Build or Buy? Text Analytics & NLP

Building from open source versus licensing a cloud API or configuring an NLP platform
Introduction

Adding new natural language processing (NLP) capabilities to a product, application or other analytics infrastructure? Answer this question before anything else:

Should I build my own system from open source components or license an NLP provider’s solution?

For NLP, this “Build or Buy” question means choosing between:

- **Building from open source**
- **Licensing a basic cloud API**
- **Customizing an NLP platform**

Today, you’re faced with dozens of open source NLP toolkits to choose from and scores of text analytics vendors vying for your attention.

This paper will help you make an informed decision by explaining the choice you face, what to expect if you choose to build, and how to select an NLP provider if you decide to integrate an API or license a platform.

If you take nothing else away from this paper, let it be this: Building your own NLP system from open source libraries can be a viable choice but is rarely the best one. Look to license and configure an API or NLP platform based on four factors: your document types, data volumes, analytical needs and other requirements.

For detailed analysis of text documents, such as multi-layered sentiment scoring, complex categorization, or recognition of custom entities, work with a configurable NLP platform. For lower-volume, simple use cases, such as document-level sentiment scoring or simple classification, go with a basic API or other NLP tool.

---

**Building from open source:**
- **System cost:** $0 (open source)
- **Expertise cost:** $81,000+ (hiring someone with NLP skills + other developers)
- **Time:** weeks to months
- **Capabilities:** limited without major additional work

**Licensing an API or NLP platform:**
- **System cost:** $10,000 (basic cloud analysis) to $40,000+ (configurable NLP application)
- **Expertise cost:** None (your vendor will do most of the work)
- **Time:** days to weeks
- **Capabilities:** both deep and broad, custom to your needs

---

**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>What is a Build or Buy Question?</td>
<td>3</td>
</tr>
<tr>
<td>What to Expect if You Choose to Build</td>
<td>4 - 5</td>
</tr>
<tr>
<td>The Limitations of Basic NLP</td>
<td>6 - 7</td>
</tr>
<tr>
<td>The Costs of Building Advanced NLP</td>
<td>8</td>
</tr>
<tr>
<td>Buy and Integrate: How to Choose a Solution</td>
<td>9 - 10</td>
</tr>
<tr>
<td>Big Tech vs. Smaller NLP Providers: Pricing</td>
<td>11</td>
</tr>
<tr>
<td>Summary</td>
<td>12</td>
</tr>
<tr>
<td>About Lexalytics</td>
<td>13</td>
</tr>
</tbody>
</table>
WHAT IS A “BUILD OR BUY” QUESTION?

A build or buy question is a choice of whether to build your own version of something or buy a prefab (“off-the-shelf”) solution from another company. People face build or buy questions every week: Cook dinner or order delivery? Manage my own investments or pay an advisor? Do my own laundry or bring it to a cleaner?

Imagine that you need a new computer. One option is to build your own. You can research what parts to buy, go shopping on Newegg, and then watch YouTube videos about how to put everything together.

This can be a lot of fun. But is building your own computer a good choice for your primary work machine? Probably not. If you have specific requirements that can’t be met by an off-the-shelf computer, it may be worth it to build one yourself using custom components. But this will add time, cost and headaches.

And in the end, when you get to the office, you need to trust that your machine will turn on and run without a fuss.

Another way to think about this is the difference between building or fixing up a car versus buying a new one. Sure, it might be fun to create your own internal combustion engine from scratch. But you have errands to run, and the time you spend fixing your car is time you could be spending getting groceries, taking your kids to daycare, or meeting up with friends.

For NLP, the Buy or Build decision means choosing between:

- Building from open source
- Licensing a basic cloud API
- Customizing an NLP platform
WHAT TO EXPECT IF YOU CHOOSE TO BUILD

Building your own NLP system from open source libraries can be a viable choice but is rarely the best one.

Why? Because even though building a very basic NLP tool is relatively simple, building something that’s actually useful is much, much harder.

Much like a car, any NLP system worth its salt involves a huge number of complex moving parts. When you buy an off-the-shelf solution, most of these are taken care of by the vendor. But when you build a text analytics system from scratch, you’re responsible for all of them.

Here are the 7 basic functions of text analytics:

1. Language Identification
2. Tokenization
3. Sentence Breaking
4. Part of Speech Tagging
5. Chunking
6. Syntax Parsing
7. Sentence Chaining

Each of these serves a vital role in accomplishing larger natural language processing features:

- Sentiment Analysis
- Named Entity Recognition
- Categorization
- Theme Analysis
- Intention Extraction
- Summarization
This chart shows a simplified view of the layers of processing an unstructured text document goes through to be transformed into structured data at Lexalytics.

All that, and machine learning models are only in the footnote. In fact, language detection, Part of Speech tagging, named entity recognition and other functions all require machine learning models to achieve reasonable accuracy. Each model must be trained on a data set consisting of hundreds or thousands of hand-tagged documents.

If you want to analyze more than one language, each and every one will require its own NLP models. And you’ll have to keep updating them as people start using words in new and weird ways.

The point is, even the most fundamental sentiment analysis system involves far more than just telling a computer “great = positive, bad = negative.” And that’s only the tip of the iceberg.

If you’re thinking about building your own system, you need to understand the misleading simplicity and hidden dangers of basic NLP features.
THE LIMITATIONS OF BASIC NLP

Rules-based sentiment analysis and query-driven categorization are pretty simple concepts. If you work from open source libraries while deliberately limiting the scope of your system, you can get a basic sentiment scorer or classifier up and running relatively quickly.

Similarly, three of the Big Tech companies (Google, Amazon, and Microsoft) offer cloud NLP services, including sentiment and topic analysis. If your analytical needs are simple, if you don’t need on-premise processing, and if you don’t have very many documents, these tools can be an efficient and cost-effective choice.

However, both of these approaches are inherently limited. Open source NLP is good for simple use cases. But the cost-benefit analysis comes out against it unless you already have an established data science program. Building up internal NLP capabilities is prohibitively expensive and time-consuming for any but the largest enterprises.

Similarly, the big cloud providers are good at solving lower-volume use cases involving one or two basic NLP features. But these tools offer narrow analytics and limited tuning. When you need more complex analyses or custom configurations, they simply won’t support you.
This is important because basic text analytics comes with a number of drawbacks. For example, document-level sentiment scores can be extremely misleading because they’re taken out of context. Imagine a restaurant review that reads:

“The linguini was great, but the room was way too dark.”

Clearly this review expresses positive sentiment towards the linguini, and negative sentiment towards the room’s ambiance. But a document-level sentiment tool will return the average of these two scores, reporting it as neutral.

This example shows how rules-based document-level sentiment analysis can be dangerous, especially if you’re making important decisions based on the results. For more perspective, read our white paper: Theme Extraction and Context Analysis.

Meanwhile, query-based categorizers search for exactly what you tell them to, not what you mean by those words. This means that the more meanings a word has, the more layers you need to add to your query topic. A query to categorize mentions of “apple” will require pages of Boolean operators. This can quickly become cumbersome, unreliable and unsustainable.

Basic text analytics features like document-level sentiment or simple categorization are useful, but only to a point. When you need deeper insights or custom analysis capabilities, you need advanced NLP features.
is prohibitive complexity, expensive and time-consuming for most companies.

THE COSTS OF BUILDING ADVANCED NLP

Advanced NLP features such as entity-level sentiment, categorization of ambiguous words, and theme analysis are powerful data analysis tools. But building these features requires you to combine NLP rules with custom machine learning models. And this adds considerable cost to your project.

According to Glassdoor, the average salary for a United States-based natural language processing engineer is more than $80,000. A data scientist to train NLP machine learning models will run you well into the six figures in salary, plus benefits and bonuses.

What’s more, you’ll need to gather, clean and annotate data to train models on. A single NLP categorization model, for example, requires at least a few hundred pieces of training content. Language-specific sentiment models or Part of Speech taggers require thousands more. And if you don’t have a good pipeline for processing data and managing the complexity as your models grow over time, you’ll quickly run into problems.

In sum: To build your own NLP system that’s actually capable of delivering deep, useful insights, you’re looking at upwards of $100,000 or more in talent and data costs. And that’s in addition to the thousands of people-hours that your team will have to spend building, tuning and re-training the system.
BUY AND INTEGRATE: HOW TO CHOOSE A SOLUTION

To reiterate: Building your own NLP system from open source libraries can be a viable choice but is rarely the best one.

In most cases, it’s far more cost-effective to license an API or NLP platform. Which approach and what vendor is best for you? Start by understanding your own needs.

Make your choice based on these four factors:

1. **Document types**
2. **Data volumes**
3. **Analytical needs**
4. **Other requirements**

First, consider the **type of documents** you’re dealing with. Simple survey verbatims? Try an open source NLP model or the text analytics capabilities of your survey tool. Unstructured content like customer reviews, free-text surveys, and social media comments? You’ll need a dedicated NLP tool to handle them. Contracts, invoices, emails, medical files or other complex documents? Look for a fully customizable NLP solution supported by an established vendor.

Next, evaluate the **volume of data** you process per month, on average. A few hundred or a couple of thousand documents can easily be handled by an open source model or cloud API. Beyond that, open source runs into performance issues and cloud solutions get very expensive. If you’re handling tens of thousands of documents or more, look for a scalable NLP tool with predictable pricing and stable, scalable architecture.

Thirdly, evaluate your **analytical needs**. Document-level sentiment analysis, simple categorization, or recognition of standard entities can usually be handled by open source models or even an application-specific solution such as a survey tool. When you want to understand why people feel the way they do, define custom entities or sort content into complex buckets, however, you’ll need an NLP platform with tuning and configuration tools.
Finally, list out your other requirements, such as private data storage, on-premise processing, semi-structured data parsing, a high level of support or specific services. Open source NLP models can process documents on-premise but leave you to fend for yourself with training. Cloud analytics providers may offer private storage, but you can’t know where your data actually goes when you call their API.

Meanwhile, the Big Tech companies don’t offer much in the way of services and training – after all, they’re not in the NLP business, they’re in the cloud business. Only dedicated NLP companies like Lexalytics combine all of the technology necessary to meet your requirements with the expertise and know-how needed to help you actually meet your goals.
BIG TECH VS. SMALLER NLP PROVIDERS: PRICING

The NLP services offered by Amazon, Google and Microsoft are an attractive choice for basic text analytics, particularly if you’re already familiar with their ecosystems. Indeed, these tools can be great for addressing simple use cases involving lower data volumes, simpler analytical needs or one-off analyses.

But all three Big Tech NLP offerings have major limitations and drawbacks. One of these is that their pricing models are complicated and get very steep at high data volumes.

Amazon prices their Comprehend NLP services by units of 100 characters at three different tiers, charging extra for custom entities, classification and topic modeling.

Google uses two different pricing structures for their Cloud Natural Language API and AutoML products and counts documents of more than 1,000 characters as multiple units.

Microsoft’s pricing model is more comprehensible but still counts each NLP feature separately and charges high overages if you exceed your monthly document limit.

Worse, all three providers charge separately for each NLP feature. With them, calling the API for both sentiment analysis and categorization on 1,000 documents costs the same as doing just sentiment analysis on 2,000 documents.

Compare this to smaller NLP providers like Lexalytics who offer simple and predictable pricing. We use fixed tiers with set numbers of API calls per month. What’s more, for every document you analyze we return entities, categories, themes, phrases, topics and multilayered sentiment scores, all as one API call.

All of this means that Big Tech can be cheaper for low-volume, simple use cases. But smaller NLP providers like Lexalytics are better for high data volumes and use cases involving multiple NLP features.
SUMMARY

So, should you build or buy for text analytics and NLP?

Building from open source

**PROS:** Free NLP libraries; complete control over the build; on-premise processing; build up internal capabilities

**CONS:** Takes time to build; no support or services; requires expertise; limited analytics without major additional work

Licensing an API or NLP platform

**PROS:** Start analyzing immediately; more analytics capabilities; tune and configure as you need; someone else maintains the core technology; full support and services

**CONS:** licensing costs; need to tune the engine; choosing the wrong provider leads to big headaches down the road

And perhaps most importantly: Choosing to buy, rather than build from scratch, frees you up to focus on actually achieving your goals: improving products and customer experiences, reducing employee turnover, managing regulatory compliance, automating business processes, or something else entirely.
Lexalytics processes BILLIONS of text documents every day, GLOBALLY.

We help our clients unlock the full value of their text data through the Lexalytics Intelligence Platform® and “semi-custom” enterprise applications.

Our solutions combine natural language processing, semi-structured data parsing, and machine learning to reveal context-rich patterns and insights within comments, reviews, contracts, medical files and other complex text documents.

Enterprises and data analysts rely on Lexalytics to improve customer experiences, reduce employee turnover, manage regulatory compliance and more.

For more information, visit www.lexalytics.com or call 1-800-377-8036