# Says who? Automatic Text-Based Content Analysis of Television News

Carlos Castillo
Qatar Computing Research Institute
Doha, Qatar
chato@acm.org

Gianmarco De Francisci Morales
Yahoo! Research
Barcelona, Spain
gdfm@yahoo-inc.com

Marcelo Mendoza
Yahoo! Research
Santiago, Chile
mendozam@yahoo-inc.com

Nasir Khan
Al Jazeera
Doha, Qatar
nasir.khan@aljazeera.net

## ABSTRACT

We perform an automatic analysis of television news programs, based on the closed captions that accompany them. Specifically, we collect all the news broadcasted in over 140 television channels in the US during a period of six months. We start by segmenting, processing, and annotating the closed captions automatically. Next, we focus on the analysis of their linguistic style and on mentions of people using NLP methods. We present a series of key insights about news providers, people in the news, and we discuss the biases that can be uncovered by automatic means. These insights are contrasted by looking at the data from multiple points of view, including qualitative assessment.

**Categories and Subject Descriptors:** H.2.8 [Information Systems]: Database Application - *Data Mining*
**General Terms:** Algorithms, Theory, Measurement.

## 1. INTRODUCTION

Television is a dominant source of news today, wielding an enormous influence over many aspects of our life. The ascent of the Web has caused a significant drop in newspaper and radio audiences, but television remains the number one source for news in the US [2].

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). To the best of our knowledge, this work is the first to perform text analysis of content in television news with such broad goals and in such an ambitious scale.


*UnstructureNLP'13,* October 28, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2415-1/13/10
http://dx.doi.org/10.1145/2513549.2513558 ...$15.00.

We introduce a NLP-based pipeline of tools to process the input stream of data. While the specific choice of each tool is task-dependent, the pipeline itself and its components represent a minimum number of necessary steps to extract useful information from the data. Any other data mining task that uses closed captions, such as text segmentation, classification, or clustering, can build upon a pre-processing pipeline similar to the one described here.

Our NLP-based 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 tackle the following tasks:

- characterize news providers and news programs in terms of style, news coverage and timeliness (Section 3);
- characterize newsmakers with respect to their popularity, profession and similarity to other people (Section 4);
- characterize biases and framing in the coverage of newsmakers from different providers (Section 5).

## 2. RELATED WORK

One of the oldest references on mining television content is a DARPA-sponsored workshop in 1999 with a topic detection and tracking challenge [3]. Higher-level applications have been emerging in recent years. For example, [9] describe a system for finding web pages related to television content, and test different methods to synthesize a web search query from a television transcript. [15] classify videos based on a transcription obtained from speech recognition. [22] describe a system to rate the credibility of information items on television by looking at how often the same image is described in a similar way by more than one news source.

Information from closed captions has also been combined with other data sources. [11] present a system for video classification based on closed captions as well as content-based attributes of the video. [13] also combine closed captions and multimedia content to improve video segmentation. [18] align the closed captions during a live event (a presidential debate) with social media reactions. [5] combine closed captions with real-time television audience measures to detect ads – which typically are accompanied by sudden drops in viewership ("*zapping*").

**Quantitative analysis of media.**
Most content analysis of news reports is based on a time-consuming process of manual coding; automatic methods are less frequently used. [7] use an indirect measure of television bias by manually counting references to think tanks in each network, and then by scoring each think tank on a left-right political spectrum by automatically counting their appearances in congressional records of speeches by politicians of different political leaning. [4] study news articles available on the web, and analyze the prevalence of different topics and the distribution of readability and subjectivity scores.

**Topics and perspectives.** The PhD thesis of [10] and related papers (e.g. [12]), introduce a joint probabilistic model of the topics and the perspectives from which documents and sentences are written. The parameters of the model are learned from a training corpus for which the ideological perspectives are known. In contrast, the methods we study on this paper are based on NLP algorithms.

Lin et al. process news videos using content-based methods. First, news boundaries are detected automatically using visual clues, such as scene and background changes. Second, each news segment is annotated with concepts such as *sky*, *person*, *outdoor*, etc. which are inferred automatically from shots in the video. This approach is effective for settings in which closed captions are not available.

## 2.1 Text pre-processing

We use closed captions provided by a software developed by us and recently presented in the software demonstration session of SIGIR [1].

We collected data that consist of all the closed captions from January 2012 to June 2012 for about 140 channels. On average, each TV channel generates ≈ 2MB of text every day.

The closed captions are streams of plain text that we process through a series of steps. First, to segment the text stream into sentences we use a series of heuristics which include detecting a change of speaker, conventionally signaled by a text marker (">>"), using the presence of full stops, and using time-based rules. We remark that there exist methods to join sentences into passages [19, 13], but for our analysis we use single sentences as basic units of content, and we only group them when they match to the same news item, as described in Section 2.2.

Second, we recognize and extract named entities by using a named entity tagger that works in two steps: entity resolution [23] and "aboutness" ranking [16]. We focus on the **person** type in the remainder of this paper, and whenever we find a given entity in the closed captions of a news provider, we count a **mention** of that person by the provider.

Third, we apply the Stanford NLP tagger [21] to perform part-of-speech tagging and dependency parsing.[1] As a last step of the text pre-processing, we apply sentiment analysis to each sentence by using SentiStrength [20].

**Example.** A brief example can illustrate the main parts of our text pre-processing. The input data is similar to this:

```
[1339302660.000]    WHAT MORE CAN YOU ASK FOR?
[1339302662.169]    >> THIS IS WHAT NBA
[1339302663.203]    BASKETBALL IS ABOUT.
```

The TV channel maps to a network name (`CNN`), and the time is used to look-up in the programming guide to determine the type (*CNN World Sports*, which is about `sports` news). Hence, this news provider is identified as `CNN[spt]`. Finally, the text is tagged to generate the following output:

```
What/WP more/JJR can/MD you/PRP ask/VB for/IN ?/. This/DT
 is/VBZ what/WDT NBA/NNP [entity: National_Basketball_
 Association] basketball/NN is/VBZ about/IN ./.
```

## 2.2 News matching

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. The matching task is the same as the one described by [9], but the approach is based on supervised learning rather than web searches.More details can be found in [1].

The matching is performed in two steps. In the first step, a per-genre classification model trained on thousands of examples labeled by editors is applied. In this model, the two classes are "same story" and "different story" and each example consists of a sentence, a news story, and a class label. The features for the classifier are computed from each sentence-story pair by applying the named entity tagger described in the previous section on both elements of the pair, and then by looking at entity co-occurrences. The models are fine-tuned to have high precision.

In the second step, recall is improved by aggregating multiple sentence-level matchings that occur in a short time period to form a "qualified matching".

# 3. NEWS PROVIDERS

In this section we examine news providers, and try to answer the following research questions:

**Q1: Styles and genres.** Are there NLP-based attributes of newscasts that correlate with the genre of each provider?

**Q2: Coverage.** Do major networks have more coverage of news events compared to minor ones?

**Q3: Timeliness.** To what extent "breaking" a news story depends on covering a large number of stories?

## 3.1 Styles and genres

We apply factor analysis to obtain a soft bi-clustering of news providers according to the words they use more frequently (we ignore for now named entities, which are considered in Section 4). The output is shown in Table 1, where we have included descriptive providers and words for each cluster.

Most of the cohesive clusters are "pure", i.e., they have a single genre. Additionally, the three most popular networks are clustered together in the sixth cluster. Among the non-pure clusters, the third one is the most interesting. `E[ent]`, an entertainment news service, and `CNN Headln[gen]` cluster around descriptive words such as *wearing*, *love*, and *looks*, terms strongly related to the world of `entertainment` news. While `E[ent]` is expected to use such words, its similarity

[1]Further details on the tag set are available in the manual `http://nlp.stanford.edu/software/dependencies_manual.pdf`

**Table 1:** Provider-word bi-clustering results. Clusters are sorted by cohesiveness (top to bottom), and the main descriptive providers and words for each cluster are shown. Clusters may overlap.

| Providers | Words |
|---|---|
| `Fox News[gen]`<br>`CNN[gen]` | the, said, has, says, was,<br>saying, former, do, get, made |
| `MLBN[spt]`<br>`ESPN Classic[spt]`<br>`NBC-w[spt]` | save, get, facing, was, goes<br>hit, got, coming, run, getting, see<br>out, here, catch, enjoys, leading |
| `E[ent]`<br>`CNN Headln[gen]` | wearing, have, was, it, has<br>had, think, love, looks, know |
| `ESN2[spt]`<br>`ESPN News[spt]` | later, taking, wins, rose, hitting<br>get, that, got, looking, won, win |
| `NFL[spt]`<br>`ESPN News[spt]` | throw, suspended, free, be, get<br>threw, played, said, one, traded |
| `Fox News[gen]`<br>`MSNBC[gen]`<br>`CNN[gen]` | reverend, vote, said, endorsed, voted<br>saying, elected, support, attacking, running<br>defeat, attack, wants, calling, conservative |
| `NBA[spt]`<br>`ESPN News[spt]`<br>`ESN2[spt]` | finds, missing, knocking, scoring, shot<br>taking, playing, later, finding, play<br>passing, driving, out, shoot, lays |
| `MSNBC[gen]`<br>`NHL[spt]`<br>`CNN Headln[gen]` | save, stopped, played, gets, makes<br>helping, comes, playing, goes, ends<br>shut, one, traded, beats, scoring |
| `Fox Business[biz]`<br>`ESN2[spt]` | later, taking, finds, facing, wins<br>passing, beats, scoring, visiting, pitched |
| `Fox Business[biz]`<br>`NHL[spt]` | save, later, finds, taking, beats<br>scoring, looking, hosting, stopped, scored |

to `CNN Headln[gen]` can be attributed to at least two factors: (i) a deliberate stylistic and content choice made to compete with other fast-paced headline-oriented providers, and (ii) intrinsic aspects of the headline format, which is less formal, less deep, and short in nature, which inevitably leads to a more superficial coverage.

Finally, the grouping of `business` and `sports` providers at the bottom of Table 1 is also of interest. They use similar polysemic terms such as *beats*, *scoring*, *wins*, *loses*. The use of a shared terminology stems from the common nature of competition associated with both `sports` and `business`.

**Part-of-speech classes and dependencies.** 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.

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

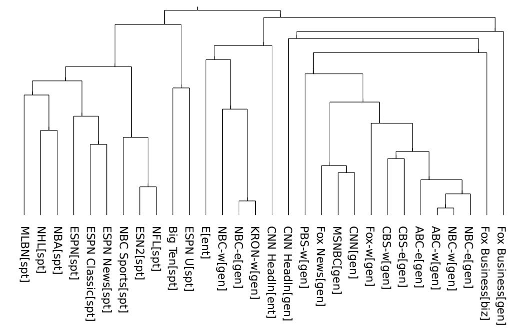



**Figure 1:** Hierarchical clustering of providers based on the prevalence of different linguistic classes.

Next we proceed to aggregate the linguistic categories at the level of genres. Figure 2 presents the results, where we have also included the type of dependency found.

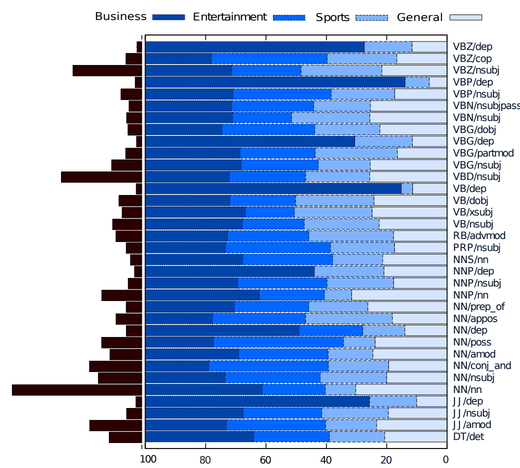



**Figure 2:** Distribution of morphological and dependency types per provider genre.

In a large number of cases for `business` providers, the dependency parser cannot extract the correct dependency (label "`/dep`"), while for `entertainment` providers the incidence is very small. A possible interpretation of this difference may be due to a different complexity of phrases. As observed by [4] for online news, politics, environment, science, and business use a more complex language; sports, arts, fashion and weather a simpler one. The analysis of other variables in our data seems to support this hypothesis. A typical sentence has 9.1 words in `sports` news, 9.0 words in `entertainment` news, 11.5 in `general` news, and 13.2 in `business` news. The Fog readability index [8] (which estimates the years of formal education needed to understand a text on a first reading) is 6.7 for `sports`, 7.2 for `entertainment`, 9.1 for `general`, and 9.4 for `business`.

For the other linguistic categories, `entertainment` has the largest relative prevalence of `NN/poss` (singular common noun, possession modifier, such as "Kristen Bell struggled to work with her *fiancé*"), `sports` has the largest value for `NN/appos` (singular common noun, appositional modifier, such as "Kevin Love's 51 *points*, a Minnesota Timberwolves team record"), and `general` news has the largest value for `NNP/nn` (singular proper noun, compound modifier, such as "President *Obama* is refocusing his campaign").

**Sentiments.** We analyze the distribution of sentiments expressed by each provider on each caption. The result is shown in Figure 3, in which we have included the number of negative words and positive words, as well as the distribution of sentiment scores.

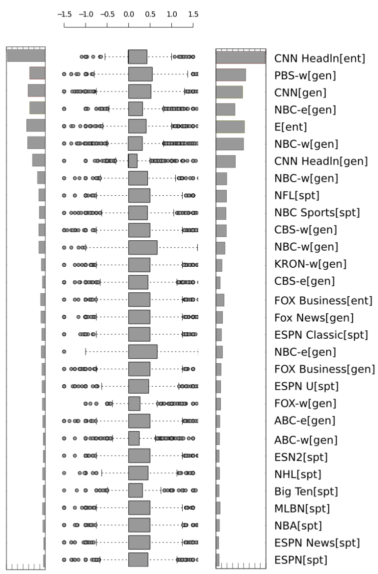



**Figure 3:** Polarity distribution of sentences per provider, shown as box plots in the center of each line. Bars on the left and right indicate the number of negative and positive words, respectively. Providers are sorted by average sentiment score, from negative (top) to positive (bottom).

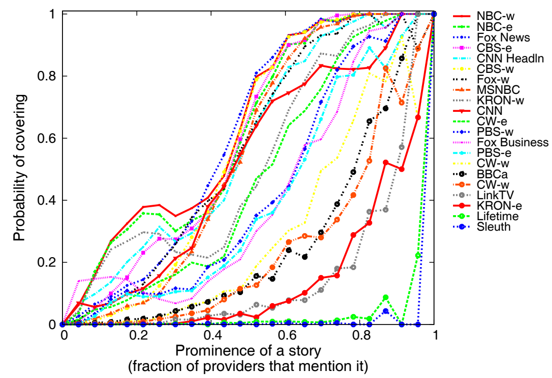
The bulk of sentiments on news seem to range from neutral to positive. All of the seven most positive providers are of the `sports` genre. `CNN Headln[ent]` is an outlier in at least two senses: it is the most negative and polarized provider, and it has many sentiment-loaded words (e.g. it has more sentiment-loaded words than `CNN Headln[gen]`, even when its constitutes the minority of programs in `CNN Headln`). This can be attributed to the "attention-grabbing" needs of the headlines format in the case of `entertainment` news.

## 3.2 Coverage

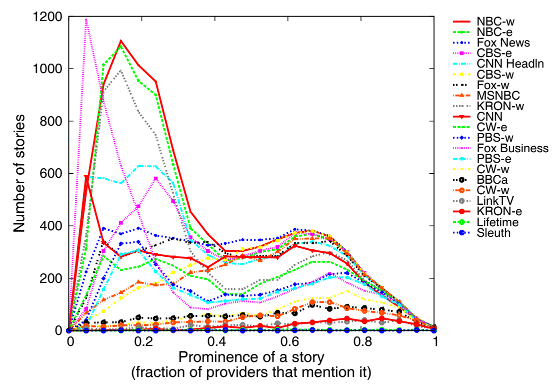
In this section and in the next one, we make use of the news matchings described in Section 2.2. A provider *covers* a story if it has at least one qualified matching for it.

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 – which is quite rare.

Figure 4(a) shows the probability that a provider of `general` news covers a story for different levels of prominence. Some providers such as `NBC`, `Fox News`, `CBS` and `CNN Headln`, offer more extensive news coverage than others. This wider selection of stories is likely due to having access to a larger




(a) Probability that a provider of `general` news covers a story as a function of its prominence.




(b) Distribution of prominence for `general` news stories. The distribution is bimodal.

**Figure 4:** Relationship between coverage and prominence for different providers (best seen in color).

pool of resources (e.g., employees, budget, affiliates) compared to the other `general` news providers. This result is expected, however the data also suggests two other relevant findings. `NBC` and `CNN Headln` seem to have a non-trivial coverage ($\approx 0.3-0.4$) of relatively niche stories (prominence $\approx 0.2$), content that is not covered by `Fox News`. However, `Fox News` has a wider coverage of stories having a prominence of $\approx 0.4$ and over, which means it reports on a higher number of stories than either `NBC` or `CNN`.

Figure 4(b) also shows coverage on `general` news, this time 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.

## 3.3 Timeliness and duration

In this section we examine how different providers cover a story over time. The life cycle of a prominent news story is exemplified by Figure 5, which depicts the coverage of an abuse probe involving US marines during two days on January 2012. Each dot represents a matching of this news story with a provider.

Providers are sorted by earliest matching: in this case, `Fox News[gen]`, `CNN Headln[gen]` and `MSNBC[gen]` are the first to broadcast the story, within minutes of each other. There are two time-dependent variables of interest that we can

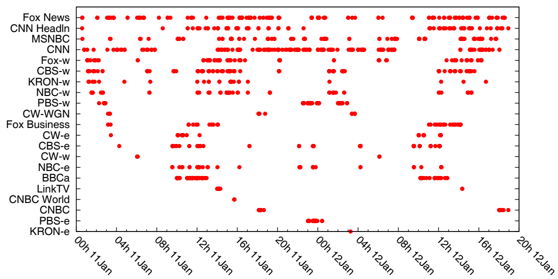



**Figure 5:** Example of matchings for a news story during two days on January 2012. The major channels are the fastest to break the news, and the minor ones follow them. The major providers have the highest density of matchings.

measure. The most important one is how often a provider is among the first ones to broadcast a story (i.e., "breaks" a story). The second is how long the providers keep broadcasting a given story.

**Timeliness.** Given that many news networks operate in cycles of 60 minutes, we consider that a provider "breaks" a story if it is among the ones that broadcast the story within the first 60 minutes since the first story matching. Figure 6 plots providers `general` news along two axes: how many qualified story matchings they provide and how many of those correspond to "breaking" a story. Most of the providers lie along the diagonal (with a ratio of $\approx 1/10$), with the exception of `Fox Business` and `NBC`. While these providers do not cover as many stories as `CNN` and `Fox News`, they are clearly better at breaking stories.

**Duration.** We define the *duration* of a story as the interval between the first and the last matching. This interval is bounded by the lifetime of the stories, which is three days. Table 2 reports the average duration per provider.

The longest duration is found in `sports` providers, followed by `business`, then `general` ones. Indeed, a game that occurs over the week-end can be commented during several days. For the major `general` news providers, the typical duration of a story is from 8 to 12 hours.

# 4. NEWSMAKERS

This section presents an analysis of *newsmakers*, i.e., people who appear in news stories on TV. We consider the following research questions:

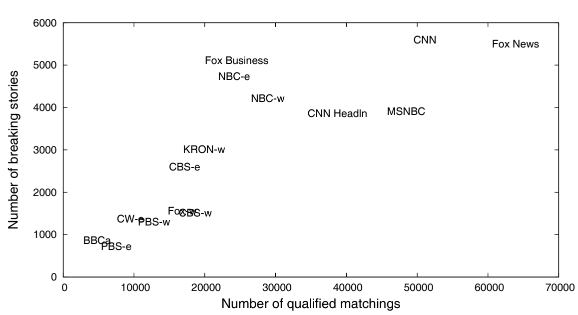



**Figure 6:** Number of breaking stories vs. number of qualified matchings, for providers of genre `general`.

**Table 2:** Average duration of a story on the top-24 providers with the most qualified matchings, i.e. the time span between the first and the last matching (hours).

| Provider | Duration (h) | Provider | Duration (h) |
|---|---|---|---|
| `ESPN Classic[spt]` | 26.6 | `CNBC[biz]` | 13.0 |
| `ESPN News[spt]` | 24.5 | `Fox Business[gen]` | 12.1 |
| `NFL[spt]` | 23.9 | `MSNBC[gen]` | 11.8 |
| `ESN2[spt]` | 23.0 | `Fox News[gen]` | 11.0 |
| `NBA[spt]` | 22.5 | `CNN[gen]` | 11.0 |
| `NHL[spt]` | 21.6 | `E[ent]` | 10.9 |
| `MLBN[spt]` | 19.3 | `CBS-w[gen]` | 9.8 |
| `NBC Sports [spt]` | 18.2 | `KRON-w[gen]` | 9.6 |
| `ESPN U[spt]` | 17.9 | `NBC-e[gen]` | 8.4 |
| `ESPN[spt]` | 17.4 | `NBC-w[gen]` | 8.4 |
| `CNBC World[biz]` | 15.7 | `CNN Headln[gen]` | 7.6 |
| `Bloomberg[biz]` | 15.4 | `PBS-w[gen]` | 5.7 |

**Q4: Newsmakers by profession.** Are there observable differences in the way different professions are covered and portrayed by news programs that can be detected by using our NLP-based pipeline?

**Q5: Newsmaker groups.** To what extent can our NLP-based pipeline identify groups of similar newsmakers?

## 4.1 Newsmakers by profession

The named entity tagger we use [23] resolves entities to Wikipedia pages, 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, and obtain a coverage of 98.2% of the mentions in our data. Table 3 shows the 24 most mentioned professions. The table spans a large range of prominence, from the most mentioned profession having more than 400k mentions to the last one shown in the table having only 3k.

**Concentration of mentions per profession.** The distribution of mentions per person in each profession varies substantially. Politicians and government officials are represented by a few key individuals who attract the majority of mentions, which is consistent with the findings of [17]. Our dataset spans across the US presidential campaign period, which may cause mentions to be even more concentrated around the top candidates. A high level of concentration of mentions is observed also in businesspersons, dominated by Warren Buffett, and in individual sports such as golf, tennis, and wrestling.

**Sentiments.** We first focus on individuals and select those that have at least 10k mentions. The persons most associated with *negative* words are: Osama bin Laden ($-0.92$), Whitney Houston (who passed away during the observation period, $-0.25$), and George W. Bush ($-0.21$). The most associated with *positive* words are three football stars: Andrew Luck[2] (1.7), Eli Manning (0.24) and Peyton Manning (0.11).

In terms of professions, Democratic Party politicians get a more negative treatment than Republican Party politicians. We observe that while the former are incumbent in the US government, the latter were undergoing their presidential primary during the first four months of our study. At each primary or caucus (there were tens of them) a number of winners or groups of winners were declared. Overall,

[2]This is likely to be to some extent, but not entirely, an artifact of "luck" being in the dictionary of the sentiment analysis software used.

**Table 3:** Top-24 occupations with the most mentions, including average sentiment and example persons.

| Activity | Sent. | Most mentioned people |
|---|---|---|
| Sports/Basketball | -0.05 | LeBron James 11%, Kobe Bryant 5%, Dwyane Wade 4% |
| Sports/Football | 0.06 | Tim Tebow 12%, Peyton Manning 9%, Tom Brady 5% |
| Politics/US GOP | -0.09 | Mitt Romney 39%, Newt Gingrich 17%, Rick Santorum 13% |
| Politics/US DEM | -0.21 | Barack Obama 76%, Hillary Clinton 9%, Joe Biden 4% |
| Art/Music | -0.01 | Whitney Houston 14%, Neil Young 12%, Jennifer Lopez 4% |
| Sports/Baseball | -0.01 | Albert Pujols 4%, Justin Verlander 3%, Bryce Harper 3% |
| Art/Actor | 0.04 | George Clooney 4%, Kim Kardashian 4%, Brad Pitt 3% |
| Sports/Ice hockey | 0.06 | Martin Brodeur 11%, Jonathan Quick 5%, Ilya Bryzgalov 4% |
| Sports/Golf | -0.02 | Tiger Woods 31%, Phil Mickelson 8%, Jack Nicklaus 8% |
| Media/ Showbiz | -0.10 | Rush Limbaugh 19%, Nicole Polizzi 13%, Al Sharpton 7% |
| Other/Business | -0.10 | Warren Buffett 21%, Jim Irsay 12%, Mark Zuckerberg 10% |
| Sports/Racing | -0.02 | Dale Earnhardt 13%, Danica Patrick 9%, Jimmie Johnson 8% |
| Sports/Tennis | -0.06 | Rafael Nadal 26%, Novan Djokovic 25%, Roger Federer 13% |
| Media/Journalist | 0.11 | Matt Lauer 12%, Wolf Blitzer 11%, Ann Curry 6% |
| Sports/Martial arts | -0.29 | Nick Diaz 13%, Nate Diaz 8%, Wanderlei Silva 5% |
| Art/Comedian | 0.07 | Stephen Colbert 10%, Bill Maher 8%, Jay Leno 8% |
| Sports/Boxing | -0.26 | Muhhamad Ali 25%, Manny Pacquiao 23%, Mike Tyson 15% |
| Mil./SA | -0.92 | Osama bin Laden 100% |
| Official/US | -0.50 | Leon Panetta 43%, Eric Holder 26%, Jay Carney 13% |
| Other/Other | -0.22 | Hilary Rosen 8%, Jeremiah Wright 8%, Andrea Yates 7% |
| Sports/Soccer | 0.01 | Lionel Messi 40%, David Beckham 27%, Wayne Rooney 13% |
| Art/Writer | -0.07 | Ernest Hemingway 17%, Andrew Breitbart 10%, William Shakespeare 7% |
| Sports/Wrestling | 0.62 | Dwayne Johnson 81%, Brock Lesnar 9%, Hulk Hogan 6% |
| Politics/US D Spouse | 0.02 | Michelle Obama 100% |

the most positive average score (0.62) is attained by professional wrestlers. Note that Dwayne "The Rock" Johnson and '80s popular culture icon Hulk Hogan also have an important career as entertainers, with "The Rock" staging a much-publicized come back in early 2012. The second most positive sentiment (0.11) is attained by journalists, thus indicating that they often refrain from criticizing or speaking in negative terms on air about their colleagues.

## 4.2 Automatic clustering of newsmakers

Table 4 shows a clustering based on linguistic attributes for each of the top providers per genre. Interestingly, `entertainment` and `sports` programs tend to conflate all politicians in one cluster, whereas `business` and `general` providers tend to separate them. For instance, `CNN Headln[gen]` generates a clear cluster with all the primary candidates, and `Fox News[gen]` separates primary candidates from final presidential candidates. This would suggest that `Fox News` has a more nuanced coverage of Republican Party politics than the other networks. Along the same lines, `Fox Business[biz]` refers to a mixture of entertainment people (George Clooney, Kim Kardashian) and sports people (LeBron James, Kobe Bryant) using a similar style, while `E[ent]` exhibit stylistic differences in the way it speaks about male celebrities (George Clooney, Justin Bieber) and female ones (Lindsay Lohan, Britney Spears): `E` displays a greater nuance in covering celebrity news.

In general, these differences suggest that providers are more discerning when covering people in their area of expertise, than when speaking about people outside it.

**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 7 shows the result of projecting a three-way interaction model on two dimensions while fixing the linguistic tags.

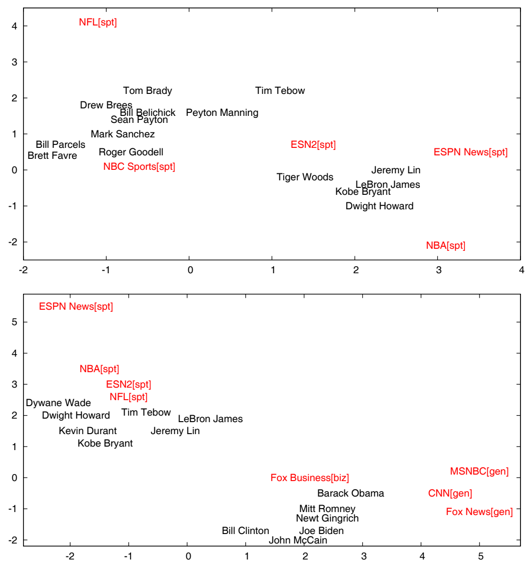



**Figure 7:** Provider-newsmaker projections. The first component (top) separates football from basketball players, the second component (bottom), sportspeople from politicians.

This model is obtained by a three-way decomposition [6], which estimates a core tensor of the data. This technique is a natural extension of principal component analysis (PCA) for two-dimensional matrices. We find that the dimensionality of the data is 3 (for the newsmakers dimension), 2 (for the linguistic dimension), and 3 (for the providers dimensions), and the tensor decomposition achieves a 78% accuracy in reconstructing the original data.

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.

# 5. FRAMING AND BIAS

In this section we analyze how different news providers frame different people. Specifically, we focus on the following research questions:

**Table 4:** Stylistic clustering per provider. Clusters are sorted by internal similarity, in decreasing order from top to bottom.

| ESPN[spt] | Fox News[gen] | CNN Headln[gen] | Fox Bus.[biz] | E[ent] | KRON[ent] |
|---|---|---|---|---|---|
| LeBron James | Neil Young | Dwyane Wade | Mitt Romney | Jimmy Carter | Mitt Romney |
| Tim Tebow | Newt Gingrich | Kobe Bryant | Newt Gingrich | Lindsay Lohan | Newt Gingrich |
| Peyton Manning | Rick Santorum | Joe Biden | Rick Santorum | Britney Spears | Tom Brady |
| Dwyane Wade | Ron Paul | Blake Griffin | Ron Paul | Bill Clinton | Bobby Brown |
| Kobe Bryant | Barack Obama | Carmelo Anthony | Bill Clinton | Brad Pitt | Matt Lauer |
| Jeremy Lin | Mitt Romney | Tom Brady | Neil Young | Mike Tyson | Whitney Houston |
| Kevin Durant | Whitney Houston | Drew Brees | Ronald Reagan | Eli Manning | Oprah Winfrey |
| Tiger Woods | Bill Clinton | Derrick Rose | Joe Biden | Hillary Clinton | George Clooney |
| Neil Young | Ronald Reagan | Eli Manning | Barack Obama | Matt Lauer | Kim Kardashian |
| Barack Obama | Joe Biden | Ronald Reagan | Tim Tebow | Michael Jackson | Justin Bieber |
| Bill Clinton | John McCain | Chris Bosh | Eli Manning | Whitney Houston | Jimmy Carter |
| Eli Manning | John Edwards | LeBron James | Jimmy Carter | Tony Romo | Brad Pitt |
| Bill Belichick | LeBron James | Tim Tebow | John McCain | Jennifer Hudson | Lindsay Lohan |
| Jimmy Carter | Dwyane Wade | Peyton Manning | John Edwards | John Travolta | Britney Spears |
| Mitt Romney | Tom Brady | Jeremy Lin | Donald Trump | Sarah Palin | Will Smith |
| Newt Gingrich | Drew Brees | Whitney Houston | John Boehner | Barack Obama | Sarah Palin |
| Joe Biden | Martin Brodeur | Tiger Woods | Hillary Clinton | Mitt Romney | Rick Santorum |
| Ronald Reagan | Andrew Luck | Jimmy Carter | Tiger Woods | Dwyane Wade | Jeremy Lin |
| Whitney Houston | Eli Manning | John McCain | Sarah Palin | Kobe Bryant | John McCain |
| Joe Biden | Tiger Woods | Barack Obama | LeBron James | Jeremy Lin | John Edwards |
| Tom Brady | John Elway | Mitt Romney | Peyton Manning | Oprah Winfrey | Hillary Clinton |
| Drew Brees | Mark Sanchez | Newt Gingrich | Dwyane Wade | Donald Trump | Donald Trump |
| Derrick Rose | Alex Smith | Rick Santorum | Tom Brady | Magic Johnson | Michelle Obama |
| Martin Brodeur | Sean Payton | Ron Paul | Kobe Bryant | Muhammad Ali | Mark Zuckerberg |
| Dwight Howard | Tim Tebow | Martin Brodeur | Jeremy Lin | Will Smith | Ann Romney |
| Chris Paul | Peyton Manning | Bill Clinton | John Elway | George Clooney | Charles Barkley |
| Blake Griffin | Kobe Bryant | Dwight Howard | Mark Sanchez | Justin Bieber | Tim Tebow |
| Chris Bosh | Jeremy Lin | Kevin Durant | Oprah Winfrey | Tim Tebow | Jennifer Hudson |
| Andrew Luck | Jimmy Carter | Andrew Luck | George Clooney | Tom Brady | Barack Obama |
| Carmelo Anthony | Michael Jordan | Bill Belichick | Kim Kardashian | Bobby Brown | Betty White |

**Q6: Positive vs negative coverage.** Are there significant differences in the sentiment polarity used by different providers when mentioning a given person?

**Q7: Outliers.** Are there people who are treated differently in a given provider compared to the rest?

## 5.1 Positive vs negative coverage per provider

Table 5 shows the average sentiment of the four most mentioned people (two politicians and two sportsmen) across the providers with the largest number of mentions of people. Previous works based on manual coding have observed clear variations in the coverage of different candidates by the major networks [14]. Our automatic analysis confirms this finding: while Obama and Romney are treated equally by `CNN` and `MSNBC`, `CNN Headln` and `Fox News` give Romney a more positive treatment. An even larger difference favoring Romney is exhibited by `Fox Business`.

With respect to `sports` news, we notice an interesting phenomenon. `NBA`, specialized in basketball, speaks more positively about the football player (Tim Tebow); conversely, `NFL`, specialized in football, speaks more positively about the basketball player (LeBron James). On average, Tim Tebow receives a more positive coverage than LeBron James, who among other things is still criticized for a team transfer in 2010, and according to USA Today became in 2012 one of the most disliked athletes in the US.[3]

## 5.2 Outlier analysis

We show outliers by vocabulary in Table 6. For each provider, we present the people whose distribution of words in sentences mentioning them differs the most from the background distribution.

With the exception of `Fox News[gen]`, other `general` news providers use a more specific vocabulary when speaking about

[3] `http://usat.ly/xgiJ25`

**Table 5:** Average sentiment scores of the 4 most mentioned persons in the top-15 providers with most mentions, grouped by type of provider. Empty cells mean no mentions.

| Provider | Barack Obama | Mitt Romney | LeBron James | Tim Tebow |
|---|---|---|---|---|
| Fox News[gen] | 0.24 | 0.33 | 0.28 | 0.37 |
| MSNBC[gen] | 0.28 | 0.28 | -0.02 | 0.15 |
| CNN[gen] | 0.30 | 0.30 | 0.01 | 0.44 |
| CNN Headln[gen] | 0.22 | 0.35 | 0.14 | 0.25 |
| ESPN News[spt] | 0.10 | 0.20 | 0.06 | 0.15 |
| NFL[spt] | 0.66 | - | 1.22 | 0.20 |
| ESN2[spt] | 0.46 | 0.14 | 0.13 | 0.20 |
| NBA[spt] | 0.21 | -0.08 | 0.06 | 0.14 |
| NHL[spt] | 0.08 | - | 0.08 | 0.37 |
| NBC Sports[spt] | 0.48 | 0.30 | 0.24 | 0.32 |
| ESPN Classic[spt] | 0.25 | - | 0.12 | 0.15 |
| MLBN[spt] | -0.04 | 0.25 | 0.10 | 0.33 |
| ESPN[spt] | 0.18 | - | 0.07 | 0.04 |
| Fox Business[biz] | 0.18 | 0.31 | 0.32 | 0.22 |
| E[ent] | 0.34 | 0.37 | - | 0.18 |

current US president Barack Obama. Interestingly, for `ESPN News[spt]` and `KRON[ent]` the most significant outliers (Junior Seau and Whitney Houston, respectively) passed away during the observation period.

# 6. CONCLUSIONS

New domains provide both new challenges for NLP methods and new insights to be drawn from the data. Closed captions data for television programs, now available from the Internet Archive, brings the television domain within the realm of data mining research. As already happened with the Web, we expect that mining this new source of data will provide a variety of interesting results.

We outlined the main results of an ambitious study on this large collection of closed captions, focusing on the domain of

**Table 6:** Top five outliers per channel according to the distribution of words in their mentions.

| CNN Headln[gen] | E[ent] | ESPN News[spt] |
|---|---|---|
| Barack Obama | Britney Spears | Junior Seau |
| Mitt Romney | Whitney Houston | Hines Ward |
| Newt Gingrich | George Clooney | Ryan Braun |
| Whitney Houston | Tim Tebow | Jerry Jones |
| Tim Tebow | Justin Bieber | Joe Philbin |
| **ESPNU[spt]** | **Fox Business[gen]** | **Fox News[gen]** |
| John Calipari | Barack Obama | Jimmy Carter |
| Brittney Griner | Mitt Romney | Mitt Romney |
| Andrew Luck | Rick Santorum | Whitney Houston |
| Rick Pitino | Ron Paul | Barack Obama |
| Jared Sullinger | Newt Gingrich | Dianne Feinstein |
| **KRON[ent]** | **KTVU[gen]** | **NFL[spt]** |
| Whitney Houston | Barack Obama | John Elway |
| Oprah Winfrey | Mitt Romney | Brett Favre |
| Britney Spears | Newt Gingrich | Joe Montana |
| John Travolta | Rick Santorum | Eli Manning |
| Jessica Simpson | Ron Paul | Sam Bradford |

news. We demonstrated the richness of this dataset by analyzing several aspects such as the relationship between genres and styles, coverage and timeliness, and sentiments and biases when mentioning people in the news. The NLP-based pipeline proposed in this paper breaks the stream of text into sentences, extracts entities, annotates the text with part-of-speech tags and sentiment scores, and finally finds matching online news articles. These steps provide the building blocks for any future task that leverages closed caption data, e.g., stream segmentation in topically homogeneous and semantically meaningful parts, classification and clustering of television news or characterization of people and events.

# 7. ACKNOWLEDGMENT

The authors wish to thank Ajay Shekhawat from Yahoo! Carlos Castillo did most of this work while he was at Yahoo!

# 8. REFERENCES

[1] C. Castillo, G. De Francisci Morales, and A. Shekhawat. Online matching of web content to closed captions in intonow. In *SIGIR demonstration papers*, 2013.

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

[3] C. Cieri, D. Graff, M. Liberman, N. Martey, S. Strassel, et al. The TDT-2 text and speech corpus. In *Proc. of DARPA Broadcast News Workshop*, pages 57–60, 1999.

[4] I. Flaounas, O. Ali, T. Lansdall, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Research methods in the age of digital journalism. *Digital Journalism*, 1(1):102–116, 2013.

[5] D. Gibbon, Z. Liu, E. Zavesky, D. Paul, D. F. Swayne, R. Jana, and B. Shahraray. Combining content analysis of television programs with audience measurement. In *Proc. of CCNC*, pages 754–758. IEEE, 2012.

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

[7] T. Groseclose and J. Milyo. A measure of media bias. *Quarterly Journal of Economics*, 120(4):1191–1237, 2005.

[8] R. Gunning. *The technique of clear writing.* McGraw-Hill New York, 1952.

[9] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-Free news search. *WWWJ*, 8(2):101–126, 2005.

[10] W.-H. Lin. *Identifying Ideological Perspectives in Text and Video.* PhD thesis, CMU, Oct. 2008.

[11] W. H. Lin and A. Hauptmann. News video classification using SVM-based multimodal classifiers and combination strategies. In *Proc. of ACM MM*, pages 323–326, 2002.

[12] W. H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proc. of CoNLL*, CoNLL-X '06, pages 109–116. ACL, 2006.

[13] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. Jose. TV news story segmentation based on semantic coherence and content similarity. In *Proc. of MMM*, volume 5916 of *LNCS*, pages 347–357, 2010.

[14] 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.

[15] S. Oger, M. Rouvier, and G. Linares. Transcription-based video genre classification. In *Proc. of ICASSP*, pages 5114–5117. IEEE, 2010.

[16] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *Proc. of CIKM*, pages 365–374. ACM Press, 2009.

[17] 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.

[18] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proc. of WSM*, pages 3–10. ACM, 2009.

[19] A. F. Smeaton, H. Lee, N. E. O'Connor, S. Marlow, and N. Murphy. TV news story segmentation, personalisation and recommendation. In *AAAI Spring Symp. on Intelligent Multimedia Knowledge Management*, 2003.

[20] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *JASIST*, 61(12):2544–2558, 2010.

[21] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL*, pages 173–180. ACL, 2003.

[22] L. Xu, Q. Ma, and M. Yoshikawa. Credibility-oriented ranking of multimedia news based on a material-opinion model. In *Proc. of WAIM*, pages 290–301, 2011.

[23] 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.