Classifying Data-Driven Messages and Estimating the Effect of Data on Message Sharing

Justin Murphy

11/8/2017

Abstract. Does the use of formal data effect the likelihood that a message will be shared? To answer this question, we pursue a two-stage approach. First, we build a classifier able to distinguish data-driven messages from non-data-driven messages. We do this using two different approaches. First, we train a classifier using hand-coded instances of data-driven messages. Then, alternatively, we train a classifier using messages drawn from the #dataviz hashtag. Either approach allows us to estimate the degree to which any arbitrary message is data-driven. We then apply this classifier to a sample of tweets mentioning "Brexit," to study the relationship between data-driven messages and message sharing, controlling for follower counts. This short document is primarily a proof of concept. None of the models are fully optimal (e.g. the regression models are not the correct models, they are purely heuristic) and all data samples are relatively small. Nonetheless in this first exploration we find no substantial relationship between the estimated data-richness of a message and its likelihood of being shared.

Study 1: Training a classifier on data-driven messages by data journalists

I scraped between 200 and 300 tweets each from 20 diverse data journalists with the largest followings on Twitter. This produced about 2600 status updates in English. I then manually coded this sample for "data-driven" status updates.

Our sample contains 461 status updates that were found by qualitative analysis to be data-driven (17.72%). Note that this is not intended to represent the prevalence of data-driven messages among data journalists: For this stage of the research, purposely sought to increase the number of data-driven messages through different means, e.g., filtering messages by data-oriented terms.

Examples of hand-coded data-driven status updates (before cleaning)

"That surge in full: -0.04% vs $; +0.25% vs ___ https://t.co/AckIl5M9Ck"

"This is astonishing: FT points out that HUGOBOSS has submitted improbable gender pay gap figures (of 0%!) and Boss revises___"

"This is an amazing chart, by my great colleague LaurenLeatherby: Amazon's growth is blotting out the rest of US retail… ___"

"Our long-running UK economic dashboard makes its print debut https://t.co/3WqIvdBNlF https://t.co/dZV4PKqLVB"

"Nice to see BillyEhrenberg and alekswis___ gender pay gap data-crunching on the front today:___ https://t.co/1mgG4K4bIl"

Are data-driven messages more likely to be shared (within the sample)?

The answer is no, not really. Below find simple bivariate visualization (Figure 1) and a regression looking only at original messages (non-RTs, and non-replies) and controlling for followings (Figure 2).

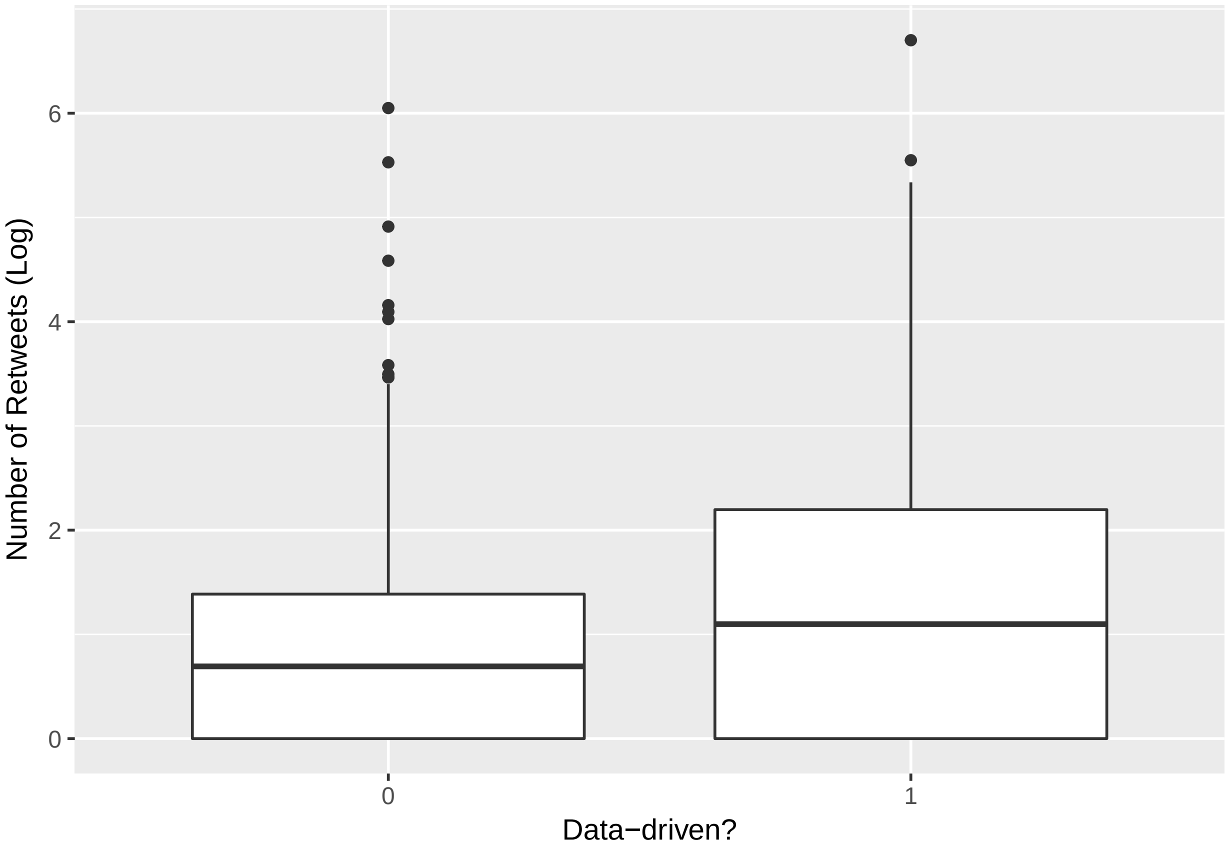

Figure 1: Data-driven tweets and Retweets

```
## Model:
##
## Call:
## z5$zelig(formula = retweet_count ~ followers_count + data, data = dj)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5016  -1.2335  -0.6218  -0.1621   6.6162
##
## Coefficients:
##                     Estimate   Std. Error z value            Pr(>|z|)
## (Intercept)      1.721339593  0.081163678  21.208 <0.0000000000000002
## followers_count -0.000024895  0.000002369 -10.507 <0.0000000000000002
## data             1.288065394  0.141354434   9.112 <0.0000000000000002
##
## (Dispersion parameter for Negative Binomial(0.266) family taken to be 1)
##
##     Null deviance: 1369.8  on 1207  degrees of freedom
## Residual deviance: 1148.2  on 1205  degrees of freedom
## AIC: 5444.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2660
##           Std. Err.:  0.0131
##
##  2 x log-likelihood:  -5436.2600
## Next step: Use 'setx' method
```

A classifier for data-driven messages

Can the hand-coded data-driven messages be used to classify a larger sample into data-driven and non-data-driven tweets?

I used the doc2vec algorithm to learn the difference between data-driven and non-data-driven messages. I split the data 70/30 into training/test data. I used logistic regression with cross-validation to build a classifier capable of predicting whether a message is data-driven or not. The model accuracy rate was about 75%.

I then scraped a random sample of tweets mentioning Brexit, and used the classifier to predict the probability a given message is data-driven. I then used these predicted values to estimate whether data-driven messages are shared more than non-data-driven messages. I did this with simple regression, looking only at non-RTs and controlling for follower counts.

Again there is not much evidence that data-driven messages predict number of retweets. I should note this linear regression model is certainly not the correct model, it is only a first exploratory look.

```
## Model:
##
## Call:
## z5$zelig(formula = retweet_count ~ followers_count + datacont,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -22.84   -2.38   -2.38   -1.39 1363.35
##
## Coefficients:
##                      Estimate    Std. Error t value      Pr(>|t|)
## (Intercept)      2.4473446952  0.5496065476   4.453 0.00000855898
## followers_count  0.0000036762  0.0000006333   5.804 0.00000000665
## datacont        -0.4673532277  3.0213768958  -0.155         0.877
##
## Residual standard error: 22.41 on 10526 degrees of freedom
## Multiple R-squared:  0.003194,   Adjusted R-squared:  0.003004
## F-statistic: 16.86 on 2 and 10526 DF,  p-value: 0.00000004885
##
## Statistical Warning: The GIM test suggests this model is misspecified
##  (based on comparisons between classical and robust SE's; see http://j.mp/GIMtest).
##  We suggest you run diagnostics to ascertain the cause, respecify the model
##  and run it again.
##
## Next step: Use 'setx' method
```

Approach 2: train a classifier using #dataviz messages

An alternative appraoch to hand-coding data-driven messages is to use a sample of tweets already known to represent data-driven messages. One candidate for this is the set of status updates containing the hashtag #dataviz. Most of these message are, in some sense, data-driven. The advantage is convenience, and a large sample. The drawback is that #dataviz also contains a lot of non-data-driven content, especially marketing and tutorials. Still we might plausibly think the language from #dataviz is a good proxy for data-driven messages in other domains.

I followed the same approach as above. The classifier had a higher accuracy, of about 87%.

The results are similar: no appreciable relationship between RTs and data-driven messages.

Approach 3: Explicit references to data, Trump-Russia and Climate Change

```
## Model:
##
## Call:
## z5$zelig(formula = formula, data = data, by = by)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135.53   -6.03   -6.01   -6.01 2131.28
##
## Coefficients:
##                     Estimate   Std. Error t value         Pr(>|t|)
## (Intercept)     6.0118819692 0.8646278159   6.953 0.00000000000391
## followers_count 0.0000047326 0.0000007448   6.354 0.00000000022323
## data1           5.1613404560 2.6634503343   1.938           0.0527
##
## Residual standard error: 66.56 on 6644 degrees of freedom
## Multiple R-squared:  0.006543,   Adjusted R-squared:  0.006244
## F-statistic: 21.88 on 2 and 6644 DF,  p-value: 0.0000000003379
##
```

```
## Statistical Warning: The GIM test suggests this model is misspecified
##  (based on comparisons between classical and robust SE's; see http://j.mp/GIMtest).
##  We suggest you run diagnostics to ascertain the cause, respecify the model
##  and run it again.
##
## Next step: Use 'setx' method
```

```
## Model:
##
## Call:
## z5$zelig(formula = formula, data = data, by = by)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.38   -1.84   -1.83   -1.83  880.73
##
## Coefficients:
##                     Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)     1.8267964297 0.1204774920  15.163 <0.0000000000000002
## followers_count 0.0000065243 0.0000002537  25.712 <0.0000000000000002
## data1           2.6547597303 1.1995050981   2.213              0.0269
##
## Residual standard error: 18.18 on 23061 degrees of freedom
## Multiple R-squared:  0.02807,    Adjusted R-squared:  0.02798
## F-statistic:   333 on 2 and 23061 DF,  p-value: < 0.00000000000000022
##
## Statistical Warning: The GIM test suggests this model is misspecified
##  (based on comparisons between classical and robust SE's; see http://j.mp/GIMtest).
##  We suggest you run diagnostics to ascertain the cause, respecify the model
##  and run it again.
##
## Next step: Use 'setx' method
```

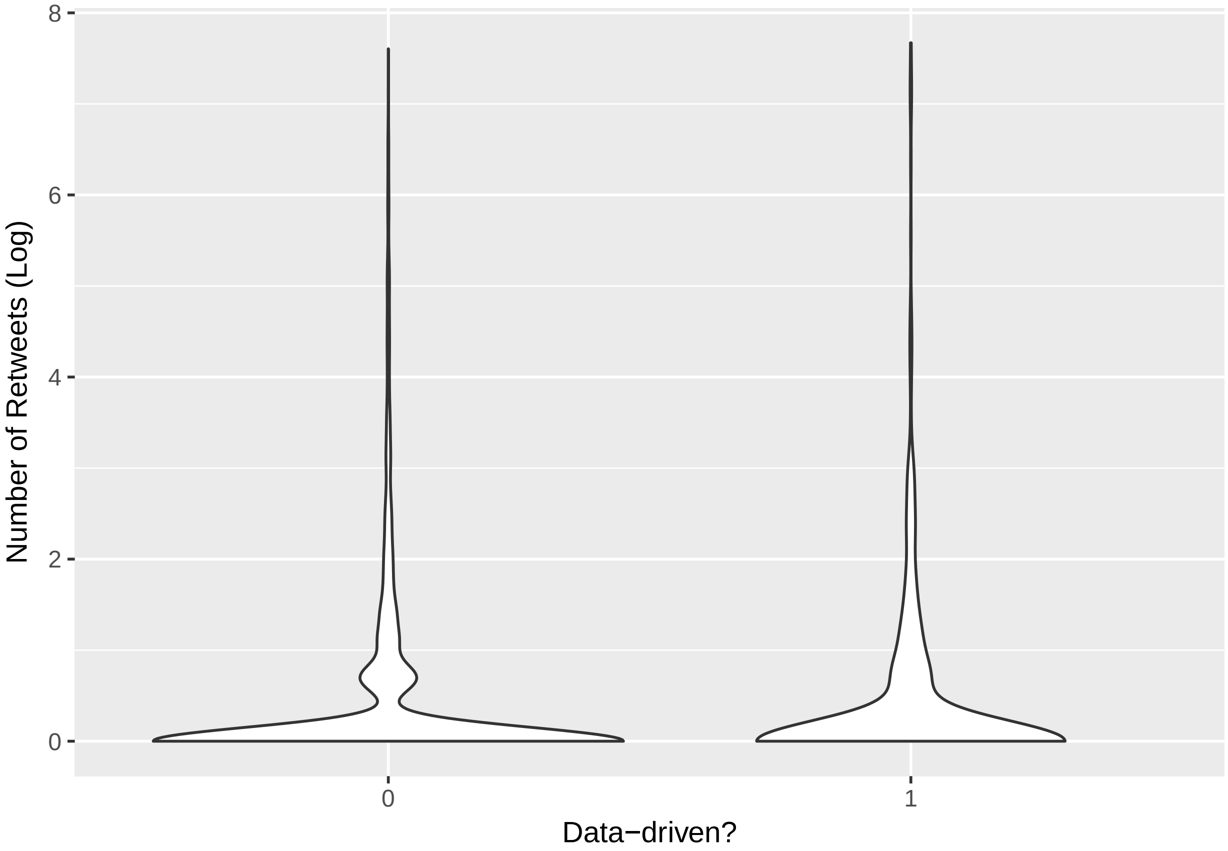

Figure 2: Estimated effect of data mentions on number of retweets (Trump/Russia)

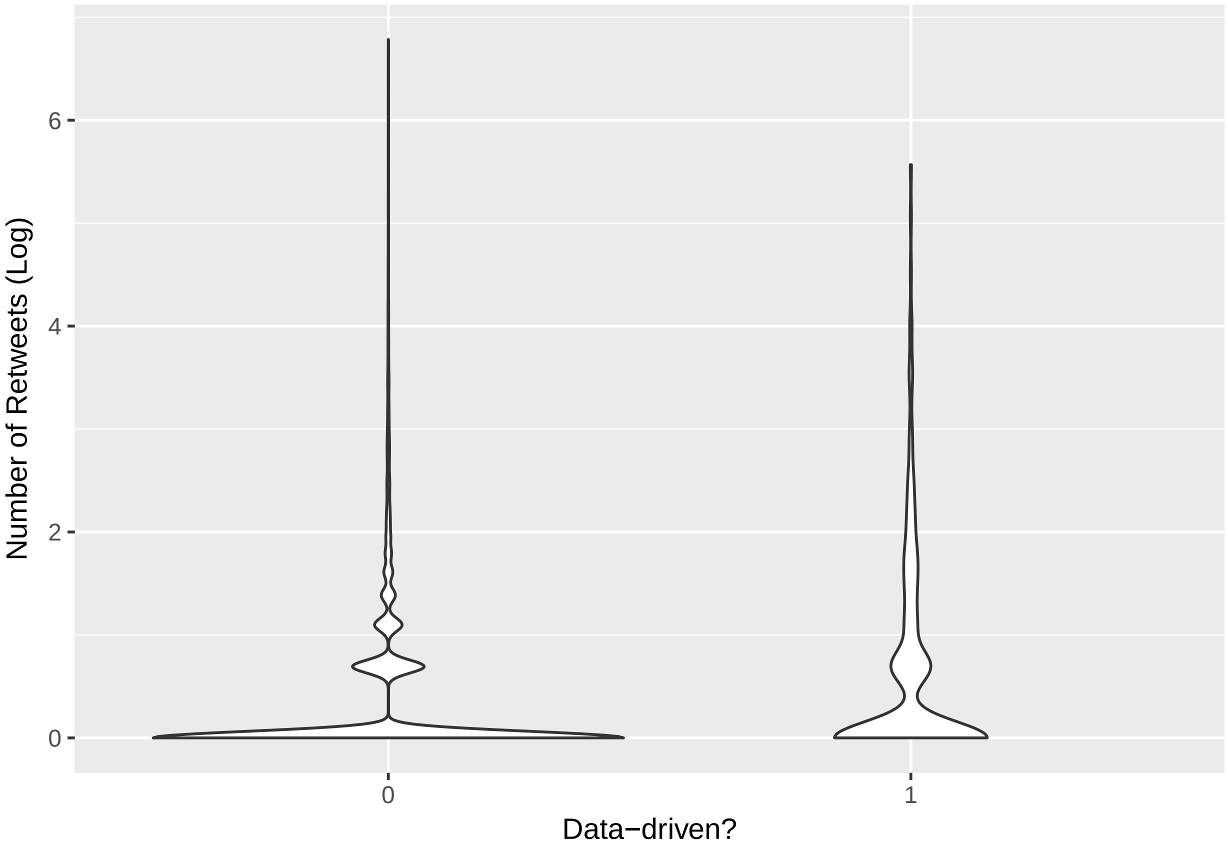

Figure 3: Estimated effect of data mentions on number of retweets (climate change)