

# The Pulse of News in Social Media: Forecasting Popularity

Roja Bandari\*

Sitaram Asur<sup>†</sup>

Bernardo Huberman<sup>‡</sup>

## Abstract

News articles are extremely time sensitive by nature. There is also intense competition among news items to propagate as widely as possible. Hence, the task of predicting the popularity of news items on the social web is both interesting and challenging. Prior research has dealt with predicting eventual online popularity based on early popularity. It is most desirable, however, to predict the popularity of items prior to their release, fostering the possibility of appropriate decision making to modify an article and the manner of its publication. In this paper, we construct a multi-dimensional feature space derived from properties of an article and evaluate the efficacy of these features to serve as predictors of online popularity. We examine both regression and classification algorithms and demonstrate that despite randomness in human behavior, it is possible to predict ranges of popularity on twitter with an overall 84% accuracy. Our study also serves to illustrate the differences between traditionally prominent sources and those immensely popular on the social web.

## 1 Introduction

News articles are very dynamic due to their relation to continuously developing events that typically have short lifespans. For a news article to be popular, it is essential for it to propagate to a large number of readers within a short time. Hence there exists a competition among different sources to generate content which is relevant to a large subset of the population and becomes virally popular.

Traditionally, news reporting and broadcasting has been costly, which meant that large news agencies dominated the competition. But the ease and low cost of online content creation and sharing has recently changed the traditional rules of competition for public attention. News sources now concentrate a large portion of their attention on online mediums where they can disseminate their news effectively and to a large population. It is therefore common for almost all major news sources to have active accounts in social media services like Twitter to take advantage of the enormous reach these services

provide.

Due to the time-sensitive aspect and the intense competition for attention, accurately estimating the extent to which a news article will spread on the web is extremely valuable to journalists, content providers, advertisers, and news recommendation systems. This is also important for activists and politicians who are using the web increasingly more to influence public opinion.

However, predicting online popularity of news articles is a challenging task. First, *context* outside the web is often not readily accessible and elements such as local and geographical conditions and various circumstances that affect the population make this prediction difficult. Furthermore, *network properties* such as the structure of social networks that are propagating the news, influence variations among members, and interplay between different sections of the web add other layers of complexity to this problem. Most significantly, intuition suggests that the *content* of an article must play a crucial role in its popularity. Content that resonates with a majority of the readers such as a major world-wide event can be expected to garner wide attention while specific content relevant only to a few may not be as successful.

Given the complexity of the problem due to the above mentioned factors, a growing number of recent studies [1], [2], [3], [4], [5] make use of early measurements of an item's popularity to predict its future success. In the present work we investigate a more difficult problem, which is prediction of social popularity without using early popularity measurements, by instead solely considering features of a news article *prior* to its publication. We focus this work on observable features in the content of an article as well as its source of publication. Our goal is to discover if any predictors relevant only to the content exist and if it is possible to make a reasonable forecast of the spread of an article based on content features.

The news data for our study was collected from Feedzilla<sup>1</sup> –a news feed aggregator– and measurements of the spread are performed on Twitter<sup>2</sup>, an immensely

---

\*UCLA.

<sup>†</sup>HP Labs.

<sup>‡</sup>HP Labs.

---

<sup>1</sup>[www.feedzilla.com](http://www.feedzilla.com)

<sup>2</sup>[www.twitter.com](http://www.twitter.com)

popular microblogging social network. Social popularity for the news articles are measured as the number of times a news URL is posted and shared on Twitter.

To generate features for the articles, we consider four different characteristics of a given article. Namely:

- The news source that generates and posts the article
- The category of news this article falls under
- The subjectivity of the language in the article
- Named entities mentioned in the article

We quantify each of these characteristics by a score making use of different scoring functions. We then use these scores to generate predictions of the spread of the news articles using regression and classification methods. Our experiments show that it is possible to estimate ranges of popularity with an overall accuracy of 84% considering only content features. Additionally, by comparing with an independent rating of news sources, we demonstrate that there exists a sharp contrast between traditionally popular news sources and the top news propagators on the social web.

In the next section we provide a survey of recent literature related to this work. Section 3 describes the dataset characteristics and the process of feature score assignment. In Section 4 we will present the results of prediction methods. Finally, in Section 5 we will conclude the paper and discuss future possibilities for this research.

## 2 Related Work

Stochastic models of information diffusion as well as deterministic epidemic models have been studied extensively in an array of papers, reaffirming theories developed in sociology such as diffusion of innovations [6]. Among these are models of viral marketing [7], models of attention on the web [8], cascading behavior in propagation of information [9] [10] and models that describe heavy tails in human dynamics [11]. While some studies incorporate factors for content *fitness* into their model [12], they only capture this in general terms and do not include detailed consideration of content features.

Salganik et al performed a controlled experiment on music, comparing quality of songs versus the effects of social influence[13]. They found that song quality did not play a role in popularity of highly rated songs and it was social influence that shaped the outcome. The effect of user influence on information diffusion motivates another set of investigations [14], [15],[16], [5].

On the subject of news dissemination, [17] and [18] study temporal aspects of spread of news memes online,

with [19] empirically studying spread of news on the social networks of digg and twitter and [20] studying facebook news feeds.

A growing number of recent studies predict spread of information based on early measurements (using early votes on digg, likes on facebook, click-throughs, and comments on forums and sites). [1] found that eventual popularity of items posted on youtube and digg has a strong correlation with their early popularity; [2] and [3] predict the popularity of a discussion thread using features based on early measurements of user comments. [4] propose the notion of a virtual temperature of weblogs using early measurements. [5] predict digg counts using stochastic models that combine design elements of the site -that in turn lead to collective user behavior- with information from early votes.

Finally, recent work on variation in the spread of content has been carried out by [21] with a focus on categories of twitter hashtags (similar to keywords). This work is aligned with ours in its attention to importance of content in variations among popularity, however they consider categories only, with news being one of the hashtag categories. [22] conduct similar work on social marketing messages.

## 3 Data and Features

This section describes the data, the feature space, and feature score assignment in detail.

**3.1 Dataset Description** Data was collected in two steps: first, a set of articles were collected via a news feed aggregator, then the number of times each article was linked to on twitter was found. In addition, for some of the feature scores, we used a 50-day history of posts on twitter. The latter will be explained in section 3.2 on feature scoring.

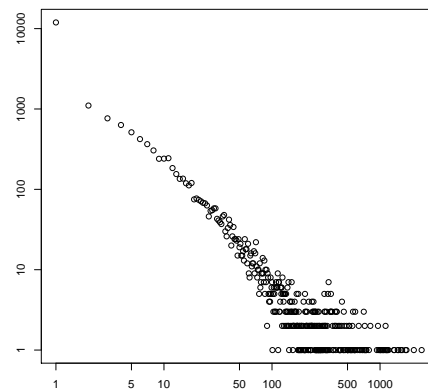


Figure 1: Log-log distribution of tweets.

Online news feed aggregators are services that col-

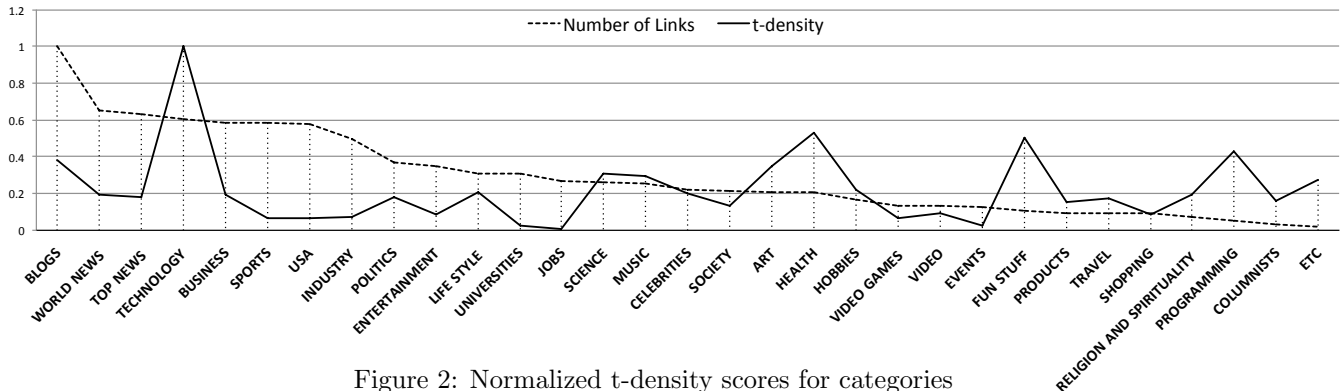


Figure 2: Normalized t-density scores for categories

lect and deliver news articles as they are published online. Using the API for a news feed aggregator named Feedzilla, we collected news feeds belonging to all news articles published online during one week (August 8th to 16th, 2011). The feed for an article includes a title, a short summary of the article, its url, and a time-stamp. In addition, each article is pre-tagged with a category either provided by the publisher or in some manner determined by Feedzilla. A fair amount of cleaning was performed to remove redundancies, resolve naming variations, and eliminate spam through the use of automated methods as well as manual inspection. As a result over 2000 out of a total of 44,000 items in the data were discarded.

The next phase of data collection was performed using Topsy<sup>3</sup>, a Twitter search engine that searches all messages posted on Twitter. We queried for the number of times each news link was posted or reshared on Twitter (tweeted or retweeted). Earlier research [17] on news meme buildup and decay suggest that popular news threads take about 4 days until their popularity starts to plateau. Therefore, we allowed 4 days for each link to fully propagate before querying for the number of times it has been shared.

The first half of the data was used in category score assignment (explained in the next section). The rest we partitioned equally into 10,000 samples each for training and test data for the classification and regression algorithms. Figure 1 shows the log distribution of total tweets over all data, demonstrating a long tail shape which is in agreement with other findings on distribution of Twitter information cascades [23]. The graph also shows that articles with zero tweets lie outside of the general linear trend of the graph because they did not propagate on the Twitter social network.

Our objective is to design features based on content to predict the number of tweets for a given article. In

the next section we will describe these features and the methods used to assign values or scores to features.

**3.2 Feature Description and Scoring** Choice of features is motivated by the following questions: Does the category of news affect its popularity within a social network? Do readers prefer factual statements or do they favor personal tone and emotionally charged language? Does it make a difference whether famous names are mentioned in the article? Does it make a difference who publishes a news article?

These questions motivate the choice of the following characteristics of an article as the feature space: the category that the news belongs to (e.g. politics, sports, etc.), whether the language of the text is objective or subjective, whether (and what) named entities are mentioned, and what is the source that published the news. These four features are chosen based on their availability and relevance, and although it is possible to add any other available features in a similar manner, we believe the four features chosen in this paper to be the most relevant.

We would like to point out that we use the terms article and link interchangeably since each article is represented by its URL link.

**3.2.1 Category Score** News feeds provided by Feedzilla are pre-tagged with category labels describing the content. We adopted these category labels and designed a score for them which essentially represents a prior distribution on the popularity of categories. Figure 2 shows a plot of categories and the number of article links in each category. We observe that news related to Technology has a more prominent presence in our dataset and most probably on twitter as a whole. Furthermore, we can see categories (such as Health) with low number of published links but higher rates of tweet per link. These categories perhaps have a niche following and loyal readers who are intent on posting and

<sup>3</sup><http://topsy.com>

retweeting its links.

Observing the variations in average tweets per link from Figure 2 we use this quantity to represent the prior popularity for a category. In order to assign a value (i.e. score) to each category, we use the the first 22,000 points in the dataset to compute the average tweet per article link in that category. We call this average tweet per link the *t-density* score and we will use this measure in score assignments for some other features as well.

**3.2.2 Subjectivity** Another feature of an article that can affect the amount of online sharing is its language. We want to examine if an article written in a more emotional, more personal, and more subjective voice can resonate stronger with the readers. Accordingly, we design a binary feature for subjectivity where we assign a zero or one value based on whether the news article or commentary is written in a more subjective voice, rather than using factual and objective language. We make use of a subjectivity classifier from LingPipe [24] a natural language toolkit. Since this requires training data, we use transcripts from well-known tv and radio shows belonging to Rush Limbaugh <sup>4</sup> and Keith Olberman <sup>5</sup> as the corpus for subjective language. On the other hand, transcripts from CSPAN <sup>6</sup> as well as the parsed text of a number of articles from the website FirstMonday <sup>7</sup> are used as the training corpus for objective language. The above two training sets provide a very high training accuracy of 99% and manual inspection of final results confirmed that the classification was satisfactory. Figure 3 illustrates the distribution of average subjectivity per source, showing that some sources consistently publish news in a more objective language and a somewhat lower number in a more subjective language.

**3.2.3 Named Entities** In this paper, a named entity refers to a known place, person, or organization. Intuition suggests that mentioning well-known entities can affect the spread of an article, increasing its chances of success. For instance, one might expect articles on Obama to achieve a larger spread than those on a minor celebrity. And it has been well documented that fans are likely to share almost any content on celebrities like Justin Bieber, Oprah Winfrey or Ashton Kutcher. We made use of the Stanford-NER <sup>8</sup> entity extraction tool to extract all the named entities present in the title and summary of each article. We then assign scores to over

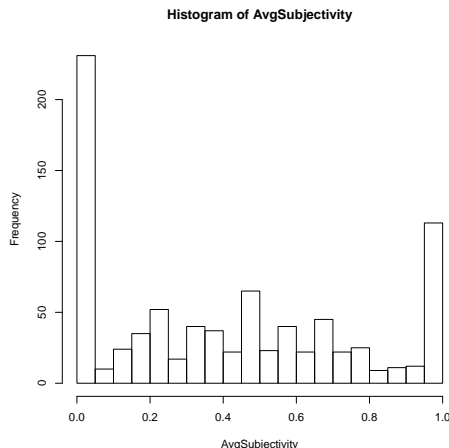


Figure 3: Distribution of average subjectivity of sources.

40,000 named entities by studying historical prominence of each entity on twitter over the timeframe of a month. The assigned score is the average t-density (tweet per link) of each named entity. To assign a score for a given article we use three different values: the number of named entities in an article, the highest score among all the named entities in an article, and the average score among the entities.

**3.2.4 Source Score** The data includes articles from 1350 unique sources on the web. We assign scores to each source based on the historical success of each source on Twitter. For this purpose, we collected the number of times articles from each source were shared on Twitter in the past. We used two different scores, first the aggregate number of times articles from a source were shared, and second the t-density of each source (average number of times each article belonging to a source was shared). The latter proved to be a better score assignment compared to the aggregate.

To investigate whether it is better to use a smaller portion of more recent history, or a larger portion going back farther in time and possibly collecting outdated information, we start with the two most recent weeks prior to our data collection and increase the number of days, going back in time. Figure 5 shows the trend of correlation between the t-density of sources in historical data and their true t-density of our dataset. We observe that the correlation increases with more datapoints from the history until it begins to plateau near 50 days. Using this result, we take 54 days of history prior to the first date in our dataset. We find that the correlation of the assigned score found in the above manner has a correlation of 0.7 with the t-density of the dataset. Meanwhile, the correlation between

<sup>4</sup><http://www.rushlimbaugh.com>

<sup>5</sup><http://www.msnbc.msn.com/id/32390086>

<sup>6</sup><http://www.c-span.org>

<sup>7</sup><http://firstmonday.org>

<sup>8</sup><http://nlp.stanford.edu/software/CRF-NER.shtml>

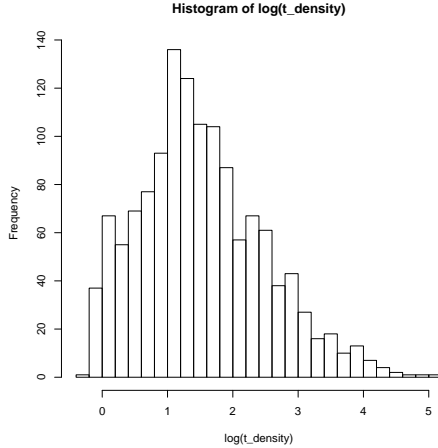


Figure 4: Distribution of log of source t-density scores

the source score and number of tweets of any given article is 0.35, suggesting that information about the source of publication alone is not sufficient in predicting popularity. Figure 4 shows the distribution of log of source scores (t-density). Taking the log of source scores produces a more normal shape, leading to improvements in regression algorithms.

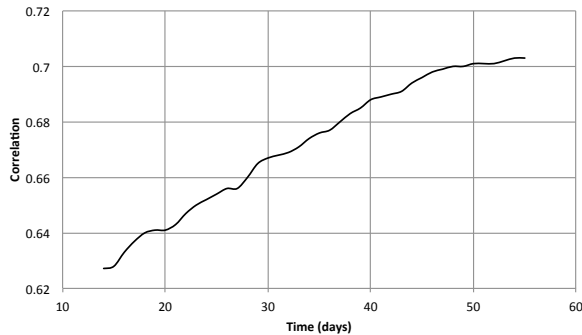


Figure 5: Correlation trend of source scores with t-density in data. Correlation increases with more days of historical data until it plateaus after 50 days.

We plot the timeline of t-densities for a few sources and find that t-density of a source can vary greatly over time. Figure 6 shows the t-density values belonging to the technology blog *Mashable* and *Blog Maverick*, a weblog of prominent entrepreneur, Mark Cuban. The t-density scores corresponding to each of these sources are 74 and 178 respectively. However, one can see that *Mashable* has a more consistent t-density compared to *Blog Maverick*.

In order to improve the score to reflect consistency we devise two methods; the first method is to smooth the measurements for each source by passing them

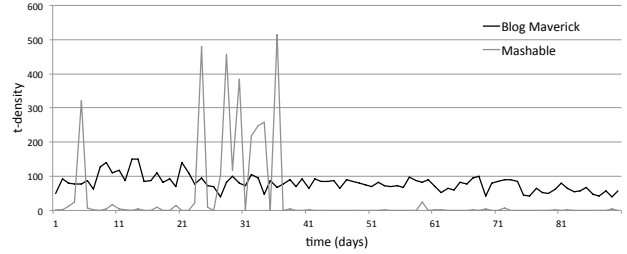
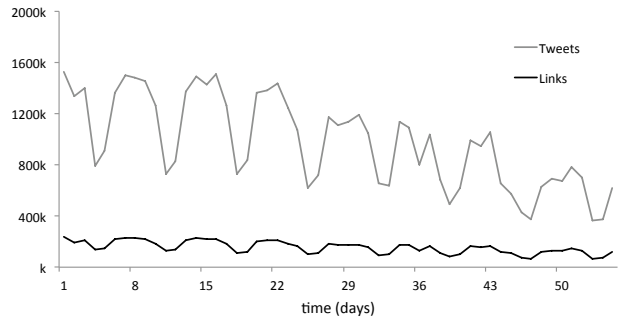
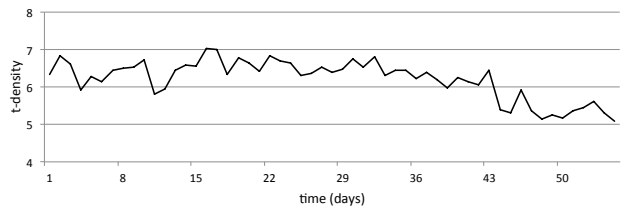


Figure 6: Timeline of t-density (tweet per link) of two sources.

through a low-pass filter. Second is to weight the score by the percentage of times a source’s t-density is above the mean t-density over all sources, penalizing sources that drop low too often. The mean value of t-densities over all sources is 6.4. Figure 7 shows the temporal variations of tweets and links over all sources. Notice that while both tweets and links have a weekly cycle, the t-density (tweets over links) does not have this periodic nature.



(a) tweets and links



(b) t-density

Figure 7: Temporal variations of tweets, links, and t-density over all sources

**3.2.5 Are top traditional news sources the most propagated?** As we assign scores to sources in our dataset, we are interested to know whether sources successful in this dataset are those that are conventionally considered prominent.

Google News<sup>9</sup> is one of the major aggregators and providers of news on the web. While inclusion in Google news results is free, Google uses its own criteria to rank the content and place some articles on its homepage, giving them more exposure. Freshness, diversity, and rich textual content are listed as the factors used by Google News to automatically rank each article as it is published. Because Google does not provide overall rankings for news sources, to get a rating of sources we use NewsKnife<sup>10</sup>. NewsKnife is a service that rates top news sites and journalists based on analysis of article’s positions on the Google news homepage and sub-pages internationally. We would like to know whether the sources that are featured more often on Google news (and thus deemed more prominent by Google and rated more highly by NewsKnife) are also those that become most popular on our dataset.

|             | Total Links | Total Tweets | t-density |
|-------------|-------------|--------------|-----------|
| Correlation | 0.57        | 0.35         | -0.05     |

Table 1: Correlation values between NewsKnife source scores and their performance on twitter dataset.

Accordingly we measure the correlation values for the 90 top NewsKnife sources that are also present in our dataset. The values are shown in Table 1. It can be observed that the ratings correlate positively with the number of links published by a source (and thus the sum of their tweets), but have no correlation (-0.05) with t-density which reflects the number of tweets that each of their links receives. For our source scoring scheme this correlation was about 0.7.

Table 2 shows a list of top sources according to NewsKnife, as well as those most popular sources in our dataset. While NewsKnife rates more traditionally prominent news agencies such as Reuters and the Wall Street Journal higher, in our dataset the top ten sources (with highest t-densities) include sites such as Mashable, AllFacebook (the unofficial facebook blog), the Google blog, marketing blogs, as well as weblogs of well-known people such as Seth Godin’s weblog and Mark Cuban’s blog (BlogMaverick). It is also worth noting that there is a bias toward news and opinion on web marketing, indicating that these sites actively use their own techniques to increase their visibility on Twitter.

While traditional sources publish many articles, those more successful on the social web garner more tweets. A comparison shows that a NewsKnife top source such as The Christian Science Monitor received

an average of 16 tweets in our dataset with several of its articles not getting any tweets. On the other hand, Mashable gained an average of nearly 1000 tweets with its least popular article still receiving 360 tweets. Highly ranked news blogs such as The Huffington Post perform relatively well in Twitter, possibly due to their active twitter accounts which share any article published on the site.

|                 |  |
|-----------------|--|
| NewsKnife       | <i>Reuters, Los Angeles Times, New York Times, Wall Street Journal, USA Today, Washington Post, ABC News, Bloomberg, Christian Science Monitor, BBC News</i> |
| Twitter Dataset | <i>Blog Maverick, Search Engine Land, Duct-tape Marketing, Seth’s Blog, Google Blog, Allfacebook, Mashable, Search Engine Watch</i>                          |

Table 2: Highly rated sources on NewsKnife versus those popular on the Twitter dataset

## 4 Prediction

In this work, we evaluate the performance of both regression and classification methods to this problem. First, we apply regression to produce exact values of tweet counts, evaluating the results by the R-squared measure. Next we define popularity classes and predict which class a given article will belong to. The following two sections describe these methods and their results.

| Variable    | Description                        |
|-------------|------------------------------------|
| $S$         | Source t-density score             |
| $C$         | Category t-density score           |
| $Subj$      | Subjectivity (0 or 1)              |
| $Ent_{ct}$  | Number of named entities           |
| $Ent_{max}$ | Highest score among named entities |
| $Ent_{avg}$ | Average score of named entities    |

Table 3: Feature set (prediction inputs)

**4.1 Regression** Once score assignment is complete, each point in the data (i.e. a given news article) will correspond to a point in the feature space defined by its category, subjectivity, named entity, and source scores. As described in the previous section, category, source, and named entity scores take real values while the subjectivity score takes a binary value of 0 or 1. Table 3 lists the features used as inputs of regression algorithms. We apply three different regression algorithms - linear

<sup>9</sup><http://news.google.com/>

<sup>10</sup><http://www.newsknife.com>

regression, k-nearest neighbors (KNN) regression and support vector machine (SVM) regression.

|                | Linear | SVM  |
|----------------|--------|------|
| All Data       | 0.34   | 0.32 |
| Tech Category  | 0.43   | 0.36 |
| Within Twitter | 0.33   | 0.25 |

Table 4: Regression Results

Since the number of tweets per article has a long-tail distribution (as discussed previously in Figure 1), we performed a logarithmic transformation on the number of tweets prior to carrying out the regression. We also used the log of source and category scores to normalize these scores further. Based on this transformation, we reached the following relationship between the final number of tweets and feature scores.

$$\ln(T) = 1.24\ln(S) + 0.45\ln(C) + 0.1Ent_{max} - 3$$

where  $S$  is the source t-density score,  $C$  is the category t-density score, and  $Ent_{max}$  is the maximum t-density of all entities found in the article. Equivalently,

$$T = S^{1.24}C^{0.45}e^{-(0.1Ent_{max}+3)}$$

with coefficient of determination  $R^2 = 0.258$ . Note that the  $R^2$  is the coefficient of determination and relates to the mean squared error and variance:

$$R^2 = 1 - \frac{MSE}{VAR}$$

Alternatively, the following model provided improved results:

$$T^{0.45} = (0.2S - 0.1Ent_{ct} - 0.1Ent_{avg} + 0.2Ent_{max})^2$$

with an improved  $R^2 = 0.34$ . Using support vector machine (SVM) regression [25], we reached similar values for  $R^2$  as listed in Table 4.

In K-Nearest Neighbor Regression, we predict the tweets of a given article using values from its nearest neighbors. We measure the Euclidean distance between two articles based on their position in the feature space [26]. Parameter  $K$  specifies the number of nearest neighbors to be considered for a given article. Results with  $K = 7$  and  $K = 3$  for a 10k test set are R-sq=0.05, with mean squared error of 5101.695. We observe that KNN performs increasingly more poorly as the dataset becomes larger.

**4.1.1 Category-specific prediction** One of the weakest predictors in regression was the Category score.

One of the reasons for this is that there seems to be a lot of overlap across categories. For example, one would expect *World News* and *Top News* to have some overlap, or the category *USA* would feature articles that overlap with others as well. So the categories provided by Feedzilla are not necessarily disjoint and this is the reason we observe a low prediction accuracy.

To evaluate this hypothesis, we repeated the prediction algorithm for particular categories of content. Using only the articles in the Technology category, we reached an  $R^2$  value of 0.43, indicating that when employing regression we can predict the popularity of articles within one category (i.e. Technology) with better results.

**4.2 Classification** Feature scores derived from historical data on Twitter are based on articles that have been tweeted and not those articles which do not make it to Twitter. As discussed in Section 3.1 this is evident in how the zero-tweet articles do not follow the linear trend of the rest of datapoints in Figure 1. Consequently, we do not include a zero-tweet class in our classification scheme and perform the classification by only considering those articles that were posted on twitter.

Table 5 shows three popularity classes A (1 to 20 tweets), B (20 to 100 tweets), C (more than 100) and the number of articles in each class in the set of 10,000 articles. Table 6 lists the results of support vector machine (SVM) classification, decision tree, and bagging [27] for classifying the articles. All methods were performed with 10-fold cross-validation. We can see that classification can perform with an overall accuracy of 84% in determining whether an article will belong to a low-tweet, medium-tweet, or high-tweet class.

In order to determine which features play a more significant role in prediction, we repeat SVM classification leaving one of the features out at each step. We found that publication source plays a more important role compared to other predictors, while subjectivity, categories, and named entities do not provide much improvement in prediction of news popularity on Twitter.

**4.2.1 Predicting Zero-tweet Articles** We perform binary classification to predict which articles will be at all mentioned on Twitter (zero tweet versus nonzero tweet articles). Using SVM classification we can predict –with 66% accuracy– whether an article will be linked to on twitter or whether it will receive zero tweets. We repeat this operation by leaving out one feature at a time to see a change in accuracy. We find that the most significant feature is the source, followed

by its category. Named entities and subjectivity did not provide more information for this prediction. So despite one might expect, we find that readers overall favor neither subjectivity nor objectivity of language in a news article.

It is interesting to note that while category score does not contribute in prediction of popularity within Twitter, it does help us determine whether an article will be at all mentioned on this social network or not. This could be due to a large bias toward sharing technology-related articles on Twitter.

| Class name | Range of tweets | Number of articles |
|------------|-----------------|--------------------|
| A          | 1–20            | 7,600              |
| B          | 20–100          | 1,800              |
| C          | 100–2400        | 600                |

Table 5: Article Classes

| Method             | Accuracy |
|--------------------|----------|
| Bagging            | 83.96%   |
| J48 Decision Trees | 83.75%   |
| SVM                | 81.54%   |
| Naive Bayes        | 77.79%   |

Table 6: Classification Results

## 5 Discussion and Conclusion

In this work we predicted the popularity of news items on Twitter using features extracted from the content of news articles. We have taken into account four features that cover the spectrum of the information that can be gleaned from the content - the source of the article, the category, subjectivity in the language and the named entities mentioned. Our results show that while these features may not be sufficient to predict the exact number of tweets that an article will garner, they can be effective in providing a range of popularity for the article on Twitter. We achieved an overall accuracy of 84% using classifiers. It is important to bear in mind that while it is intriguing to pay attention to the most popular articles –those that become viral on the web– a great number of articles spread in medium numbers. These medium levels can target highly interested and informed readers and thus the mid-ranges of popularity should not be dismissed.

Interestingly we have found that in terms of number of retweets, the top news sources on twitter are not necessarily the conventionally popular news agencies and various technology blogs such as Mashable and the

Google Blog are very widely shared in social media. Overall, we discovered that one of the most important predictors of popularity was the source of the article. This is in agreement with the intuition that readers are likely to be influenced by the news source that disseminates the article. On the other hand, the category feature did not perform well. One reason for this is that we are relying on categories provided by Feedzilla, many of which overlap in content. Thus a future task is to extract categories independently and ensure little overlap. Combining other layers of complexity described in the introduction opens up the possibility of better prediction. It would be interesting to incorporate network factors such as the influence of individual propagators to this work.

## 6 Acknowledgements

We would like to thank Vwani Roychowdhury for his support of the project.

## References

- [1] Gábor Szabó and Bernardo A. Huberman, “Predicting the popularity of online content,” *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.
- [2] Jong Gun Lee, Sue Moon, and Kavé Salamatian, “An approach to model and predict the popularity of online contents with explanatory factors,” in *Web Intelligence*. 2010, pp. 623–630, IEEE.
- [3] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida, “Predicting the popularity of online articles based on user comments,” in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA, 2011, WIMS ’11, pp. 67:1–67:8, ACM.
- [4] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho, “Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity,” in *IEEE 11th International Conference on Computer and Information Technology (CIT)*, 31 2011-sept. 2 2011, pp. 449–454.
- [5] Kristina Lerman and Tad Hogg, “Using a model of social dynamics to predict popularity of news,” in *WWW*. 2010, pp. 621–630, ACM.
- [6] E.M. Rogers, *Diffusion of innovations*, Free Pr, 1995.
- [7] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman, “The dynamics of viral marketing,” *TWEB*, vol. 1, no. 1, 2007.
- [8] Fang Wu and Bernardo A. Huberman, “Novelty and collective attention,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, pp. 17599–17601, 2007.
- [9] Daniel Gruhl, David Liben-Nowell, Ramanathan V. Guha, and Andrew Tomkins, “Information diffusion



- through blogspace,” *SIGKDD Explorations*, vol. 6, no. 2, pp. 43–52, 2004.
- [10] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst, “Cascading behavior in large blog graphs,” in *In SDM*, 2007.
- [11] Alexei Vázquez, João Gama Oliveira, Zoltán Dezső, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási, “Modeling bursts and heavy tails in human dynamics,” *Phys. Rev. E*, vol. 73, pp. 036127, Mar 2006.
- [12] M. V. Simkin and V. P. Roychowdhury, “A theory of web traffic,” *EPL (Europhysics Letters)*, vol. 82, no. 2, pp. 28006, Apr. 2008.
- [13] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts, “Experimental study of inequality and unpredictability in an artificial cultural market,” *Science*, vol. 311, no. 5762, pp. 854–856, 02 2006.
- [14] David Kempe, Jon M. Kleinberg, and Éva Tardos, “Maximizing the spread of influence through a social network,” in *KDD. 2003*, pp. 137–146, ACM.
- [15] Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, Xiangyang Lan, and Siddharth Suri, “Sequential influence models in social networks,” in *4th International Conference on Weblogs and Social Media*, 2010.
- [16] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu, “Identifying the influential bloggers in a community,” in *Proceedings of the international conference on Web search and web data mining*, New York, NY, USA, 2008, WSDM ’08, pp. 207–218, ACM.
- [17] Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in *KDD. 2009*, pp. 497–506, ACM.
- [18] Jaewon Yang and Jure Leskovec, “Patterns of temporal variation in online media,” in *WSDM. 2011*, pp. 177–186, ACM.
- [19] Kristina Lerman and Rumi Ghosh, “Information contagion: An empirical study of the spread of news on digg and twitter social networks,” in *ICWSM. 2010*, The AAAI Press.
- [20] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M. Lento, “Gesundheit! modeling contagion through facebook news feed,” in *ICWSM. 2009*, The AAAI Press.
- [21] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg, “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter,” in *Proceedings of the 20th international conference on World wide web*, New York, NY, USA, 2011, WWW ’11, pp. 695–704, ACM.
- [22] Bei Yu, Miao Chen, and Linchi Kwok, “Toward predicting popularity of social marketing messages,” in *SBP. 2011*, vol. 6589 of *Lecture Notes in Computer Science*, pp. 317–324, Springer.
- [23] Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury, “Information resonance on twitter: watching iran,” in *Proceedings of the First Workshop on Social Media Analytics*, New York, NY, USA, 2010, SOMA ’10, pp. 123–131, ACM.
- [24] Alias-i., “Lingpipe 4.1.0,” <http://alias-i.com/lingpipe>, 2008.
- [25] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [26] T. Hastie, R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer series in statistics. Springer, 2008.
- [27] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, “The weka data mining software: an update,” *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.