Correlations

[Sophie Thesis Update Mon Jun 13]

Corpus

2.6 K stories remain*

Independent Variables (X)

- Mean sentiment of tweets about a story (where -1 is negative and 1 is positive)
  - (-1, 1) continuous
-
-

Dependent Value (Y)

- Number of tweets

Descriptive Statistics

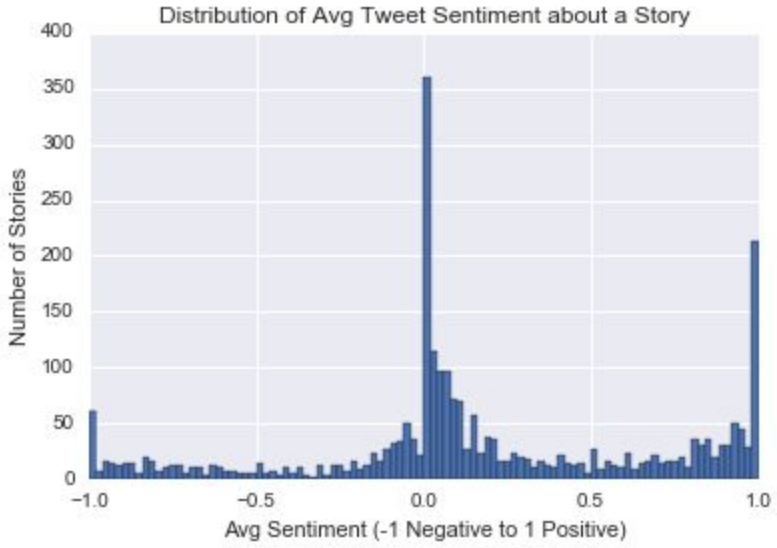
Number of Tweets

Mean Sentiment

Note: the tweets tend to be preoccupied mostly by the title of the story so it is not necessarily indicative of the reaction sentiment.

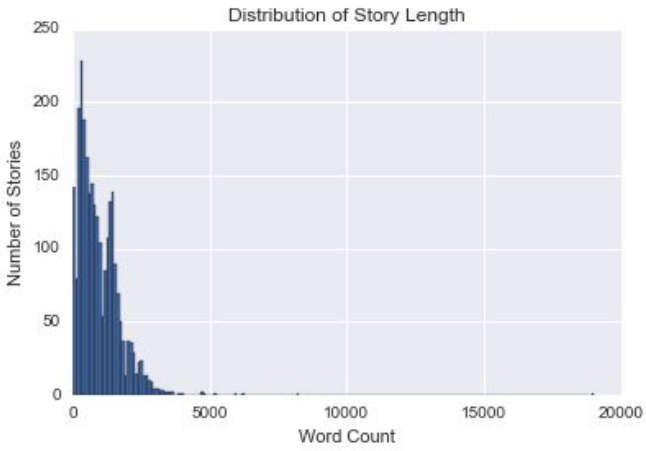
Range: -1 to 1

mean	0.172562
std	0.530127
min	-1.000000
25%	-0.025158
50%	0.071429
75%	0.615385
max	1.000000



Word Count

mean	968.708035
std	812.935257
min	0.000000
25%	392.500000
50%	797.000000
75%	1414.000000
max	19008.000000



Testing for Correlations: Simple Regressions
(no mixed effects or multivariates-- next step)

Story Length x Tweet Volume

Significant negative correlation between length of story and number of shares.
Possible explanation: people have short attention spans

```
Call:
lm(formula = num_tweets ~ wc, data = stories)

Residuals:
   Min     1Q  Median     3Q    Max
-40.32  -30.90  -20.22    5.99 1215.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.425959   1.818541  27.729  < 2e-16 ***
wc          -0.004180   0.001438  -2.907  0.00368 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.16 on 2648 degrees of freedom
Multiple R-squared:  0.003182,          Adjusted R-squared:  0.002805
F-statistic: 8.452 on 1 and 2648 DF,  p-value: 0.003677
```

Emotionality x Tweet Volume

Significant positive correlation between emotionality and number of shares.
Possible explanation: high emotional content gets more shares!

```
Call:
lm(formula = num_tweets ~ emotionality, data = stories)

Residuals:
   Min     1Q  Median     3Q    Max
-50.89  -30.98  -20.32    5.47 1214.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    40.349      2.685  15.027   <2e-16 ***
emotionality  305.229    122.427   2.493   0.0127 *
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Positivity X Tweet Volume

Significant negative correlation between positivity and number of shares.
Possible explanation: negative content gets more shares! This supports the hate-linking/negativity of the internet research Q/hypothesis.

lm(formula = num_tweets ~ positivity, data = stories)

Residuals:

Min	1Q	Median	3Q	Max
-46.27	-31.08	-20.38	5.41	1215.44

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	47.078	1.221	38.572	<2e-16	***
positivity	-281.010	139.546	-2.014	0.0441	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.21 on 2648 degrees of freedom
Multiple R-squared: 0.001529, Adjusted R-squared: 0.001152
F-statistic: 4.055 on 1 and 2648 DF, p-value: 0.04414

Number of unique URLS: 2726
2408 stories with full text in *our DB*
I filled in the rest-- 8 broken links, (dropped)
2718 remaining stories

NOTES

***Comparison to Milkman’s Corpus:**
6,956 articles total, timespan Aug 30-Nov 30 (2008) [**Us:** 6,496]
Subset of 2,566 manually coded

** “A computer program (LIWC) counted the number of positive and negative words in each article using a list of 7630 words classified as positive or negative by human readers (Pennebaker, Booth, and Francis 2007). We quantified positivity as the difference between the percentage of positive and negative words in an article. We quantified emotionality as the percentage of words that were classified as either positive or negative.”