Automatic Analysis of Television News: Media, People, Framing and Bias

Carlos Castillo, Gianmarco De Francisci Morales, Marcelo Mendoza, Nasir Khan
Qatar Computing Research Institute, Yahoo Labs, Universidad Federico Santa Maria, Al Jazeera
chato@acm.org, gdfm@yahoo-inc.com, mmendoza@inf.utfsm.cl, nasir.khan@aljazeera.net

Television is a dominant source of news today, wielding an enormous influence over many aspects of our life. We analyze the closed captions of newscasts, which are provided by the news networks themselves. By using these streams of text, we study how to characterize each news network, each person-type named entity mentioned in the news (*newsmaker*), and the relationship between news networks and newsmakers (e.g., the biases of networks in the coverage of news related to a person).

We propose a pipeline of tools to process the input stream of data. Our pipeline filters out non-news programs, segments the captions into sentences, detects named entities (specifically people), applies a part-of-speech tagger to find words and qualifiers used together with each entity, and labels automatically each sentence with an overall sentiment score. These tools extract a set of measurable dimensions from the text, which we employ to characterize news providers, newsmakers, and framing in their coverage.

Data and tools. We use closed captions provided by a software company that develops *second-screen experiences*, i.e., applications that display extra information about a TV show on a smartphone or tablet. Our dataset consist of all the closed captions from January 2012 to June 2012 for about 140 channels. Each channel is mapped to the conventional name of its network; for some networks we have data from the East Coast and West Coast of the US, and denote them as, e.g., `ABC-w` and `ABC-e`. The three major news networks according to the number of self-declared regular viewers are `Fox News` (21%), `CNN` (16%) and `MSNBC` (11%).

We obtain in advance a television schedule with the title of each program, the channel on which it will be aired, the beginning and ending times, and a category and sub-category. We consider all the programs of type **newscast**, and four sub-categories: `general` news, `sports` news, `entertainment` news and `business` news. Given that one channel may have more than one type of program , we define a **news provider** as a combination of network and genre, e.g., `Fox Business[biz]` and `Fox Business[gen]`.

We match the processed captions to recent news stories, which are obtained from a major online news aggregator. Captions are matched in the same genre, e.g., sentences in `sports` are matched to online news in the `sports` section of the news aggregator. News in the website that are older than three days are ignored. More details in [1].

Style. We use part-of-speech and dependency tags to analyze the differences in style among providers. We represent each provider as a distribution over linguistic categories (e.g., number of verbs, number of adjectives), and apply hierarchical agglomerative clustering with euclidean distance to this representation. Figure 1 shows the resulting clustering of the top-30 providers with most mentions.

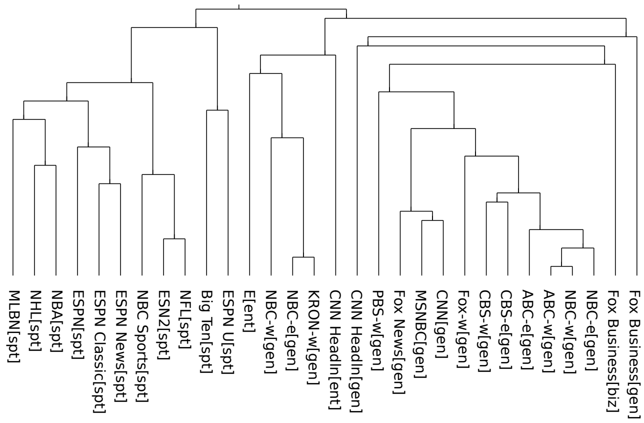


Figure 1: Hierarchical clustering of providers based on the prevalence of different linguistic classes.

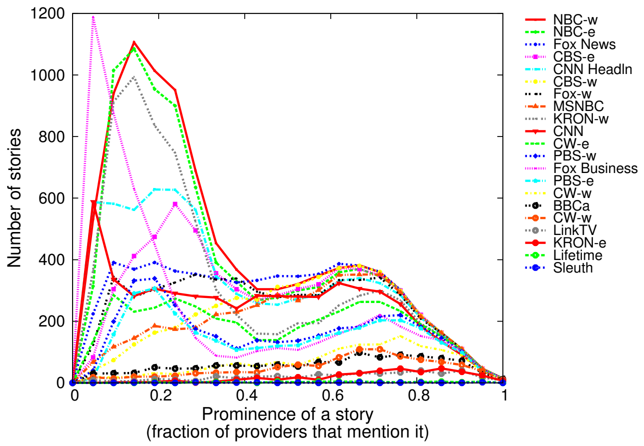


Figure 2: Distribution of prominence for `general` news stories. The distribution is bimodal.

The clustering presents three clear super-groups: `sports` news on the left, `entertainment` news in the middle, and `general` and `business` news on the right. Thus, while `business` providers share their vocabulary with `sports` providers, their style is closer to `general` providers.

`Fox News` and `MSNBC` are often considered antagonistic organizations with polarizing conservative and liberal views. However, from the perspective of style they are similar, and also similar to `CNN`. Therefore, the three most popular networks are similar both in their vocabulary and style. One outlier is `PBS`, essentially a public broadcaster whose style is quite different from the major networks. Finally, both `KRON`

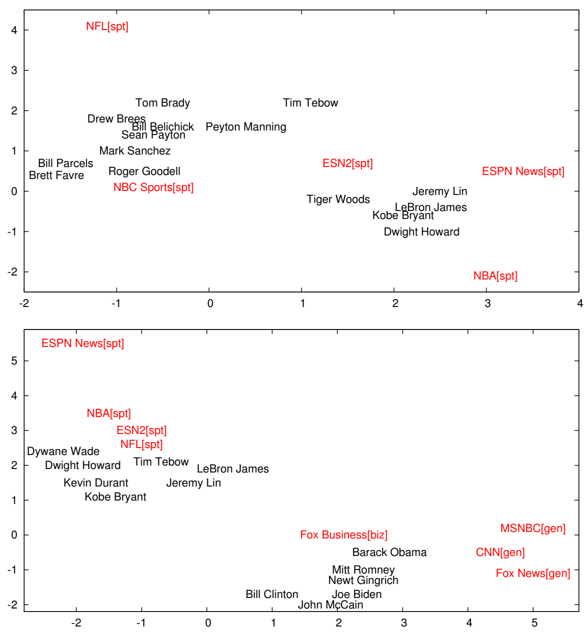


Figure 3: Provider-newsmaker projections. The first component (top) separates football from basketball players, the second component (bottom), sportspeople from politicians.

and NBC (which are affiliates and share several programs) show stylistic similarities to `entertainment` providers even when broadcasting `general` news.

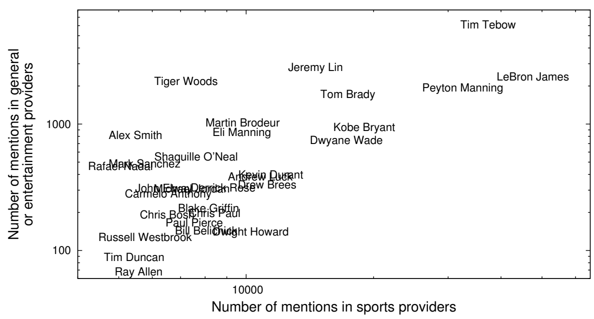
Coverage. A provider *covers* a story if it has at least one matching for it. When measuring coverage, we have to consider that some news stories are more prominent than others. We denote by *prominence* the fraction of providers of a given genre that covers a story, so a story has prominence 1.0 if it is covered by all the providers of a genre.

Figure 2 shows coverage on `general` news from the point of view of the distribution across different levels of prominence. The distribution is clearly bimodal, with the first mode around 3, and the second one around 14. Most news are covered by just a handful of providers, while a few manage to catch the attention of many providers.

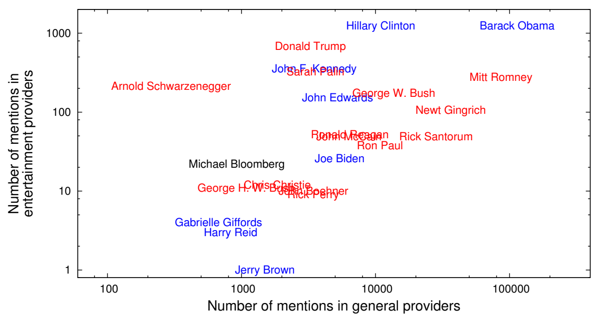
`Fox Business` has a large share of unique stories, probably due to the introduction of more niche stories from the `business` domain, even in their `general` news programs. In addition, a person watching `NBC`, `KRON`, and to some extent `CNN Headln` has a relatively higher chance to be exposed to news that are not shown by many other providers.

Newsmakers by profession. The named entity tagger we use resolves entities to Wikipedia pages [12], thus allowing us to obtain more information about people from those pages. We scrape data from Wikipedia *infoboxes* to categorize newsmakers according to professional areas and activities. The five most prominent professions in the news are basketball player, football player, Republican Party politician, Democratic Party politician, and musician. The rest of the list is dominated by other sportspeople, artists, and entertainers. The relative prominence of different sports probably varies during the year: our observation period covers the key games of the basketball and football season, but only the initial part of the baseball season. Further down the list, businesspeople appear at the 11th place, journalists at the 14th place, and government officials at position 19.

Tensor decomposition. We further explore the stylistic relationship between newsmakers and news providers via a multi-way analysis of the principal components extracted from newsmaker-tag-provider triads. Figure 3 shows the result of projecting a three-way interaction model on two dimensions while fixing the linguistic tags.



(a) Athletes



(b) Politicians

Figure 4: (a) Mentions in `general` and `entertainment` vs `sports` providers for the top-30 most mentioned athletes. (b) Mentions in `general` vs `entertainment` providers for the top-30 most mentioned politicians. Democrats in blue, Republicans in red and independents in black.

The first component neatly separates football from basketball players, which are the two most prominent professions in our dataset. The sport-specific providers `NBA[spt]` and `NFL[spt]` appear near the axes, as naturally they cover more in depth their main sport. Generalist `sports` news providers such as `ESPN News[spt]` and `ESN2[spt]` appear towards the top-right corner, while `ESPN News[spt]` seems to have a slight bias towards basketball.

The second component clearly separates sportspeople from politicians (the second most mentioned area in our dataset), together with the providers that mention each the most.

Athletes, politicians, or entertainers? Sportspeople in general are often in the news for their off-field behavior more than their performance on the field. Other times, the media focuses their attention on them in a way that emphasizes their celebrity status as a media figure over their professional accomplishments. In Figure 4(a), we display the number of mentions in `sports` stories vs the number of mentions in `general` and `entertainment` news stories for the most mentioned sportspeople. Those above the diagonal are mentioned more in `entertainment` and `general` stories than other equally-famous people in `sports` news.

We draw a similar graph for politicians in Figure 4(b). Arnold Schwarzenegger, Donald Trump, John F. Kennedy, Sarah Palin, and to some extent Hillary Clinton have a relatively strong presence in `entertainment` news. We note that California governor Arnold Schwarzenegger still has a career as an actor, Donald Trump appeared in a reality show, and Sarah Palin was a former beauty pageant.

1. REFERENCES

[1] C. Castillo, G. De Francisci Morales, and A. Shekhawat. Online matching of web content to closed captions in IntoNow. In *Proc. of SIGIR (posters)*, 2013.

[2] P. R. Center. In Changing News Landscape, Even Television is Vulnerable. Trends in News Consumption., Sep. 2012.

[3] K. Coe, D. Tewksbury, B. J. Bond, K. L. Drogos, R. W. Porter, A. Yahn, and Y. Zhang. Hostile news: Partisan use and perceptions of cable news programming. *J. of Comm.*, 58(2):201–219, 2008.

[4] R. M. Entman. Framing: Toward clarification of a fractured paradigm. *J. of Comm.*, 43(4):51–58, 1993.

[5] P. Giordani, H. Kiers, and M. Del Ferraro. ThreeWay: An R package for three-way component analysis, 2012.

[6] M. Montemurro and D. Zanette. Universal entropy of word ordering across linguistic families. *PLoSOne*, 6(5), 2011.

[7] J. S. Morris. Slanted objectivity? perceived media bias, cable news exposure, and political attitudes. *Social Science Quarterly*, 88(3):707–728, 2007.

[8] J. S. Morris and P. L. Francia. From network news to cable commentary: The evolution of television coverage of the party conventions. In *State of the Parties Conference*, 2005.

[9] D. Niven. Objective evidence on media bias: Newspaper coverage of congressional party switchers. *Journalism & Mass Communication Quarterly*, 80(2):311–326, 2003.

[10] S. Reese and P. Shoemaker. Mediating the message: Theories of influence on mass media content. *Journalism Quarterly*, 74:2, 1996.

[11] K. Schoenbach, J. De Ridder, and E. Lauf. Politicians on TV news: Getting attention in dutch and german election campaigns. *European Journal of Political Research*, 39(4): 519–531, 2001.

[12] Y. Zhou, L. Nie, O. Rouhani-Kalleh, F. Vasile, and S. Gaffney. Resolving surface forms to Wikipedia topics. In *Proc. of COLING*, pages 1335–1343. ACL, 2010.