April 2021

Four Ways

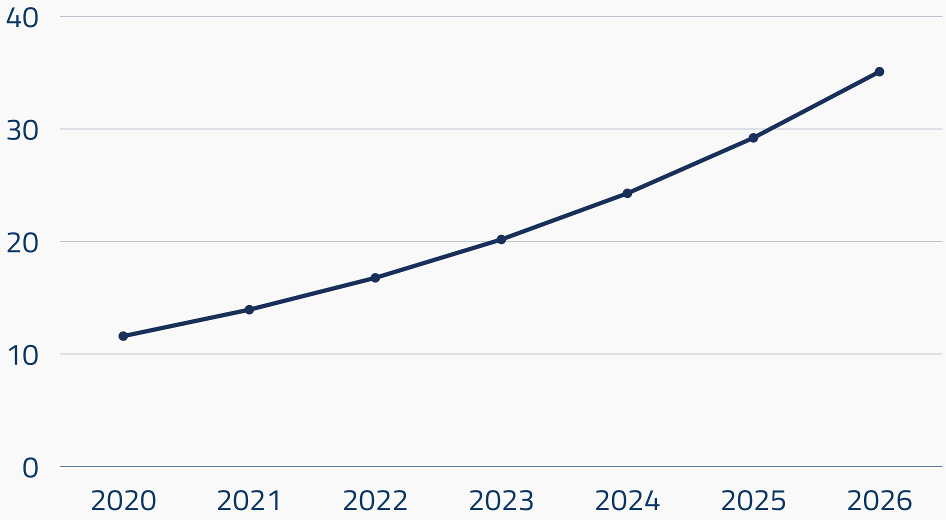
Technical Leaders Are Structuring Text To Drive Data Transformations

We structure the world's knowledge

Intro

Natural and unstructured language is how humans largely communicate. For this reason, it's often the format of organizations' most detailed and meaningful feedback and market intelligence.

NLP Market Size (USD Billions)



Historically impractical to parse at scale, natural language processing has hit mainstream adoption. The global NLP market is expected to grow 20% annually through 2026. Analysts expect that -- for data teams of all sizes -- if you aren't presently working on a game plan for NLP, you will be in the next five years.

As a benchmark-topping natural language processing API provider, Diffbot is in a unique position to survey cutting-edge NLP uses. In this paper, we'll work through the state of open source, cloud-based, and custom NLP solutions in 2021, and lay out four ways in which technical leaders are structuring text to drive data transformations.

Overview

In common use cases, NLP is used to:

- Mine user reviews
- Identify and track entities and relationships in the news
- Structure medical, scientific, or legal documents
- Serve contextual ads based on an understanding of the surrounding text
- Market research
- Categorize support tickets
- Summarize documents
- Among other uses

One of the primary choices for tech leaders looking to implement NLP solutions includes discerning what level of service that meets their requirements.

In the landscape of NLP today, we have three basic tiers of service, each with distinct trade-offs. In particular, organizations are choosing between the following:

- Building out an open source NLP software
- Licensing a “basic” commercially-available NLP API
- Leveraging a custom NLP solution

In this guide we will work you through the trade offs of these methods, as well as provide real world examples of how we’re seeing organizations build off of “basic” and custom NLP solutions in game-changing ways.

Building Out An Open Source NLP Software

There are a plethora of open source NLP projects. Many of these projects provide some collection of the following functions needed to build up to more complex natural language tasks.

Language identification is important for routing unstructured text to the proper structuring functions

Tokenization is the process of breaking a document, sentence, or paragraph into smaller segments for analysis. These segments are called tokens

Identification of parts of speech provides the building blocks for extracting entities, relationships, and facts

Chunking builds off of identification of parts of speech to assemble phrases that provide the bedrock for understanding how tokens relate to each other

Machine learning expertise may well be the largest expense if choosing to build your own solution. Machine learning models are necessary for building any of the above basic NLP functions into more advanced solutions that come standard in most cloud-based natural language processing services.

For example, for new document types, topics, languages, and even entity recognition all require machine learning models and the expertise to hone your accuracy over time. Unless your NLP needs are limited or won't be changing over time, building off of open source NLP solutions often takes the form of "reinventing the wheel"

Common Features Of Cloud-Based Natural Language Processing API Services

Cloud-based NLP services let you rely on organizations who have already heavily invested in building out a robust NLP functionality. To some degree these options will work "right out of the box" but are still likely trained and benchmarked on a specific unstructured text type.

As opposed to the most basic building blocks of NLP provided by open source projects, cloud-based natural language processing APIs tend to provide some collection of the following features:

Named entity recognition is the process of locating and classifying entities that are mentioned in unstructured text. By classification, what is typically meant is that entities are placed into parent categories like “person names,” “location,” “product,” and “stock ticker,” among others.

Sentiment analysis provides a sentiment score either for the entire unstructured document or individual entities within the document (depending on service provider). Scores range from -1 (very negative) to 1 (very positive).

Salience analysis provides a view into how central an entity is to understanding a given unstructured document. For example, one mention of IBM in a book about Apple would likely hold very low salience.

Analysis of themes bundles entities into broader themes to provide absolute sentiment values or relative sentiment ranks. While

Annotation or redaction of sensitive content is a feature of some NLP tools centered around health records, accounting data, or more. This feature may also be trained in a custom build of cloud-based NLP services.

Topic categorization involves providing an overall category for a natural language document.

Intention extraction involves applying pretrained models to natural language documents to return buyer intent data. Though not a common feature of many NLP services, some providers focus exclusively on intention extraction.

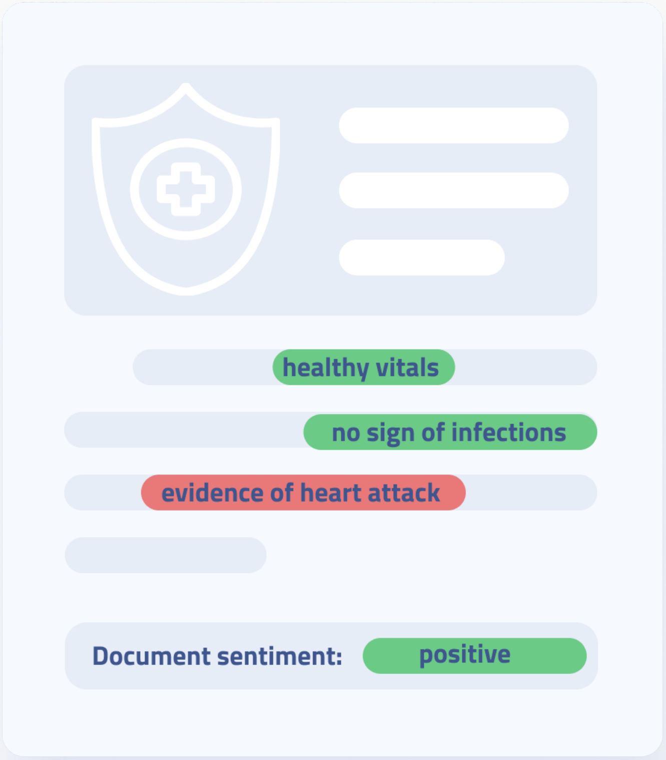
Summarization is provided by some NLP providers as a way to prep large natural language texts for quicker human interpretation.

A subset of NLP API services are also "generative" in the sense that they can provide you with some form of "creative" output. This is most commonly seen in services that translate, or mimic the writing style of natural language document input.

Basic NLP Can Be Misleading

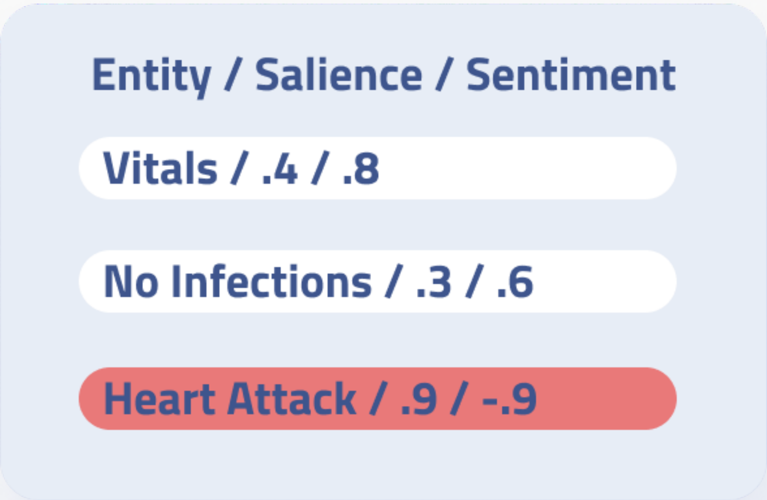
Basic text analytics and NLP can be achieved through out-of-the-box solutions as well as rolling your own NLP tools from open source projects. But at the end of the day, robust testing should occur to understand if an NLP solution that isn't customised for your particular use is providing verifiably accurate insights.

Take document-level sentiment, which is a standard offering in many "out of the box" NLP solutions. Given particularly informal or specialized language, document-level sentiment doesn’t know how to weight the range of positive and negative statements within a document.



Document-level sentiment can be woefully misleading

In contrast, entity-level sentiment, and particularly more advanced NLP services that can provide properties (relationship "lines" on a graph) can provide the granularity you need for your information systems to perceive the whole meaning of rich unstructed natural language.



Entity / Salience / Sentiment

Vitals / .4 / .8

No Infections / .3 / .6

Heart Attack / .9 / -.9

Salience -- how central an entity is to understanding a document -- and entity sentiment provide clarity

Additionally, even NLP services with named entity recognition training sets as large as the entire public web can have trouble identifying particularly technical or specialized language entities. Custom NLP that tracks entities, relationships, salience, and sentiment around topics from your own domain can remedy this.

Leave Highly-Tuned NLP To The Pros

> *In our tests, the Diffbot Natural Language API outperformed IBM Watson and DBpedia Spotlight by over 30%.*
> *Ron Snyder, Director of R&D, JSTOR Labs*

Customizing your NLP solution can lead to powerful benefits including entity-level sentiment, more accurate results, and the recognition of domain-specific entities and properties of entities. But there's a large trade off. Just as it takes substantial resources to build from open source to something like a "basic" cloud-based NLP solution, each customization of an NLP product will require custom machine learning models. To build your own customization you're likely looking at needing a machine learning engineer, a data scientist as well as the cost to gather, clean, and annotate training data for models.

Running with a cloud-based NL API provider who also offers customization should at the very least be considered unless your organization already holds substantial machine learning and data expertise.

Even in the event you do end up customizing your NL solution, you'll need to sort out robust structured training data for every language and new feature you would like to support. Diffbot is one of only three North American organizations to crawl the entire web. And the only one of these three to provide subscription access to the structured data that is returned.This makes Diffbot uniquely positioned to provide massive natural language corpora for training in almost any domain (and in any language).

Additionally, we've built in the ability to quickly train custom NLP models of up to 1M domain-specific entities in as little as a single day. We'll explain the importance of being able to rapidly train (and re-train) natural language models in the following section. But for now we should simply note that for most organizations building out training data on their own is orders of magnitude more time and money consuming.

The Bleeding-Edge: Entity Sentiment, KG Construction, And Custom Property Training

Use One: A KG For An Entire Nation

A tricky aspect of NLP is that performant models need training data that has already been processed to some extent. That's where Diffbot's Natural Language API comes into play for researchers at the University of Alberta. Their mission is to create a central knowledge graph for all things related to Canadian culture, scholarly works, and literary works.

In the past, the University of Alberta has utilized Diffbot's web data extraction APIs to find structure data at scale. In this project they extracted over 240 million quotes from news articles across the web. Our NLP-aided extraction can be used almost like a "re-tweet count" for the entire web. And in this case comes with additional contextual data like the speaker of the quote, topical tags, and information about what publication the quote was presented in.

Today these same researchers utilize Diffbot's Natural Language API to structure unstructured natural language for their own NLP models. This is a hybrid approach in terms of "build, buy, or tune" that we discussed above. And we see similar uses regularly in which Diffbot web extraction data (or our Knowledge Graph) is used to customize our own natural language service.

Use Two: Entity Sentiment For Ad Placement

> *If you could measure my sentiment, it would be 0.99999, so I am delighted with this.*
> *Stefano Costella, Data and Finance Manager at Dianomi*

Dianomi is the largest native ad network in finance with over 350 publishers. They use Diffbot to monitor topics being discussed in the pages of their network to find the best pages for each ad. Dianomi uses Diffbot's sentiment analysis to make sure their ad placements are brand safe.

We routinely see use of our natural language API for entity level sentiment analysis. As opposed to most competing products, which tend to supply document-level sentiment, Diffbot's Natural Language API provides a reading of which entities are most central to understanding a given text, as well as what the sentiment of that entity is in and of itself. This allows for sentiment tracking over time as well as being able to filter out details that may be highly negative or or positive but aren’t central to understanding a given text.

In our own analysis, we’ve been able to use this functionality to determine where diners feel most safe eating out during the Covid-19 pandemic by parsing review data with our NL API.

Similarly, financial services and VC firms have found uses for monitoring sentiment related to products or organizations in their portfolios. Paired with large volumes of news data such as Diffbot’s Knowledge Graph (50x the size of Google News), or Dianomi’s network, our NL API provides ways to derive entity-level insight previously not possible.

Use Three: Custom Properties For Fraud Detection

Thus far we’ve touched on entity identification, salience, and sentiment. But a great deal of the value in unstructured data is found in identifying modifiers of or relationships between entities.

The crucial step of building out custom properties enables organizations to hand off an additional analysis step to their information systems. In particular, custom properties can be trained that sift through complex or convoluted relationships detailed in natural language documents. At scale, this can be incredibly valuable and allows for custom market intelligence or news monitoring uses that can be applied to many sites at once.

Some of the most impactful uses of our NLP with customization include uses in which unnamed organizations monitor news and rely on our NL API to infer relationships between entities in the form of “Org A defrauded Org B” or “Org C stole Entity Z from Org D.” The savings from this sort of automation becomes even more drastic when considering the commensurate amount of paid specialist hours it would cost to cover as many technical, legal, health-related, or regulatory documents manually.

Use Four: 1M Custom Entities In A Day

As an NLP solution becomes more performant, results often surface that hint at how the solution could be improved even more for a given use case. Patterns emerge that push for analysis of new entity or property types. Or results are validated so NLP workflows can be expanded. At each of these junctures the need for large training data sets and the ability to quickly retrain a model are pressing.

Access to our research and customer solutions teams at Diffbot have aided numerous clients as they tune their NLP solutions even more over time. While our tech is always evolving, presently we can train up to a million custom entity types within roughly a calendar day. This is a function of having a structured feed of a majority of the web through our Knowledge Graph, performant custom web scraping tools, and a rapidly retrainable NL API.

Recap

As a benchmark-topping NL API provider and one of only a handful of entities to crawl the entire web for machine learning data, Diffbot is a unique position to advise on the best ways to set up your own NLP workflows.

While there are examples of building from open source, licensing an NLP platform, or buying a custom NLP solution that work, there are distinct trade offs for each route.

"Rolling Your Own" From Open Source

+ build internal expertise. Can process NL documents on premises. Wide range of open source NLP libraries for basic processing tasks.

− substantial costs and talent needs. With work leads to the equivalent of a commercially available basic NLP solution. Basic non-customized NLP can be misleading.

Licensing an API or NLP Service

Offered by Diffbot

+ piggyback on organizations who have already invested heavily in machine learning and NLP. Much faster set up time than rolling your own.

− doesn't help build your orgs expertise. Reliant on provider. Subscription fees or licensing costs. Still not custom trained for your domain of interest.

Buying a Custom NLP Solution

Offered by Diffbot

+ Substantially faster than “rolling your own.” Interface with NLP experts so you can focus on your org goals. Massive savings when compared to manual human analysis at scale.

− Higher costs than licensing an API. Third party reliance. Need to line up substantial training data sets for each customization.

Want to talk about your custom NLP needs?

Reach out at sales@diffbot.com or sign up for a free 14-day trial.