StatFury LLC | June 2019 - January 2020

Whitepaper

Impact With Context:

Tool To Measure News Impact

Reynolds Journalism Institute

Non-Residential Fellowship 2019-2020

Leezel Tanglao

Abstract

Measuring impact is subjective and disorganized and often lacks context. There's no unified way to look at the true (contextualized) impact of a story other than searching for citations, syndications and potency for popularity on social media. Impact, however, must be seen in a more holistic way through the lens of the diversity of views and its implication for underserved communities and business models. For example, gene-editing has a tremendous impact on science. What does its impact look like for those who cannot afford it and who gets left behind in the conversation? Impact includes not only the depleted and narrow heuristics of hits, views and ads but also the semantic information of related media and the views espoused and/or insinuated. There is not just one narrative, there are multiple. To highlight these narratives implies improved news articulation and the restoration of trust from the public in our ability to dispel fake news and highlight holistic reporting.

It's not enough to publish a story and distribute it across all platforms and hope it gets seen by the audience you are targeting. If you don't measure impact with context, how can you better understand your audience and take the next actionable steps? The tool I've built during my Reynolds Journalism Institute Non-Residential fellowship, Impact With Context, will help complete the storytelling cycle by understanding the magnitude of a story's reach, engagement, diversity of sentiment and views by breaking down what resonated with the readers and what triggers prompted readers/viewers to take a certain action in real-life. It aims to take a step further by uncovering opportunities for potential story follow-ups, profiles and events. It categorizes the impact in sectors (financial, social, politics etc) and groups (gender, race, disabilities, etc).

The Impact With Context tool is intended to serve a more objective, actionable, holistic and contextualized view into how a story impacts the community by ingesting several streams of information and analyzing it with context. The tool addresses this challenge by tapping into multiple online databases and APIs to measure a story's reach (who/where), engagement (what), history (when) and sentiment (how). This tool differs from tools like Chartbeat because it provides a context to better understand how a story affects readers' lives versus just tracking clicks, referrals and recirculation.

Further iterations of the tool include plans to create a news consumer view. By showing the work that went into a story, the public gets closer to understanding why a free and open press is crucial in combating disinformation and fake news.

What We'll Cover

In this whitepaper, we'll cover the goals and challenges of this project, current landscape and technical next steps to continue to improve and iterate better ways to measure impact of news stories after they are published in the real world, beyond vanity metrics.

- Impact With Context Project
- Literature review
- Competitive review
- Case Study
- Methodology
- Tech specs of the tool
- Feedback
- Next Steps
- References

Impact With Context Project

Challenges, Goals and Deliverables

There are many tools that measure "impact" from a quantitative perspective -- usually including company KPIs (page views, uniques, video starts, ratings etc). But these tools do not capture impact, a highly qualitative idea, with sufficient granularity. What is lacking is a tool to identify and track impact in a more holistic way which would include actions beyond vanity metrics. However, there are many variables and challenges to this. How would you track interactions beyond what is posted online such as phone calls and on the ground feedback? Most newsrooms today are lean in staffing and organizations simply do not have time to keep track of every interaction unless a staffer was assigned solely to this. Also the limitations to certain APIs that are cost-prohibited and contain restrictions are impediments in collecting all potential impact.

My hypothesis was that we can build a tool using the power of semantic relationships to scour APIs and databases to better represent the impact of articles after their publication. What currently exists relies heavily on keyword(s) aggregation, which is the foundation for search results but is not the only factor that needs to be incorporated when extracting "impact" in a fast and automated way.

My goal for this project was to answer the following question: Can we build a tool to better identify impact (in the real world) from news articles after their publications? Impact being defined as an action taken in the real world (public comments in city council meetings, people starting crowdsourcing campaigns) - or also subsequent written articles?

In pursuit of this challenge, I partnered with UCLA's Samueli School of Engineering to help me conceptualize and build out the technical end of the tool as I served as the project manager and provided the editorial and product direction. The project started in June of 2019 as we kicked off weekly video sessions to go over product requirements and operated on one to two-week development sprints to first identity best approaches to tackle this problem using artificial intelligence (AI) and natural language processing (NLP) to scour news articles, identify impact and use those identified relationships to crawl all available APIs to surface more relevant results around impact after a story's publication.

As part of this eight month fellowship, we are able to deliver the following and future plans:

1) **Whitepaper**
2) **Published article industry media** (CJR/Nieman lab/poynter/API), outside media where applicable and possibly in an academic paper.
    a) RJI article was published on the tool's goal: https://www.rjionline.org/stories/rji-fellow-developing-tool-to-measure-real-life-impact-beyond-clicks-and-so
    b) Currently looking into opportunities to publish in journals and present at conferences
3) **MVP/prototype** of the metric tool with sample data from test newsroom and consumer focus groups
4) **First live iteration** of metric in beta - This would be the next step pending more funding to continue development, testing and research.

Timeline:

- June-July 2019 - Research and work begins
- August-Sept 2019 - Backend work begins
- October 2019 - Version 1 released for feedback
- November 2019 - Feedback for version 1, bugs fixed
- December 2019 - Version 2 released, bugs fixed
- January 2020 - Version 2, sent for testing and feedback, wrapped up any bugs and plan for next steps to continue development

Literature Review

The Merriam Webster dictionary defines "impact" as the "force of impression of one thing on another: a significant or major effect." For the purposes of this fellowship project, "impact" is defined as "an action taken in the real world after a news article has been published. Real world action can mean public comments at a city council meeting or initiating a crowdfunding campaign. Social media sentiment and engagement on platforms is not included in the tool at this time due to the focus on surfacing impact pulled from other primary sources. In the tool's roadmap, social media sentiment and engagement will be integrated and added to the assessment.

Measuring impact has been a challenge well documented in the journalism industry. In "Measuring Impact: The art, science and mystery of nonprofit news," by Charles Lewis and Hilary Niles defines impact as "a social outcome proves a complicated proposition that generally evolves according to the constituency attempting to define it." Lewis and Niles point to the many variables that add to the complexities of this assessment: many stakeholders with diverse reasons for measuring impact. However, even more importantly before trying to measure impact, just defining it, there's no unified agreed upon system across the industry.

Lindsay Green-Barber of the Center of Investigative Reporting tries to dissect this in a 2014 Nieman lab article, "How can journalists measure the impact of their work? Notes toward a model of measurement," where four areas that "need to be addressed: the need for standards, the relationship between audience engagement and impact, the significance of online analytics, and the difficulty of understanding how qualitative offline outcomes figure in to the long-term impact of media." Green-Barber points to what has been considered the definition that has been widely accepted: "large-scale changes, such as the adoption of a new law or a government investigation." While this is true to what impact is, it isn't the only definition that should be considered. Only relying on large scale changes to prove impact undercuts other types of impact that can result from the publication of the story.

The lack of tracking all types of impact, both small and larger can prove problematic when telling the impact narrative to both funders and the general public. "All around the world, media outlets are learning that some funders are uncomfortable with supporting journalism merely as a "public good." They want to see proof of impact," Anya Schiffrin and Ethan Zuckerman wrote in their article, "Can We Measure Media Impact? Surveying the Field" for Stanford Social Innovation Review in 2015. Schiffrin and Zuckerman rightly point to the other challenge of introducing measuring impact as part of the journalist's workflow. "The task of "proving impact" doesn't come naturally to most journalists. They reject a utilitarian view of their worth, preferring to believe that news is a public good that merits support for its own sake. They view themselves not as campaigners for a cause but as fair and impartial observers," they wrote. To that end, there are organizations and movements to integrate impact as part of a journalist's role such as the Solutions Journalism Network, which hopes to spread the "practice of solutions journalism: rigorous reporting on responses to social problems."

Competitive Landscape

Measuring impact isn't a new concept. Nonprofits like Reveal created an impact tracker and startups like Chalkbeat have made their impact metric open source (Chalkbeat MORI - Measuring our Reporting Influence) and free for other groups and organizations to build upon. NewsLynx, an "open-source platform to track, store, and analyze impact," still requires some manual tracking. In terms of an automated solution, ICFJ Knight Fellow Pedro Burgos came up with Impacto, which crawls social networks and government pages for mentions of the news outlet. However, there is still not one tool that is able to capture more than just searching for keywords.

The challenge of the similar impact metric trackers out there is that they are manual and subjective to the person making the call on what type of "impact" a story has. The Impact With Context tool takes it a step further by pulling in APIs and data sets that measure sentiment, historical data by looking at relationships and surface up relevant results that can prompt recommended follow up stories for newsrooms to consider. Impact isn't limited to one big investigation or breaking news story. Impact is ongoing but often after the big initial publication and push on bigger stories, the impact isn't always revisited on a regular basis or looked at through a different lens, particularly when it comes to underserved and underreported communities. Unless a newsroom had a team just decided to that one area all the time, impact is typically monitored heavily the first few weeks after a story has been published.

Case Studies

User Problem Narratives

The Associated Press, the world's largest news wire agency, distributes about 2,000 pieces of content a day to its customers. Its customers include print, broadcast, radio and online publications. The AP also serves a number of non-journalism related customers as well. However a major challenge for many wire services such as The AP, is tracking the impact of its stories. Due to the many different ways customers receive AP content, it is a challenge to track what's being used, how it's being used and how well it does in terms of traffic under one system.

In the fall of 2018, I tried to best collate the many sources of information and tracking available to the AP as the company's first Audience Development Lead. I was embedded in the Health and Science department and I was tasked to understand who and what our audience (in this case our customers), wanted and what stories performed well. Due to the different delivery methods and platforms, this proved to be a challenge not only on the day to day news cycle but also measuring longer term impact of stories and major investigations. Because of the dilemma, I started exploring if there were ways to better automate some of the sources were pulling from. This is where the formation of Impact With Context tool came into focus.

In previous roles at CNN, I had access to many tools that measured high level vanity metrics such as page views, video starts and unique visitors. At the time and even now, there's not a more automated way to search for impact in short of keeping manually records. When the challenge presented itself during my time at the AP, I knew this was the time to get serious in finding a better way to scrap this information. There was another incentive -- grant funding reporting.

As news organizations as a whole look for diverse ways to fund their journalism, many have turned to grants from foundations. As standard practice, many of these grants have reporting requirements that often include the impact of the funding provided. This was the case with the AP as well as at HuffPost.

HuffPost's impact team wanted to track the interactions and actions around articles published as part of their series, This New World, which includes "stories of progress toward building an economy that works for everyone." This team had a small staff but still needed to keep track of mentions, link backs and ways HuffPost's report made an impact. In both cases, the user need was clear: measuring impact is not straightforward as vanity metrics and that certain information that would signal impact are not easily scrapable or available online. With this in mind, I started a quest to try to alleviate this pain point in many newsrooms by leveraging technology to better surface "impact" content up so that organization don't miss out on reporting the impact of their work more holistically.

Methodology

In completing this project for my non-residential RJI fellowship, I employed the following methodology and processes: literature research and review, product and development sprints to flesh out requirements with my technical partners at UCLA's engineering school (Professor Vwani Roychowdury and UCLA graduate student Pavan Holur), recruitment of beta testers and interviews with feedback test participants. After the tool was scooped and groomed, development began for both the front end consumer facing view and the back end algorithms. After the initial round of feedback, more improvements to the tool were deployed and a second round of feedback was initiated and documentation for further development.

Tool Tech Specs

Define the Categorical Impact Metrics

Before technical work began, we first had to define "impact" metrics. In partnership with UCLA's Professor Vwani Roychowdury and UCLA graduate student Pavan Holur, we looked at impact in categories of news coverage.

At a high level, we defined actionable "impact" of a news article as a semantic trajectory traversed in journalism and/or other media after the publication and/or interaction with the story.

What is considered an "action" in the real world?

Action includes but not limited to:

- Protest
- Rallies
- Op-ed
- Art creation in response to
- Stock market reaction

- Investment in or divestment
- Crowdfunding campaigns
- Creation of new product or service
- New research
- New laws/legislations
- Public comments
- Petitions (example: change.org)
- Increased media coverage
- Increased or decreased in graduation rates
- Increased or decreased in applications (jobs, university etc)
- Increased or decreased in crime rates
- Increased or decreased in population rates
- Increased or decreased in property values
- New/increased/decreased jobs
- Reaction to Weather or natural disaster

Once we had defined the categories of “impact” we reviewed and accessed APIs and databases, in order to build a corpus of knowledge from which the impact could be parsed. Some of these APIs included weather, stock market data, traffic and government databases.

Next I came up with a vocabulary bank of “impact” phrases used in stories. Every news story has some type of “impact.” Some of these include, but limited to words such as:

- Voted to ban
- Voted to approve
- Approved ban
- Approved new law
- Invested
- Divested
- Arrested
- Filed lawsuit
- Indicted
- Turned down
- Discovered
- Vote of confidence, vote of no confidence
- Raised/Lowered prices, wages, requirements
- Scored low/high
- After report by
- Citing report by
- Call for investigation
- In response to
- Issued a statement questioning report
- Refuted report
- Pledged to/not to
- Challenged
- Started petition
- Opened inquiry
- Spoke out against

With these impact terms, the team explored default summarizers based on word aggregation to extract the impact of a story in the semantic sense. We experimented with real world examples to improve and to adjust the summarizers. I provided the typical story workflow to help my tech partners understand the story creation process from beginning and end. This served as the first step in developing the pipeline for categorical impact identification.

Typical Story workflow

1) Reporter pitches/editor assigns story →
2) Reporter writes and reports story →
3) Editor edits story →
4) Story published →
5) Story posted on site, social platforms →
6) Story shared, read, liked by audience →
7) Reporter moves on to the next assignment →
8) Reporter may or may not follow up depending on resources and time

With this in mind, it's clear that there's a need for smarter tools to assist with the storytelling process from the instant when the story is published to when/if a follow up is published. Across the industry, newsrooms are lean and typically can't afford to allocate a reporter to solely a single topic and they are often juggling many stories at the same time. But impact doesn't stop after a story's publication. How can we make it easier for editorial teams to track and uncover the impact of their reporting and their stories?

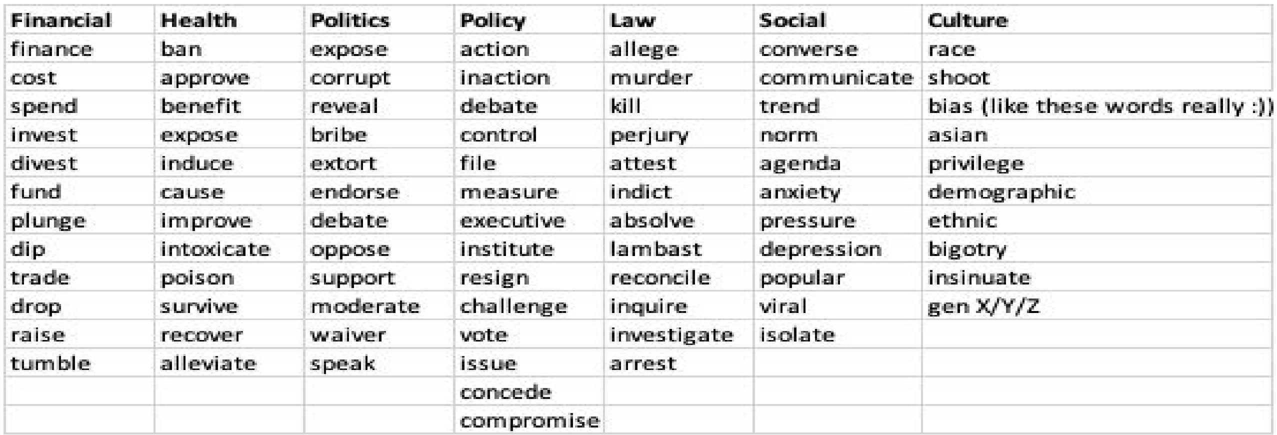
Often, because resources are so scarce, a decision has to be made whether to cover one story over another. If journalists had a tool to better see the impact of their work, they can make better decisions on their time and efforts vs. relying solely on vanity metrics like page views, video starts and unique visitors. Those are good to have but they don't paint a holistic view impact. It's best to look at both quantitative and qualitative data.

Impact Word Categorization

After several practice runs through summarizers, we came up with several impact categories along with keywords to create the framework for algorithms to be used in the tool.

Impact:
Financial: (stocks, earnings, bankruptcy, openings, revenue, profit, loss)
Health: (medicine, research,)
Politics: (new candidates, elections)
Policy: (legislation, new laws)
Law: (Indictments, lawsuit)
Social: (sentiment, conversations)
Culture: (impact in underrepresented communities - black, asian, latino etc)

The purpose of this would be to make it easier to surface related content and threads that reporters and editors can explore and make a decision whether to do a follow up etc..



Financial	Health	Politics	Policy	Law	Social	Culture
finance	ban	expose	action	allege	converse	race
cost	approve	corrupt	inaction	murder	communicate	shoot
spend	benefit	reveal	debate	kill	trend	bias (like these words really :))
invest	expose	bribe	control	perjury	norm	asian
divest	induce	extort	file	attest	agenda	privilege
fund	cause	endorse	measure	indict	anxiety	demographic
plunge	improve	debate	executive	absolve	pressure	ethnic
dip	intoxicate	oppose	institute	lambast	depression	bigotry
trade	poison	support	resign	reconcile	popular	insinuate
drop	survive	moderate	challenge	inquire	viral	gen X/Y/Z
raise	recover	waiver	vote	investigate	isolate	
tumble	alleviate	speak	issue	arrest		
			concede			
			compromise			

Relationships vs Keywords

Rather than relying on keywords to guide our results, we took a different approach in tackling this challenge. Leveraging the power of artificial intelligence and natural language processing to identify impact and then look at relationships around “impact” trigger words in the story to instruct the computer to look for those relationships in the APIs and databases.

Pavan created the following model and compared it to the keyword search provided by the default summarizer.

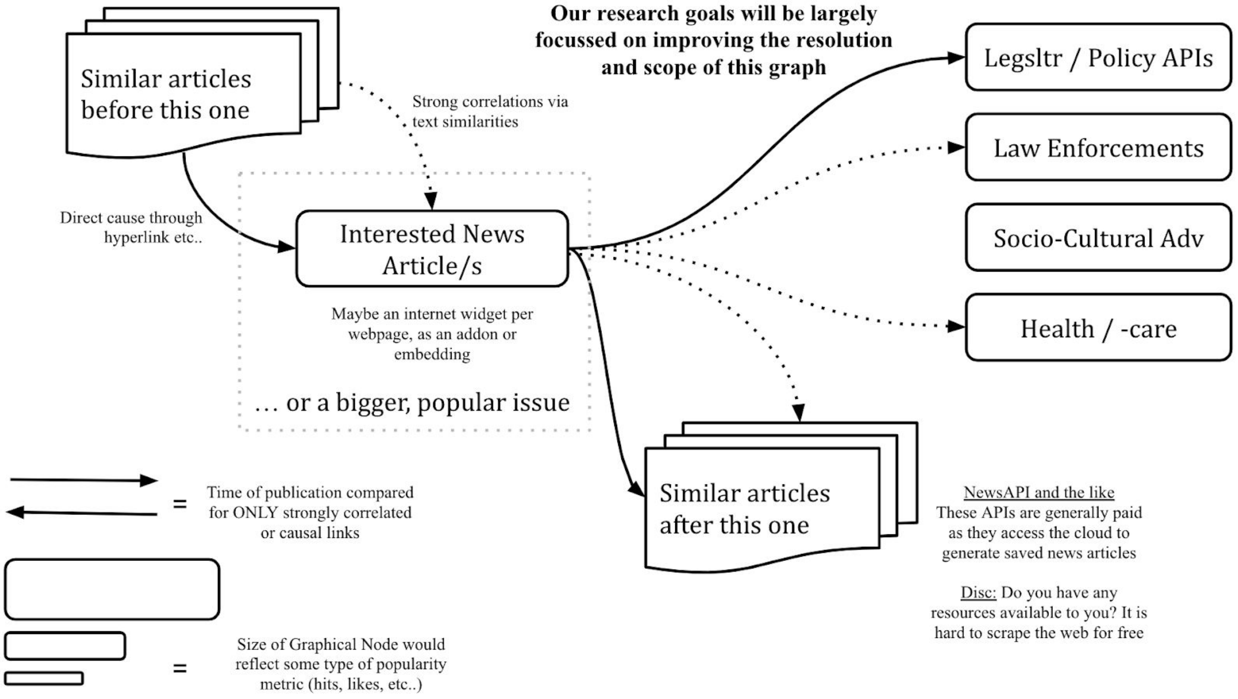
1. “Seed” words were automatically searched on Google, and the first 50 news articles were extracted, scraped (to just body + title for now) and lemmatized.
2. A simple TF-IDF aggregator was trained on this data and the provided link was tested on it to return the top keywords (I can also return a similarity metric with the fit-transform)

Note: The TF-IDF classifier on 50 articles was sensitive to the news articles scanned and was at most as good as a filtered keyword-aggregated search. Instead, for the relationship extraction phase, we used a custom feedback mechanism whereby we improved the quality of relationships extracted by improving our search for the “right” keywords through repeated iteration through the semantic impact classes shown above.

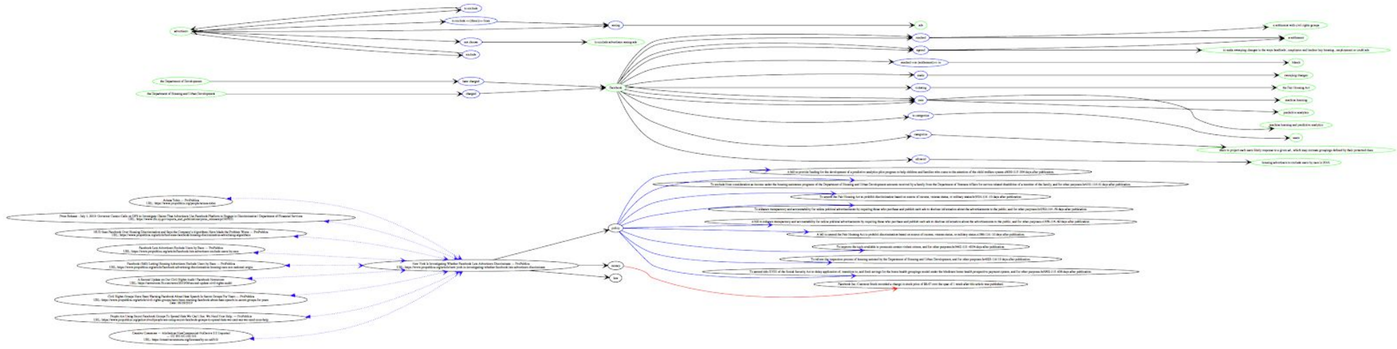
Example:
An example of improved keywords versus an off-the-shelf approach is provided below for the following article:
https://www.huffpost.com/entry/julie-burr-federal-contract-worker-gofundme-shutdown_n_5c346577e4b05d4e96bbbee0

Top 5 keywords:

1. Improved Keywords: ['shutdown', 'gofundme', 'paycheck', 'government', 'federal’]
2. Default Keywords: [‘help', 'contracted', 'shutdown', 'page', 'contract’]


Our research goals will be largely focussed on improving the resolution and scope of this graph
Similar articles before this one
Strong correlations via text similarities
Direct cause through hyperlink etc..
Interested News Article/s
Maybe an internet widget per webpage, as an addon or embedding
… or a bigger, popular issue
Legsltr / Policy APIs
Law Enforcements
Socio-Cultural Adv
Health / -care
Similar articles after this one
=
Time of publication compared for ONLY strongly correlated or causal links
=
Size of Graphical Node would reflect some type of popularity metric (hits, likes, etc..)
NewsAPI and the like
These APIs are generally paid as they access the cloud to generate saved news articles
Disc: Do you have any resources available to you? It is hard to scrape the web for free

From this initial model, Pavan worked on refining the model even further to better extract better results: (zoom in to view graphic)

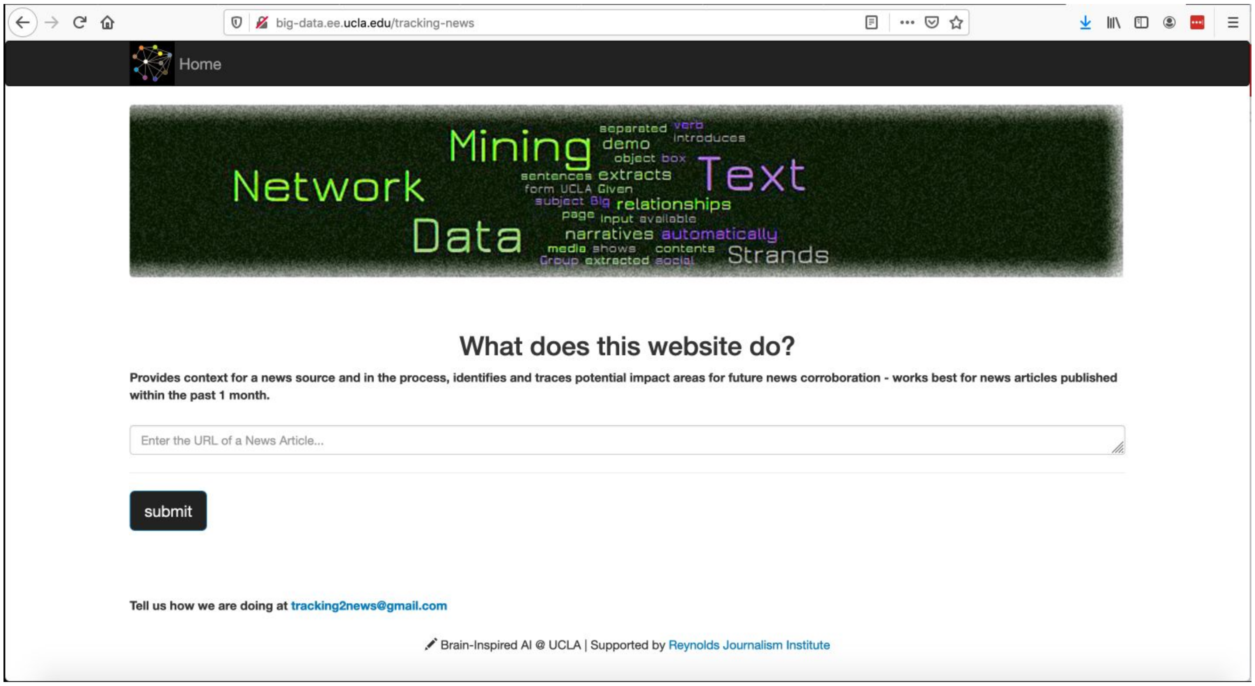


Beta Tests

With these models at hand, I continued to provide more test urls to see if the model would produce the expected results in order to continue to teach the algorithm Pavan worked on. After several weeks of running test urls and blind tests, we released the following version for user feedback. This version was sent to more than 30 people from both large and small local news organizations as people in adjacent industries (marketing and PR).

Version #1: (Released end of October 2019)
Specs: Flask / uWSGI framework - serial | APIs include: BeautifulSoup | Written in Python, Java backend with manual database manager | HTML, CSS - Bootstrap | Port access required for relationship extraction

URL: http://big-data.ee.ucla.edu/tracking-news



What does this website do?

Provides context for a news source and in the process, identifies and traces potential impact areas for future news corroboration - works best for news articles published within the past 1 month.

Enter the URL of a News Article...

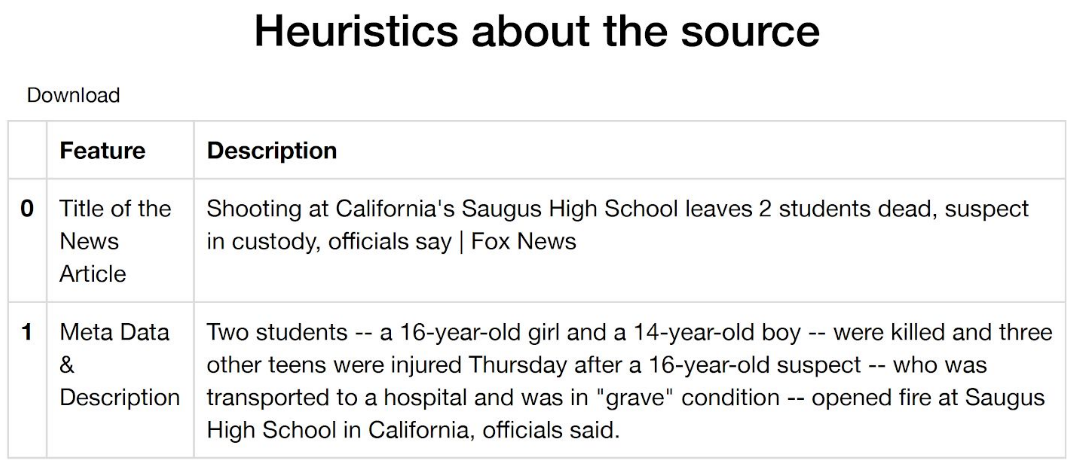
Tell us how we are doing at tracking2news@gmail.com

Brain-Inspired AI @ UCLA | Supported by Reynolds Journalism Institute

This version identified impact areas, impact sentences and surfaced related content if available. It also included a visual representation of the impact in the story broken down by actors and actants.

Sections of the tool

Heuristics About The Source



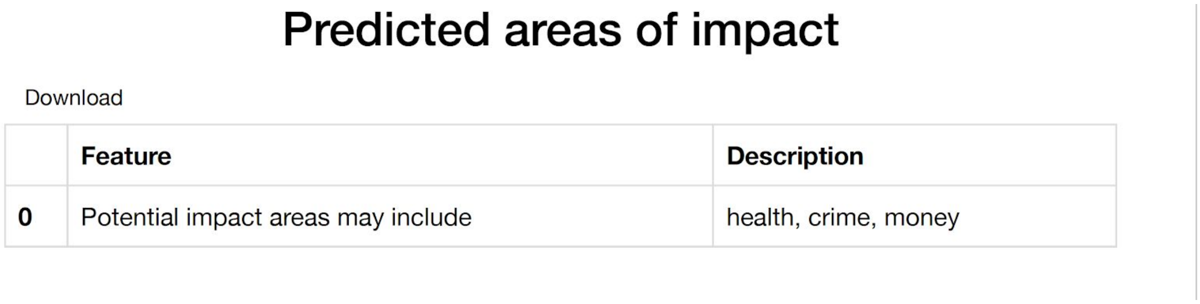
Heuristics about the source

Download

	Feature	Description
0	Title of the News Article	Shooting at California's Saugus High School leaves 2 students dead, suspect in custody, officials say | Fox News
1	Meta Data & Description	Two students -- a 16-year-old girl and a 14-year-old boy -- were killed and three other teens were injured Thursday after a 16-year-old suspect -- who was transported to a hospital and was in "grave" condition -- opened fire at Saugus High School in California, officials said.

After a user pastes a URL into the tool, the first section they encounter is “Heuristics about the source” which pulls the metadata of the URL. This is useful in seeing the description, and the title of the story without having to do a view source action.

Predicted Areas of Impact



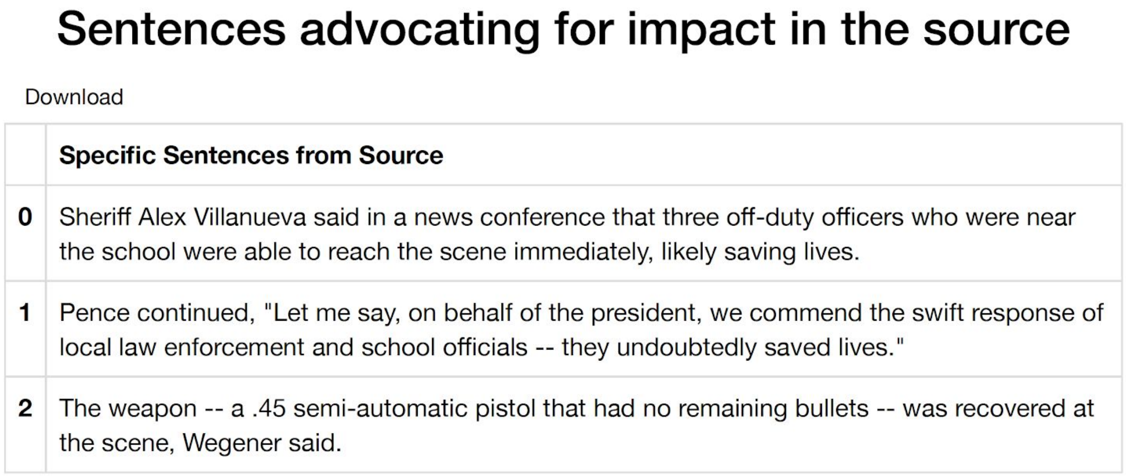
Predicted areas of impact

Download

	Feature	Description
0	Potential impact areas may include	health, crime, money

This is pulled from the vocabulary bank of impact terms (with additional classes) I had given to Pavan and we categorized the keyword clusters into 14 areas: 'sports', 'weather', 'law', 'health', 'medicine', 'policy', 'crime', 'technology', 'money', 'disaster', 'science', 'race', 'entertainment', 'politics'. Some of these impact areas were further clustered while accessing APIs. The detail of extracting impact within these areas is an area to be explored in future versions.

Sentences Advocating For Impact In The Source



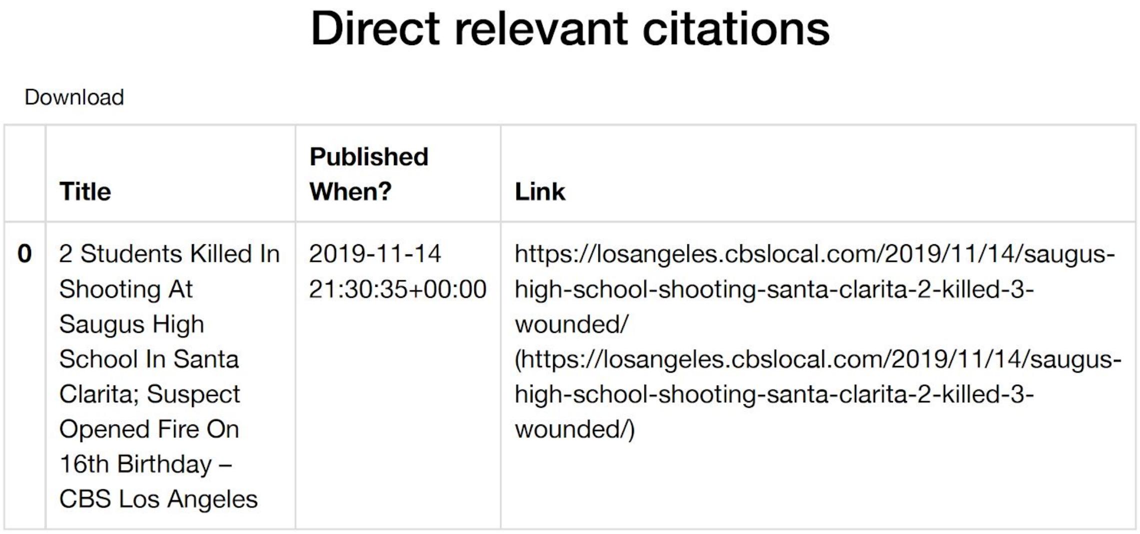
Sentences advocating for impact in the source

Download

	Specific Sentences from Source
0	Sheriff Alex Villanueva said in a news conference that three off-duty officers who were near the school were able to reach the scene immediately, likely saving lives.
1	Pence continued, "Let me say, on behalf of the president, we commend the swift response of local law enforcement and school officials -- they undoubtedly saved lives."
2	The weapon -- a .45 semi-automatic pistol that had no remaining bullets -- was recovered at the scene, Wegener said.

This section is intended to pull out sentences around the “impact” words. If successful, this section would identify the nutgraf and main sentences of the article. This saves the user time reading through the article if pressed for time.

Direct Relevant Citations



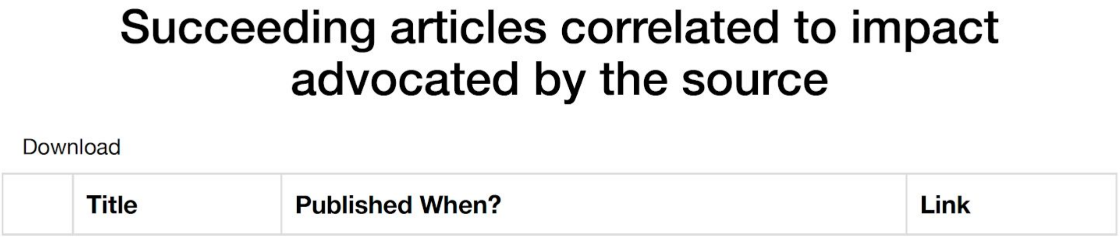
Direct relevant citations

Download

	Title	Published When?	Link
0	2 Students Killed In Shooting At Saugus High School In Santa Clarita; Suspect Opened Fire On 16th Birthday – CBS Los Angeles	2019-11-14 21:30:35+00:00	https://losangeles.cbslocal.com/2019/11/14/saugus-high-school-shooting-santa-clarita-2-killed-3-wounded/ (https://losangeles.cbslocal.com/2019/11/14/saugus-high-school-shooting-santa-clarita-2-killed-3-wounded/)

This section pulls out any links embedded in the article. This was intended to make it easier for the user to easily identify any links in the story. The next phase of this section includes citations from other publications besides what is contained in the article.

Succeeding Articles Correlated To Impact Advocated By The Source



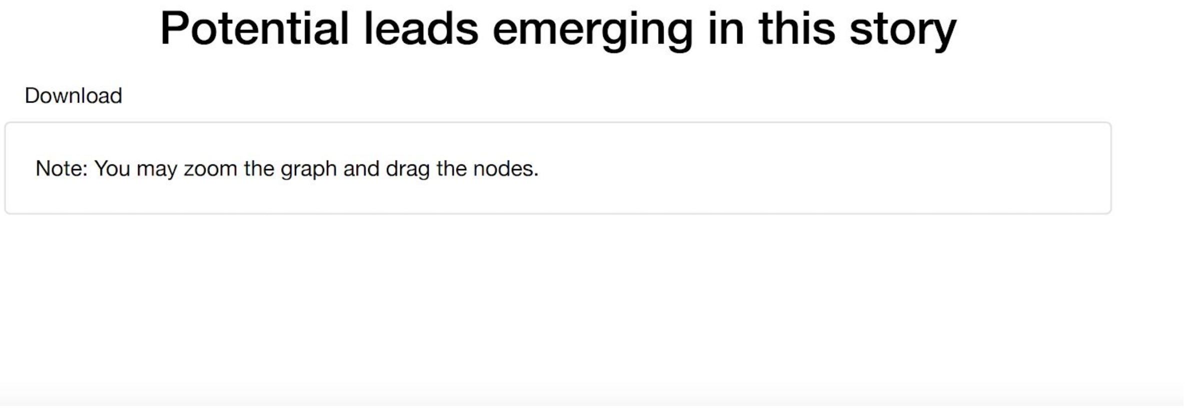
Succeeding articles correlated to impact advocated by the source

Download

	Title	Published When?	Link

This section would show any relevant articles published outside this source and other outside news media articles from connected APIs. For example, if the story involved a publicly traded company, this section would show any stock movement after the story publication date. If the story mentioned any law of policy, if there was a result in the government database, it would show the latest

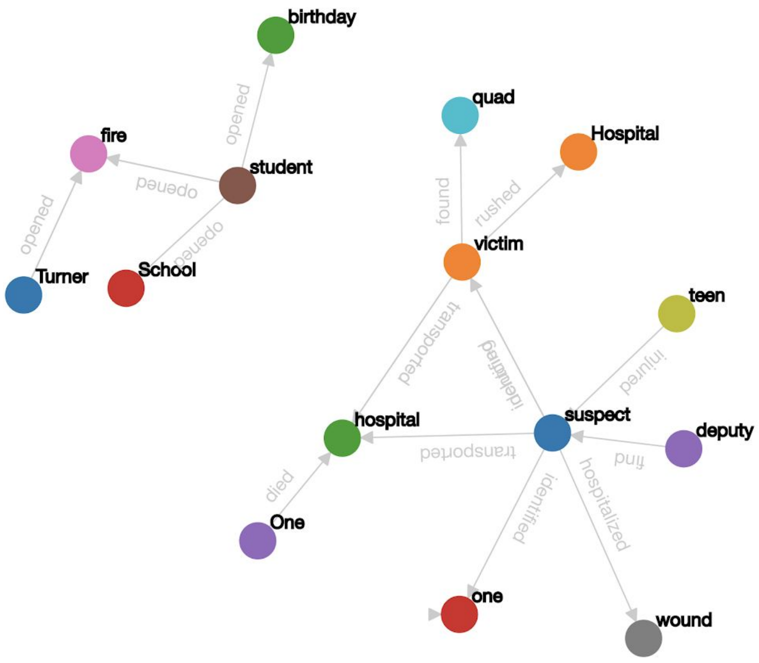
Potential Leads From Story



Potential leads emerging in this story

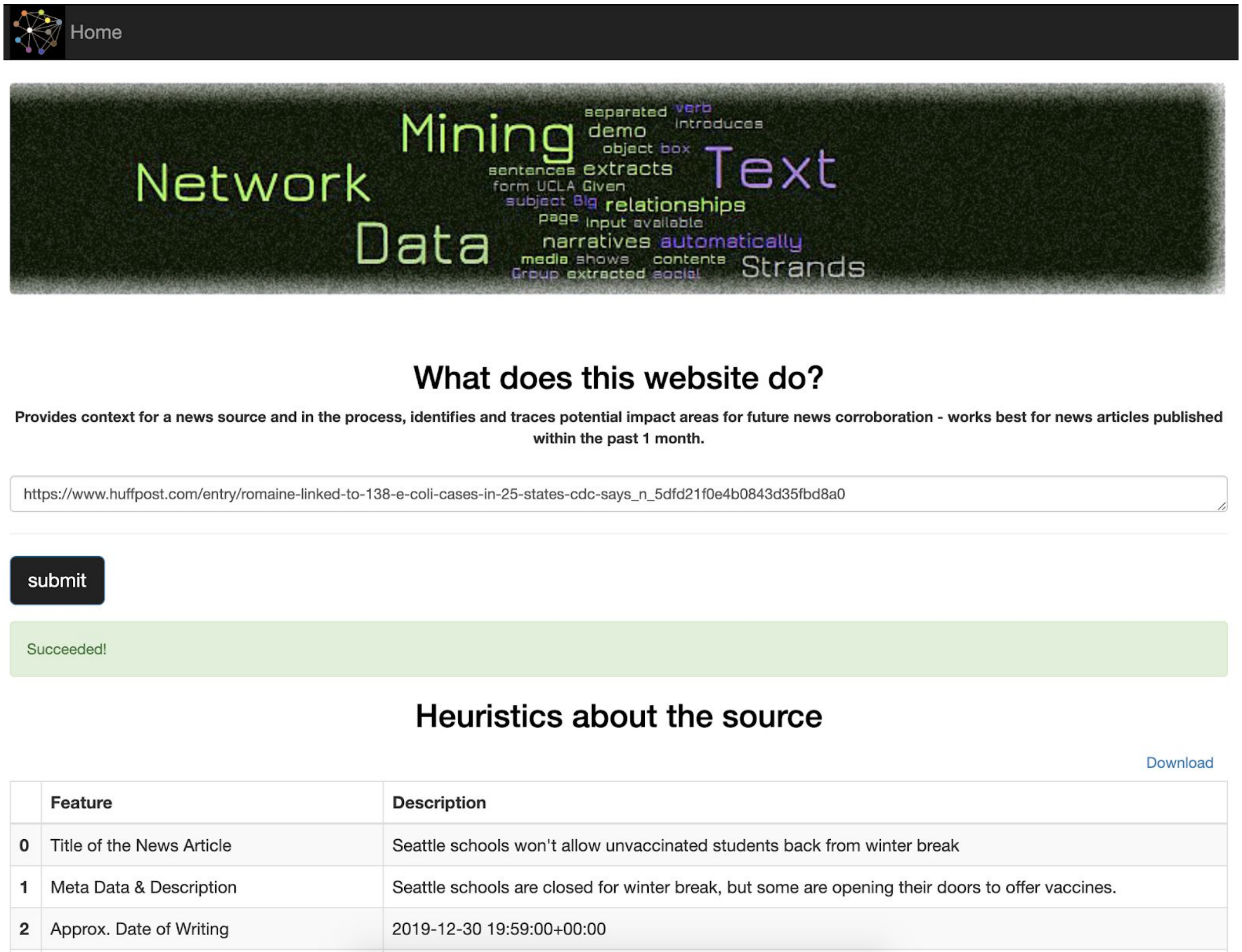
Download

Note: You may zoom the graph and drag the nodes.



The intention for this section is to present a visual representation of the impact in the sentence. This is useful for more complex long-form stories with many threads. The more lines that come out of the center, the more connections.

Version #2 (Released end of December 2019):
Flask / uWSGI framework - parallel | APIs include: BeautifulSoup, Article Date extractor, NewsAPI, policyAPI | Written in Python, Java, Shell backend with manual database manager | HTML, CSS - Bootstrap | Port access required for relationship extraction


Home
Mining
Network
Data
Text
Strands
separated verb introduces demo object box sentences extracts form UCLA Given subject Big relationships page input available narratives automatically media shows contents Group extracted social
submit
Download

What does this website do?

Provides context for a news source and in the process, identifies and traces potential impact areas for future news corroboration - works best for news articles published within the past 1 month.

https://www.huffpost.com/entry/romaine-linked-to-138-e-coli-cases-in-25-states-cdc-says_n_5dfd21f0e4b0843d35fbd8a0

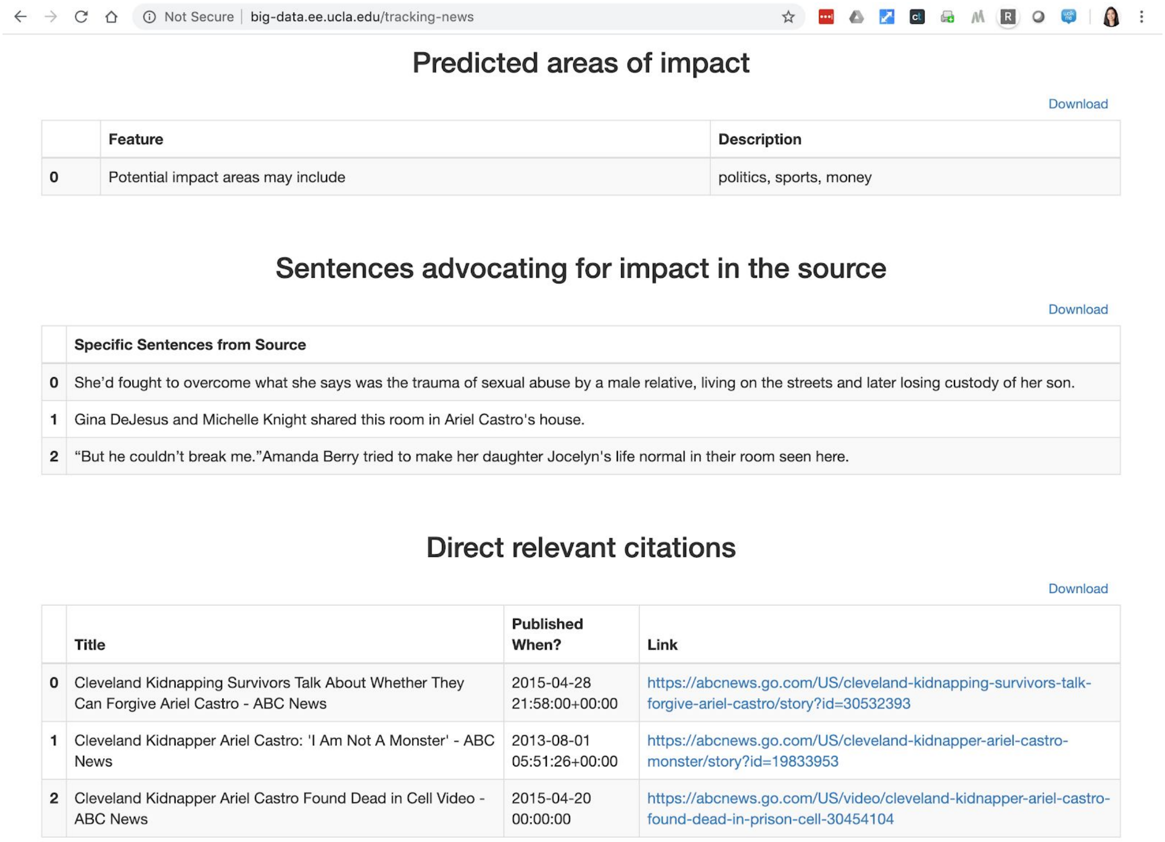
Succeeded!

Heuristics about the source

	Feature	Description
0	Title of the News Article	Seattle schools won't allow unvaccinated students back from winter break
1	Meta Data & Description	Seattle schools are closed for winter break, but some are opening their doors to offer vaccines.
2	Approx. Date of Writing	2019-12-30 19:59:00+00:00

Sections Of The Tool

In version #2, the majority of the sections remain the same on the front end, with improvements more on the backend to surface up more relevant and refined results.



Predicted areas of impact

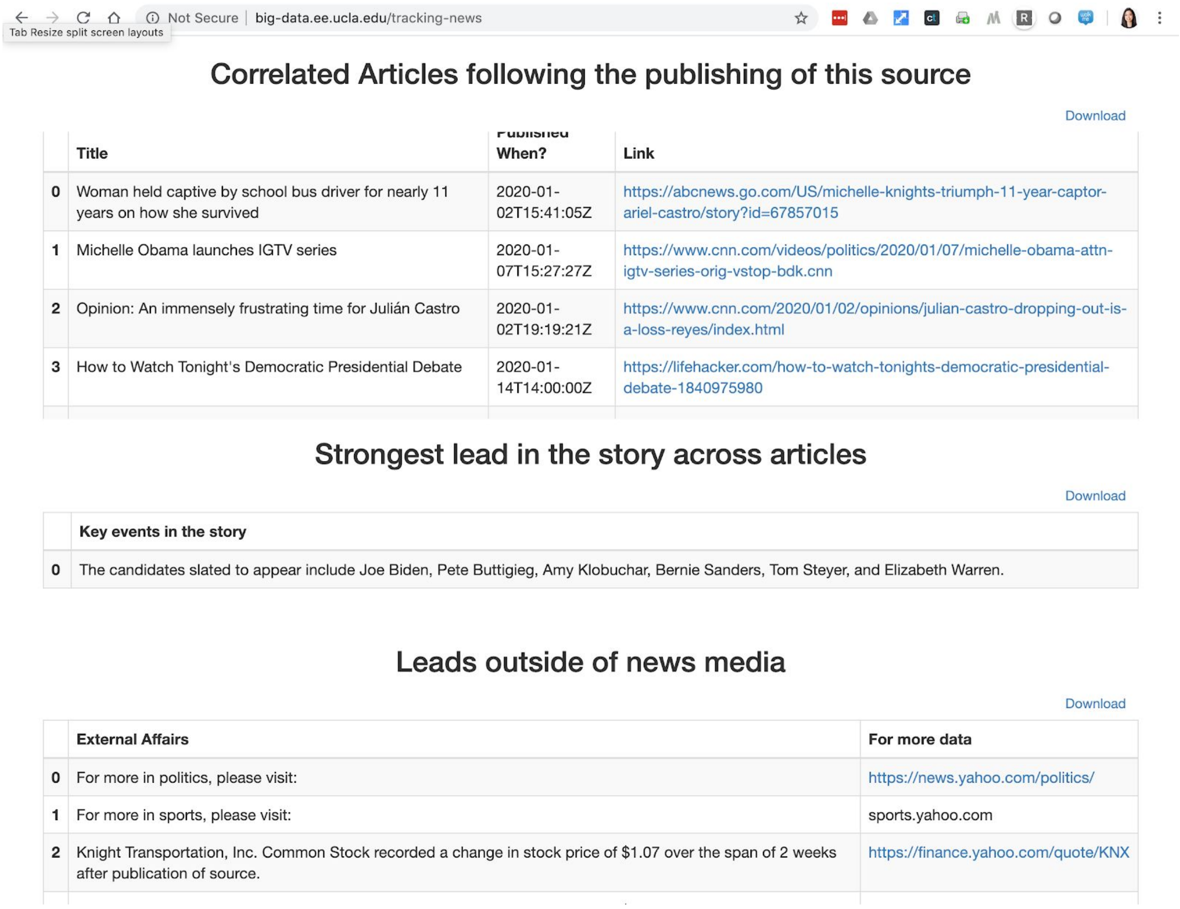
	Feature	Description
0	Potential impact areas may include	politics, sports, money

Sentences advocating for impact in the source

	Specific Sentences from Source
0	She'd fought to overcome what she says was the trauma of sexual abuse by a male relative, living on the streets and later losing custody of her son.
1	Gina DeJesus and Michelle Knight shared this room in Ariel Castro's house.
2	“But he couldn't break me.”Amanda Berry tried to make her daughter Jocelyn's life normal in their room seen here.

Direct relevant citations

	Title	Published When?	Link
0	Cleveland Kidnapping Survivors Talk About Whether They Can Forgive Ariel Castro - ABC News	2015-04-28 21:58:00+00:00	https://abcnews.go.com/US/cleveland-kidnapping-survivors-talk-forgive-ariel-castro/story?id=30532393
1	Cleveland Kidnapper Ariel Castro: 'I Am Not A Monster' - ABC News	2013-08-01 05:51:26+00:00	https://abcnews.go.com/US/cleveland-kidnapper-ariel-castro-monster/story?id=19833953
2	Cleveland Kidnapper Ariel Castro Found Dead in Cell Video - ABC News	2015-04-20 00:00:00	https://abcnews.go.com/US/video/cleveland-kidnapper-ariel-castro-found-dead-in-prison-cell-30454104


Not Secure | big-data.ee.ucla.edu/tracking-news
Tab Resize split screen layouts

Correlated Articles following the publishing of this source

Download

	Title	Published When?	Link
0	Woman held captive by school bus driver for nearly 11 years on how she survived	2020-01-02T15:41:05Z	https://abcnews.go.com/US/michelle-knights-triumph-11-year-captor-ariel-castro/story?id=67857015
1	Michelle Obama launches IGTV series	2020-01-07T15:27:27Z	https://www.cnn.com/videos/politics/2020/01/07/michelle-obama-attn-igtv-series-orig-vstop-bdk.cnn
2	Opinion: An immensely frustrating time for Julián Castro	2020-01-02T19:19:21Z	https://www.cnn.com/2020/01/02/opinions/julian-castro-dropping-out-is-a-loss-reyes/index.html
3	How to Watch Tonight's Democratic Presidential Debate	2020-01-14T14:00:00Z	https://lifehacker.com/how-to-watch-tonights-democratic-presidential-debate-1840975980

Strongest lead in the story across articles

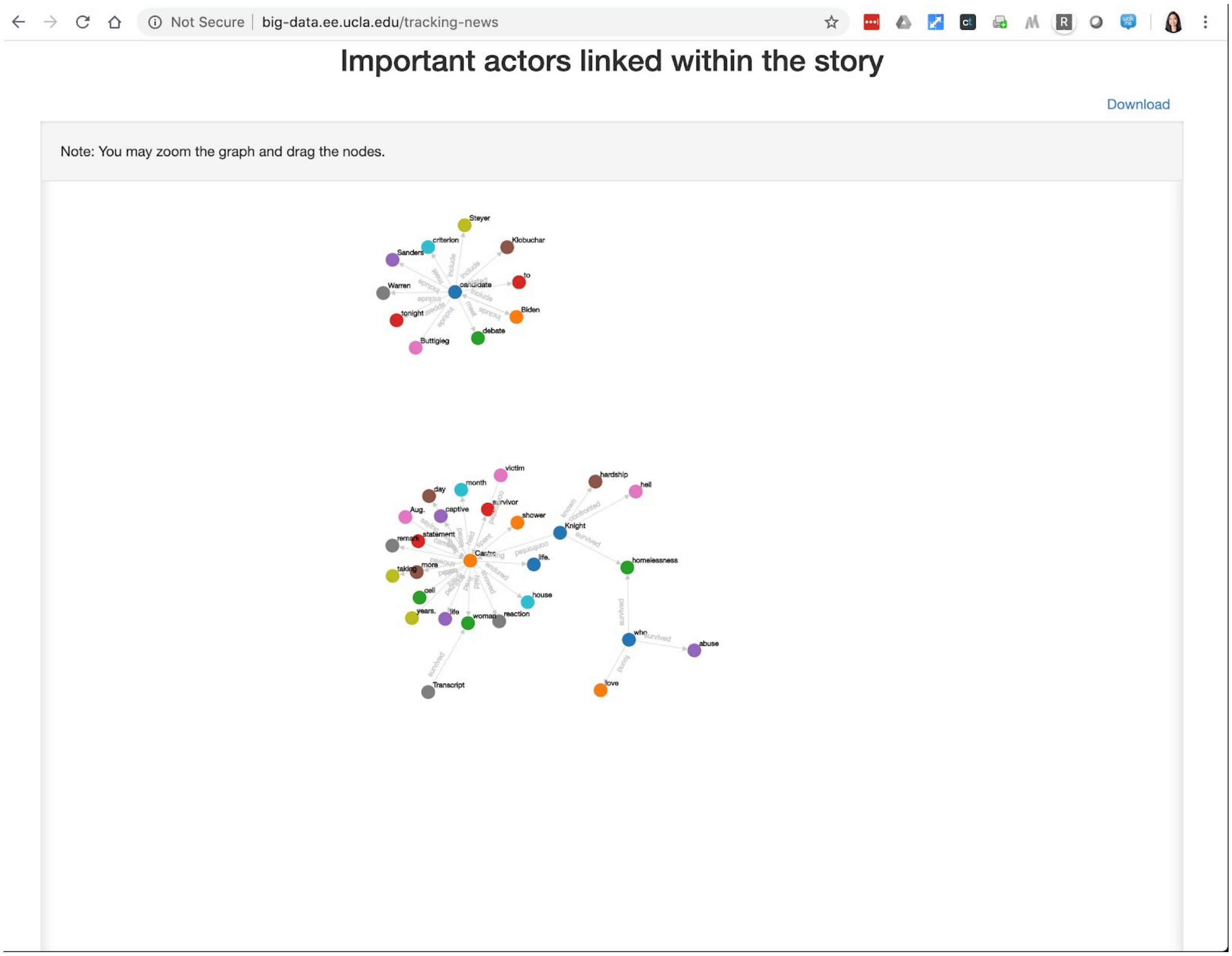
Download

	Key events in the story
0	The candidates slated to appear include Joe Biden, Pete Buttigieg, Amy Klobuchar, Bernie Sanders, Tom Steyer, and Elizabeth Warren.

Leads outside of news media

Download

	External Affairs	For more data
0	For more in politics, please visit:	https://news.yahoo.com/politics/
1	For more in sports, please visit:	sports.yahoo.com
2	Knight Transportation, Inc. Common Stock recorded a change in stock price of $1.07 over the span of 2 weeks after publication of source.	https://finance.yahoo.com/quote/KNX


Not Secure | big-data.ee.ucla.edu/tracking-news

Important actors linked within the story

Download

Note: You may zoom the graph and drag the nodes.

Feedback

Version #1 (Released October 2019)

We learned several things to improve as obtained from user feedback.

We put a call out for beta testers in September 2019. We received 30 responses. Out of 30, 8 testers returned feedback.

The questions testers were asked:

- Most/Least Useful Features (Sections)
- What are your biggest KPIs or OKRs?
- Do you use any other impact tools? If so, which ones?

From the editorial end, several users ran into issues of timeouts and lag time.
Overall, users had about a 50/50 result on the urls they tested in surfacing relevant results.

Some feedback included:

- Interesting, took about an hour to do. Might want to have 5 articles from a larger group of people.
- For the "most useful" and "least useful," I picked based on the assumption that the functions will be refined over time rather than as they're currently working.
- The site timed out (504) this afternoon so I'm submitting the links and the feedback from the articles I was able to research. I hope this is helpful!
- I received several error messages and instructed to try back again or later. Thus, I was only able to get through four articles in this round.

There are many factors into why:

- Not all available APIs have been connected to the tool
- Some urls on special templates made it difficult to scrap
- Thresholds were too broad or too narrow, therefore non relevant results were being surfaced

From the backend:

  - Cleaning the UI and improving representation methods
  - Sanitized output to user and better informing the choices displayed on the website (including removing false positives, repeats, URLs, date of writing, adding complete sentences)
  - Reorganized the elements on the URL

- Held calls to filter through stack to result in quicker user update
  - Added timeouts to blocking methods that stall the program until completion

- Connected the various data sources together to create more reliable output
  - Direct URLs and related URLs were scraped as well with the pipeline running on these to collaborate findings and verify output.

- Extracted actor-actant relationships across articles to improve the fidelity of impact recognition
  - Limited to first and nut paragraph to capture key events in different articles to limit time taken and yet get a reliable understanding of the story underneath.

- Added more APIs to get better access on impact around news
- Resolved particular website crashes
  - Included improving error handling try - except clauses, and particular domain failures including Huff post (for example, date is a string not a datetime object)

Version #2 (Released end of December 2019)

We learned several things to improve as obtained from user feedback.

We put a second call out for beta testers in January 2020. We received 4 responses and all 4 testers returned feedback.

The questions testers were asked:

- Most/Least Useful Features (Sections)
- What are your biggest KPIs or OKRs?
- Do you use any other impact tools? If so, which ones?

From the editorial end, users are still reporting lag times and timeouts. However, with improvements to the modeling, we've been seeing more relevant results.

Some feedback included:

- For future versions I'd suggest your developers try reconfiguring this to work asynchronously. Specifically, someone would submit a URL and, instead of waiting for it process in the browser window, you could leave and come back later. Setting a cookie or using local storage to save the job ID, as long it's not cleared by the user, would mean the user could even close the tab and see the results when they return. Also, I'm not sure if they're using some kind of task queue, but that could also help and allow them to set retries at a set interval in case anything failed initially.

- I called the context-relevant features least useful because they did not provide relevant information. If they HAD provided relevant information, I suspect they WOULD be useful features.

- The features that might be most useful seem to need the most work, which is understandable since they're more complicated to track (articles correlated to impact, leads outside news media & actors linked within story). Looking forward to seeing how the end product works. For my searches, the results mostly weren't actually tied to the news story. But a big part of this may be because the URLs I entered were all for AP's home site, apnews.com, which doesn't get a lot of traffic since the AP is a B2B news provider. The "actors linked in story" category also seemed to have changed from the first version to be more narrow? If so, think I preferred when the results were more expansive.

Overall the biggest changes to Version #2 from user perspective:

- sanitized output, correlated and referenced output
- relevant relationship extracted
- impact sentences across articles
- user layout is changed
- Domain specific and famous domains work well

From the backend:

Version #2 (Released end of December 2019):
Flask / uWSGI framework - parallel | APIs include: BeautifulSoup, Article Date extractor, NewsAPI, policyAPI | Written in Python, Java, Shell backend with manual database manager | HTML, CSS - Bootstrap | Port access required for relationship extraction

Things to improve:

- Higher resistance to DDoS (Distributed Denial of Service) attacks
    - Server is mounted on a low-cost platform causing high-usage to serialize and delay response when traffic is high
    - Solutions:
        - Adding bandwidth to server (typically involves more expenditure)
        - Adding higher-cost API access to improve the rate of data retrieval

- UI to be made more interactive
    - Dynamically allow users to interact with website and orient the results according to their on-line preference.
    - Considerations of various more sophisticated presentation tools

- Adding stronger APIs for out-of-news access
    - Current APIs are good but can be improved to access further non-news but related impact news
- Feedback on results
    - Tracking results that are not relevant to improve algorithm - particularly in the leads outside news media

Next Steps

There’s much more work to be done on Impact With Context tool. We’ve been able to take an idea to better automate and use AI and NLP to identify impact in stories and get it to a place where we are seeing some insightful early results. However, as NLP modeling gets more sophisticated and advances, in order to keep improving this tool, more funding and time is needed to fully flesh out what we started.

Ideally, I’d like to build out the front end of it to have a more user-friendly UX/UI which also returns results faster. As the data gets updated, so does the results - ultimately making it easier to show the progression of impact. This would be a game-changer for journalism on many fronts. In addition, I’d like to incorporate more ways to provide feedback on both the backend and frontend directly on the tool itself.

For the for-profit newsrooms, it can make it clear to investors why doing that investigative series was worth spending money despite not running ads. For the nonprofit newsrooms, it can clearly show the impact with context that’s in alignment with their foundation’s grant funding guidelines. For the news consumer, it can localize an international story and prompt them to find out more.

My plan is to use this MVP/prototype and seek more funding so I can move this tool into further development and release it to a larger audience.

Conclusion

Constant iteration and improvement are needed to make sure this impact metric tool achieves the goal of providing the context to better understand how a story affects readers' lives versus just tracking clicks and recirculation. As technology evolves, so does the data. This tool will need to keep evolving and keep redefining what is "impact" - and as a result, help journalists see the unseen and for news consumers come away better informed about the world around them.

References

Burgos, Pedro. “How Do We Measure the ‘Real-World’ Impact of Journalism?” *Medium*, Fellow Journalists, 10 July 2015, medium.com/fellow-journalists/how-do-we-measure-the-real-world-impact-of-journalism-dab42f16b6bf.

Colmery, Ben. “How Can We Measure the Impact of Journalism?” *International Journalists' Network*, 15 Mar. 2013, ijnet.org/en/story/how-can-we-measure-impact-journalism.

Lewis, Charles, and Hilary Niles. “Measuring Impact: The Art, Science and Mystery of Nonprofit News Assessment by Charles Lewis.” *Goodreads*, Goodreads, 5 Nov. 2013, www.goodreads.com/book/show/32203549-measuring-impact, https://irw.s3.amazonaws.com/uploads%2Fmeasuring-impact-final-pdf.pdf

Green-Barber, Lindsay. “How Can Journalists Measure the Impact of Their Work? Notes toward a Model of Measurement.” *Nieman Lab*, 19 Mar. 2014, www.niemanlab.org/2014/03/how-can-journalists-measure-the-impact-of-their-work-notes-toward-a-model-of-measurement/.

Schiffrin, Anya, and Ethan Zuckerman. “Can We Measure Media Impact? Surveying the Field (SSIR).” *Stanford Social Innovation Review: Informing and Inspiring Leaders of Social Change*, ssir.org/articles/entry/can_we_measure_media_impact_surveying_the_field.

Acknowledgments

Innovation is a work in progress, it's neither static nor new. It's about understanding needs and priorities and connecting the dots between elements that push the boundaries even more than they have before. This project would not be possible without several groups, individuals and support networks that helped bring an idea into reality.

Foremost, I'd like to thank the Reynolds Journalism Instititute's Randy Picht, Mike McKean, Kat Duncan, Jennifer Nelson and Rueben Stern for their guidance, support throughout my fellowship. Thank you for believing in my project and giving me the means to build, test and iterate it. I would not have been able to take it to the point it is now without your support. In addition, I'd like to thank my 2019-2020 RJI Fellows cohort: Virginia Arrigucci, Michael Epstein, Krystal Knapp, Jim MacMillan, Neil Mara and Carolyn Robinson, who were supportive throughout the process.

Secondly, I'd like to thank UCLA's Samueli School of Engineering's Professor Vwani Roychowdury and graduate student Pavan Holur who were incredible technical partners in this project. This partnership is an example of how such a partnership should work: both parties learn from each while working toward a common goal. The work we were able to accomplish in just eight months sets the foundation for further work in this space and will hopefully lead to better leveraging AI and NLP in the news space. A big thank you to Matthew Chin, who connected me to Professor Roychowdury and graduate student Shadi Shahsavari, who provided guidance early in the process.

Thirdly, I'd like to thank my employers, both past and present -- The Associated Press and HuffPost -- who supported me through this process as I balanced both my day role and this fellowship during off hours. Thank you to the AP's Jon Fahey, Alicia Chang, Sarah Nordgren, and Lisa Gibbs for your support. Thank you to HuffPost's Jennifer Kho for the time and space to continue my work on this tool.

Lastly, I'd like to thank all the beta tool testers who signed up and made time to test and provide feedback on my tool. Your feedback has been invaluable in the development and further development of this project and only helps our industry at large.