OfSentences	Ordinal	.068‡	.148‡	.047‡	.154‡	.020†	.116‡	.059‡	.104‡	.025‡	.133‡
NLP	FracOfNouns	Interval	−.023‡	−.078‡	−.012	−.082‡	.003	−.078‡	−.031‡	−.061‡	−.023‡	−.075‡
	FracOfAdjectives	Interval	.103‡	.096‡	.000	.051‡	−.031‡	.016	.146‡	.139‡	.063‡	.077‡
	FracOfAdverbs	Interval	−.020†	.067‡	−.002	.077‡	−.002	.070‡	−.028‡	.043‡	−.022†	.084‡
	FracOfVerbs	Interval	.113‡	.195‡	.103‡	.200‡	.065‡	.167‡	.089‡	.121‡	.093‡	.151‡
	FractOfOthers	Interval	−.081‡	−.088‡	−.043‡	−.065‡	−.027‡	−.033‡	−.059‡	−.068‡	−.032‡	−.055‡
Genre	IsTopStory	Binary	.127‡	.091‡	.059‡	.078‡	.016	.058‡	.252‡	.264‡	.141‡	.119‡
	isLaw	Binary	.010	.002	.001	−.003	−.001	−.005	−.007	.005	.038‡	.025‡
	isSports	Binary	−.091‡	−.071‡	−.035‡	−.063‡	−.016	−.050‡	−.214‡	−.184‡	−.119‡	−.111‡
	isEconomy	Binary	−.007	−.023‡	−.014	−.020†	−.008	−.014	.021†	−.017	−.018†	−.032‡
	isHealth	Binary	.049‡	.012	.012	.008	.009	.005	.192‡	.115‡	.029‡	.015
	isEducation	Binary	−.006	−.006	−.005	−.007	−.003	−.004	.022†	.020†	−.009	.002
	isTechnology	Binary	−.017†	−.020†	−.019†	−.017	−.009	−.021†	.127‡	.080‡	−.021†	−.007
	isBusiness	Binary	−.035‡	−.033‡	−.016	−.034‡	−.007	−.025‡	−.006	−.024‡	−.024‡	−.032‡
	isEntertainment	Binary	.072‡	.012	.024‡	.011	.006	.005	.237‡	.141‡	.014	.045‡
	isScience	Binary	.022†	.136‡	.019†	.117‡	.002	.083‡	−.007	.061‡	.059‡	.144‡
	isPolitics	Binary	.052‡	.037‡	.036‡	.044‡	.027‡	.049‡	.063‡	.066‡	.129‡	.042‡
	isComputers	Binary	.012	−.009	−.004	−.015	−.003	−.012	.127‡	.087‡	.006	.005
	isLife	Binary	.023‡	.022†	−.008	.018†	−.004	.009	.051‡	.037‡	−.000	.063‡
	isRegional	Binary	.022†	.010	−.007	.005	−.007	.003	.030‡	.036‡	.055‡	.031‡
	isWorld	Binary	.094‡	.032‡	.002	.022†	−.003	.014	.225‡	.188‡	.126‡	.061‡
Sentiment analysis	Sentimentality	Ordinal	.103‡	.168‡	.055‡	.164‡	.022†	.120‡	.081‡	.113‡	.052‡	.148‡
	Polarity	Ordinal	−.160‡	−.213‡	−.116‡	−.208‡	−.054‡	−.158‡	−.156‡	−.163‡	−.153‡	−.174‡
Entity extraction	#OfLocEntities	Ordinal	.019†	.011	.009	.015	−.021†	−.001	.001	−.022†	.066‡	.025‡
	#OfOrgEntities	Ordinal	.033‡	.062‡	.016	.068‡	.011	.046‡	.051‡	.062‡	−.012	−.022‡
	#OfPerEntities	Ordinal	−.032‡	.006	.003	.038‡	.020†	.045‡	−.046‡	−.007	−.058‡	−.025‡
	#OfAllEntities	Ordinal	.021†	.041‡	.016	.055‡	.005	.041‡	.012	.021†	−.005	−.021†
Wikipedia	WikiPop	Interval	.025‡	.042‡	.031‡	.054‡	.007	.045‡	.021†	.023‡	.028‡	.055‡
Twitter logs	TwitterPop-H	Interval	.014	.051‡	.058‡	.072‡	.036‡	.071‡	.015	.020†	.085‡	.113‡
	TwitterPop-D	Interval	.018†	.065‡	.046‡	.086‡	.022‡	.075‡	.019†	.030‡	.081‡	.117‡
	TwitterPop-W	Interval	.026‡	.080‡	.052‡	.104‡	.024‡	.088‡	.026‡	.042‡	.058‡	.109‡
Search logs	SearchPop-H	Interval	.047‡	.130‡	.087‡	.154‡	.058‡	.135‡	.061‡	.095‡	.041‡	.105‡
	SearchPop-D	Interval	.040‡	.127‡	.075‡	.154‡	.047‡	.132‡	.059‡	.097‡	.015	.083‡
	SearchPop-W	Interval	.034‡	.123‡	.074‡	.160‡	.045‡	.136‡	.063‡	.102‡	−.054‡	.026‡

†Correlation is significant at the 0.05 level (2-tailed). ‡Correlation is significant at the 0.01 level (2-tailed).

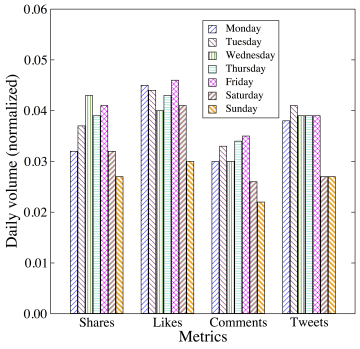


Figure 4: Daily volume distribution for different metrics.

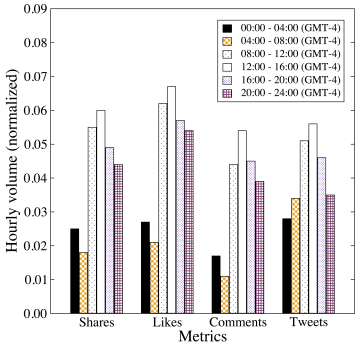


Figure 5: Hourly volume distribution for different metrics.

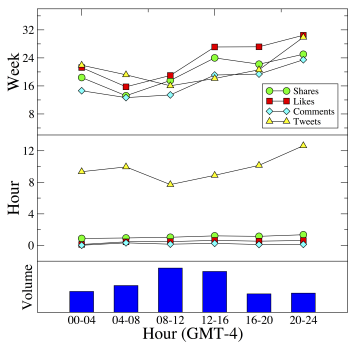


Figure 6: Impact of publication time on the metrics.

compute the point-biserial correlation coefficient (r_{pb}). In the case of r_{pb}, the sign of the correlation depends on the way the coding of the variables was made and we, therefore, ignore all information about direction. Finally, for the nominal features, we compute eta (η), which is the correlation coefficient of non-parametric variables with a non-linear relationship. Given that the predictor variable is nominal, negative correlations are not applicable. Hence, η varies between 0.0 and 1.0.

In Table 3, we observe several statistically significant correlations between our features and the popularity metrics. However, if we consider Cohen's conventions for the interpretation of effect size, we conclude that all reported correlation coefficients represent weak or small associations. This leaves a large percentage of the variability, for each pair-wise comparison we perform between features and popularity metrics, still to be explained by other factors. In what follows, we introduce each feature and discuss their correlation with the popularity metrics.

Time: First, we examine the temporal features of online news consumption. Our choice is motivated by [Ahmed et al., 2013, Marujo et al., 2011], where the authors successfully employ date and time information as features for their prediction tasks. As the correlation coefficient scores shown in Table 3 suggest, this category of features is weakly correlated with the `Tweets` and `Pageviews` metrics.

Figure 4 shows the normalised metric values observed on different days of the week, and suggests a 24-hour periodic, circadian variation.[3] Moreover, the figure indicates that the news consumption is lower at the beginning and at the end of the week, with a significant drop during the weekend, while there is a notable peak mid-week. Although this pattern is shared among Facebook metrics, `Tweets` do not show high volatility on the weekdays despite the fact that they exhibit a decrease of approximately 25% during the weekend.

Figure 5 shows the normalised metric values observed in certain time periods of the day. The figure implies an increase in the amount of news consumption between 08:00 and 16:00. However, this increase may be simply because more articles are published in this time period. Indeed, as seen in the bottom part of Figure 6, on a given day, more than half of the news articles are published during this time period. In Figure 6, we show the average (per article) metric values observed one hour and one week after an article is published. The figure indicates that during

[3]The values are normalised by the total volume across all metrics and all days.

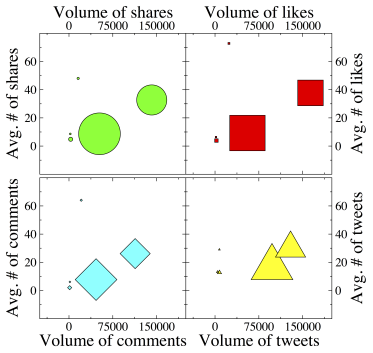


Figure 7: Impact of news source on the metrics.

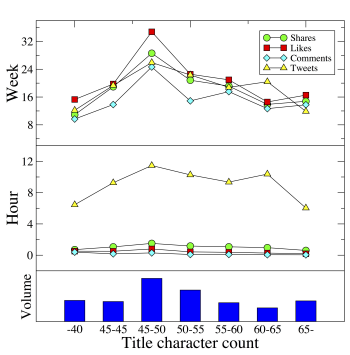


Figure 8: Impact of title length on the metrics.

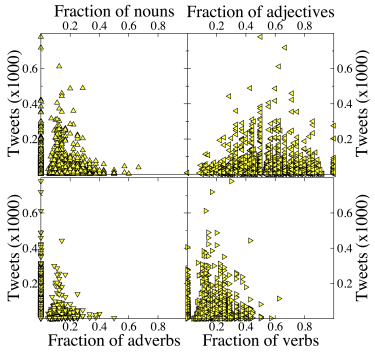


Figure 9: Impact of part-of-speech tags on the metrics.

late hours (16:00–24:00) online social activity and news consumption is actually higher. That is, the peaks in the total and average volume of news consumption occur in different time windows.

News source: Similar to [Bandari et al., 2012], we use the news source as a feature. The coefficients reported in Table 3 suggest a small effect size, which improves as we move from the hourly to the weekly counts of the popularity metrics. In Figure 7, we display the average and total metric values associated with the articles obtained from five news distributors. In this figure, the size of the symbols indicate the share of the news distributor in our news collection. We observe that a large portion of the articles are delivered by two major distributors, Reuters and Associated Press, while the share of the remaining agencies in the total news volume is much smaller. We observe a slightly different behaviour for the two biggest news distributors. Although Associated Press has a lower volume of articles, its articles result in larger metric values. The gap between the two distributors is consistent for `Shares`, `Likes`, and `Comments`, while the gap is smaller for `Tweets`.

Length: Our length features include the number of characters, words, and sentences in the body of the news articles. The first two features are computed also for the titles of articles. The correlation coefficients shown in Table 3 assume the features that were extracted using the body of the articles. As the results suggest, the length of an article is positively correlated with the attention it receives. In other words, longer articles are associated with larger metric values although it is not possible to comment on the direction of the relationship.

A different effect is observed when title length is considered. Figure 8 shows that the Facebook metrics exhibit a subtle positive relationship with the number of characters in articles' titles, while `Tweets` appear to be positively correlated. However, this relationship becomes negative once the character count exceeds a certain threshold. We hypothesise that, due to the length constraint, Twitter users are less inclined or cannot post the URLs of news articles that exceed a certain character length. For the weekly observations, the Facebook metrics show a clear positive correlation with the character count with all length features.

NLP: We examine the effect of linguistic features on the metrics. Our approach involves computing the distribution of nouns, adverbs, and verbs in the title and body of news articles.

Table 4: Distribution of (S)hares, (L)ikes, (C)omments, and (T)weets with respect to genre

		Share in volume (%)								Avg. indicator value per article							
		Hour				Week				Hour				Week			
Genre	Freq.	S	L	C	T	S	L	C	T	S	L	C	T	S	L	C	T
Business	0.11	0.05	0.05	0.06	0.08	0.05	0.04	0.04	0.08	0.7	0.3	0.1	9.2	10.0	9.6	7.4	17.7
Computers	0.02	0.02	0.01	0.01	0.03	0.01	0.01	0.01	0.03	1.5	0.3	0.1	20.0	12.5	7.5	5.1	39.9
Economy	0.05	0.03	0.01	0.01	0.04	0.02	0.02	0.02	0.03	1.0	0.2	0.0	10.5	9.0	9.9	8.3	17.4
Education	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.7	0.1	0.0	13.0	10.0	9.0	9.2	28.9
Entertainment	0.04	0.07	0.08	0.06	0.07	0.04	0.04	0.04	0.06	2.6	1.1	0.3	23.0	26.6	30.9	19.8	42.4
Health	0.02	0.03	0.03	0.05	0.04	0.02	0.02	0.02	0.03	2.6	1.0	0.5	25.5	29.4	31.5	21.3	46.6
Law	0.02	0.02	0.02	0.01	0.01	0.02	0.01	0.01	0.01	1.4	0.5	0.2	8.7	21.3	19.9	11.7	20.9
Life	0.03	0.04	0.02	0.01	0.03	0.05	0.05	0.04	0.03	1.6	0.3	0.1	12.5	32.4	37.1	22.8	26.2
Politics	0.14	0.16	0.23	0.35	0.13	0.17	0.20	0.23	0.13	1.6	1.0	0.5	11.1	29.4	37.8	32.0	24.8
Regional	0.12	0.11	0.08	0.06	0.10	0.11	0.11	0.11	0.10	1.3	0.4	0.1	10.3	22.6	24.8	17.7	22.9
Science	0.02	0.02	0.04	0.02	0.01	0.09	0.09	0.08	0.02	1.8	1.3	0.3	8.7	138.7	157.5	104.2	35.4
Sports	0.14	0.02	0.01	0.00	0.04	0.01	0.01	0.00	0.03	0.2	0.0	0.0	3.4	1.2	1.2	0.7	5.7
Tech	0.06	0.04	0.01	0.01	0.08	0.03	0.03	0.02	0.07	0.8	0.1	0.0	14.9	11.4	13.8	6.2	29.6
TopStory	0.17	0.26	0.34	0.30	0.22	0.29	0.29	0.29	0.24	2.2	1.2	0.3	15.5	40.6	46.5	32.7	37.7
World	0.07	0.12	0.06	0.05	0.11	0.09	0.09	0.08	0.11	2.6	0.5	0.1	18.9	32.7	34.5	23.2	41.9
Overall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.4	0.6	0.2	12.1	23.7	26.9	19.2	25.9

Our motivation for applying text analysis, even at this basic level, is that linguistic features can provide insights into certain aspects of the textual meaning or the impact on the reading experience. The scatterplot in Figure 9 shows the distribution of nouns, adjectives, adverbs, and verbs in news articles against the `Tweets`. The positively skewed distribution reveals that the news articles that are associated with the highest number of `Tweets` contain a low fraction of the aforementioned lexical structures. On the contrary, adjectives appear to be normally distributed, suggesting a random dispersal with a high concentration close to the mean. Here, we can observe that certain linguistic features are more prominent than the others and also convey a different degree of semantic information. However, the correlation coefficients shown in Table 3 suggest a weak association between the frequencies of corresponding linguistic features and our metrics.

Genre: In [Bandari et al., 2012], the authors use meta-information about the article category (i.e., genre) as one of the features. The authors observe that news related to certain genres have a more prominent presence in their dataset and most likely in the social media as well. Based on their results, we look further into specific genres. Table 4 shows the distribution of popularity metric values across different news genres. The second column shows the relative contribution of each genre to the total article count. Columns from 3 to 10 show the share of each genre in the total metric volume while columns from 11 to 18 show the average metric value per article for each genre.

Table 4 reveals some distinct patterns for genres. The most notable one involves genres that have a low share in the total volume but, on average, their articles are associated with high metric values. Such genres are probably related to more niche articles that have loyal readers, who tend to share in social media their links (e.g., `Science` articles). Another distinguishable pattern is that few genres dominate the others in terms of their share in the total volume and metric values (e.g., the `TopStory` genre). Yet another pattern includes genres whose articles have a low share in the total volume of metrics and are also associated with low metric values per article.

The `TopStory`, `Entertainment`, `Politics`, and `Science` appear to be the most prominent genres. This is further supported by the correlation analysis of the frequency data (column 2) and the volume data for all popularity metrics (columns 3 to 10). In all cases, we observe a significant positive relationship ($r > .95$, one-tailed), i.e., genres with high frequencies are associated with

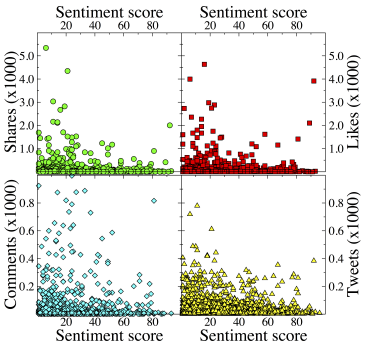


Figure 10: Impact of sentiment score on the metrics.

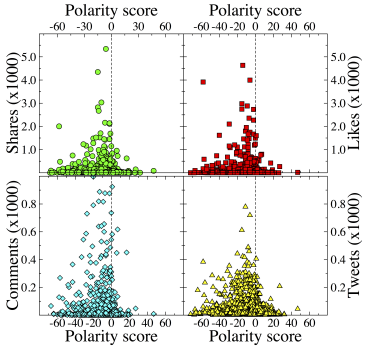


Figure 11: Impact of polarity score on the metrics.

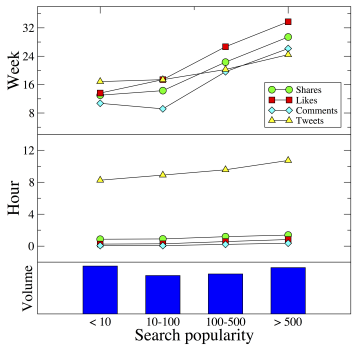


Figure 12: Impact of search popularity on the metrics.

high shares in the total volume of metrics. The correlation coefficients in Table 3 also support this finding (e.g., the `TopStory` genre has the highest correlation with the metrics).

Sentiment analysis: For sentiment analysis, we use SentiStrength, a lexicon-based sentiment analysis tool [Thelwall et al., 2012]. For every sentence in the article, the SentiStrength tool generates a positive and a negative sentiment score by analysing the sentiments associated with the words in the sentence. The positive scores are integers within the range of $+1$ (neutral) to $+5$ (extremely positive) while the negative scores are integers within the range of -1 (neutral) to -5 (extremely negative). We use the positive/negative score returned by SentiStrength to compute a sentimentality score and a polarity score for the sentence [Kucuktunc et al., 2012]. The sentimentality score of a sentence is computed as the sum of the absolute values of the positive and negative scores associated with the sentence. The polarity score of a sentence is a sum of its positive and negative scores. The sentimentality and polarity scores of an article are then computed by summing the corresponding scores associated with the sentences in the article. We compute the same scores also for the title of the article, treating the title as a single sentence.

Figure 10 shows a scatterplot of the four metrics plotted against the sentimentality scores assigned to each article. As observed, the distribution is asymmetric and very similar across all features, revealing signs of positive skew. We also observe that the most popular news articles tend to be those with neutral or subtly sentimental content, with a few outliers at the opposite end. The scatterplot in Figure 11 shows the distribution of metric values with respect to the polarity of news articles. Again, the shapes of the distributions appear to be very similar across all metrics, showing a tendency for scores to cluster in the peak rather than the tails of the distribution. The bulk of articles are concentrated very close to the y-axis and the negative side of the x-axis, suggesting that the content of popular news articles is characterised by a negative polarity. Finally, according to Table 3, highly sentimental and negative news appear to be more strongly correlated with online popularity. This is somewhat expected since news articles are known to be negative compared to other types of content [Wu et al., 2011].

Entity extraction: Similar to [Bandari et al., 2012], we use an in-house software to extract named entities from the news articles. In particular, we are interested in knowing if the number of named entities (people, locations, and organisations) in a news article affects its popularity.

In general, we observe that articles that mention at least one entity are more likely to become popular than articles that do not mention any.

Wikipedia: For each named entity mentioned in the article, we retrieve the popularity of the entity in Wikipedia.[4] Title- and body-level popularity values are then computed by summing the popularity values of all entities extracted from the title and article body, respectively. Other aggregation techniques, like averaging, yield inferior performance. These features give an idea about the past popularity of news articles although the correlation coefficients in Table 3 indicate small effects.

Twitter logs: As a proxy for the short-term popularity of articles, we use the popularity of their named entities in Twitter. For each entity, we track the volume of tweets referring to the entity starting one hour, one day, and one week before the article's publication date.[5] Similar to Wikipedia, the Twitter popularity feature also exhibits a weak correlation with the popularity metrics.

Web search logs: We repeat the same technique above on a large sample of queries submitted to the front-end of Yahoo search and compute the frequency of entities in the sample. Again, the popularity of an entity is computed at three different time intervals (one hour, one day, and one week before) and the aggregate search popularity for an article is determined as before. The correlation coefficients in Table 3 show a significant positive relationship between the search log features and the popularity metrics, in most cases of small effect size. In Figure 12, we observe an increasing trend in metric values with growing search popularity. This suggests that, even before an article is published, there is a potential increase in the number of associated entities that are queried in web search engines.

In this section, we investigated the existence of relationships between several popularity metrics and a large set of features that are commonly associated with or can be extracted from news articles or other sources. These features have been used in many prediction related tasks. Although some relationships have been identified, none of them, at least when considered individually, are significant, suggesting that the cold-start problem of predicting news article popularity is indeed a challenging task. In the rest of this paper, we dive deep into this prediction task by building learning models that combine these weak features and investigate further if such models have better predictive power.

6 Experiments

We conduct a series of experiments to understand the feasibility of predicting the future popularity of news articles at cold start. To this end, we learn models using the features presented in the previous section. We report our results under two different headings, classification and regression, since we modeled the prediction problem, separately, as a classification task and a regression task.

[4] `http://www.mediawiki.org/wiki/API:Main_page`

[5] We use Topsy's Otter API, available at `http://code.google.com/p/otterapi/`

6.1 Predicting article popularity through classification

As a first step, we try to reproduce the classification results presented in [Bandari et al., 2012] by Bandari et al. for `Tweets`. To this end, we split two weeks of articles (13,319 articles in total) into three classes based on their tweet counts: `A` (low popularity), `B` (medium popularity), and `C` (high popularity). Adopting the choice made in [Bandari et al., 2012], the tweet count ranges are set to $[1, 20]$, $(20, 100]$, and $(100, \infty)$ for the `A`, `B`, and `C` classes, respectively. Articles that are not tweeted are removed from the data and not included in set `A`. For the remaining indicators (`Shares`, `Likes`, and `Comments`), which are evaluated only in our work, we split the articles into three classes using the same ranges so that the results are comparable across different indicators. We experiment with the same classifiers used in [Bandari et al., 2012]: naive Bayesian (`NB`), bagging (`Bagging`), decision trees (`J48`), and support vector machines (`SVM`). Moreover, for comparison purposes, we include a baseline classifier `Baseline` that always predicts the majority class in the training data. We predict the popularity values one hour, one day, and one week after the articles are published. We perform logarithmic transformation on all features that exhibit a skewed distribution.

Despite our efforts to create a similar experimental setup, there are two minor differences between our setup and the setup used in [Bandari et al., 2012]. First, the articles used in [Bandari et al., 2012] (10,000 articles in total) are obtained from a large number of news sites while our collection is obtained from a single, relatively major news site. Second, in [Bandari et al., 2012], the popularity of articles are assumed to saturate after four days. In our case, as the closest value, we can use the popularity values obtained after one week. Nevertheless, since the features used in our study form a powerful superset of the features used in [Bandari et al., 2012], we expect to attain better or at least similar classification performance.

In Table 5, we report the classification performance in terms of the accuracy metric, i.e., the fraction of test articles whose class is correctly predicted by the classifier.[6] The reported results are obtained by performing cross-validation with ten folds, again adopting the choice made in [Bandari et al., 2012]. According to the table, for the (`Tweets`, `Week`) combination, the best performing classifier (`SVM`) achieves an accuracy of 79.7%, which is a bit lower than the best accuracy value (83.96%) reported in [Bandari et al., 2012] (achieved by `Bagging`). However, when we observe the relative improvement with respect to the baseline ($79.7\% - 70.3\% = 9.4\%$), we find it to be slightly higher in our case. Although not directly reported in [Bandari et al., 2012], the relative improvement in their case can be estimated to be $83.96\% - 76\% = 7.96\%$ using the data the authors provided in Tables 5 and 6. In either case, the reported results are comparable, and hence we believe that we were able to reproduce the results reported in [Bandari et al., 2012] to a certain extent.

A potential issue in the previous experiment is the use of cross-validation, which implies that the learned models may exploit future data about news popularity. Therefore, the observed classification performance may not be realistic. In a real-life setting, a model would be trained at a fixed point in time using only the features extracted from previously seen articles and then it would be applied to predict the popularity of newly received articles. Hence, instead of performing cross-validation, we repeat the previous prediction experiment by splitting our data into a training and test set in the time dimension (all other parameters of the experiment remaining the same). The training set contains articles published in the first week and the test set contains articles of the following week. The two sets are roughly equal in size.

[6] We do not report the classification accuracies for the `Pageviews` metric as this may reveal confidential information about the distribution of page views.

Table 5: Classification accuracy (ten-fold cross validation, excluding zero-popularity articles)

	Shares			Likes			Comments			Tweets		
Technique	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
Baseline	0.988	0.842	0.820	0.978	0.816	0.788	0.970	0.776	0.746	0.840	0.710	0.703
NB	0.848	0.708	0.700	0.818	0.684	0.665	0.910	0.633	0.625	0.693	0.581	0.574
Bagging	0.988	0.837	0.814	0.978	0.810	0.778	0.970	0.767	0.729	0.858	0.749	0.741
J48	0.988	0.833	0.801	0.978	0.785	0.751	0.965	0.766	0.690	0.856	0.781	0.775
SVM	0.988	0.842	0.820	0.978	0.815	0.786	0.965	0.776	0.746	0.859	0.802	0.797

Table 6: Classification accuracy (training/test split, excluding zero-popularity articles)

	Shares			Likes			Comments			Tweets		
Technique	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
Baseline	0.985	0.812	0.786	0.982	0.822	0.793	0.978	0.778	0.752	0.839	0.706	0.698
NB	0.852	0.732	0.716	0.877	0.721	0.672	0.944	0.635	0.647	0.735	0.589	0.584
Bagging	0.985	0.809	0.786	0.982	0.812	0.778	0.978	0.769	0.737	0.858	0.737	0.74
J48	0.985	0.808	0.784	0.982	0.781	0.758	0.978	0.778	0.723	0.852	0.779	0.774
SVM	0.985	0.812	0.786	0.982	0.822	0.793	0.955	0.778	0.752	0.861	0.803	0.798

We note that, in this work, we are not able to create more than one training/test set because of the following two reasons. First, the data that is available to us spans exactly two weeks. Therefore, we do not have the opportunity to slide a time window over multiple weeks to obtain different training/test sets. Second, in our scenario, splitting the data in a more fine-grain manner (e.g., at the level of days instead of weeks) creates a bias. This is because i) we use certain temporal features related to the publication times of articles (e.g., the day of the week the article is published), ii) the publication rate of the articles varies in time (e.g., more articles are published during the week days), and iii) the reader activity varies in time. Therefore, we believe that splitting the data at the week level allows us to obtain a more fair distribution in the training and test sets (at the expense of not being able to conduct statistical significance tests).

Table 6 reports the results of this new experiment. According to the table, we do not see a major change in the results. The NB classifier seems to perform better than before in certain classification scenarios (e.g., (Likes, Hour) and (Comments, Hour)). But, for the remaining classifiers, the accuracies are somewhat similar to those observed in Table 5. In the remaining experiments, we continue to use the same setting, which is based on a training/test split.

Another issue that we observe in the methodology followed in [Bandari et al., 2012] is the artificial manipulation of the original news collection. Before conducting their experiments, the authors remove from the data every article that is not tweeted at all after it is published. This manipulation may lead to unfair results because, in a real-life setting, it is not possible to know whether an article will be tweeted or not before it is published. Hence, herein, we repeat the previous experiment after including zero-popularity articles in the A class. The results are reported in Table 7. We observe that the classification problem is now easier than before as the accuracy of the best performing classifier has increased in all scenarios. In particular, the best accuracy increases from 79.8% to 82.5% in case of the (Tweets, Week) scenario. On the other hand, the performance gap between the best performing classifiers and Baseline gets smaller. As an example, in case of (Tweets, Week), the improvement drops from 10.0% to 8.5%. In the remaining experiments, we always include zero-popularity articles in the A class.

Table 7: Classification accuracy (training/test split, including zero-popularity articles)

	Shares			Likes			Comments			Tweets		
Technique	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
Baseline	0.995	0.903	0.885	0.998	0.927	0.908	0.999	0.951	0.937	0.871	0.746	0.740
NB	0.907	0.805	0.774	0.943	0.788	0.709	0.991	0.799	0.784	0.772	0.642	0.633
Bagging	0.995	0.902	0.884	0.998	0.924	0.904	0.999	0.949	0.934	0.886	0.780	0.769
J48	0.995	0.903	0.883	0.998	0.922	0.905	0.999	0.951	0.931	0.883	0.805	0.804
SVM	0.995	0.903	0.885	0.998	0.927	0.908	0.999	0.951	0.937	0.890	0.829	0.825

Table 8: Fraction of instances in each of the three popularity classes

	Shares			Likes			Comments			Tweets		
Class	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
A	0.995	0.903	0.885	0.998	0.927	0.908	0.999	0.951	0.937	0.871	0.746	0.740
B	0.004	0.066	0.080	0.002	0.046	0.060	0.001	0.032	0.041	0.125	0.227	0.231
C	0.001	0.031	0.036	0.000	0.027	0.032	0.000	0.018	0.022	0.004	0.027	0.029

Table 9: Confusion matrix for (Tweets, Week)

	Predicted		
Actual	A	B	C
A	4,698	247	0
B	728	812	0
C	98	96	0

All of the results reported so far indicate high classification accuracies. But, how meaningful or useful are these results in practice? Can we really distinguish article popularity through classification? The answer lies in the surprisingly good performance of the Baseline classifier, which always predicts the label of the majority class in the training data. This implies that high accuracy values could be due to the highly skewed nature of the popularity distribution and the resulting class imbalance. In such scenarios, the classifiers are biased to learn and predict the majority class, leading to superficial accuracies (a known issue in machine learning).

But, how skewed is the class distribution at hand? In Table 8, we display the fraction of articles in the test set for each of the three classes (confirming Figure 2). As we can see, the collection is dominated by the unpopular articles in class A. In all cases, class C (the class of most popular articles) constitutes less than 4% of the sample. In a real-life setting, it is much more important to distinguish the articles in class C from the rest. The question is then how good are we in predicting class C articles. To answer this question, one can look at the confusion matrices, showing the true and false positive rates per class. In Table 9, as a representative case, we provide the confusion matrix for the (Tweet, Week) scenario (using the best performing classifier, SVM). According to the table, the classifier does quite well in correctly identifying class A articles. However, it fails to distinguish the articles in the most important class C from class A and B articles since no test instances are labeled as class C. This result indicates that the accuracy numbers reported in [Bandari et al., 2012] are very likely to be not useful either. We omit the results for the remaining scenarios (e.g., (Shares, Hour)) since the main finding is the same (i.e., the classifier does not label any test instance as class C).

Table 10: Root mean squared error

	Shares			Likes			Comments			Tweets		
Technique	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
BaselineR	1.087	2.088	2.191	0.583	1.895	2.038	0.306	1.623	1.792	1.701	1.931	1.950
LR	0.929	1.702	1.774	0.554	1.617	1.722	0.305	1.479	1.616	1.132	1.270	1.305
kNNR	1.113	2.162	2.266	0.769	2.188	2.321	0.434	2.040	2.213	1.537	1.720	1.753
SVM	1.051	1.897	1.947	0.605	1.769	1.838	0.308	1.713	1.902	1.135	1.278	1.315

Table 11: R^2

	Shares			Likes			Comments			Tweets		
Technique	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
LR	0.325	0.376	0.381	0.102	0.272	0.285	0.026	0.169	0.186	0.555	0.566	0.551
kNNR	0.105	0.144	0.154	0.021	0.110	0.120	0.003	0.037	0.050	0.327	0.341	0.331
SVM	0.192	0.318	0.341	0.010	0.210	0.237	0.015	0.033	0.041	0.552	0.560	0.543

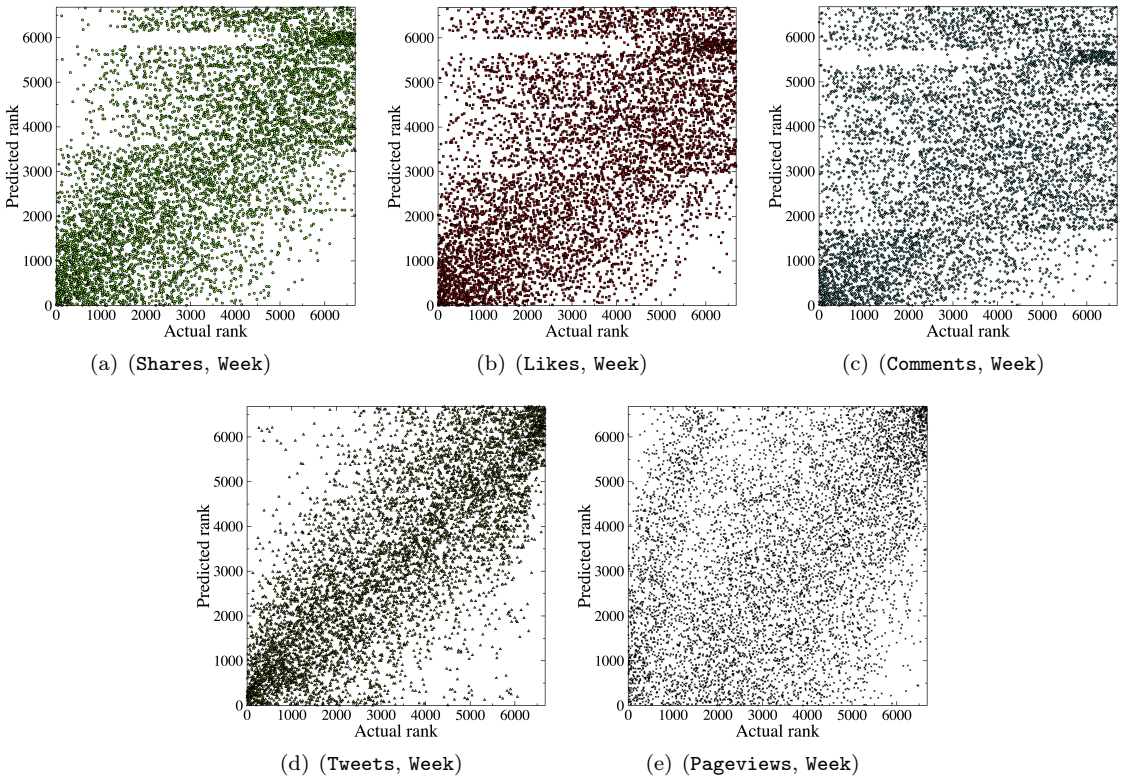
6.2 Predicting article popularity through regression

Given that classification does not yield meaningful performance, we turn our attention to regression and observe the performance in predicting the actual popularity values rather than the popularity class values. To this end, we evaluate three regression approaches: linear regression (LR), k-nearest neighbor regression (kNNR), and support vector machines (SVM). For comparison purposes, we also use a simple baseline (BaselineR) that always predicts the mean value in the training data. We perform some logarithmic transformation on the target popularity values before regression. To check whether collinearity of the attributes affects the performance, we performed some preliminary experiments and identified seven attributes whose correlation values were larger than 0.90. When training the linear regression algorithm, we turned off/on the parameter for eliminating collinear attributes. These experiments showed that the results were almost identical and thus collinearity of the attributes does not harm the regression performance.

In Tables 10 and 11, the regression performance is reported in terms of the root mean squared error and R^2 (coefficient of determination) measures, respectively. According to Table 10, LR appears to be the best performing regression technique in all cases. Overall, the calculated errors are low, and also there is considerable improvement with respect to BaselineR. As we go from Hour to Week, the error tends to increase due to the larger variation in popularity values of articles. However, we observe that the improvement with respect to BaselineR increases as well. This is because predicting late-stage popularity is easier than predicting early-stage popularity. According to Table 11, the explanatory power of learned regression models are quite low (perhaps, with the exception of Tweets). The results support those in Tables 10 in that LR is the best performing regression technique.

Although the regression results give an idea about the prediction quality, they still do not tell us whether the predictions are biased towards unpopular articles. Moreover, in practice, accurate ranking of articles (in decreasing order of popularity) is more important than accurate prediction of their exact popularity. That is, given a popular and an unpopular article, the difference between the predicted popularity values is not of high importance as long as we can correctly rank the popular article above the unpopular article.

To visualize the ranking performance, in Figure 13, we display the actual versus predicted popularity ranks of the articles (e.g., the article with the highest popularity is ranked first). The



(a) (`Shares`, `Week`) (b) (`Likes`, `Week`) (c) (`Comments`, `Week`)

(d) (`Tweets`, `Week`) (e) (`Pageviews`, `Week`)

Figure 13: Actual versus predicted ranks.

strongest correlation is observed between the actual and predicted tweet counts. For the remaining metrics, we observe relatively weaker correlation. The Kendall Tau (`KT`) correlation values (Tau-a) reported in Table 12 are consistent with the plots. For the `Week` case, the correlation values are 0.561, 0.441, 0.358, 0.287, 0.229 for `Tweets`, `Shares`, `Likes`, `Pageviews`, and `Comments`, respectively.

Finally, we evaluate the performance focusing on the top ranked articles. This is important because, as we mentioned before, only a small fraction of articles gain visibility due to the limited space in web pages and the limited attention span of users. Therefore, it is vital to get the popularity ranking right especially at the high ranks by correctly identifying the most popular articles. To evaluate the performance at top ranks, we define the recall@k (`R@k`) metric (this is different than the traditional recall metric in information retrieval). Our metric basically selects the articles that are placed in the top k ranks by the prediction algorithm and then computes what fraction of them also appear within the top k ranks in the actual popularity ranking.

We report `R@k` values for $k \in \{10, 100, 1,000\}$ in Table 12. For $k = 1,000$, even with the best prediction scenario (i.e., `Tweets`), we observe that about 45% of the articles in the top 1,000 ranks are not ranked among the top 1,000 articles in the actual popularity ranking. The results are even worse as the k value decreases. For $k = 100$, in the best case (i.e., (`Likes`, `Week`)), only 12 out of 100 most popular articles could be retrieved. These final results illustrate the real difficulty of the

Table 12: Kendall Tau (Tau-a) and recall@k

	Shares			Likes			Comments			Tweets			Pageviews		
Metric	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week	Hour	Day	Week
KT	0.324	0.427	0.441	0.116	0.334	0.358	0.041	0.208	0.229	0.551	0.569	0.561	0.078	0.286	0.287
R@10	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R@100	0.110	0.080	0.100	0.070	0.070	0.120	0.090	0.090	0.140	0.240	0.110	0.090	0.010	0.020	0.060
R@1000	0.467	0.479	0.476	0.357	0.452	0.444	0.212	0.426	0.420	0.578	0.557	0.548	0.212	0.173	0.245

problem and indicates the superficial nature of the previous results obtained through classification and regression [Bandari et al., 2012].

We note that we also conducted experiments with two different learning-to-rank software (SVM-Light with the "-z p" option and Sofia-ML with the "--loop_type rank" option), both implementing pairwise learning algorithms, but the attained results were inferior compared to those attained by the standard regression techniques. This is probably because the learning-to-rank techniques are effective when many rankings are available in the training phase. In our scenario, we had only a single ranking for training (and a single ranking for testing). Hence, the learning-to-rank techniques turned out to be not so useful.

7 Concluding Discussion

In this work, we investigated the cold-start news popularity prediction problem. We measured the popularity of articles in terms of several sociometrics as well as page views. Using the content of news articles and external sources such as Wikipedia and search logs, we engineered a large number of features that may indicate the future popularity of news articles. We used these features in both classification and regression frameworks for popularity prediction, two well-known machine learning techniques used in prediction tasks.

Our work revealed that predicting news popularity at cold start is not yet a solved problem. This is mainly a consequence of the highly skewed distribution of popularity metrics. We observed that the classifiers were biased to learn unpopular articles due to the imbalanced distribution. Our results indicated that the promising performance reported by Bandari et al. [Bandari et al., 2012] may be somewhat superficial. By focusing only on the top of the rankings, we could show that highly popular articles cannot be accurately detected, rendering the predictions not useful for most practical scenarios.

We observed weak correlations between news popularity and certain features such as genre and news source with the remaining features making little or no contribution. Our findings suggest that popularity is disconnected from the inherent structural characteristics of news content and cannot be easily modelled. We believe that news popularity can be accurately predicted only if early-stage popularity measurements are incorporated into the prediction models as features. In general, increasing the duration of such measurements will increase the accuracy of predictions but decrease their importance (news are often ephemeral), leading to an interesting trade-off.

As a side result, our analysis showed that each sociometric has a different behaviour, characterised by its rate of growth, range of values, and other temporal features. We observed that not all popularity metrics are necessary to estimate a measure of popularity, due to a high level of covariance and a varying degree of association with page access. Hence, one could strategically

employ the popularity metrics depending on the end goal, e.g., `Tweets` for assessing early-stage popularity or `Shares` for long-term popularity.

Although we reported a "negative" result, we believe that there is value in replicating and validating previous work. Beside the detailed correlation analysis, the identification of additional features to be used in the prediction can be seen as a novel contribution. Perhaps, even more importantly, we questioned aspects of the typical research methodology and whether the obtained results are truly valuable in practice. While some of these issues are known in different communities, way too often these questions are not asked.

References

[Agarwal et al., 2012] Agarwal, D., Chen, B.-C., and Wang, X. (2012). Multi-faceted ranking of news articles using post-read actions. In *Proc. 21st ACM Int'l Conf. Information and Knowledge Management*, pages 694–703.

[Ahmed et al., 2013] Ahmed, M., Spagna, S., Huici, F., and Niccolini, S. (2013). A peek into the future: Predicting the evolution of popularity in user generated content. In *Proc. 6th ACM Int'l Conf. Web Search and Data Mining*, pages 607–616.

[Arapakis et al., 2014] Arapakis, I., Cambazoglu, B., and Lalmas, M. (2014). On the feasibility of predicting news popularity at cold start. In Aiello, L. and McFarland, D., editors, *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 290–299. Springer International Publishing.

[Bandari et al., 2012] Bandari, R., Sitaram, A., and Huberman, Bernardo, A. (2012). The pulse of news in social media: Forecasting popularity. In *Proc. 6th Int'l Conf. Weblogs and Social Media*.

[Brody et al., 2006] Brody, T., Harnad, S., and Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 57:1060–1072.

[Freyne et al., 2010] Freyne, J., Berkovsky, S., Daly, E. M., and Geyer, W. (2010). Social networking feeds: recommending items of interest. In *Proc. 4th ACM Conf. Recommender Systems*, pages 277–280.

[Garimella and Castillo, 2014] Garimella, V. R. K. and Castillo, C. (2014). Fast: Forecast and analytics of social media and traffic. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW Companion '14, pages 13–16, New York, NY, USA. ACM.

[Givon and Lavrenko, 2009] Givon, S. and Lavrenko, V. (2009). Predicting social-tags for cold start book recommendations. In *Proc. 3rd ACM Conf. Recommender Systems*, pages 333–336.

[Gupta et al., 2012] Gupta, M., Gao, J., Zhai, C., and Han, J. (2012). Predicting future popularity trend of events in microblogging platforms. *Proc. American Society for Information Science and Technology*, 49(1):1–10.

[Jamali and Rangwala, 2009] Jamali, S. and Rangwala, H. (2009). Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proc. 2009 Int'l Conf. Web Information Systems and Mining*, pages 32–38.

[Kim et al., 2011] Kim, S.-D., Kim, S.-H., and Cho, H.-G. (2011). Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *Proc. 2011 IEEE 11th Int'l Conf. Computer and Information Technology*, pages 449–454.

[Kucuktunc et al., 2012] Kucuktunc, O., Cambazoglu, B. B., Weber, I., and Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! Answers. In *Proc. 5th ACM Int'l Conf. Web Search and Data Mining*, pages 633–642.

[Lehmann et al., 2012] Lehmann, J., Gonçalves, B., Ramasco, J. J., and Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proc. 21st Int'l Conf. World Wide Web*, pages 251–260.

[Lerman and Hogg, 2010] Lerman, K. and Hogg, T. (2010). Using a model of social dynamics to predict popularity of news. In *Proc. 19th Int'l Conf. World Wide Web*, pages 621–630.

[Levi et al., 2012] Levi, A., Mokryn, O., Diot, C., and Taft, N. (2012). Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In *Proc. 6th ACM Conf. Recommender Systems*, pages 115–122.

[Liu et al., 2011] Liu, N. N., Meng, X., Liu, C., and Yang, Q. (2011). Wisdom of the better few: cold start recommendation via representative based rating elicitation. In *Proc. 5th ACM Conf. Recommender Systems*, pages 37–44.

[Manduchi and Picard, 2009] Manduchi, A. and Picard, R. (2009). Circulations, revenues, and profits in a newspaper market with fixed advertising costs. *Journal of Media Economics*, 22(4):211–238.

[Marujo et al., 2011] Marujo, L., Bugalho, M., da Silva Neto, J. P., Gershman, A., and Carbonell, J. (2011). Hourly traffic prediction of news stories. In *3rd Int'l Workshop on Context-Aware Recommender Systems*.

[Mathioudakis et al., 2010] Mathioudakis, M., Koudas, N., and Marbach, P. (2010). Early online identification of attention gathering items in social media. In *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining*, pages 301–310.

[Phelan et al., 2009] Phelan, O., McCarthy, K., and Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proc. 3rd ACM Conf. Recommender Systems*, pages 385–388.

[Pinto et al., 2013] Pinto, H., Almeida, J. M., and Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of youtube videos. In *Proc. 6th ACM Int'l Conf. Web Search and Data Mining*, pages 365–374.

[Quercia et al., 2010] Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., and Crowcroft, J. (2010). Recommending social events from mobile phone location data. In *2010 IEEE 10th Int'l Conf. Data Mining*, pages 971–976.

[Ruan et al., 2012] Ruan, Y., Purohit, H., Fuhry, D., Parthasarthy, S., and Sheth, A. P. (2012). Prediction of topic volume on twitter. In *Proc. 4th Int'l ACM Conf. Web Science*.

[Shamma et al., 2011] Shamma, D. A., Yew, J., Kennedy, L., and Churchill, E. F. (2011). Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proc. 5th Int'l Conf. Weblogs and Social Media*.

[Szabo and Huberman, 2010] Szabo, G. and Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, 53:80–88.

[Tatar et al., 2011] Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., and Fdida, S. (2011). Predicting the popularity of online articles based on user comments. In *Proc. Int'l Conf. Web Intelligence, Mining and Semantics*, pages 67:1–67:8.

[Thelwall, 2006] Thelwall, M. (2006). Bloggers during the London attacks: Top information sources and topics. In *Proc. 15th Int'l Conf. World Wide Web*.

[Thelwall et al., 2012] Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social Web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

[Thorndike, 1953] Thorndike, L. R. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

[Tsagkias et al., 2010] Tsagkias, M., Weerkamp, W., and de Rijke, M. (2010). News comments: exploring, modeling, and online prediction. In *Proc. 32nd European Conf. Advances in Information Retrieval*, pages 191–203.

[Tumasjan et al., 2010] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proc. 4th Int'l Conf. Weblogs and Social Media*, pages 178–185.

[Vasconcelos et al., 2014a] Vasconcelos, M., Almeida, J., and Gonçalves, M. (2014a). What makes your opinion popular?: Predicting the popularity of micro-reviews in foursquare. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pages 598–603, New York, NY, USA. ACM.

[Vasconcelos et al., 2014b] Vasconcelos, M., Almeida, J., Gonçalves, M., Souza, D., and Gomes, G. (2014b). Popularity dynamics of foursquare micro-reviews. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, pages 119–130, New York, NY, USA. ACM.

[Wu et al., 2011] Wu, S., Tan, C., Kleinberg, J. M., and Macy, M. W. (2011). Does bad news go away faster? In *Proc. 5th Int'l Conf. Weblogs and Social Media*.

[Yu et al., 2011] Yu, B., Chen, M., and Kwok, L. (2011). Toward predicting popularity of social marketing messages. In *Proc. 4th Int'l Conf. Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 317–324.

[Zhang et al., 2011] Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26:55–62.