Headlines as networked language

A study of content and audience across 73 million links on Twitter

David McClure

January 25, 2019

Abstract

Abstract.

1 Introduction

Imagine that someone showed you a headline from a news article, but in complete isolation, stripped of all context – just a sequence of words. All you were told is that the headline came from either the New York Times or Fox, and you were asked to guess which one. In some cases, this might be fairly easy. For example, if it's a recipe – we might remember that The New York Times has a large cooking section:

- Chicken Thighs With Cumin, Cayenne and Citrus

Or, if it's about (both) New York baseball teams:

- In Early Going, the Yankees Steal the Mets' Thunder

(Though, of course, Fox also does plenty of sports reporting.) Meanwhile, we might associate a story about MS 13 with Fox, to the extent that right-leaning outlets have focused attention on immigration and crime:

- East Coast MS 13 gang leader admits racketeering conspiracy

But, other things might be significantly harder. For example – one of these headlines came from Fox, the other from The New York Times:

- Zambia's 1st female fighter pilot says she "doesn't feel like a woman" in her job
- 4 freed from Thailand cave, but rescuers face "war with water and time" to get to others

Here, to my eye, there aren't any obvious "tells" – there are certainly things that might seem to tip the scaled one direction or the other, but it's not clear.[1] In trying to guess where the headline came from, we'd have to bring to bear a wide set of intuitions about what might be thought of as the "voice" of the outlet – the set of issues, locations, people that the outlet tends to focus on. And, beyond the raw content of what's being – *how* it's being covered, the style, intonation, attitude, affect. Trying to guess the outlet, in other words, would force us to formalize a kind of mental model about precisely how the two outlets are similar or different.

It also, indirectly, gives a way to reason about the degree to which they're similar or different. Now, imagine that instead of just doing this once, we did it for 100 headlines, and counted up the number

[1] The answer – NYT, Fox.

of correct guesses. We'd likely do better than random – but how much better? 60%, 70%, 95%? How differentiable are NYT and Fox, at a purely linguistic level? And, how does this compare to other pairs of outlets? What if we took headlines from NYT and CNN, instead of NYT and Fox, for example, and repeated the experiment. We might guess that NYT and CNN are more similar, and thus harder to tell apart. But, how true is this, exactly? Say we got 80 headlines right when guessing between NYT and Fox comparison – would we get, perhaps, 70 right for NYT and CNN? Or 60, 55? In a rough sense, we could start to reason about the relative proximities between different pairs of outlets.

Of course, doing this manually, it would be hard to scale beyond a few outlets and a couple hundred headlines. But – what if we could do this at a much larger scale, across dozens of different media organizations and millions of headlines? This thesis explores this question as a *language engineering task*, working with a corpus of 73,198,274 tweets harvested from the Decahose over an 18 month period – to what degree is it possible to train machine learning models to differentiate between headlines produced by different news outlets? Unlike other studies that have explored the tractability of language inference tasks on news data, though, goal here isn't to solve a applied engineering task (flagging "clickbait" headlines, optimizing click-through rates) – instead, we use the predictive models as a descriptive and interpretive tool. By training models to differentiate between content from different sources, we can then examine the representations that are induced by the models, making it possible to "map" headlines as a kind of conceptual space.

Most fundamentally, beyond this ability to "read" headlines at a scale that would otherwise be impossible – these models make it possible to start filling in a kind of ground-truthed understanding of the degree to which different outlets are similar and different, where we might otherwise have to rely on intuition and anecdote. Modeling a complete matrix of pairwise similarities across 15 major media organizations, we can construct a kind of "content graph," a fully-connected representation how, and to what degree, outlets produce (dis)similar headlines:

Once we have this empirical view of the linguistic proximities among the outlets, though, there's a way in which this immediately generates a new set of questions. Most of which, really, are some version of just one – where is this graph *not what we would expect*? To return to the analogy of voice – are there outlets that are "speaking" in ways that don't map onto assumptions about the role that they play in a broader political, commercial, or historical context?

The second half of the thesis takes up this question, and explores the degree to which these linguistic relationships correspond (and don't correspond) to two different backdrops. First – we explore the degree to which these similarities at the level of content align with similarities at the level of audience. Parallel to the "content graph" defined by the headline similarities, can imagine an "audience graph" that represents similarities at the level of the people who read and distribute content from each pair of outlets.

This question of the interaction between content and audience cuts to the core of a wide range of recent discussions about processes of fragmentation and polarization in the news ecosystem. For example, the widely popularized notions of "filter bubbles" and "echo chambers," where the concern is that these two graphs are tightly connected in a self-reinforcing (and socially problematic) way – media sources that are highly differentiated at the level of content allow people with different perspectives to self-select into highly differentiated information ecosystems, which, in a vicious circle, then aggravates the underlying social and ideological fragmentation. Stark divisions at the level of content get mirrored by stark differences in audience, which in turn incentivize even more differentiated content, etc.

The graph above, though, suggests there might be places where this alignment between content and audience breaks down. For example, it looks like headlines from Fox are about as similar to headlines from AP as from Breitbart; whereas, we might guess that the audience overlap between Fox and Breitbart would be significantly higher. This suggests a rich set of questions – are there outlets that "sound" similar to one another, but that have highly non-overlapping audiences? Or, zooming back to the complete ecosystem of media organizations, can we identify outlets that act as "mediators" or "bridges" between different subsets of audiences? For example – are there outlets that produce content that "sounds like" content produced by left-leaning outlets, but that have relatively high audience overlap with right-leaning outlets – that is, outlets that produce content that has the effect of exposing right-leaning readers to content that's more typical of left-leaning news feeds? Or vice versa?

These incongruities, misalignments, places where the language graph doesn't map easily onto the underlying social graph – do they exist at all? If so – what do they look like? What kind of content does this? Looking beyond this study, which is fundamentally descriptive – if we're interested in exploring the degree to which different topics, styles, idioms, and ways of speaking might have the effect of working against patterns of fragmentation in media consumption habits, these sites of mismatch could be of particular interest. They would represent, in a sense, naturally-occurring short-circuits between otherwise entrenched associations between particular audiences and particular types of content, places where audiences move (or get moved) out of their typical "lanes."

Second, and last – building on this question of movement or change, we explore the degree to which the matrix of linguistic similarities has evolved over historical time – the degree to which it does and doesn't correspond to previous versions of itself. Instead of rolling up the entire 18-month window of data into a single, synchronic unit, we analyze the set of 73 million links as a chronological sequence, covering roughly the first two years of the Trump presidency. Are the linguistic similarities between outlets basically stable, or have there been significant changes over this period? Do outlets drift away from each other, towards each other, back and forth? How fixed or flexible are the major organizations in the media landscape, both in terms of the type of content that they produce and the audiences that read it? And, in the same way that we could explore the overall correlation (and lack thereof) between content and audience – to what extent do changes in content correspond to changes in audience, when spooled out over historical time? If Huffington Post starts producing headlines that sound more like The Hill – do the audiences also start to converge?

In summary – first we quantify the level of similarity between headlines produced by different outlets, and then attempt to contextualize these differences by exploring (1) the degree to which they have evolved over time and (2) the degree to which they align with the composition of the audiences that interact with the content. We find:

1. Modeled as a text classification task, headlines from 15 major US media sources highly differentiable, even after an aggressive cleaning process that strips out any kind of "paratext" in the headline that directly or indirectly reveals the source. (For example, "... – CNN Video" or "AP Breaking News: ...") The strongest model is a bidirectional LSTM, which gets to 41% accuracy in a 15-label model, where random would be 7%. Though, standard non-neural baselines also perform very well – SVM and logistic regression models over ngram count features are competitive at 36-38% accuracy (and beat the weaker neural models), suggesting that a majority of the differentiability among outlets comes from which words and phrases are being used – roughly, what might be thought of as "topic" or "content" – and less from the composition or sequencing of words, which can be modeled with more power by the recurrent models.

2. Pulling sentence embeddings out of the top layer of the LSTM, we can explore the structure of the linguistic space in terms of the high-dimensional representations learned by the model, both in terms of the relationships among different outlets are the internal structure

of the headlines within the outlets themselves. This reveals significant differences in what might be thought of as the "shape" of different outlets in the high-dimensional space. Specifically, we focus on what we call the "diameter" of the embedding family for each outlet – the average cosine distance between randomly sampled pairs of headlines – which, we argue, roughly captures the degree to which outlets are narrowly focused (low diameter) or more broad and varied (high diameter). Second, we also explore the clustering structure of headlines within outlets, the degree to which content from an outlet can be cleanly decomposed into a set of categories, verticals, or "desks." On both metrics, we find significant variation, with BuzzFeed, Bloomberg, and Daily Kos registering the lowest "diameters" and smallest cluster counts (most focused), and CNN, The Washington Post, and Fox showing the highest diameters / largest cluster counts (most broad, varied, diverse).

3. Aggregating over the full set of individual articles in each outlet, we can use the classifiers to model a complete "content graph" among the 15 organizations, a fully-connected graph of pairwise similarity scores, which are extracted from the behavior of the classifiers in a collection of different ways – pairwise accuracy scores, probability mass correlations in multi-label models, and confusion counts. With this content graph in hand, we then compare to the underlying "audience graph," derived from correlations in the users who tweet links from each pair of outlets. From these two graphs, for each outlet we can construct two normalized rankings of similarity with all other outlets, one based on headline similarity, and the other based on audience similarity. We then explore the correlation of these two rankings on a per-outlet level – that is, the degree to which an outlets edge weights in the "content graph" are similar to its edge weights in the "audience graph."

   In general, we find that these two sets of weights tend to be positively correlated, often highly so – in general, outlet A "sounds like" outlet B, it also has a high level of audience overlap. But, there is a wide range in the strength of this correlation, and there are some strong exceptions – namely The Associated Press and The Hill, both of which show a significant level of misalignment between their content and audience graphs, with correlations of roughly 0. For AP, this is driven by Fox, which both syndicates a sizable number of articles from AP and also produces in-house content that closely resembles AP. But, by and large, this AP-style content is dramatically different from the rest of the coverage produced by Fox, and represents arguably the most significant content-audience mismatch in the dataset. After the AP (which is anomalous in that it's a wire service, and AP content gets literally "duplicated" by other outlets), The Hill seems to be the closest thing to a naturally "bridging" or "bipartisan" outlet – it has high audience overlap with left-leaning outlets, but high headline similarity with the right-leaning Daily Caller and Breitbart; but also with the left-leaning MSNBC.

4. We then explore the evolution of these relationships over historical time – instead of treating the corpora as a monolithic bundle of data, the 2-year data window is broken into 100 temporal bins, and a 10-bin rolling window is moved across this space, producing 91 10-bin windows. In each of these windows, we train fully independent A-versus-B models on each unique pair of outlets, and then track changes in these accuracies over time. This reveals significant changes in the structure of the content graph in the last two years. First, The Huffington Post and BuzzFeed have moved away from each other, due to mirror-image changes – HuffPo has moved away from "clickbait" content (advice colums, diet recommendations) and towards more narrowly-focused political reporting, similar to The Hill, DailyKos, and CNN; whereas BuzzFeed doubled down on clickbait content (specifically, a certain type of "quiz" article), pushing down the proportion of content from BuzzFeed that consists of serious political and investigative reporting. Second, a cohort of politically-focused outlets have become more similar – The Hill, The Daily Kos, CNN, AP, NYT, WaPo, NPR – suggesting

a tendency towards convergence in political coverage since a high-water mark of differentiability in the spring of 2017, in the first months of the Trump administration. Last – Fox has consistently moved away from almost all other outlets, including other right-leaning outlets like Breitbart and The Daily Caller; it has become much more highly differentiated *individually*, compared to everything else. By examining headlines that typify the overall movement of Fox in the high-dimensional linguistic space modeled by the neural classifier, it appears that this shift is driven by a move in the direction of a highly sensationalized, tabloid style of headline, many of which involve violent crime and socially-charged political issues.

2 Headlines on Twitter

To explore this interaction between content and audience, we need what might be thought of as a "networked" corpus – a collection of individual pieces of content that can be linked directly to specific patterns of engagement and distribution. With this goal in mind, this study built around a very particular data set – posts on Twitter that contain links to news articles, which we extract from the Decahose, a 10% sample of all activity on Twitter. For example, if someone posts a tweet with a link to a New York Times article, Twitter identifies the link and provides structured information about it in the JSON payload that appears in the Decahose:

```
"urls": [
    {
       "url": "https://t.co/MOZgWRzhRK",
       "expanded_url": "https://www.nytimes.com/2017/09/27/us/politics/navy-
           orders-safety-operational-standards.html",
       "display_url": "nytimes.com/2017/09/27/us…/",
       "indices": [66, 89]
    }
 ],
```

We can then parse the raw URL string and extract the registered domain – here, `nytimes.com` – which makes it possible to associated URLs with different media organizations.

Critically, though, our lab also has access to the "Enhanced URLs" add-on product available through Gnip, the vendor that distributes data from Twitter. The "Enhanced URLs" metadata provides two pieces of information that are essential to this study – first, the "expanded" URL, which is formed by following redirects from the raw URL that appears on Twitter. This valuable here because a sizable percentage of links on Twitter are shortened by services like `bit.ly`, meaning that they have to be "unrolled" in order to identify out the real domain – for example, `nytimes.com` instead of `bit.ly`. In theory, we could do this ourselves; but, especially when working with data that's more than a few weeks old, this can be difficult. Many shorteners like `bit.ly` automatically expire the links, meaning that, for example, if we tried to expand a link from spring of 2017, it would likely fail.

Second, and most important, "Enhanced URLs" also provides scraped copies of the Open Graph `title` and `description` tags of the page that the link points to. At a product level, Twitter uses the title and description in the link "preview boxes" for the article that automatically get displayed with tweets that contain links.