An Insight Into News Popularity

Jinchang Fan, Yujuan Qiu, Maansi Vatsan, Xiaoyu Qiao, Yizhuo Han | December 15, 2018

Introduction

With over 3.2 billion people using the internet worldwide, there is a tremendous amount of social media data that is collected daily. There is an immense potential for social media to be used as a useful tool to gather and disseminate important news, but in order to do so we must first better understand how consumers interact with news on social media sites on a daily basis.

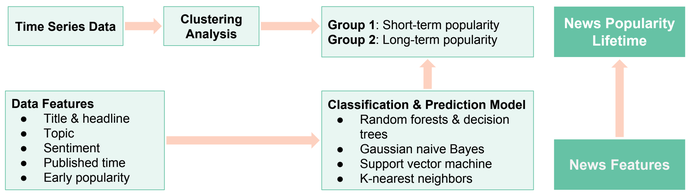
- **Hypothesis:** News popularity will vary significantly based on the time at which it is published, and different topics will gain maximum popularity on different social media sites.

This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn.

- The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: the economy, Microsoft, Obama and Palestine. Each of these topics was selected based on two major factors: their worldwide popularity and constant activity, and the fact that they relate to different types of entities (sector, company, person and country, respectively).

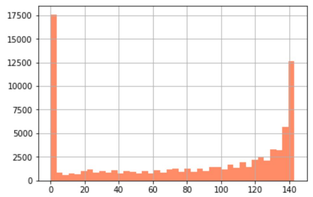
Variables	Description	Type	Additional Notes
IDLink	Unique identifier of news items	Numeric	The popularity of each article was measured over a period of two days
Topic	Query topic used to obtain the items in the official media sources	String	N/A
PublishDate	Date and time of the news items' publication	Timestamp	N/A
SentimentTitle	Sentiment score of the text in the news items' title, based on opinion mining algorithms.	Numeric	Scores generated using the **qdap** R package with default parameterization.
SentimentHeadline	Sentiment score of the text in the news items' headline, based on opinion mining algorithms.	Numeric	Scores generated using the **qdap** R package with default parameterization.
Early 10 & Early 20	Popularity of the news item measured at the 10th and 20th time points, respectively	Numeric	N/A
Facebook	Final value of the news items' popularity according to the social media source Facebook	Numeric	Popularity measure: number of shares. In 12.4% of cases, it was not possible to obtain this number.
Google+	Final value of the news items' popularity according to the social media source Google+	Numeric	Popularity measure: number of "+1." In 6.2% of cases, it was not possible to obtain this number.
LinkedIn	Final value of the news items' popularity according to the social media source LinkedIn	Numeric	Popularity measure: number of shares. In 6.2% of cases, it was not possible to obtain this number.

Analysis overview



Preprocessing

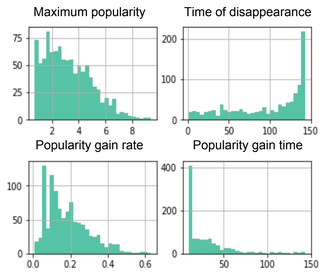
1. Selected a subset of the size 1,000 data points
2. Assigned weights to each topic proportional to the topic's frequency
3. Excluded any missing data
4. Randomly selected 1,000 data points based on the weights for each topic
5. Converted datetime information to a categorical variable consisting of three different date ranges, then converted to a numerical value via one-hot encoding
6. Converted the categorical variable 'topic' to a dummy variable via one-hot encoding



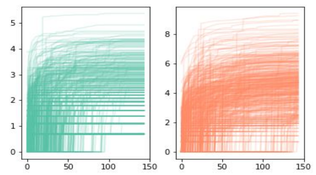
Time range including missing values (0)

Clustering analysis

Selected features	Feature description
Maximum popularity	The maximum popularity level that a news article attains
Time of disappearance	The time at which a news article stops gaining popularity, i.e. the time the article "disappears" on social media
Popularity gain rate	The maximum rate at which a news article gains popularity, separated into 10 periods
Popularity gain time	The time at which a news article begins to rapidly gain popularity

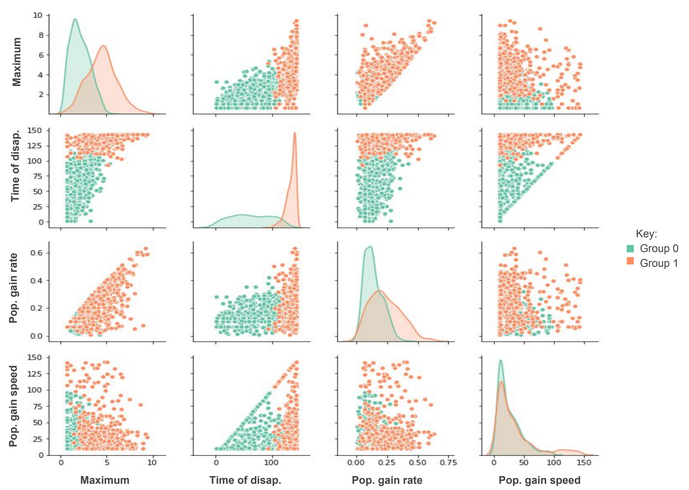


Clustering of features based on time series data



Left: Clustering analysis conducted on the time series data shows a division between news articles that gain popularity in the "short term" and "long term"

Below: Two component Gaussian mixture models showing the relationships between each of the four features, separated into short term (group 0, teal) and long term (group 1, orange) popularity



Logistic regression

Variable	Logistic regression coefficient
Title sentiment	-0.0976
Headline sentiment	0.0905
Topic: Microsoft	0.0653
Topic: Obama	-0.0568
Topic: Palestine	-0.0620
Topic: Economy	0.0347

Variable	Logistic regression coefficient
Early 10	0.1028
Early 20	-0.6111
Published Nov 2015 through Jan 2016	-0.0055
Published Feb 2016 through April 2016	-0.0204
Published May 2016 through July 2016	0.0302

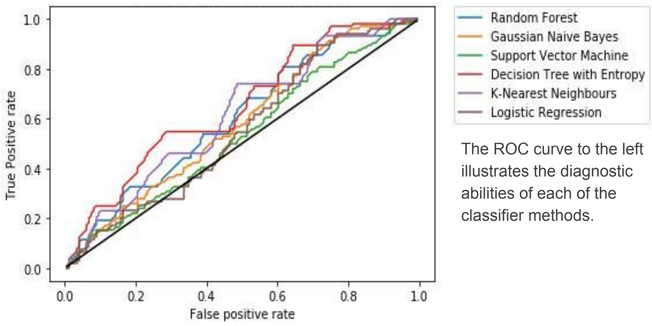
Classification

Variables for classification: For classification, we examined the influence of six variables:

- Sentiment score for the article headline
- Sentiment score of the article title
- Entity that the article belongs to (assigned a dummy variable)
- Time at which the article is published (assigned a dummy variable)
- Popularity of the news at the 10^{th} measured time point
- Popularity of the news at the 20^{th} measured time point

Classification method	Method accuracy	Standard deviation
Random forest	0.580	0.025
Gaussian naive Bayes	0.544	0.014
Support vector machine	0.527	0.021
Decision tree (Gini index)	0.599	0.042
Decision tree (Entropy index)	0.600	0.044
K-nearest neighbors	0.549	0.028
Logistic regression	0.530	0.018

Receiver operating characteristic curve



The ROC curve to the left illustrates the diagnostic abilities of each of the classifier methods.

Conclusion & further study

In order for a news article to reach the most people on social media, it is important that it maintain long-term popularity. Through the use of multiple classifying methods, we sought to determine which features of a news article contribute most to the article gaining long-term popularity. We found that the decision tree using the entropy index as a measure of accuracy provided us with the most accurate estimation of which group an article would fall into - long-term or short-term. Though we initially predicted that the topic of an article and the time at which it was published would play a prominent role in its eventual popularity, we found that the most important characteristic was in fact the article's popularity as measured at an earlier time point.

Further study:

Further study of this data set could involve an analysis of the popularity of different topics on each of the social media sites to see if a particular topic does better on a particular site. The data we used also only followed four specific topics on three social media sites. This research could be expanded to examine the same topics on other popular social media sites such as Twitter and Instagram.