2

JOURNALISTIC DATA MINING

If there's one person you should be able to trust—whom you have to be able to trust with your body—it's your doctor. That's why it was such a startling revelation when the *Atlanta Journal Constitution* newspaper published an investigative report uncovering more than 2,400 doctors across the United States who had betrayed that trust. These doctors had been disciplined for sexual misconduct in their practice, but about half of them still had licenses and were seeing patients.[1]

The story began to come to light after reporter Danny Robbins, who had been doing other investigations into doctors in the state of Georgia, noticed a pattern of doctors continuing to practice after being accused of sexually violating patients. He compiled those cases and found about seventy in the state of Georgia alone. To see how big the issue was nationally, his team then tried to use public records requests, but "nearly all [states] said they didn't keep such data, and only a few provided other information addressing our requests."[2] Instead they had to turn to scraping—collecting medical board documents using automated scripts written by data journalist Jeff Ernsthausen so that they could crawl each state's website and download the documents about disciplinary actions against doctors. The scraping was productive, yielding more than 100,000 documents in which those 2,400 disturbed doctors were buried. But reading 100,000 documents would take thousands of hours. How could the team possibly realize their ambition of a national look at the issue?

To cope with the scale of documents, the team turned to a data-mining technique. They used a logistic regression classifier on the text of those documents to score each for its likelihood of containing a case that was relevant to the investigation. Ernsthausen iteratively built the model by reading documents and noting key terms that were both indicative and not indicative of sexual misconduct. For instance, a document mentioning "breast" could be referring to negligence in a case of "breast cancer," rather than sexual misconduct. He and other reporters tagged hundreds of documents as "interesting" or "not interesting" based on their close readings. From the set of tagged documents the model then learned to weight the selected terms and score a document based on their presence. With 84 percent accuracy the statistical model could tell the journalists whether a document was likely to yield a substantive case for their investigation.[3] And, all of a sudden, the classifier melted the 100,000 documents down to a tenth the size—still formidable, but also surmountable by the team in the given timeframe allotted for the project. From there it was still months of effort to review documents, report out hundreds of cases, and flesh out the overall story. Without the data-mining technique the investigation would have needed to scale back: "There's a chance we would have made it a regional story," explained Ernsthausen.

Advanced data-mining and machine-learning techniques are now used throughout the news industry to extract business and editorial value from massive stores of data. Beyond the investigative scenario developed by the *Atlanta Journal Constitution*, there is a growing array of journalistic uses in which algorithms can be employed for editorial purposes. In this chapter I detail five use cases that demonstrate how data mining enables editorial orientation and evaluation of information. These include discovering stories, detecting and monitoring events, making predictions, finding the truth, and curating content. I then look across these use cases to examine the larger economic rewards, subsidies, and ways in which data-mining algorithms may shape coverage and the production of journalistic knowledge. But before all that, it will be useful to understand what the capabilities of data mining really are. What can data mining do for journalism?

What Data Mining Can Do

Data mining describes a process for discovering new, valuable, and nontrivial knowledge from data.[4] With roots in statistics, machine learning, and databases, it has come to describe the entire process of knowledge production from data, including data collection, data cleaning, statistical model learning, and interpretation and application of those models. Data mining broadens data journalism by incorporating the idea that some aspects of the data-analysis phase of news production can be automated. There are six primary data-mining capabilities that enable different types of knowledge to be produced. These include: classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.[5] Oftentimes these are coupled to interactive visual analytic interfaces that augment human analysis of data.

Classification involves assigning a data item to a set of predefined classes, such as “newsworthy” or “not-newsworthy,” as the investigative journalists at the *Atlanta Journal Constitution* did. Regression entails mapping a data item into a predicted numerical variable, such as a veracity score for an image found on social media. Clustering is a process that seeks to divide data into a set of emergent categories. Tweets can be clustered into groups that represent events of interest, for example. Summarization entails the description of a set of data in a more compact form. This includes something as simple as taking the mean to represent the average value of a set of numbers, as well as more complex operations such as curating a representative set of comments to reflect debate around a news issue. Dependency modeling describes the process of finding associations (both their existence and their strength) between variables of interest. In news investigations, knowing that there is an association between two people could imply something newsworthy depending on the type and implication of the association. Finally, change and deviation detection is about discovering significant changes in data as compared to expected values.

Data mining is often enabled by machine-learning (ML) techniques—algorithms that allow for computers to learn patterns from data. ML algorithms are often distinguished by the amount and type of human feedback given to the system during the learning process: *supervised* learning depends on labeled data (such as a “newsworthy” tag for a document from which it can infer the properties associated with that label), whereas *unsupervised* learning doesn’t expect such a label. Supervised learning is useful for building classifiers or regressions, while unsupervised learning can be useful for uncovering the structure of data such as how it clusters together according to some definition of similarity between items. Another type of ML, called “reinforcement learning,” does not need labeled data but rather attempts to maximize some reward function as the algorithm makes decisions over time and observes the results. Headline testing uses this method, with a click on a headline providing positive reinforcement and feedback (a reward) for the algorithm to learn from.[6] Technically, there are a variety of specific ML algorithms that can be used for unsupervised, supervised, or semisupervised learning.

The six core data-mining capabilities form a palette of computational possibilities that news organizations can use to increase efficiencies. For instance, classifiers can tag content such as text or photos so they can be found more easily, saving time for photo editors, reporters, or archivists.[7]

Recommendation systems, virality prediction, and the timing of content are all used to optimize the reach of news content.[8] Audience analytics help to characterize users based on their past website interactions, thereby helping to optimize the funnel of new subscribers by reducing churn.[9] Data-mining techniques can clearly lead to efficiency gains throughout a news organization, but in this chapter my focus is more squarely on editorial scenarios, particularly those that involve finding and assessing news information prior to publication.

One of the defining advantages of automating the analytic stage of news production is in coping with the scale of data now generated in the world. The volume of user-generated content shared on a daily basis is staggering, creating a challenge for news organizations that recognize that in the noisy stream of content from platforms there are newsworthy events waiting to be found and reported. Data mining can help in two primary ways here, by (1) *orienting* the limited attention of professional journalists toward the subset of data or content that is likely to be journalistically interesting, and (2) *evaluating* the credibility, veracity, and factuality of sources, content, and statements to inform a degree of trust in information mined from potentially unreliable channels. In both of these cases, automation is often combined with human operators in order to ensure a high degree of quality in the information that is ultimately published.

Editorial Orientation and Evaluation

In their book *The Elements of Journalism,* Bill Kovach and Tom Rosenstiel posit the idea of "The Awareness Instinct"—the drive to know what's occurring beyond the direct experience that any individual has of the world.[10] Data mining enables this awareness instinct to operate at scale over vast swaths of data representing what's going on in the world. You could think of it as a data-driven sixth sense that orients attention. Digging through information to look for stories is an essential journalistic task that operates across a variety of data and source inputs and scenarios, from numerical data streams, to social media platforms, to leaked document sets. Triggers, alerts, and leads based on automated analysis can help to orient a journalist's attention to events likely to be newsworthy. Woven throughout are a host of editorial evaluations including decisions of newsworthiness, veracity, and credibility that help journalists not only find the news, but also ensure the truth of that news. Data-mining algorithms can help with both editorial orientation and evaluation, as I show through the lens of five use cases: discovering stories, detecting and monitoring events, making predictions, finding the truth, and curating content. Various specific journalistic uses, along with the enabling data-mining capability, and specific illustrative examples are shown in Table 2.1.

Discovering Stories: Finding the News in Data

A lot happens in the world. Most of it is unremarkable, but some of it includes events that lots of people want to know about. For a journalist confronted with an overwhelming array and scale of information, an important question then becomes how to surface the things that are interesting and newsworthy to a wide variety of people, or at least to a subset of people in a particular audience. The question "What is news?" is a difficult one to answer because it is contingent on a range of individual, organizational, social, ideological, cultural, economic, and technical forces.[11] Certain news values have, however, been repeatedly observed in journalistic news selection and are manifest in the types of stories journalists report and publish.

A recent review of news values for contemporary news practices presented a typology of fifteen possibilities including exclusivity, conflict, surprise, bad news, good news, audio-visuals, shareability, entertainment, drama, follow-up, the power elite, relevance, magnitude, celebrity, and news organization's agenda.[12] Other newsworthiness factors include proximity, novelty, salience, and societal significance (whether it be political, economic, cultural, or scientific).[13] Different domains of reporting, such as investigative or breaking news, may weigh the importance of these values in the selection of stories differently. Moreover, a story needs to fit with a publication's editorial focus, agenda, or other organizational constraints, as well as with audience expectations. Newsworthiness is not intrinsic to an event. It arises out of a judgment process that humans, and increasingly algorithms, contribute to.

The six data-mining capabilities noted earlier—classification, regression, clustering, summarization, dependency modeling, and change and deviation detection—offer new possibilities for helping to detect and discover what's newsworthy within data and documents. But each data-mining capability varies in the affordances and utilities it offers for finding different types of stories. Depending on what dimension of newsworthiness a journalist is going after, he or she might want to draw on data mining in different ways. So, for instance, dependency modeling, which entails finding associations in data, can inform investigative journalism oriented toward stories concerning connections or influence among people and organizations. Clustering, on the other hand, can be useful for collapsing the many social media posts about a particular event to find the one that has the greatest magnitude of participation. Classification can orient attention to documents likely to have salient individual stories to exemplify a broader trend. And change and deviation detection can detect anomalies and outliers that are surprising. In general, data mining enables the news value of "exclusivity" because it permits journalists to find and reveal news stories in ways that would not otherwise have been possible. It also amplifies the news value of "magnitude" as it expands the ability to find stories that are greater in scope. In the following subsections I consider specific approaches to finding stories with data mining.

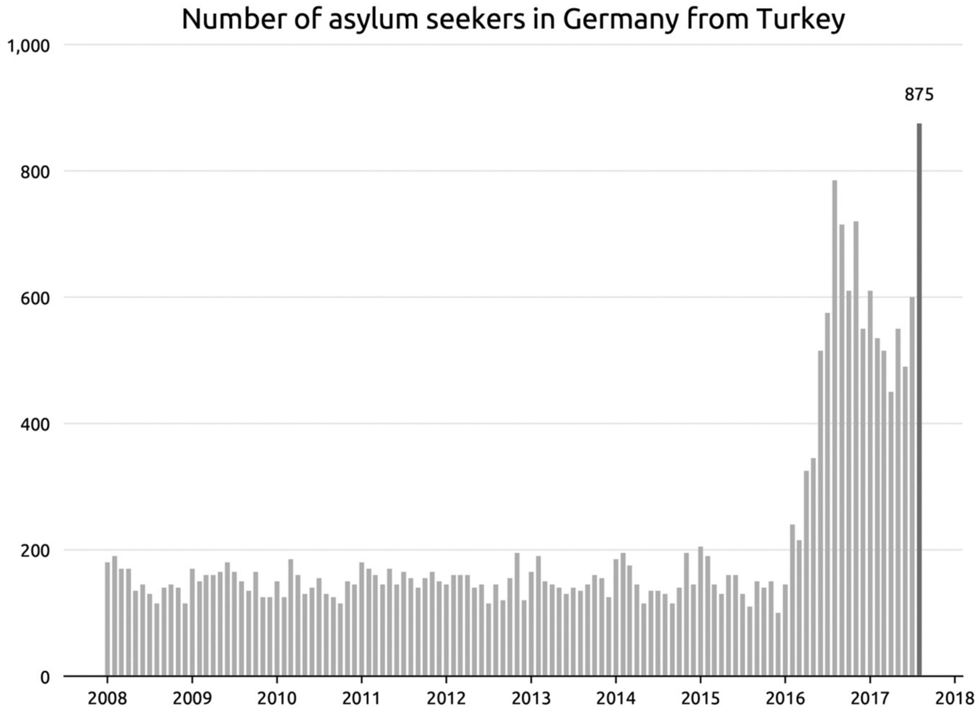
Detecting Surprises, Anomalies, Changes, and Deviations

Stories often emerge where there's a mismatch between the expectation of a reporter and a measurement of the world provided by data—a surprise. There's a lot of statistical machinery available to help find surprises in data. In science expectations are called "hypotheses," and statistical hypothesis testing can be used to answer the question of whether what the data shows is a valid, reliable, and nonspurious reflection of there being a surprise (or not) in expectations. While not widely practiced, this form of lead generation has contributed to recent investigative journalism projects such as "The Tennis Racket" published by BuzzFeed News in 2016.[14] The investigation examined the issue of fraud in professional tennis, specifically as it relates to the potential for a player to intentionally lose a match and therefore enrich people who had bet against the player. By analyzing betting odds at both the beginning and ending of a match, journalists were able to identify instances where there was an unlikely large swing (in other words, a violation of expectation, or surprise). BuzzFeed's analysis provided statistical evidence that some deviations in betting odds were anomalous. One way to think about these statistical aberrations is as leads deserving of further investigation.

Change and deviation detection techniques can also help find story leads in the routine monitoring of streams of information or data. For instance, the British Broadcasting Corporation's (BBC) research and development lab developed a prototype called "Data Stringer" that monitors numerical datasets and triggers alerts when various rules are matched.[15] Journalists could set a rule, for instance, for whether the crime rate in a particular neighborhood saw a substantial increase in comparison to last month—an interesting local story. In addition to rule-based methods, statistical anomaly detection can be used to trigger alerts. The Newsworthy project (previously called "Marple" in its prototype phase) from Journalism++ Stockholm detects statistical anomalies, or outliers, as well as trends in numerical data streams localized to different municipalities.[16] It monitors dynamic Swedish and European datasets such as employment numbers, crime reports, real-estate prices, and asylum applicants with the goal of increasing the ability of local reporters to develop news articles about any of the anomalies detected. (See Figure 2.1 for a news lead produced by the system.) In its pilot deployment the prototype system distributed thirty to one hundred news leads per dataset per month to journalists who were part of the partner network in Sweden. According to Jens Finnäs, who runs the project, anywhere from 10 to 50 percent of the leads result in some form of published story, though this conversion rate depends on factors such as the topic of the data, the day of the week, the time of year (such as summer when it’s a slower news cycle), and the availability of a local journalist’s time to chase the lead.

The number of asylum applicants from Turkey to Germany reach highest level since 2008

Published: 2017-11-03



Source: Eurostat
Prepared and analyzed by Newsworthy.se

At the end of August 875 citizens from Turkey had pending asylum applications in Germany, according to new data from Eurostat. That is more than any month since 2008.

On average Germany has seen 213 asylum applications from Turkey per month since 2008.

In the whole European Union Germany received the most asylum applications from Turkey, 875 in August, while Bulgaria, Denmark, Hungary, and Slovakia received fewest (0 applications).

None

Get the chart

As png (web) As svg (web/print) As eps (print)

Make chart with Localfocus

Get the data

As excel file As csv file

The data has been gathered from Eurostat. Get the original data here.

Figure 2.1. A news lead produced by the Newsworthy system using data from Euro Stat about asylum seekers to Germany. The system detected a spike in November 2017. Note the combination of headline, descriptive text, and annotated graph, as well as links to high resolution or interactive versions of the graph, and to the original data source. *Source:* Newsworthy (https://www.newsworthy.se/en/).

Swiss media company Tamedia sees document and data monitoring as a competitive advantage and is investing $1.5 million over three years in its Tadam project.[17] The system monitors and ingests data from social media, scraped websites, press releases, or document dumps, and triggers alerts when something newsworthy shows up. The system accepts a range of documents that it first puts through optical character recognition (OCR) in order to digitize and make them indexable for textual searches. According to Titus Plattner, who helps manage the project, it supports about twenty journalists on a regular basis and perhaps one hundred of the organization's more than one thousand journalists use it on a more ad hoc basis. Because the system must support journalists working throughout Switzerland, it automatically translates documents (and can thus trigger alerts according to language rules) into German, French, and English. This leads to interesting new capabilities for monitoring. For instance, reporter Hans-Ulrich Schaad configured the system to scrape the twenty to seventy Swiss federal court decisions posted online at noon every day and get alerts within minutes for the handful that impact the local canton he reports on. In Switzerland the language of court proceedings can vary depending on the preferred language of the defendant. Because of the automatic translation, the system was able to alert Schaad to a locally relevant decision that he could then report on for his German-speaking canton, even though the original court proceeding was in French.

Monitoring software allows journalists to expand the scope of their surveillance for potentially interesting stories. The Marshall Project, which covers criminal justice issues in the United States, developed an open-source web-monitoring tool called "Klaxon," which they use to monitor the web sites of Departments of Correction and Supreme Courts in all fifty states. Tom Meagher, the deputy managing editor, says it allows him to cover a lot more terrain than he would otherwise be able to. "It's just easier for me to sort of be aware of what's going on and to then determine where to target my energy," he explained. The monitoring capability of Klaxon has directed attention to newsworthy material that the Marshall Project has then been able to share as leads with partner organizations. But Meagher is careful to note that they really use the tool just for orienting attention—any final decisions about editorial significance are reserved for reporters and editors.

Derek Willis takes a similar approach to monitoring election information for ProPublica's Election DataBot news app, which tracks campaign finance contributions as well as other sources of election data.[18] "You cannot report on presidential campaigns, on the campaign finance aspect of it, without software. … The filings are too large. You'll miss things. You won't understand things. You won't even get the numbers right," explained Willis. At any given time he runs eight to ten data-driven rules, or simple classification algorithms, that send an alert when something of interest in the stream of campaign finance disclosures pops up. The rules are set to trigger on patterns of activity that deviate in some way from Willis's expectations; a few even lead directly to stories.[19]

Filtering for Known Patterns

Once a journalist is aware of a newsworthy pattern in a dataset, that journalist can write rules or train classifiers to scale up their ability to monitor or search for that pattern. At the *LA Times* reporters use an email monitoring system to track police reports of who was arrested the previous day. The system parses the email attachments sent from the local police departments every day and reconciles them in a database. Some simple newsworthiness rules are then run across the database to identify potential news leads, such as by sorting by the biggest bail amount or highlighting if anyone with certain occupations such as teacher or judge was arrested. Another version of this idea is the Local News Engine, which scans civic data from courts, housing developments, and business licenses in several London boroughs.[20] If it detects any newsworthy people, places, or companies, these get sent as leads to local news media.

Beyond simple rules or triggers (which can be quite effective), ML techniques can also be employed for finding patterns in large datasets. BuzzFeed News used this approach to sift through the flight data of 20,000 planes it had collected in order to identify planes exhibiting patterns of movement resembling those of FBI or DHS planes involved in surveillance (hint: they fly in circles). Using a machine-learned classifier, they were able to uncover a host of aircraft involved in suspect surveillance activities buried in the data.[21] In Ukraine ML was applied to satellite imagery to identify areas where illegal amber mining was taking place.[22] Amber mining produces a characteristic pockmarked appearance in the images that makes them relatively easy for an algorithm to distinguish from ordinary natural or urban terrain.

The Associated Press (AP) used an unsupervised data-mining technique to help find additional instances of unintentional child shootings within the Gun Violence Archive (GVA) dataset.[23] The GVA is full of data entry errors such as misspelled names, incorrect ages, or missing tags, which made it difficult to comprehensively find the cases the reporters were interested in for the story. Using the principal components analysis (PCA) technique for dimensionality reduction on the tags data, reporters were able to see how incident entries with messy data aligned with the patterns in the data associated with child shootings. "Incidents with less detail in them could thus be fit into the more generalized patterns," Francesco Marconi of the AP told me. "After the PCA was computed, every incident was either determined to be 'definitely in our scope,' 'could be, but errors / incomplete information put it on the fence,' or 'definitely not in our scope.' Deeper analysis could proceed immediately on incidents that were definitely in our interest, while vetting and research efforts could be targeted to incidents that were on the fence."

Establishing Associations and Connections

Data-mining capabilities have also proven valuable to journalists looking to find connections and associations in large document and data sets. For instance, the "follow the money" style of journalism practiced by the Organized Crime and Corruption Reporting Project (OCCRP) hinges on an ability to link one person or company to another in order to trace the influence of money in politics or to track it across borders. This allows journalists to detect fraud, corruption, or other criminal schemes. Such investigations often need to synthesize across many data sets to find connections between people of interest and other leaked documents or databases. Journalists can then leverage graph databases to query relationships by type, such as finding an individual's connections to corporations.[24] In some cases networks of relationships can be further mined using network centrality metrics to help identify and prioritize the most important figures for investigation.[25]

The challenge with these techniques is that the same people or corporations mentioned in one database might not be referred to in the same way in another database. There's lots of noise from missing data or typos that confound journalists' ability to easily match one record or document to another. Friedrich Lindenberg, a software developer with OCCRP, has been working on several data-mining solutions to cope with these challenges. Using their Linkage tool, reporters can upload a set of companies that are then matched to all of the various international databases that OCCRP curates, which then yields a list of potential associations between companies and other actors. The tool helped find a company that was smuggling US electronics components to Russian defense technology firms and was paid using laundered Russian money. The company had already been put on an international sanctions list for electronics smuggling, and so the Linkage tool could match it to OCCRP's investigation on Russian money laundering.[26] Another OCCRP tool enriches an uploaded dataset of companies or people by connecting information about ownership, control, or other relationships.[27] The tool spits out a network diagram that the journalist can visualize to see tentative links and similarity scores between entities.

Record linkage is prevalent throughout document- and data-heavy investigative journalism. Chase Davis, the former head of interactive news at the *New York Times*, explained how they built a custom data-mining system to make the task of campaign finance reporting more tractable. Although campaign finance contributions must be filed and are open to journalists to analyze, the Federal Elections Commission (FEC) does not link the records from multiple donations by the same donor. The data is really messy. One record might list a donation by "Robert L. Mercer," while another omits a middle initial and lists "Robert Mercer," and a third abbreviates a first name and transposes characters in a last name as "R. Mrecer." In order to understand the story of how money influences politics, it's necessary to be able to add up all of the contributions from a single donor. Doing this requires linking the variations on a name to a single unique identifier for that donor. As Chase told me, the *New York Times* uses an algorithm that can "discern with some probability whether this Robert Mercer, at this address or that, lives in this city, with this stated occupation, and employer … is the same as this other Robert Mercer who you know has maybe something slightly similar or slightly different listed in all of those fields."

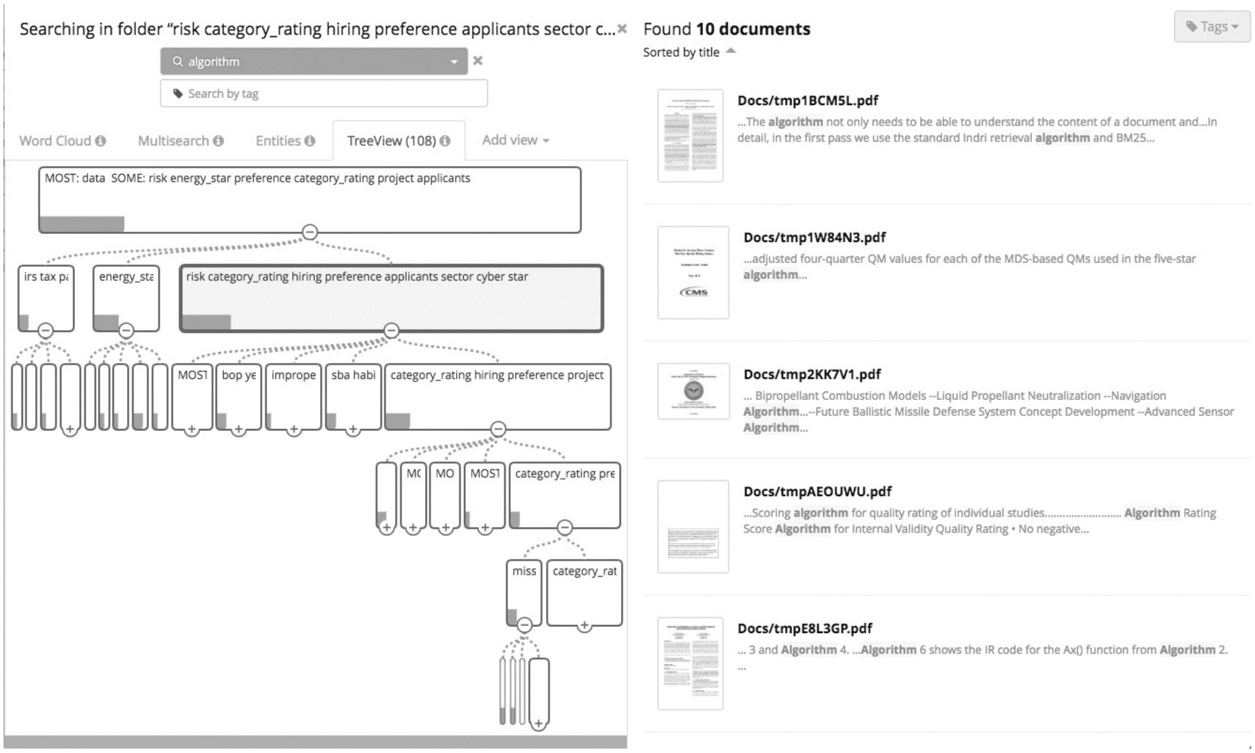
The software development time and expertise needed to use tools developed by OCCRP and the *New York Times* is still fairly high, but record linkage is slowly becoming more accessible. MaryJo Webster at the *Minneapolis Star Tribune* told me about her use of Dedupe.io, a commercial tool that has a straightforward interface that journalists can use to train an ML classifier to find record matches across messy datasets.[28] After interactively training it on her specific data, she felt confident it was finding matches she would not have found with other techniques. To be sure, each of those matches was then reported out and verified to ensure accuracy and also to flesh out the stories for each person in the dataset.

Data mining can also uncover associations that allow journalists to find and tell entirely new types of stories. Turning back to campaign finance, Nikolas Iubel and Derek Willis wanted to better characterize the different types of relationships between donors and recipients based on patterns of donations. To do so, they built a tool called “Bedfellows,” which includes a set of six association metrics that connect legislators with political action committee (PAC) donors along different dimensions such as exclusivity or duration of the donor-recipient relationship.[29] The tool led to at least one story showing that donors to Republican leaders like Paul Ryan are more similar to donors to Democratic leaders than to some other Republican members of Congress.[30] The only limit to the development of new stories and angles like this is the creativity of computational journalists and their ability to write code that can measure the existence and strength of some meaningful relationship of interest.

Counting

Counting things is a tried and true way for data journalists to find a story. Whether it’s the number of serious crimes committed, a statistic about unemployed workers, or the count of votes cast in an election, the absolute and relative rates of occurrence of various quantities in the world can trigger newsworthy observations about the magnitude or distribution of counts, or more generally about deviations and anomalies based on prior expectations.

Classification algorithms can help count items into different categories, which then provides interesting new angles on large document sets. For instance, the *LA Times* uses a classifier to tabulate the rate of campaign finance contributions from different sectors, such as “unions” or “entertainment,” and this sometimes triggers follow-on reporting if investigators notice breaks in expectations for those counts. Classification was also used to help inform the story in “The Echo Chamber,” an investigation into the nature and characteristics of US Supreme Court petitions that were ultimately heard by the court.[31] The story looked at both the type of petitioner (such as business or individual) and the topic of each petition to understand trends and patterns between type of petitioner and topic. The distribution of topics in the petitions was tabulated using a data-mining technique called “Latent Dirichlet Allocation” (LDA), which identified words most associated with any of forty different topics. This allowed the journalists to count the types of petitions that different lawyers typically submitted and to find, for example, that large firms tend to take cases pro bono only when they relate to criminal law or social issues such as same-sex marriage.


Searching in folder "risk category_rating hiring preference applicants sector c...
algorithm
Search by tag
Word Cloud
Multisearch
Entities
TreeView (108)
Add view
MOST: data SOME: risk energy_star preference category_rating project applicants
irs tax pa
energy_sta
risk category_rating hiring preference applicants sector cyber star
MOST
bop ye
imprope
sba habi
category_rating hiring preference project
MC
MO
MOST
category_rating pre
miss
category_rat
Tags

Found **10 documents**

Sorted by title

Docs/tmp1BCM5L.pdf

...The **algorithm** not only needs to be able to understand the content of a document and...In detail, in the first pass we use the standard Indri retrieval **algorithm** and BM25...

Docs/tmp1W84N3.pdf

...adjusted four-quarter QM values for each of the MDS-based QMs used in the five-star **algorithm**...

Docs/tmp2KK7V1.pdf

... Bipropellant Combustion Models --Liquid Propellant Neutralization --Navigation **Algorithm**...--Future Ballistic Missile Defense System Concept Development --Advanced Sensor **Algorithm**...

Docs/tmpAEOUWU.pdf

...Scoring **algorithm** for quality rating of individual studies......................... **Algorithm** Rating Score **Algorithm** for Internal Validity Quality Rating • No negative...

Docs/tmpE8L3GP.pdf

... 3 and **Algorithm** 4. ...**Algorithm** 6 shows the IR code for the Ax() function from **Algorithm** 2. ...

Figure 2.2. The Overview tool visual analytic interface showing a hierarchical tree view of clustered PDF documents. When the user searches within the document set the tree is highlighted and excerpts of documents are extracted and highlighted in the reading pane. *Source:* Overview (https://www.overviewdocs.com/).

Interactive Exploration

Leads from data-mining techniques provide a form of information subsidy to reporters looking for stories—they orient attention to something *potentially* interesting. If a reporter is convinced there may be something newsworthy to write about, then he or she can do additional research and reporting or engage in a variety of verification tasks before moving toward publication. This is precisely how the *Atlanta Journal Constitution* used classification to save time in its investigation of doctors and sex abuse. But pure data mining can often provide only the most basic of starting points: "look here." Story selection is an iterative and contingent process that can benefit from interactivity. Coupling human perceptual abilities in a tight feedback loop with data-mining capabilities can help move the process from "looking" for stories to interactively selecting them.[32] Interactive exploration of data using visualizations integrated with automated data-analysis capabilities is known as *visual analytics,* a growing field of research that is also increasingly being used to find news stories.

Overview is a visual analytic tool that incorporates document clustering to help investigative journalists hone in on stories of interest.[33] As of 2017, journalists had found and published about two dozen stories as a direct result of using the tool's interactive data-mining and visualization capabilities.[34] The visual clustering interface allows journalists to see groups of documents that are related based on their contents (see Figure 2.2). These groups may correspond to document subsets of interest to an investigation. RevEx (for "Review Explorer") is another example of a visual analytic interface successfully used to enable journalism. New York University and ProPublica developed it in collaboration in order to investigate more than 1.3 million Yelp reviews of healthcare providers.[35] A similar academic-plus-industry collaboration between Technische Universität Darmstadt and *Der Spiegel* developed the new / s / leak tool, which focuses on the visualization of networks of entities.[36] The goal was to help answer the question "Who does what to whom?" for investigative journalists examining large document sets. Interactive tools like Overview, RevEx, and new / s / leak acknowledge that making sense of huge document sets requires iterative slicing, dicing, and visual representation of data, exploration of what the data might contain that the analyst may not have initially considered, and an ability to iteratively question and hypothesize about what the story may really be.

Detecting and Monitoring Events

Twitter, Facebook, YouTube, and other platforms produce an immense scale and velocity of social media content. Eyewitnesses upload photos and videos and share real-time observations of events, creating an opportunity for journalists to monitor the platforms for newsworthy material. Twitter can even act as a sort of distributed social sensor system that feeds algorithms to detect physical events such as earthquakes within minutes after they occur.[37] A growing number of algorithmically augmented social media listening tools are now commercially available and are used extensively by newsrooms.[38]

Monitoring Social Platforms

Reuters is big enough that they decided to build their own social monitoring tool. The tool, called "Tracer," knits together many of the six core data-mining capabilities to help journalists keep track of social media at scale.[39] Tracer provides a set of data-mining algorithms that feed an interactive user interface for journalists. The algorithm monitors and filters a sample of tweets from Twitter, clusters and detects events by grouping tweets likely to be about the same thing, labels those clusters using text summarization, scores events based on a newsworthiness prediction, and then rates the veracity of the event. Professional journalists can then interactively adjust event curation according to search terms and automatically generated facets relating to location, recency, and magnitude of impact.[40] For instance, a journalist could set up the tool to monitor for “disaster” events related to the topic of “oil” along the Gulf Coast, and it would alert them if an explosion happens at an oil refinery there. The journalist can then independently verify the event through his or her own contacts and sources before deciding to publish.

Tracer enables a whole new scale of social media monitoring that simply wouldn’t be possible without data-mining algorithms. The tool churns through huge piles of tweets—filtering through about 12 million per day, clustering those down to about 16,000 events, and then winnowing further down to about 6,600 based on the newsworthiness prediction. The tool provides a competitive advantage when it comes to speed: it detected the 2016 Brussels airport bombing two minutes before local media, eight minutes before the standard Reuters alert was sent, and ten minutes before the BBC reported it. An evaluation of thirty-one events found that the tool would accelerate the speed of Reuters’ news alerts in 84 percent of cases.[41] The tool orients Reuters journalists to breaking news events, often giving them a head start in their reporting and providing welcome hints about the veracity of the event.

The increased scale and speed Tracer affords to Reuters’ journalists comes at a cost though. Machine-learned classifiers for evaluating events according to newsworthiness and veracity are far from perfect. In an evaluation, researchers used a test set of 300 events in which 63 were newsworthy according to journalists. Ranking those 300 events by their newsworthiness score and then looking at only the top 50 resulted in finding 36 events deemed newsworthy by journalists (and 14 non-newsworthy events). Such a benchmark is fairly impressive given that an idea like newsworthiness depends on reporting context and is subjective and difficult for many journalists to articulate. A corpus of more than 800 test events evenly split between true and false events was used to evaluate the veracity scoring algorithm. The method could debunk “75% of rumors earlier than news media with at least 80% accuracy.”[42] The accuracy improves over time, reaching 83 percent after twelve hours and almost 88 percent after seventy-two hours. These evaluations point out the main trade-off involved in employing a system like Tracer: it helps with scale, but introduces potential for errors. Predicted newsworthiness and veracity scores have uncertainty associated with them.

On the academic side of research on social media monitoring, the City Beat project at Cornell Tech used geotagged Instagram posts to locate neighborhood-level events in New York City.[43] Deployment of the tool to four local newsrooms led to several key insights about the design of such monitoring tools for journalists. For one, journalists are aware that popularity can be a red herring: just because a lot of people are sharing information about an event doesn't make it important news. Newsworthiness is a construct whose definition varies from newsroom to newsroom, so tools like City Beat also need to have configurable definitions of newsworthiness detection algorithms in order to meet those different needs. Also, because Instagram was the source of media used by City Beat, many of the events detected were conferences, concerts, festivals, gallery openings, sports events, and so on, as well as a few emergencies such as fires. Most of these (with the exception of fires) are planned events, which are almost always announced via other channels in advance, limiting the utility of the tool.

Monitoring News Media

Other forms of media, such as photos or videos from the news media itself, can also be monitored automatically using data-mining techniques.[44] For instance, Global Data on Events, Location, and Tone (GDELT) monitors the world's media in over one hundred languages and can detect three hundred different categories of events such as protests or acts of aggression.[45] Such systems are typically bootstrapped using linguistic rules, but can then be refined using statistical techniques to classify an event type and associate a source and target of the event.[46] So the sentence "Postal service workers stage protest over cutbacks by federal government" could be classified as representing a protest event with a source of "postal service workers" and a target of "federal government." Extracting typed events from media reports using data-mining techniques has the potential to allow journalists to go beyond the initial alert about the existence of the event to learn about who might be involved, which could in turn inform decisions about newsworthiness.

Media monitoring can also be useful for building datasets that enable tracking of events over time. The Documenting Hate News Index, produced by ProPublica, uses a machine learned classifier to identify instances of hate crimes reported in the news media; in effect, it acts as a type of aggregator on the topic.[47] Another recurring issue warranting the development of ongoing monitoring has been the use of fatal force by police. News organizations such as the *Guardian* and the *Washington Post* create datasets to track the issue by, for instance, submitting public records requests to learn more about each case. Research in data mining is also making progress in tabulating such events automatically. A recently published algorithm was able to detect thirty-nine cases of fatal police shootings missed by a manually created police fatality database.[48] Yet overall accuracy benchmarks are still lackluster. False alarms, incorrectly associated actors (such as a police officer being fatally shot rather than doing the fatal shooting), and events that may happen in the future or are hypothetical all confound such techniques. But complete automation is the wrong way to think about the application of these techniques in journalism practice. The results of automated event detection can be coupled to human knowledge and intellect. The algorithm can detect fatal shooting events that are moderately to highly likely and highlight why that may be the case, and then a human can verify each instance.

Making Predictions

While detecting or extracting events based on vast troves of media can help journalists stay on top of breaking events and tabulate events of interest over time, the frontier of data mining online media is in predicting events in the future. US intelligence services are already deploying the technology to get a jump on geopolitical events. For instance, the Early Model Based Event Recognition Using Surrogates (EMBERS) project funded by the Intelligence Advanced Research Projects Activity (IARPA) has the goal of predicting events such as protests, offering alerts that provide advanced warning for planning event responses. By monitoring an array of information from news sites, blogs, and other social media the system can forecast when a protest will happen, in what city it will happen, which subgroups of the population will be protesting, and why they’re protesting. And it provides these forecasts an average of almost nine days in advance with an accuracy of 69 percent. It successfully forecast the June 2013 protests in Brazil and the February 2014 protests in Venezuela.

Although news organizations have not yet deployed such event prediction systems, it’s not hard to envision their utility. If a news organization were to know ahead of time that a major protest was likely to happen in a particular city, they could deploy reporters or video equipment ahead of time in order to be ready for any breaking news. This is not entirely unlike how forecasts for extreme weather events such as hurricanes work. At the same time, predicting social events is quite different from predicting weather events. Social events are composed of individuals who have agency and can affect outcomes, whereas weather is a physical process. The adoption of prediction for socially oriented events faces interesting ethical questions given that a news organization's own behavior and actions could influence how the predicted events unfold. By publishing a prediction that a protest event is likely to happen, does it send a signal to potential participants that makes it *even more* likely to take place? For social events, the act of publication may create a feedback loop that amplifies (or dampens) the likelihood of the event. Could a news organization truly represent its actions as independent in such cases, given that it may be co-constructing the event?

News organizations are not yet publishing predictions of social unrest, but they are deploying prediction in other editorial scenarios. FiveThirtyEight has become well-known for its forecasts in the domain of sports, including for American football, basketball, and tennis.[49] Other news organizations such as the *New York Times* also publish predictions about sports. Predictions range from how likely a football team is to make it to or even win the Super Bowl to whether a team should punt or go for a goal, to how good an NBA player will be (and how much that player will be worth) several years from now. Data-driven sports coverage is a genre that attracts a lot of attention, and die-hard sports fans may appreciate the rankings such predictions create. Sports has the advantage of being a relatively low-stakes environment where published predictions have limited potential to influence the system they're predicting.

Politics is another domain where news organizations have started to employ prediction. For several years now, the *Atlanta Journal Constitution* has integrated a score into its online news app, the Georgia Legislative Navigator, which reflects a prediction of whether a state bill is likely to pass or not. The model uses features such as which party sponsored the bill, how many cosponsors it had, and a range of content-related features based on keywords and phrases. It achieves a respectable prediction accuracy that updates over the course of the legislative session.[50] The predictions offer an interesting signal to audience members following at arm's length, although a diligent reporter might produce more accurate predictions by making a few phone calls to gather nonquantified social knowledge about a bill's chances. In one case the predictions led to some community controversy. A bill proposing the creation of a new town got some people in favor of it hopeful and others opposed to it angry when they saw the relatively high predicted probability of the legislation passing. At the end of the day the bill didn’t pass—people, pro and con, had gotten excited for nothing. This example raises questions about how to make the uncertainty of predictions more clearly understood, particularly to people who may themselves have limited statistical knowledge.[51] More broadly, new ethical treatments may be needed to grapple with feedback loops and the potential for predictions to impact social behavior and reactions, particularly around high-stakes election predictions.[52]

For several years now FiveThirtyEight, as well as other news outlets, have been active in publishing predictions relating to electoral politics in the United States. The founder of FiveThirtyEight, Nate Silver, made a name for himself by using statistical models to accurately predict the presidential vote outcome in forty-nine out of fifty states in 2008 and for all fifty states in 2012. But in 2016 Silver’s (and others’) predictions turned out differently. Their models indicated that Hillary Clinton had a greater chance of winning the election than Donald Trump (she did win the popular vote, but not the Electoral College). The methods Silver uses are complex and extremely wonky—the 2016 forecast was accompanied by a link to an almost 5,000-word “user guide” to understanding the model.[53] The failure of the predictive model (or perhaps of the attempt to convey an apt interpretation of that model) prompted an eleven-part post mortem on “The Real Story of 2016,” which unpacked some of the contributing factors, including overhyped early voting, “invisible” undecided voters, electoral college weakness due to concentration of liberal voters in cities, and the impact of then–FBI Director James Comey’s letter to Congress suggesting new evidence had surfaced in a case related to Hillary Clinton’s private email server.[54] Predictive models may be able to capture some of these aspects of the world in future iterations, but nonquantified events will still impact outcomes—a fundamental weakness of relying too heavily on data-driven prediction.

Finding the Truth

Journalistic reports not only need to be newsworthy, they also need to be true. Editorial evaluations of newsworthiness are pervasive throughout the process of finding and selecting stories, but evaluations of veracity are just as important, and are tightly integrated into the overall workflow as journalists find and select stories.[55] In an interview study of twenty-two journalists, support for helping to verify content was ranked as the number two desired feature for a social media tool, just behind alerts for breaking news.[56] As journalists follow a lead to assess its newsworthiness, they need to ensure that their information sources—whether social media contacts, documents, or databases—are trustworthy. In other journalistic scenarios, such as in fact-checking the statements and claims of politicians or other powerful elites, veracity assessment is an end in and of itself. Data mining can help journalists make more effective decisions about the trustworthiness and utility of information sources as they produce news.

Source and Content Verification

Data journalism relies on data as a source. But as with any other source, journalists need to evaluate its credibility and veracity: Who produced the data, and is it complete, timely, and accurate? Given that government data can be prone to mistakes and errors, sometimes data quality itself is the story. A good example comes from the *LA Times* where reporter Anthony Pesce built a machine-learned classifier to evaluate data received from a public records request to the Los Angeles Police Department (LAPD) about crimes in the city. The classifier allowed Pesce to extend the manual evaluation of a smaller slice of city data he had undertaken as part of a previous investigation. The story showed that the LAPD had misclassified an estimated 14,000 serious assaults as minor offenses, thus artificially lowering the city's crime rate.[57] To arrive at this conclusion the classifier learned which keywords in the crime description data were associated with serious versus minor assaults and then compared this to how the crime had been officially categorized. Despite the classifier itself having an error rate, the technique was able to identify errors in the LAPD's data and to quickly show reporters that the misclassification rate was stable over time.[58] Most likely there hadn't been a blatant attempt to manipulate the crime rate; otherwise the errors would have changed over time.

Evaluating sources is increasingly important on social media, too. Platforms like Twitter make it possible to expand the set of sources available to journalists to include more nonofficial and alternative sources while reducing reliance on mainstream or institutional elite sources.[59] Yet relying on more nonofficial (and perhaps unfamiliar) sources poses challenges to source verification. Journalists need to quickly vet sources for credibility and trustworthiness while coping with the often overwhelming scale of social media content and the time pressure of unfolding events. Journalists need to be able to quickly assess whether any particular Twitter user could be a valuable source for additional information. Different events may demand different evaluations of credibility. The data-mining capability of summarization can help with the scale of this problem by crunching data about account activity and history into scores.[60] But ultimately credibility is a construct that relies heavily on contextual information that may not be available to algorithms. Some stories may call for identifying experts who can speak reliably to a topic or issue, so-called cognitive authorities. In other situations, such as in breaking news, which involve readily perceivable information (fires, crimes, storms, bombings, and the like), cognitive authorities are less useful, at least initially, than eyewitnesses. By nature of their proximity and their ability to report on an event using their own perceptions of the world, eyewitnesses have increased credibility in such situations.

It was against this backdrop that, with collaborators at Rutgers University, I designed a tool called "Seriously Rapid Source Review" (SRSR) in 2011 to integrate data-mined signals about sources on Twitter into an interactive interface journalists could use to search, filter, and evaluate potential sources during a breaking news event.[61] The prototype provided interface cues gleaned from various data-mining routines to show additional source context such as their likely location, their social network connections, and their user type (such as institution, journalist, or other), as well as whether they were identified as a probable eyewitness. Our evaluations with professional journalists looking at data from both a tornado and a riot event demonstrated that these contextual cues could help journalists make quicker and more effective judgments about potential sources.

The eyewitness classification algorithm we developed was built with the understanding that people who see, hear, or know by personal experience and perception are coveted sources for journalists covering breaking news events. The algorithm was simple, relying on a dictionary-based technique to analyze the content of tweets and look for the presence of any of 741 different words that related to categories of perception such as seeing or hearing. The classifier would mark someone as a likely eyewitness if the user had used any of these words in tweets about the event. In evaluating the technique against manually tagged accounts, we found that the classifier had a high precision, where 89 percent of the time if it said someone was an eyewitness then that person was, and a lower recall indicating it only found about 32 percent of eyewitnesses overall (it missed quite a few). More sophisticated ML techniques can classify event witnesses with an overall accuracy of close to 90 percent.[62] Such techniques rely on additional textual features relating to crisis-sensitive language such as "near me" and expressions of time awareness such as "all of a sudden."

Finding credible sources is just a small slice of the search for truth on social media. In the wake of the 2016 US presidential elections the topic of “fake news” reached a fever pitch as media scholars struggled to understand the impacts of misleading or manipulated information, false context, rumors, or even completely fabricated media circulating on social platforms.[63] The key evaluations journalists need to make are whether a piece of content is authentic (that is, it is what it says it is) and that what it claims is true. Those aren’t easy tasks with social media. A study of rumor propagation on Twitter quantified the difference between the amount of time it takes to resolve a true rumor (two hours) versus a false one (fourteen hours) and concluded that “proving a fact is not accurate is far more difficult than proving it is true.”[64] Data mining is not a silver bullet for automatic content verification or rumor debunking, but it can provide additional signals that human evaluators might take into consideration, such as cues about information provenance or credibility.[65] Such a hybrid approach is what Reuters’ Tracer system implemented.

Evaluating content for credibility and veracity is a tall order for data-mining techniques, but initial results have been promising. Research has demonstrated machine-learned classifiers that can rate whether a tweet is credible or not using text from the post as well as features such as sentiment and the use of links and question marks.[66] Accuracy was 86 percent across 608 test cases. More recent research in the InVid project (“In Video Veritas”) has developed automated techniques to aid with debunking fake images and videos online, reaching an accuracy of 92 percent with sophisticated ML processes.[67] Users can interactively access the algorithm’s results on video content using Chrome or Firefox browser plugins.[68] From year to year international competitions with names like MediaEval and RumourEval promulgate structured tasks, evaluation metrics, and open data sets that challenge researchers to try different approaches for advancing the accuracy of automated content verification.[69]

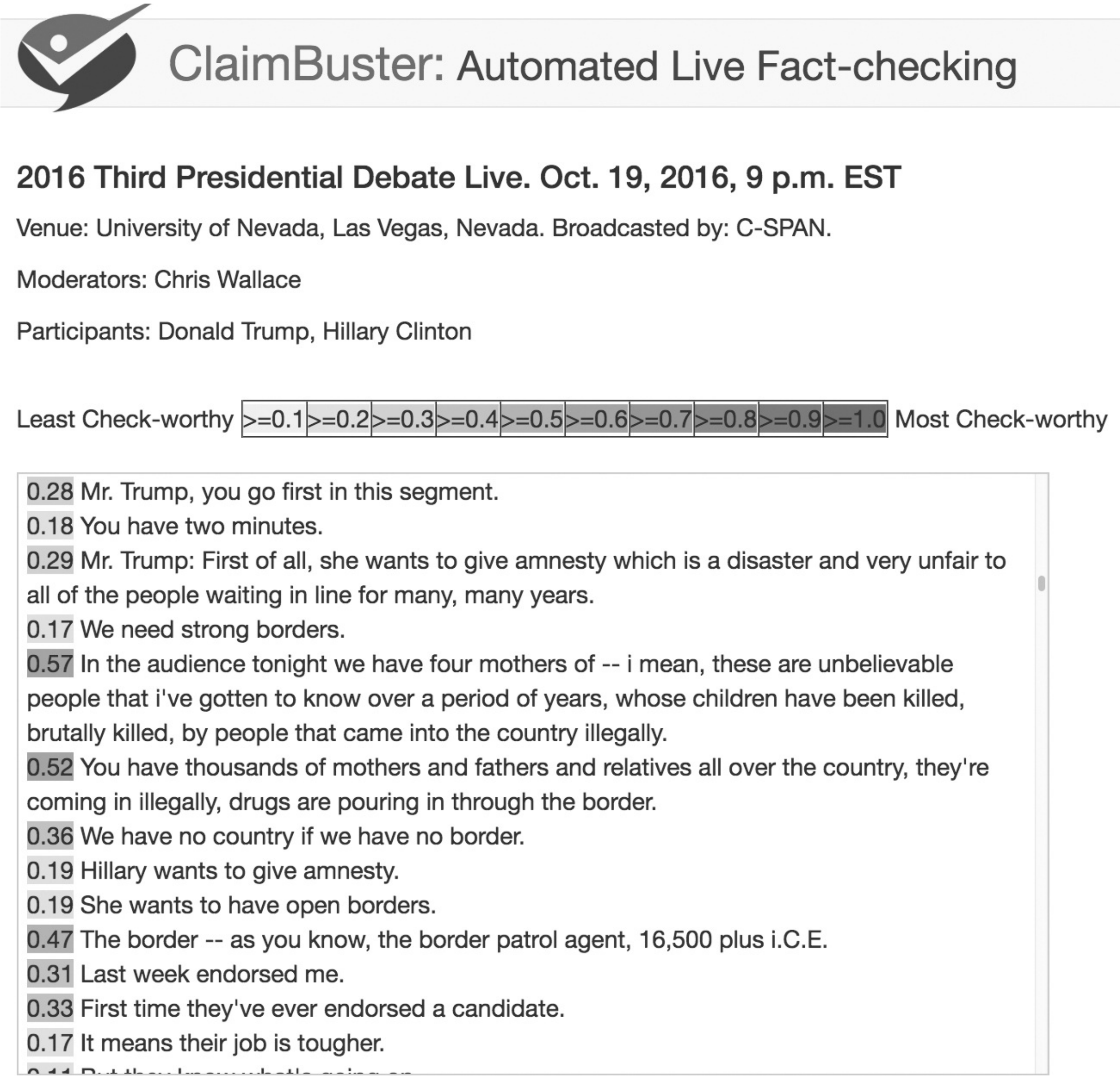
Fact-Checking

Fact-checking is another type of verification task that tackles the evaluation of statements and claims made by information sources. Traditionally publishers would check facts before publication, subjecting all the names, stats, and other claims in a story to a rigorous internal verification process.[70] More recently sites such as PolitiFact, FactCheck.org, and FullFact have made a name for themselves by pursuing fact-checking as a public activity, which produces its own form of coverage and content. National Public Radio (NPR) published a near real-time fact-checked transcript of the 2016 presidential debates drawing on the expertise of more than thirty newsroom staffers and attracting record traffic to the website on the day of the first debate.[71] Journalists for these organizations tirelessly research and assess the accuracy of all kinds of statements and claims from politicians, think tanks, and other sources. Drawing on background knowledge, context, and a healthy understanding of how trained communicators try to spin, hype, or reframe facts,[72] the task of parsing out the real facts from the opinions, the matters of taste, and the ambiguously misleading is a painstaking one.

A 2016 white paper from FullFact, a UK-based fact-checking organization, outlined several ideas for computational tools to aid with monitoring claims, spotting claims to check, doing the check, and then publishing the check.[73] Claim-spotting is one of the initial areas of focus since it's more computationally tractable. Claim-spotting can be broken down further into tasks such as detecting claims in new text that have already been checked, identifying new claims that have not yet been checked, prioritizing claims editorially so that human fact-checkers can attend to the more important first, and coping with different phrasings of the same claims. There are a number of challenging data-mining problems here, but FullFact frames the solution as a hybrid that takes advantage of computing to spot claims and surface relevant context while humans take on the sometimes nuanced interpretation and arbitration of statements whose truth values span shades of gray and can be evaluated only with access to nonquantified context and the synthesis of information from multiple sources. Computer-assisted fact-checking appears to be the most productive course of action for scaling up fact-checking activities.[74]

One of the earliest research systems for claim spotting, called ClaimBuster, monitors live interviews, speeches, debates, and social media to identify factual claims that a person might look at more closely.[75] Its classifier can distinguish between nonfactual sentences, unimportant factual sentences, and so-called check-worthy factual sentences. Check-worthy sentences tend to use numbers more often and are written using the past tense of a verb. Trained on more than 20,000 hand-labeled sentences from past US presidential debates, the system rates each sentence it sees on a scale from 0 to 1, with a 1 being more check-worthy. For instance, the last sentence has a score of 0.72, likely due to its use of numbers. (See Figure 2.3 for further examples.)[76] In general, 96 of the top 100 sentences that received a check-worthy score were claims that human raters also agreed should be checked. The score also correlated well with claims that CNN and PolitiFact checked, thus showing good external validity in terms of ability to rank statements that professional fact-checkers find important to assess. News organizations are already making use of automation to help spot claims for human fact-checkers to focus on. Duke University's Reporters' Lab uses the ClaimBuster scores to monitor CNN transcripts for checkable claims on a daily basis. The claims are automatically sent to newsrooms such as the *Washington Post* and PolitiFact, where professional fact-checkers decide if they want to actually check a claim.[77] In the first eight months of 2018 at least eleven fact checks were published as a result of these alerts. FullFact has developed a claim spotting algorithm that they use to highlight checkable claims for fact-checkers in real time during television broadcasts. By using more sophisticated representations of language their algorithm achieves a 5 percent relative performance improvement in comparison to ClaimBuster, in particular by missing fewer checkable claims in the texts it scans.[78]

Another aspect of claim-spotting is the identification of textual claims that have already been checked.[79] Politicians tend to repeat their talking points all the time, so why repeat a fact-check if you've already got a database of checks that simply need to be associated to the various versions of the statement coming across the wire? Matching a checked statement to a new statement is actually harder to automate than you might think. There are a lot of different ways of saying the same thing, which confounds natural language understanding by algorithms. Moreover, the tiniest change in context could alter the meaning of a statement and make it difficult to assess the equivalence of statements. A statement such as "The employment rate in New York rose to record levels last year" depends on what year the statement was written; the truth might be different depending on whether we're talking about 2017 or 2018 as "last year." Instead of trying to do this whole process automatically FullFact's tool surfaces context for each claim it matches to its database, giving the fact-checker a chance to verify the match before publishing.



ClaimBuster: Automated Live Fact-checking

2016 Third Presidential Debate Live. Oct. 19, 2016, 9 p.m. EST

Venue: University of Nevada, Las Vegas, Nevada. Broadcasted by: C-SPAN.

Moderators: Chris Wallace

Participants: Donald Trump, Hillary Clinton

Least Check-worthy >=0.1 >=0.2 >=0.3 >=0.4 >=0.5 >=0.6 >=0.7 >=0.8 >=0.9 >=1.0 Most Check-worthy

0.28 Mr. Trump, you go first in this segment.
0.18 You have two minutes.
0.29 Mr. Trump: First of all, she wants to give amnesty which is a disaster and very unfair to all of the people waiting in line for many, many years.
0.17 We need strong borders.
0.57 In the audience tonight we have four mothers of -- i mean, these are unbelievable people that i've gotten to know over a period of years, whose children have been killed, brutally killed, by people that came into the country illegally.
0.52 You have thousands of mothers and fathers and relatives all over the country, they're coming in illegally, drugs are pouring in through the border.
0.36 We have no country if we have no border.
0.19 Hillary wants to give amnesty.
0.19 She wants to have open borders.
0.47 The border -- as you know, the border patrol agent, 16,500 plus i.C.E.
0.31 Last week endorsed me.
0.33 First time they've ever endorsed a candidate.
0.17 It means their job is tougher.

Figure 2.3. ClaimBuster scores for an excerpt of the 2016 third presidential debate. *Source:* Claimbuster: http://idir-server2.uta.edu/claimbuster/

Nascent research efforts are also developing algorithms that can not only identify claims to check, but also automatically assess the truth value of the claim itself. For instance, one effort has focused on assessing the factuality of simple numerical claims by using a knowledge base that can corroborate or refute the claims.[80] So the statement “Germany has about 80 million inhabitants” could be compared against a knowledge-based entry <Germany; Population; 2017; 82,670,000> and show that the statement is quite close to being true. The algorithm first matches entities in a sentence to the knowledge-base entries, then these candidates are filtered using a machine-learned classifier that assesses the relevance of each entry to the claim, and finally the value in the statement is compared to the value in the knowledge base in order to label the claim as true or not. The approach is limited by the coverage of the knowledge base and is also unable to deal with sentences with more than one property, such as comparisons. Such scores could be productively woven into the workflow of human fact-checkers to make them more efficient and effective. But for now, fully automated claim-checking remains quite challenging, with systems able to deal only with simple statements that lack implied claims, comparisons, or any real degree of linguistic complexity.

Curating Content

A 2013 survey found that 100 percent of top national news outlets and more than 90 percent of local news outlets in the United States allowed for users to write comments published below news articles.[81] While extremely prevalent online, such comments can be both a boon and a bane to news organizations. At their best they offer a space for users to exchange additional information, develop opinions through debate, and interact socially while building loyalty for the news brand. But they also raise concerns over the potential for vitriol and off-putting interactions that could push people away.[82] Some journalists see it as within their purview to act as conversational shepherds, shaping these online spaces to develop positive experiences for participants.

Moderating online news comments is a particularly challenging task due to the overwhelming volume of content. A recent visit to washingtonpost.com revealed articles like “Republicans Fear Political Risk in Senate Races as House Moves to Extend Tax Cuts” which had more than 2,300 comments. The challenge of scale is combined with the nuance and finesse moderators sometimes need in order to make effective decisions without stifling discussion. A variety of editorial decisions confront moderators, but perhaps most significant are those that reflect the exclusion of damaging, hateful, harassing, trolling, or otherwise toxic comments that could easily derail a debate, mislead people’s perceptions, or erupt into a war of words.[83] Here we see a problem suited to the deployment of automation to cope with the scale of content moderation. Rule-based auto-moderation has been in use for some years on sites such as Reddit,[84] but 2017 saw the emergence of data-mining-based classifiers to distinguish acceptable from unacceptable news comments. Almost simultaneously both the *Washington Post* and the *New York Times* deployed machine-learned models to help them automatically make moderation decisions about individual comments.

The *Post* has been collecting data for years on the actions of their moderators: as of late-2018 it was receiving somewhere on the order of about 1.5 million comments every month, of which about 70,000 received some form of attention from moderators. This data provides the raw material that the *Post* mined for signals, such as what types of words tend to reflect a comment flagged by the community as inappropriate. The system, called "ModBot," was then further trained by the comments editor, who provided feedback on cases in which the classifier disagreed with human moderation decisions. ModBot's classifier provides both a threshold and a certainty score for each comment it reads.[85] "When ModBot is extremely certain that comments should be deleted, we allow it to automatically delete those comments. When it's very certain that a comment should be approved, we allow ModBot to approve those comments. And then for anything in between we send comments on to the moderators," Greg Barber, the director of newsroom product, told me. A performance test on a sample of 3,796 comments demonstrated an accuracy rate of 88 percent. The classifier was initially deployed mainly in the "new user" queue for first-time comments made by newly registered users, which require moderation before appearing on the site. According to Barber, a good chunk of the comments in that queue are fine and don't violate any rules or norms, and so, he says, "We can rely on ModBot more heavily there [because] it requires less human decision-making skill." In the initial roll-out the *Post* has seen the amount of time that moderators spend in the new users' queue drop significantly, allowing them to redirect attention to comments that have been flagged as problematic by the community, or to pay closer attention to strategically important or controversial stories.

Much like the *Post*, the *New York Times* was sitting on years' worth of tagged comments data—more than 16 million of them. But unlike the *Post*, the *Times* has historically employed comment moderators to read each and every comment *before* it's published to the website. A staff of fourteen people makes this possible. Because of the heavy reliance on human labor, the *Times* could never allow commenting on more than about 10 percent of articles, otherwise staff would be totally overwhelmed. As of mid-2017, that figure had increased to 25 percent of articles, while using the same or even slightly less staff time, due to the *Times'* decision to partner with Alphabet's subsidiary company Jigsaw to develop an aptly named system called "Moderator." This system uses machine-learned classifiers to predict the tags human moderators would have applied. The algorithm grades each sentence of each comment with a score for each tag, such as “inflammatory,” “obscene,” or “off-topic,” plus a score Jigsaw developed called “toxicity,” and an overall summary score that aggregates the reject likelihood, obscenity, and toxicity scores. The worst scoring sentence for a comment becomes the overall score for the comment. When I spoke to *New York Times* Community Desk Editor Bassey Etim, he told me that his desk currently uses the scores only to make automated comment approval decisions. Rejection decisions are still made by people, although moderators’ work is greatly accelerated because the user interface highlights low-scoring sentences, which allows for quick scanning and rejection by human eyes. The interjection of human oversight mitigates concerns over technical limitations, such as an inability to discern profanity from harmful ideas dressed in the trappings of civil language.[86] Etim is aware that the algorithm isn’t perfect—there are at least a few dozen falsely approved comments each day—but also thinks it’s a fair trade-off and not terribly different from the types of mistakes human moderators have always made.

Sifting out the dreck addresses only one side of the online discussion quality issue. Top publishers are also interested in selecting and highlighting high-quality comments that set the tone for the site. The *New York Times* calls these “NYT Picks,” and they are meant to represent the “most interesting and thoughtful” comments on the site. The *Times* is still struggling to implement a data-mining solution to help automatically identify comments likely to be NYT Picks, but research I have conducted shows that there are a variety of quality-related scores that could enable this capability. For instance, there is a strong correlation between the rate at which a comment is selected as a NYT Pick and that comment’s relevance to the article or to the rest of the conversation.[87] Other dimensions of comment quality discernable in NYT Picks comments include argument quality, criticality, internal coherence, personal experience, readability, and thoughtfulness.[88] Scores such as readability, personal experience, and relevance were beneficial to journalists when presented in a prototype visual analytic interface called “CommentIQ,” which was designed to assess the metrics’ utility in interactively finding high-quality comments.[89] An ongoing challenge for computational linguists is to develop reliable data-mining techniques for numerically scoring a comment by a broader set of high-quality indicators.

Table 2.1. Journalistic uses of data mining, with supporting capabilities, and specific examples.*

Journalistic Uses	Data-Mining Capability	Examples
Find story lead via statistical surprise, anomaly, change, or deviation	Change and deviation detection (e.g., anomaly detection)	BuzzFeed Tennis Racket; Newsworthy Project; *Data Stringer;* Tadam; Klaxon; Election DataBot
Find or expand story via filtering for known patterns	Classification (e.g., interesting vs. not interesting)	Atlanta Journal Constitution Doctors & Sex Abuse; LA Times police reports; Local News Engine; BuzzFeed Spy Planes; Ukrainian Amber Mining; AP Gun Violence
Find story via connections between entities	Dependency modeling (e.g. find associations); clustering (e.g., find groups of related records)	Linkage; Dedupe; Bedfellows
Find story by counting	Classification (e.g., item to be counted or not)	LA Times campaign finance; The Echo Chamber
Find story via interactive exploration of data and documents	Clustering (e.g., grouping related documents); summarizing (e.g., labeling groups); dependency modeling (e.g., visualizing networks)	Overview; RevEx; */ new / s / leak*
Event detection in social media	Clustering (e.g., grouping related posts); summarizing (e.g., labeling groups); prediction (e.g., of newsworthiness, credibility)	Tracer; *City Beat*
Media monitoring	Classification (e.g., count instances via text)	Documenting Hate; *Police Shootings*
Event Prediction	Regression (e.g., predicted score)	Georgia Legislative Navigator; 538 U.S. Elections
Evaluate Source Data	Classification (e.g., compare predicted data category to official data category)	LA Crime Rates
Evaluate Source Credibility	Classification (e.g., type of source such as eyewitnesses); Regression (e.g., predicted credibility score)	*Seriously Rapid Source Review (SRSR);* Tracer
Evaluate Content Verity	Classification (e.g., of manipulated image or video)	InVid
Claim Spotting	Classification (e.g., identify claims worth fact-checking); Dependency modeling (e.g., find related claims that were already checked)	FullFact; ClaimBuster
Claim Checking	Regression (e.g., predict likelihood a claim is true)	*Knowledge base comparison* (see note 80, Chapter 2)
Comment Curation	Classification (e.g., appropriate vs. inappropriate)	ModBot; Moderator; *CommentIQ*

* Examples shown in italics are prototypes that could inform practice, while other examples are already in use in journalism practice.

Making Data-Mining Work for Journalism

Understanding the opportunities that data mining offer for story finding, event detection, prediction, verification, and curation is essential background for drilling into the consequences of data mining for the practice of journalism and for the broader provision of content by the news media. Four areas warranting further reflection on how to harness data mining for journalism include: the economics of content provision via automated analysis, the interface between journalists and data mining, gatekeeping and the role data mining plays in shaping coverage, and how journalistic routines absorb the evidence derived from data mining. I consider each of these in turn, as they relate in particular to core themes of sustainability, changes to practices, and journalistic values.

Ironing Out Economics

The deployment of data mining and automated analysis technologies in editorial tasks raises questions about the economics and sustainability of content provision. How is labor redistributed across human and algorithmic actors? What is the cost structure of story discovery? And how can news organizations use data mining to gain a competitive advantage in increasingly commodified news markets? The utility data mining provides can act both to complement human labor, which can increase the scope and quality of content, as well as to substitute human labor, which can speed production up or decrease overall human time spent.

Data mining can expand the ambition, scope, and quality of stories, and thus amplify the potential impact of journalism. At the *LA Times*, for example, campaign finance coverage was able to move beyond city elections to include more ambitious federal campaign finance stories, in ways that, given the volume of campaign contributions, would not have been possible without automation. Similarly, the crime classification algorithm at the *LA Times* allowed reporters to expand a manual pilot analysis back in time to cover almost a decade's worth of data. The *Atlanta Journal Constitution*'s "Doctors & Sex Abuse" story might have remained a state-level or perhaps regional story without the help of the attention-orienting classifier that allowed reporters to expand the story nationally. Here we see the value of deploying data mining to increase the geographic scope and ambition of stories, as well as to improve the quality of journalistic output, particularly in investigative scenarios where it can enhance the comprehensiveness of the investigation. For instance, the *New York Times*'s and the *Minneapolis Star Tribune*'s use of machine-learned algorithms to do record linkage resulted in more robust and complete investigations. In such complementary deployments of data mining, the level of human effort is more or less constant, but the scope and quality of output are enhanced, which in turn could lead to greater impact and more unique stories. Data mining can offer a competitive advantage to news organizations seeking to develop original and unique content—a valuable asset for building a brand in a crowded marketplace.[90] Chase Davis underscored this utility, “These techniques enable us to look at data in ways that get us stories we just couldn’t otherwise get.”

On the other hand, substituting human effort with data mining can speed things up faster than a person could ever go, as well as save the effort of people having to continuously monitor information sources. Again, Reuters’ Tracer sped up its news alerts in about 84 percent of cases, which is valuable given that it competes on breaking news information. In other cases, such as the Marshall Project or Tamedia’s monitoring of government websites, journalists can have information pushed at them via notifications rather than spending time every day looking at information sources and finding that there’s nothing new to see. The news leads sent by Newsworthy every month save time by identifying interesting patterns for follow-up. In some cases outlets might simply copy and paste the text from leads directly into stories. Labor substitution can also compress the timeframe for completing investigative projects. “It makes it quicker on something we probably would have done already,” noted Anthony Pesce at the *LA Times*, while also turning “a task that before we were using these techniques would have taken a year and ten people and a bunch of interns” into “three or six months with a handful of people and no interns.” An interesting question for future research will be whether the average amount of time it takes to complete an investigation decreases as data-mining tools become more widespread.[91] Rigorous evaluations of the tradeoffs in efficiency and accuracy between automated and manual or crowdsourced workflows will also be needed.[92] Depending on the specifics of an investigation, it may not always make sense to use a data-mining approach.

In curating comments, the *Washington Post* deliberately deployed ModBot to optimize for labor saving and, since moderators were spending most of their time looking at relatively benign comments in the “new user” queue, rolled out the technology there first. As a result of saving effort, moderator time was reallocated to comments that had been flagged by the community on the site or on stories that were considered strategic. In other words moderator attention could be shifted to where it was more needed and more valuable. Ideally, the *Post* would like to reallocate moderator attention to highlight quality comments on the site. “If we would somehow find the magic bullet that took care of all of the comments that we needed to get off the site, I would still want to spend exactly the same amount of time and money finding good stuff,” explained Greg

Barber. At the *New York Times*, Bassey Etim explained that as a result of rolling out their Moderator tool, “We’ve lost a few hours to some other efforts that got redirected, so we’ve probably got a little bit less staffing.” He estimates the amount of staff time being siphoned into other projects at about 10 percent but notes this may continue to fluctuate. A key question that should concern media management is how human effort is reallocated in light of tasks that may be substituted by automation.

Whether the combined human-computer system allows for human time to be invested in increasing scope or quality of current tasks or to be saved and reallocated to other tasks, it’s important to acknowledge that new costs and forms of labor also accrue in creating such systems. A full accounting of costs must take into consideration the time to develop or adapt data-mining techniques, learn how to use them, maintain them, and assure the accuracy of their output. Development costs are amortizable, but it’s not always clear up front how long a system will be useful or how large a scale in output it will enable. Development often entails doing a trial project, which may be costly in terms of human effort. “If it turns out to be useful in a more general way then we basically try to productize it,” explained Friedrich Lindenberg. For the Newsworthy lead generation service, Jens Finnäs noted that each new data domain where they want to provide news leads demands an upfront investment of time in order to do topic-specific analysis. Otherwise the leads won’t be meaningful or useful. Productizing data mining takes a substantial amount of effort because it involves generalizing and parameterizing techniques so that they can be used across various stories. News organizations that are able to build data-mining capabilities into products that are applicable to different scenarios will be able to amortize the cost of development of these tools across many stories.

Costs are also introduced as other forms of labor that emerge. Datasets must be labeled in order to train algorithms such as classifiers—tedious work that also requires careful consideration and rigorous application of classification definitions and criteria. In some cases datasets may be categorized as part of ongoing data journalism efforts and these sunk costs can be leveraged into other stories by applying data mining to the labeled dataset. Anthony Pesce notes that this labor is often spread around, so that “when other reporters pitch ideas they’re asked to pitch in on some of the manual classification work.” But these are also prime tasks for lower-skilled workers, such as interns, who are overseen by experienced reporters and editors.

Building Smart Interfaces

Reaping the rewards of data mining will require the adaptation of news production practices. In order to integrate data mining effectively into human processes, designers and developers need to create interfaces to support journalists in their various workflows. Relevant user interface issues include how to signal whether a news lead is worth the time and energy to pursue; how to filter for the most relevant leads; how to engender appropriate trust, reliance, and provide context to enable expert editorial decision-making based on the lead; and how to ameliorate information overload and alert fatigue while encouraging autonomy of human workers in their interactions with such systems.

News leads should enable journalists to make an informed decision about whether a given lead is worth their time and energy to pursue. They must be salient, credible, and framed in a way that highlights their potential value and impact while including essential context. The data-mining algorithm is essentially pitching a story to an editor, much as a freelancer would. A typical pitch might quickly communicate what the story is about, why it's significant, what the take-away might be, and what the sources are. An important aspect of story discovery tool interfaces is the degree to which they enable configurability in terms of what data and documents are monitored as well as how leads are filtered according to interests, topics, aspects of newsworthiness, or domain-specific concerns. For instance, one issue for journalists using the fact-checking leads provided by the Duke Reporters' Lab is that the system doesn't allow users to filter leads based on the type or identity of the speaker of the claim. For fact-checkers, *who* made a claim is an important modulator of newsworthiness that could be made more salient or filterable in the interface.

Some lead generation systems have begun to grapple with the challenge of presentation by developing multimodal and interactive interfaces. Leads from Newsworthy, for example, include several summary sentences of text about the anomaly, plus a line chart to provide visual evidence and context, and perhaps most importantly the data so that reporters can interpret results for themselves (see Figure 2.1). For the fact-checking leads from Reporters' Lab, a link provides easy access to the source material including a transcript which allows reporters to assess the context of a statement before pursuing it further. Lead presentation is not only scenario dependent, but it can also be topic-dependent: the relevant context could change if a reporter is looking at crime, unemployment, or education data. The Stacked Up project, for example, embeds specific domain knowledge about the appropriate provision of educational textbooks in the city of Philadelphia in order to suggest mismatches between an expectation based on regulation and a reality as conveyed through public data. It produces and presents leads as interactive data visualizations "designed to answer the most common questions a reporter might ask in order to assess whether a story might be found at a particular school."[93] What's clear from all of these early efforts is that ongoing research will be needed to design and evaluate effective information displays and user interfaces for automatically produced news leads in different journalistic contexts.

Another challenge related to the interface between data mining and journalists has to do with information overload and alert fatigue. Friedrich Lindenberg cuts right to the chase: "A lot of our reporters are already incredibly busy. And if we now give them 10,000 possible cases of corruption they're just going to tell us to fuck off." It's clear that there's value to having a data-mining algorithm identify a potential case of corruption, but a balance is required in how many leads to offer, and of course how to present them so they don't feel overwhelming but rather enabling. If there are too many false positive leads, reporters might even be habituated to ignore them. One recipient of the Newsworthy leads explained that their newsroom could probably only handle one good lead per month. Even one or two leads per week felt like too many, since each could take up to a week's worth of effort from a reporter to mature into a story. The leads are really just statistical observations, and so time is needed to do the reporting and local sourcing necessary to find impacted individuals and turn the lead into something compelling that people would want to read about. The appropriate volume of leads to provide may be some function of how much human effort it takes to chase a lead in relation to how much human effort is available, which itself can vary according to the news cycle.

The Newsworthy project is experimenting with ways to ameliorate story fatigue by using additional editorial logic to ensure news leads are sufficiently novel. As Finnäs explained, "If there's the same story happening multiple months in a row, we won't deliver it for several months in a row, but wait at least three months before we send it out again." He says it would be easy to create one news lead per municipality per month, but the service intentionally doesn't do that in order to avoid alert fatigue. Instead there's roughly a 10 to 30 percent chance of a news lead for any particular municipality turning up in a given month, and so if a journalist is monitoring ten municipalities, then he or she would get no more than a handful of story ideas per month. Another way to cope with the eventuality of too many leads could be to lower the amount of skill it takes for a human journalist to assess a lead as worthy of pursuit. In other words, if less skilled workers can evaluate news leads as a first step, the most promising leads could then be passed on to seasoned journalists for further investigation.

Shaping News Coverage Algorithmically

Gatekeeping describes the idea that some bits of information make it into the news, while others don't. Different forces impact the flow of information to end-consumers including influences at the level of individuals (cognition or background), routines (patterns of work), organizations (media ownership), social institutions (forces outside an organization), and social systems (cultural or ideological factors).[94] These are the "enduring features of the social, physical, and digital worlds" that shape the gatekeeping function.[95] Gatekeeping theorists acknowledge that "any technological innovation, once adopted, offers routine paths for news organizations to select and shape the news."[96] How will newsroom adoption of data-mining technologies, in particular, shape coverage in significant ways?

By orienting attention and reducing the costs of finding certain types of events or stories, data-mining algorithms provide information subsidies that can influence how journalists end up covering various topics and beats, which in turn will shape the news available for public consumption. Journalists should remain cognizant of how the design and development of data-mining algorithms may affect coverage and consider how journalistic news values and, ideally, public interest values may or may not be reflected in those algorithms. Because algorithms will shape the attention of journalists and ultimately of coverage, society should also be vigilant with regard to the ownership concentration and diversity of editorial perspectives exuded through those algorithms. One hundred or one thousand variations of story-discovery algorithms will be preferable to having "one algorithm to rule them all."[97]

The ClaimBuster system provides an illustration of how data mining could impact coverage. The system was evaluated by comparing the topic of claims identified for fact-checking by the algorithm to claims manually selected for fact-checking by CNN and PolitiFact. Findings showed that ClaimBuster identified more claims about the economy and fewer claims about social issues (see Figure 2.4). If reporters were to solely rely on ClaimBuster to identify and guide attention toward check-worthy factual claims, this could decrease the attention fact-checkers give to social issues—an outcome that may not be desirable from a public interest standpoint. One possible solution is to train claim-spotting algorithms on data from different news organizations, allowing them to more easily align results to the predilections of various editorial outlets.[98] Another ClaimBuster evaluation on the twenty-one transcripts of US presidential debates in 2016 showed that Donald Trump had fewer check-worthy factual claims than Hillary Clinton. Combined with the observation that the ClaimBuster system heavily weights the presence of numbers and figures in its selection of claims, this suggests that Trump's rhetoric and mode of communication may have made his statements less susceptible to being highlighted by the algorithm. As automated fact-spotting techniques become adopted in practice, it will be important to assess how they impact the coverage of various types of stories, claims, events, and modes of political rhetoric. Journalists will need to become more cognizant of the ways such algorithms orient (or divert) attention in characteristic ways and be able to fill in the gaps as needed.

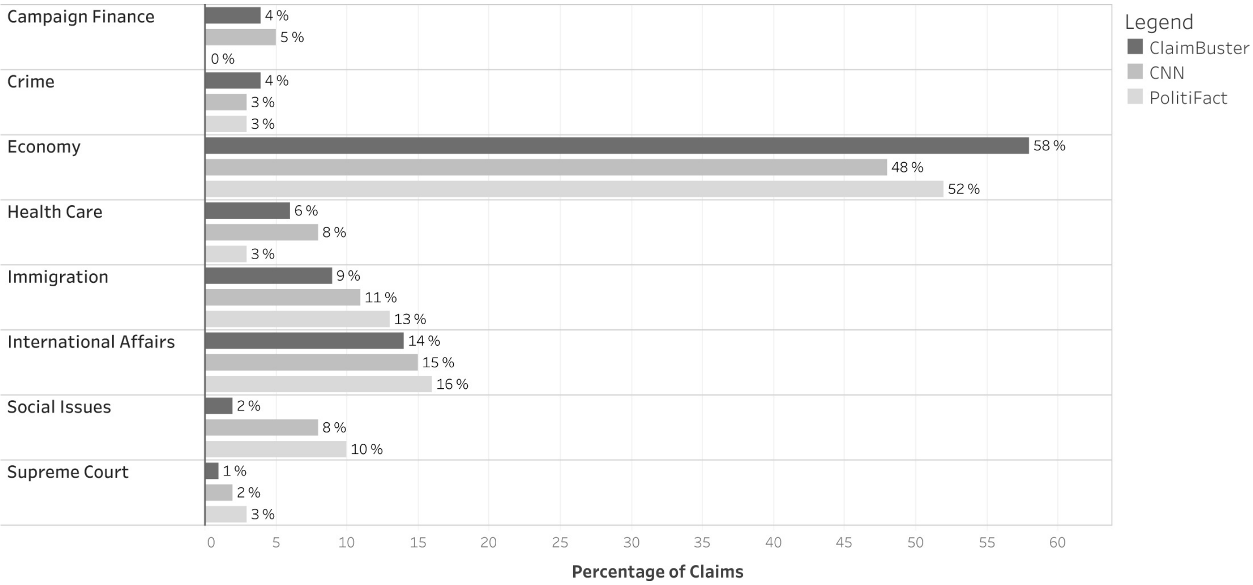


Figure 2.4. A comparison of the topics of claims spotted by ClaimBuster, CNN, and PolitiFact. Note that ClaimBuster spots fewer claims related to social issues, immigration, international affairs, and the Supreme Court, whereas it spots more claims related to the economy. *Source:* Original chart produced by author using data published in N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster," in *Proceedings of the International Conference on Knowledge Discovery and Datamining (KDD)* (New York:, ACM, 2017).

Definitions and their computational operationalization are important factors in how data mining influences attention and ultimately the shape of coverage.

Redefining a metric could lead to a system surfacing entirely different facts.[99] Defining what sources an event monitoring system tracks leads to sourcing biases. The Social Sensor project, for instance, found an inclination toward male or mainstream sources based on the algorithm that had been defined for finding sources to track.[100] Definitions can also shift from use-case to use-case. The Newsworthy lead generator varies the definition of newsworthiness by topic because the semantics of the data impact whether, for instance, a peak is more interesting (as for crime) than a trend (as for unemployment). Geographic context can also impact newsworthiness—a trend in one city might not become interesting until it's put into contrast with the trend in a neighboring city.[101] Often an appropriate definition of newsworthiness requires a good bit of domain knowledge. "A lot of it depends … [on] my experience and knowledge of what political fundraising is like at the federal level," explained Derek Willis. Definitions can limit the adoption of a data-mining system if the definition embedded in the system by its designer does not align with the definition of someone who wants to use the system.

All too often we measure what is easy to measure rather than what we really want to (or should) measure. As prototypes such as City Beat have shown, simply defining newsworthiness according to popularity may be straightforward to do computationally, but isn't so popular with journalists.[102] Definitions other than popularity may be preferable, but if they are not technically feasible to encode into an algorithm, then simpler definitions may prevail. Consider for a moment the challenging proposition of trying to monitor government for stories that are newsworthy by investigative standards. Investigative journalists are typically oriented to a range of problems related to the breakdown of delegated decision-making.[103] These include issues of effort (waste, mismanagement, neglect), money (bribery, embezzlement, theft, corruption), advantage (nepotism, patronage, conflict of interest, rent seeking, influence peddling, favoritism), power (abuse, harassment, misconduct, discrimination, misuse), and information (fraud, deception, and misleading). In some cases, such as fraud, statistical and machine-learning techniques can and have been used to help detect and alert investigators to the issue.[104] But other issues such as political patronage (that is, using state resources to reward people for their political support) might present a greater challenge to a classification algorithm. The information needed to detect something like patronage may simply not be quantified or straightforward to represent in such a way that an algorithm can detect it. Some types of newsworthy stories may ultimately be harder to define in ways that can be written into code, with the result that those types of stories might receive less coverage because algorithms can't pick them up.

Creating Journalistic Knowledge

Epistemology refers to the philosophical understanding of how knowledge is produced by combining evidence and justification to arrive at claims about the world.[105] More simply: How do journalists know what they know? Traditionally journalists have created knowledge by drawing on sources such as eyewitness accounts, interviews with experts, and document inspection. These sources play into a sort of credibility calculus as journalists piece together evidence and seek justifications for what they eventually report as the news. Data-mining algorithms provide a new source of evidence—tips, leads, relationships, and predictions—for journalists to incorporate into their epistemological practices of seeking the truth. But just as with other sources, the knowledge provided by algorithms needs to be evaluated and appropriately weighed in relation to other modes of evidence. While some scholarship has explored the epistemological implications of the use of "Big Data" in journalism, the focus here is more squarely on the role that algorithms play.[106] In particular, how do data-mining algorithms change knowledge production practices for journalists?

Coping with Uncertainty

For many of the projects surveyed in this chapter, the journalists I spoke to expressed the need for ongoing skepticism and verification of the outputs of data mining. A data-mining algorithm is a source like any other, and reliance on it can increase over time as journalists gain familiarity. Jeff Ernsthausen explained how he used a continuous quality assurance (Q / A) process applied to the output of the *Atlanta Journal Constitution*'s sexual-misconduct document classifier by periodically sampling one hundred documents with low "interestingness" scores and reading them manually to make sure the classifier wasn't systematically missing important leads. Evaluating the reliability of results was a challenging aspect of the process: "The hard part is understanding how to generate the things that tell you whether it's working well or not," explained Ernsthausen. Anthony Pesce concurred: "We spent probably a couple weeks looking at the code, going back and forth tweaking the model making sure it was working right, re-testing it on our test and our training data. And then went back through and just looked at a huge sample of them." For their campaign finance project, which they have been pursuing on an ongoing basis, the *LA Times* put in a lot of time upfront to build confidence in the system, but "now that we've used it so much, we're very comfortable with it," explained Pesce. The *Washington Post* does periodic "spot checks" on the output of ModBot to manually examine the comments it has deleted or approved. In other words, substantive effort must be applied toward evaluating evidence to convince editors that model results are reliable. Journalists must be convinced that the outputs of a system are reliable and in line with their own goals. This effort can, however, be amortized if the software is reused by different newsrooms or the same newsroom over time.

In other cases, data mining is presented to journalists not for their acceptance per se but instead to continually remind them that it should be treated with uncertainty. At the *New York Times*, the campaign finance disclosure record-linkage algorithm presents its results in a sorted list. The user interface provides an initial set of matches that have high match certainty, but less certain matches can only be accessed via a link. "We don't want reporters to look at the output of this and assume that it's correct. We want to use it as a tool where they can kind of like get a little bit more information but the interface also makes clear that this is not vetted," explained Chase Davis. "The role of the algorithm … is essentially to make the vetting easier," he added.

Journalists acknowledge that data-mining algorithms make mistakes and have inherent statistical uncertainty, but that there are conditions in which that is tolerable or can be overcome. An acceptance of false negatives—documents that should have been classified as sexual misconduct but weren't—meant that the team at the *Atlanta Journal Constitution* knew there would be cases that they missed. This in turn meant that whatever claims they reported publicly about the magnitude and scope of the problem would be a lower bound in terms of the actual number of sexual misconduct cases in the country. At the same time, they also knew that the effectiveness of the classifier varied from state to state based on the quality and verbosity of documents produced by different states. This put limits on their ability to make certain types of claims, such as comparing the rate of sexual misconduct cases between different states. Davis explained the idea with respect to the *Times'* campaign finance project: "It's very unlikely that no matter how fastidious you are that you're going to catch absolutely every last thing.… Whenever stories like this are written, there's always some amount of hedging just understanding how messy the data is and how hard it is to be able to get absolute certainty out of it." Uncertainty from data mining can also subtly steer the editorial focus of an investigation. Ernsthausen described a real estate investigation where the focus shifted to larger entities that were less sensitive to geocoding errors. Accepting the limitations of data-mining techniques means modulating the nature, strength, and presentation of claims pursued and ultimately published.

Contingencies on Claims

In their book *Custodians of Conscience* James Ettema and Theodore Glasser note the contextual variability of what constitutes an adequate justification for knowing within the epistemology of journalism: “The criteria for adequate justification may vary from one context to another” and “What counts as sufficient grounds for a knowledge claim varies from one domain of inquiry to another.”[107] How then do variations in journalistic context impact the uptake of information produced via data mining? Relevant variations in context include whether a knowledge claim from data mining is made publicly or is used only internally in a newsroom, whether it’s possible to corroborate a piece of predicted information, and what the implications and consequences are for individuals or other stakeholders if a particular fact is published.

There are greater demands on justification and verification the closer an algorithm is to direct publication of information. Public claims must be justifiable as judgments that are fair and accurate. “It gets a bit tougher the closer the … data mining technique or the algorithm is to a publishable product, the more I think you have to understand about it,” explained Davis. In comparison, claims that are used internally in a newsroom or are leads that will not be directly published can be assessed by other reporters or an editor before any claim is made in public. If the data mining is wrong or produces a high degree of uncertainty, but the public doesn’t see it, then at worst it wastes some internal newsroom effort. Certain types of ML are easier to justify to editors or the public—often referred to as “explainable models.” An example of an explainable model is a decision tree: its classification decisions can be expressed in terms of simple rules that apply to each dimension of data.[108] Davis related some of the benefits of explainability in justifying a model: “I’d use things like decision trees if I was modeling something out and where you can actually at the end of the day if you really really wanted to you can print the decision tree on paper and see exactly how it works. And I could walk an editor through that in human terms.” Explainable models can be useful aids for communicating and building credibility for knowledge claims stemming from the model.

In some cases journalists can justify the output of data mining by corroborating it with evidence from other reporting methods. At the *LA Times* they contacted the LAPD about the result of their model—that serious assaults had been systematically underreported in Los Angeles—and got confirmation that the model was correct. “We felt pretty good about it at that point. If they’re saying, ‘We’re not going to contest this, … that’s pretty much all we needed,” explained Pesce. The LAPD had just finished an internal audit of its data and had found the misclassification error was even higher than the *LA Times’* model suggested. The fact that the result of the model was corroborated helped justify its use as evidence in the story.

However, sometimes it’s simply not possible to confirm, corroborate, or deny evidence from data mining prior to publication. Take, for instance, the various predictions relating to outcomes of elections or legislative activities. These predictions, by definition, cannot be confirmed until the predicted event takes place. In these types of cases, transparency of the data-mining method is often provided, offering descriptions of how the method works, what data it operates from, how constructs are operationalized, what the performance and error rates are on test data, and sometimes even including open source code repositories to facilitate reproducibility.[109] For instance, for the *Atlanta Journal Constitution*’s Legislative Navigator App, a series of blog posts described in great detail how the predictive model performed on past data. In attempts to reflect the provisional nature of predicted information, some journalists are experimenting with communicating uncertainty directly to the audience. The *New York Times* took this approach in its 2016 US election prediction. As voting results were tallied on election day, a dynamic prediction was depicted as a dial showing the chance that either candidate would win. To convey the uncertainty in the underlying statistical prediction, the dial’s needle jittered, especially early in the evening when every new precinct reporting could cause a swing in the model’s prediction of the close race.[110]

In the BuzzFeed investigation into professional tennis cheats, the statistical anomalies could likewise not be corroborated through other forms of evidence. As a result the story did not include the names of the players identified. In essence, the evidence from the statistical tests was only suggestive, justified through an open-source code release and anonymized data, but ultimately deemed not solid enough to publicly label any particular players as cheaters. As this case illustrates, the potential consequences of public claims also factors into how journalists come to rely on information produced by data mining. When publishing a claim that impacts an individual in a negative way, it must be clearly justified based on the available evidence. If a classifier indicates that a politician is 80 percent likely to be engaged in banking fraud, that may not be a high enough level of confidence to move forward with a public statement, given that such an indictment could have severe consequences for the individual. When “we’re talking about people getting in trouble, it’s more important for us to ensure that the 500 people we’re putting in there are legit,” explained MaryJo Webster, underscoring the additional consideration given to validity and justifiability when publication could negatively impact individuals. This starkly contrasts with typical scientific knowledge claims, which often result in empirically developed theories describing central tendencies of a sample rather than assertions about individuals that can produce negative social consequences. On the other hand, if a model is being used to produce information for entertainment purposes, as is the case in FiveThirtyEight’s sports predictions, the consequences of being wrong are far less onerous. Chase Davis contrasts these two situations: “For the really hard investigative stuff that some of this [data mining] is occasionally being applied to … you’ve got a different standard that you’re trying to meet.”

When journalists have low tolerance for statistical error and need absolute certainty because of the import of claims, they mostly fall back on the manual verification of data-mined results. A “no false positives” mantra is actualized by applying a manual check to any evidence or claim supplied by data mining. In investigative journalism, “The reporters are ultimately going to want to vet everything themselves by hand to ensure that it’s correct [and] to ensure that they understand it,” noted Davis. For cases that their campaign finance model couldn’t classify, or where the classifier had low confidence, the *LA Times* went back through each of them one by one to see if they could be classified by hand. By manually checking each outcome, journalists are able to catch errors, such as if a person with a common name was accidentally associated with the wrong person in another database. The reliance on manual methods to justify making socially consequential claims in public also has an impact on the way data-mining models are parameterized. “We can essentially tune so that it would give us either more false positives or more false negatives just depending on the threshold of confidence that we’re searching for,” explained Davis. “Knowing that we are going to be reviewing things anyway we essentially wanted the algorithm to be a little bit overzealous.” Algorithms can be tuned to minimize false negatives, increasing confidence in the comprehensiveness of an investigation, while knowing there will be a manual step at the end to catch any false positives.

Data mining offers a whole host of opportunities that are only just beginning to be explored and exploited for editorial purposes. From finding stories to monitoring or predicting events, evaluating content and sources, and helping to curate discussions, the editorial utility of data mining is gaining increasing purchase in newsrooms. Data mining has the potential to transform how leads are developed, to alter the economics of content production, and to change how knowledge itself is produced. As it becomes more woven into practice, data mining will ultimately shape coverage, reflecting whatever values of newsworthiness its designers have thought to include. Data mining is still just a part of the equation for automated news production though. Algorithmic analysis may be helpful for finding the story, but algorithmic content production will be needed for telling it.