Deloitte.
Insights

A REPORT FROM THE DELOITTE CENTER
FOR GOVERNMENT INSIGHTS

Using AI to unleash the power of unstructured government data

Applications and examples of natural language processing (NLP) across government

About the Center for Government Insights

The Deloitte Center for Government Insights shares inspiring stories of government innovation, looking at what's behind the adoption of new technologies and management practices. We produce cutting-edge research that guides public officials without burying them in jargon and minutiae, crystalizing essential insights in an easy-to-absorb format. Through research, forums, and immersive workshops, our goal is to provide public officials, policy professionals, and members of the media with fresh insights that advance an understanding of what is possible in government transformation.

About the authors

WILLIAM D. EGGERS is the executive director of Deloitte's Center for Government Insights, where he is responsible for the firm's public sector thought leadership. His most recent book is *Delivering on Digital: The Innovators and Technologies that Are Transforming Government* (Deloitte University Press, 2016). His other books include *The Solution Revolution*, *Washington Post* best-seller *If We Can Put a Man on the Moon*, and *Governing by Network*. He coined the term Government 2.0 in a book by the same name. His commentary has appeared in dozens of major media outlets including the *New York Times*, *Wall Street Journal*, and *Washington Post*. He can be reached at weggers@deloitte.com or on twitter @wdeggers. He is based in Rosslyn, Virginia.

NEHA MALIK is an assistant manager with the Deloitte Center for Government Insights. She researches issues related to public–private partnerships and innovation at the federal, state, and local government levels. Malik is based in Mumbai, India.

MATT GRACIE is a managing director in the Strategy & Analytics team at Deloitte. He leads Deloitte's NLP/Text Analytics practice, which supports civilian, defense, national security, and health sector agencies in gaining insights from unstructured data to better serve their missions. With over 30 years of experience in financial services and consulting, Gracie is a thought leader with global and national experience in strategy, analytics, marketing, and consulting. He is based in McLean, Virginia.

Contents

Our Cognitive Advantage offerings are designed to help organizations transform through the use of automation, insights, and engagement capabilities. We help clients seize the insight-driven advantage with cognitive capabilities every day, around the world. What makes our approach different? Our cognitive offerings are tailored for issues that are unique to individual industries and can be integrated with other Deloitte solutions. Plus, we help our clients tap into an ecosystem of vendors and other collaborators in the industry, giving them access to leading technology, solutions, and talent that would be difficult to find otherwise. To learn more, visit our webpage on Deloitte.com.

Introduction

Tom is an analyst at the US Department of Defense (DoD).[1] All day long, he and his team collect and process massive amounts of data from a variety of sources—weather data from the National Weather Service, traffic information from the US Department of Transportation, military troop movements, public website comments, and social media posts—to assess potential threats and inform mission planning.

WHILE SOME OF the information Tom's group collects is structured and can be categorized easily (such as tropical storms in progress or active military engagements), the vast majority is simply unstructured text, including social media conversations, comments on public websites, and narrative reports filed by field agents. Because the data is unstructured, it's difficult to find patterns and draw meaningful conclusions. Tom and his team spend much of their day poring over paper and digital documents to detect trends, patterns, and activity that could raise red flags.

In response to these kinds of challenges, DoD's Defense Advanced Research Projects Agency (DARPA) recently created the Deep Exploration and Filtering of Text (DEFT) program, which uses natural language processing (NLP), a form of artificial intelligence, to *automatically* extract relevant information and help analysts derive actionable insights from it.[2]

Across government, whether in defense, transportation, human services, public safety, or health care, agencies struggle with a similar problem—making sense out of huge volumes of unstructured text to inform decisions, improve services, and save lives. Text analytics, and specifically NLP, can be used to aid processes from investigating crime to providing intelligence for policy analysis.

Think of unstructured text as being "trapped" in physical and virtual file cabinets. The promise is clear: Governments could improve effectiveness and prevent many catastrophes by improving their ability to "connect the dots" and identify patterns in available data.

This article explores NLP, its capabilities, and critical government issues it can address.

The evolution of NLP

NATURAL LANGUAGE PROCESSING dates back to the 1940s, when Roberto Busa, an Italian Jesuit priest and a pioneer of computational linguistics, analyzed the complete works of St. Thomas Aquinas, the 13th-century Catholic priest and philosopher. In 1949, Busa met with IBM founder Thomas J. Watson and persuaded him to sponsor the *Index Thomisticus*, a computer-readable compilation of Aquinas' works. The project took more than 30 years and eventually was published in 56 volumes based on more than 11 million computer punch cards, one for every word analyzed.[3]

NLP tools encompass the entire cycle of recognizing human speech, understanding and processing natural language, and generating text that can be read and interpreted by humans.

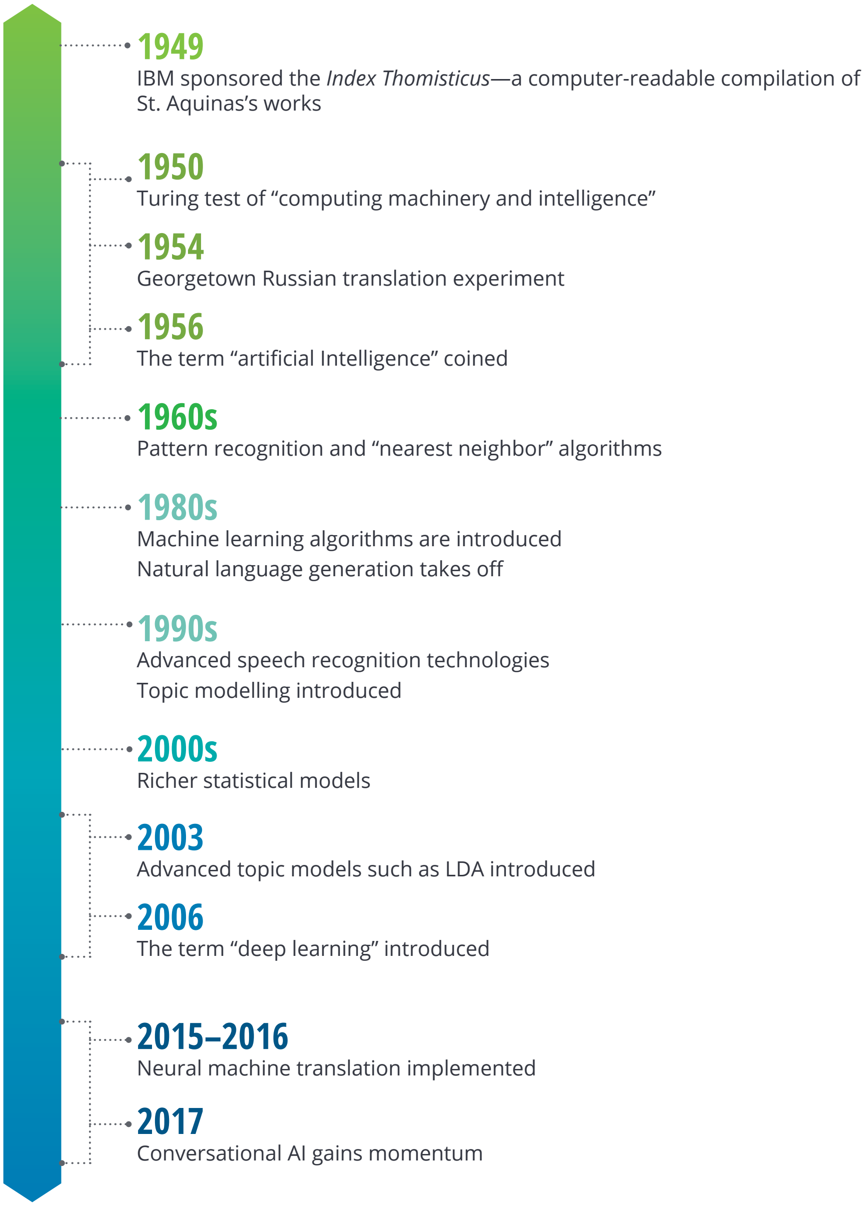
NLP first received widespread recognition in the 1950s, when researchers and linguistics experts began developing machines to automate language translation.[4] In the 1990s, historian Sharon Block used *topic modeling*, one facet of NLP, to conduct a quantitative analysis of the *Pennsylvania Gazette*, one of the most prominent American newspapers of the 18th century.[5]

Today's biggest advances in NLP tend to be guided by deep-learning methods based on neural networks (figure 1). Such methods are designed to mimic the function of the neurons in a human brain to ensure better performance.[6] Unlike applications based on algorithms or statistical methods, deep-learning NLP can automatically learn from examples.[7] According to the International Data Corporation, organizations deriving insights from structured and unstructured data could achieve an additional US$430 billion in productivity gains by 2020.[8]

The US Department of Defense has long been a pioneer in the application of NLP and Natural Language Generation (NLG) to government. NLP tools encompass the entire cycle of recognizing human speech, understanding and processing natural language, and generating text that can be read and interpreted by humans. NLG is a subset of the NLP tools that translate data into interpretable natural language narratives.[9] Back in the 1990s, the second of the three waves of AI at DARPA focused largely on developments in natural language technologies to accelerate advances in pattern recognition. And today, Natural Language Understanding (NLU), a crucial component of NLP that helps comprehend unstructured text, as well as Natural Language Generation, form a core part of DARPA's latest AI campaign to promote the development of machines that can mimic human reasoning and communication.[10] Of the DoD's total AI spend, NLP has emerged as one of the larger investments with a budget totaling close to US$83 million in 2017, reflecting a jump of nearly 17 percent from 2012 spending.[11]

FIGURE 1

The evolution of NLP and underlying algorithms



1949
IBM sponsored the *Index Thomisticus*—a computer-readable compilation of St. Aquinas's works

1950
Turing test of "computing machinery and intelligence"

1954
Georgetown Russian translation experiment

1956
The term "artificial Intelligence" coined

1960s
Pattern recognition and "nearest neighbor" algorithms

1980s
Machine learning algorithms are introduced
Natural language generation takes off

1990s
Advanced speech recognition technologies
Topic modelling introduced

2000s
Richer statistical models

2003
Advanced topic models such as LDA introduced

2006
The term "deep learning" introduced

2015–2016
Neural machine translation implemented

2017
Conversational AI gains momentum

Sources: Thomas N. Winter, "Roberto Busa, S.J., and the inventor of the machine-generated concordance," University of Nebraska-Lincoln, January 1999; Bhargav Shah, "The power of natural language processing: Today's boom in artificial intelligence," *Medium*, July 13, 2017; Chris Smith et al., *The history of artificial intelligence*, University of Washington, December 2006; Eric Eaton, "Introduction to machine learning," presentation, University of Pennsylvania; Kendall Fortney, "Pre-processing in natural language machine learning," *Towards Data Science,* November 28, 2017; Clark Boyd, "The past, present, and future of speech recognition technology," *Medium*, January 10, 2018; Regina Barzilay, "Natural language processing," lecture, MIT, September 2012; Build with Google Cloud, "A history of machine learning," accessed December 19, 2018; Thomas Hofmann, "Probabilistic latent semantic indexing," proceedings of the twenty-second Annual International SIGIR Conference on Research and Development in Information Retrieval, accessed December 19, 2018; Robert Dale, Barbara Di Eugenio, and Donia Scott, "Introduction to the special issue on natural language generation," *Computational Linguistics* 24, no. 3 (September 1998); *Medium*, "History and frontier of the neural machine translation," August 17, 2017; Ram Menon, "The rise of the conversational AI," *Forbes*, December 4, 2017.

Seven key technical capabilities of NLP

GOVERNMENT AGENCIES AROUND the world are accelerating efforts to abandon paper and modernize how they handle data. According to the National Archives and Records Administration, the US federal government has already digitized more than 235 million pages of government records and plans to reach 500 million pages by fiscal 2024.[12]

While digitizing paper documents can help government agencies increase efficiency, improve communications, and enhance public services, most of the digitized data will still be unstructured. That's where NLP comes in.

With recent technological advances, computers now can read, understand, and use human language. They can even measure the *sentiment* behind certain text and speech (see sidebar, "Applications of natural language technologies").[13] These capabilities (figure 2) allow government agencies to recognize patterns, categorize topics, and analyze public opinion.

NLP has seven key capabilities:

Topic modeling

Topic modeling is a method based on statistical algorithms to help uncover hidden topics from large collections of documents. Topic models are unsupervised methods of NLP; they do not depend on predefined labels or ontologies. A popular method within topic modeling is Latent Dirichlet Allocation (LDA), which is used to discover latent patterns in a sea of unstructured data. The US Securities and Exchange Commission (SEC), for example, made its initial foray into natural language processing in the aftermath of the 2008 financial crisis. The SEC used LDA to identify potential problems in the disclosure reports of companies charged with financial misconduct.[14]

With recent technological advances, computers now can read, understand, and use human language. They can even measure the sentiment behind certain text and speech.

The UK government uses the same technique to better understand public comments on GOV.UK. With LDA, the government can see how customer complaints and comments relate to one another; for example, that mortgage complaints often contain allegations of racial discrimination. Uncovering such topics allows the government to address them.[15]

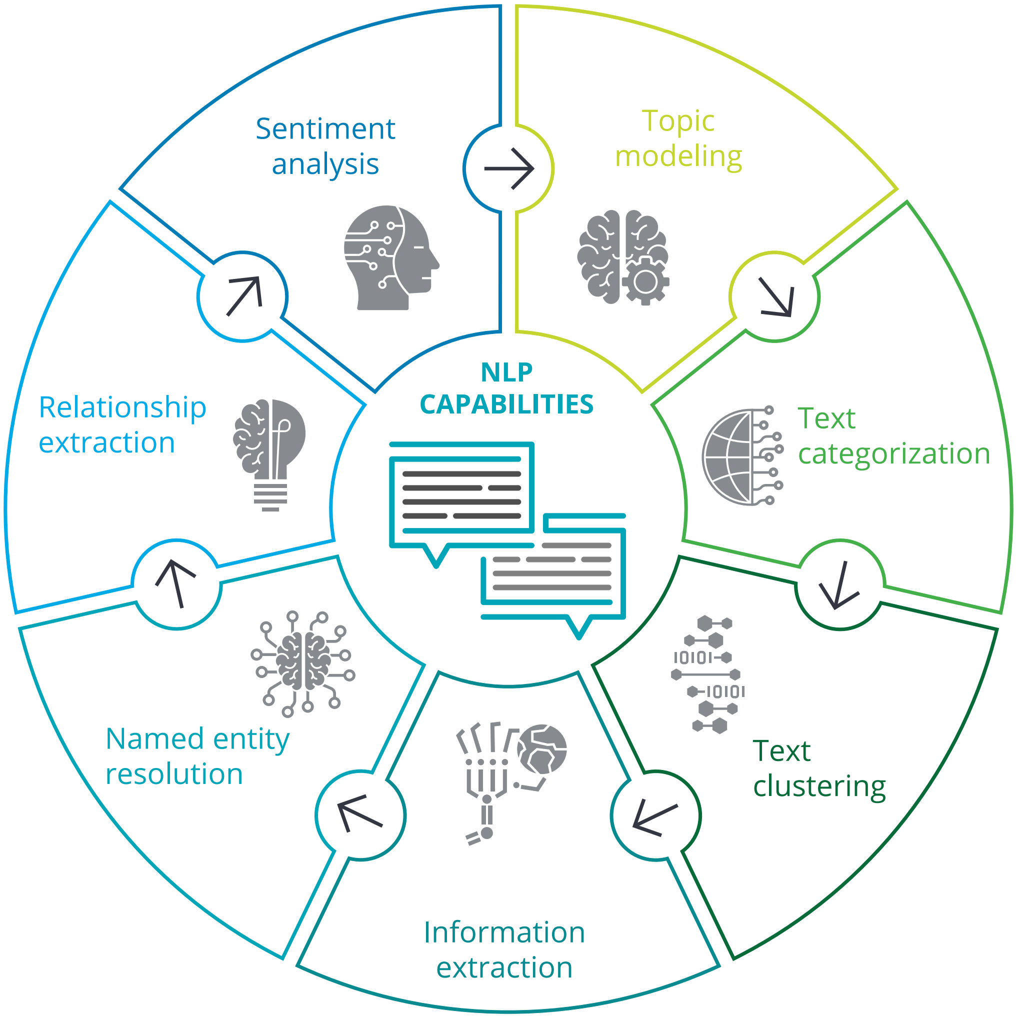
Text categorization

This method sorts text into specific taxonomies, typically after training by humans. For instance, complaints can be automatically filed into specific categories; tweets can be categorized as pro candidate A or against candidate B.[16]

One use concerns the classification of sensitive information. A research study used NLP and

FIGURE 2

Key NLP capabilities



Source: Deloitte analysis.

machine learning on nearly one million declassified diplomatic cables from the 1970s to identify the features of records that had a higher chance of being classified, such as those concerning international negotiations and military matters.[17]

In another study, researchers underscore the benefits of automating security classification by classifying US security documents using NLP. The researchers also propose using text categorization to classify over a 100,000 declassified government records available on the Digital National Security Archive.[18]

Text clustering

Text clustering is a technique used to group text or documents based on similarities in content. It can be used to group similar documents (such as news articles, tweets, and social media posts), analyze them, and discover important but hidden subjects.[19]

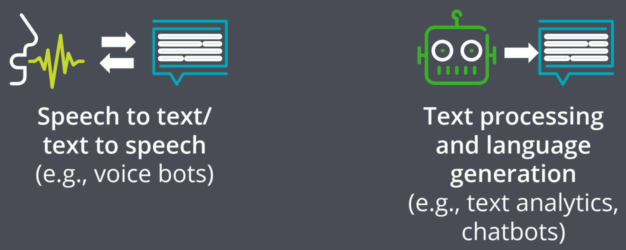
The Center for Tobacco Products (CTP), part of the US Food and Drug Administration, uses text clustering and topic modeling to group documents based on specific key terms. For instance, documents related to the topic "menthol" could form one cluster, with those concerning menthol usage among youths representing a subset.[20] Text clustering helps the CTP organize and glean insights from documents—from FDA submissions for new tobacco products to advertising claims—to better understand the impact of the manufacture, marketing, and distribution of tobacco products on public health and help inform policy-making, particularly concerning the implicit marketing of tobacco products to youths.[21]

APPLICATIONS OF NATURAL LANGUAGE TECHNOLOGIES

FIGURE 3

Core applications of natural language technologies

NLP processors eliminate ambiguities



Source: Deloitte analysis.

Some of the most common applications (figure 3) of natural language processing are:

- **Speech to text.** Popularly known as speech recognition, this application serves as a foundation for the other applications of NLP. As the name suggests, its basic purpose is to convert spoken language into text that can be processed by other applications.[22]
- **Text to speech.** Also called speech synthesis, this application converts text into speech.[23]
- **Text processing.** Deriving text from narratives is one thing; the critical step is to derive *value* from of the text. Text processing identifies and extracts bits of information that can be used by data scientists to derive meaningful insights. The entire process involves little to no human intervention.[24]
- **Natural language generation (NLG).** While text processing translates unstructured text into useful data, NLG can *summarize* insights directly from a dense data dashboard, eliminating the need to spend analysts' time on manual summarization.[25]
- **Chatbots.** Bots backed by NLP technology can understand a great deal of human communication. While chatbots cannot be expected to catch every nuance, they *can* be customized to address and answer specific questions relevant to a particular organization.[26]
- **Machine translation.** This involves the automatic translation of text from one human language to another. Machine translation has advanced from being predominantly rules-based to using statistical models, and most recently, to the human-like "thinking" of neural networks.[27]

Information extraction

Information extraction is used to automatically find meaningful information in unstructured text. One potential application can improve transparency and accuracy in crime reporting. Often, police reports are written in haste, while crime witnesses' and victims' accounts can be incomplete due to embarrassment or fear of repercussions. Researchers at Claremont Graduate University found that NLP technology could comb through witness and police reports and related news articles to identify crucial elements such as weapons, vehicles, time, people, clothes, and locations with high precision.[28]

Named entity resolution

This method can extract the names of persons, places, companies, and more, and classify them into predefined labels and link the named entities to a specific ontology.[29] For instance, a text may contain references to the entity "Georgia," which is both a nation and a US state. With the help of entity resolution, "Georgia" can be resolved to the correct category, the country or the state.

Government agencies can extract named entities in social media to identify threat perpetrators of cybercrime, for instance, as well as their future prospects.[30] The more ontologies are defined in the NLP tool, the more effective the outcome.

Relationship extraction

This capability helps establish semantic relations between entities. For instance, if a document mentions the Office of Management and Budget and the US federal government, relationship extraction identifies and creates a parent–agency relationship between them.[31]

Sentiment analysis

Sentiment analysis decodes the meaning behind human language, allowing agencies to analyze and interpret citizen and business comments on social media platforms, websites, and other venues for public comment. Washington, DC's sentiment analysis program (GradeDC.gov), for example, examines citizens' feedback by analyzing their comments on social media platforms. The district was the first municipal government in the United States to adopt such an initiative.[32]

Another study used sentiment analysis to examine the experiences of patients with various health care providers throughout the United States. The authors used an exhaustive dataset of more than 27 million tweets related to patient experience over a period of four years. A principal objective of the study was to examine the variation in such experiences across the country. The findings suggested a higher proportion of positive experiences for metropolitan areas compared to the nonmetropolitan areas.[33]

NLP can address critical government issues

THE DELUGE OF unstructured data pouring into government agencies in both analog and digital form presents significant challenges for agency operations, rulemaking, policy analysis, and customer service. NLP can provide the tools needed to identify patterns and glean insights from all of this data, allowing government agencies to improve operations, identify potential risks, solve crimes, and improve public services. Ways in which NLP can help address important government issues are summarized in figure 4.

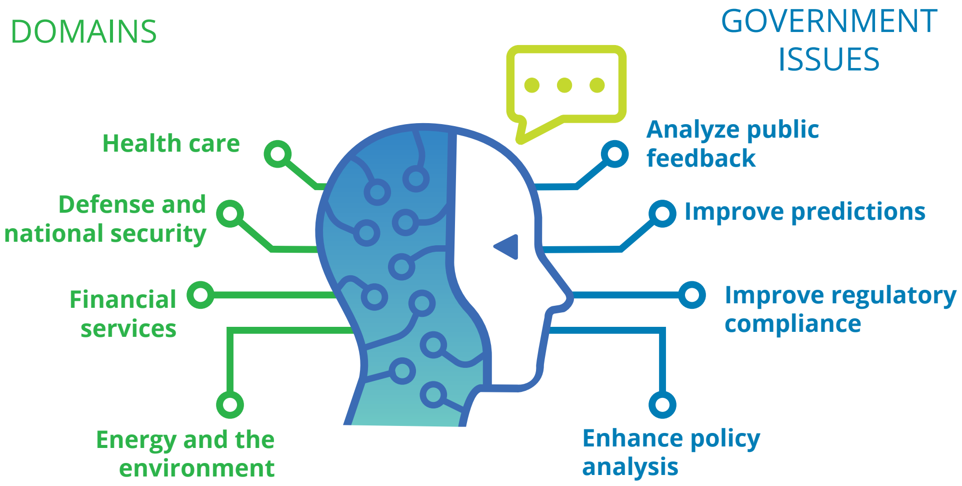
Analyzing public feedback

Whether it's a comment dropped into a suggestion box at the Department of Motor Vehicles, an online survey on the Internal Revenue Service website, or various grievances posted on social media, public comments help government agencies to understand citizen and business concerns and better serve the public. NLP can analyze feedback, particularly in unstructured content, far more efficiently than humans can. Many organizations today are monitoring and analyzing consumer responses on social media with the help of sentiment analysis.

FiscalNote, a data and media company, used NLP to analyze more than 22 million comments on the FCC's proposal to repeal net neutrality. A major challenge was sifting through millions of responses to separate genuine comments from fake ones generated by bots. To identify fakes, the company used NLP techniques to cluster the comments and identify similarities in sentences and paragraph structures.[34]

FIGURE 4

NLP can address important issues across government domains



Source: Deloitte analysis.

NLP also can help governments engage with citizens and provide answers to their questions. Singapore's government used NLP to create "Ask Jamie," a virtual assistant that can be placed on agency websites and trained to respond to questions posed by citizens. For questions with multiple answer options and permutations, Ask Jamie can ask questions to narrow down to an answer relevant to the query posed.[35]

Improving forensics investigations

A growing number of government agencies are using NLP-based solutions to improve investigations in critical areas such as law enforcement, defense, and intelligence. The DoD's DEFT program referenced above uses NLP to uncover connections implicit in large text documents. Its objective is to improve the efficiency of defense analysts who investigate multiple documents to detect anomalies and causal relationships.[36]

NLP can provide the tools needed to identify patterns and glean insights from data, allowing government agencies to improve operations, identify potential risks, solve crimes, and improve public services.

The European Union's Horizon 2020 program launched an initiative called RED (Real-time Early Detection) Alert, aimed at countering terrorism by using NLP to monitor and analyze social media conversations. RED Alert is designed to provide early alerts of potential propaganda and signs of warfare by identifying online content posted by extremists. To comply with the General Data Protection Regulation (GDPR), this analysis uses homomorphic encryption, a method that allows mathematical operations to be performed on encrypted text without disturbing the original encryption.[37]

Improving predictions to aid decision-making

One of the most striking characteristics of NLP is its ability to facilitate better predictions, which can help agencies design preemptive measures. The police department of Durham, North Carolina, uses NLP in crimefighting by enabling the police to observe patterns and interrelations in criminal activities and identify pockets with a high incidence of crime, thus allowing for quicker interventions. This contributed to a 39 percent drop in violent crime in Durham from 2007 to 2014.[38]

NLP also is being used to combat child trafficking. About 75 percent of child trafficking involves online advertisements. DARPA, in collaboration with commercial technology experts, has developed a platform that monitors and draws connections among the dubious content of online advertisements. Virginia's Fairfax County Police Department and New Orleans's Homeland Security investigations both use this advanced software to identify high-risk web advertisements and detect code words used by traffickers.[39]

Similarly, the Australian Securities and Investments Commission (ASIC) is piloting the use of NLP applications to identify dubious product promotions and sales malpractice.[40]

Enhancing policy analysis

The World Bank's Poverty and Equity Global Practice Group used LDA topic modeling to measure changes in policy priorities by examining presidential speeches in 10 Latin American countries and Spain from 1819 to 2016. Using LDA, the authors could identify the main topics for each document and indicate the variation in their significance across countries and over time. In Peru, for instance, topics on infrastructure and public services diminished in importance over time. With the help of topic modeling, the authors were able to establish, for each nation, a negative correlation between policy volatility and long-term growth.[41]

Improving regulatory compliance

NLP can engender stricter adherence to regulations. One case in point is a pilot launched by the General Services Administration's (GSA's) Office of Government-wide Policy (OGP). Solicitations posted on the Federal Business Opportunities website (fbo.gov) must comply with Section 508 of the federal Rehabilitation Act, which requires federal agencies "to make their electronic and information technology accessible to people with disabilities."[42] A Solicitation Review Tool (SRT) piloted by OGP uses NLP to automatically check for compliance with Section 508 with 95 percent accuracy, which allows the GSA to redeploy some of its human resources to other important tasks.[43]

Another pertinent example relates to the use of AI-based pattern recognition technology by the World Anti-Doping Agency (WADA). The WADA plans to deploy AI algorithms that can instantaneously search through data collected by global anti-doping agencies to identify any breach of conduct by athletes, allowing it to monitor athletes effectively while maximizing its limited resources.[44]

NLP capabilities across different domains

NLP CAPABILITIES HAVE the potential to be used across a wide spectrum of government domains. In this chapter, we explore several examples that exemplify the possibilities in this area.

Health care

The US National Library of Medicine's Lister Hill National Center for Biomedical Communications uses NLP to "de-identify" clinical information in narrative medical reports, protecting patient privacy while preserving clinical knowledge.[45]

Topic modeling has been used to deconstruct large biomedical datasets for medical surveillance. The National Center for Toxicological Research, for instance, used topic modeling on 10 years of reports extracted from the FDA's Adverse Event Reporting System (FAERS) to identify relevant drug groups from more than 60,000 drug adverse event pairs, i.e., pairs of drugs and adverse events in which the adverse reaction is caused by the drug. The objective: to better predict potential adverse drug reactions.[46]

Defense and national security

DARPA's DEFT program uses NLP to automatically extract operationally relevant information from unstructured text to help defense analysts derive actionable insights from data.[47]

The Institute for Strategic Dialogue in the United Kingdom developed NLP-based solutions to monitor signs of extremism and radicalization. Analysts used NLP capabilities to examine comments on select public pages and flag instances of violent or aggressive language. Of the total sample of 42,000 individuals identified online, nearly 800 were found to indicate signs of extremism.[48]

Energy and the environment

The US Department of Energy's Oak Ridge national laboratory is leveraging NLP capabilities to extract data on energy ecosystem components to rank the top clean energy innovation ecosystems in the United States.[49] They used NLP to transform text and numerical data into metrics on clean energy innovation activity and geography. This helps investors, researchers, and corporations rapidly identify, quantify, and characterize clean energy innovation.[50]

Researchers at the Environmental Defense Fund are working to develop a system backed by NLP that can analyze applications for oil and gas permits submitted under the national Environmental Protection Act. The system would provide a deeper analysis of filed applications, thereby helping local regulators and other stakeholders determine whether a project may pose a threat to wildlife, water, or cultural heritage sites.[51]

Getting started with NLP

GOVERNMENT AGENCIES CAN build NLP capabilities by following the steps elaborated below.

Defining the problem

The first step is to define the problems the agency faces and which technologies, including NLP, might best address them. For example, a police department might want to improve its ability to make predictions about crimes in specific neighborhoods. After mapping the problem to a specific NLP capability, the department would work with a technical team to identify the infrastructure and tools needed, such as a front-end system for visualizing and interpreting data.

Building the team

It's important for agencies to create a team at the beginning of the project and define specific responsibilities. For example, agency directors could define specific job roles and titles for software linguists, language engineers, data scientists, engineers, and UI designers. Data science expertise outside the agency can be recruited or contracted with to build a more robust capability. Analysts and programmers then could build the appropriate algorithms, applications, and computer programs. Technology executives, meanwhile, could provide a plan for using the system's outputs. Building a team in the early stages can help facilitate the development and adoption of NLP tools and helps agencies determine if they need additional infrastructure, such as data warehouses and data pipelines.

Identifying the data

Next, organizations should identify the relevant data and determine its accessibility. Some data may be easily acquired; others may not be in a machine-readable format, or may be unlabeled or of poor quality. If necessary, agencies can use optical character recognition (OCR) to convert the data into a machine-readable format, clean it, create a labeled data set, and perform exploratory analysis.[52]

Developing models

Initiative leaders should select and develop the NLP models that best suit their needs. The final selection should be based on performance measures such as the model's precision and its ability to be integrated into the total technology infrastructure. The data science team also can start developing ways to reuse the data and codes in the future.

Testing and deploying the model

The next step is to amend the NLP model based on user feedback and deploy it after thorough testing. It is important to test the model to see how it integrates with other platforms and applications that could be affected. Additional testing criteria could include creating reports, configuring pipelines, monitoring indices, and creating audit access. Another key element is training end users.

Managing, changing, scaling, and reiterating

The postdeployment stage typically calls for a robust operations and maintenance process. Data scientists should monitor the performance of NLP models continuously to assess whether their implementation has resulted in significant improvements. The models may have to be improved further based on new data sets and use cases. Government agencies can work with other departments or agencies to identify additional opportunities to build NLP capabilities.

Staying ahead of the curve

As the digitization of information accelerates, government agencies around the world will increasingly face an onslaught of unstructured text—social media posts, user comments on public websites, emails, narrative reports from government employees, or applications for permits or new products—that new technologies can analyze in ways never before possible. By taking steps *now* to harness the power of NLP and other machine-learning capabilities, agencies can stay ahead of the curve and derive meaningful insights from the data *before* they are overwhelmed.

Endnotes

1. Tom is a hypothetical example.
2. Boyan Onyshkevych, “Deep exploration and filtering of text (DEFT),” Defense Advanced Research Projects Agency, US Department of Defense, accessed December 19, 2018.
3. Thomas N. Winter, “Roberto Busa, S.J., and the invention of the machine-generated concordance,” University of Nebraska-Lincoln, Faculty publications, Classics and religious studies department, January 1999.
4. Stanford University, “Natural language processing, overview: History,” accessed December 19, 2018.
5. Sharon Block, “Doing more with digitization—An introduction to topic modeling of early American sources,” Tales from the Vault, *Common-Place* 6, no.2, January 2006.
6. Jianfeng Gao, “An introduction to deep learning for natural language processing,” Presentation at the Microsoft AI & Research International Summer School on Deep Learning 2017, July 20–21, 2017; Robert D. Hof, “Deep learning: With massive amounts of computational power, machines can now recognize objects and translate speech in real time; Artificial intelligence is finally getting smart,” *MIT Technology Review*, accessed December 19, 2018.
7. Quan Wan, Ellen Wu, and Dongming Lei, “Deep learning for natural language processing,” University of Illinois at Urbana-Champaign, accessed December 19, 2018.
8. International Data Corporation, “Information transformation initiatives create information-based organizations,” accessed December 19, 2018.
9. Chrissy Kidd, “NLU vs NLP: What’s the difference?,” bmc blogs, May 28, 2018.
10. Ibid.
11. Amber Corrin, “DoD spent $7.4 billion on big data, AI and the cloud last year. Is that enough?,” C4ISRNET, December 6, 2017.
12. National Archives, *Strategic plan 2018–2022*, February 2018.
13. SAS Institute Inc., “Natural language processing: What it is and why it matters,” accessed December 19, 2018.
14. Scott W. Bauguess, “The role of big data, machine learning, and AI in assessing risks: A regulatory perspective,” Champagne keynote address, US Securities and Exchange Commission, June 21, 2017.
15. Dan Heron, “Understanding more from user feedback,” Data in Government blog, Gov.UK, November 9, 2016.
16. Fabrizio Sebastiani, “Text categorization,” *Encyclopedia of Database Technologies and Applications*, 2005.
17. Renato Rocha Souza et al., “Using artificial intelligence to identify state secrets,” Semanticsscholar.org, accessed December 19, 2018.
18. J. David Brown, “Developing an automatic document classification system—A review of current literature and future directions,” Defence Research and Development Canada, January 2010; National Security Archive, accessed December 19, 2018.
19. Meaning Cloud, “What is text clustering?,” accessed December 19, 2018.
20. Hesha J. Duggirala et al., “Data mining at FDA,” US Food and Drug Administration, August 20, 2018.
21. See, for instance, US Food and Drug Administration, *Tobacco regulatory science research program at FDA’s Center for Tobacco Products: Summary and highlights, Fiscal Years 2010–2017*, June 2018.
22. Ayn de Jesus, “AI for speech recognition: Current companies, technology, and trends,” emeRJ, August 2018.

23. Chris Lin, Qian (Sarah) Mu, and Yi Shao, "iTalk: A 3-component system for text-to-speech synthesis," Stanford University, 2017.

24. Niels Kasch, "Text analytics and natural language processing in the era of big data," *Pivotal Blog*, October 23, 2014.

25. *Medium*, "The ultimate guide to natural language generation," January 30, 2018.

26. David Gilbert, "Machine learning text analysis unlocks the power of natural language processing," Samsung Insights, September 8, 2017.

27. Microsoft Corporation, "What is a neural machine translation (NMT)?," accessed December 19, 2018.

28. Justin Ku, Alicia Iriberri, and Gondy Leroy, "Natural language processing and e-government: Crime information extraction from heterogeneous data sources," conference paper at the 9th Annual International Digital Government Research Conference, January 2008.

29. Christopher Dozier et al., "Named entity recognition and resolution in legal text," chapter from *Semantic Processing of Legal Texts* (Berlin: Springer, 2010).

30. Paul A. Watters, "Named entity resolution in social media," chapter 2 from *Automating Open Source Intelligence* (Elsevier, 2016).

31. Nguyen Bach and Sameer Badaskar, "A review of relation extraction," May 2011.

32. Matthew McClellan and Stephen Goldsmith, "From comment cards to sentiment mining: The future of government service rating," Data-Smart City Solutions, September 18, 2013.

33. Kara C Sewalk et al., "Using Twitter to examine Web-based patient experience sentiments in the United States: Longitudinal study," *Journal of Medical Internet Research*, Vol. 20 (10), October 2018.

34. Issie Lapowski, "How bots broke the FCC's public comment system," November 28, 2017.

35. Government Technology Agency, "'Ask Jamie' virtual assistant," accessed December 19, 2018.

36. Onyshkevych, "Deep exploration and filtering of text (DEFT)."

37. John Stevenson, "EU-funded project uses artificial intelligence to tackle terrorist cyber-propaganda," City University of London, July 11, 2017.

38. Kevin Ebi, "How Durham, N.C. fights crime with data—and wins," Smart Cities Council, September 17, 2014.

39. Kristin Quinn, "Modern slavery: Cognitive computing and geospatial technology help law enforcement track, locate, and rescue human trafficking victims," *Trajectory*, November 1, 2016.

40. Justin Hendry, "ASIC eyes AI to crack down on dodgy financial practices," *itnews*, February 22, 2018.

41. Oscar Calvo-González, Axel Eizmendi, and Germán Reyes, "Winners never quit, quitters never grow: Using text mining to measure policy volatility and its link with long-term growth in Latin America," policy research working paper 8310, World Bank Group, January 2018.

42. US General Services Administration, "IT accessibility laws and policies," November 2018.

43. Keith Nakasone, "Game changers: Artificial intelligence part II; artificial intelligence and the federal government," US General Services Administration, March 7, 2018.

44. Hong Bui, "Artificial intelligence takes on doping," Veritone, April 10, 2018.

45. US National Library of Medicine, "De-identification tools," accessed December 19, 2018.

46. Mitra Rocca, *Lessons learned from NLP implementations at FDA*, US Food and Drug Administration, June 15, 2017, p. 24.

47. Onyshkevych, "Deep exploration and filtering of text (DEFT)."

48. Jacob Davey, Jonathan Birdwell, and Rebecca Skellett, "Counter conversations: A model for direct engagement with individuals showing signs of radicalisation online," Institute for Strategic Dialogue, 2018, p. 6.

49. Jessica Lin et al., "Ecosystem discovery: Measuring clean energy innovation ecosystems through knowledge discovery and mapping techniques," Oak Ridge National Laboratory, study abstract, accessed December 19, 2018.

50. Ibid.

51. Nasdaq, "How artificial intelligence is taking over oil and gas," August 10, 2018.

52. Certain NLP-specific cleaning tasks such as tokenization, stemming/lemmatization, part of speech tagging, and stopword removal also may have to be carried out to ensure a clean data set.

Acknowledgments

Subject matter experts made significant contributions to this report. In particular, the authors would like to thank **Aleks Ontman**, **Fallon Farmer**, **Matthew Gracie**, **Peter Viechnicki**, **Shahvaiz Hanif**, **Vishal Kapur**, **Will Frankenstein**, and **Scott Buchholz** of Deloitte Consulting, LLP.

Pankaj Kishnani from the Deloitte Center for Government Insights also contributed to the research of the project, while **Mahesh Kelkar** from the Center provided thoughtful feedback on the drafts. **David Noone** played a critical role in project coordination.

Contacts

William D. Eggers
Executive director
Center for Government Insights
Deloitte Service LP
+1 571 882 6585
weggers@deloitte.com

Matt Gracie
Managing director
Strategy & Analytics team
Deloitte Consulting LLP
+1 571 766 7327
magracie@deloitte.com

Deloitte.
Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights.

Follow @DeloitteInsight

Deloitte Insights contributors
Editorial: Karen Edelman, Blythe Hurley, and Rupesh Bhat
Creative: Sonya Vasilieff
Promotion: Alexandra Kawecki
Cover artwork: David Owens

About Deloitte Insights

Deloitte Insights publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee (“DTTL”), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as “Deloitte Global”) does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the “Deloitte” name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.