GRADIENT FLOW

2020 NLP Industry Survey Report

How Companies Use Natural Language Processing Technologies

by Ben Lorica and Paco Nathan

Sponsored by John Snow LABS

Executive Summary

The Natural Language Processing (NLP) Industry Survey was an online survey which ran for 41 days (July 5 to August 14, 2020). A total of 571 respondents from more than 50 countries completed the survey. A quarter of all respondents held technical leadership roles. Respondents were recruited via social media, online advertising, the Gradient Flow Newsletter, and through industry partners and contacts. We hope you find this survey informative and helpful; we plan to repeat this survey in the years to come.

Here are some highlights:

- **NLP budgets are growing:** 53% of respondents who are Technical Leaders stated their NLP budget was at least 10% higher compared to 2019. 31% of respondents who are Technical Leaders stated their NLP budget was at least 30% higher compared to 2019. The same trend applies when we examined respondents from large companies (firms with 5,000 or more employees). 39% of respondents who worked at large companies stated their NLP budget was at least 10% higher compared to 2019. 21% of respondents who worked at large companies stated their NLP budget was at least 30% higher compared to 2019.

- **Document Classification and NER are the most popular applications of NLP technologies:** The four most popular applications of NLP are Document Classification, Named Entity Recognition (NER), Sentiment Analysis, and Knowledge Graphs. Respondents from the Healthcare industry cited de-identification as another common NLP use case. Document Classification and NER are by far the most popular use cases among respondents who worked in organizations further along the NLP adoption curve. Note that the accuracy of these kinds of use cases can be enhanced by tools and practices for annotating (labeling) text.

- **Users value accuracy:** More than 40% of all respondents cited accuracy as the most important criteria they used when they evaluate NLP libraries. A quarter of respondents cited accuracy as the main criteria they used when evaluating NLP cloud services. Accuracy refers to pre-trained models that get used in multi-stage pipelines in NLP libraries. These models let users input text to get common outputs (e.g., tokens, lemmas, part-of-speech (POS), similarity, and entity recognition).

(Executive Summary cont.)

- **NLP libraries:** Half of all respondents (53%) used at least one of the top two libraries: Spark NLP and spaCy. A third (33%) of all respondents stated they use Spark NLP, making it the most popular NLP library in our survey. More than a quarter (26%) of all respondents stated they use spaCy. AllenNLP, a newer PyTorch-based library for NLP research, was the third most popular library. We also looked at the most popular libraries in a few key industry sectors: Healthcare (Spark NLP), Technology (spaCy), and Financial Services (nltk).

- **NLP cloud services:** 77% of all survey respondents indicated they use at least one of the four NLP cloud services we listed (Google, AWS, Azure, and IBM), with Google's service garnering the most users. Google Cloud is particularly popular among respondents who are still in the early stages of adopting NLP. Cloud usage rate drops slightly when we look at companies that have more experience in deploying NLP: 65% of respondents working at companies further along the NLP adoption curve use at least one of the NLP cloud services we listed (Google, AWS, Azure, and IBM). Respondents cited cost as the key challenge they face when using NLP cloud services. There are general concerns about extensibility since so many NLP applications depend on domain-specific language use, and the cloud providers have, in general, been slow to service these market needs.

- **Data sources:** Data from files and databases top the list of data sources used in NLP projects. 61% of all Technical Leaders stated they used Files (e.g., PDF, TXT, and DOCX) for their NLP systems. More than a third (36%) of all Technical Leaders stated their organization used a text annotation tool for labeling training data for NLP.

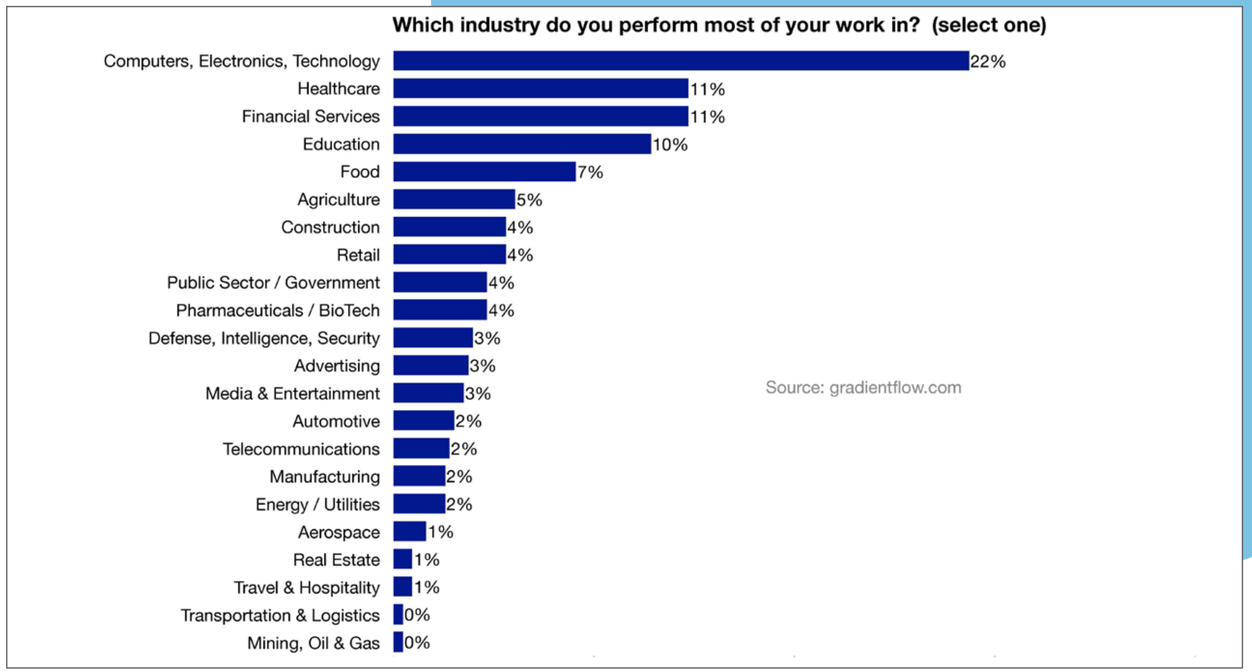
Introduction

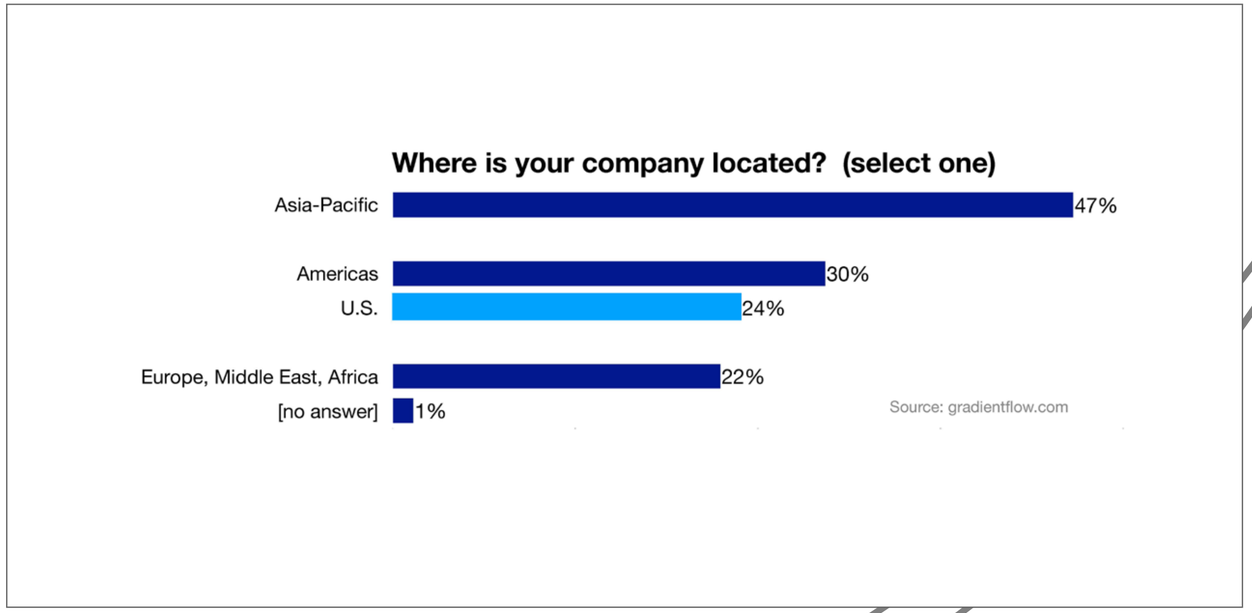
The term “NLP”—or natural language processing—encompasses a wide range of business use cases that are mostly text based. Consider that people use text to record and transmit their communications in general, and, as such, it’s one of the most widely available and “interoperable” data formats. While some industry sectors such as finance and healthcare have long used text mining approaches, much more sophisticated use cases have been gaining traction in industry, especially during the latter 2010s.

Our industry survey results consider several important contrasts, including: organizations with years of history deploying NLP applications in production compared to those which are exploring NLP, Technical Leaders versus general respondents, company size, scale of documents, and geographic region. We draw insights and indicate trends based on those contrasts.

The range and implications of NLP applications also tends to track the formality of text used as input data. Some sectors, such as healthcare and finance, have formal implications (contractual, regulatory) about their uses of text, so their NLP applications bring much value. Other sectors, such as social media and ecommerce, may have relatively casual uses of text—on the surface; even so, there can be serious implications downstream. For example, bias in recommender systems often depends on how the features used to train models or score results were extracted from large volumes of text.

When evaluating the suitability of NLP libraries for industry use cases, it’s important to understand that these libraries typically provide pipelines where machine learning models get applied at each stage. One must gauge the effectiveness of a multi-stage pipeline for a given application. Since late 2017, the use of deep learning in models used in NLP pipeline stages has dramatically increased accuracy of these libraries overall, making a much broader range of business use cases more practical.





Demographics and Key Segments

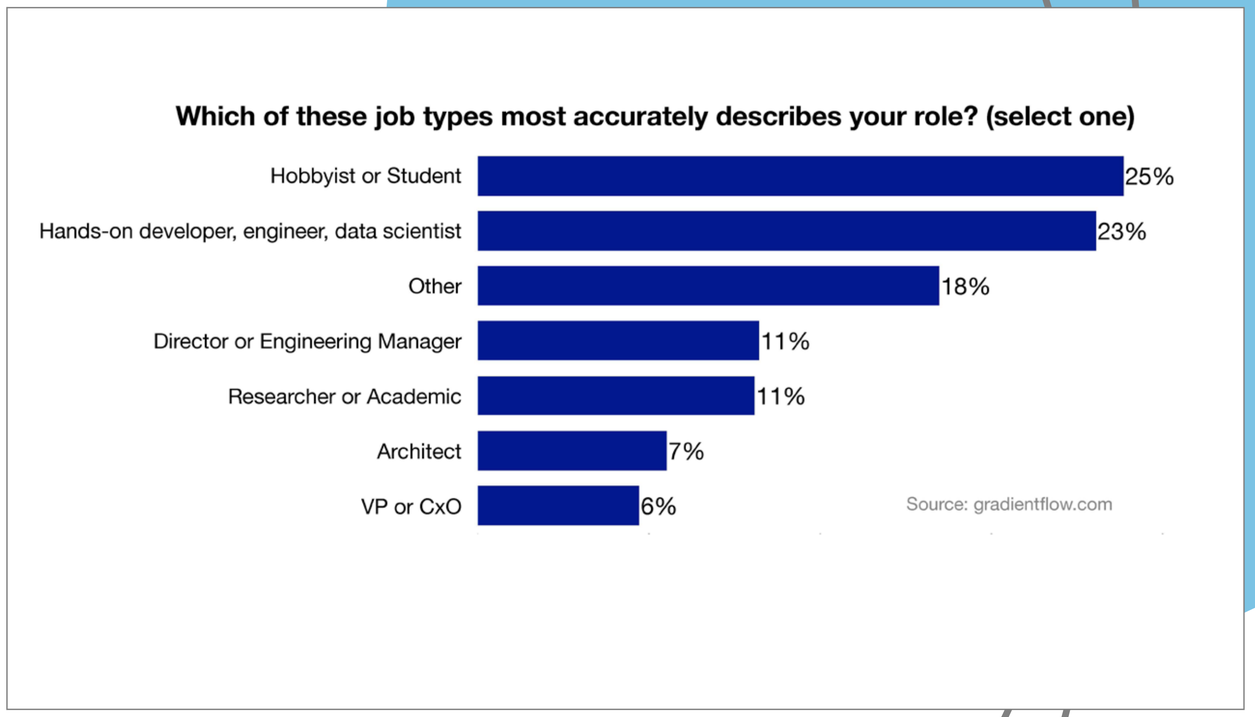
Respondents came from a wide variety of industry sectors. Along with the Technology sector, companies in Healthcare and Financial Services have long used text mining technology. We will occasionally highlight responses from these three key sectors (Technology, Healthcare, and Finance) in the remainder of this report.

Respondents came from over 50 countries: close to half came from the Asia-Pacific region, and about a quarter came from the US.

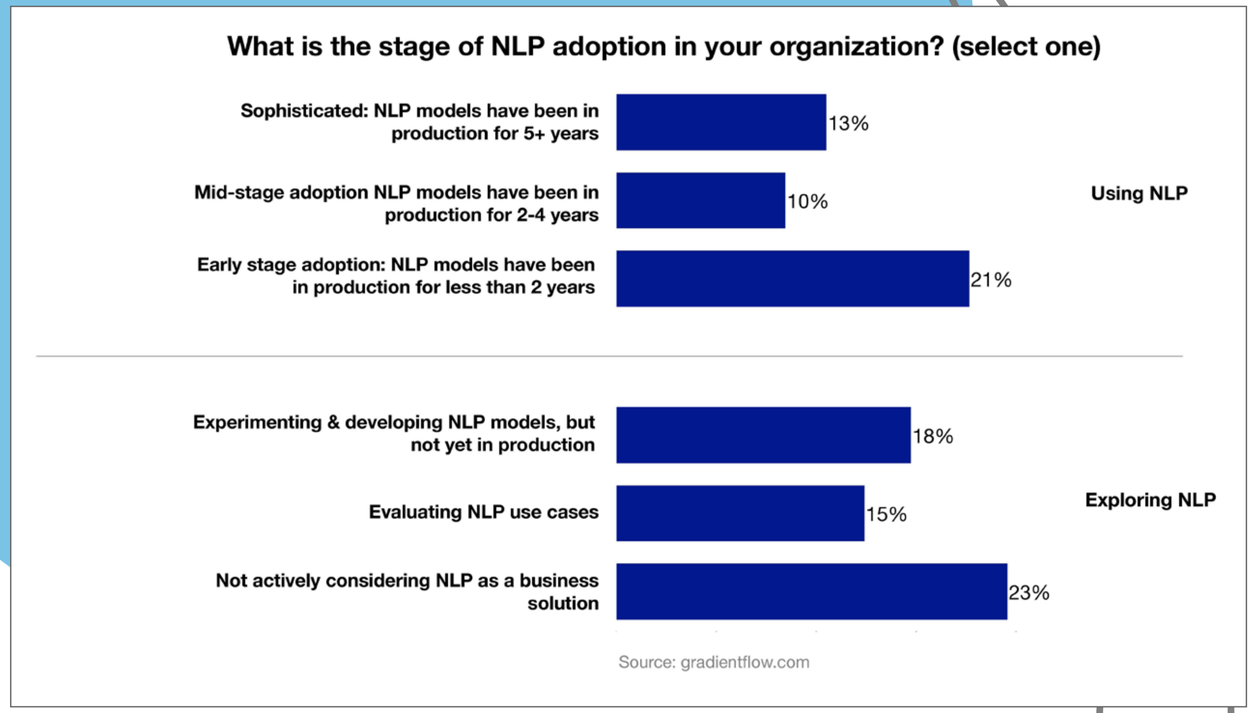
A quarter of respondents have “hands-on” technical roles, but Architects and other technical leadership roles were also well represented in the survey.

In the remainder of this report, we use the term **Technical Leaders** to refer to the group of respondents who indicated they hold one of these job types: “Director or Engineering Manager,” “Architect,” “VP or CxO.” Technical Leaders comprise about a quarter of all survey respondents.

We will occasionally group respondents into audience segments based on their response to a few questions in the survey.



GRADIENT FLOW



Stage of Adoption

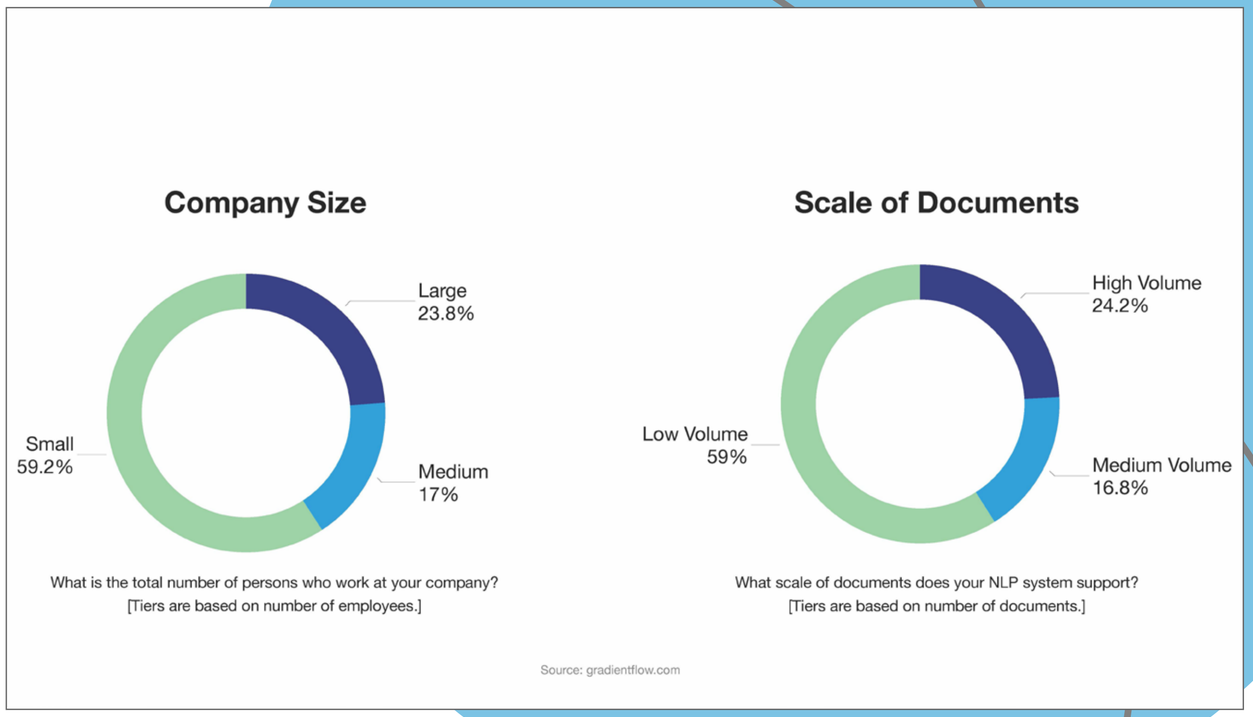
We group respondents based their response to a question that attempts to measure their stage of adoption of NLP technologies:

- **Using NLP**: respondents who have deployed NLP to production (44%)
- **Exploring NLP**: respondents who have not yet deployed NLP to production (56%)

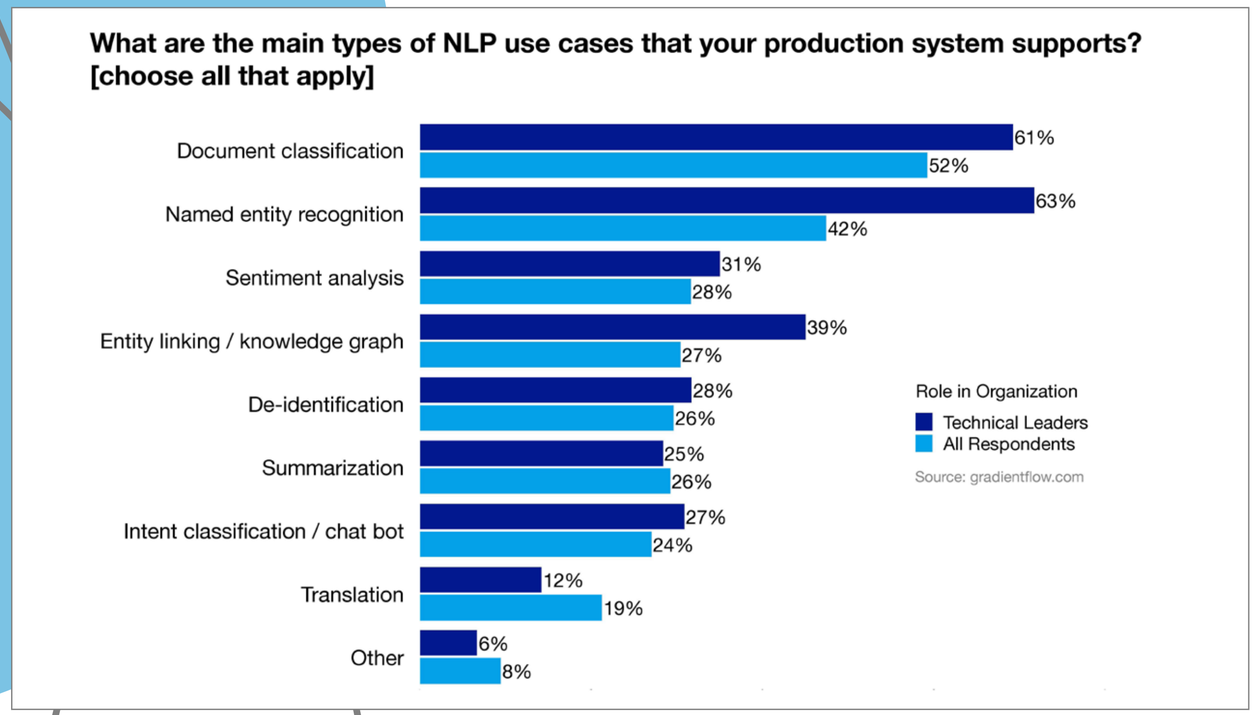
Company Size and Volume of Documents

We also group respondents using two additional variables:

- **Company Size** refers to the number of employees in a company (Small = 500 or fewer employees; Medium = 501 to 5,000 employees; Large = more than 5,000 employees).
- **Scale of Documents** refers to the number of documents an NLP system supports (Low = fewer than 50,000 documents monthly or overall; Medium = between 50,000 to 500,000 documents per month; High = more than 500,000 documents per month).



GRADIENT FLOW

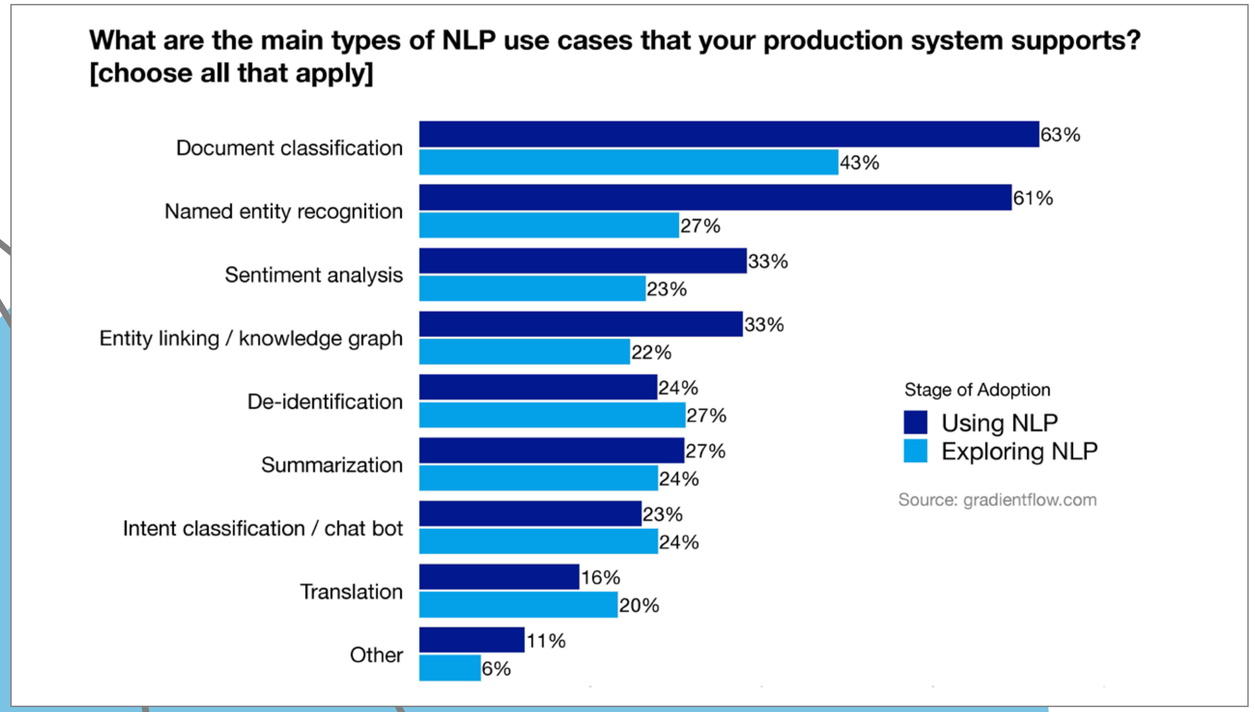


Use Cases

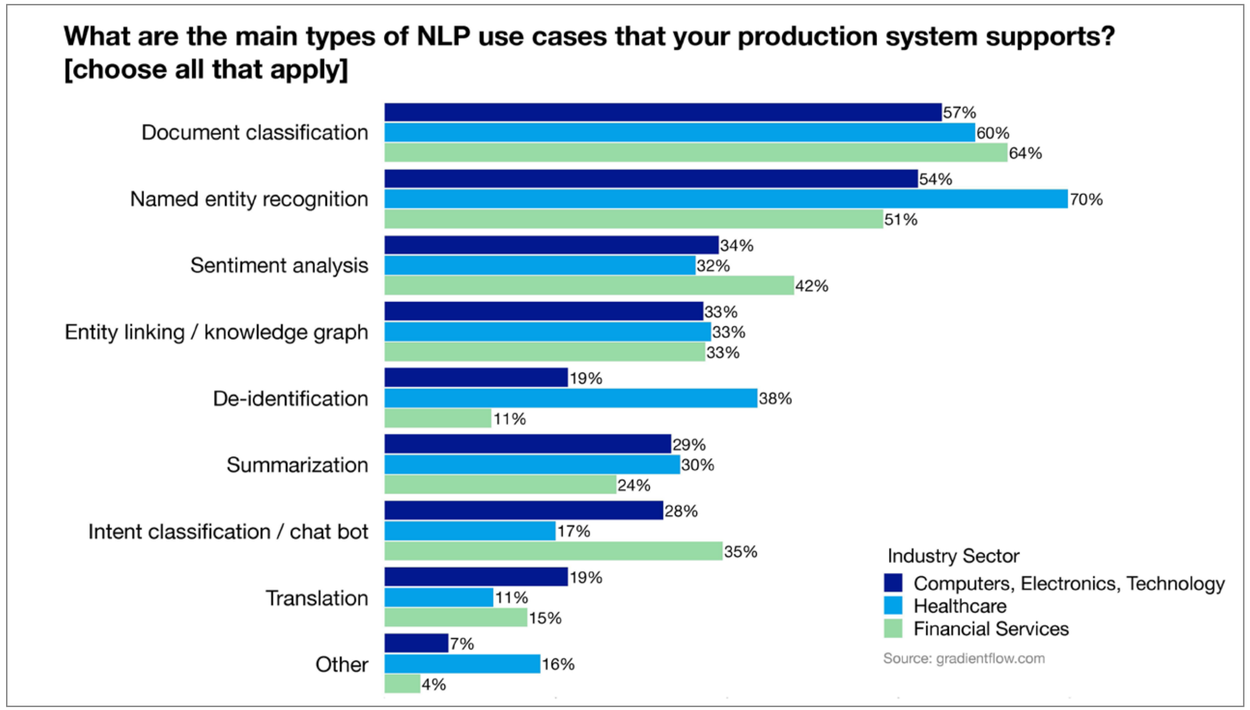
Given the relative mix of Healthcare and Finance, it's no surprise that document classification, named entity recognition (NER), and sentiment analysis topped the list of use cases cited by all respondents. NER models seek to automatically extract named entities (for example, "company name" or "location") from unstructured text. More than a third (39%) of all Technical Leaders stated they also use NLP for entity linking and knowledge graphs.

All of these use cases are classic applications of NLP. Since late 2017, breakthroughs in the use of deep learning for natural language have raised the level of accuracy for predicting the next word or character in a text sequence, as well as training leverage based on transfer learning. You may read about these approaches described as embedded language models, transformers, and Sesame Street (ELMo, BERT, ERNIE). Use cases such as translation, NER, and summarization have undergone significant advances, while others such as question answering and link prediction have now become more practical. We expect to see the latter deployed in industry applications more frequently.

Respondents who work at companies that already use NLP in production ("Using NLP") signaled that document classification and NER were by far the use cases they are most likely to have. The chart to the right displays the share of respondents within a given group: 63% of respondents who work at companies Using NLP indicated they are using NLP for document classification, while 27% of respondents at companies that are Exploring NLP are using it for NER.



GRADIENT FLOW

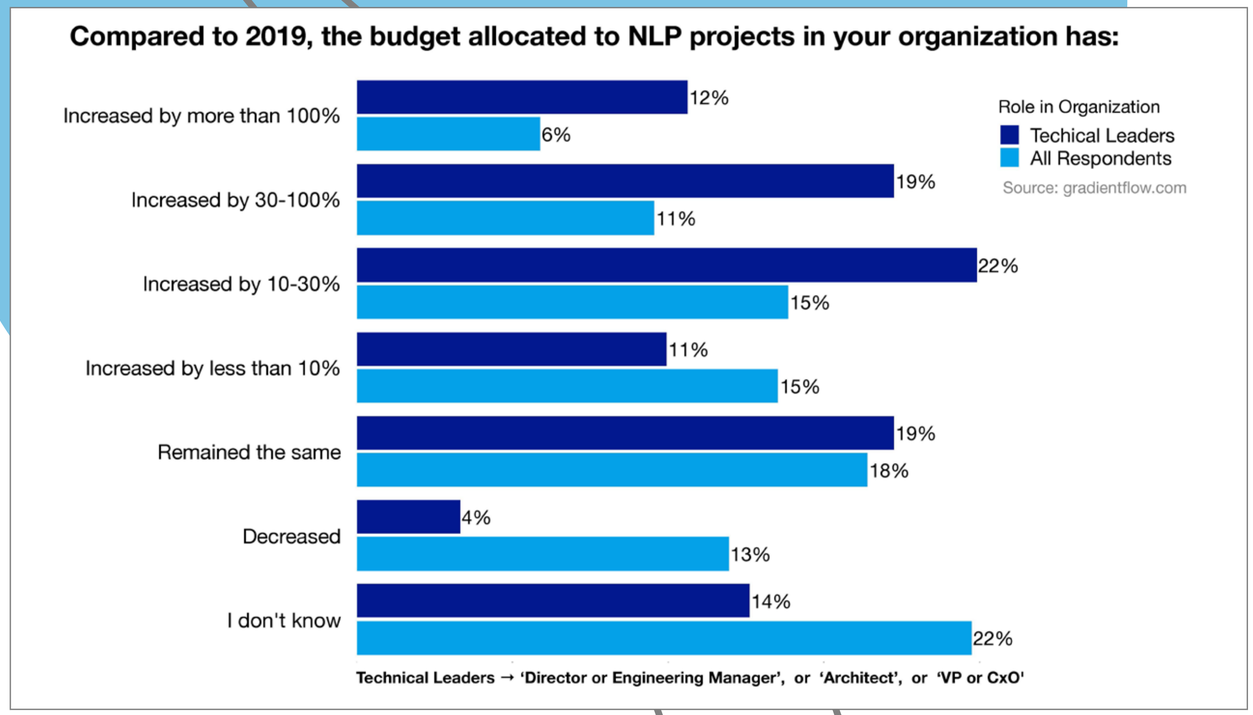


NER and de-identification rated higher among respondents in Healthcare, when compared to respondents in the Technology sector or Financial Services: 38% of all respondents who work in Healthcare stated their companies use NLP for de-identification. Privacy regulations in healthcare require that users strip (medical) records of any protected health information (PHI)—this process is known as de-identification. De-identification has largely been a manual and labor-intensive process, but in recent years, NLP software is being used to partially automate this task.

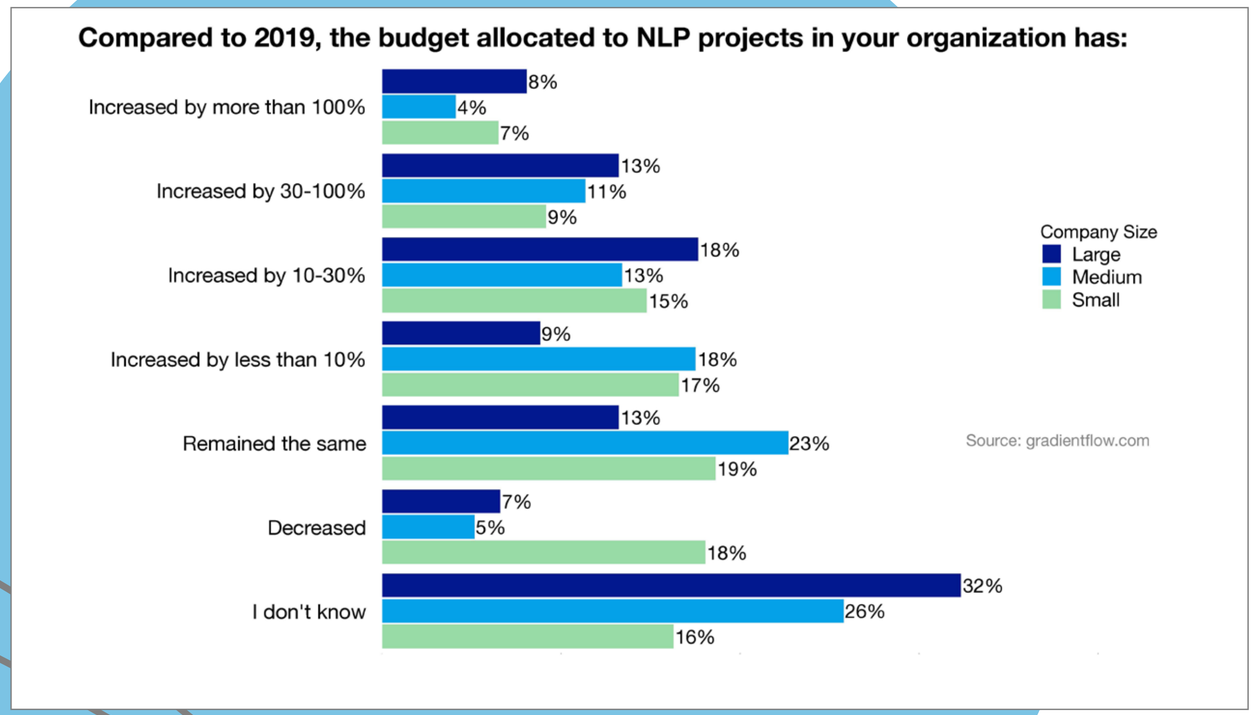
NLP Budgets

This survey took place July 5 through August 14, 2020, a period when the global pandemic disrupted IT budgets across the globe. To gauge the level of spending on NLP, we wanted to understand budgets allocated to NLP from the perspective of people who have visibility into IT spending within their organizations. Recall that *Technical Leaders* are respondents who described their role to be "Director or Engineering Manager" or "Architect" or "VP or CxO."

- 53% of respondents who are Technical Leaders stated their NLP budget was at least 10% higher compared to 2019.
- 31% of respondents who are Technical Leaders stated their NLP budget was at least 30% higher compared to 2019.
- 12% of respondents who are Technical Leaders stated their NLP budget has more than doubled compared to 2019.



GRADIENT FLOW



Respondents who worked at *large companies* (5,000 or more employees) also hinted that their budgets will grow significantly:

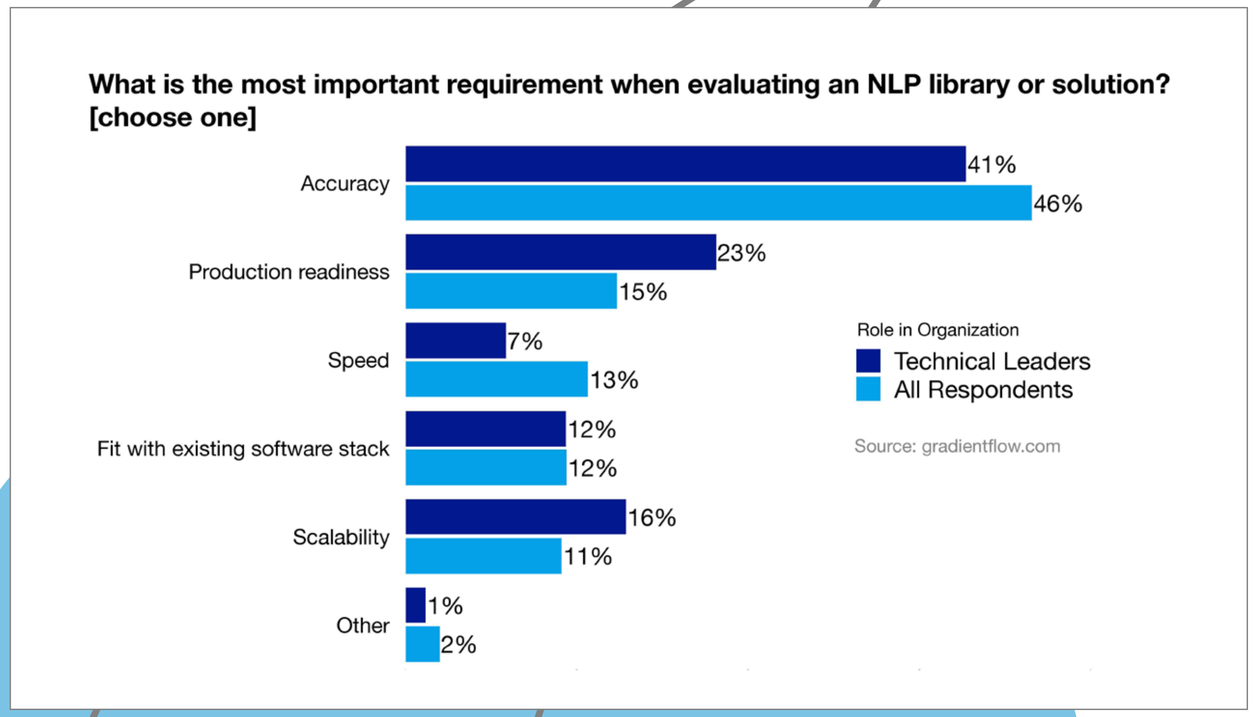
- 39% of respondents who worked at large companies stated their NLP budget was at least 10% higher compared to 2019.
- 21% of respondents who worked at large companies stated their NLP budget was at least 30% higher compared to 2019.

NLP Libraries

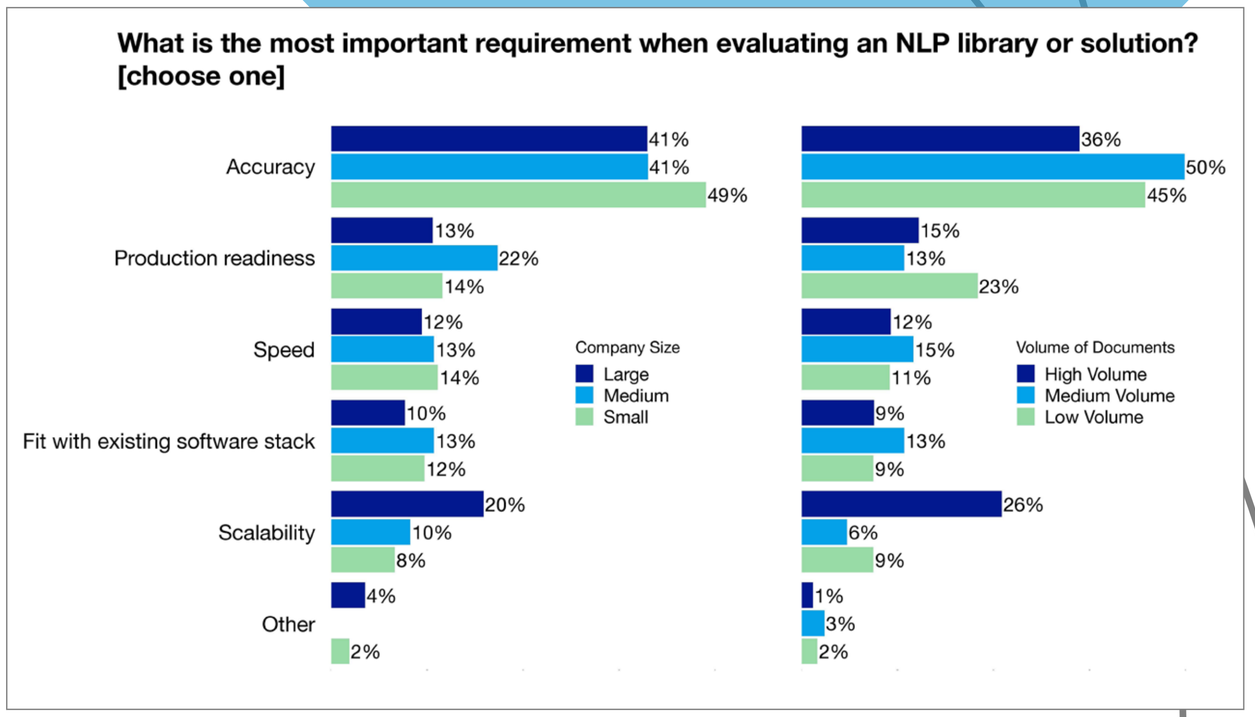
With many more libraries and models to choose from compared to five years ago, this is truly a great time to be an NLP user. What criteria do respondents use to gauge NLP tools? When evaluating an NLP library, survey respondents cite *accuracy* as the most important requirement. Accuracy refers to the effectiveness of models (typically pre-trained models) that now come with many NLP libraries. These models allow users to input text into a pipeline then get common *outputs* (e.g., tokens, lemmas, part-of-speech (POS), similarity, and entity recognition).

When reviewing these libraries, it's important to understand them as multi-stage pipelines, where a cascade of models gets used across each stage of processing. Therefore, there is no single metric for accuracy, but instead a cumulative measure for any given NLP application based on how these models get applied. According to our survey results, users are increasingly comparing libraries based on the accuracy of such pre-trained models.

Technical Leaders also put a premium on production readiness and scalability: about a quarter (23%) of all Technical Leaders cited production readiness as the most important criteria.



GRADIENT FLOW



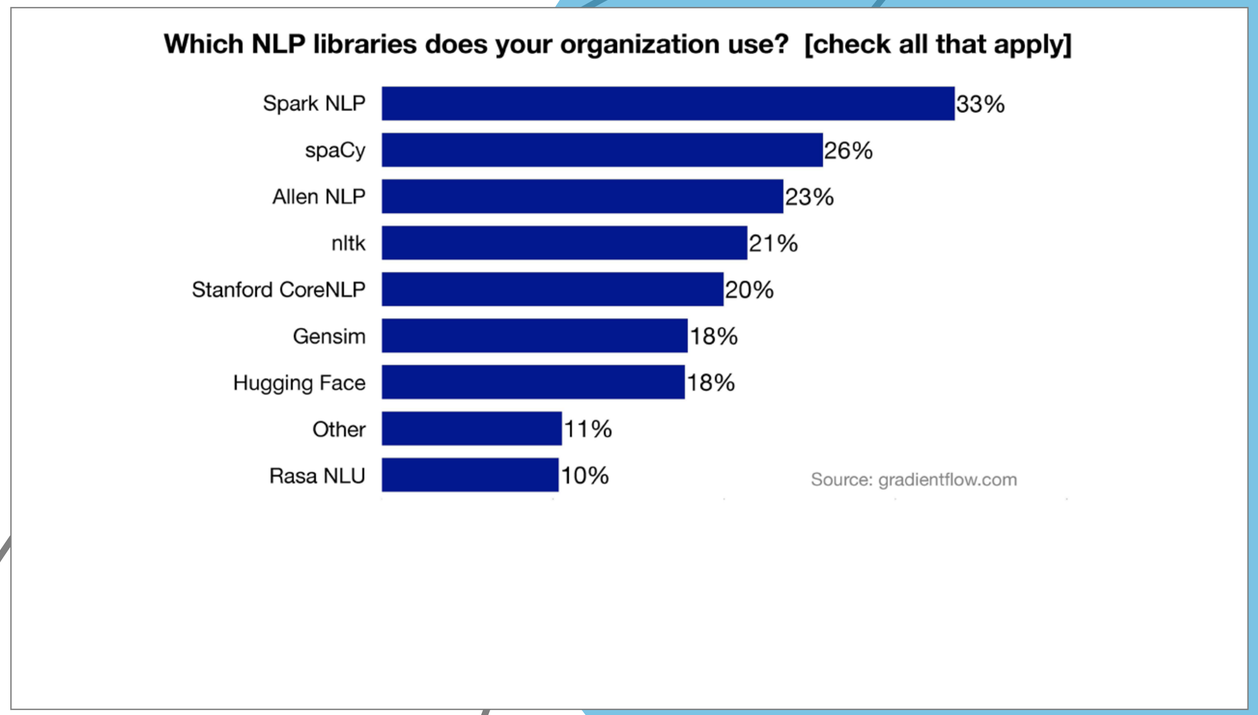
Not surprisingly, scalability is an important requirement for respondents who work at organizations that process a large number of documents each month: a quarter (26%) of all respondents who work at companies that process a Large Volume of documents each month chose scalability as the most important requirement; a fifth (20%) of all respondents who work at Large Companies chose scalability as the most important requirement.

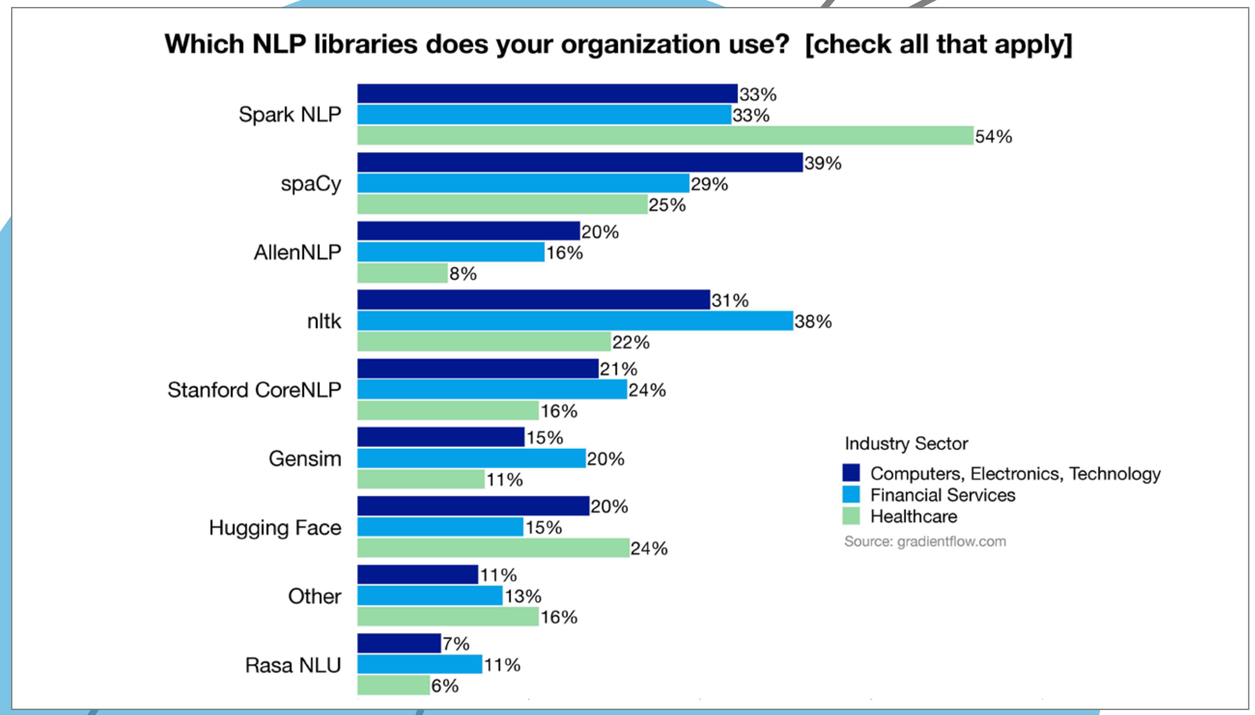
Note that scalability in NLP often differs from other areas of big data applications: typically a document must be processed as a unit of work, in order to help preserve context. In contrast, applications such as log file analysis can generally split their input, distribute, then aggregate quantitative results. In practice, this constraint tends to add complexity to the scalability of NLP applications.

Next, we turn to the NLP libraries survey respondents are using. A third of respondents (33%) indicated they use **Spark NLP**, an open source library build on top of Apache Spark, which provides Python, Java, and Scala API's. About a quarter (26%) indicated they use **spaCy**, an open source NLP library that has become one of the most popular NLP libraries in the Python ecosystem. Over half of all respondents (53%) used *at least one of these two libraries*. We have conducted surveys on *data science and machine learning tools*, and these two libraries also finished high in those earlier surveys.

Rounding out the top five most-used libraries in our survey:

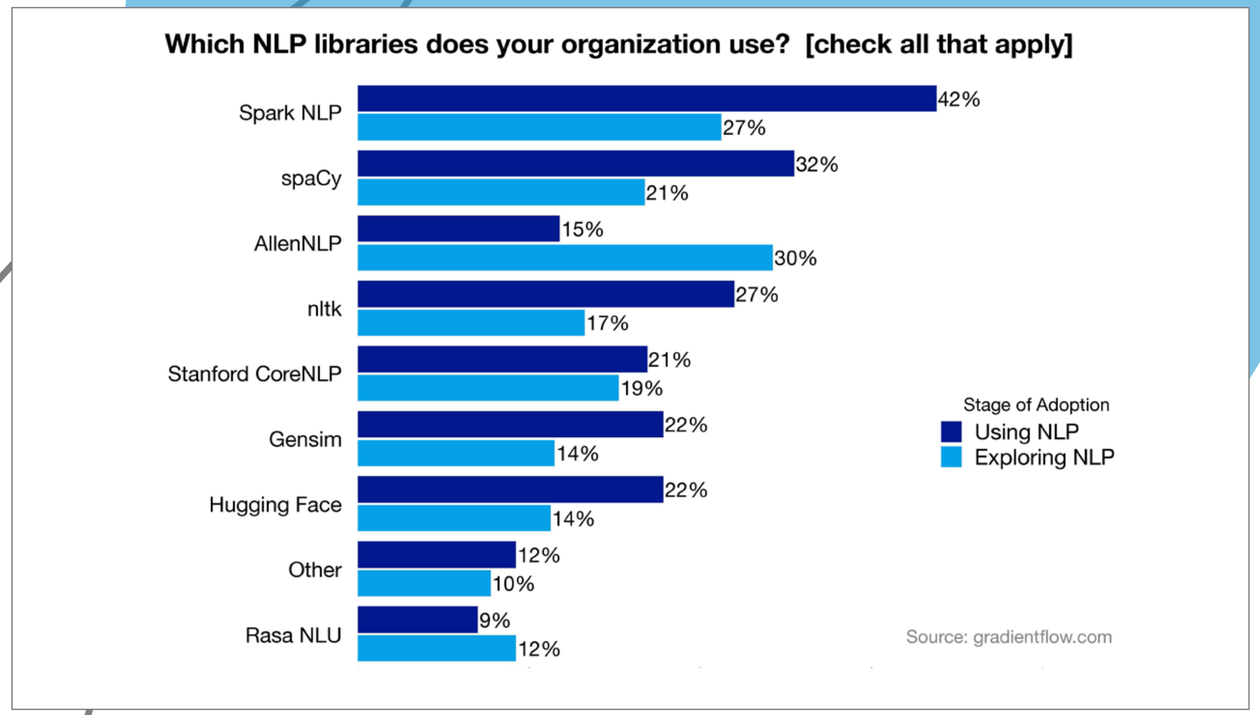
- **AllenNLP** is a newer PyTorch-based library for NLP research.
- First released in 2001, **nltk** is one of the oldest Python NLP libraries.
- **Stanford CoreNLP** is written in Java, but has a Python API.



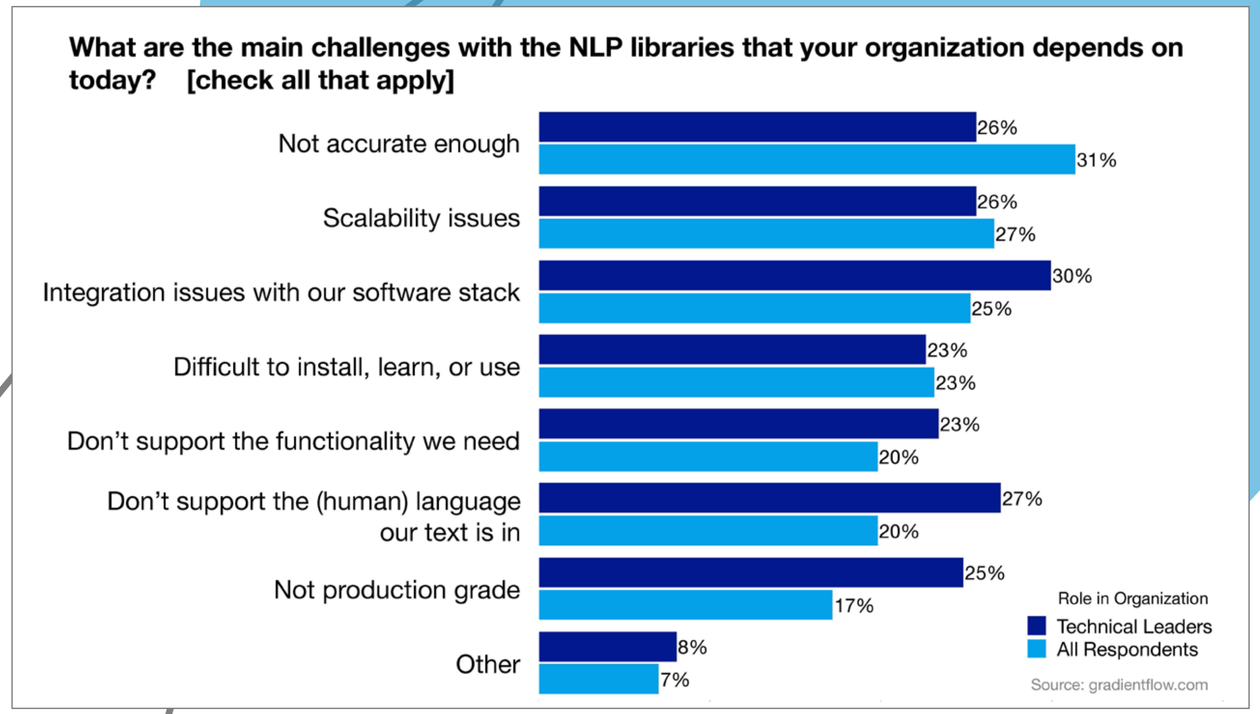


The creators of Spark NLP (John Snow Labs) have a commercial product aimed at the healthcare sector, so it's no surprise that over half of all respondents (54%) from Healthcare signaled that they use Spark NLP. Anecdotally, many people we come across in the technology sector use spaCy, something that is borne out by our survey. Since spaCy is a library and not a framework, its integration costs are often lower, and it fits well into engineering practices that appeal to software engineers. Financial Services have long used NLP and text analytics, and many have embedded/legacy applications, where nltk tends to be prevalent since it's been around much longer than the other libraries on our list of options.

AllenNLP targets researchers, and our survey confirms that it's more popular among respondents who are still in the preliminary stages of adoption.

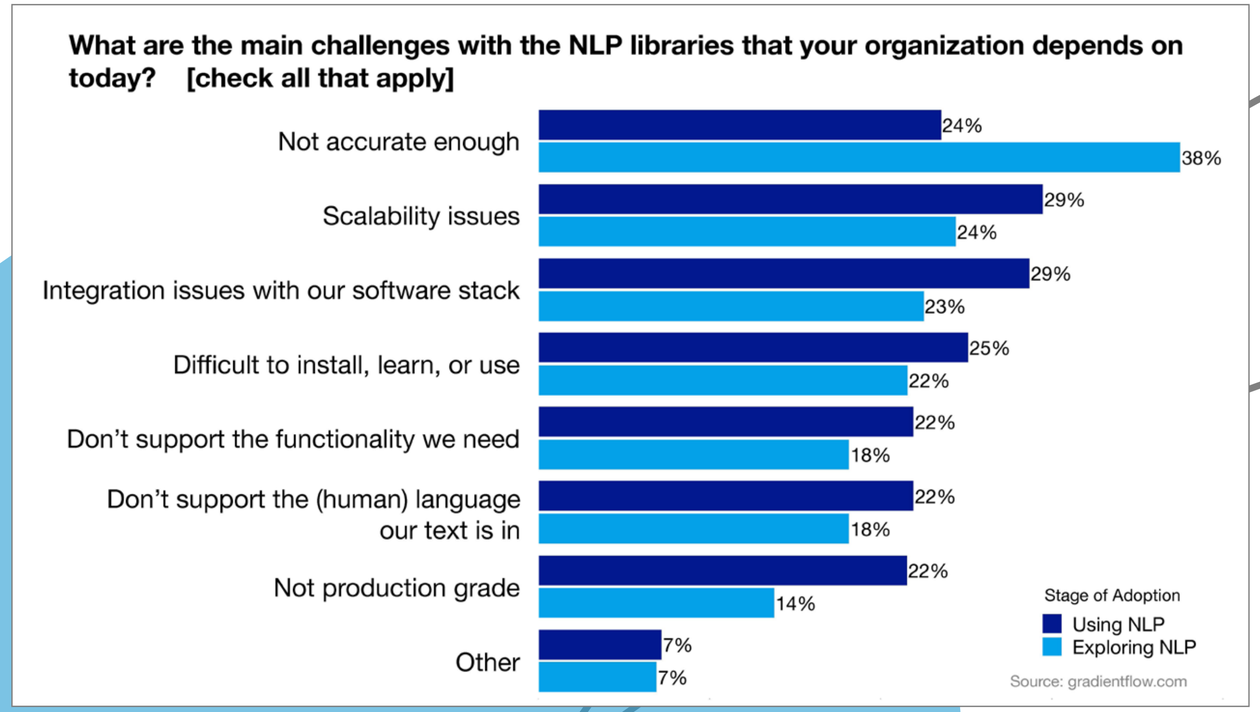


GRADIENT FLOW

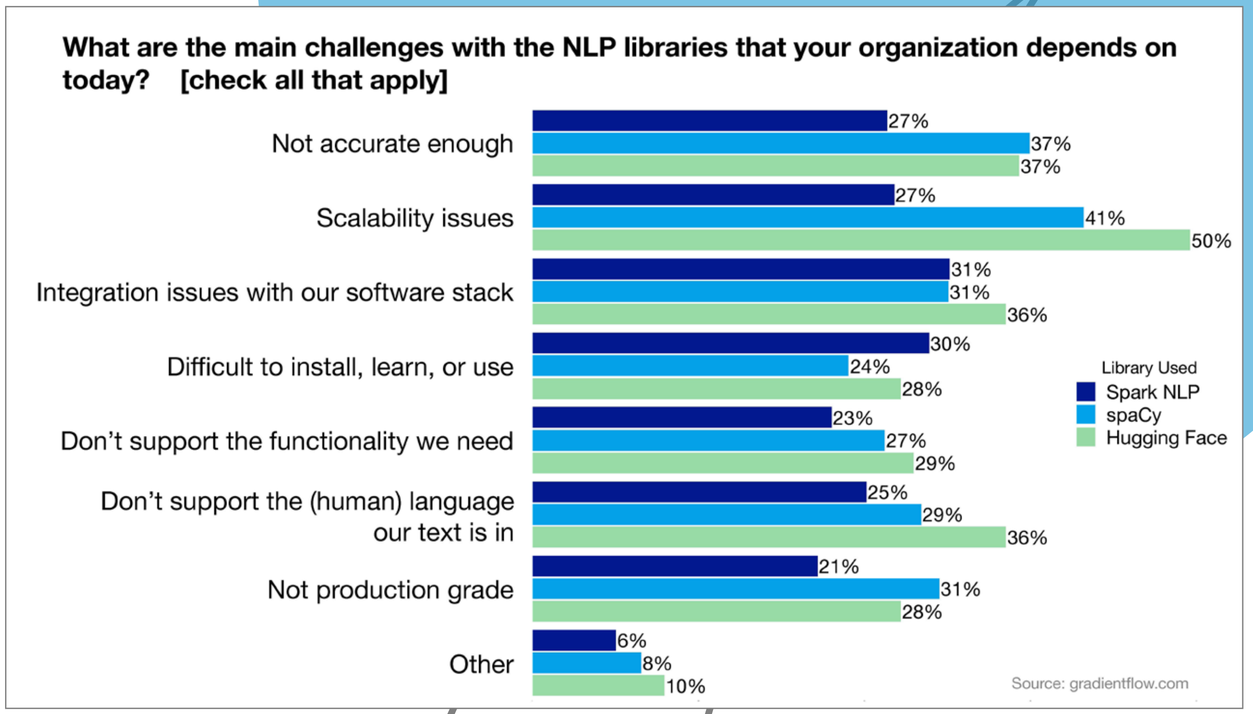


Once they select and begin using a library, what challenges do our respondents face? Accuracy is once again the most frequent item cited by *all respondents*. However, when we look at the subset of respondents who are Technical Leaders, integration issues, language support, and scalability are right there with accuracy as far as pressing challenges are concerned. The number of languages supported varies across NLP libraries (e.g., Stanford CoreNLP lists six, Spark NLP ships with models in 46 languages).

Accuracy is a particularly important issue for our respondents who work at organizations that are still Exploring NLP.



GRADIENT FLOW

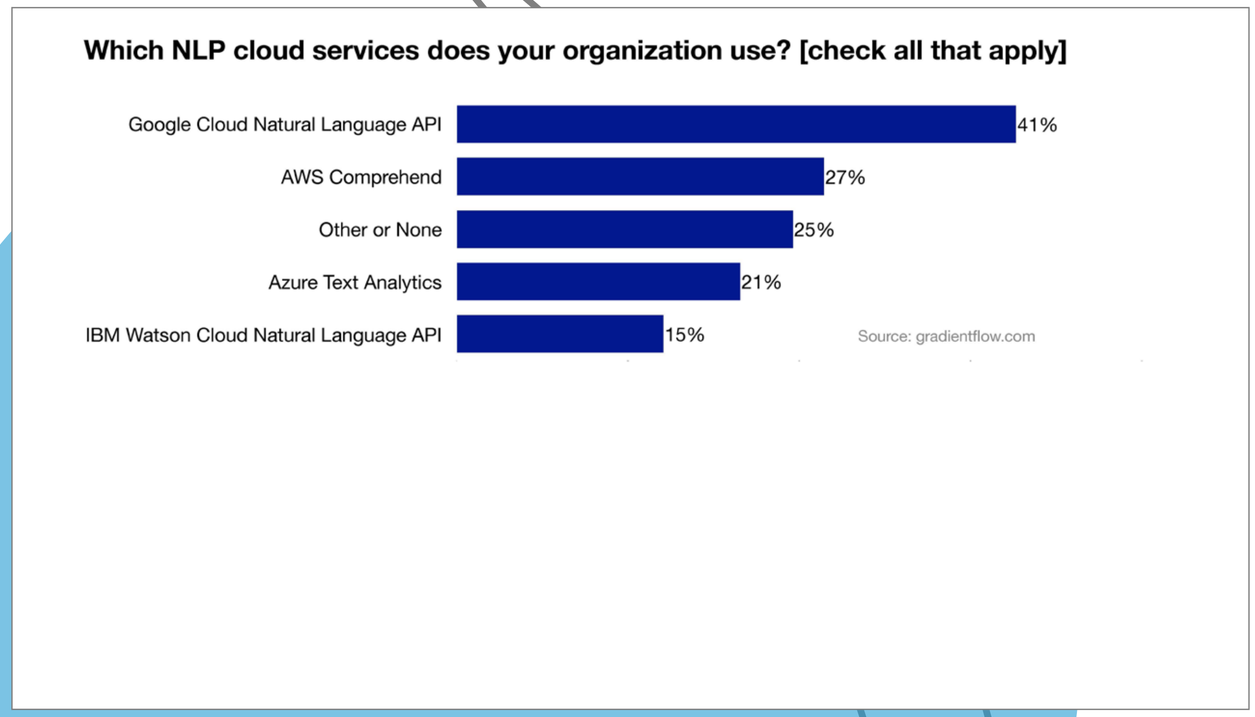


We also looked at the challenges cited by users of popular open source libraries. Spark NLP does better on scalability and accuracy compared to spaCy and Hugging Face, but less so when it comes to ease of use. Overall, the concerns about language support, integration, and lack of functionality can be addressed by the *extensibility* of a given library—e.g., how readily developers can extend or add special kinds of processing into NLP pipelines for a given application.

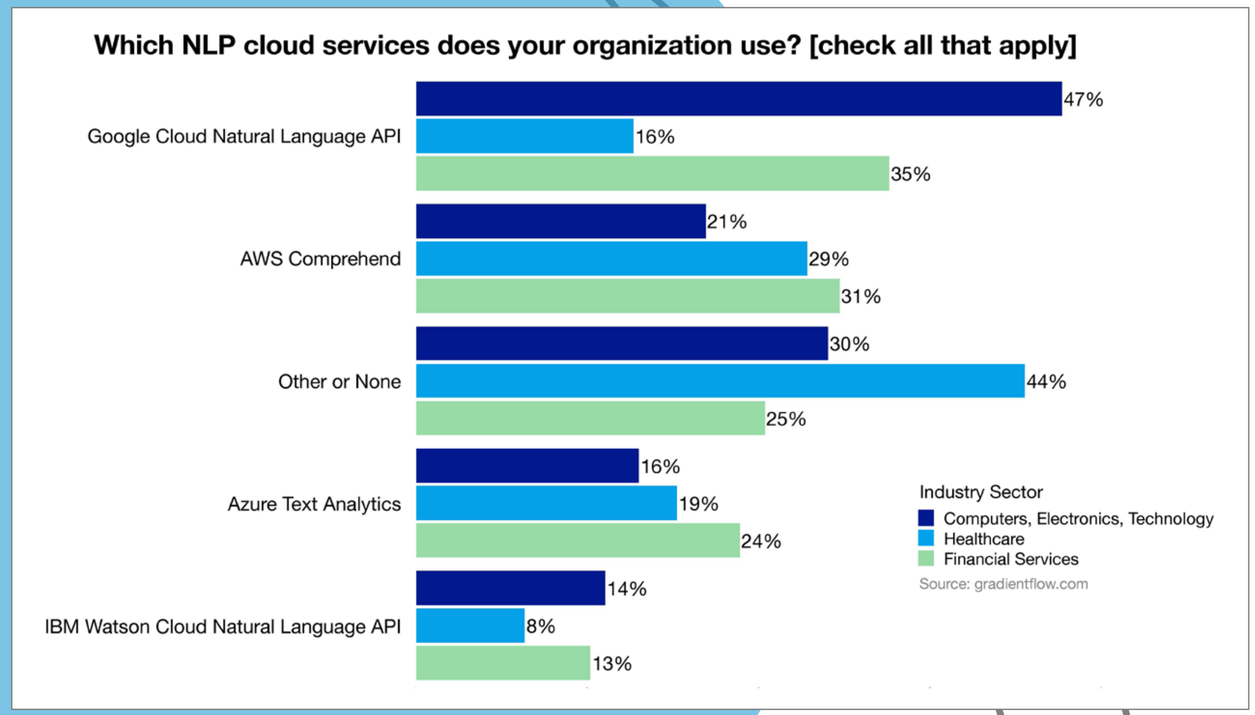
NLP Cloud Services

With the proliferation of NLP services in the cloud, companies need not install and manage NLP libraries like the ones we listed in the previous section. 77% of all survey respondents indicated that they use *at least one* of the NLP cloud services we listed (Google, AWS, Azure, IBM), with Google's service garnering the most users. The share of users drops to about two-thirds when we look at the following targeted segments:

- When we restrict to respondents who are Technical Leaders, 64% of them indicated that they use *at least one* of the NLP cloud services we listed (Google, AWS, Azure, IBM).
- When we restrict to respondents who work at companies that are Using NLP (companies further along the NLP adoption curve), 65% of them stated they use *at least one* of the NLP cloud services we listed (Google, AWS, Azure, IBM).

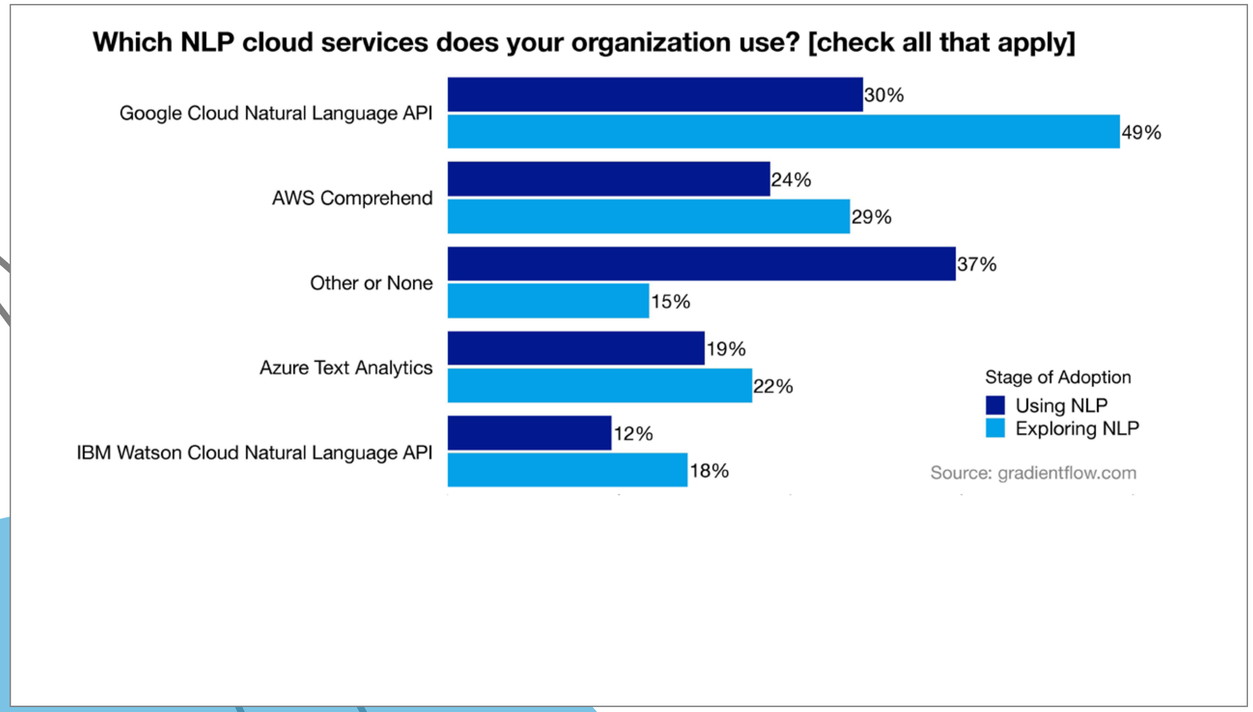


GRADIENT FLOW

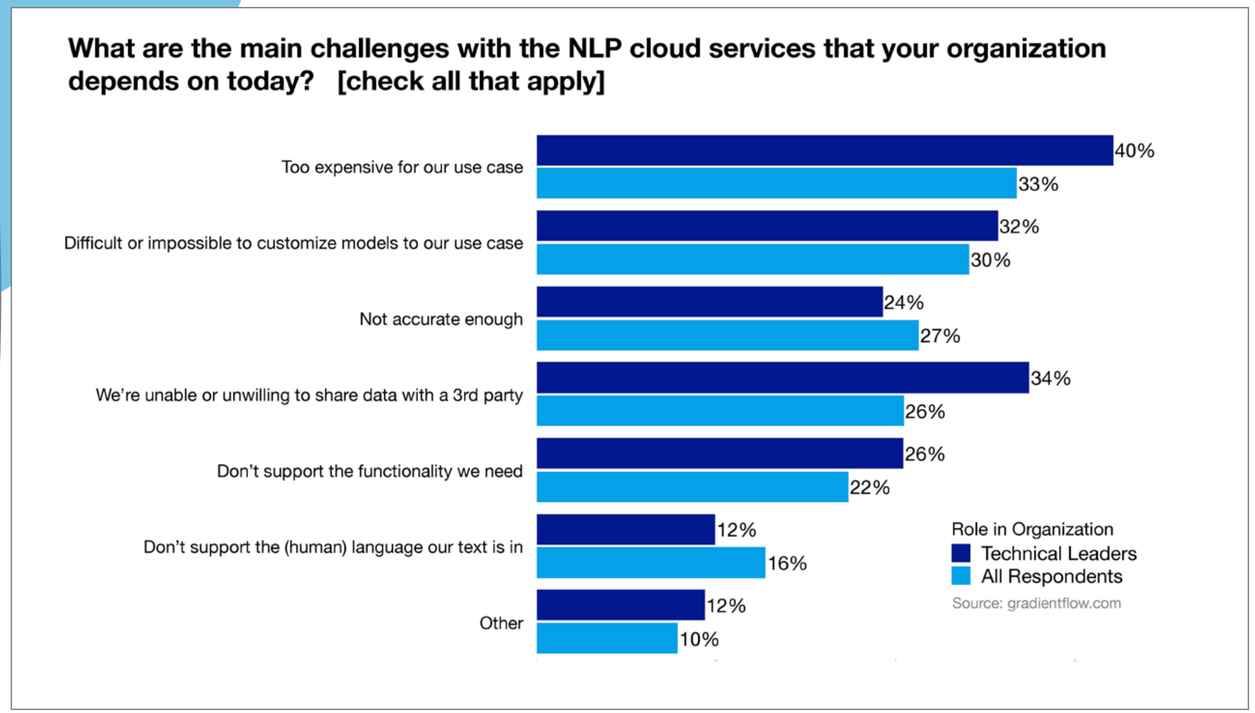


Close to half (47%) of all respondents who work in the technology sector indicated that their company used Google Cloud Natural Language API. Over a third (35%) of all respondents who work in Financial Services indicated that they, too, use Google Cloud Natural Language API.

Google Cloud Natural Language API also appears to be the favorite SaaS for those who are still Exploring NLP. About half (49%) of all respondents who worked in companies that are still Exploring NLP indicated they are using Google's cloud NLP tools.



GRADIENT FLOW



Cloud NLP services are not without their challenges, however. About a third (34%) of all Technical Leaders cited data privacy and security (*"We're unable or unwilling to share data with a 3rd party"*) as a key challenge. Another challenge cited by about a third of all Technical Leaders (32%) is the difficulty customizing models.

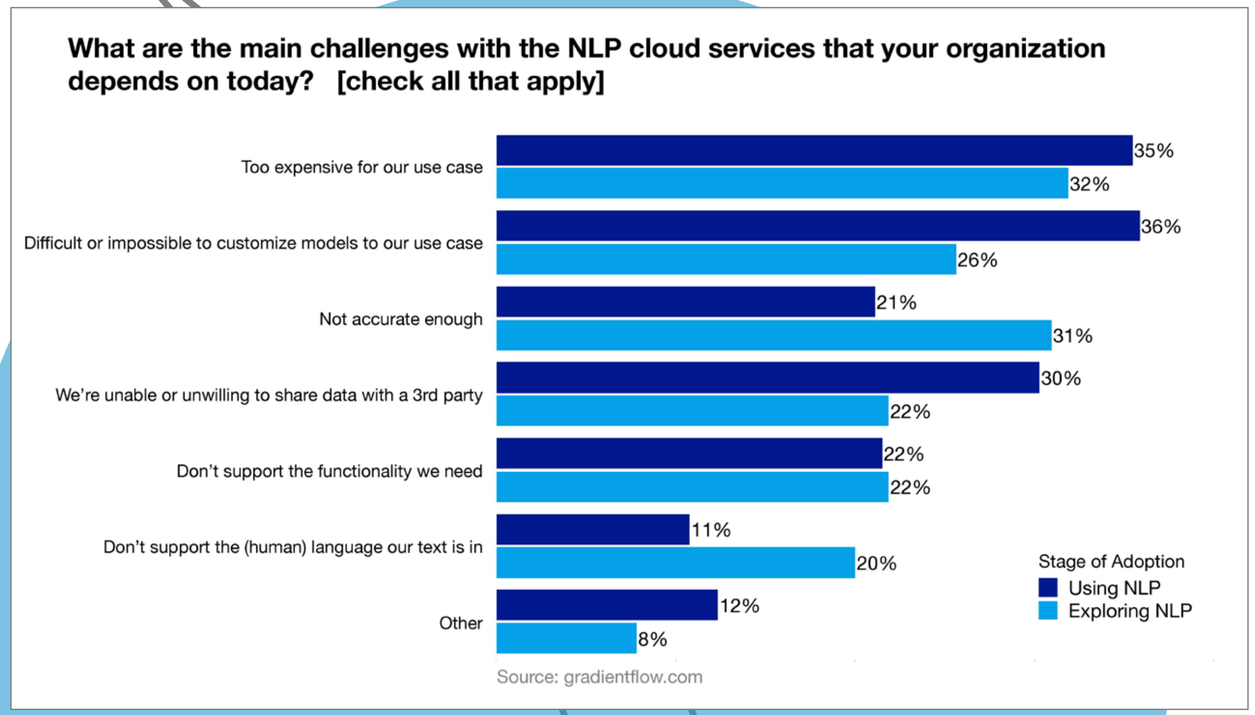
Language is very application- and domain-specific, so models often must be tuned and customized. This can be especially painful when a cloud-based service is trained for general uses of words, but does not understand how to recognize or disambiguate terms-of-art for a specific domain. For example, speech-to-text services for video transcripts from a DevOps conference might identify the word "doctor" for nearly every instance of "Docker," which degrades the accuracy of SaaS-based solutions.

The leading reason cited by all respondents (and by Technical Leaders) is the *cost* of using NLP cloud services. These services get expensive as your collection of documents grows, thus it's perfectly understandable why cost ranks high among the challenges cited by users of these services. Here's how a few of these services describe their pricing:

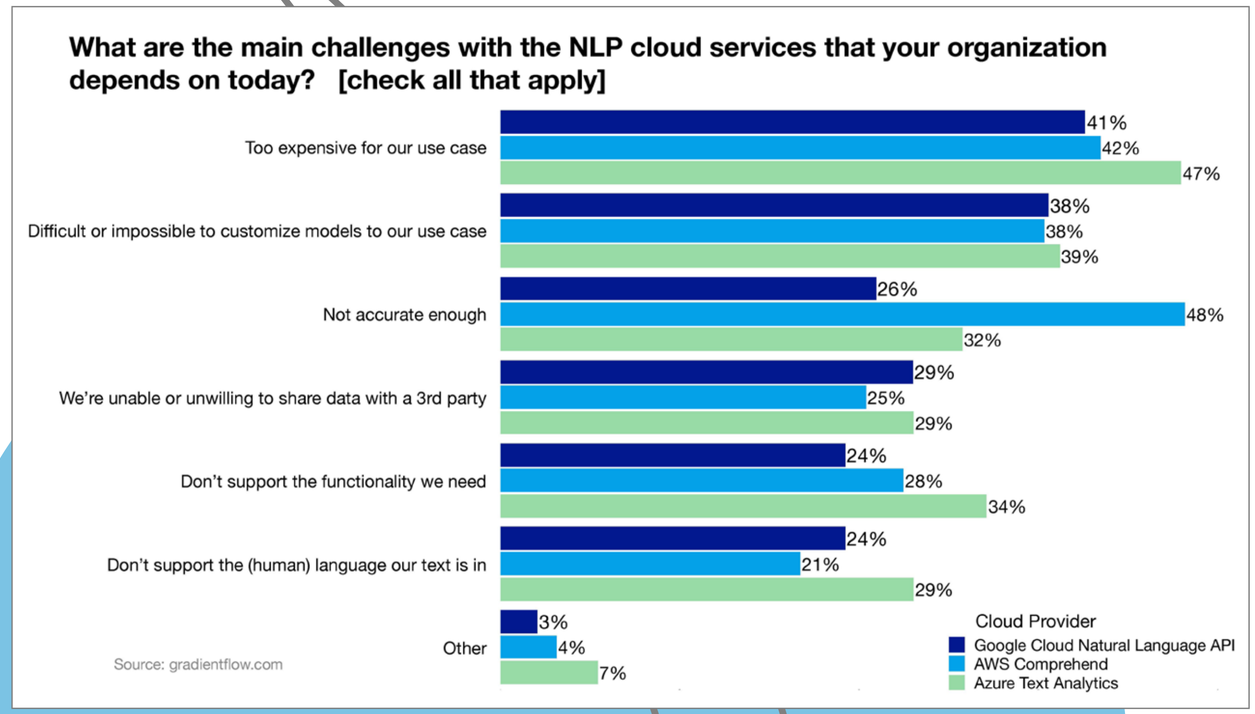
- Google: "Your usage of the Natural Language is calculated in terms of "units," where each document sent to the API for analysis is at least one unit."
- Amazon: "Amazon Comprehend requests for Entity Recognition, Sentiment Analysis, Syntax Analysis, Key Phrase Extraction, and Language Detection are measured in units of 100 characters, with a 3 unit (300 character) minimum charge per request."
- Azure: "Text Records correspond to the number of 1,000-character units within a document that is provided as input to a Text Analytics API request."

Those who are Using NLP face challenges that are somewhat symmetric to those still Exploring NLP: assuming that you addressed accuracy challenges earlier on in your adoption cycle, customizing models becomes more critical as you use a Cloud NLP service in production.

- 36% of all respondents Using NLP cite customizing models as a key challenge, compared to 26% of those Exploring NLP.
- 21% of all respondents Using NLP cite accuracy as a key challenge, compared to 31% of those Exploring NLP.



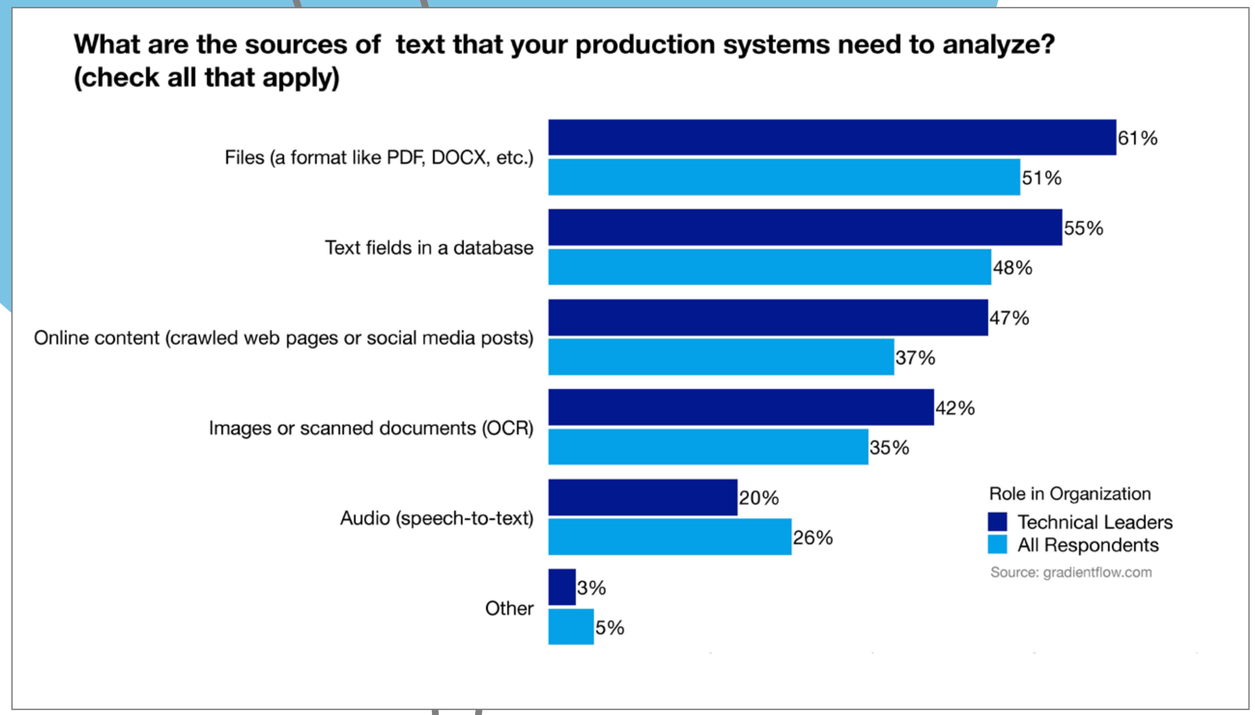
GRADIENT FLOW



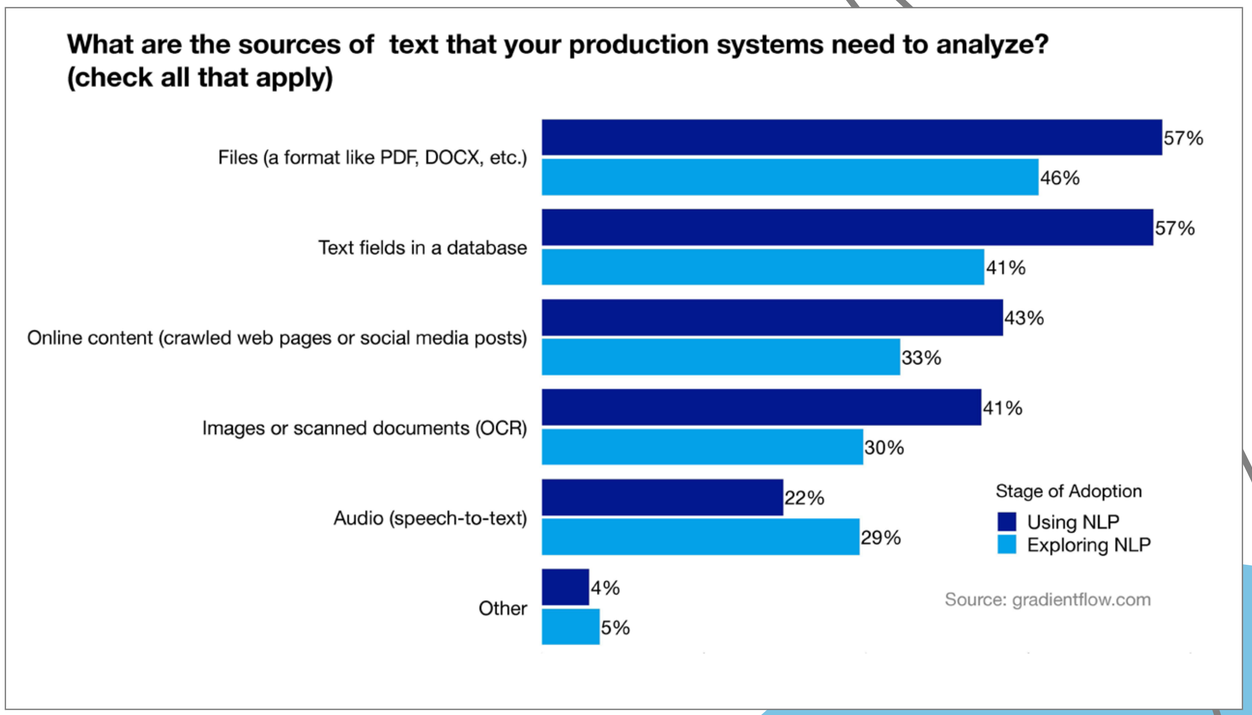
As with the case of NLP libraries, about a quarter of Technical Leaders (24%) and all respondents (27%) cited accuracy as a key challenge facing users of NLP cloud services. About half (48%) of respondents who use AWS Comprehend cited accuracy as one of their main challenges.

Data for NLP

Finally, we examine data resources used in NLP projects. Data from files and databases top the list of data sources used in NLP projects: 61% of all Technical Leaders stated that Files (e.g., documents) were a source of text for their NLP systems.



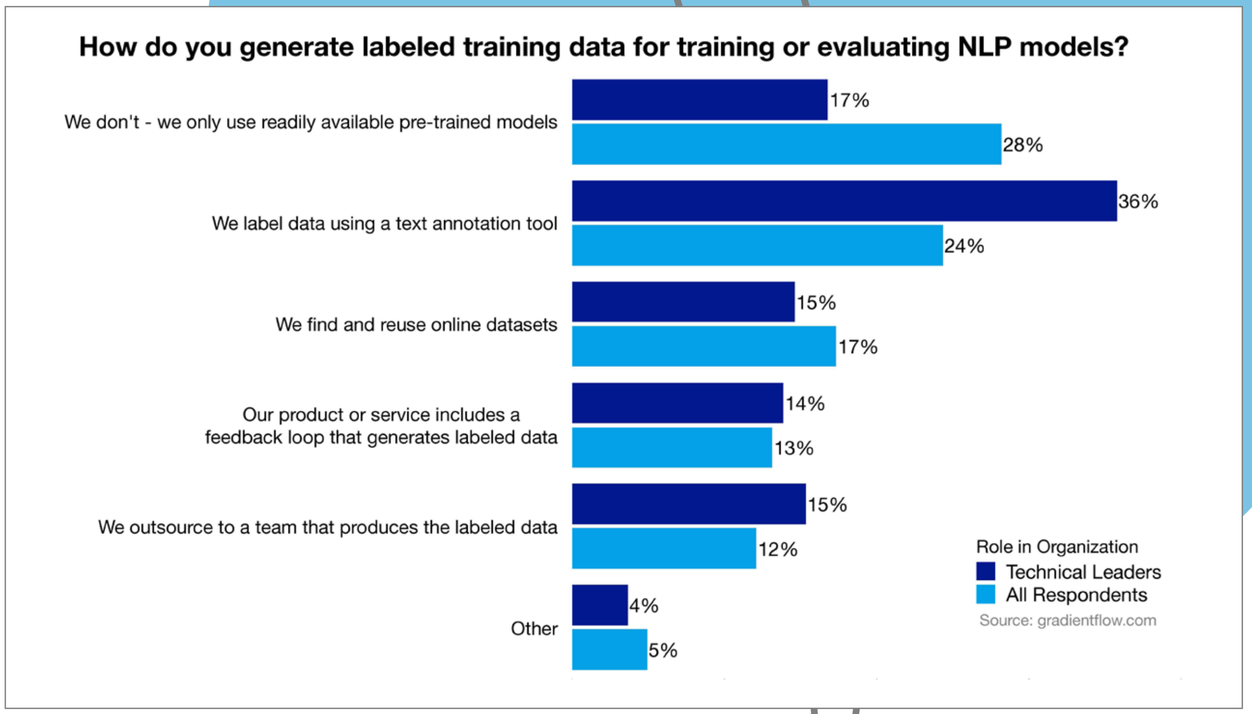
GRADIENT FLOW

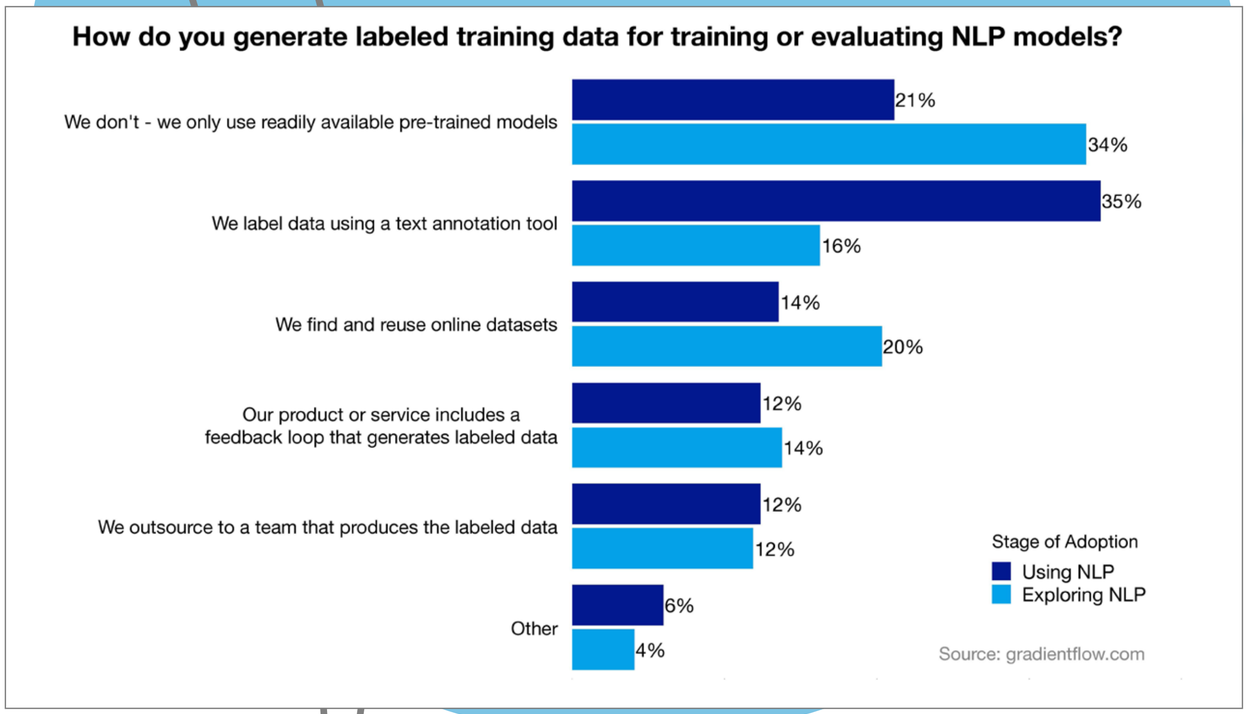


Note that input documents are often stored as PDFs: contracts, articles, reports, filings, and medical records. However, there are many challenges and data-quality issues inherent with attempts to extract text from PDFs, and while deep learning models have made advances over the past two years, these issues persist. Often, it becomes significantly more cost-effective to scan a PDF document and then apply OCR to extract its text, prior to use of any NLP library.

Also of interest: those who are still Exploring NLP use audio data at a higher rate (29% vs. 22%) compared to respondents at companies that were further along the adoption curve.

We close by examining how respondents obtain labeled data. Data labeling is an important step in many machine learning tasks, and text applications are no exception. Looking at our entire pool of respondents, about a quarter (28%) stated they relied on pre-trained models and thus had no need to generate labeled data. But more than a third (36%) of all Technical Leaders stated their organization used a text annotation tool. Note that labeling and annotation of text intersects closely with *human-in-the-loop* approaches to enhance machine learning models in production, which will likely continue to rise in industry.





We get a similar pattern when we examine responses to this question based on the stage of adoption. A third (34%) of respondents at companies that are still Exploring NLP stated they used pre-trained models and, thus, did not need to generate labeled data. In comparison, more than a third (35%) of all respondents at companies already Using NLP stated their organization used a text annotation tool.

Closing Thoughts

Overall, an approximate formula for NLP adoption emerges: the breadth of industry applications grows as the level of accuracy increases, especially for use cases that must understand or generate text. Growth in business applications is largely due to how NLP enhances automation and scale, enabling the more complex use cases to become cost-effective.

While on the surface we talk about text-based data sources as being “unstructured” or “semi-structured,” in practice, the semantics that are embedded in text are often highly structured. This is especially the case in the more formal uses of text, including legal contracts, financial disclosures, sales reports, patent applications, policy hearings, and scientific articles. As advances in NLP lead to AI applications that can leverage those embedded meanings within text, and, in turn, generate text summaries that are more readily “consumed” by both people and machines, we will likely see much broader industry adoption of NLP.

To date, deep learning applied in NLP has been focused on predicting the next word or character in a text sequence—with ongoing improvements in overall accuracy. Related techniques are proving to be powerful in other ways—e.g., for predicting links in a graph or missing relationships between objects in a database. These additional uses of advanced NLP open a much broader range of applications in business areas such as legal practices, supply chain management, accelerated development of pharmaceuticals, material science, and even in AI used for developing public policy. For a good summary of NLP use cases on the immediate horizon and leaderboards for their respective solutions, see NLP Progress.

We should also note that the state of NLP in industry poses somewhat of a dilemma for the cloud providers. On one hand, the cloud providers tend to have among the largest AI teams and top R&D talent, they have business units which acquire highly valuable, large amounts of labeled data, plus they own the compute resources required to train very large machine learning models for NLP, which often have billions of

(Closing Thoughts cont.)

parameters. It makes sense to expose these enormous, high-performance NLP capabilities via SaaS offerings. On the other hand, so many of the NLP applications in business depend on domain-specific uses of language. The cloud providers have been slow to respond to market needs for customizing solutions and extensibility, and moreover, have pricing strategies that tend to work against other organizations' integration efforts. The net result is that cloud-based NLP services tend to be perceived as high in cost and low in accuracy. Given these conditions, plus a long-term trend toward the commodification of deep learning, open source libraries for NLP have enjoyed adoption based on their relative ease of use and extensibility per application, which result in better overall cost-effectiveness. Even so, tension persists between the power and dominance of tech giants, which tend to favor one-size-fits-all service offerings, and the effectiveness of open source libraries that address many specific business opportunities. Watch this space.

Acknowledgements

Thanks to John Snow Labs for sponsoring this inaugural survey, and thanks to Kathy Yu and Jenn Webb for providing critical assistance. This survey was conducted by **Gradient Flow**; see our statement of editorial independence.