Primer

Extracting value from unstructured enterprise data in the public sector

The problem

As a Commander in the Royal Army Julie Henrickson is responsible for emergency response related to the Royal Family. Under her charge is a cell of analysts who are responsible for synthesizing the increasing deluge of information around public events that may hold clues to prevent threats.

Julie's team combs through social media, from Twitter to Reddit to 4chan to Telegram, as well as internal reports looking for signs that potential threat actors may be coordinating an attack. The more data they can get their eyes on, the better their chances are to prevent a crisis.

Unfortunately the amount of data her team has to consume is increasing exponentially every year, and Julie's colleagues in MI-5 were recently taken to task for missing clues that the recent Manchester bomber was connected to ISIS that could have prevented that attack.

Headlines like these are painful for our public servants, but they don't adequately describe the huge challenge our nation's security organizations face when it comes to making sense of ever-growing data repositories, and taking timely and appropriate action. Simply put, having the data isn't enough if there is too much data for humans to synthesize it quickly enough to act on it.

Government agencies, like Julie's, have vast and growing volumes of data from which they have to make informed decisions. Most organizations have a strategy to capture and store data in a data lake or warehouse, and then to analyze this data, humans manually consume documents using spreadsheets, or other database software to make sense of the data. Teams like Julie's cannot keep up with the amount of data being produced with a manual approach.

And the problem is only expected to get worse with some predictions estimating the world will be producing 163 zettabytes of data annually by 2025, according to IDC data.

Getting more from your data

Forward-looking commanders like Julie are beginning to turn to advanced artificial intelligence (AI) solutions to extract value from their existing data in order to:

- Surface insights from previously unutilized information
- Accelerate time to insight and action
- Expedite workflows and gain organizational efficiency
- Elevate analysts from spreadsheet wizards to informed decision-makers

Julie is not alone. In 2023, the U.S. government dedicated $3 billion to responsibly develop, test, procure, and integrate transformative AI applications across agencies.[1]

This paper covers a type of AI called Natural Language Processing (NLP) that unlocks a new kind of data analysis at scale, enabling organizations to extract insights from millions of documents, both internal and external, in real time, with human-level accuracy.

Primer

Artificial intelligence isn't new, but applying it to what's commonly referred to as unstructured data like social media, intelligence reports, phone call transcripts, and similar is new. NLP augments human workflows, identifying known bad actors from thousands (or millions) of documents in seconds so that analysts can see a more complete picture of their information landscape, detect anomalies, and provide more accurate recommendations to leadership, faster.

This is a critical problem. IDC estimates up to 80% of organizational data is unstructured and therefore being underutilized during decision making, if it's being utilized at all.[2] Governmental organizations responsible for our nation's security must take action to leverage these types of data.

80% of organizational data is unstructured[2]

Structured vs unstructured data

Structured data is data that has a clear data schema: like numbers in a table that we can analyze in a spreadsheet. Most governmental organizations have very mature storage and analysis capabilities for structured data. Unstructured data, however, like audio transcripts or policy documents, doesn't fit into a spreadsheet or a legacy database.

Structured data	Unstructured data
Flight manifests	Records of interviews
Financial events	Research papers
Financial transactions	Audio transcriptions (telephone calls, meetings)
GPS locations tracking data	Policy documents
Citizen's personal data	Tweet analysis
	Sentiment
	Social media and dark web sources

Most organizations are storing unstructured data, but there isn't a good way with traditional methods to search and analyze that data, which also means the opportunity cost of not surfacing value from either data type is largely unquantifiable using current tools.

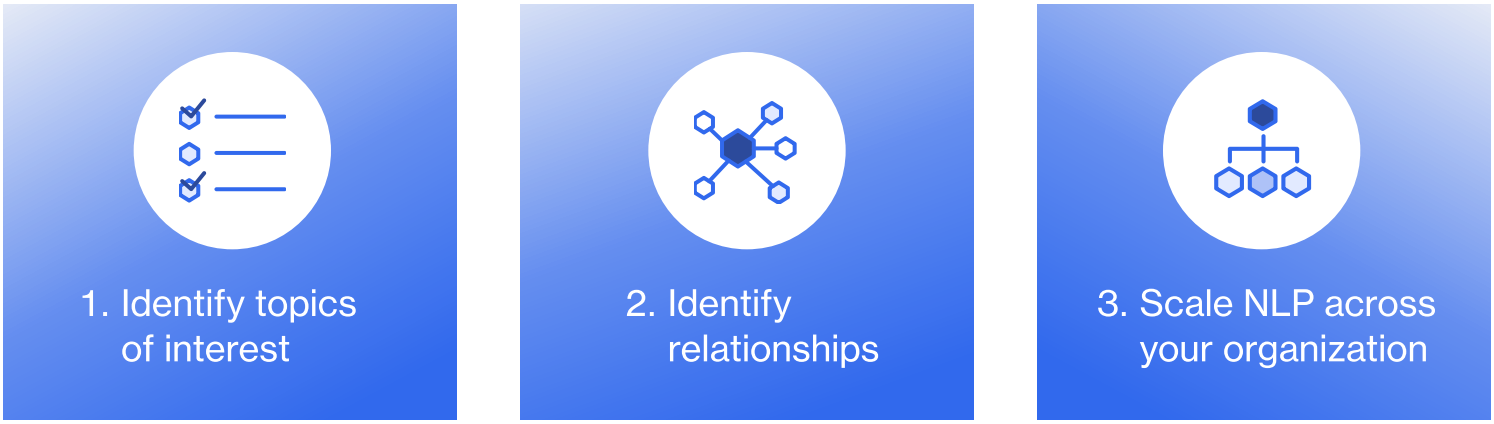
Analyzing unstructured data with NLP

NLP is the machine learning discipline that trains a machine to read millions of unstructured documents to extract critical information - like names, organizations, and their relationships - in seconds for analysts to review and triage. More simply, NLP enables machines to understand and structure human language to make it easier for people to see the big picture quickly, and then choose where to zoom into the details.

For example, Primer recently used NLP to organize and summarize information conveyed in publicly available Russian military audio feeds. The NLP models filtered out background noise, translated and converted the content to English text. It then summarized the communications and pulled out key entities of interest, like military equipment, people, and locations. This gave analysts strategic insights at machine, rather than human speed.

In addition to audio feeds, NLP works on any type of secure, internal document, as well as open source information. Leveraging NLP, analysts have power to identify patterns hiding amongst millions and millions of unstructured data. *So how should Governmental Organizations approach adopting NLP?*

Three stages of NLP adoption



Stage one - identifying topics of interest

The first step in any AI process is to look for the data types that need to be analyzed. A great use case is captured material gathered on the front lines, like interview transcripts or activity reports. Social media and dark web sources that analysts immerse themselves in every day is also unstructured.

Finding information of interest

Currently, analysts manually comb through this data to identify and highlight all the entities that might be relevant. With NLP, analysts can train a machine to process a database of documents, as well as incoming social media, and then extract names and locations, surfacing insights that relate to persons of interest around a particular event.

This means that at a glance a team of analysts can see all the documents that mention John Smith, and all the documents that mention known associates of John Smith and then only focus their attention on those documents. And once an NLP model is automatically identifying all those documents, an analyst team lead like Julie, can set up alerts that get her team's attention when new documents relating to her person of interest show up, instantly.

Adding in a layer of expertise

To take it a step further, Julie's team can take a list of thousands of names and see if those names show up inside a corpus of documents, like a large database or all the social media and news in a 24-hour period in an instant, and they can do this with human-level precision using NLP.

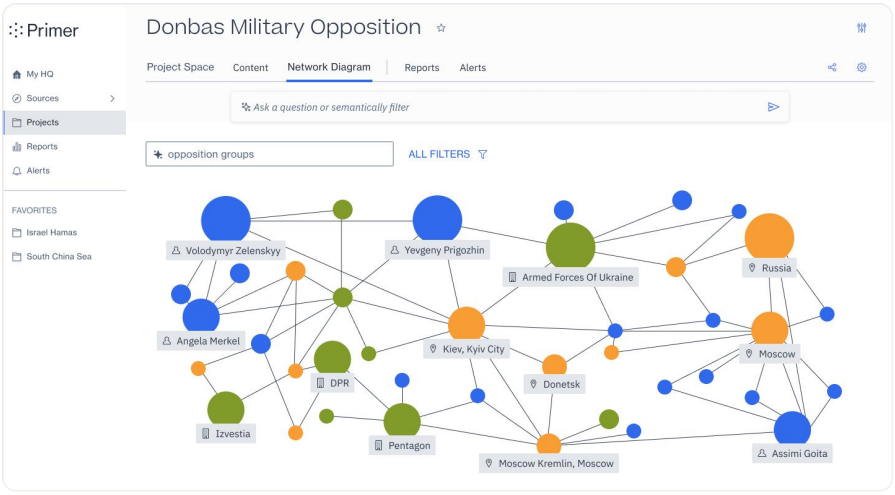
This can save Julie's team a lot of time reading so they can spend more time doing real analysis. Instead of having to read through everything manually, they can focus their attention on the documents containing persons of interest.

But identifying the person of interest is barely scratching the surface of how NLP can aid in analyst workflows. Once analysts identify their subject in a stack of documents they need to understand what they are doing, who they are associated with, and how their behavior has changed, for example.

Identifying these evolving relationships takes us to stage two of NLP adoption.

Stage two - identifying relationships

Manually discovering relationships and connections between subjects of interest temporally takes time, resources, and expertise. But with targets changing and stretching across the globe these relationships become increasingly difficult for analysts to track. Analysts can lean on NLP to streamline these time-consuming, data heavy tasks. For example, NLP algorithms can determine if John Smith's association with a particular terrorist organization has changed over the past 3 months. Or if his name has shown up more frequently with other persons of interest, or with other known organizations to take the example a few steps further.

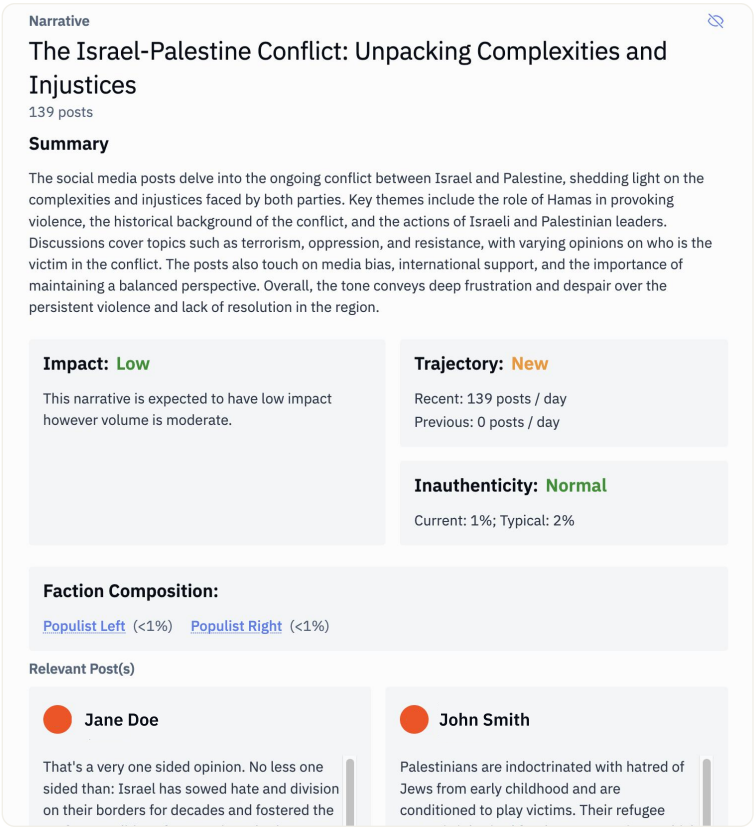


Networks behind information operations

Another facet of relationships that a team like Julie's will certainly need to grapple with is understanding the networks and state actors behind information operations. Primer has identified people and organizations who push narratives and grouped them into "factions". Primer makes it easy for analysts to see how a narrative is traversing from faction to faction to identify narratives the adversary cares most about, the propensity of these narratives to "jump" to the next faction, and who is likely to be influenced next.

For example, Primer helps analysts classify entities into groups that are known to promote narratives on behalf of foreign powers, or groups that post antivax narratives, or other groups that are known to be hostile to our organization for any reason. Primer models can identify if organizational leadership or other persons of interest are mentioned in content produced by groups known to promote disinformation.

These factions help analysts understand and predict how narratives evolve over time, so they know which ones are a high risk and which ones will likely die out. This will not only help them take appropriate action, but also show patterns that help them perform better analysis on how disinformation spreads online.



Primer has used machines to analyze a huge dataset of thousands upon thousands of unstructured pieces of content (aka, the news), and in moments create a meaningful structure that helps analysts identify patterns of disinformation. That same exercise would have been far too time consuming, expensive, and inaccurate if done by humans. But with NLP, the task becomes feasible.

Primer

Stage three - scaling

The final stage of NLP enlightenment is enabling this capability across your organization. Initial use cases like having NLP models do rote work like cull through data looking for threat actors and their activity can be used to demonstrate the value in other parts of your organization. Eventually your organization will begin to take any type of data, structured or unstructured, and put machines to the task of analyzing it. See the example below of a basic customer complaint workflow solution that could be significantly improved with NLP.

Managing customer complaints

Old way	New way
Transcoding and populating a customer complaint system.	Audio transcription happens automatically and then NLP'd into a complaint management system.

The vast majority of data is unstructured and increasing at a rate faster than humans can keep up with. To process this data effectively requires NLP so that teams across organizations can do more with less. The more data that can be processed automatically, the more you can give your teams the ability to streamline their workflows.

Because the volume of data is outpacing our ability to keep pace, this is not a matter of taking human jobs and replacing them with machines. Rather, it's about giving people the tools they need to be effective. In the next decade, organizations will begin streaming thousands and thousands of sources of unstructured data into their organizations, in real real time, from any source, to be run on NLP workflows.

And as these capabilities become widely available internally, back office teams will begin leveraging NLP to automate things like contract reviews, procurement, HR analysis and other repetitive, time-consuming tasks. Work that is currently too expensive to be done by humans will suddenly become possible, and unlock a new level of operational productivity as well as efficiency to the point that many back office resources may be reallocated to critical frontline functions for leaner, more effective operations.

More data enabled organizations will have a critical advantage, driven not only by the speed and accuracy of the analysis that they can do, but also by the data sources that they can analyze.

5 steps to getting started with NLP

- What are my data types?
- Where is my data stored?
- What questions do I want to ask of my data?
- How secure is my data?
- What are my existing data science capabilities?

Menu of NLP models

- NER - Identify people, locations, and organizations within a body of text
- NEL - Link extracted entities with your own knowledge graph of entities
- RELEX - Identify relationships between entities
- Classification - Categorize any kind of text
- Sentiment - Classify text as as positive, negative, or neutral
- Summarization - Generate a precise, tailored summary of a document
- Key Phrase Extraction - Automatically identify and extract the key information within a document
- Topic Modeling - Identify the topic that a text is talking about
- Translation - Translate texts
- Event Detection - Identity real-world events described within any piece of text
- Location Extraction - Identify all the locations in a piece of text and place them on a map
- Difference Engine - Determine and highly differences between two versions of a text

1 https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/11/fact-sheet-the-presidents-budget-advances-president-bidens-unity-agenda/#:~:text=The%20Budget%20provides%20over%20%243,of%20the%20Administration's%20Executive%20Order

2 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf