NovaceneAI® Platform
User Manual

Novacene AI Corp.
123 Slater Street, Suite 610, Ottawa, ON K1P 5H2, Canada | (800) 717-0814 | www.novacene.ai

Contents

Disclaimers

By using the NovaceneAI™ Platform, you agree to the following disclaimers:

Exclusion of Consequential and Related Damages. IN NO EVENT SHALL NOVACENE OR YOU, HAVE ANY LIABILITY TO THE OTHER PARTY FOR ANY LOST PROFITS OR FOR ANY INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES HOWEVER CAUSED AND, WHETHER IN CONTRACT, TORT, OR UNDER ANY OTHER THEORY OF LIABILITY, WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Exclusions. THE LIMITATIONS OF LIABILITY SHALL NOT APPLY TO DAMAGES ARISING FROM A PARTY'S OBLIGATIONS WITH RESPECT TO (I) THE UPLOADING OF PERSONAL DATA TO THE SERVICES WITHOUT SUFFICIENT CONSENT OR AUTHORIZATION; (II) INFRINGEMENT OF A THIRD PARTY'S INTELLECTUAL PROPERTY RIGHTS; (III) ARISING FROM A PARTY'S GROSS NEGLIGENCE, RECKLESSNESS, INTENTIONAL OR WILLFUL MISCONDUCT; (IV) BREACH OF EITHER PARTY'S OBLIGATIONS OF CONFIDENTIALITY; OR (V) VIOLATION OF ANY APPLICABLE LAW.

Disclaimer of Warranty. EXCEPT AS EXPRESSLY PROVIDED HEREIN, NOVACENE MAKES NO WARRANTIES OF ANY KIND, WHETHER EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, AND SPECIFICALLY DISCLAIMS ALL IMPLIED WARRANTIES, INCLUDING ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW. WITH RESPECT TO THE USE OF THE SERVICES, NOVACENE MAKES NO EXPRESS OR IMPLIED WARRANTY THAT SERVICES ARE OR WILL BE ENTIRELY SECURE, UNINTERRUPTED, WITHOUT ERROR, OR FREE OF PROGRAM LIMITATIONS. YOU SHALL BE SOLELY RESPONSIBLE FOR ANY AND ALL BREACHES RESULTING FROM ITS OR ITS AUTHORIZED USERS' ACCESS TO THE SERVICES FROM AN UNSECURE PLACE OR NETWORK, OR FROM A JURISDICTION THAT MONITORS NATIONAL INTERNET USE.

What is the NovaceneAI Platform?

The NovaceneAI Platform enables developers, data scientists, research analysts, and non-technical domain experts to harvest insights from their data using artificial intelligence (AI) and machine learning (ML). The process involves three steps: ingesting data, enriching data, and visualizing the insights.

Novacene AI Corp.
123 Slater Street, Suite 610, Ottawa, ON K1P 5H2, Canada | (800) 717-0814 | www.novacene.ai

Quick start

A **typical analysis process** is comprised of the following steps:

1. Prepare your data following the data preparation steps
2. Upload your data file ("dataset") following the data ingestion steps
3. Analyze your dataset using the following the autopilot enrichment steps or data enrichment steps
4. Once an enricher finishes running, load the enriched dataset (the dataset automatically created by the application when you applied the enricher) in the Studio and apply the next enricher in your enrichment sequence. Repeat this process until all enrichers in your sequence have been applied.
5. Visualize the results and create your own reports in the Stage following the data visualization steps.

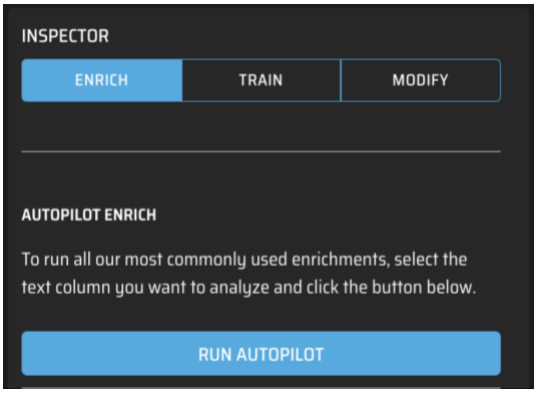
Data preparation steps

1. If you need to analyze data from different files, merge them into one single file
2. Remove any columns unnecessary to your analysis
3. Format cells containing text as '**Text**' to prevent Excel from turning values into formulas
4. Format cells containing dates to use one of the supported date formats
5. Rename columns to use short and descriptive names
6. Save the file in one of the supported file formats, using detailed data preparation steps.

Autopilot Enrichment

To run all of our most commonly used enrichers with one click, take the following steps.

1. Select the column of text you want to analyze in the Studio
2. Hit the “Run Autopilot” button



This will queue up the following enrichers. You can monitor the progress of them running on the Jobs page.

1. Language Translator
2. Clause Extraction
3. Text Pre-processing
4. Clustering
5. Cluster Theme Extraction
6. Sentiment Analysis

You can choose to run further enrichments by selecting the enrichment manually from the dropdown in the Studio.

More Enrichment Sequences

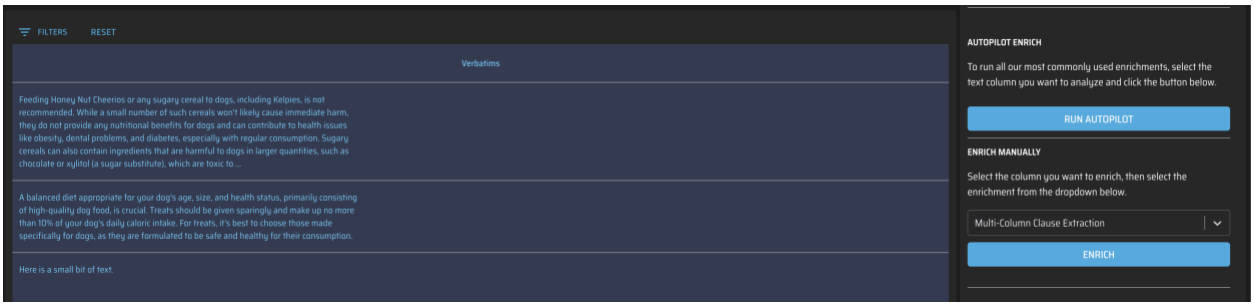
Enrichment sequences describe which enrichers and in which order are to be applied to your data. The below is a sequence for analyzing open-ended feedback, such as survey open-ends, or reviews. **Note:** Before applying the next enricher in the sequence, you need to make sure that the last-enriched dataset is loaded in the Studio.

Step #	Enricher	Target column
1	Language Translator *(if the data is multilingual)*	The column containing the data to analyze
3	Social Media Content Cleanser *(if the text contains content from social media)*	The column containing the processed text
3	Clause Extraction	The column containing the processed text (or the cleansed text if the Social Media Content Cleanser was applied).
4	Sentiment Analysis	The column containing the clauses
5	Clustering	The column containing the clauses
6	Cluster Theme Extraction	The column containing the clauses

Multi-Column Clause Extraction

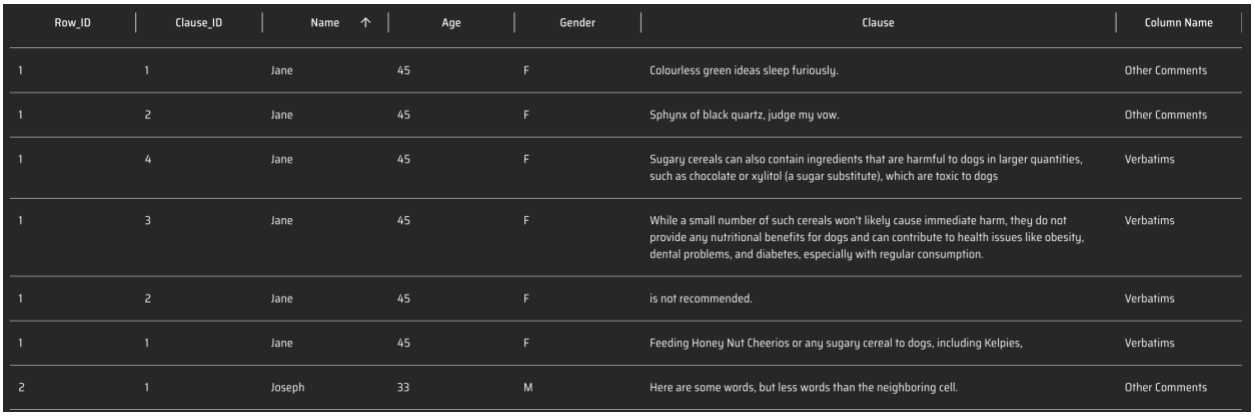
If your dataset contains multiple columns of text that need to be analyzed, the **Multi-Column Clause Extraction** enricher will merge selected columns into a single column. From there, you can run enrichments on your entire dataset at once, giving you a more holistic view of the topics in your dataset.

To use this enricher, select **Multi-Column Clause Extraction,** select a column with text you'd like to analyze, and run Multi-Column Clause Extraction. Once it has completed, open the enriched file, repeat the process on the next column until every column of interest has been run.



At the end of this process, all of your open-ended questions will be merged into two columns at the far right of your dataset. The column named "Column Name" will contain the text from the headers of your selected columns, and the column named "Clause" will contain the clause-

extracted text from your columns. From there, you can run any other enrichments you need on your clauses, including Autopilot.



Row_ID	Clause_ID	Name ↑	Age	Gender	Clause	Column Name
1	1	Jane	45	F	Colourless green ideas sleep furiously.	Other Comments
1	2	Jane	45	F	Sphynx of black quartz, judge my vow.	Other Comments
1	4	Jane	45	F	Sugary cereals can also contain ingredients that are harmful to dogs in larger quantities, such as chocolate or xylitol (a sugar substitute), which are toxic to dogs	Verbatims
1	3	Jane	45	F	While a small number of such cereals won't likely cause immediate harm, they do not provide any nutritional benefits for dogs and can contribute to health issues like obesity, dental problems, and diabetes, especially with regular consumption.	Verbatims
1	2	Jane	45	F	is not recommended.	Verbatims
1	1	Jane	45	F	Feeding Honey Nut Cheerios or any sugary cereal to dogs, including Kelpies,	Verbatims
2	1	Joseph	33	M	Here are some words, but less words than the neighboring cell.	Other Comments

Platform overview

In this section:

Signing in

> Authentication is available out-of-the-box on instances hosted by NovaceneAI, and as a paid add-on service available for air-gapped instances hosted on AWS. Authentication is not available on air-gapped environments hosted outside AWS.

With this manual, you should have received the URL to your instance and a set of user credentials. If you haven't received it or misplaced it, please contact your administrator.

You are required to set up multi-factor authentication to help keep you and your data secure. This change will require some setup steps.

We recommend you use the Microsoft Authenticator application, however you can also use Google Authenticator (Android and Apple) or LastPass Authenticator (Android and Apple) if you prefer.

Next, head to your landing page, and your login screen will look like the image below. Enter your credentials contained in the invitation email and click "Next".

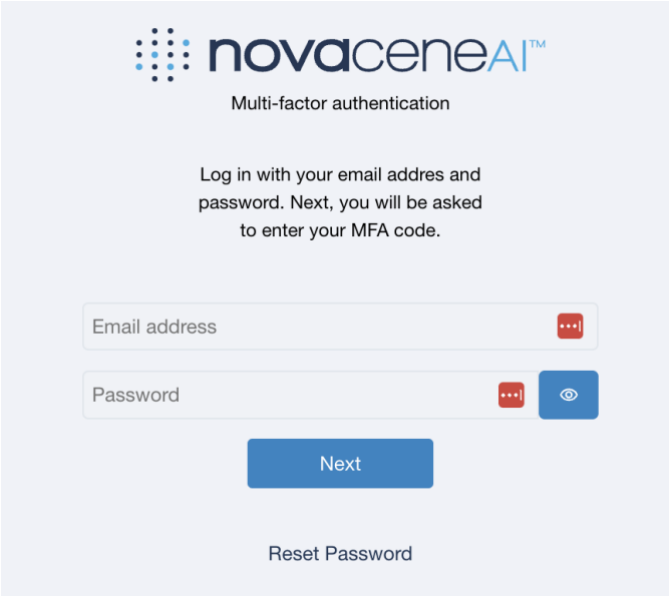


Figure 1: NovaceneAI Platform Log In screen

Next, you'll be prompted to set a new password. Your new password must:

- Be at least 8 characters in length
- Contain at least 1 number
- Contain at least 1 special character (!@#$% etc.)
- Contain at least 1 uppercase and 1 lowercase letter

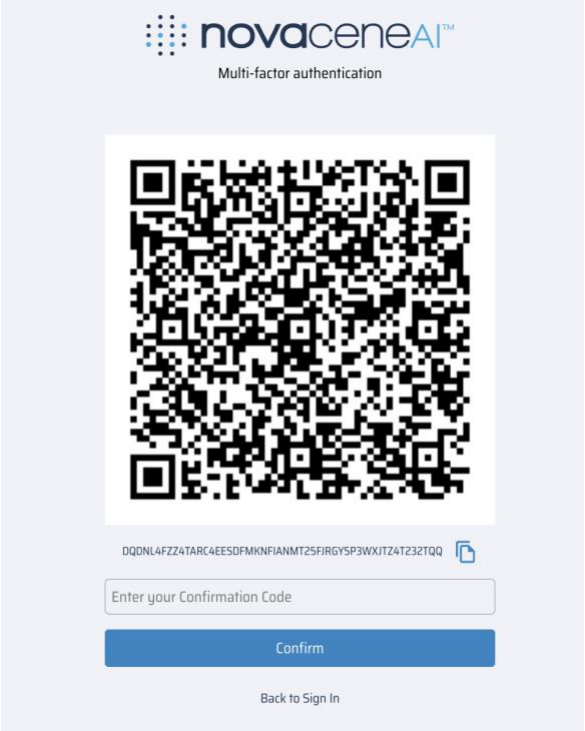


Figure 2: Multi-factor authentication setup screen

Once you select a password, you'll be presented with a QR code. Scan this QR code with your Authenticator app. This will give you a 6-digit code to enter into the platform. Once this step is complete, you're logged in! Now, each time you log in, you will need to get a 6 digit confirmation code from your Authenticator application. To balance security and usability, you will be required to log in again every 7 days.

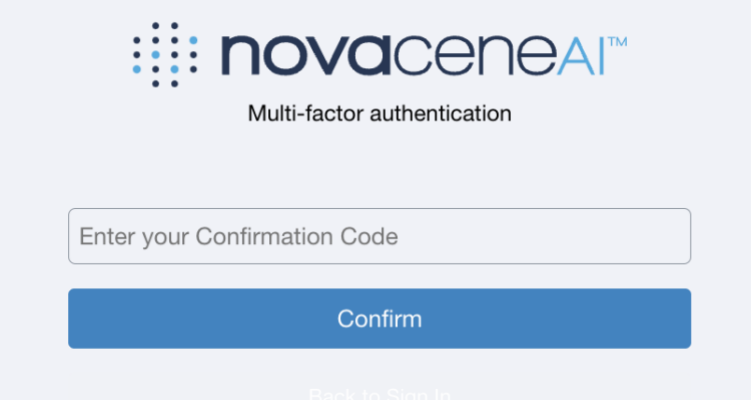


Figure 3: Confirmation code screen on login

Overview of the web user interface (UI)

The web UI is organized into four screens:

1. **Data:** Ingest data and organize it into Projects. For information on how to ingest data, consult the

2. **Data** management section.
3. **Studio:** Review and enrich data using various enrichers. For information on how to apply and train enrichers, consult the **Data enrichment** section

Stage: Visualize the data using charts. For information on how to visualize data, consult the

4. Data visualization section.
5. **Lab:** Measure accuracy of trained models. For information on how to measure the accuracy of trained models, consult the **Measuring enricher accuracy** section.
6. **Jobs:** Monitor and review enrichment and training jobs. For information on how to monitor jobs, consult the **Monitoring enrichments** section.

Detailed Data Preparation Steps

Formatting cells

MS Excel or other text editors will allow you to prepare your data for processing in the NovaceneAI platform.

Text

To ensure that cells containing text remain formatted correctly, highlight the columns containing text and select “Text”.

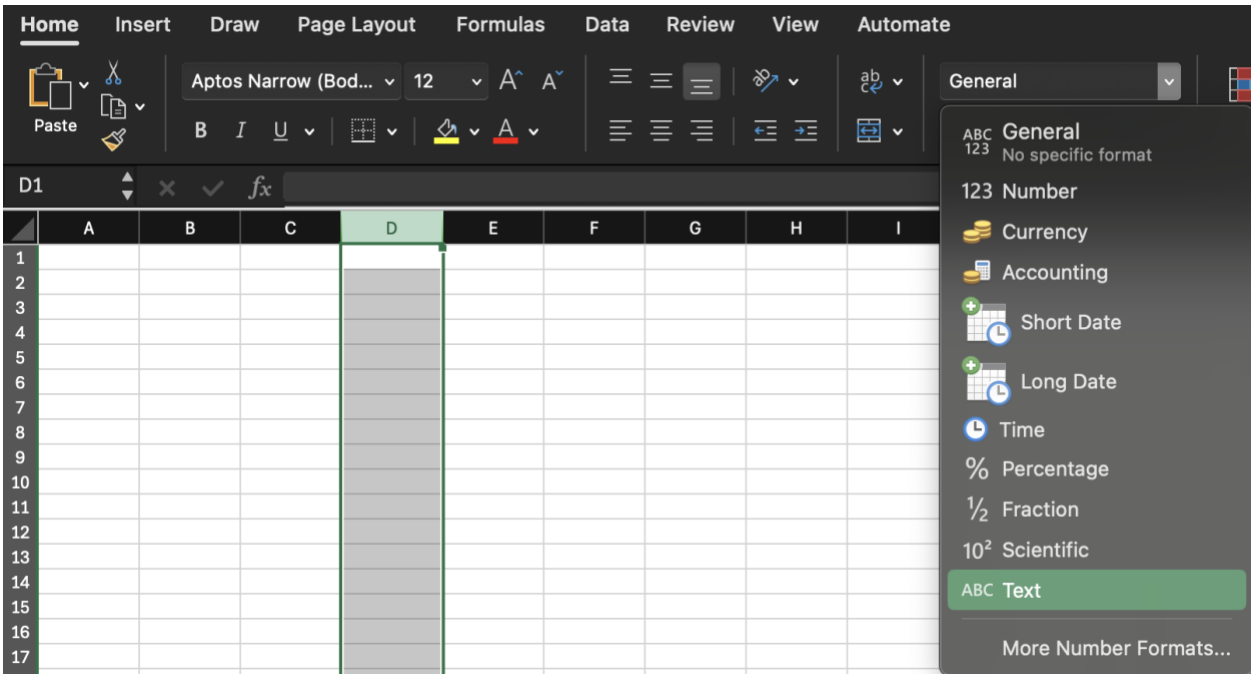


Fig. 4: Formatting a column in excel

Dates

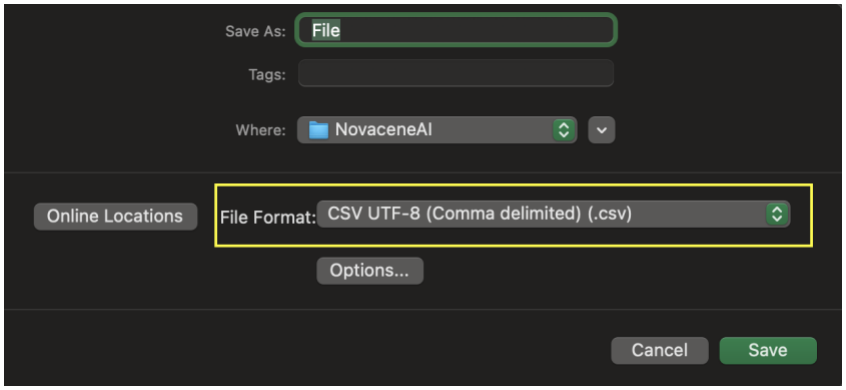
Ensure your dates are in one of the following formats:

- YYYY/MM/DD
- MM/DD/YYYY HH:MM:SS AM/PM
- YYYY-MM-DD HH:MM

MS Excel can help you automatically convert a column to this format.

Exporting data from excel into .CSV UTF-8

If your data is in .xlsx or .xls format, go to File → Save as…, which will open a dialogue where you can select .CSV UTF-8 under “File Type”, as shown below.



Data management

The Data screen acts as a repository of all the data you upload to the application, as well as the data that is created within the application through enrichments. Each of these data objects are called *Datasets.*

In this section:

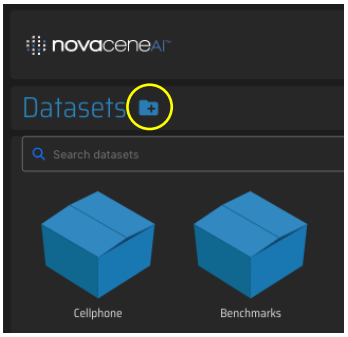
Projects

Projects act as dedicated environments where you can organize and analyze your datasets. These instructions will describe the creation, sharing, and management of these Projects, ensuring a collaborative space where team members can contribute, access, and control data analysis jobs in a secure and structured manner.

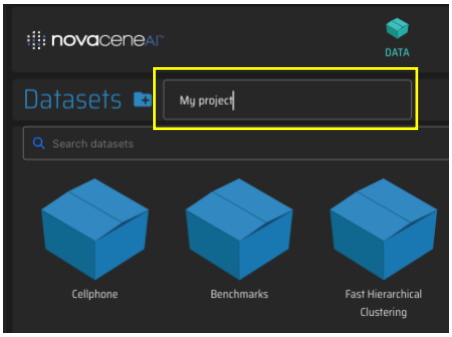
Creating a Project

Initiate your data exploration by creating a new Project. This is your first step towards building a shared workspace where you can collaborate with others to analyze your datasets.

1) Go to the Datasets screen.
2) Click the “Add Project” icon, name your Project, and click enter. Your project will appear on the screen.



1 – Click the “Add Project” icon



2 – Name your Project

3 – The Project will appear on the screen

Once you’ve created the Project, you can enter it by double-clicking the Project box.

Project access and visibility

Maintaining the integrity and confidentiality of your data analysis projects is crucial. This section guides you on how to manage access to your projects, ensuring that only authorized users can view and interact with the sensitive datasets you are working with. Note that only Project owners—the users who created the Project—and Project Collaborators can see and access the datasets saved in the Project. This access restriction extends to saved reports.

Granting and managing collaborator access

Effective data analysis often requires a team effort. Here, we detail the process for inviting Collaborators to your Projects and controlling their access levels. You'll learn how to empower your team with the right access to contribute to data analysis jobs while safeguarding your datasets from unauthorized access.

1) **Granting and revoking access:** Within the Project settings, use the 'Collaborators' section to add users by checking the box next to their email address or username.

Revoke a Collaborator's access by unchecking the box. Revoked collaborators will immediately lose access to the data housed in the Project. If they are actively in the Project during revocation, they will remain inside the Project until they leave, but they will not be able to view or interact with any datasets.

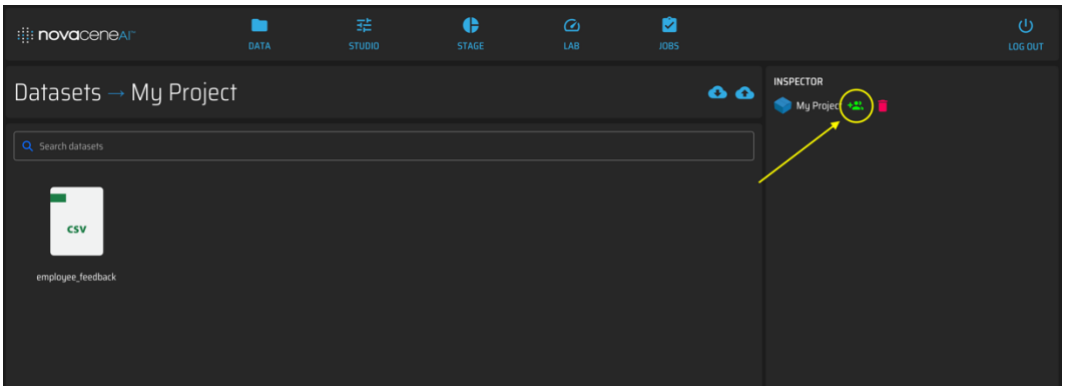


Figure 4: Click the Collaborators icon

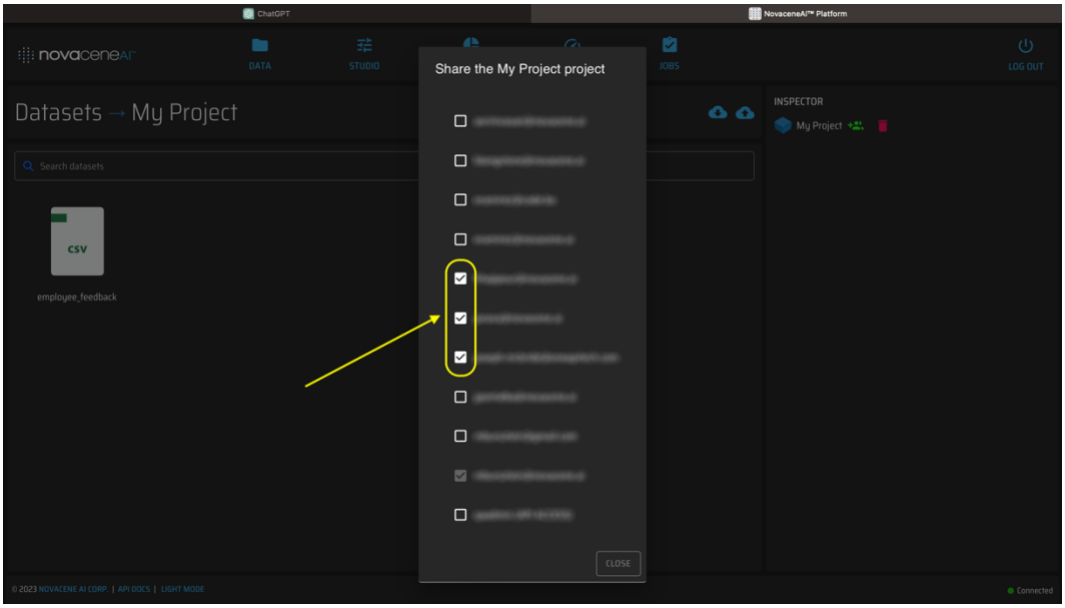


Figure 5: Select users to grant them Collaborator access.

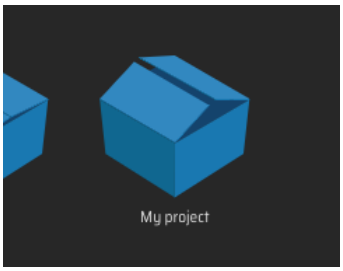
Collaborator permissions:

- Access datasets and saved reports associated with the Project.
- Further grant access to other users.
- Revoke the access of any other Collaborators. Note that Collaborators cannot revoke the access of the Project owner.
- Remove themselves from a Project. This action is irreversible through their own accord; they cannot access the Project again unless re-invited.

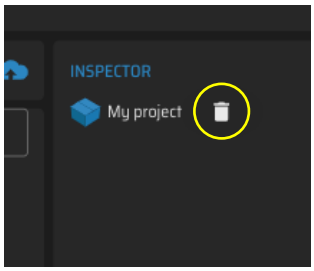
- Delete the Project.

Deleting Projects

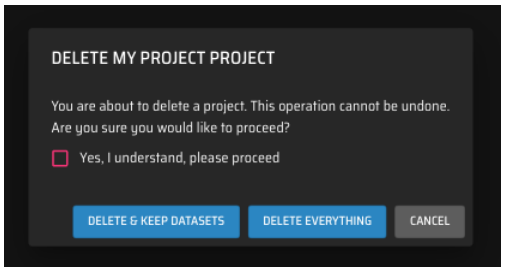
Follow these steps below to delete a Project.



1 – Select the Project you want to delete



2 – Click the Trash icon in the Inspector



3 – Provide confirmation and choose a delete option

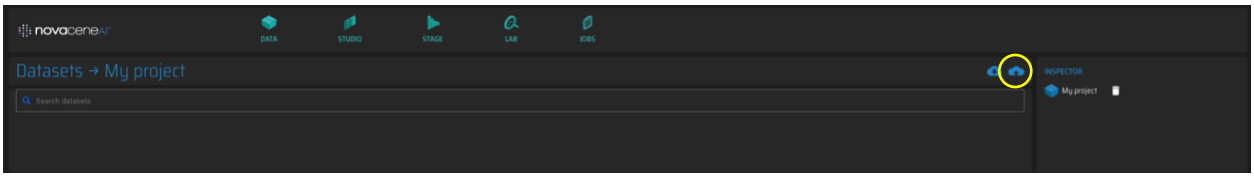
Ingestion

You can bring their data in one of two ways:

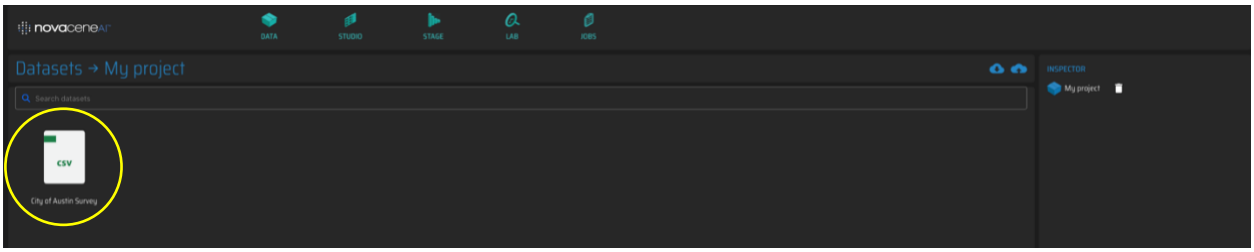
- Uploading files
- Fetching from external sources
- Add-on connectors

Upload

To upload a file, click the Upload icon located on the Data screen, and select a file from your computer.



Once uploaded, the file will appear on the screen.



File types

You can upload files to the application. Supported file formats include:

- CSV UTF-8 (Comma delimited) (.csv)
- Comma Separated values (.csv)
- Macintosh Comma Separated (.csv)
- MS-DOS Comma Separated (.csv)
- ZIP

File uploads can be performed in different ways:

- Using the web application
- Using the API
- Via SFTP

About character encodings: If your CSV file is of an unsupported encoding (i.e. “UTF-16“, UTF-16 LE”), you can convert the file using MS Excel and selecting “UTF-8” as the target saved file encoding. Once the CSV file is in “UTF-8” encoding, you will be able to upload it to the application.

Recognized data types

The application automatically recognizes the following data types:

2. **Date:**
    - YYYY/MM/DD
    - MM/DD/YYYY HH:MM:SS AM/PM
    - YYYY-MM-DD HH:MM

    Note: Using any other format may result in inaccurate visualizations.

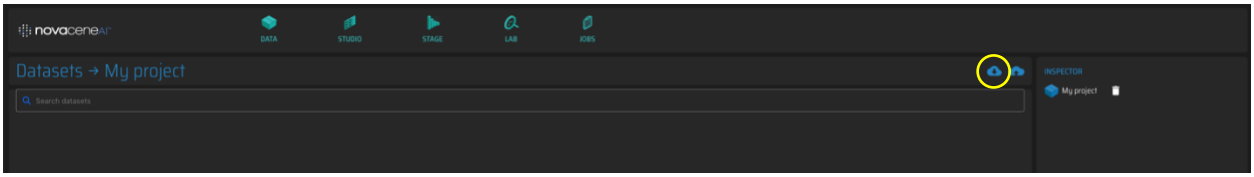
3. **Geolocation:** When the dataset contains two columns labelled: “Latitude” and “Longitude”

Fetch

The application allows you to fetch data from the following sources:

- Twitter
- The web

To fetch data, click on the Fetch icon and follow the instructions below specific to the source from which you are fetching.



Fetching from X (Twitter)

To fetch data, enter a search query in the search box. To fetch from Twitter, type in your query starting with “#”

Note: the application will omit the '#' when querying Twitter. (e.g., #bitcoin will search for bitcoin)

The application will fetch 500 tweets per request. Note that there is a limit of 100 requests per month.

Operators:

Note: depending on your subscription, some operators **may not be available.**

-is:retweet: excludes tweets that were retweeted (includes only original tweets)

Example query: #bitcoin -is:retweet

Columns in generated dataset:

Column heading	Description
Tweet ID	The ID of the Tweet
Created At	The date and time of the Tweet
Author	The Tweet’s author
Followers	The Tweet’s author’s number of followers
Text	The Tweet text
Favourites	The number of times the Tweets was favourited
Retweets	The number of times the Tweets was retweeted
Author Location	The author’s location as specified by the author in their profile
Tweet Country	The Tweet’s country, if available
Tweet Bounding Box	The Tweet’s bounding box, if available

Referenced Tweet ID	The Tweet being retweeted or “Original Tweet” of the Tweet is original and not a retweet.

Table 1: Twitter dataset columns

Fetching from the Web

The Web connector allows you to fetch content found in news sources on the web. To fetch data, enter a search query in the search box, and hit the Enter key.

The application will search the web for results matching the query, and for each result, it will scrape the content.

Add-on connectors

The following connectors are provided outside of the platform. Contact us to use these connectors.

Reddit

The Reddit connector will fetch posts (“submissions”), comments, and information about posters. To use the connectors, you must provide keywords, the subreddit to search, and the timeframe.

Option	Description
Keywords	See https://support.reddithelp.com/hc/en-us/articles/19696541895316-Available-search-features#h_01HBS25GVN59BBF1TMQCRZT3GA
Subreddit	The name of the subreddit to search (e.g., 3Dprinting)
Timeframe	Must be one of: hour, day, week, month, year, or all

The following table shows the output from this connector:

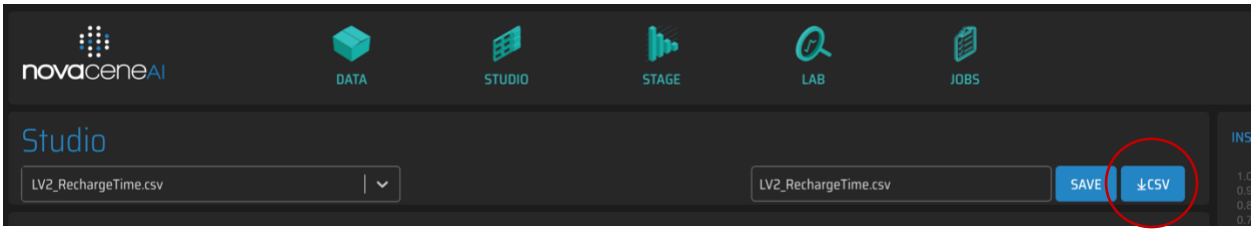
Option	Description
ID	The ID of the submission or comment (together, “content”)
Subreddit	The Subreddit where the content was posted
Keyword	The keyword used to fetch the submission
Date	The date the content was posted
Title	The title of the submission
Text	The text of the content

Votes	The number of votes on the content
Author	The author who posted the content
Karma	The author's comment karma
Parent submission	The submission against which the comment was posted
Type	Whether the content is a submission or a comment

Limits: The current connector returns up to a maximum of 250 submissions.

Export

You can export data from the application in CSV format. To export the data, go to the Studio, load the dataset you wish to export, and click the "CSV" button.



Opening exported CSV files in Excel

For best results, **import** the file instead of simply opening. When importing the file, select:

a) Comma as the delimiter
b) "UTF-8" as the "File Origin" option
c) "Text" as "Column data format" option for all columns

Purging

Manual data purging

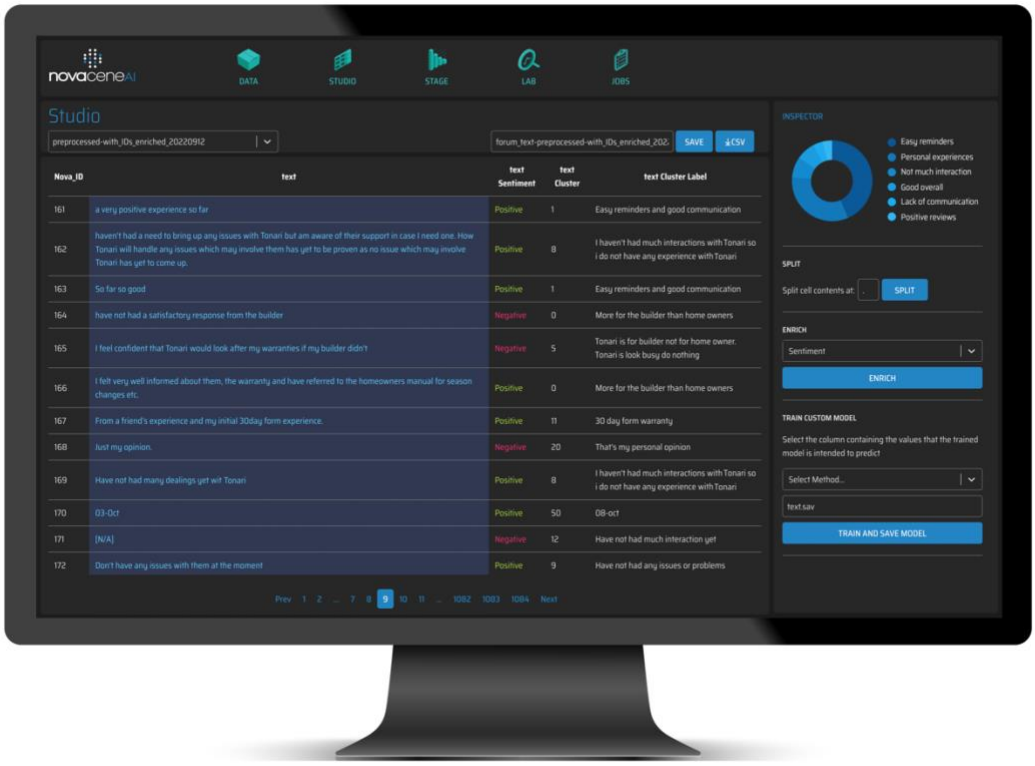
You can delete datasets manually via the web application, or by calling the appropriate API method.

Automated data purging

During initial provisioning, the application is configured to purge datasets automatically after a certain number of days (the retention period). If you process many daily jobs, it is possible that your instance will accumulate several datasets, which may degrade the speed with which the application responds. In this case, please contact support to reduce the retention period, or increase the compute power of your instance.

Data enrichment

Data enrichment represents the core functionality of the application. Enriching data is the act of inferring new data from your existing data using various enrichers. All enrichment operations are performed in the Studio screen.



In this section:

Enrichers

Enrichers are a combination of AI & ML models and algorithms available to you for performing inference and transformations on your data. The application provides built-in enrichers, and you can also train your own.

Built-in enrichers

Built-in enrichers do not require to be trained and can be used out-of-the-box. Trained enrichers are enrichers you trained using your own training data and are therefore adapted to your specific need.

To apply built-in enrichers:

1. Open your dataset in the Studio
2. Select the input column
3. Select the enricher, and click the “Enrich” button

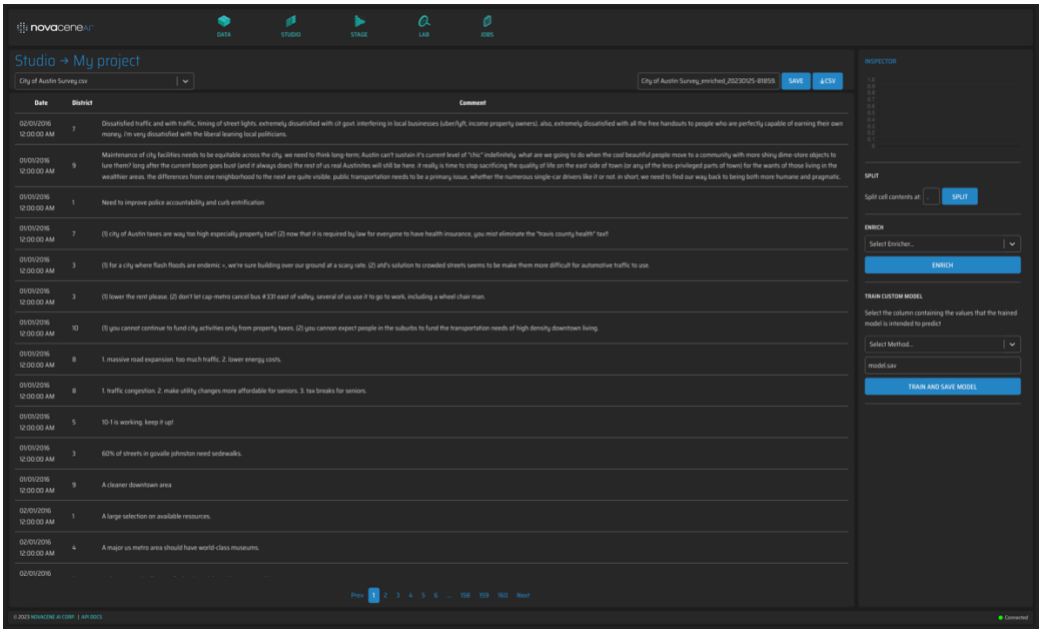


Figure 3: To open a dataset in the Studio, either click on the dataset icon on the Data screen and then click the Studio icon on the top navigation bar; or go to the Studio and select the dataset from the drop down menu on the top-left area of the screen

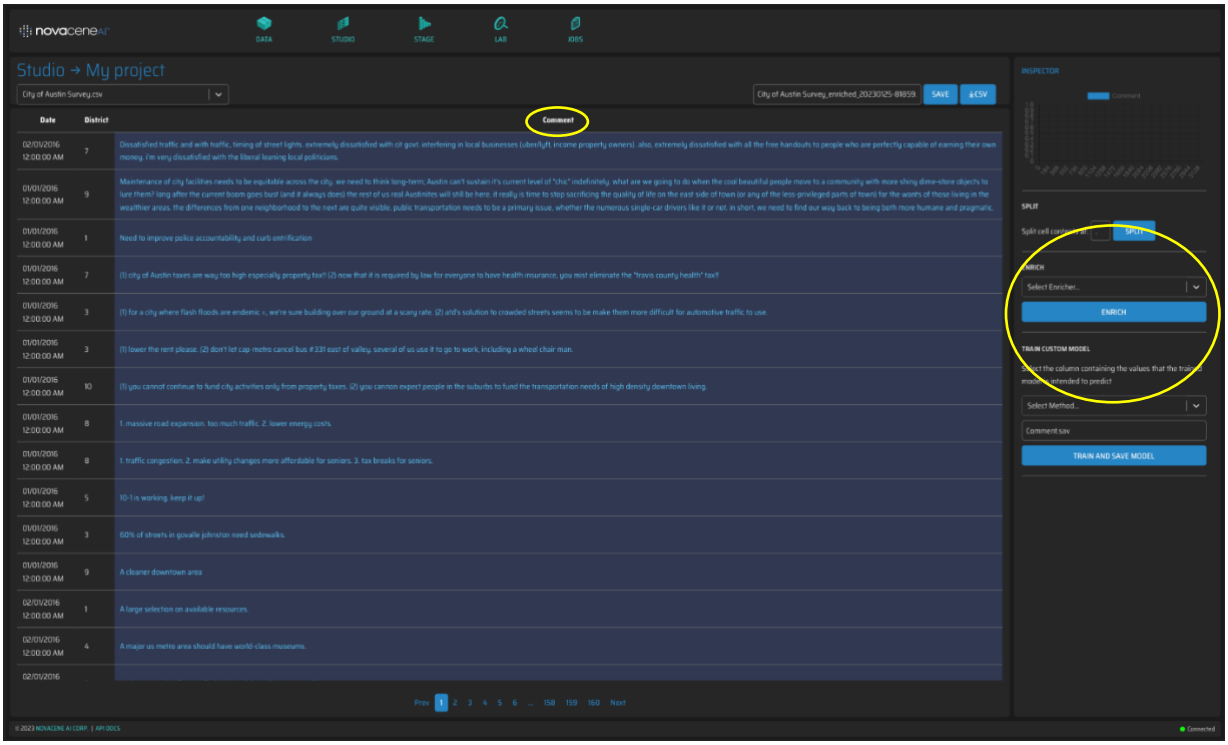


Figure 4: To select the column to which you wish to apply the enricher, click on the column heading

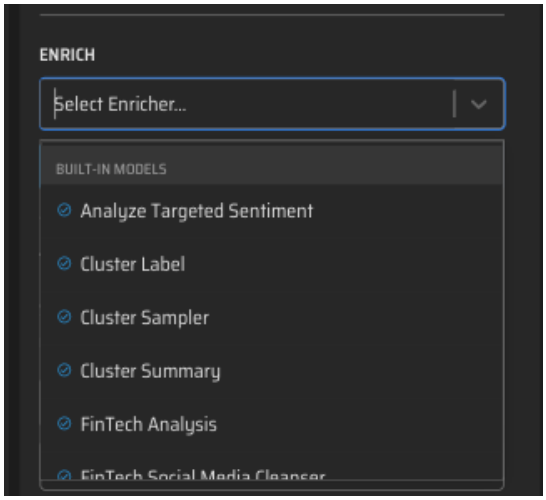


Figure 5: Select the enricher you wish to apply

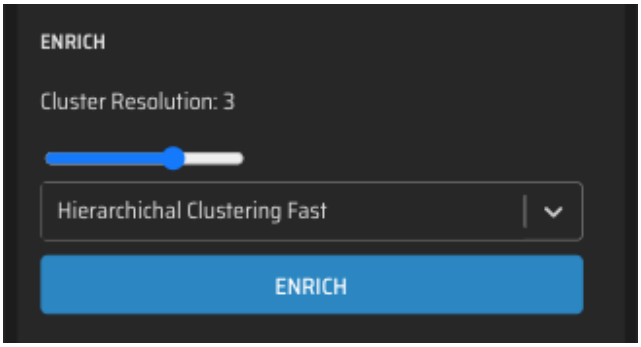


Figure 6: Adjust settings if applicable, and click the Enrich button.

Enriched datasets

Each time you apply an enrichment to a dataset, the application clones the original dataset and adds the results from the enrichment to the clone. This is done to preserve the integrity of datasets and provide you with a trail of datasets as they have been enriched.

Viewing enrichment results

You can see the results from enrichments while they're being executed. Note that the progress of enrichers that act on your entire dataset at once, like clustering enrichers for example, cannot be tracked while the enrichment runs.

As soon as you click the Enrich button, the application will add columns to your dataset where the results from the enrichment will be displayed.

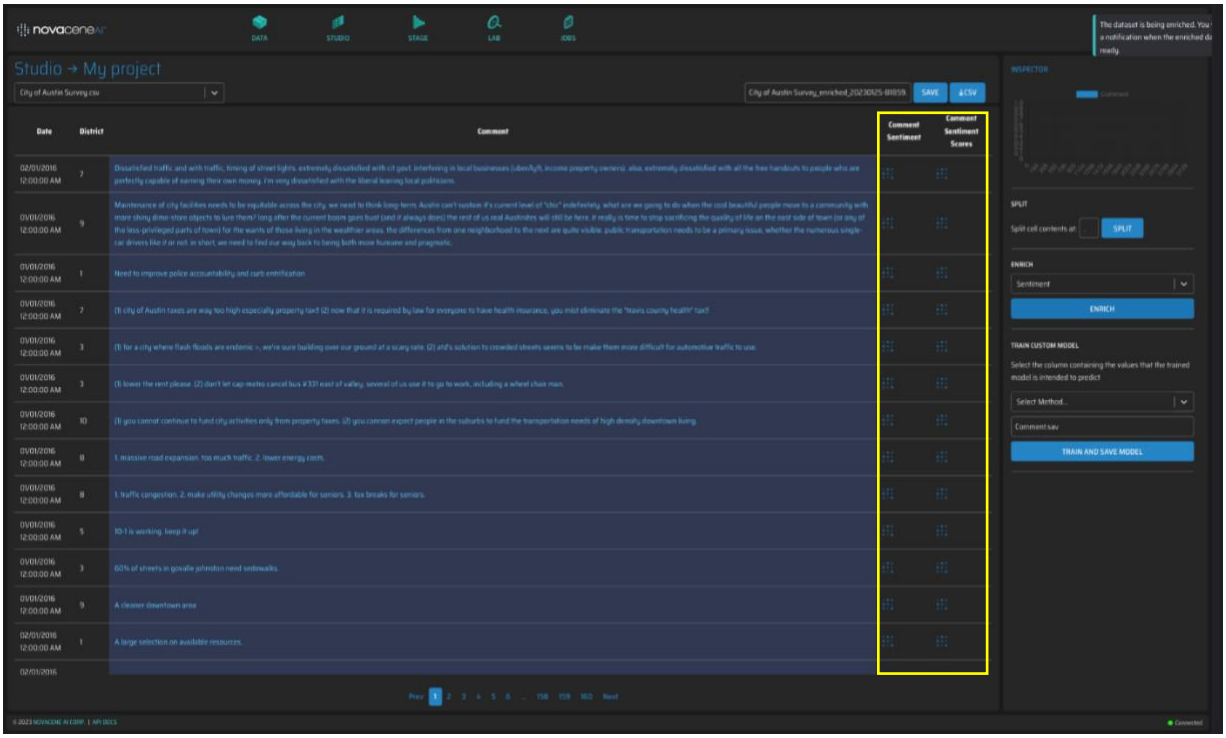


Figure 7: Additional columns added by the application to hold the results from the enrichment

As results become available, the application will "push" the results to the Studio:

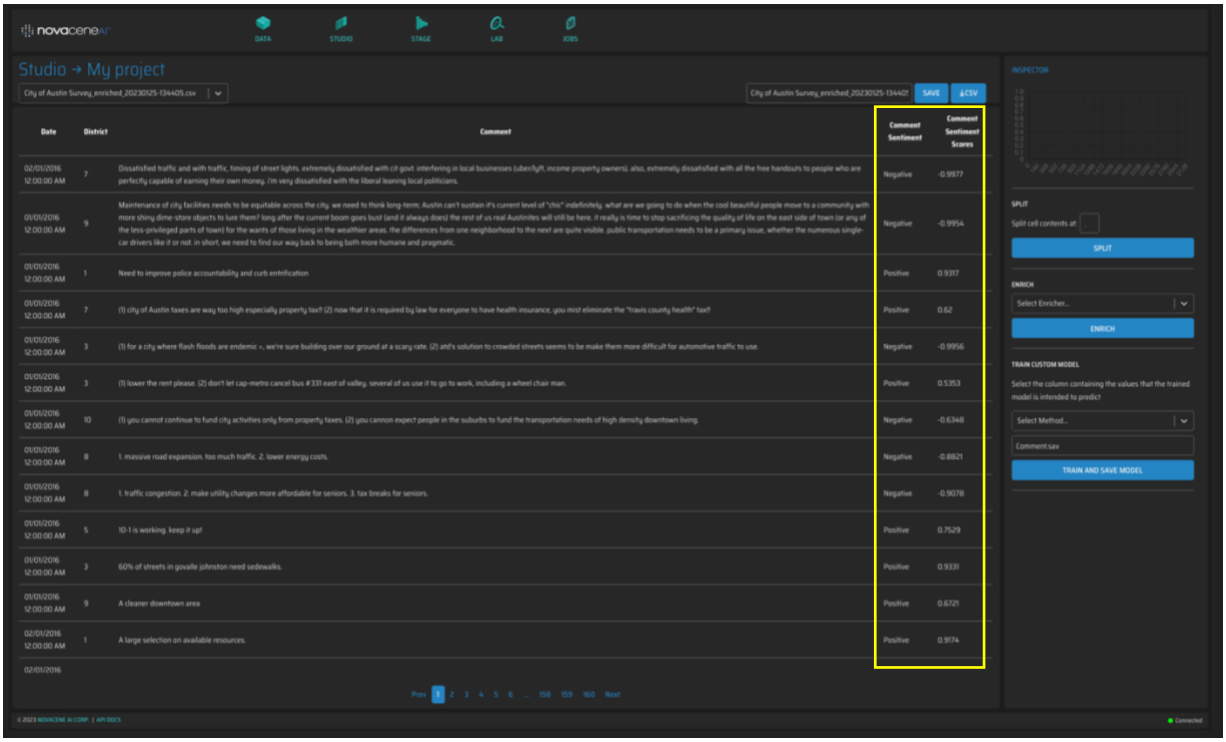


Figure 8: Results from the enrichment job are pushed to the Studio as they become available. Note that the predictions shown are only a preview, and that you will need to load the enriched dataset in order to apply further enrichments

Available built-in enrichers

Clause Extraction

The Clause Extraction enricher splits a paragraph into clauses, allowing you to apply other enrichers to your text in a more granular way.

> **Why doesn't the Clause Extraction enricher add a column to my dataset on the Studio screen when I click the Enrich button?** The Clause Extraction enricher adds both multiple rows and multiple columns to your dataset. Currently, the interface only able to add multiple columns only.

Clustering

The Clustering enricher organizes similar text into categories. Each category is represented by a number, starting with 0.

The *Cluster Resolution* setting affects the number of clusters that are formed. A higher resolution results in many clusters, while a lower resolution results in fewer clusters.

This enricher does not group unclustered documents into a common class (like "Other") but rather it forces all documents to be placed in clusters, even if some clusters end up containing one single document.

Cluster Label[1]

Outputs the single most representative sample in a cluster, providing an understanding of the topic of the samples in the cluster.

Cluster Theme Extraction[1]

Outputs a label based on extractive samples of text in the cluster to provide an idea of the themes in a given cluster. Outputs five pieces of text separated by a bar symbol.

LLM Cluster Theme Extraction[1]

Outputs a *generative* label based on samples of text in the cluster to provide an idea of the themes in a given cluster. This enricher uses the LLM selected for your environment. As this is enricher is generative, it can output different results each time it is run.

Cluster Summary[1]

Outputs a few samples formatted as a paragraph, providing context as to what the samples in the cluster are about. Requires that the input data has already been clustered.

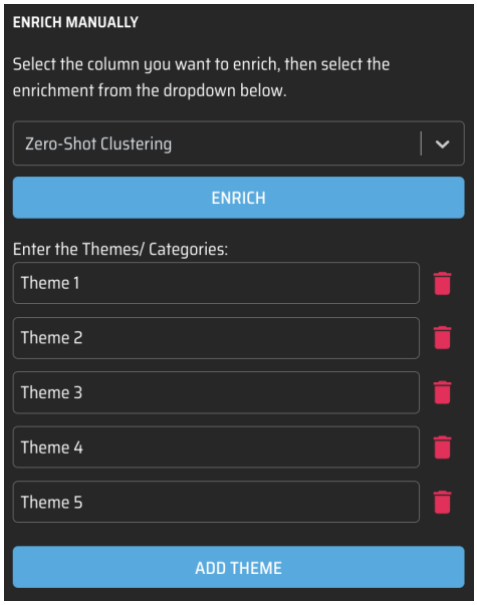
Cluster Summary Generator[1]

Generates a summary of the contents of the cluster using the OpenAI GPT API. Note that to overcome limits imposed by the API, the summary might based on a subset of the contents of each cluster. Requires that the input data has already been clustered.

[1] Requires that the input data has already been clustered.

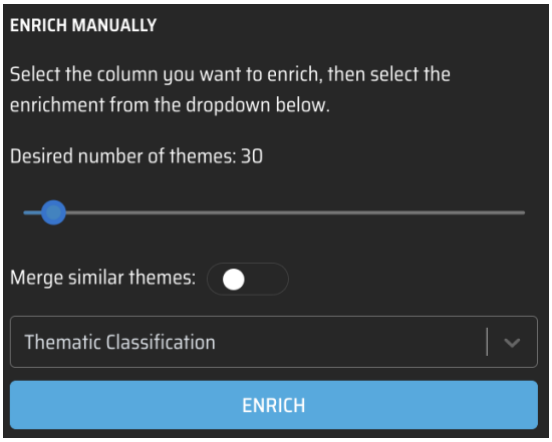
Zero Shot Clustering

This enricher uses an LLM to categorizes text according to themes input by the user. Users can add or remove themes in the Inspector.



Thematic Classification

This enricher is designed for large datasets. When using Thematic Classification, you can choose the maximum amount of themes you'd like to receive. You can also choose to merge similar themes with a toggle when running this enricher.



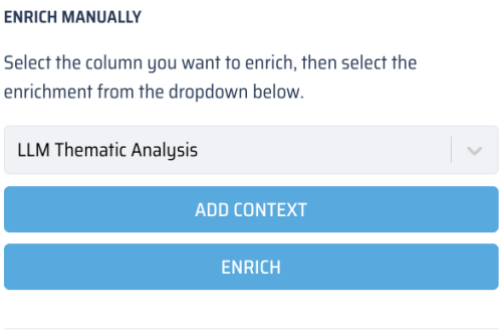
LLM Thematic Analysis

This enricher will provide theme labels for your text. It allows users to include additional context to ensure that the enriched results are informative and related to your work. When using this enricher, you can provide:

- Research Question

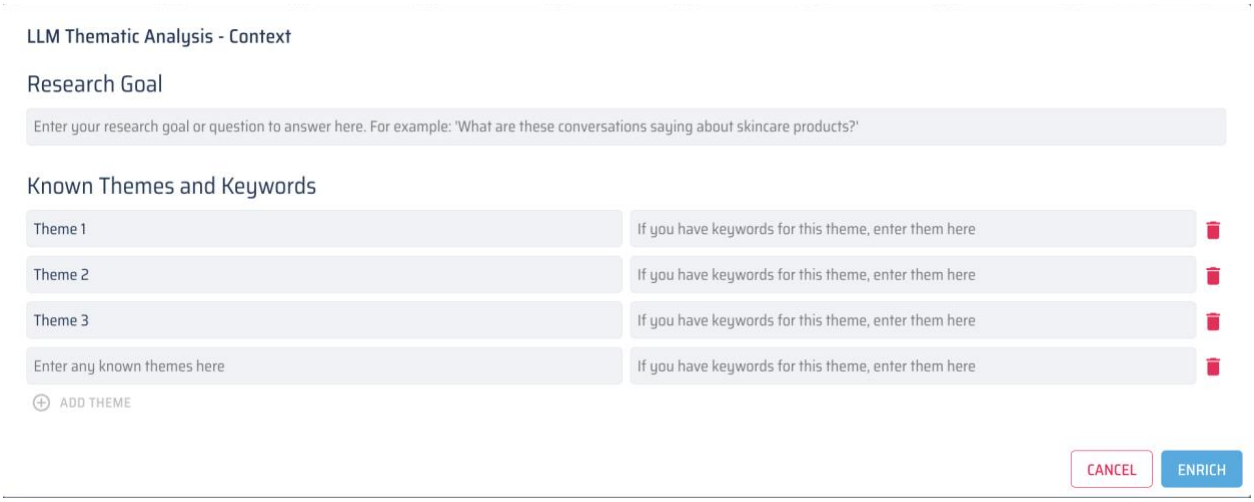
- Example Themes
- Keywords Relating to Themes

This enricher will output four columns: theme labels, justifications for selecting the theme labels, and two variations of the theme labels.



Click on the column you'd like to enrich. In the Inspector, select LLM Thematic Analysis from the dropdown. Click the "Add Context" button.

A dialog will open with space for you to enter your research question, your known themes, and any keywords that relate to your themes.



You can add as many themes as you like—the LLM will label the others that don't fit within your labels. You don't have to add a research question, or keywords. However, you can't enter keywords without first entering a theme.

Providing a **Research Goal** will help the LLM understand the **context** of your research. You can include information like "These responses are to the question "What influences your opinion on social media?" I am trying to understand how social media influences the opinions of users, and understand how cognitive biases may play a role."

If you know certain themes are present in your dataset, you can list them, and associate any relevant keywords. Each keyword should be separated by a comma. You do not need to use * operators—LLMs are good at picking up variations of a root word.

If you don't have a specific research goal, or know of any themes in your dataset, you can always run the LLM Thematic Analysis enricher without providing this context by simply hitting "Enrich".

This enricher will output 4 columns.

[Name of selected column] Themes Original	[Name of selected column] Theme Justification	[Name of selected column] Themes 0.6	[Name of selected column] Themes 0.8
These are the original themes provided by the LLM.	This is the reasoning that the LLM used to select the theme label in the column to the left.	This column contains the least theme labels. Similar themes have been merged together.	This column contains less theme labels than the original. Very similar themes have been merged together.

We suggest looking at each of the theme columns in the stage to determine which one has the best resolution for your data. If you think the original themes look like they are too similar or redundant, try at 0.8. If you still find that there are very similar labels, look at 0.6. The best selection will depend heavily on your input data, the range of topics being discussed, and the language being used in the data.

Results will include your input themes, as well as other themes found in the data. Themes will be relevant to your research goal.

Sentiment Analysis (Offline)

This enricher analyzes the sentiment of text, and classifies results as positive, neutral, or negative. This enricher also outputs a confidence score of the classification (i.e. a number between 0.00 and 0.99 indicating how certain the model is of the classification)

Sentiment Analysis

This enricher outputs three columns of enrichment. In the first column, it analyzes the sentiment of text, and classifies results as positive, neutral, or negative. In the second column, it describes the emotion of the text. In the third column, it outputs whether the text is sarcastic or sincere.

Emotion Analysis

This enricher analyzes the emotion of text, and outputs an emotion label.

FinTech Analysis

Classifies text into one of these categories: ["mining", "price", "scam", "project", "wallet"]

FinTech Social Media Cleanser

Strips tweets from potentially distracting content such as broken URLs, special symbols, Retweet (RT) tags, etc.

Hierarchical Clustering

Clusters text by grouping similar content together into categories. Further documentation available at: https://demo.novacene.ai/docs/novacene-api/redoc/#operation/studio_hierarchical_clustering_social_list

Ideas and Comments Classifier

The *Ideas and Comments Classifier* is an enricher in which input text is classified as an *idea* or *comment.* This is useful for idea extraction and analysis. This enricher was trained using an expert-annotated open dataset related to energy generation. Details on the training dataset can be found in **Deprecated enrichers**

Enricher name	Description	Deprecation reason
Sentiment Sampler	Outputs the top 5 most negative samples in the dataset in decreasing order of negative sentiment score.	Ranking to be treated as a function on the visualization system and not as part of the classification algorithm.
Cluster Summary[1]	Performs a one sentence summary using extractive summarization.	Generative methods outperform this extractive method.
Analyze Targeted Sentiment		
FinTech Analysis		
FinTech Social Media Cleanser	Strips tweets from potentially distracting content such as broken URLs, special symbols, ReTweet (RT) tags, etc.	Replaced by HTML cleanser and Social Media Content Cleanser
Hierarchical Clustering Responses	Clusters text by grouping similar content together into categories. (Optimized for some types of survey responses).	Inflexible for different data formats
Quantum Classifier	Binary classification algorithm that runs on a Quantum backend.	
Hierarchical Clustering Social	Clusters text by grouping similar content together into categories.	

	(Optimized for short texts or social media updates).	
Peer Clustering[2 3]	Clusters records that share many similar attributes.	
Sentence Segmentation	Expands the input text by segmenting input text into separate sentences.	Replaced by Clause Extraction
Sentiment Pre-Processing		Replaced by Sentiment Analysis and Emotion Analysis.
Sentiment Sampler	[TBD]	Please contact us for instructions.
Sentiment Analysis (Retail)	Specific for retail data, classifies the tone of the text as being *positive, negative, or neutral.*	Replaced by Sentiment Analysis
Topic Modelling and Clustering M1	Clustering for text	Earlier versions of Clustering
Topic Modelling and Clustering M2	Clustering for text	Earlier versions of Clustering

Appendix: Supported Languages

Arabic	Dutch	Greek	Italian	Russian
Catalan	Esperanto	Hebrew	Japanese	Spanish
Chinese	Finnish	Hindi	Korean	Swedish
Czech	French	Hungarian	Persian	Turkish
Danish	German	Indonesian	Portuguese	Ukrainian

Novacene AI Corp.
123 Slater Street, Suite 610, Ottawa, ON K1P 5H2, Canada | (800) 717-0814 | www.novacene.ai

Appendix: Ideas and Comments Classifier.

Language Translator

The Language Translator enricher detects text in languages other than English and translates it into English. Our API-based language translation service has been updated.

Why do I sometimes get different translations for the same text? Due to the nature of the automatic language translation technology, you may see slightly different translations for the same text each time you apply the enricher. This is normal, and the differences should not affect the meaning of the translation.

Online Language Translator

We support the following languages for our online Language translator:

Supported languages

Afrikaans, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bosnian, Bulgarian, Catalan, Cebuano, Chichewa, Chinese (Simplified), Chinese (Traditional), Corsican, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Filipino, Finnish, French, Frisian, Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hausa, Hawaiian, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Igbo, Indonesian, Irish, Italian, Japanese, Javanese, Kannada, Kazakh, Khmer, Korean, Kurdish (Kurmanji), Kyrgyz, Lao, Latin, Latvian, Lithuanian, Luxembourgish, Macedonian, Malagasy, Malay, Malayalam, Maltese, Maori, Marathi, Mongolian, Myanmar (Burmese), Nepali, Norwegian, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Samoan, Scots Gaelic, Serbian, Sesotho, Shona, Sindhi, Sinhala, Slovak, Slovenian, Somali, Spanish, Sundanese, Swahili, Swedish, Tajik, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese, Welsh, Xhosa, Yiddish, Yoruba, and Zulu.

Offline Language Translator

The following languages are supported in the Offline Language Translator:

Arabic, Azerbaijani, Catalan, Chinese, Czech, Danish, Dutch, English, Esperanto, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, and Ukrainian.

Named Entity Recognition

The Named Entity Recognition (NER) enricher extracts entities from text. The following entities are recognized:

Entity Type	Description
LOC	Location name
PER	Person name
ORG	Organization name
MISC	Other name

Tips for best results when using NER

When extracting organizations (ORG) from a dataset containing social media data like tweets, it is recommended that you first clean the text using the Social Media Content Cleanser enricher, followed by the Language Translator enricher to translate any text that isn't English.

Peer Clustering

The *Peer Clustering* enricher groups together cases based on the number of shared attributes amongst them. For example, it can be used to group universities by the type of software tools they use, or to group restaurant orders by the type of dishes that have been ordered.

Preparing your data

This enricher requires that data be provided in the following format:

Case id	Type	Value

For example, if you are trying to group together restaurant orders, your data may look like this:

Order id	Meal type	Dish ordered
1	Appetizer	Soup
1	Main	Chicken
1	Dessert	Cake
2	Main	Chicken
2	Main	Lasagna
3	Appetizer	Soup
3	Dessert	Ice cream

Note that you need to supply a minimum of 100 unique **case ids** to use this enricher.

Applying the enricher

To apply the enricher, load the dataset using the format shown above, open the dataset in the Studio, select the Peer Clustering enricher from the dropdown and click the Enrich button. Note that you do not need to select a column to apply this enricher.

Interpreting the results

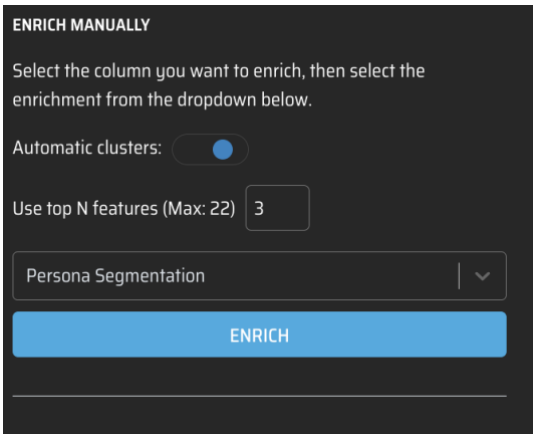
The enricher will output a flattened dataset with a number of rows matching the number of unique case ids. Each row will contain all the information related to each case, and the last

column will contain a number representing the cluster. You can then filter by the this Peer Cluster number to see the cases that have been grouped together.

Persona Segmentation

This enricher helps users understand unique cohorts of populations based on a table of features. This enricher uses feature selection and k-means clustering to assign each row to a cluster, and outputs a crosstab of the features of each cluster, along with their over/under index compared to the entire population. Additionally, this enricher outputs the scores for each potential number of clusters.

To apply the enricher, open the dataset in the Studio, select any column in your dataset, and select Persona Segmentation from the Enrichment dropdown.



Here, you are presented with options for your enrichment.

Automatic Clusters: This will automatically determine the optimal number of clusters (aka ‘k’ or segments) for segmentation with k-means clustering. If you want a specific number of clusters/segments in your output, toggle this control, and enter the desired number of clusters/segments with the slider.

Use Top N Features: This algorithm will automatically select the best features to use with clustering. You can enter the number of features you’d like to use in the text box. The maximum listed number of features is based on how many columns are in your dataset. This enricher needs numerical features to perform k-means clustering.

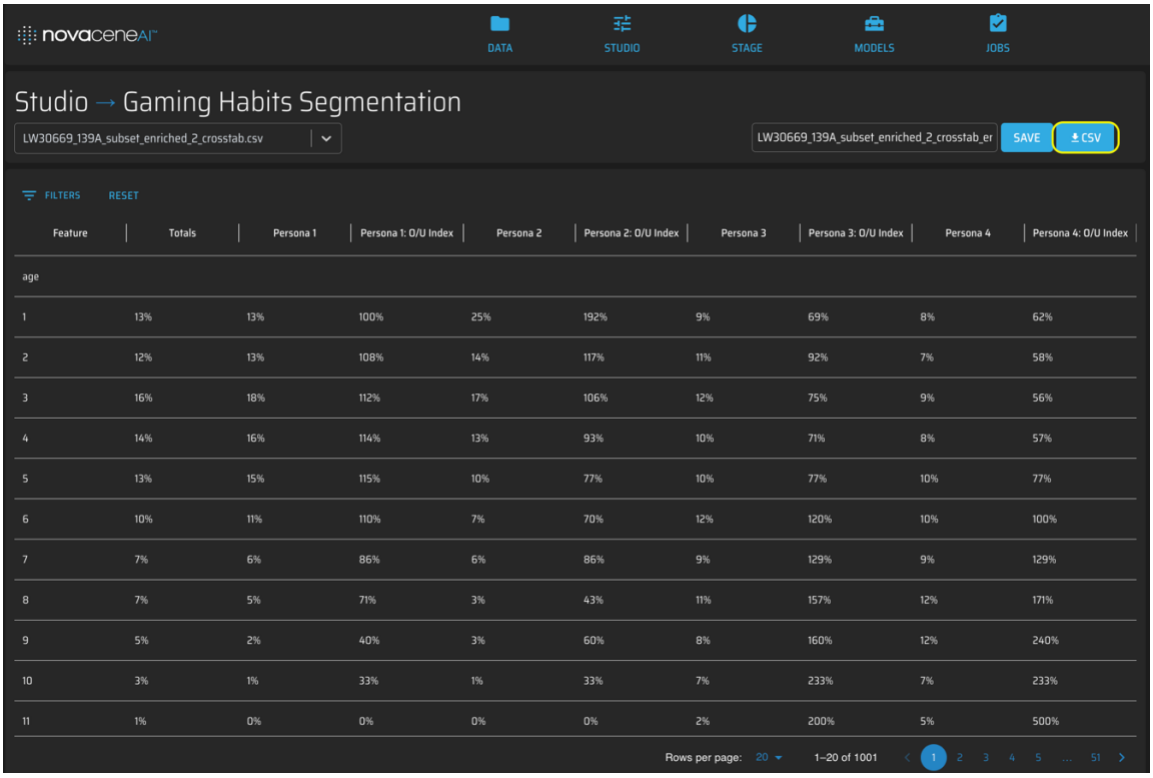
The following features will be automatically ignored in the process:

- Unique identifiers (i.e. any value that is unique in every row of a given column)
- Dates
- Text features (i.e. any features represented as text)
- Any features that are missing values (any column that is missing values in a row)

Next, click **Enrich.** Now, your job will run, and you can monitor progress in the **Jobs** screen.

When your job is completed, you will have two or three outputs:

1. An enriched dataset where each row has been assigned a cluster number in the rightmost column. This dataset will be named [originalfilename]_enriched_[# of enrichments applied].csv
2. A crosstab of the traits of each cluster persona. This dataset will be named [originalfilename]_enriched_[# of enrichments applied]_crosstab.csv



Feature	Totals	Persona 1	Persona 1: O/U Index	Persona 2	Persona 2: O/U Index	Persona 3	Persona 3: O/U Index	Persona 4	Persona 4: O/U Index
age									
1	13%	13%	100%	25%	192%	9%	69%	8%	62%
2	12%	13%	108%	14%	117%	11%	92%	7%	58%
3	16%	18%	112%	17%	106%	12%	75%	9%	56%
4	14%	16%	114%	13%	93%	10%	71%	8%	57%
5	13%	15%	115%	10%	77%	10%	77%	10%	77%
6	10%	11%	110%	7%	70%	12%	120%	10%	100%
7	7%	6%	86%	6%	86%	9%	129%	9%	129%
8	7%	5%	71%	3%	43%	11%	157%	12%	171%
9	5%	2%	40%	3%	60%	8%	160%	12%	240%
10	3%	1%	33%	1%	33%	7%	233%	7%	233%
11	1%	0%	0%	0%	0%	2%	200%	5%	500%

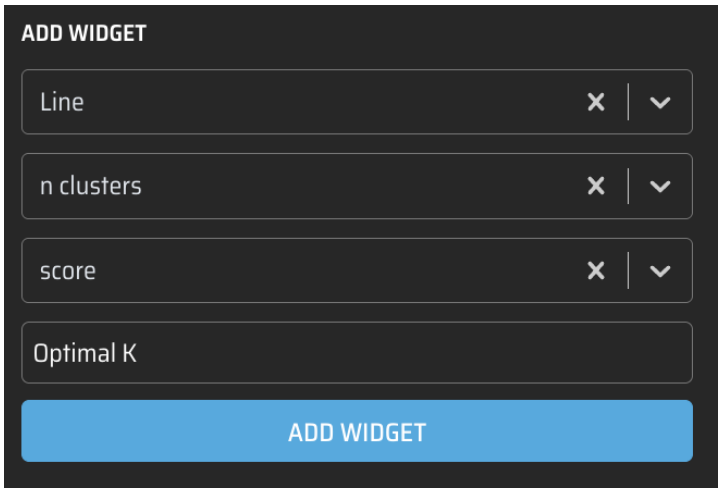
   The features listed in this crosstab are ordered by feature importance—the most impactful features in determining the clusters are listed in order. Features used for clustering are marked with an asterisk.

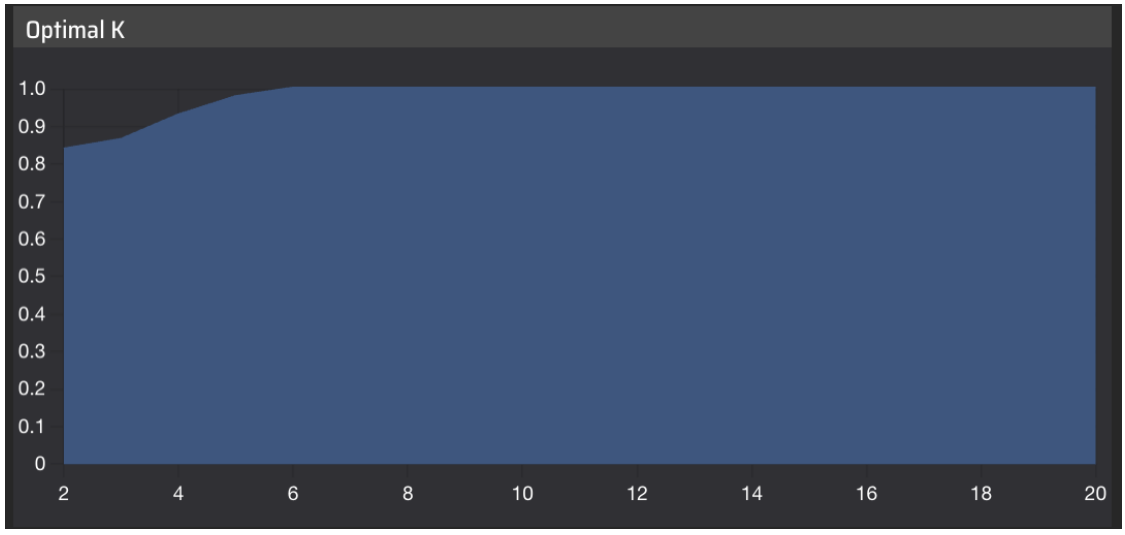
3. If you selected **Automatic Clusters,** you will get a third file that demonstrates the numbers used to select optimal K. This can be used to draw a line graph in the Stage screen. This dataset will be named [originalfilename]_enriched_ [# of enrichments applied]_cluster_scores.csv

The tabular form of this data will show a score per each potential number of clusters. The cluster number with the highest score is **Optimal K.** Optimal K is also noted in the tabular, as shown below. In this case, optimal k is 6.

N clusters	Score	Optimal k
2	0.8374985795881578	
3	0.8631273901276854	
4	0.9283475258309992	
5	0.977229545116512	
6	0.9999999999318548	optimal
7	0.9999999999318548	

To create the silhouette graph in the stage, select the above file. In the Inspector tab, select Line graph, with X axis set to n clusters, and the y axis set to score.





The resulting line chart will have its highest point at **Optimal K.** In the case above, optimal K is 6.

Please note that if you manually choose a number of clusters, this output will not be created, as you chose K instead of using an algorithm to determine optimal K.

You can find these files in the **Data** screen in your project folder. You can view these files in the **Studio**, and download them from here as well.

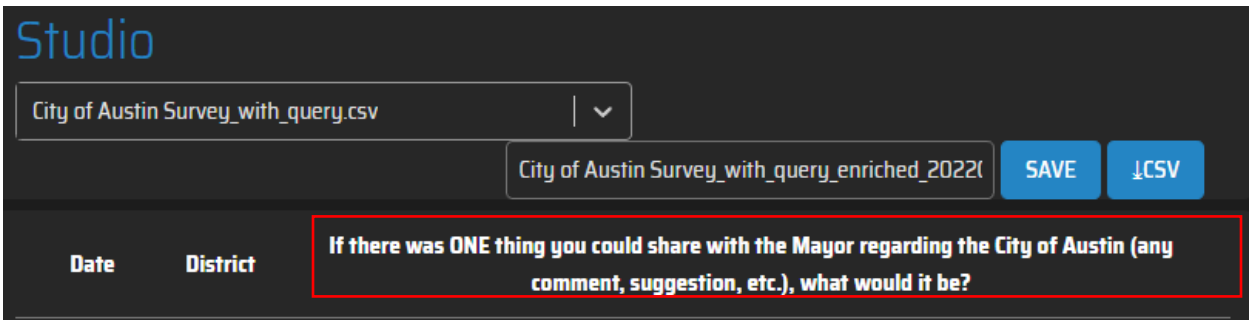
You can re-run Persona Segmentation with different settings if desired.

Term Frequency

The *Term* Frequency enricher can output two types of grams with their corresponding counts, 1) Unigrams and 2) Noun Phrases.

1) Unigrams – an n-gram that consists of a single item from a sequence (input text)
2) Noun Phrases – a part of speech pattern that contains nouns, adjectives, or verbs.

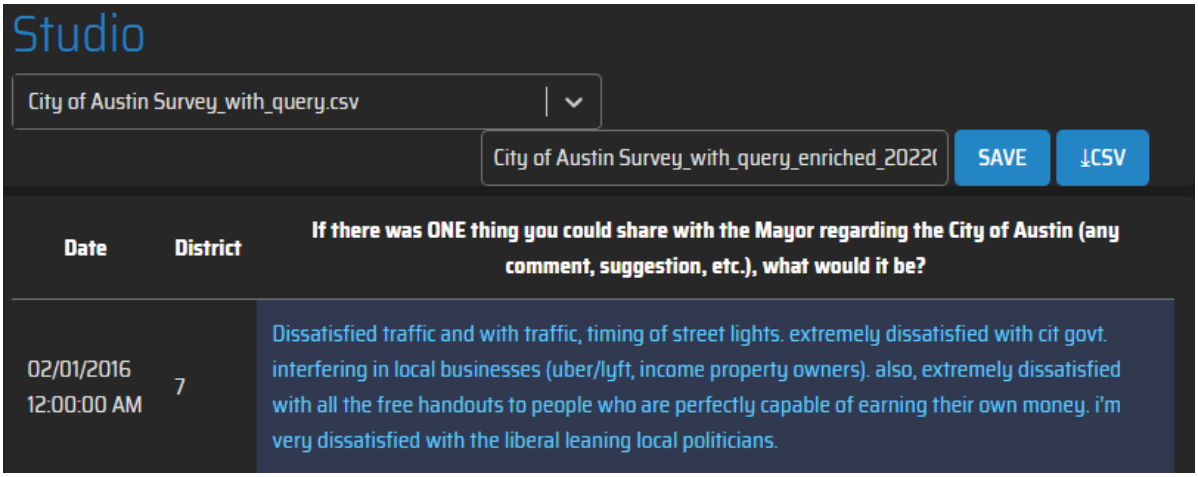
In addition, the enricher includes a parameter that can be set, that is, the use of *context-aware stop words*. This feature is beneficial when a *query* is set as the column name for the input data.



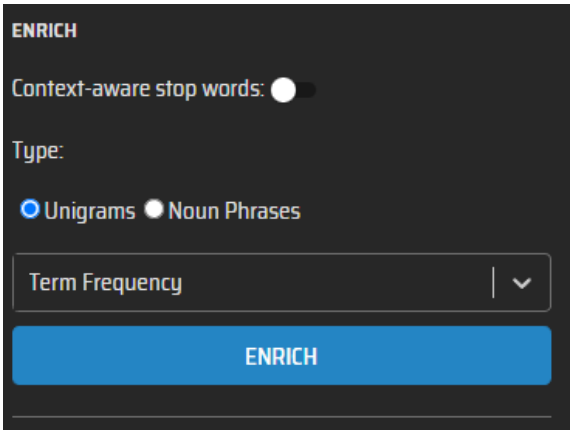
By enabling *context-aware* stop words, the enricher will remove words based on the *query* that may not be useful to the user such as *Mayor, city,* and *Austin.*

Usage

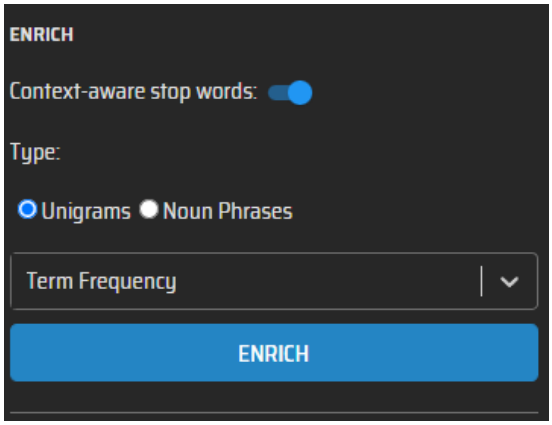
1. Select desired column



2. Select parameters based on the desired output

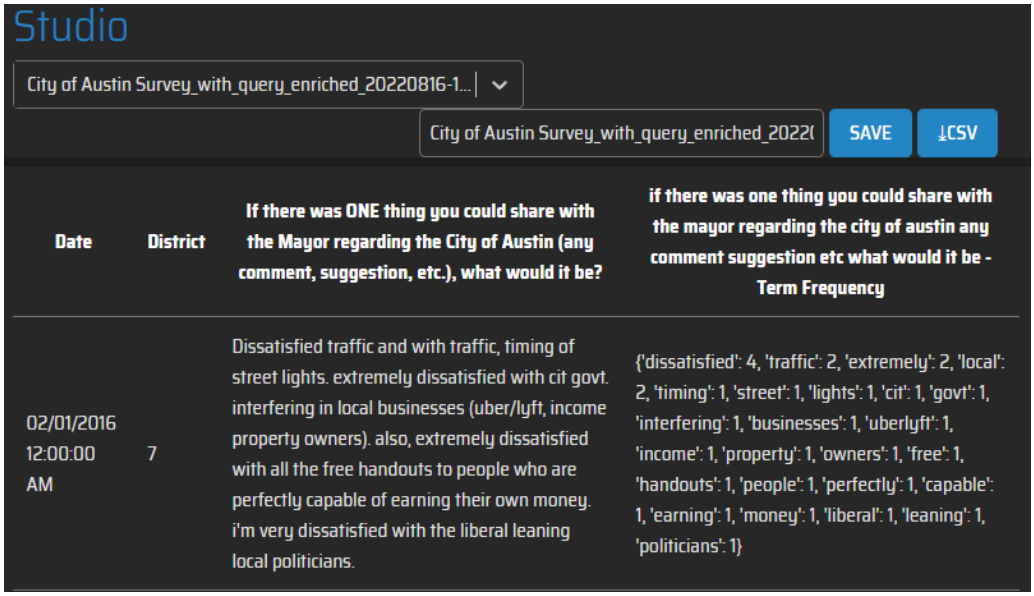


Example settings: ***Unigrams*** and **disabled** *context-aware stop words*

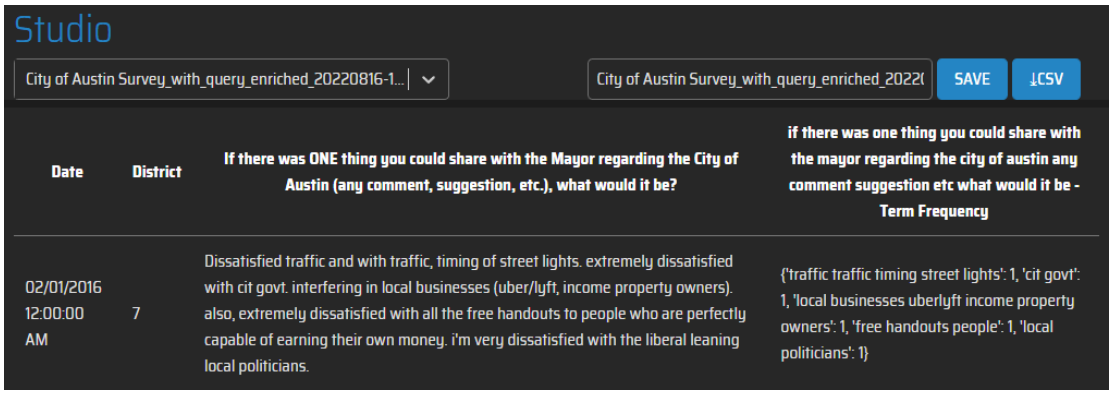


Example settings: ***Unigrams*** and **enabled** *context-aware stop words*

3. Click the *Enrich* button.



Example output of ***Unigrams*** checked



Example output of ***Noun Phrases*** checked

To visualize the term frequency data. Please see the Widgets section below.

Text Summary

The *Text Summary* enricher summarizes text, cutting down the original text by about 60% while still preserving its most salient ideas.

To use it, select the column that contains the text to summarize, select the enricher, and click the Enrich button. In cases where you're trying to summarize a document with various distinct sections, consider having each section as a separate row in the CSV in order to preserve the meaning of each, control which sections need to be summarized vs. which ones don't, and to prevent the merging of sections in the resulting summary.

JSON to CSV

The *Json to CSV* enricher aims at flattening a JSON file consisting of nested features at different levels into a CSV file with a tabular format where all the attributes of the JSON file are converted to different columns.

The enricher begins by flattening one attribute at a time. Whenever it encounters a nested JSON feature (parent), the nested attribute (child) is flattened and added as a new column to the root table. The parent attribute's name is then appended as a prefix to this new attribute. This prefix addition is applied to any attribute at any level of nesting, including cases where the parent attribute itself is nested.

Note: A tilde (~) is used to separate the parent attribute name from the nested attribute names during prefixing. Hence, it is imperative to make sure that the existing features in the input JSON file don't use '~' in their feature names.

To use the enricher, simply select the enricher and click on the '**Enrich'** button to apply it.

> **Issues with some JSON files:** Not all JSON files are created equal. If the platform fails to convert a JSON file, please contact us and we will assist you.

Custom enrichers

Creating a custom enricher

You can create your own enrichers using your own data and choosing to either train a specific base model or use the AutoML feature to let the platform choose the best algorithm for you automatically. The following methods are available for training a custom model:

Method name	Description
Category (M1-M2)	Use this method to categorize data. Both features and target variable must exhibit categorical values. The M2 version transforms empty values, while M1 does not.
Forecasting	Use this method to create predictions that output a quantifiable value, like revenue, temperature, distance, and time for example.
Text Classification (M1-M5)	Use this method to classify text using two or more classes. M1: optimized for unbalanced datasets. M2: optimized for balanced datasets for binary classification and reducing false negatives. M3: unoptimized and uses a different algorithm from M1 and M2 methods. M4: optimized for social media content and speed. M5: optimized for social media content and accuracy but slower speed than M4.
AutoML	Use this method to make predictions on tabular data and give control to the platform in choosing the best performing machine learning algorithm to make

	binary or multiclass classifications on the target variable. The platform tries Gradient Boosting, Instance-Based Learning, Deep Learning, and Ensemble Learning methods. AutoML accepts numerical, Boolean, categorical or textual data.

Creating a custom enricher:

1. Upload a dataset that will be used to train and test the model.
2. Navigate to the Models screen.
3. Select the options in the inspector to train the model.

   Note: when using any of the text classification methods, you will be asked to select the input data that will be used to test the trained model in the dropdown labelled *"Select column for test analysis"*.
4. Provide a name for your model.
5. Click the "Create Model" button.

At this point, the platform will begin training the model. Once the training is complete, a "Trained" label will be shown next to the model. At this point, you must reload the screen to trigger the testing of the newly trained model. Once tested, the platform will provide performance metrics in the Inspector section when selecting the model.

Model statuses

Status	Description
Initialized	Initial state
Training	Model in training
Trained	Model training complete
Testing	Trained model being verified
Failed	Model training failed

Training a model using AutoML

The AutoML feature on the platform is a step towards automation that lets the platform choose the algorithm or model to be trained for tabular classification tasks. This feature enables the user to upload tabular datasets consisting of different data types. The feature supports the following data types - Numeric, Boolean, Categorical or Text. The platform can ingest the data while making sure to transform it into a format appropriate for the Machine Learning algorithms to process the data.

Note: This feature is available only through the Models screen.

Creating a custom AutoML enricher via the Models screen:

1. Upload a dataset that will be used to train and test the model.

   NOTE: The training data should exclude columns/features which are irrelevant such as unique identifiers, time-related features, features that won't be available at the time of inference, etc.

2. Select the AutoML option from the model drop down Menu.
3. Choose the target variable that you want to train your model on.
    a. The platform automatically detects whether the task is one of binary or multiclass classification based on the number of classes in the target variable.
4. In the Performance requirement drop-down menu, choose the appropriate target performance metric to optimize the model based on your preference:
    a. Fewer False Negatives -> Optimizes the ML models for better Recall scores.
    b. Fewer False Positives -> Optimizes the ML models for better Precision scores.
    c. Default Metric (Balanced) -> Optimizes the ML models for better F1 scores.
5. Click the 'Create Model' button.

The platform will begin training multiple ML models and choose the best performing on to serve for inferencing. Once the training is complete, the platform will test the newly created model and provide performance metrics for the best performing model in the Inspector section of the Models screen.

You will also be able to download the feature importance rankings for all the features that were considered important during training by clicking the **'Download Train Features'** button. Upon clicking the download button, a zip file containing two files will get downloaded:

- train_features.txt - This file provides the list of features that were fed to the AutoML training after performing preprocessing.
- Optuna_MODEL_NAME_feature_importance_rankings.csv - This file provides a list of top-ranked features considered important by the best-performing model (MODEL_NAME is placeholder for the best performing model's name) post training.

Note: The performance of the AutoML model is correlated to the usability of the datasets that are being fed for training. Datasets involving complex datatypes or unclean data might require tailor-made preprocessing functions to better handle the data and achieve higher accuracy.

Any model trained as part of AutoML will also be available for retraining in the Ongoing Learning module.

Once the model is trained, you will be able to use it through the web interface, or the API. To access the model via API, follow these steps:

1. Query /studio/get_local_enrichment_methods/ to get a list of custom models, locate the one you're looking for, and copy its id.

Ongoing learning

The platform enables users to provide feedback on the performance of custom enrichers and uses this feedback to improve the enrichers over time. This functionality is done by providing corrections to misclassifications and using those corrections to retrain new versions of a custom model.

Adding Data to the Master Enriched Dataset

When you create a custom model using AutoML, three files will be created.

1. A master training dataset: This is the data that was used to train your model.
2. A master testing dataset: This is the data used to test your model, which is used to create the confusion matrix.
3. A master enriched dataset: This is data that you have enriched with your custom model. Data in this file can be used for correction and retraining.

There are two ways you can generate predictions to be used for model feedback and training Both of these methods will append your predictions to the master enriched dataset:

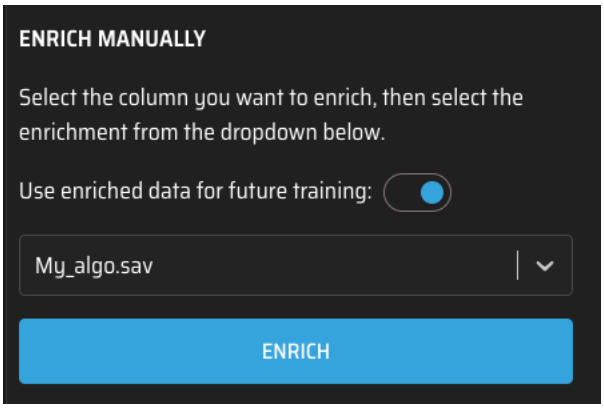
1. Via API

When sending your data for enrichment, use the below parameter set to true.

```
outputToMaster=true
```

2. Via front-end

Upload a dataset to enrich with your custom model. In the Studio, select your model from the enrichment dropdown. To include this data in your model's master enriched file, set the toggle for "Use enriched data for future training" to its blue active state on the right as shown below.



To output a separate file that will **not** be used for training (i.e. to assess model performance), set the above toggle to its gray inactive state to the left.

Once you have added predictions from your model to its master enriched dataset, you can begin correcting the outputs, and using this information to retrain your model.

Correcting misclassifications

The platform allows analysts to "correct" the mispredictions that the platform might make. The platform will then learn from these corrections and keep its AI trained with the latest data.

NOTE: You must first generate predictions using the model before you are able to correct predictions. If you try to access the corrections screen before generating predictions, you will see an empty list, since no predictions have yet been generated.

To make corrections, go to the Models screen, locate the model you'd like to make corrections for, click the "Correct" button, and follow the instructions below.

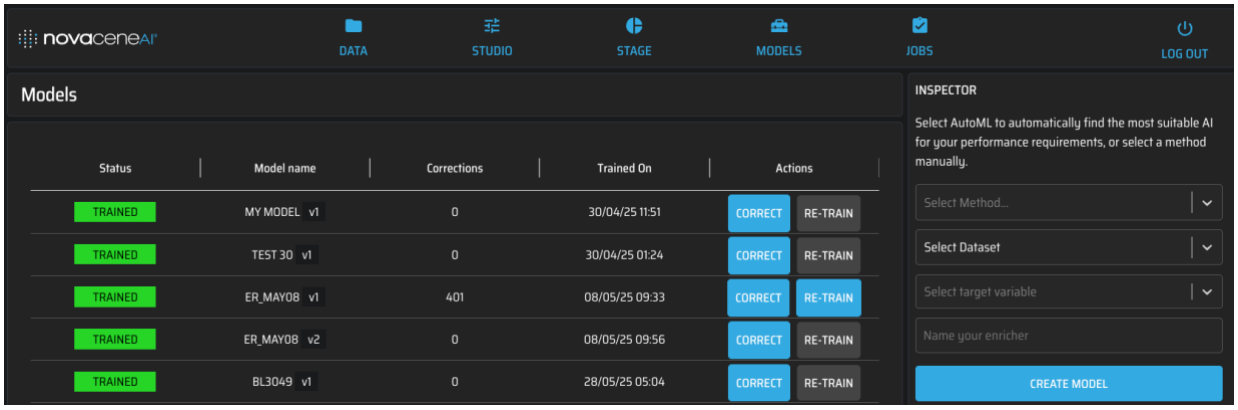


Figure 9: Accessing the corrections interface

1. **Making corrections:** You will be taken to the Studio, where you will be able to correct mispredictions. To do so, locate the row you'd like to correct, (1) click the pencil icon, (2) select the correct value from the drop-down menu, and (3) click the save icon. Repeat these steps for each correction you'd like to make.

> **NOTE:** You must first generate predictions using the model in order to be able to make corrections. If you try to access the corrections screen before generating predictions, you will see an empty list, since no predictions have yet been generated.

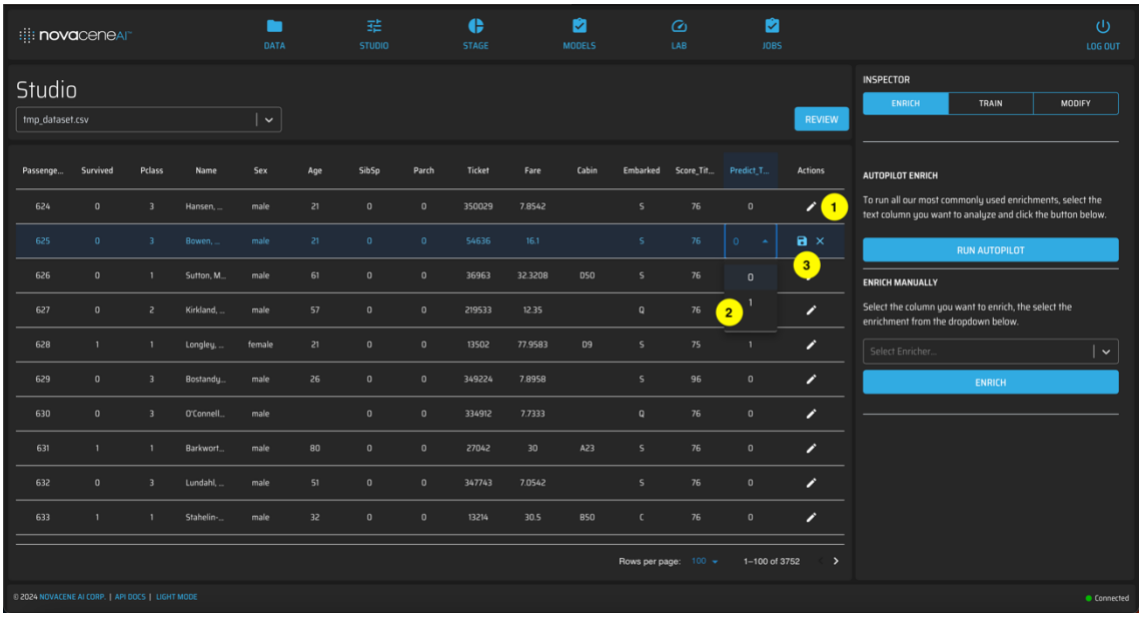


Figure 10: Correcting mispredictions

2. **Reviewing corrections:** Once you're done making corrections, you will need to merge those corrections back into the datasets that will be used to train a new version of the model. To do so, click the "Review" button to review your changes, and resolve any conflicts that may have been introduced during the correction process.

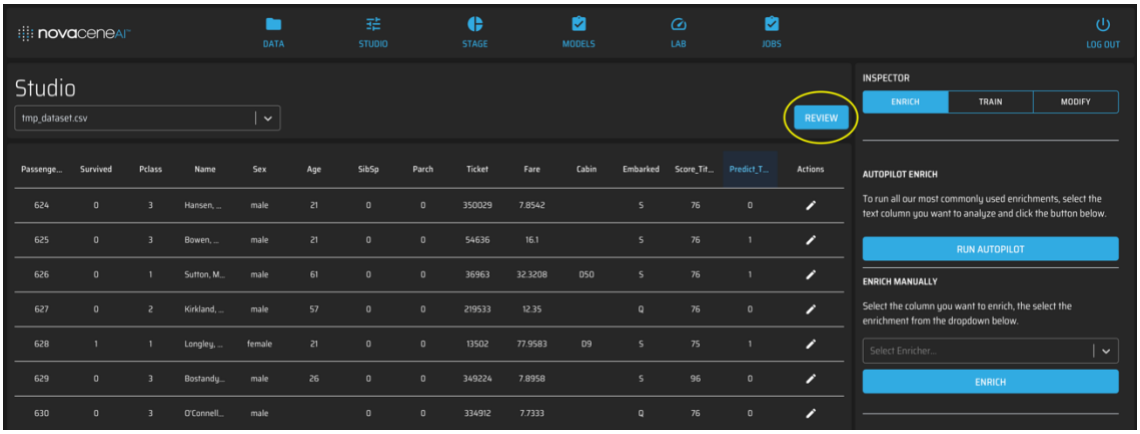


Figure 11: Accessing the Review interface

3. **Resolving conflicts:** The Review screen will show you the rows that you have corrected. The rows with green text indicate changes that can be merged without issues. The rows with red text show the rows where the corrections contradict existing identical cases. Change any values you would like to update and proceed to merge your changes.

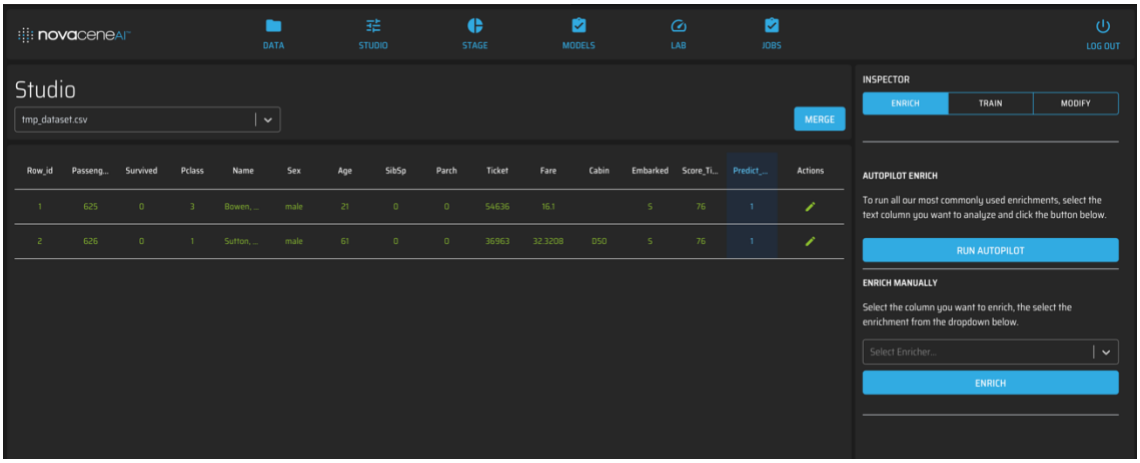


Figure 12: Reviewing changes

4. **Merging changes:** Once you're happy with all the changes, click the "Merge" button to merge those changes into the datasets that will be used by the retraining process.

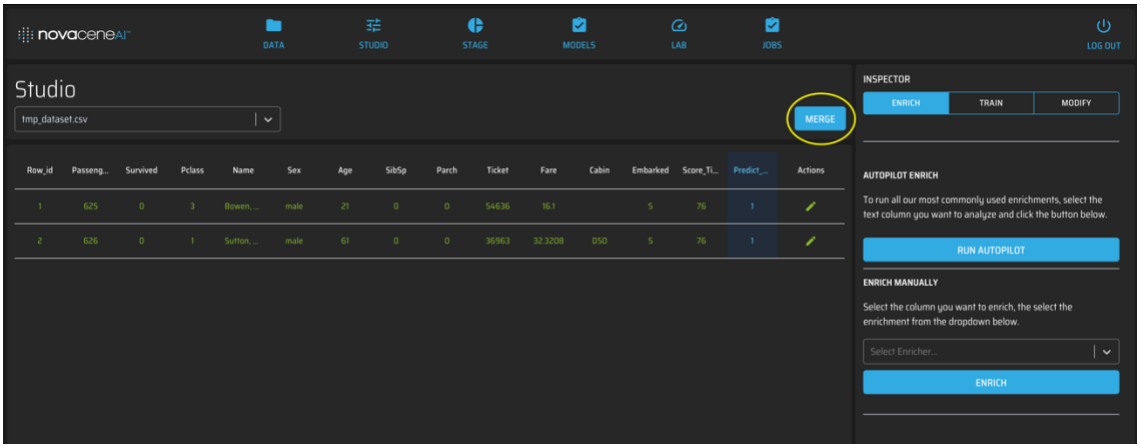


Figure 13: Merging changes

Retraining enrichers

Once enough corrections have been accumulated, the “Re-Train” button will become active. Clicking the button will trigger the training of a new version of the model. Once the new version is trained, you will be able to compare the performance of the new model against that of the old one and choose whether to start using the new version or the previous one.

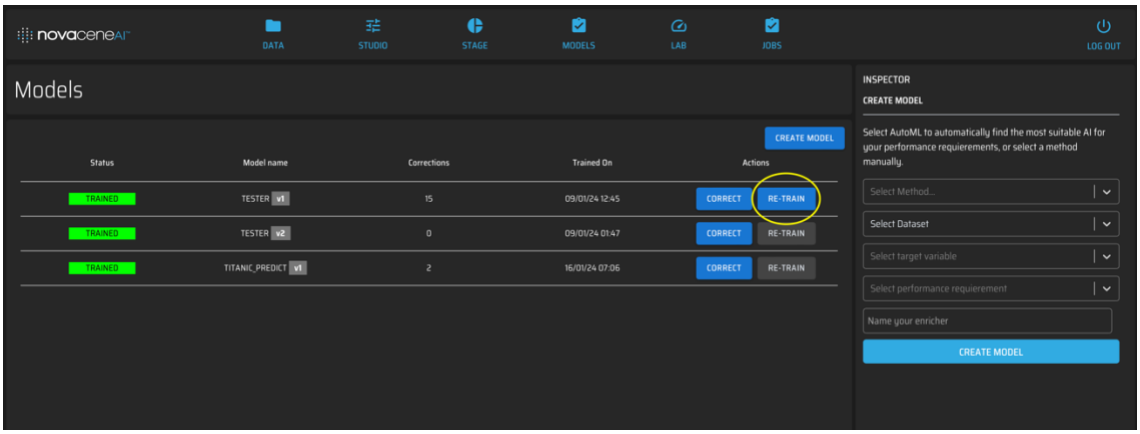


Figure 14: Triggering a retraining job

Deleting custom enrichers

Depending on where you created the custom enricher, you’ll need to follow different set of instructions to delete it.

1. **Enrichers created via the Studio screen:**
    a. Open your browser and go to your Novacene instance, /admin/ (e.g., my.novacene.ai/admin/)
    b. Go to DATASETS > **Enrichment methods**

c. Select the model(s) you want to delete
d. Select the option to "Delete..." from the dropdown and click the "Go" button
e. Log in to the server where you deleted the model
f. Cd to /opt/web_app/{env}/data/MLmodels/
g. Remove the corresponding .sav files

2. **Enrichers created via the Models screen:**
   a. Go to https://{subdomain}.novacene.ai/admin/models/trainingmodelmanager/
   b. Select the model(s) and in the “Action” menu, select “Delete selected model(s) and associated master datasets”, and click the “Go” button.
   c. Verify that that deleted models no longer appear on the Models list.

Deleting custom enrichers using the API

You can delete custom models via the API as well.

Measuring enricher accuracy

The NovaceneAI Platform allows users to test the accuracy of their models using a confusion matrix. Confusion matrices are available for binary and multiclass classification models.

Generating a confusion matrix

Please contact Novacene for instructions.

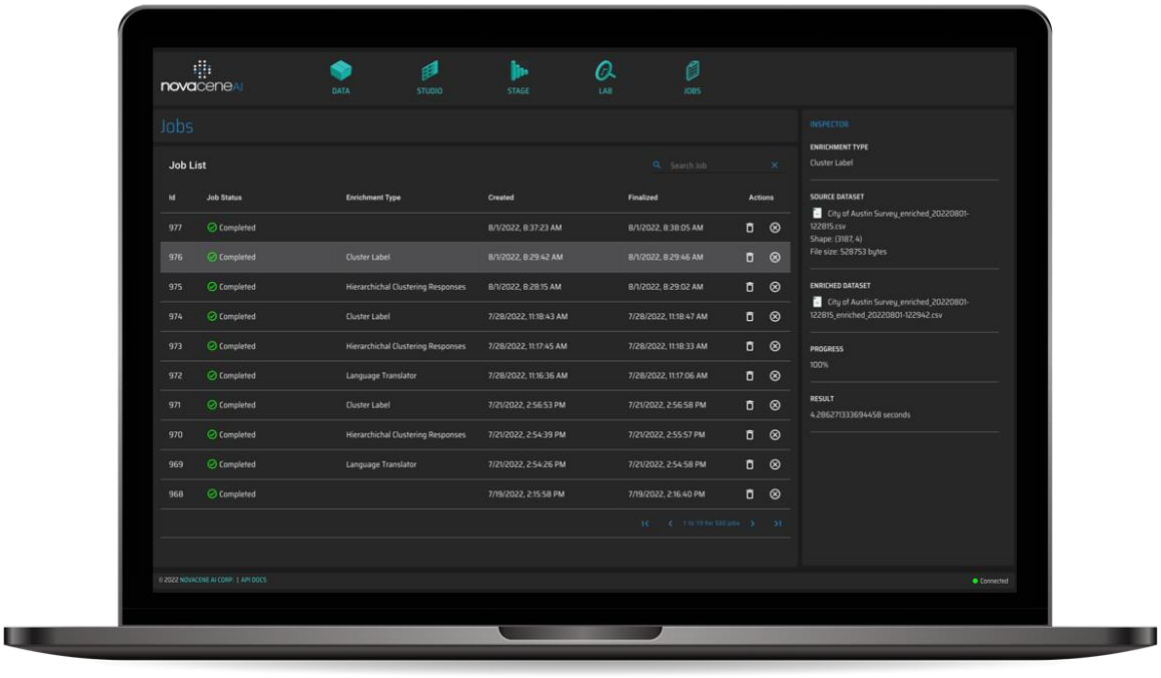
Deleting a confusion matrix

To delete all confusion matrices (both binary and multiclass):

1. Go to /admin/datasets/dataset/
2. Select the datasets:
   a. *models_confusion_matrix.csv*
   b. *models_classification_report.csv*
3. In the dropdown next to “Action:”, select “Delete selected data sets” and click the “Go” button
4. Confirm that you wish to delete the datasets
5. All the confusion matrices will be deleted

Monitoring enrichments

The application allows you to monitor the progress of both enrichments and training jobs. You can monitor both types of jobs on the Jobs screen.



Monitoring progress of enrichment jobs

In addition to being able to monitor enrichment jobs on the Jobs screen, you can also see the progress on the Datasets screen. To see the progress, locate the dataset being enriched. You will see a progress wheel overlaid on the dataset. This progress wheel will show progress from 0 to 100, denoting the percentage of completion of enrichment.

FAQs about monitoring

Why some enrichment progress indicators show a spinner instead of a percentage? Predicting the time of certain enrichments can depend on a multitude of factors, including how the data is processed, and how frequently certain tasks need to be retried.

Why do progress indicators sometime update frequently and sometimes in larger chunks? Because datasets are divided in chunks of 100 records to be processed at a time. In the case of this dataset of 400 rows, the update frequency will be every 25% (400 rows / 100 chunks = 4 intervals = 25%)

What do the different job states mean? The table below provides descriptions for each job state:

State name	Description
UPLOADED	Dataset uploaded
INITIALIZED	Job received by server
ENQUEUED	Job is in queue to be run
RUNNING	Job is currently being processed
COMPLETED	Job completed successfully
FAILED	Job failed
CANCELLED	Job cancelled by user

Notification of completed jobs

When a job finish running, you will hear a notification. This is intended to draw your attention on the finished job in case you are away from the NovaceneAI Platform when this happens. Note that to hear the notification some browsers may require that you allow for this to happen.

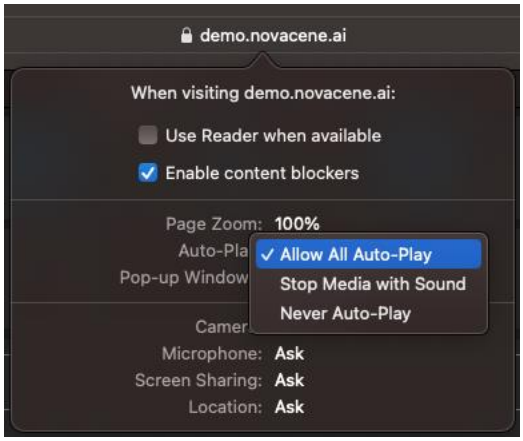
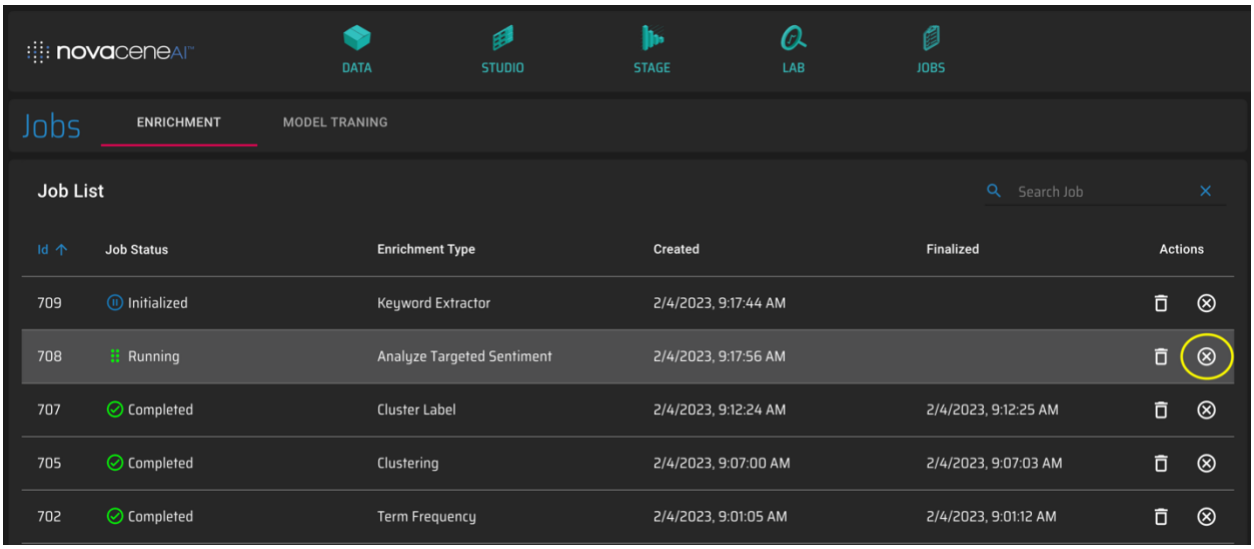


Figure 15: Some browsers require that you set permissions for notifications to be heard

Cancelling jobs

To cancel a job, click the Cancel icon corresponding to the running job.



Why does a cancelled job continue to run, or re-appears as Running on the Jobs screen? Some enrichers cannot be cancelled. You can either let the enricher finish, or contact us and we will cancel the job for you.

Data visualization

Use the *Stage* screen to visualize your data.

In this section:

1. Stage overview
2. Auto-visualization
3. Creating reports
4. Widgets

Stage overview

The Stage screen enables you to create reports comprised of data visualization widgets. You can add as many widgets as you want, size them, and lay them out in whichever arrangement makes the most sense to the analysis you are performing.

Auto-visualization

When you load a dataset in the Stage screen, the platform will automatically show widgets for each of the columns from your dataset that are likely to benefit from a visualization. The platform will automatically determine the type of visualization widget to use based on each column's data.

The platform will not produce automated visualizations for a given column when the column contains:

- a single identical value across all rows
- distinct values in every row
- only blank values

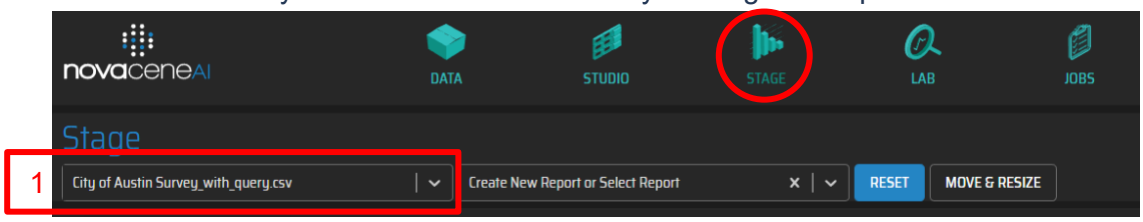
Creating reports

Reports are collections of widgets laid out in a specific arrangement. You can create as many reports as you want, and you can save them to retrieve them later. Saved reports will remember the layout, any applied filters, and the zoom and pan settings for each widget.

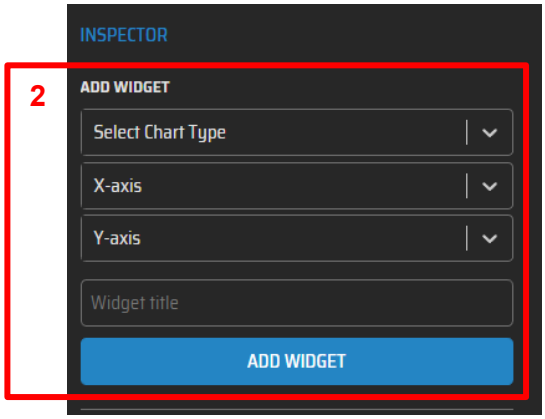
When you create a report, the report is associated with the dataset you used to create it.

To create a report:

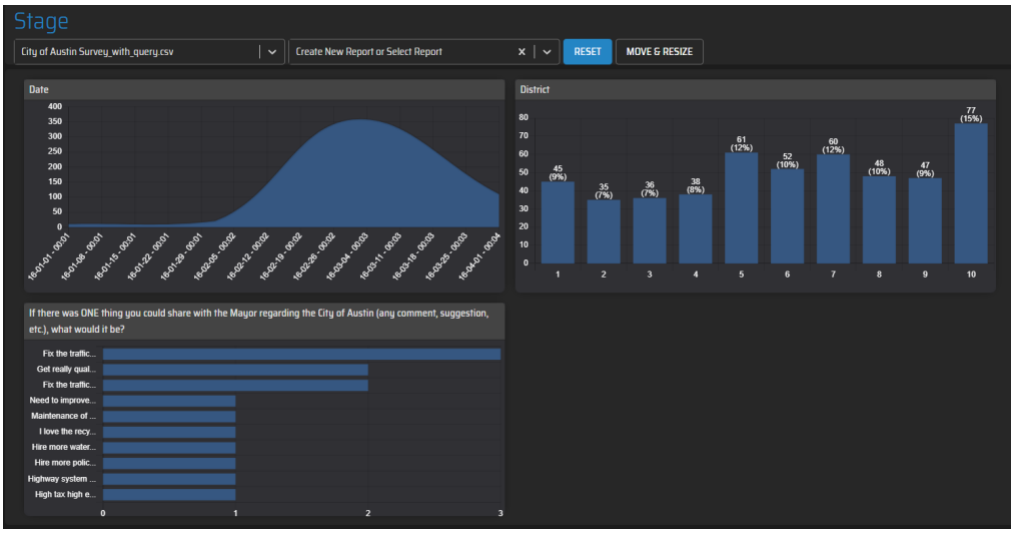
1. Load the dataset you would like to visualize by clicking the drop-down arrow.



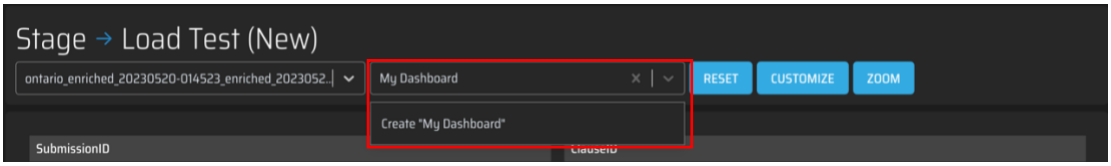
2. The application will automatically generate visualization widgets for your data. For more information, consult the Auto-visualization section.
3. Remove any widgets you don't need.
4. Add your own widgets as necessary. To add a widget:
    a. From the Inspector window on the right-hand side, use the *Select Chart Type* drop-down to choose a chart type for to use.



b. Subsequently, select the desired *X-axis Y-axis* and *Widget title*. Once the *Add Widget* button is clicked, the chart will be displayed in the main *Stage* area as shown below:



5. Customize the report to resize and re-arrange widgets in whatever layout makes the most sense for your analysis.
6. Use the “Create New Report or Select Report” drop down menu to name your report, and click the *“Create ‘My report name’”* option.



Troubleshooting saved reports

If you get an error when loading a saved report, make sure that:

- The report you are trying to load was built using the current dataset, or a dataset that contains the same columns as the current dataset
- If your report was filtered, make sure that you did not remove the widget that was used to filter the report.

Widgets

Data visualization widgets are charts that help you visualize your data, so it is easy to understand insights and spot trends. Each widget provides a set of options, including Chart Type, X-axis and Y-axis, Group-by, and Widget title.

Chart types

Bar (Vertical)

A vertical bar chart.

Bar (Horizontal)

A horizontal bar chart.

Bar (Diverging Stacked)

A horizontal bar chart where the X axis represents a value range which neutral point is centered, and the Y axis represents a breakdown by which values are reported. For example, you can visualize the distribution of sentiment along the X-axis and provide a breakdown on the Y-axis (such as by topic). This chart allows you to gain insights into the distribution of sentiment across different topics. Positive opinions are represented on the right side of the X-axis and negative opinions on left. The center of the chart indicates neutral or balanced sentiment.

Note that the Group-by option cannot be applied to this chart.

Here are sample settings to create a chart:

x-axis: sentiment
y-axis: cluster label
grouped by: [leave empty]

Doughnut

A circular chart showing the percentage of share of values in a range. Similar to a pie chart.

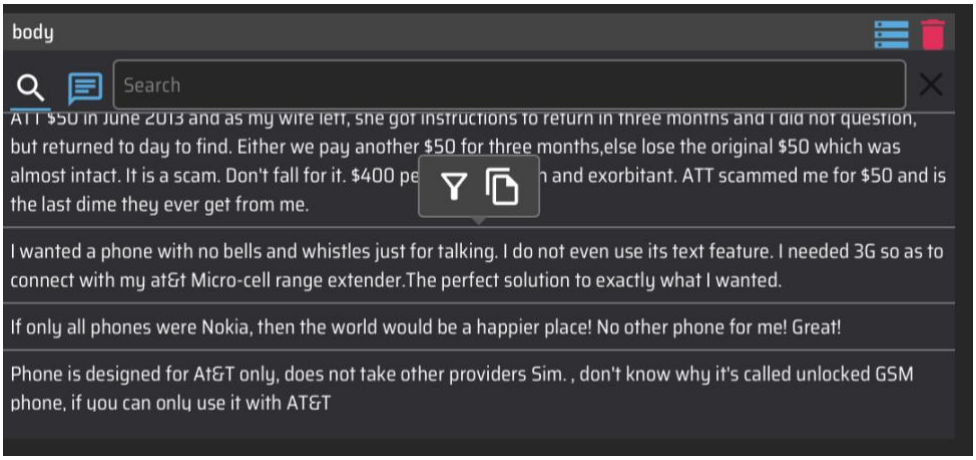
Line

A continuous line chart.

List

A list of items. This widget is a good choice for displaying text, or information you'd prefer to read. Items displayed in the list are searchable. The buttons that appear on hover shown in the

image below allow you to copy text in the list widget, and filter the stage by a particular item in the list widget.



Additionally, questions can be asked against the items in the list using Generative AI. This function will use all the items displayed in the list as context for the Generative AI, along with your query.

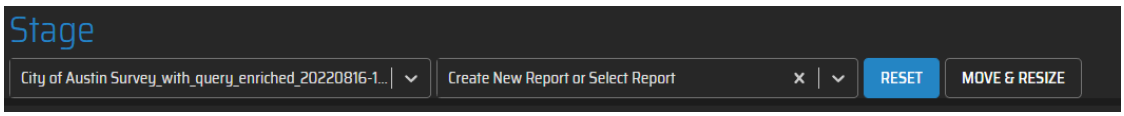
The stage will automatically create a List widget to display columns with text.

Word Cloud

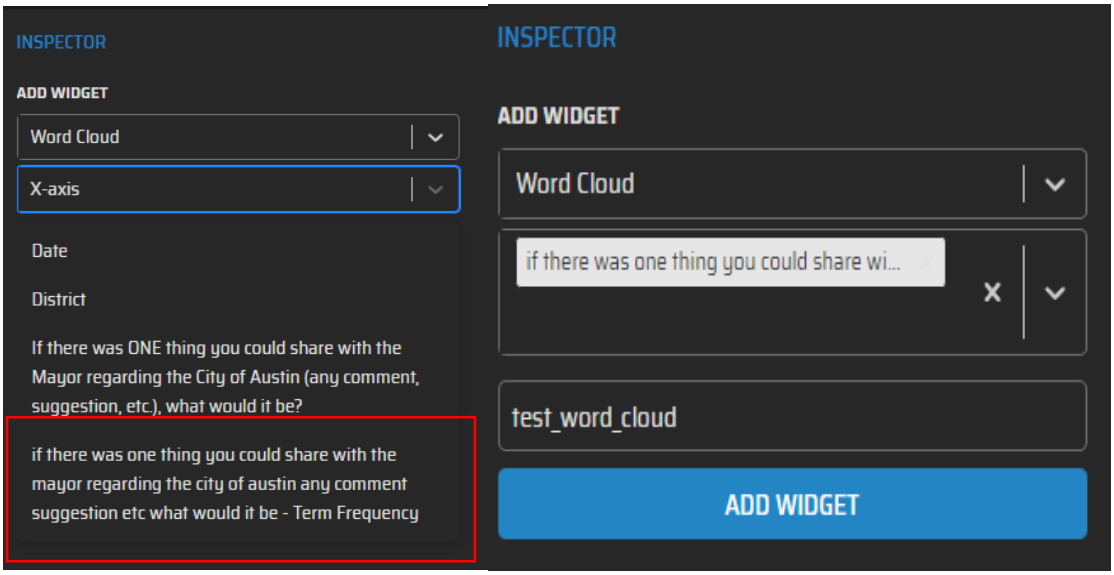
A Word Cloud is a group of words or phrases where the size of the items represents their frequency. The Word Cloud chart type requires the Term Frequency enricher to applied on the Studio page before the Word Cloud tool can be used.

The process below outlines the steps to use the *Word Cloud* visualization tool following the steps described above.

1. Go to the Stage screen
2. Select the dataset



3. Select the Chart type called Word Cloud. Please note that the selected column name includes *Term Frequency.*



4. Press the *Add Widget* button.

5. The *Word Cloud* is displayed on the main *Stage* page.

6. The image of the *Word Cloud* can be downloaded using the arrow button as shown in the picture above.

Net Sentiment

Net Sentiment shows the overall sentiment score recorded over a period of time, broken down by theme.

To create a Net Sentiment visualization:

1. Select the “Line” chart type
2. For the X axis, select a column containing date or time data
3. For the Y axis, select the “Net Sentiment” option, and then the column that contains the Themes
4. Name your widget and click the “Add Widget” button

Group-by

Use the group-by option when you want values to be grouped by values of another column. For example, let’s say you have a dataset with 1,000 reviews. When you apply clause extraction, the platform will add rows your dataset to accommodate each clause it its own row. To preserve the relationship between a clause and the original submission, the platform adds a column called *SubmissionID.* Now let’s say you want to create a widget of type vertical bar where you can see the number of unique submissions by age range. You would then select your **Bar (Vertical)** chart type, ***Age Range*** for the X-axis, ***count*** for the Y-axis, and ***SubmissionID*** as the group-by option.

You have two options to select from for Group-By:

Normalized

The normalized group-by method will count how often the selected variable occurs per variable selected as the group-by value. For example, if you are creating a horizontal bar chart and you select Count for the x axis, Themes for the Y-axis, and SubmissionID as the group-by option, you will see a count of each theme per SubmissionID. Note that the total count of themes will not always match the total count of unique SubmissionIDs, as each submission may have more than one theme associated with it.

Co-occurrence

The co-occurrence group-by method will count how often unique combinations of the selected variable occur per variable selected for the group-by value. For example, if you are creating a horizontal bar chart and you select Count for the x axis, Themes for the Y-axis, and SubmissionID as the group-by option, your chart will have a count of every combination of themes in a single SubmissionID. Note that the total count of theme combinations will not match the total count of unique SubmissionIDs, as each submission may have more than one theme associated with it.

Widget title

The widget title option allows you to give your widget a name.

Chat with your Data

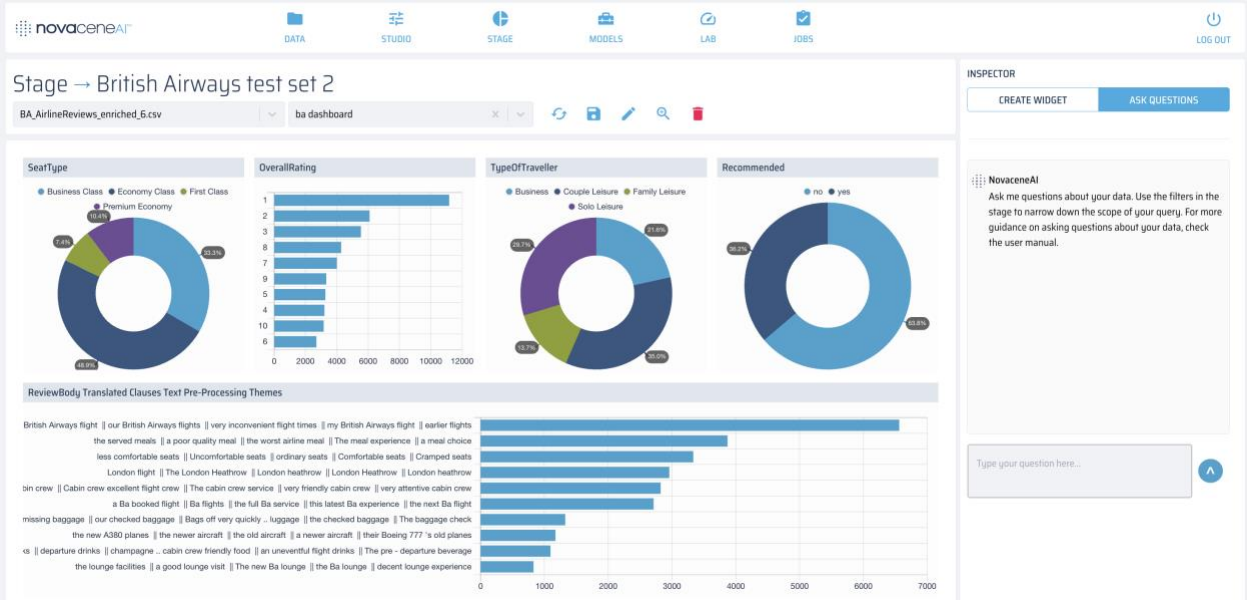
Using the “Ask Questions” Feature

Release 3.6.2 includes a new version of the Ask Questions feature, which can be controlled with a toggle. The user experience of both versions is very similar, but the mechanism and data inputs are different.

With the **new** version, we’ve enhanced Ask Questions to deliver smarter, more adaptive responses. The new Ask Questions leverages contextual understanding to dynamically identify the most relevant columns for each question, enabling broader use cases and delivering faster, more accurate, and more relevant insights. As before, it remains fully integrated with your dashboards, interactively filtering visualizations to reflect the data underpinning each answer.

The **original** Ask Questions feature requires that you run text embedding or clustering on the column you’d like to ask questions about. These steps happen automatically when you run Autopilot. You can also select these enrichments from the Enrich Manually dropdown in the Studio.

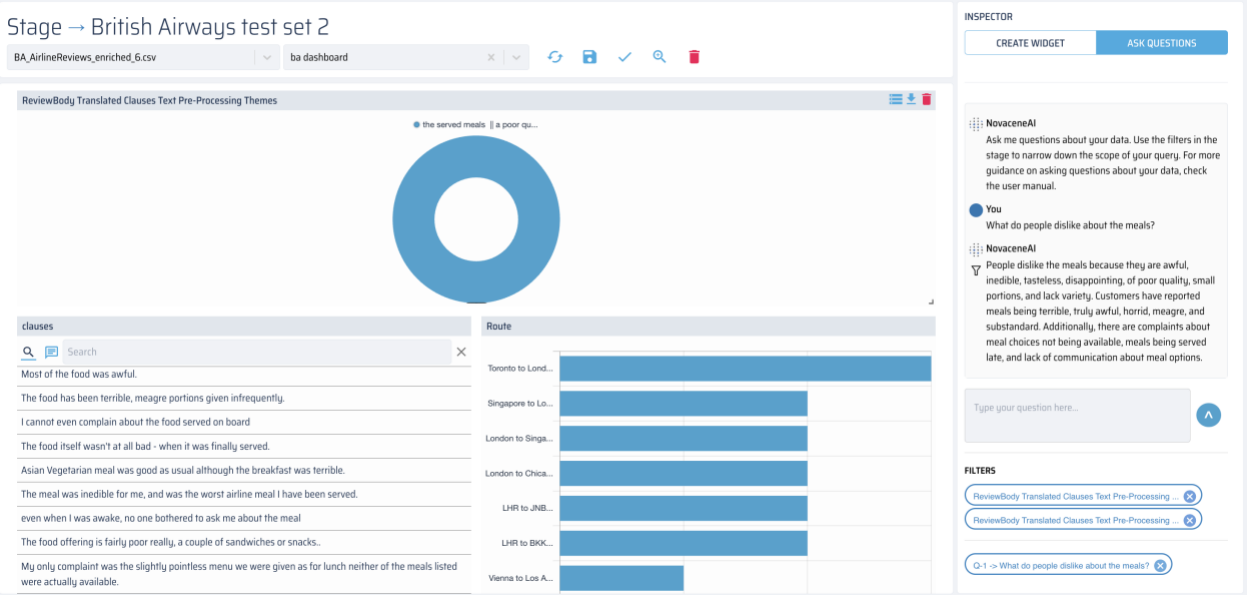
In the stage, you can collaborate with generative AI to analyze your data.



Now, when you ask a question about your data, you will be able to view exactly what pieces of text in your dataset were used to generate the answer. These pieces of your dataset are also visualized in the dashboard.

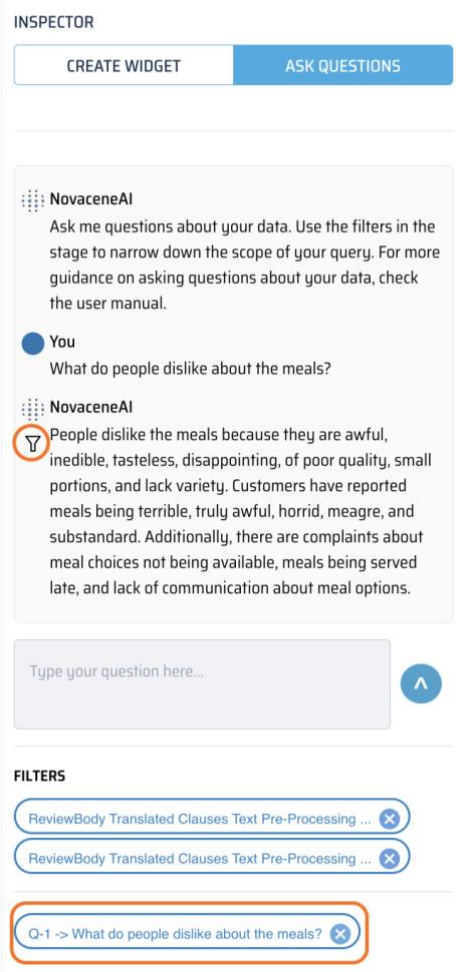
For example: You have a dataset of a survey about experiences on an airline. You're interested understanding how the meal experience can be improved. So, you click on the “Meal Experience” bar in the “Themes” chart, and then in the Sentiment doughnut chart, you click on “Negative” to filter your dashboard.

Now, we can ask in chat-- “What do reviewers dislike about the meals?”



Next, an answer is generated in the chat, and the dashboard is **filtered** to show you the exact text that was used to generate the answer in the text box. This allows you to ensure that generated answers are accurate, and to potentially find patterns in the visualization widgets. Take a look at the generated answer above, and the List widget with the clauses used to generate this answer.

You can add or remove these filters as shown in the image below.



You can add or remove Question filters using the filter toggle next to the generated answer, or remove them by hitting the “X” on the filter with your question text.

Tips for writing good questions:

- **Collaborate** with genAI by using the filters to help narrow down the scope of your query. For example, if you are interested in negative comments, use the negative sentiment enrichment filter. Let’s say we’re interested understanding why there are so many negative comments about the meals in this dataset—we'd use the topic filter “Meals” and sentiment filter “negative” to get the most relevant data for the genAI.
  You get the best results when you work together.
- **Verify** your generated answers quickly and easily using the list widget. All generative AI is vulnerable to hallucinations. With this feature, you can see exactly what text has been sent to the AI to answer your question in your dashboard.

Supported systems

The application is programmed to work on most modern browsers. For the best experience, please use Chrome or Safari on PC or Mac.

Contact us

To contact Novacene, send an email to support@novacene.ai

Novacene AI Corp.
123 Slater Street, Suite 610, Ottawa, ON K1P 5H2, Canada | (800) 717-0814 | www.novacene.ai

Appendix: Built-in enrichers

The following is a partial list of built-in enrichers available on the application. Additional built-in enrichers may be described in section: Available built-in enrichers.

Note: Some enrichers require a previously-clustered datasets; these enrichers have been marked with a **superscript ([1]).**

Enricher name	Column suffix	Description
Clause Extraction	Clauses	Extracts clauses from input text.
Multi-Column Clause Extraction	Clauses and Columns	Extracts clauses from input text, and merges all enriched columns into two: clauses and column.
Cluster Label[1]	Themes	Extracts the most central sample from each cluster.
Cluster Sampler[1]	Cluster Sampler	Outputs a set of the topmost representative samples in a cluster.
Cluster Summary Generator[1]	Cluster Summary	Generates a summary of the contents of the cluster using the OpenAI GPT API. Note that to overcome limits imposed by the API, the summary is based on a subset of the contents of each cluster. Requires that the input data has already been clustered.
Cluster Theme Extraction[1]	Themes	Outputs key phrases that describe the contents of a cluster. Extractive method.
LLM Cluster Theme Extraction[1]	Themes	Outputs generated labels based on samples of each cluster. Generative method, uses your environment's LLM.
Clustering	Cluster	Clusters text by grouping similar content together into categories (Does not include an unclustered category). Increasing the threshold produces a higher number of clusters with more members, while decreasing the threshold produces a lower number of clusters with less members.
LLM Thematic Analysis	Themes	Outputs theme labels per row. Can consider context, known themes, and keywords added by the user.
Custom Theme Classification	Themes	Uses an LLM to sort text into user-defined themes. Generative method, uses your environment's LLM.

Emotion Analysis	Emotion Analysis	Classifies the emotion of the text as: *Anger, Annoyance, Anticipation, Compassion, Concern, Confusion, Contempt, Disappointment, Discomfort, Disgust, Embarrassment, Fear, Frustration, Gratitude, Happiness, Hatred, Joy, Neutral, Relief, Sadness, Satisfaction, Surprise*
HTML Content Cleanser		Removes HTML tags from text.
Ideas and Comment Classifier	Idea and Comment Classifier	Classifies input text as an *idea* or *comment.* Useful for idea extraction and analysis.
Image Quality		Assigns a score based on image quality factors such as contrast, sharpness and more. (Lower scores equal higher image quality).
Language Translator (Online)	Translated	Detects non-English text and translates into English. (Supports French, Spanish, and Chinese).
Language Translator (Offline)	Translated	Detects non-English text and translates into English. (Supports 25 languages, see Appendix: Supported Languages).
Names Entity Recognition	NER	Detects and highlight people, places, organizations, and other known entities found in text.
Public Support Detector	Public Support Detector	Assigns one of five classes on a 5-point Likert Scale ranging from *strongly approve* to *strongly disapprove.*
Public Support Sampler	Public Support Sampler	Outputs the top 5 most disapproving samples in decreasing stance score.
Sentiment Analysis (Offline)	Sentiment Analysis	Outputs a *Positive, Negative,* or *Neutral* label for each sample. Score ranges from 0 – 1.0 and represents the confidence in the prediction. Can be run offline.
Sentiment Analysis	Sentiment Analysis	Outputs 3 columns: Sentiment, Emotion, and Sincerity. Generative method, uses your environment's LLM.
Social Media Content Cleanser	Cleansed	Removes special characters commonly used in social media updates such as hashtags, @mentions, and more.
Term Frequency	Term Frequency	Outputs the number of times a *unigram* or *noun-phrase* appears in an input text.
Thematic Analysis	Themes	Outputs a theme group number and a label. Users can select a maximum amount of

		themes. Works best on datasets with more than 150k rows.
Threats / Abuse	Toxicity	Detects threatening or abusive language.

Table 2: Built-in enrichers

[1] Requires a previously-clustered dataset.

[2] Order of source dataset columns matter.

[3] Number of source dataset column matter.

Deprecated enrichers

Enricher name	Description	Deprecation reason
Sentiment Sampler	Outputs the top 5 most negative samples in the dataset in decreasing order of negative sentiment score.	Ranking to be treated as a function on the visualization system and not as part of the classification algorithm.
Cluster Summary[1]	Performs a one sentence summary using extractive summarization.	Generative methods outperform this extractive method.
Analyze Targeted Sentiment		
FinTech Analysis		
FinTech Social Media Cleanser	Strips tweets from potentially distracting content such as broken URLs, special symbols, ReTweet (RT) tags, etc.	Replaced by HTML cleanser and Social Media Content Cleanser
Hierarchical Clustering Responses	Clusters text by grouping similar content together into categories. (Optimized for some types of survey responses).	Inflexible for different data formats
Quantum Classifier	Binary classification algorithm that runs on a Quantum backend.	
Hierarchical Clustering Social	Clusters text by grouping similar content together into categories. (Optimized for short texts or social media updates).	
Peer Clustering[2] [3]	Clusters records that share many similar attributes.	
Sentence Segmentation	Expands the input text by segmenting input text into separate sentences.	Replaced by Clause Extraction

Sentiment Pre-Processing		Replaced by Sentiment Analysis and Emotion Analysis.
Sentiment Sampler	[TBD]	Please contact us for instructions.
Sentiment Analysis (Retail)	Specific for retail data, classifies the tone of the text as being *positive, negative, or neutral.*	Replaced by Sentiment Analysis
Topic Modelling and Clustering M1	Clustering for text	Earlier versions of Clustering
Topic Modelling and Clustering M2	Clustering for text	Earlier versions of Clustering

Appendix: Supported Languages

Arabic	Dutch	Greek	Italian	Russian
Catalan	Esperanto	Hebrew	Japanese	Spanish
Chinese	Finnish	Hindi	Korean	Swedish
Czech	French	Hungarian	Persian	Turkish
Danish	German	Indonesian	Portuguese	Ukrainian

Novacene AI Corp.
123 Slater Street, Suite 610, Ottawa, ON K1P 5H2, Canada | (800) 717-0814 | www.novacene.ai

Appendix: Ideas and Comments Classifier

The Generation Energy Idea and Comment Submissions dataset is comprised of 1779 instances. Text from the dataset included a wide range of topics which can be observed in the table below.

Theme contained in dataset	
Forum Related Themes	**Number of Related Ideas**
Affordability/Abordabilité	64
Biomass/Biomasse	34
Communities/Commuautés	71
Energy Efficiency/Efficacité énergétique	4
Electricity/Électricité	87
Finance/La finance	183
Geothermal/Géothermie	22
Governance/Gouvernance	47
Heating /Le chauffage	20
Hydro/Hydroélectricité	36
Hydrogen/D'hydrogène	15
Information/Information	11
Innovation/Innovation	98
International/Internationale	107
Labour Markets/Marché du travail	47
Natural Gas/Gaz Naturel	71
Nuclear/Énergie Nucléaire	26
Petroleum/Pétrolier	149
Pipelines/Pipelines	66
Public Confidence/La confiance du public	3
Remote Communities/Les collectivités éloignées	11
Renewables/Énergies renouvelables	280
Security/Sûreté	10
Solar/Solaire	150

Storage/Stockage d’énergie	34
Tidal/Énergie Marémotrice	15
Transportation/Transports	166
Wind/Éolienne	65
Youth/jeunesse	5