Predicting Popularity in News Media

Jess Beering, Molly Carmody, Bella Hutchins,
Ryan Nicholson, Catalina Sanchez-Carrion

The Problem

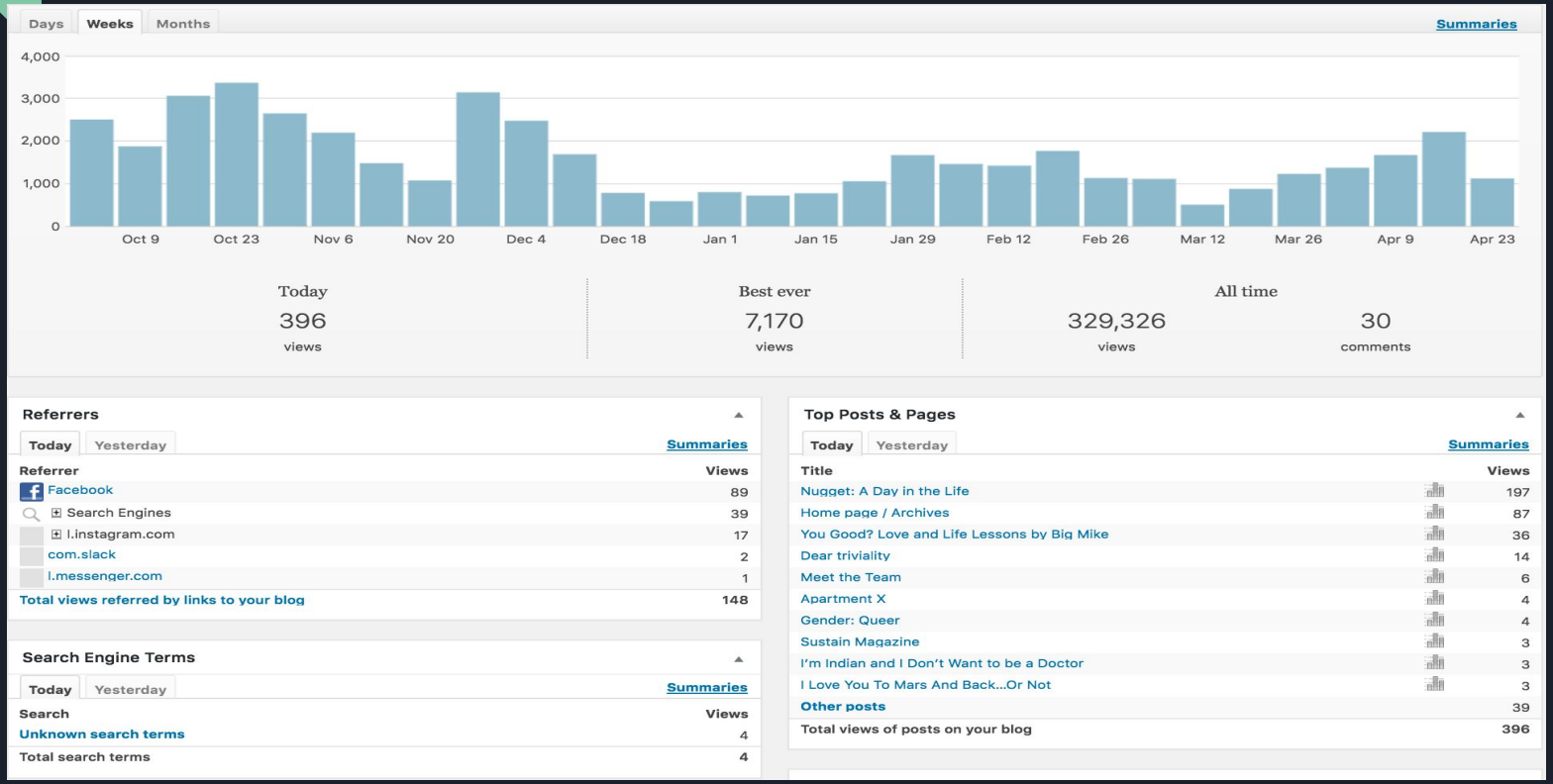
News and media outlets aim for their posts to reach the largest number of people as possible, therefore have a high popularity among social media users. We wanted to determine how the different factors that determine popularity impact the total number of shares an article gets.

What makes an article popular?

POPULARITY = NUMBER OF SHARES

Why did we choose this topic?

Why did we choose this topic?



Related Work - HP Labs

HP predicted Tweet popularity considering various features including:

- News category
- Whether the article is subjective or objective
- What named entities are mentioned
- What is the source of the news

Predictive models they used: Regression

Their conclusion: The source that published the news had the highest “importance” in their prediction → it most impacted the number of times the article was Tweeted

HP Labs - How our work compares

Similarities:

- Both used news category and presence of trends as a feature to determine popularity
- Both focused on an overall smaller number of features

Differences :

- HP failed to incorporate the readability of articles as a feature, which we learned has a strong impact on the popularity of an article
- They only used Regression as a predictive model, while we also incorporated Naive Bayes and Random Forest

Main Goals for the Project

- Explore attributes that best predict popularity/number of shares for an article
- Explore accuracy of three predictive models
  - Naïve Bayes
  - Random Forest
  - Regression
- Importance/relevance of number of Google Trending Topics in an article and its correlation with number of shares

Data Sets Used

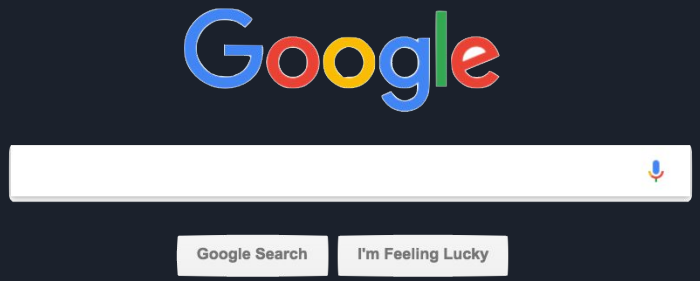
- UCI Online News Popularity Data Set including over 39,000 Mashable Articles
- Google Trending Topics from 10 categories per month (2013 and 2014)
- Google Trending Search Topics (top 50 trending in the US)

Features Used in Our Classifiers

- Week day posted
- Number of visual aides
- Article category:
  - Lifestyle
  - Entertainment
  - Business
  - Social Media
  - Technology
  - World
- Trending Word Count
- Automated Readability Index (ARI)

Features: Trending Words

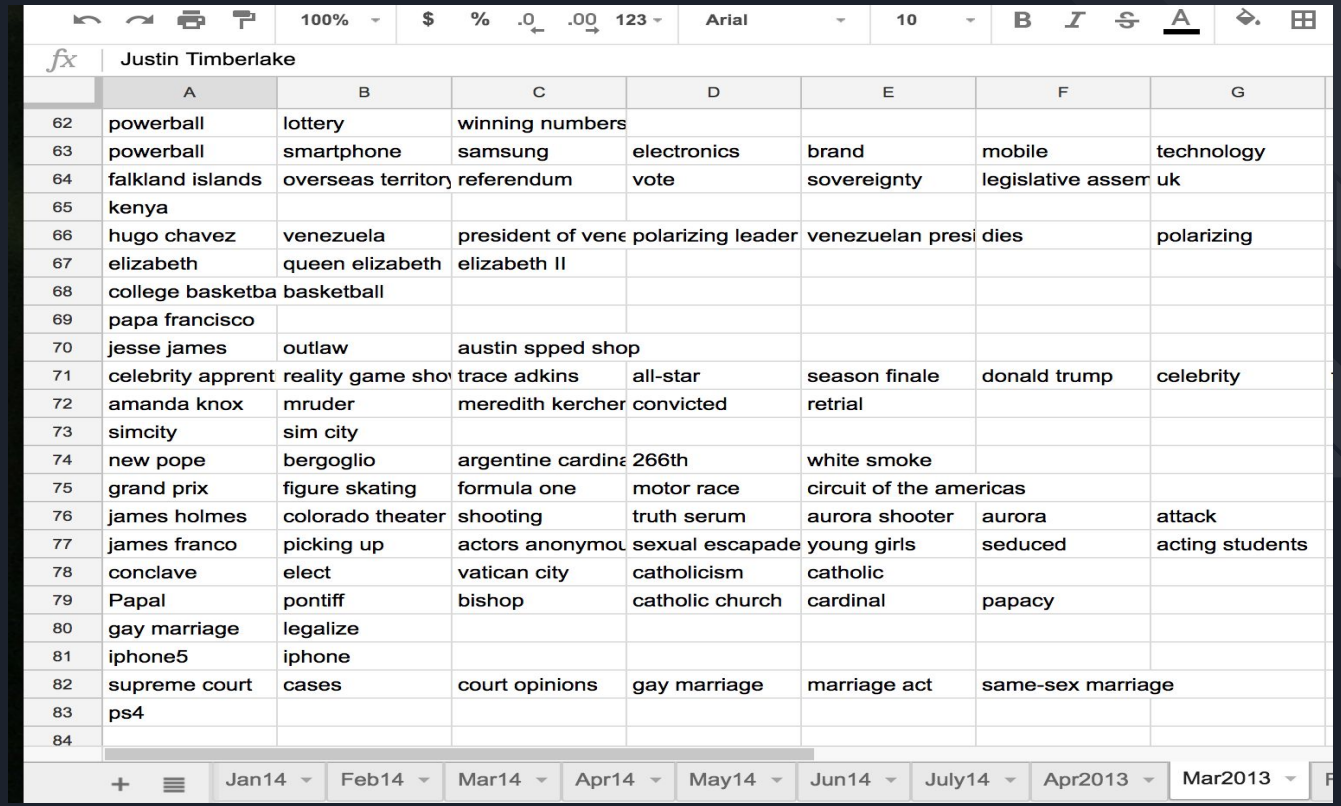
Wanted to mimic someone searching for an article online, based on what was relevant in the world at that time



COLLECTING TRENDING WORDS

1. Collected words from Google Trends for 10 categories, including books, actors, movies, among others
2. Collected words from Google Trending searches (top 50 in the US)
3. Added related words/synonyms

EXAMPLE OF TRENDING WORDS SPREADSHEET



100% $ % .0 .00 123 Arial 10

fx Justin Timberlake

	A	B	C	D	E	F	G
62	powerball	lottery	winning numbers				
63	powerball	smartphone	samsung	electronics	brand	mobile	technology
64	falkland islands	overseas territory	referendum	vote	sovereignty	legislative assem	uk
65	kenya						
66	hugo chavez	venezuela	president of vene	polarizing leader	venezuelan presi	dies	polarizing
67	elizabeth	queen elizabeth	elizabeth II				
68	college basketba	basketball					
69	papa francisco						
70	jesse james	outlaw	austin spped shop				
71	celebrity apprent	reality game sho	trace adkins	all-star	season finale	donald trump	celebrity
72	amanda knox	mruder	meredith kercher	convicted	retrial		
73	simcity	sim city					
74	new pope	bergoglio	argentine cardina	266th	white smoke		
75	grand prix	figure skating	formula one	motor race	circuit of the americas		
76	james holmes	colorado theater	shooting	truth serum	aurora shooter	aurora	attack
77	james franco	picking up	actors anonymou	sexual escapade	young girls	seduced	acting students
78	conclave	elect	vatican city	catholicism	catholic		
79	Papal	pontiff	bishop	catholic church	cardinal	papacy	
80	gay marriage	legalize					
81	iphone5	iphone					
82	supreme court	cases	court opinions	gay marriage	marriage act	same-sex marriage	
83	ps4						
84							

Jan14 Feb14 Mar14 Apr14 May14 Jun14 July14 Apr2013 Mar2013

Features: Trending Words

COUNTING THE TRENDING WORDS

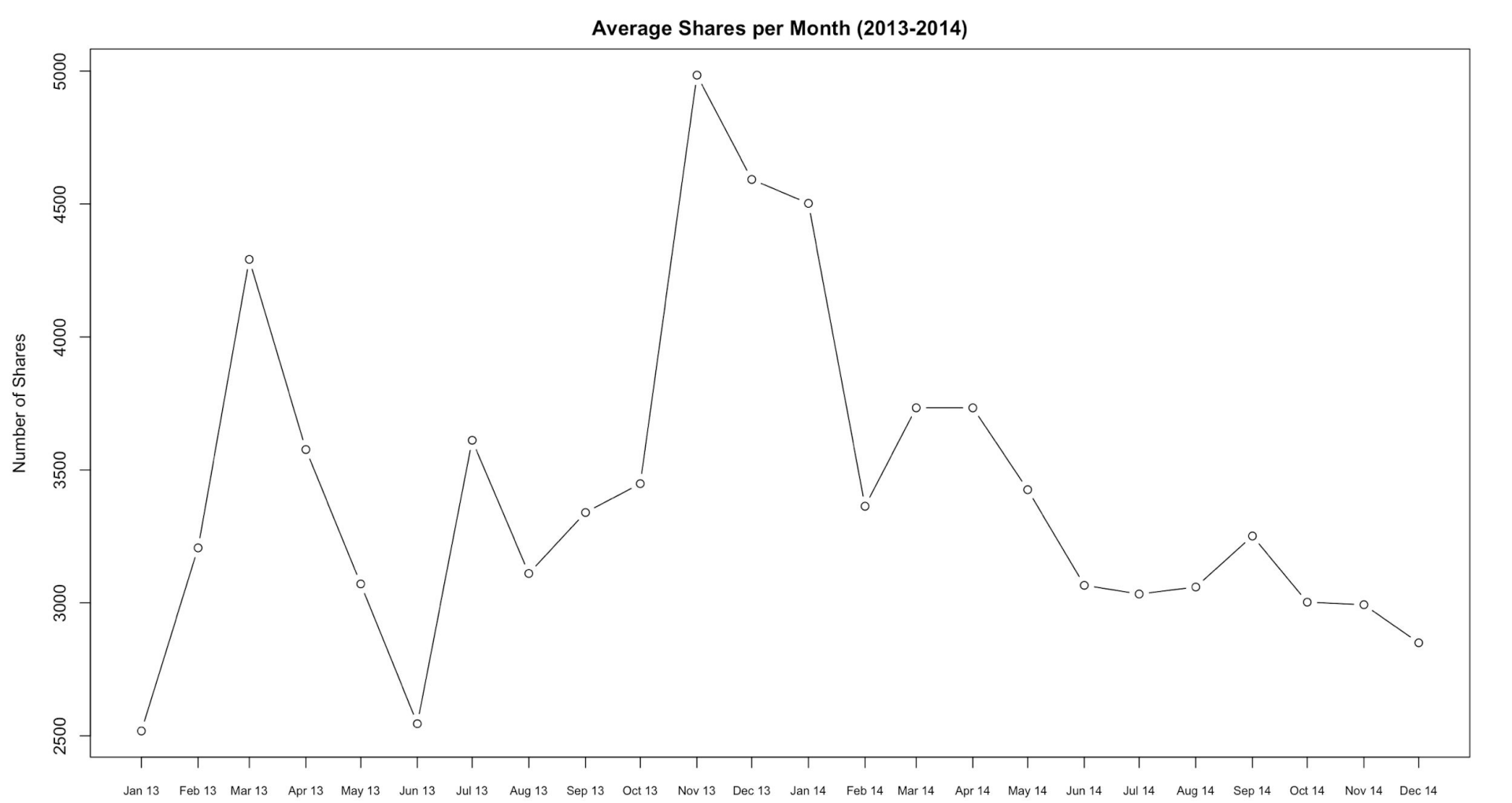
1. **Obtaining the content of each article and the day it was published**
2. **Ran countVectorizer on all the articles**
    i. Vocabulary = trending words for the month that article was published
    ii. Text = that articles content
    iii. Returned = array of article vs. total number of trending words in article
3. **Put into csv file to be used as feature**

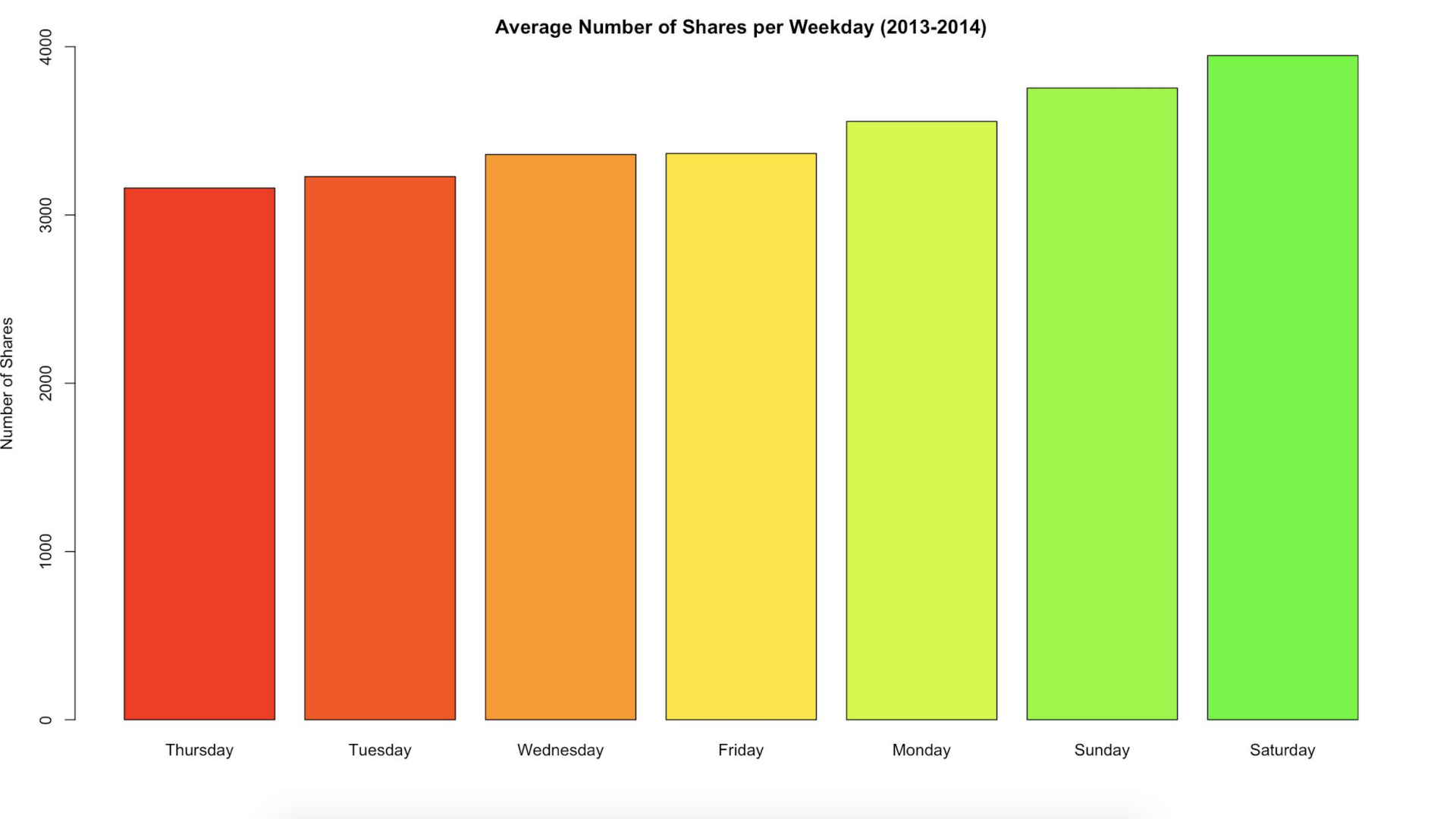
Automated Readability Index

- Designed to gauge the understandability of an article
- Approximate representation of age needed to comprehend the article
- Used python to parse article content for:
  - Number of characters
  - Number of words
  - Number of sentences

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$

ANALYSIS & PREDICTIVE MODELS


Average Shares per Month (2013-2014)
Number of Shares
5000
4500
4000
3500
3000
2500
Jan 13
Feb 13
Mar 13
Apr 13
May 13
Jun 13
Jul 13
Aug 13
Sep 13
Oct 13
Nov 13
Dec 13
Jan 14
Feb 14
Mar 14
Apr 14
May 14
Jun 14
Jul 14
Aug 14
Sep 14
Oct 14
Nov 14
Dec 14


Average Number of Shares per Weekday (2013-2014)
Number of Shares
0
1000
2000
3000
4000
Thursday
Tuesday
Wednesday
Friday
Monday
Sunday
Saturday

Naïve Bayes

0%	25%	50%	75%	100%
1	946	1400	2800	843300

Popularity Score (Labels)

Unpopular Article: less than 946 shares

[0]

Normal Article: between 946 and 2800 shares

[1]

Popular article: greater than 2800 shares

[2]

Naïve Bayes

38%

Overall accuracy of Naïve Bayes Classifier using all features

Target Values

[1. 1. 0. 2. 0. 2. 2. 0. 2. 2.]

Naive Bayes Prediction Values

[2. 2. 1. 1. 0. 2. 2. 0. 1. 0.]

Naïve Bayes

We performed Naïve Bayes on each separate feature category

Features	Accuracy
Visual Score	49.43%
Data Channel	39.71%
Day of Week	32.38%
Trending Word Score	50.68%
ARI score	50.32%

Random Forest

Popularity Score

Same as Naive Bayes for purposes of comparability

Unpopular Article: less than 957 shares [1]

Normal Article: between 957 and 2800 shares [2]

Popular article: greater than 2801 shares[3]

Random Forest

63.18%

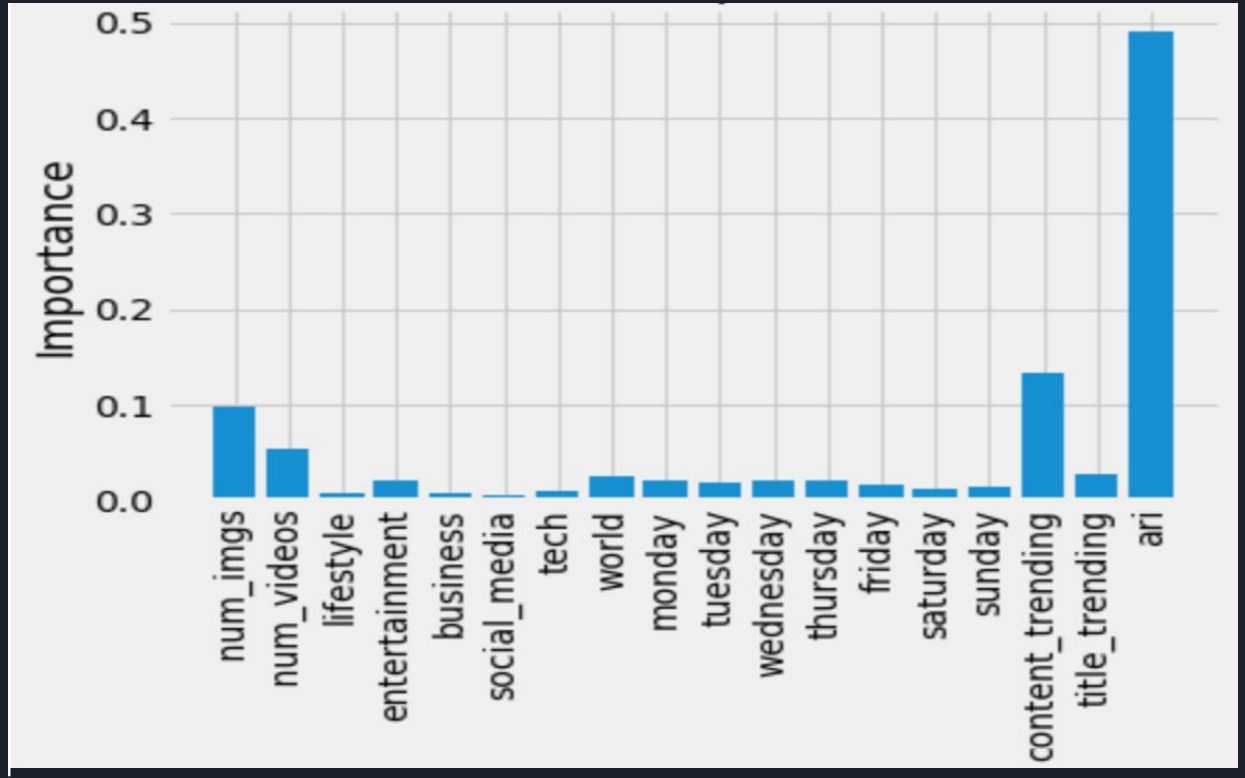
Overall accuracy of Random Forest Classifier using all features

Random Forest

We calculated the “importance” of each feature

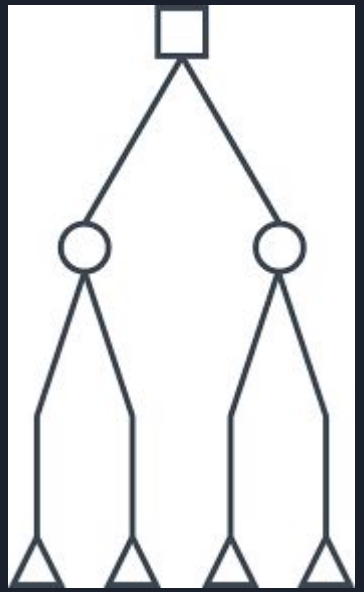
Features	Importance
ARI	49%
Trending words in content	13%
Number of images	10%
Number of videos	5%
Trending words in title	3%
Data channel	2% or 1%
Day of the week	2% or 1%

Feature Importance Graph



Random Forest - Our Model

Team 23's Forest



X 1000

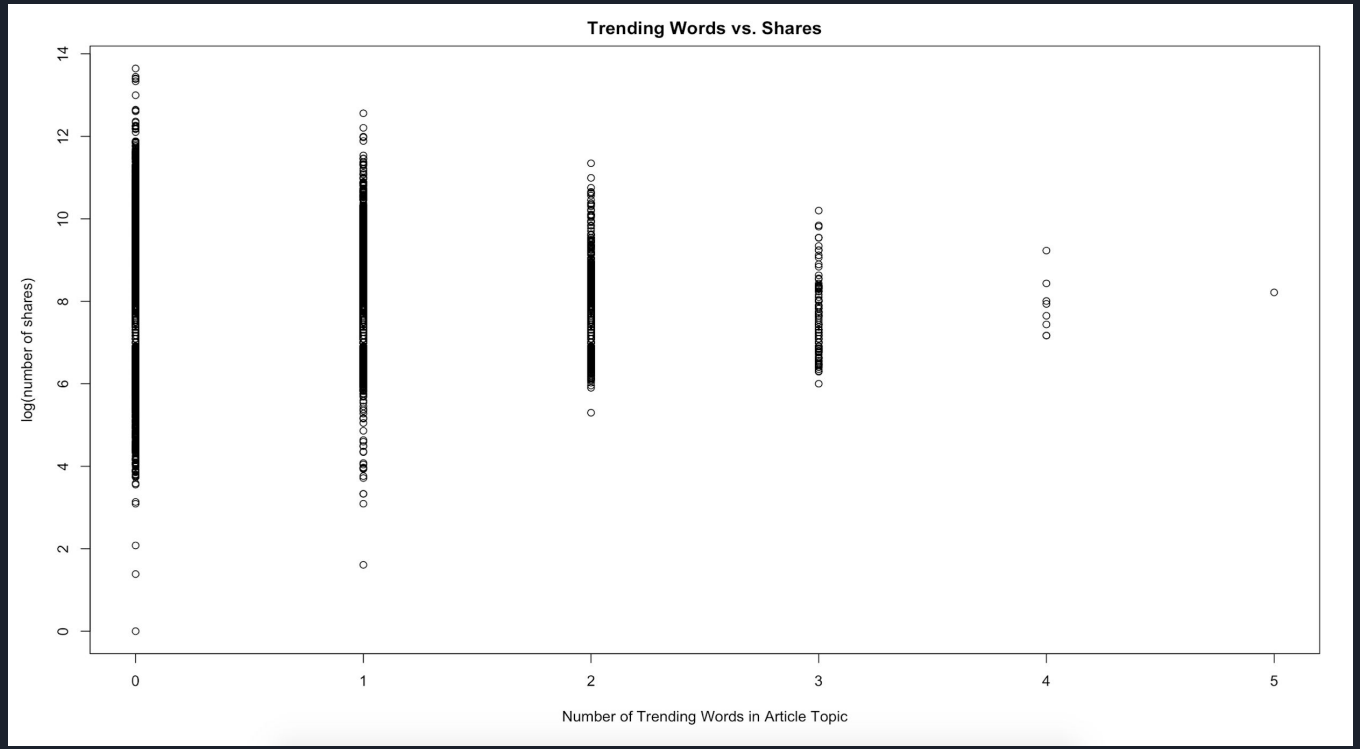
Number of estimators: 1000

- With 100 estimators, the accuracy only dropped by .05%
- Due to broadness of popularity index

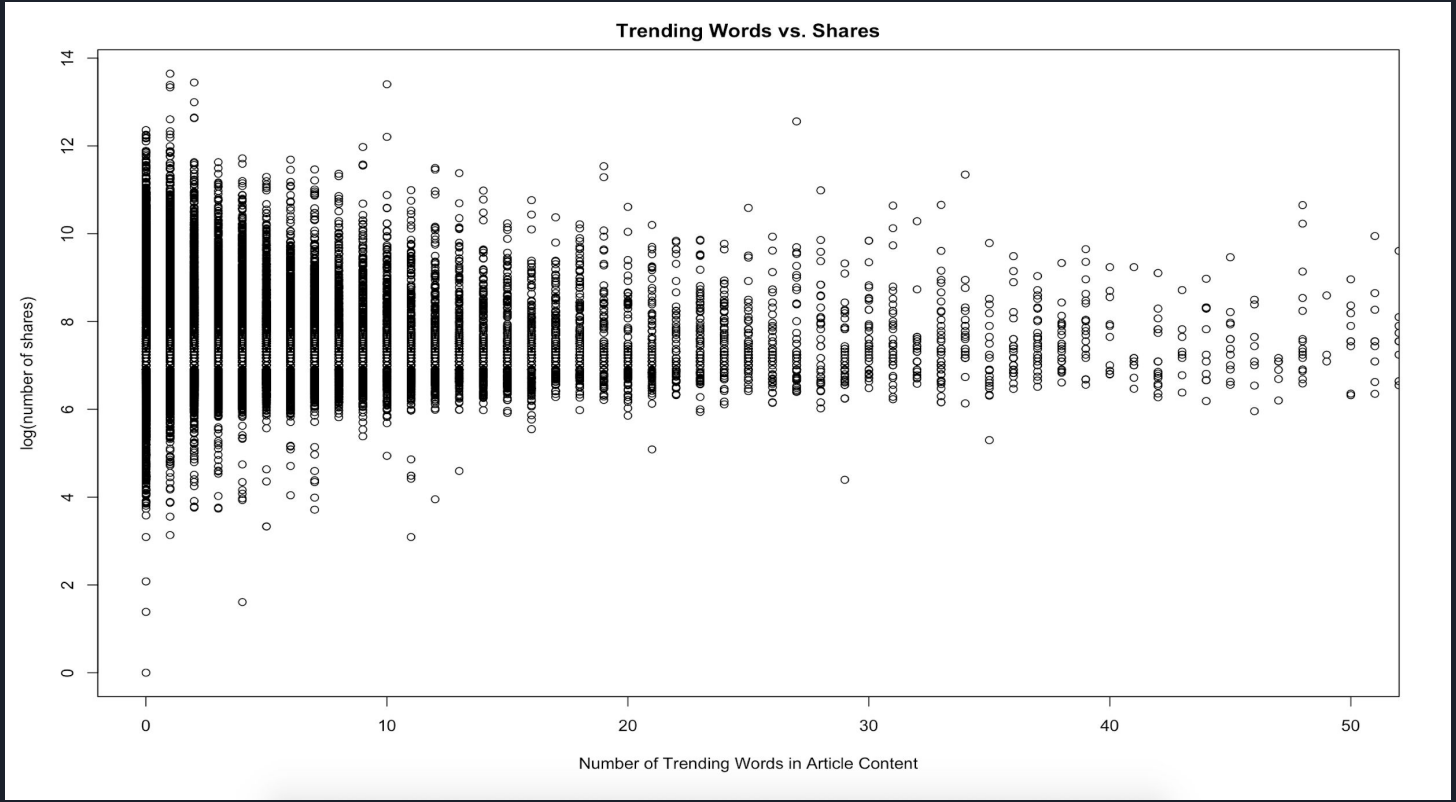
Test size: 0.25

Random state (default): 42

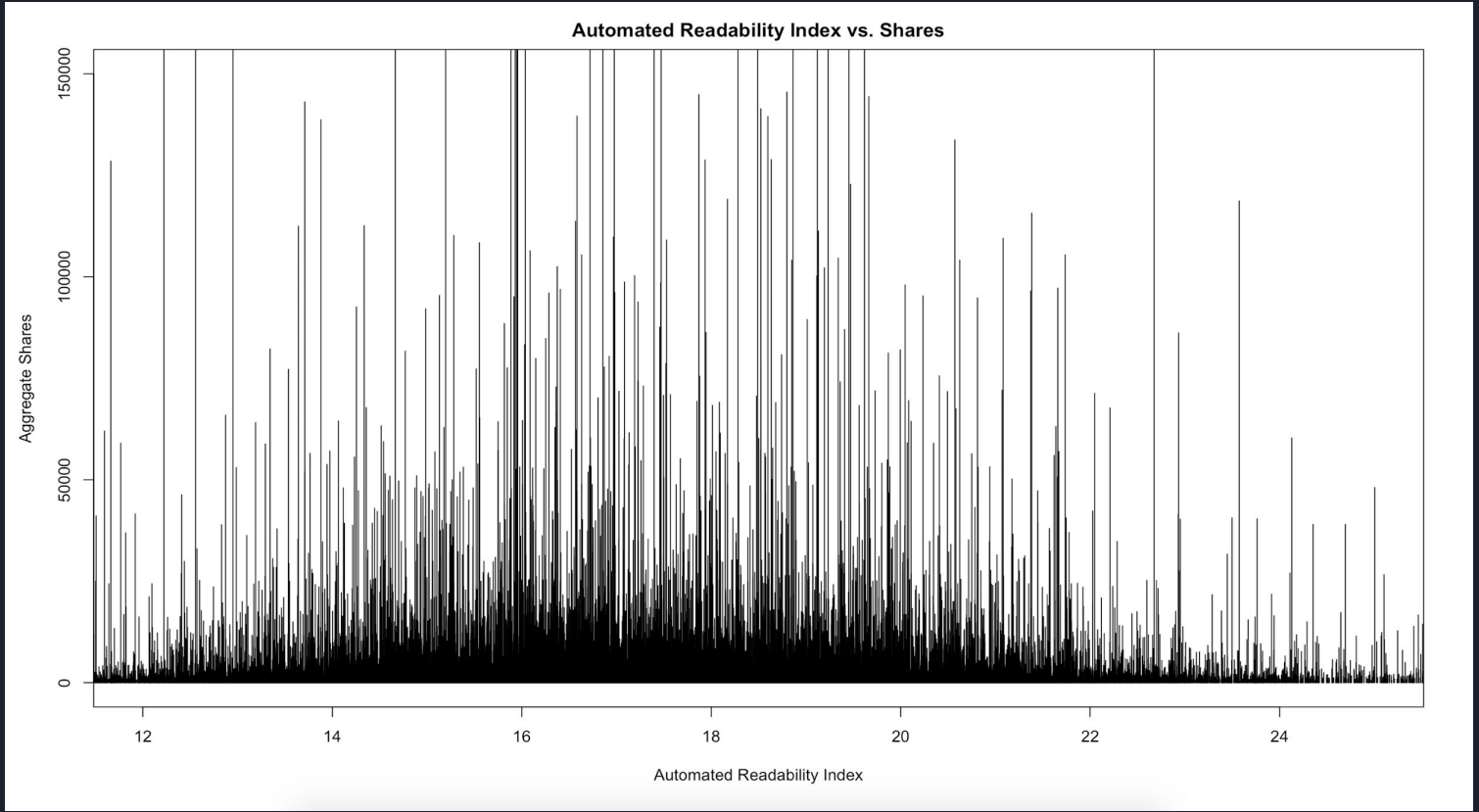
Regression


Trending Words vs. Shares
log(number of shares)
0
2
4
6
8
10
12
14
0
1
2
3
4
5
Number of Trending Words in Article Topic

Regression


Trending Words vs. Shares
log(number of shares)
0
2
4
6
8
10
12
14
0
10
20
30
40
50
Number of Trending Words in Article Content

Regression


Automated Readability Index vs. Shares
Aggregate Shares
0
50000
100000
150000
12
14
16
18
20
22
24
Automated Readability Index

Regression

6%

Adjusted R-squared value using all features (after several transformations)

Improvements

- Add more trending words
- Considering that people may not share certain/sensitive content
- Look at trending words per day
- Use twitter hashtags (couldn't before because earliest archive went was 2015)
- Other news sources
  - Different news source have different audiences (WSJ vs. Buzzfeed), affecting what type of content is most shared
  - Data wasn't available

Mashable
All That's New on the Web

WSJ

The Ideal Article

Readability Index - Between 16 and 22

images in the content - more is better

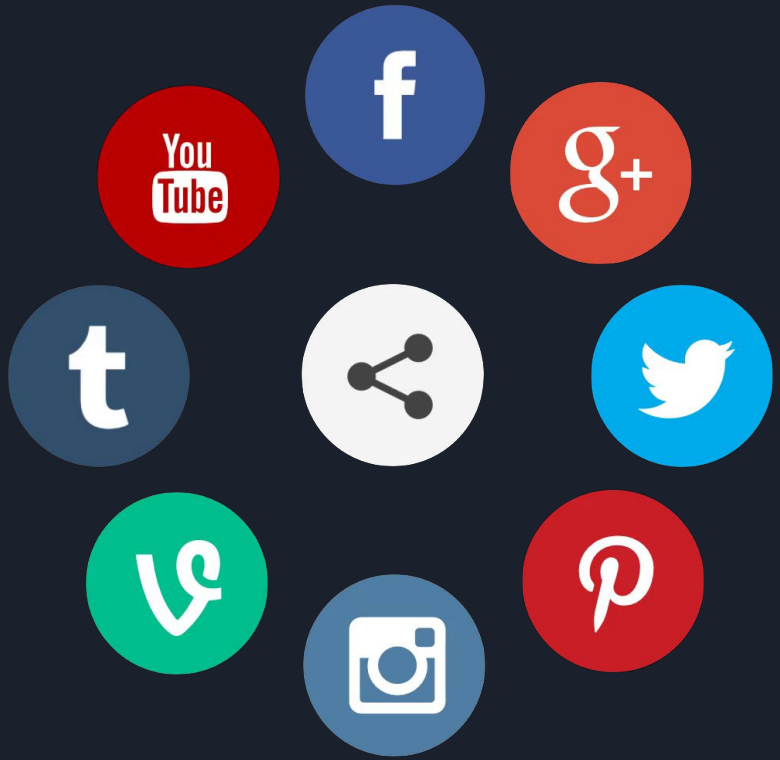
videos in the content - more is better

trending words in the article - more is better

trending words in the title - more is better

Section: Lifestyle

Day of the week : Saturday



Thank you!