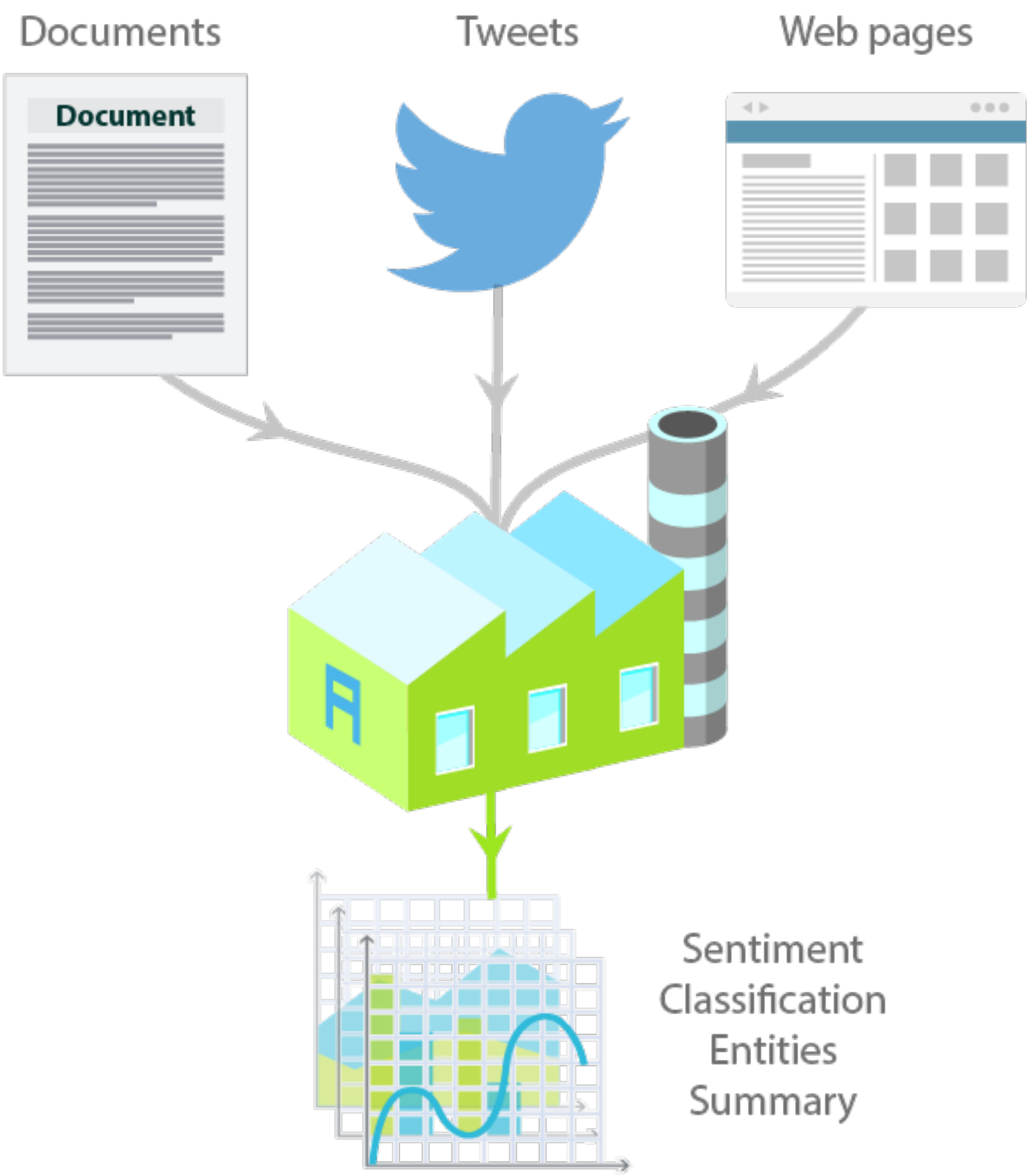
AYLIEN

Using NLP to understand textual content at scale

Parsa Ghaffari, CEO & Founder

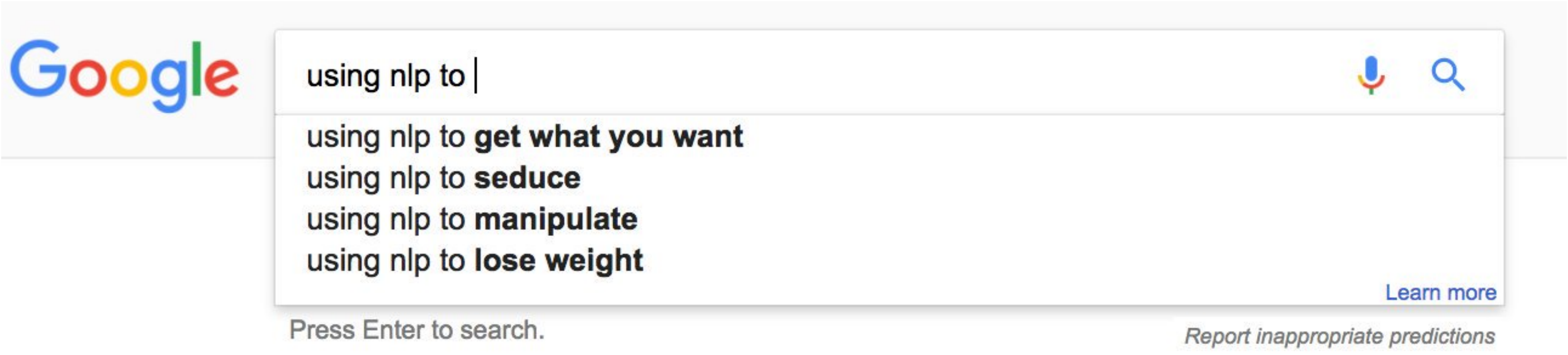
aylien.com / @_aylien

Introduction



AYLIEN is a platform for **aggregating** large volumes of textual content and **analyzing** and **understanding** it using **NLP.**

What is NLP?



using nlp to

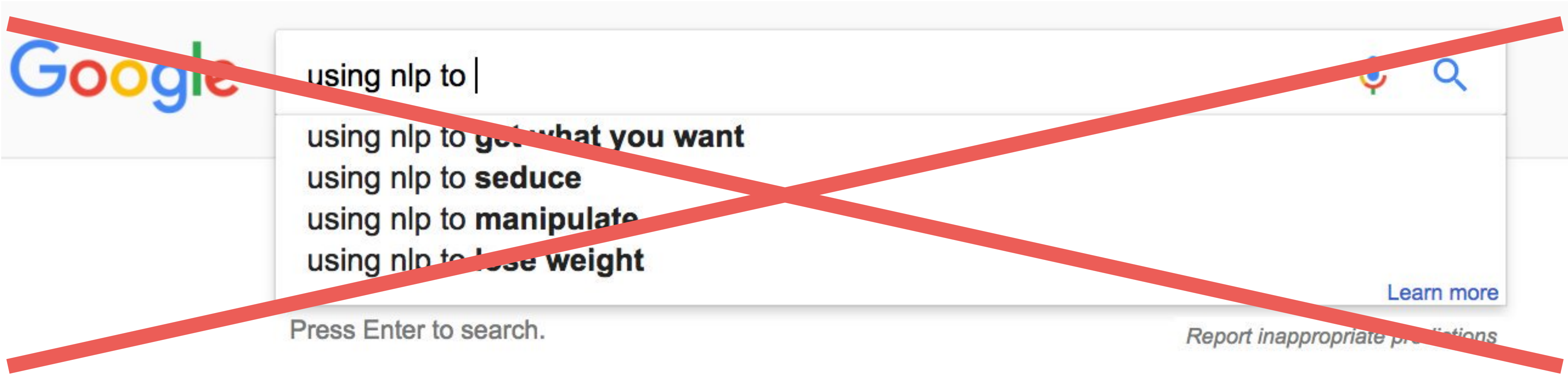
using nlp to **get what you want**

using nlp to **seduce**

using nlp to **manipulate**

using nlp to **lose weight**

What is NLP *not*?



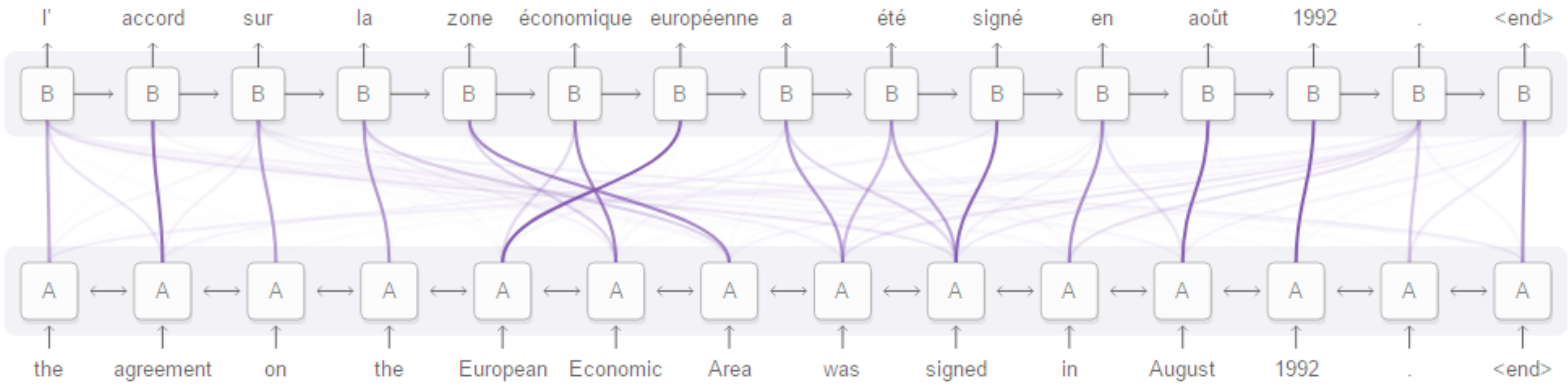
~~using nlp to~~

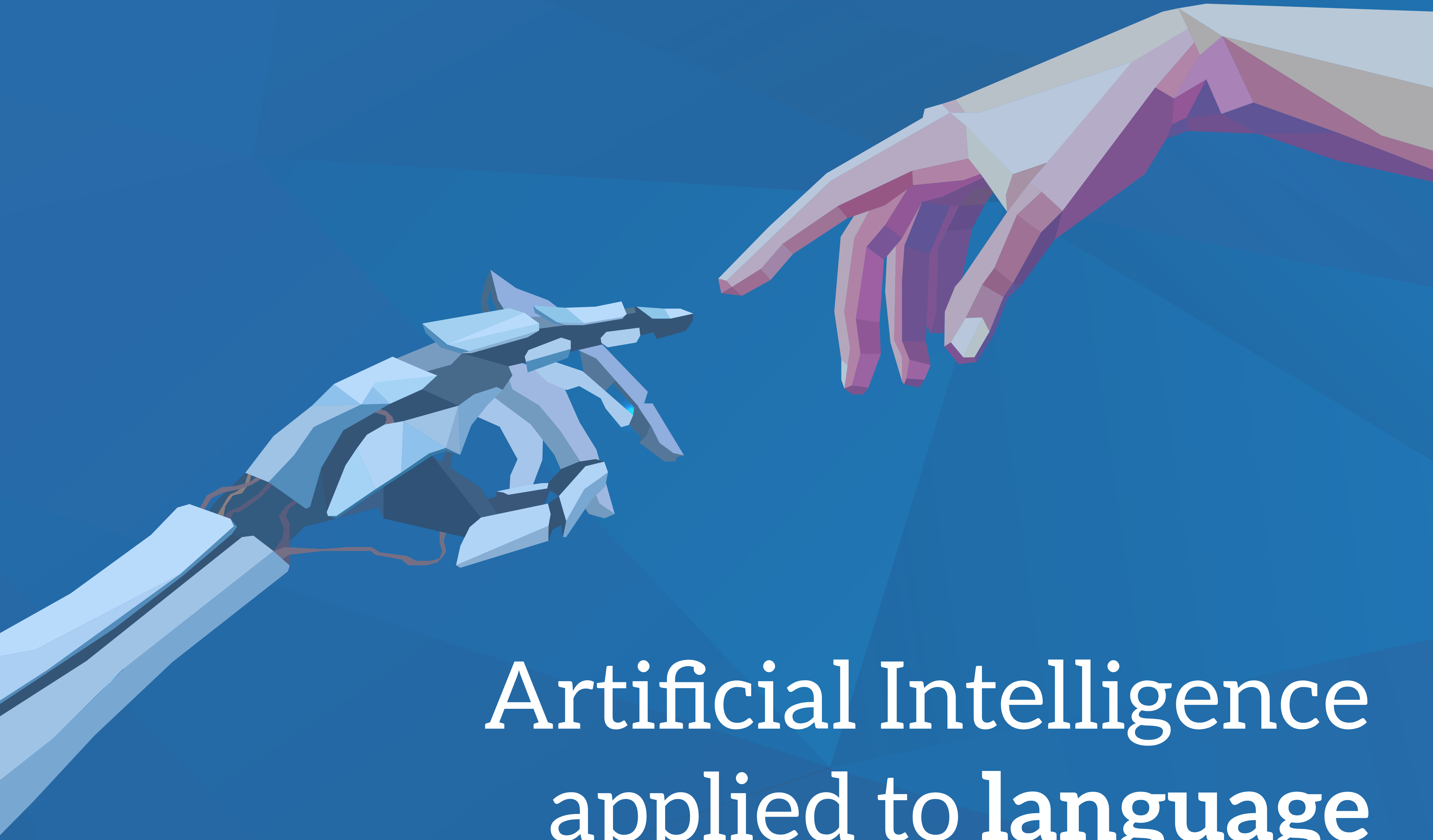
~~using nlp to **get what you want**~~

~~using nlp to **seduce**~~

~~using nlp to **manipulate**~~

~~using nlp to **lose weight**~~





Artificial Intelligence applied to **language**

Why NLP?

Language is a proxy to people's thoughts, emotions and feelings, and intentions

↓

We can understand people by understanding language

↓

Computers can help people accomplish things more efficiently, if they understand language

↓

Wildly applicable technology capable of creating $B’s in value

Challenges of NLP

Challenge #1

The inherent complexity of language

What does it mean when someone says:
“I made her duck”?

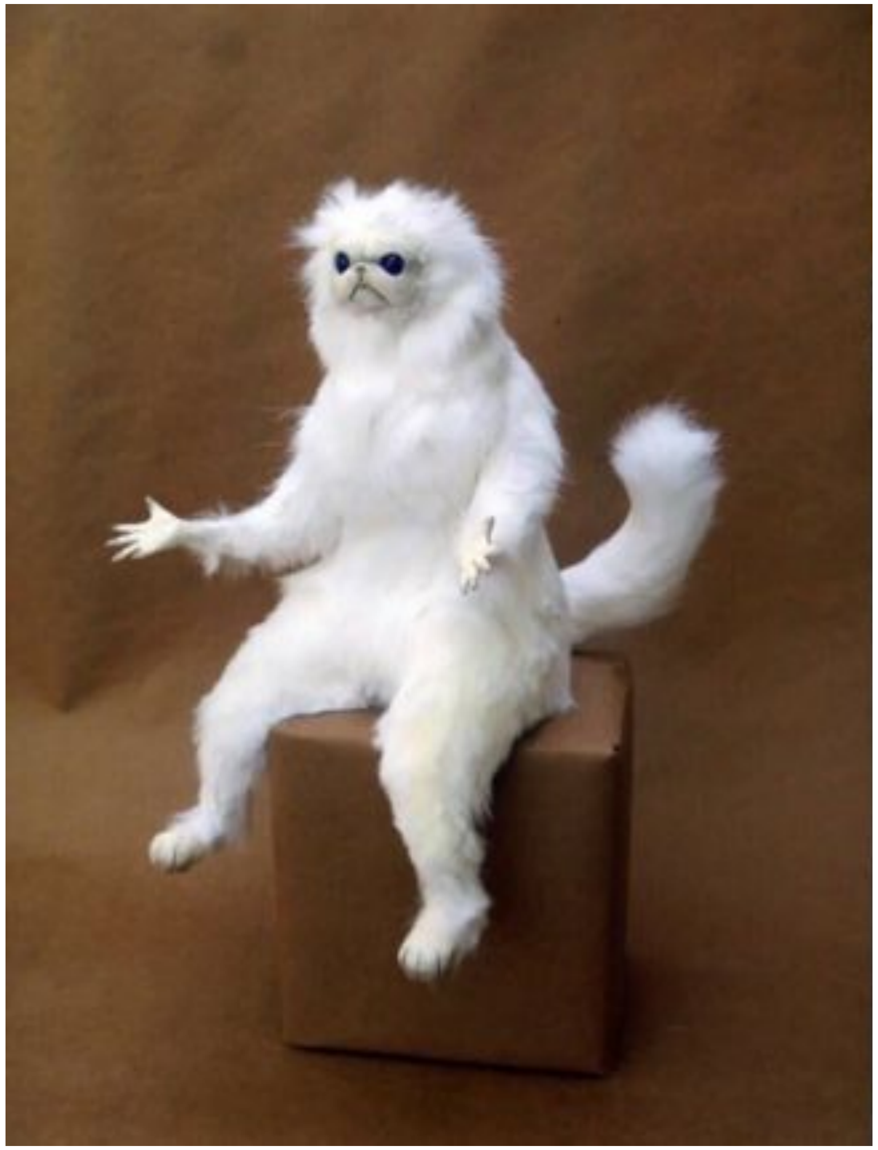
How many distinct interpretations can you count?

Challenge #1

The inherent complexity of language

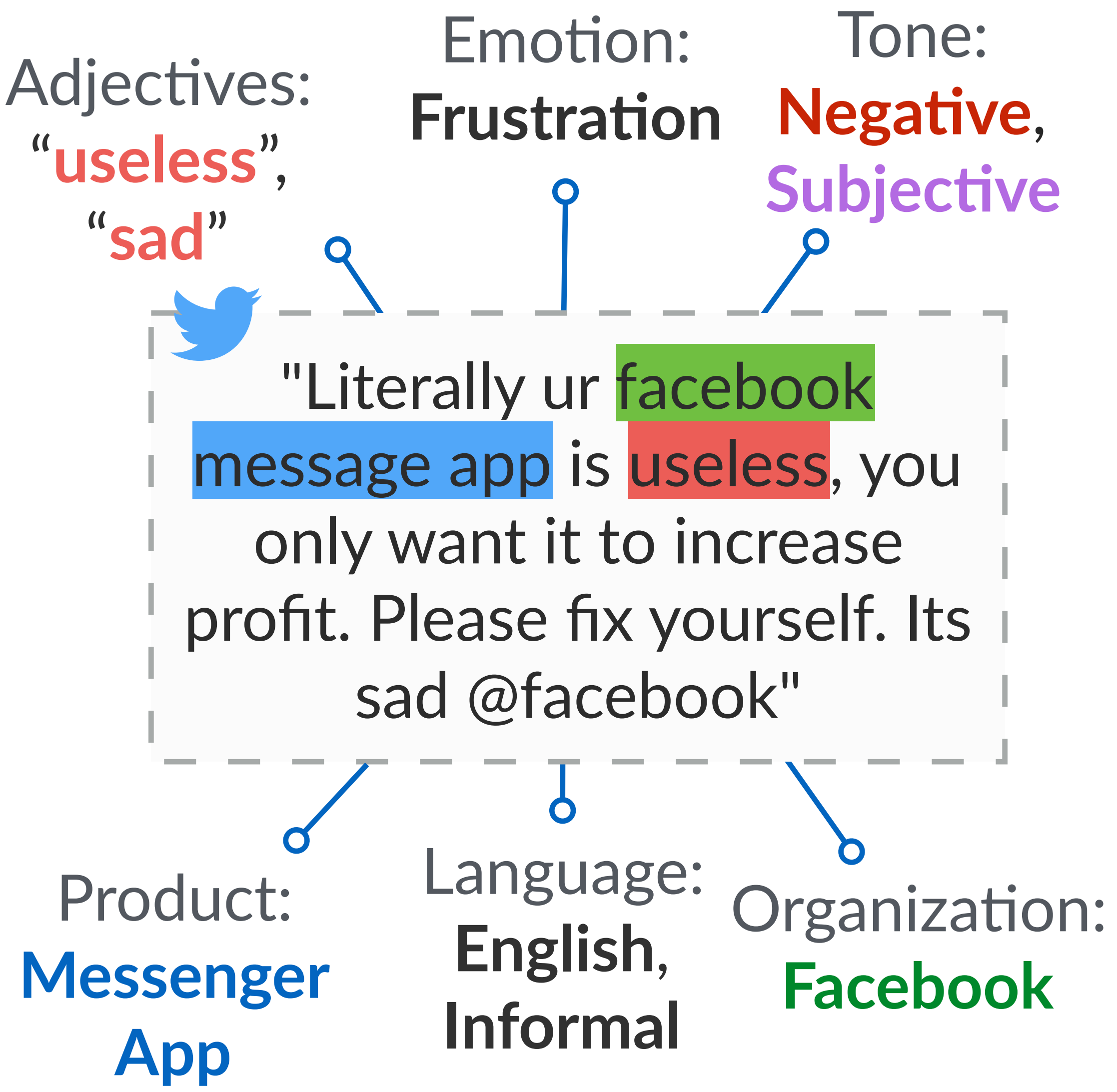
"I made her duck" could be interpreted as:

- I cooked a duck for her
- I cooked a duck belonging to her
- I created a duck for her
- I created a duck that now belongs to her
- I caused her to lower her head
- I turned her into a duck (!)



Challenge 1:

Complexity of Language



Adjectives:
"useless",
"sad"

Emotion:
Frustration

Tone:
Negative,
Subjective

"Literally ur facebook message app is useless, you only want it to increase profit. Please fix yourself. Its sad @facebook"

Product:
Messenger App

Language:
English,
Informal

Organization:
Facebook

Challenge 1:

Complexity of Language

And it's not just **English**!

- **German:** "Donaudampfschifffahrtsgesellschaftskapitän" (5 "words")
- **Chinese:** 50,000 different characters (2-3k to read a newspaper)
- **Japanese:** 3 writing systems
- **Thai:** Ambiguous word boundaries and sentence concepts
- **Slavic:** Different word forms depending on gender, case, tense

Challenge 1:

Complexity of Language

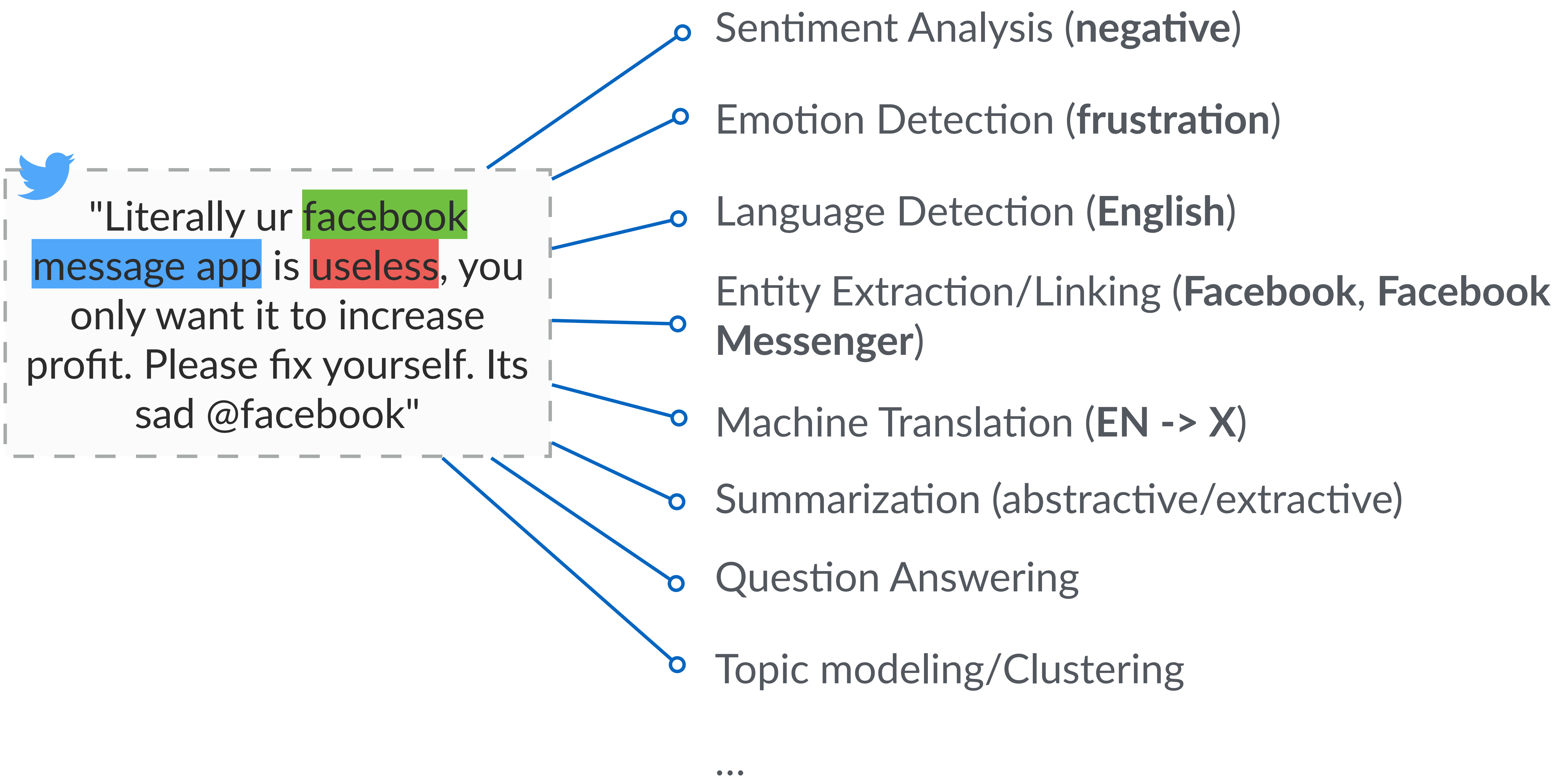
Machine Learning is the most suitable toolbox available to us for dealing with this complexity.

However, there are still challenges:

1. Variety of tasks
2. Diversity of domains
3. Data preparation
4. Model training
5. Model evaluation
6. Workflow issues

Challenge 1.1:

Variety of Tasks



"Literally ur facebook message app is useless, you only want it to increase profit. Please fix yourself. Its sad @facebook"

- Sentiment Analysis (**negative**)
- Emotion Detection (**frustration**)
- Language Detection (**English**)
- Entity Extraction/Linking (**Facebook**, **Facebook Messenger**)
- Machine Translation (**EN -> X**)
- Summarization (abstractive/extractive)
- Question Answering
- Topic modeling/Clustering

...

Challenge 1.2:

Diversity of Domains

Genres & sources:

- News articles
- Social media updates
- Reviews
- …

Languages & dialects:

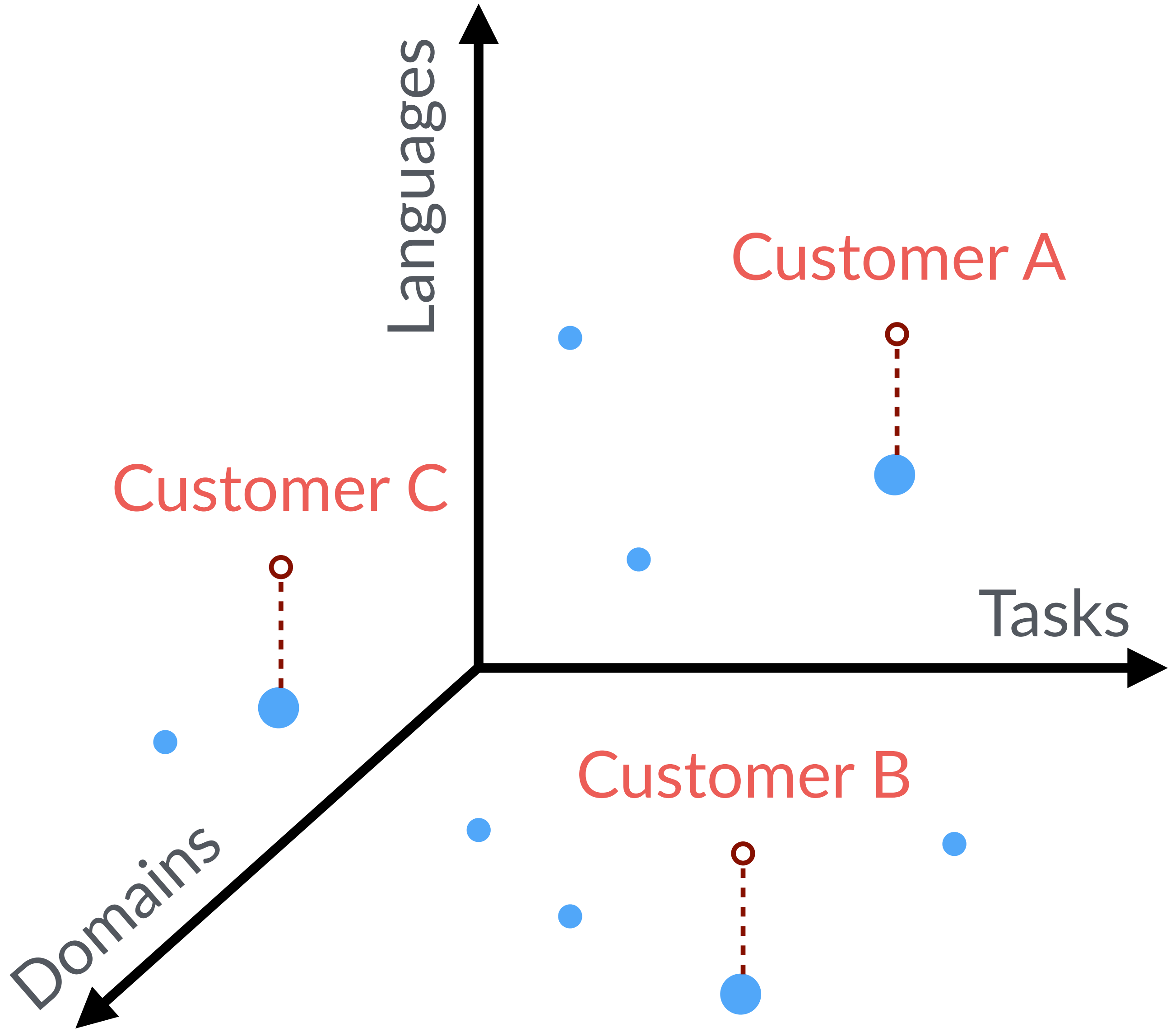
- English:
    - British:
        - Northern:
            - Cheshire
            - …
        - …
    - American:
        - …
    - …
- …

Other types of domains:

- Industries
- Users
- Time
- …

Challenges 1.1-2:

Tasks, Domains & Languages



Challenge 1.3:

Data Preparation

- Gathering data for training/testing
- Defining labels/a taxonomy
- Cleaning up the data
- Getting the data in the right format
- Annotating the data
- Splitting the data

Challenge 1.4:

Model Training

- Selecting the right algorithm/NN architecture
- Picking the right representations (e.g. pre-trained word embeddings)
- Hyper-parameter tuning
- Pre-training on other data (e.g. pre-training on sentiment data before emotion detection)
- Domain adaptation/transfer learning

Challenge 1.5:

Model Evaluation

- Picking/defining the right metrics
- Creating a test set to evaluate models on
- Comparing models
- Qualitative evaluation
- Explaining predictions

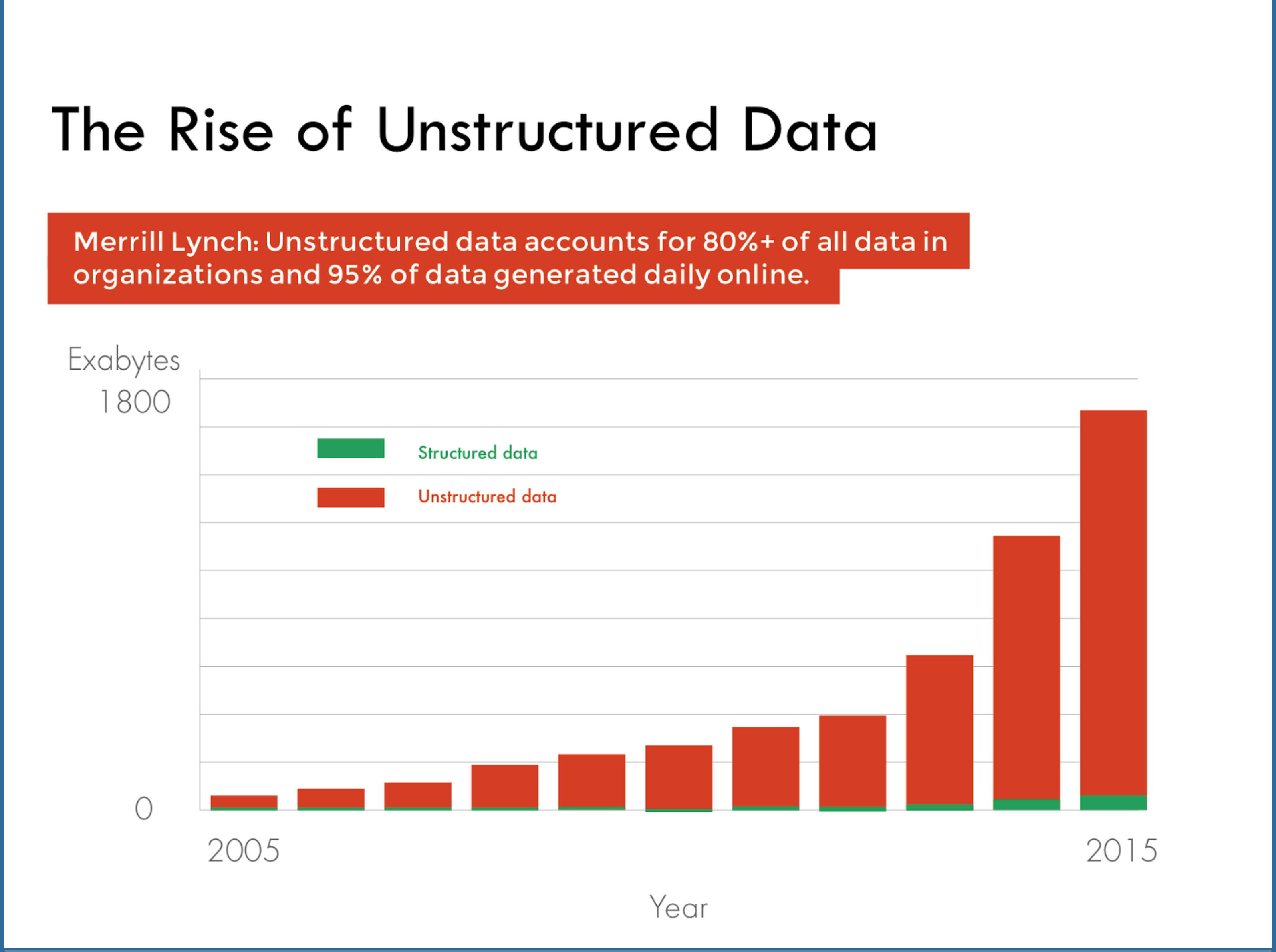
Challenge 1.6:

Workflow-related Issues

- Continuously updating models and getting feedback
- Active learning
- Domain adaptation
- Dataset visualization
- Consuming trained models

Challenge #2

The scale and production rate of unstructured data



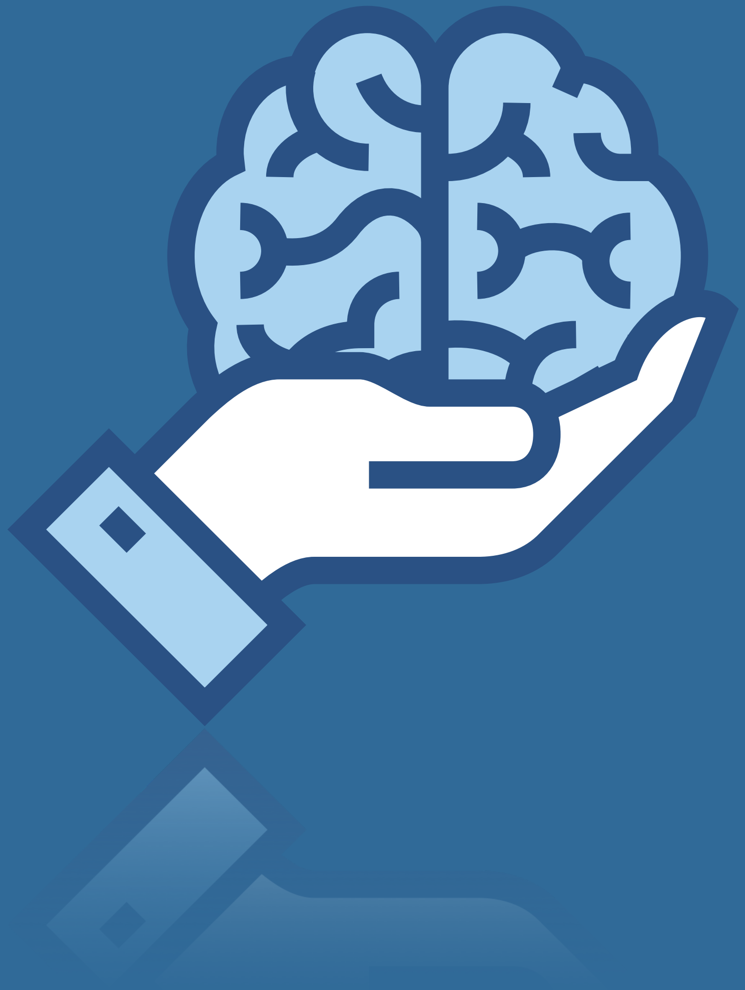
The Rise of Unstructured Data

Merrill Lynch: Unstructured data accounts for 80%+ of all data in organizations and 95% of data generated daily online.

Research from International Data Corporation (IDC) shows that unstructured content accounts for 90% of all digital information.

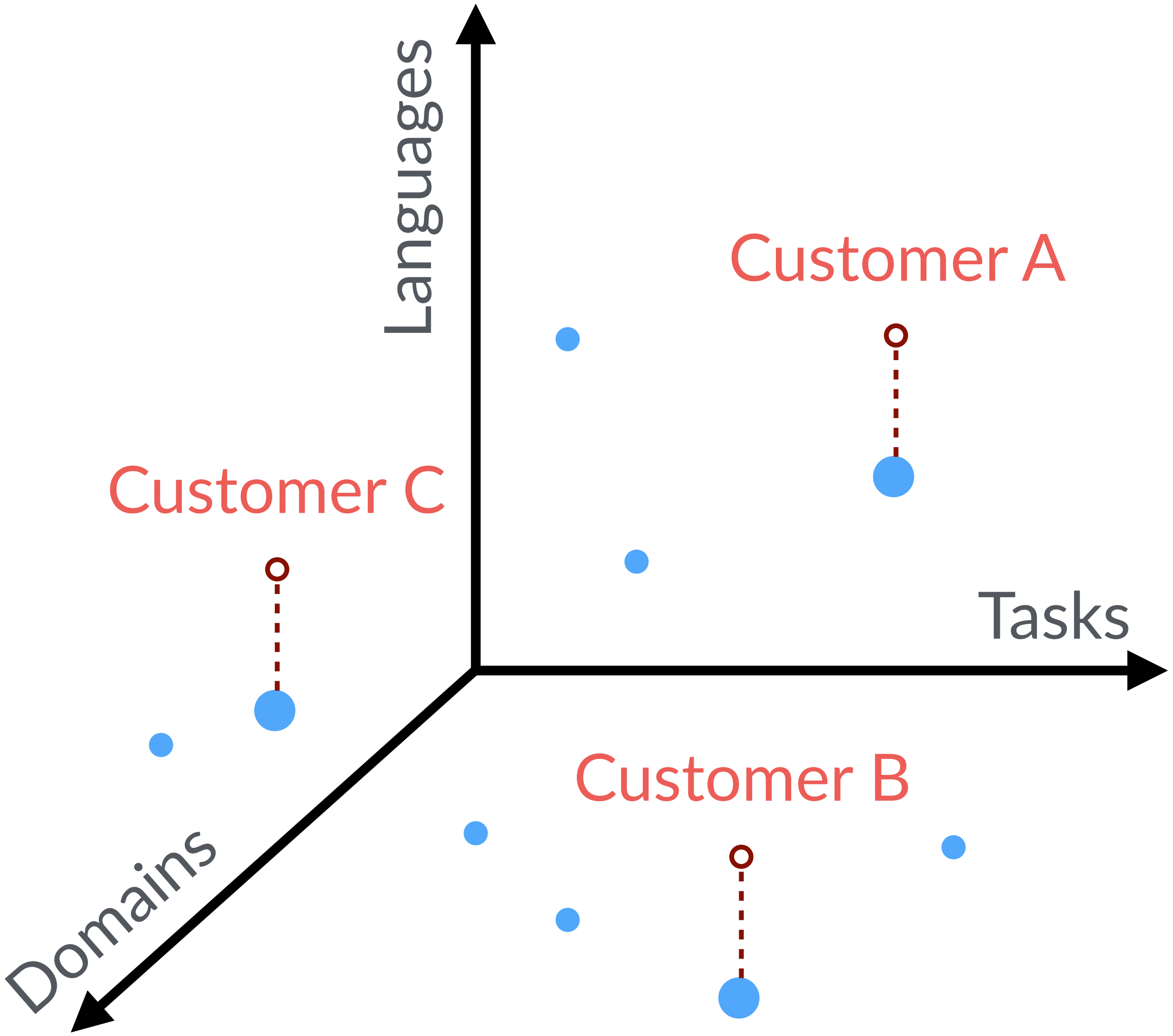
“More content was uploaded yesterday than any one human could ever consume in their entire life.” – Condé Nast

Our Solution



Challenges 1.1-2:

Tasks, Domains & Languages



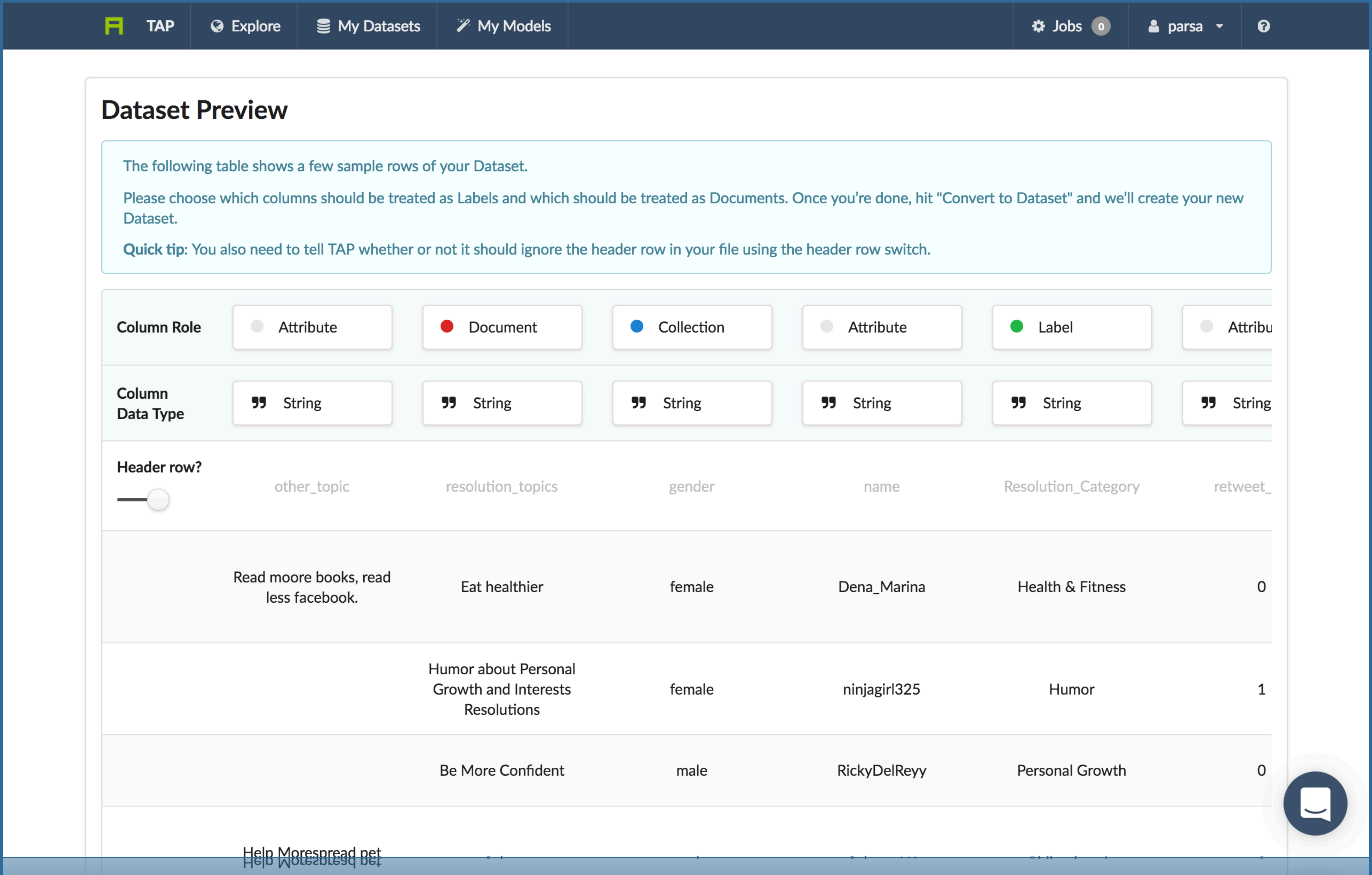
WHOA
WHOA WHOA!
WHOA WHOA!
quickmeme.com
WHOA.

Text Analysis Platform (TAP)ᵝ

An in-browser environment for building **custom NLP models**

Build Datasets

From **CSV** files, using our **Knowledge Base** or by forking other datasets



Dataset Preview

The following table shows a few sample rows of your Dataset.

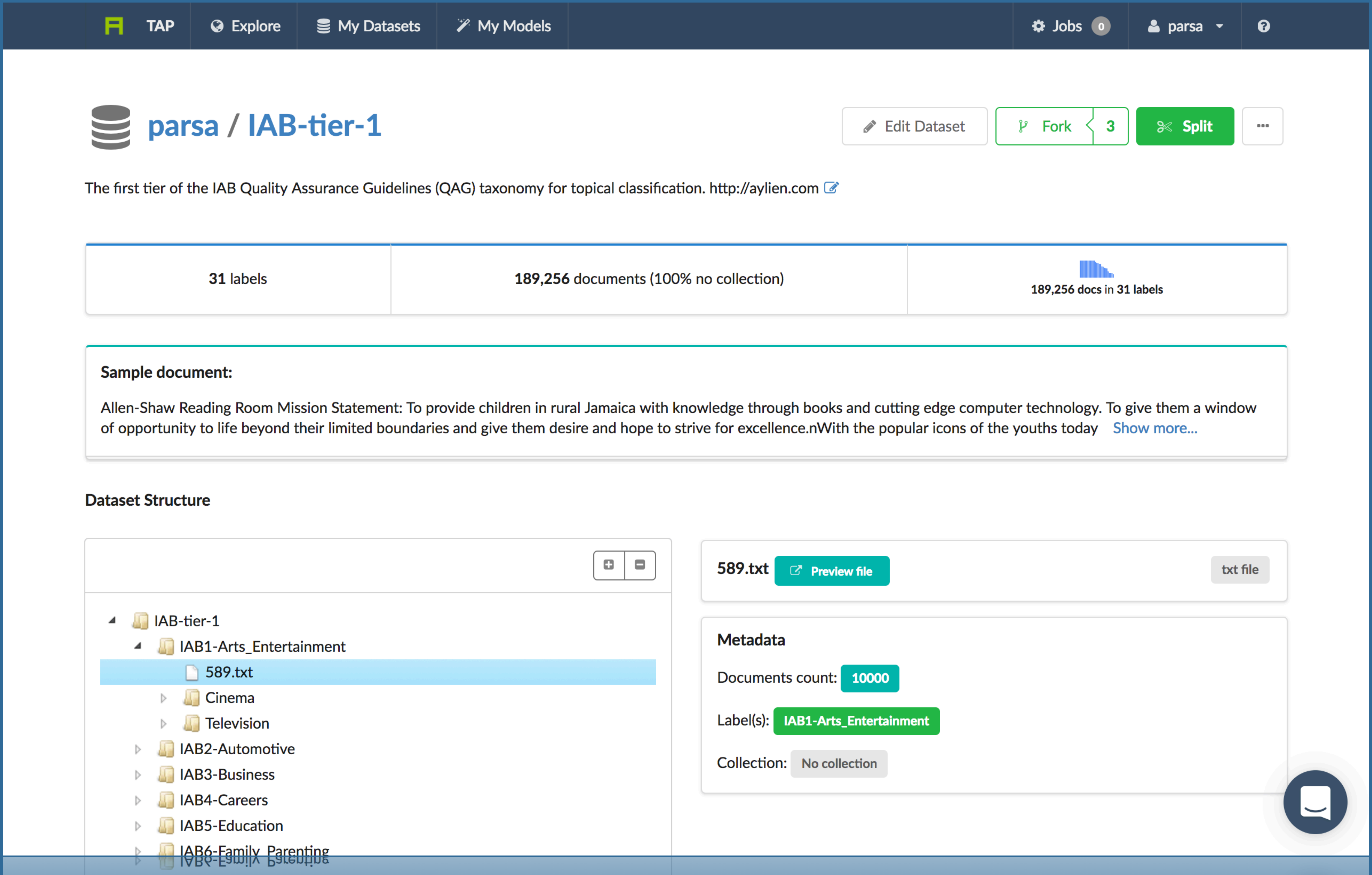
Please choose which columns should be treated as Labels and which should be treated as Documents. Once you're done, hit "Convert to Dataset" and we'll create your new Dataset.

Quick tip: You also need to tell TAP whether or not it should ignore the header row in your file using the header row switch.

Column Role	○ Attribute	● Document	● Collection	○ Attribute	● Label	○ Attribu
Column Data Type	❞ String	❞ String	❞ String	❞ String	❞ String	❞ String
Header row?	other_topic	resolution_topics	gender	name	Resolution_Category	retweet_
	Read moore books, read less facebook.	Eat healthier	female	Dena_Marina	Health & Fitness	0
		Humor about Personal Growth and Interests Resolutions	female	ninjagirl325	Humor	1
		Be More Confident	male	RickyDelReyy	Personal Growth	0
	Help Morespread pet					

Build Datasets

From **CSV** files, using our **Knowledge Base** or by forking other datasets



parsa / IAB-tier-1

The first tier of the IAB Quality Assurance Guidelines (QAG) taxonomy for topical classification. http://aylien.com

31 labels	**189,256** documents (100% no collection)	**189,256 docs** in **31 labels**

Sample document:

Allen-Shaw Reading Room Mission Statement: To provide children in rural Jamaica with knowledge through books and cutting edge computer technology. To give them a window of opportunity to life beyond their limited boundaries and give them desire and hope to strive for excellence.nWith the popular icons of the youths today Show more...

Dataset Structure

- IAB-tier-1
  - IAB1-Arts_Entertainment
    - 589.txt
    - Cinema
    - Television
  - IAB2-Automotive
  - IAB3-Business
  - IAB4-Careers
  - IAB5-Education
  - IAB6-Family_Parenting

589.txt Preview file txt file

Metadata

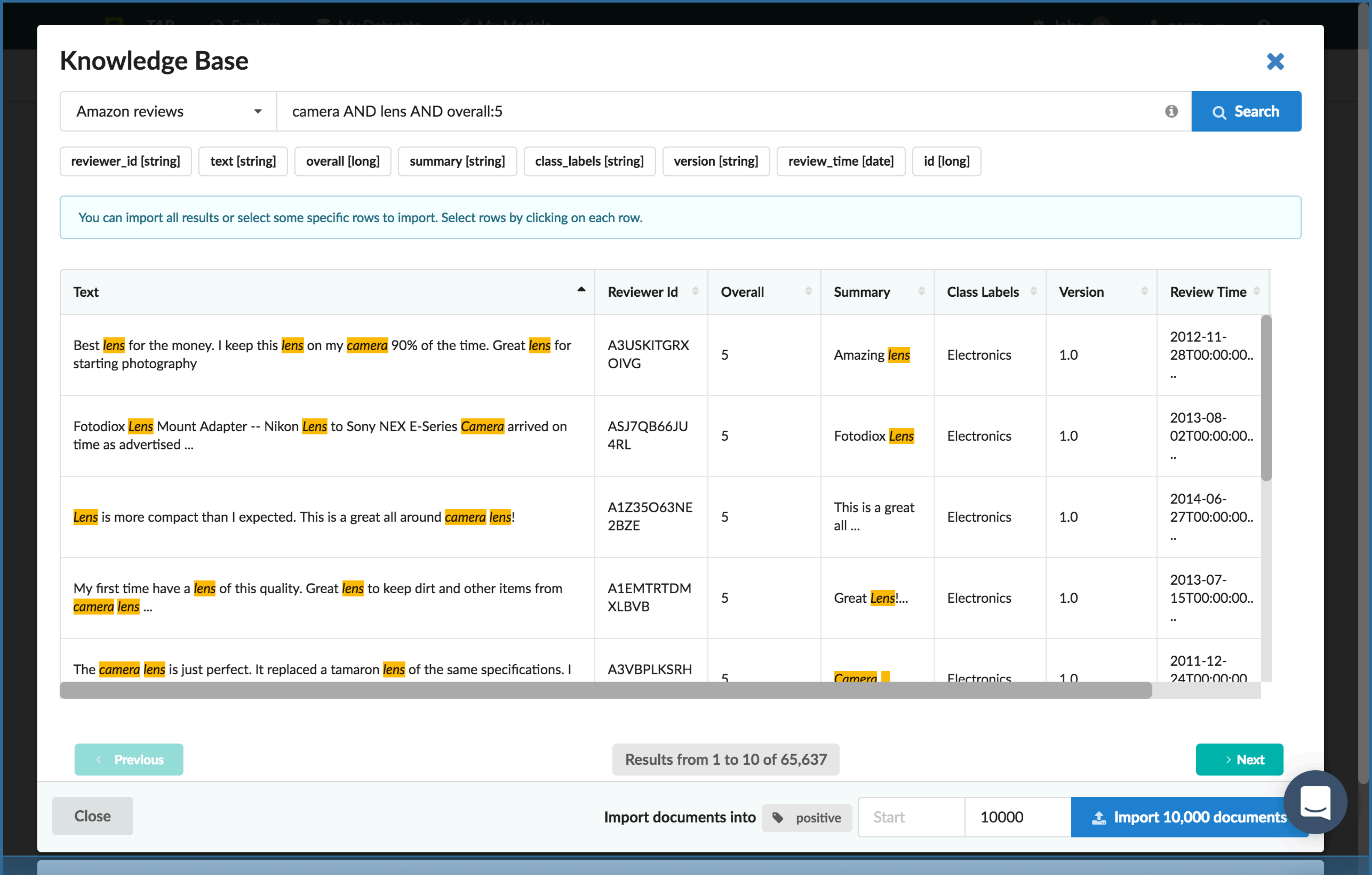
Documents count: 10000

Label(s): IAB1-Arts_Entertainment

Collection: No collection

Build Datasets

From **CSV** files, using our **Knowledge Base** or by forking other datasets



Knowledge Base

Amazon reviews ▾ | camera AND lens AND overall:5 | Search

reviewer_id [string] | text [string] | overall [long] | summary [string] | class_labels [string] | version [string] | review_time [date] | id [long]

You can import all results or select some specific rows to import. Select rows by clicking on each row.

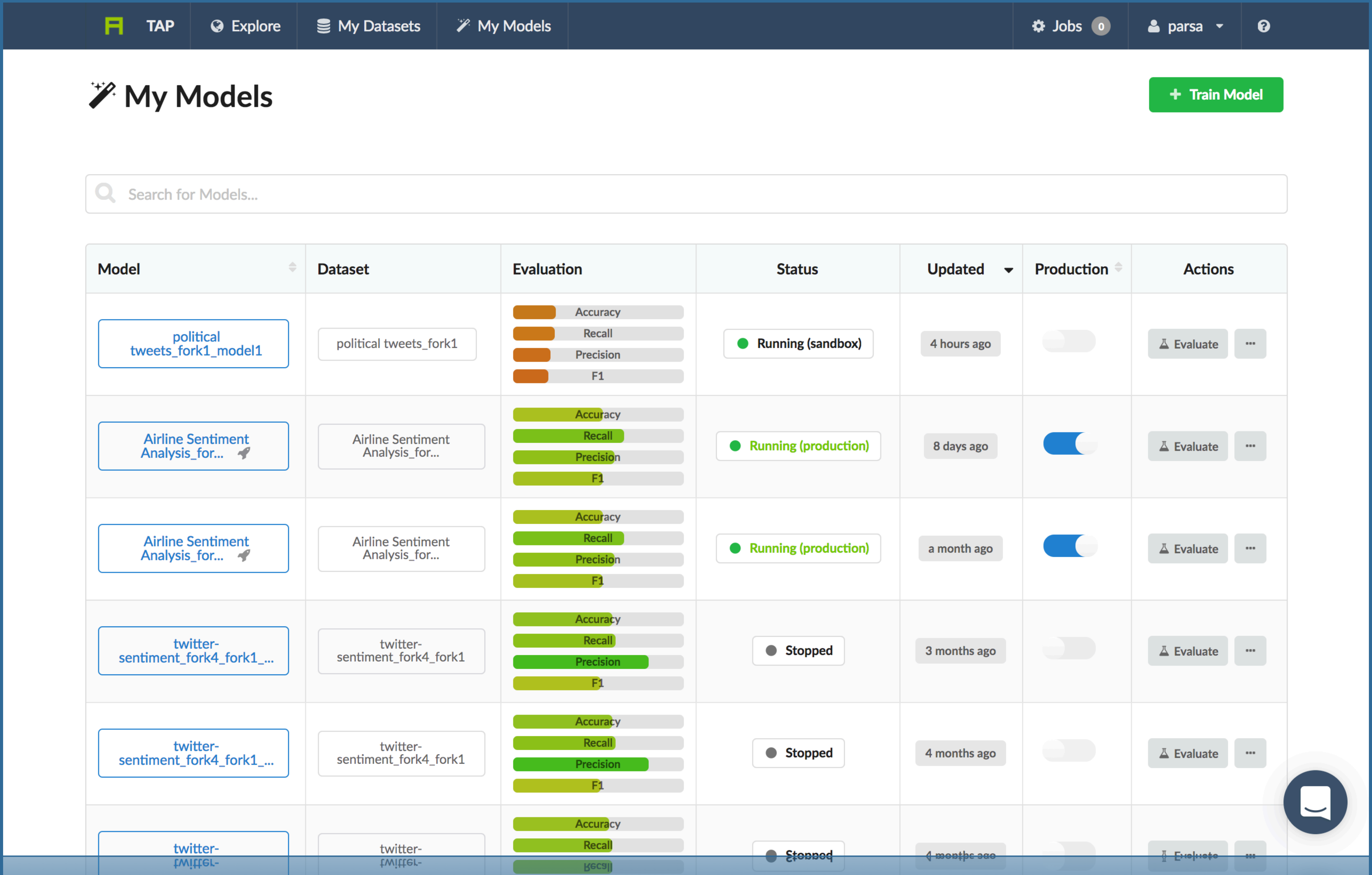
Text	Reviewer Id	Overall	Summary	Class Labels	Version	Review Time
Best *lens* for the money. I keep this *lens* on my *camera* 90% of the time. Great *lens* for starting photography	A3USKITGRX OIVG	5	Amazing *lens*	Electronics	1.0	2012-11-28T00:00:00.. ..
Fotodiox *Lens* Mount Adapter -- Nikon *Lens* to Sony NEX E-Series *Camera* arrived on time as advertised ...	ASJ7QB66JU 4RL	5	Fotodiox *Lens*	Electronics	1.0	2013-08-02T00:00:00.. ..
Lens is more compact than I expected. This is a great all around *camera* *lens*!	A1Z35O63NE 2BZE	5	This is a great all ...	Electronics	1.0	2014-06-27T00:00:00.. ..
My first time have a *lens* of this quality. Great *lens* to keep dirt and other items from *camera* *lens* ...	A1EMTRTDM XLBVB	5	Great *Lens*!...	Electronics	1.0	2013-07-15T00:00:00.. ..
The *camera* *lens* is just perfect. It replaced a tamaron *lens* of the same specifications. I	A3VBPLKSRH	5	*Camera*	Electronics	1.0	2011-12-24T00:00:00

‹ Previous | Results from 1 to 10 of 65,637 | › Next

Close | Import documents into positive | Start | 10000 | Import 10,000 documents

Train Models

Compare various models/parameters and pick the best one



TAP | Explore | My Datasets | My Models | Jobs 0 | parsa | ?

My Models

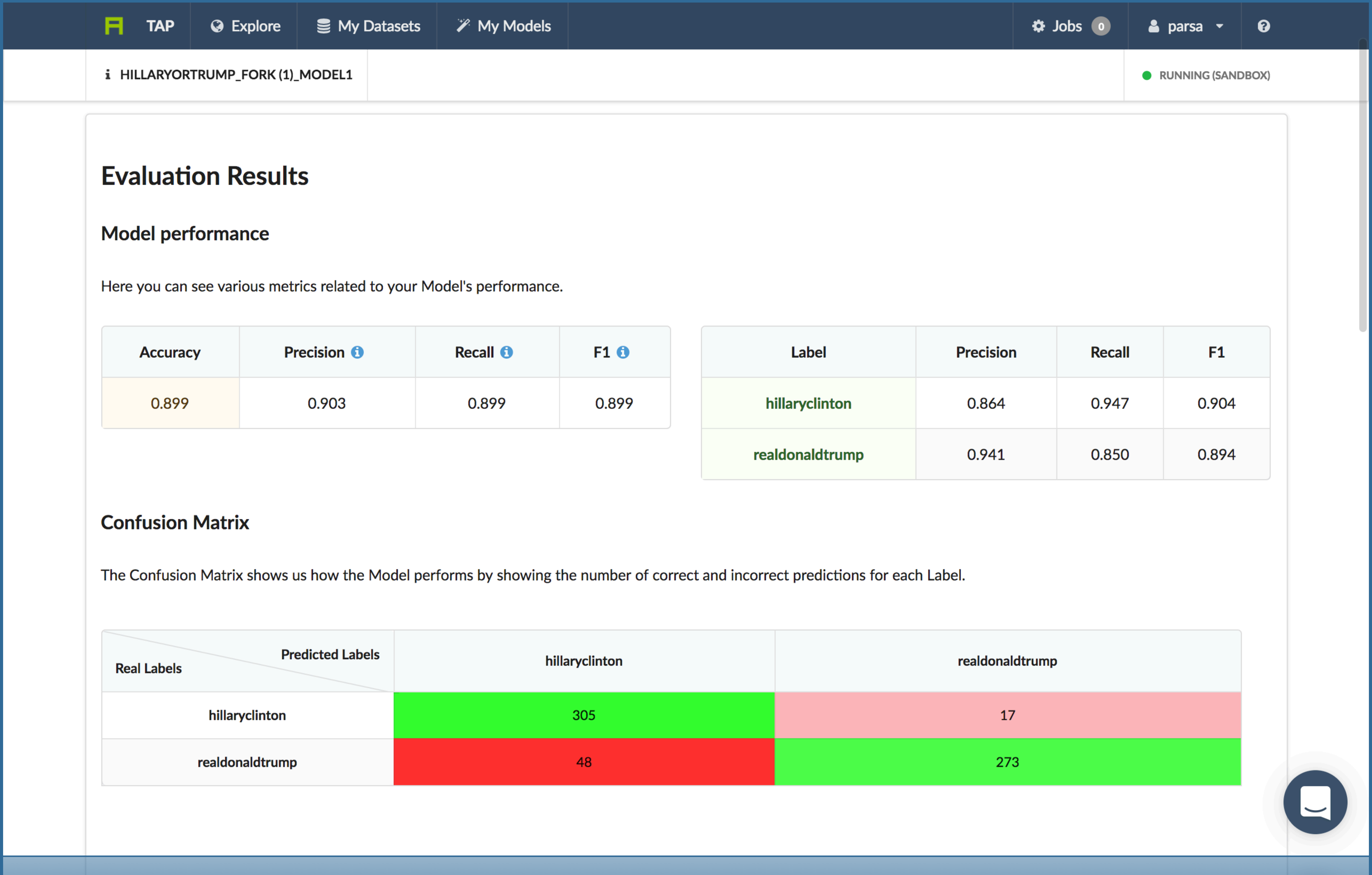
+ Train Model

Search for Models...

Model	Dataset	Evaluation	Status	Updated	Production	Actions
political tweets_fork1_model1	political tweets_fork1	Accuracy Recall Precision F1	● Running (sandbox)	4 hours ago	○	Evaluate ···
Airline Sentiment Analysis_for...	Airline Sentiment Analysis_for...	Accuracy Recall Precision F1	● Running (production)	8 days ago	●	Evaluate ···
Airline Sentiment Analysis_for...	Airline Sentiment Analysis_for...	Accuracy Recall Precision F1	● Running (production)	a month ago	●	Evaluate ···
twitter-sentiment_fork4_fork1_...	twitter-sentiment_fork4_fork1	Accuracy Recall Precision F1	● Stopped	3 months ago	○	Evaluate ···
twitter-sentiment_fork4_fork1_...	twitter-sentiment_fork4_fork1	Accuracy Recall Precision F1	● Stopped	4 months ago	○	Evaluate ···
twitter-	twitter-	Accuracy Recall	● Stopped	4 months ago	○	Evaluate ···

Evaluate Models

Understand different aspects of each model



Evaluation Results

Model performance

Here you can see various metrics related to your Model's performance.

Accuracy	Precision	Recall	F1
0.899	0.903	0.899	0.899

Label	Precision	Recall	F1
hillaryclinton	0.864	0.947	0.904
realdonaldtrump	0.941	0.850	0.894

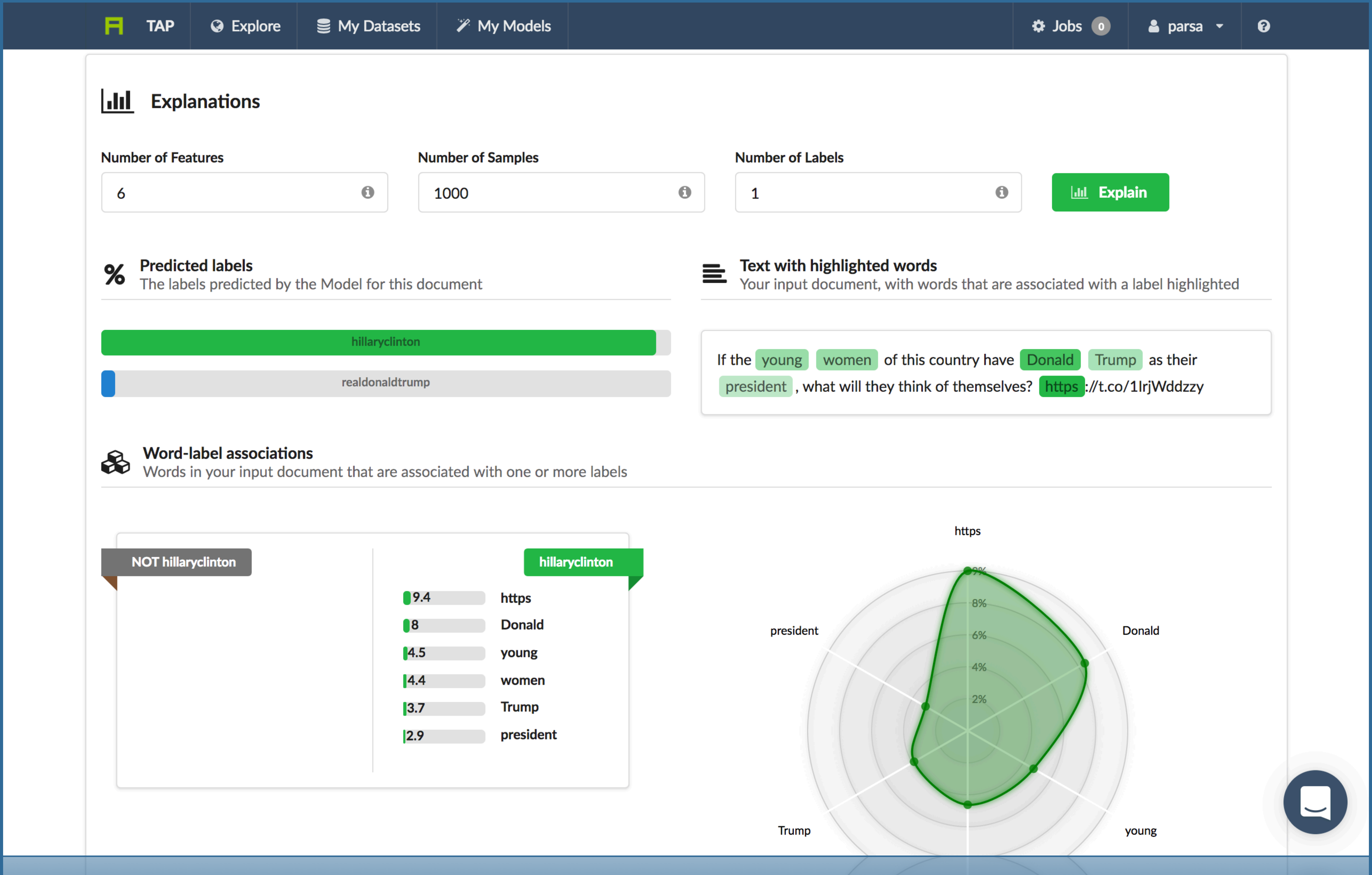
Confusion Matrix

The Confusion Matrix shows us how the Model performs by showing the number of correct and incorrect predictions for each Label.

Predicted Labels Real Labels	hillaryclinton	realdonaldtrump
hillaryclinton	305	17
realdonaldtrump	48	273

Evaluate Models

Explain predictions made by a model



Explanations

Number of Features: 6

Number of Samples: 1000

Number of Labels: 1

Predicted labels
The labels predicted by the Model for this document

hillaryclinton

realdonaldtrump

Text with highlighted words
Your input document, with words that are associated with a label highlighted

If the young women of this country have Donald Trump as their president , what will they think of themselves? https ://t.co/1lrjWddzzy

Word-label associations
Words in your input document that are associated with one or more labels

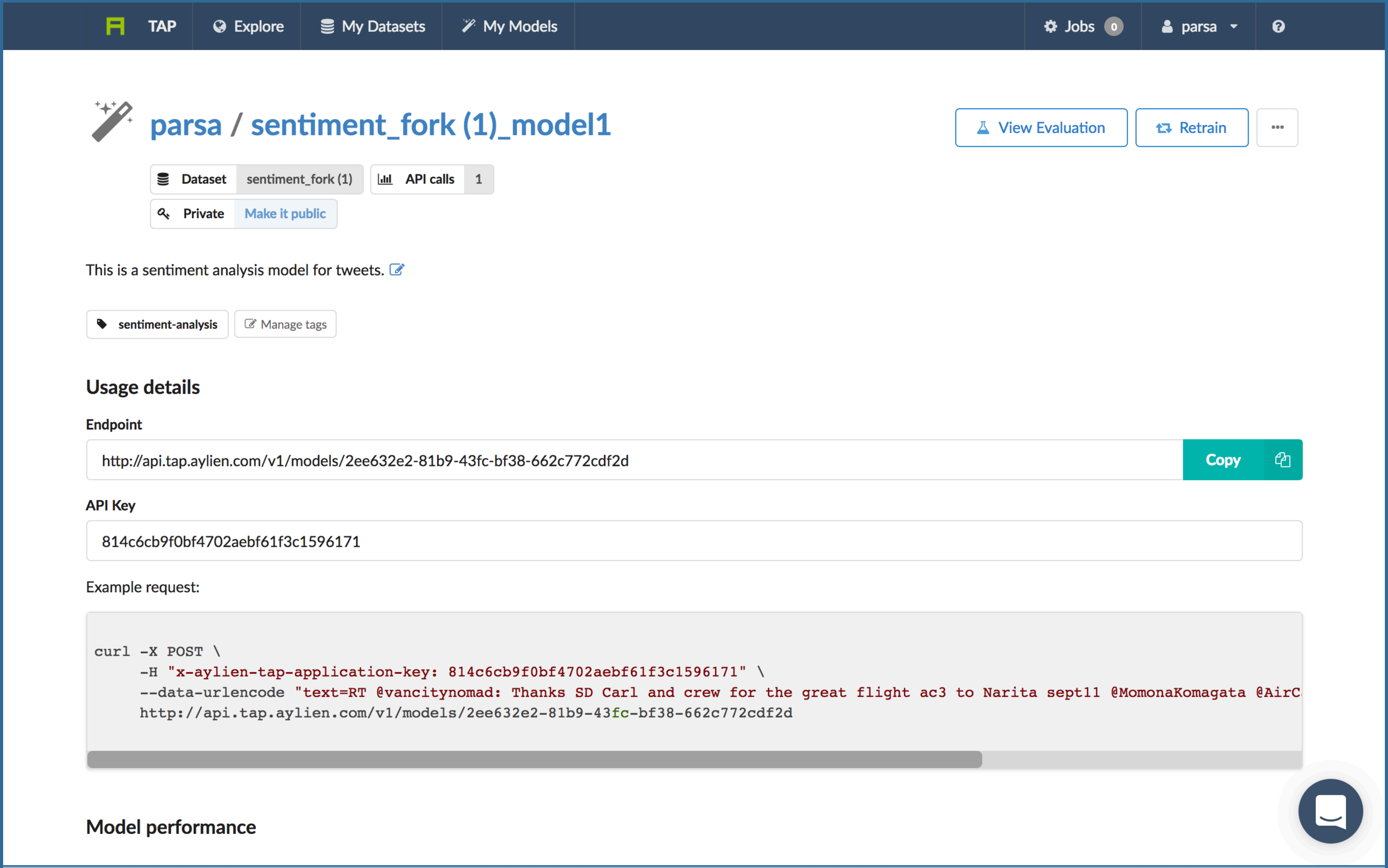
NOT hillaryclinton

hillaryclinton

9.4	https
8	Donald
4.5	young
4.4	women
3.7	Trump
2.9	president

Use Models

Deploy the best version of your model to production, and use it as an **API**



parsa / sentiment_fork (1)_model1

Dataset	sentiment_fork (1)	API calls	1

Private Make it public

This is a sentiment analysis model for tweets.

sentiment-analysis

Usage details

Endpoint

http://api.tap.aylien.com/v1/models/2ee632e2-81b9-43fc-bf38-662c772cdf2d

API Key

814c6cb9f0bf4702aebf61f3c1596171

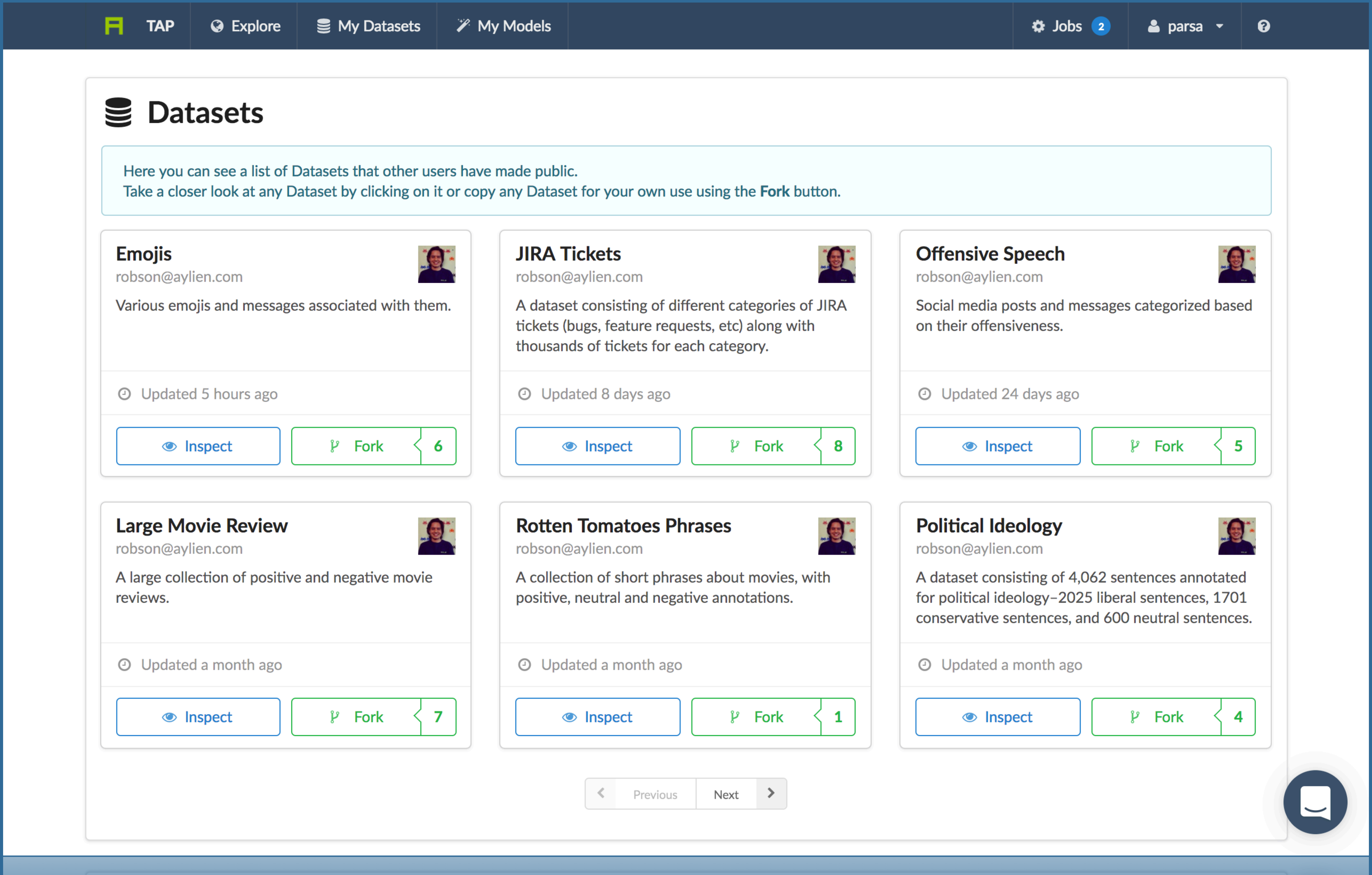
Example request:

```
curl -X POST \
     -H "x-aylien-tap-application-key: 814c6cb9f0bf4702aebf61f3c1596171" \
     --data-urlencode "text=RT @vancitynomad: Thanks SD Carl and crew for the great flight ac3 to Narita sept11 @MomonaKomagata @AirC
     http://api.tap.aylien.com/v1/models/2ee632e2-81b9-43fc-bf38-662c772cdf2d
```

Model performance

Leverage the **Marketplace**

Share your **models** and **datasets** with other users, or leverage theirs



Datasets

Here you can see a list of Datasets that other users have made public.
Take a closer look at any Dataset by clicking on it or copy any Dataset for your own use using the **Fork** button.

Emojis
robson@aylien.com
Various emojis and messages associated with them.
Updated 5 hours ago

JIRA Tickets
robson@aylien.com
A dataset consisting of different categories of JIRA tickets (bugs, feature requests, etc) along with thousands of tickets for each category.
Updated 8 days ago

Offensive Speech
robson@aylien.com
Social media posts and messages categorized based on their offensiveness.
Updated 24 days ago

Large Movie Review
robson@aylien.com
A large collection of positive and negative movie reviews.
Updated a month ago

Rotten Tomatoes Phrases
robson@aylien.com
A collection of short phrases about movies, with positive, neutral and negative annotations.
Updated a month ago

Political Ideology
robson@aylien.com
A dataset consisting of 4,062 sentences annotated for political ideology–2025 liberal sentences, 1701 conservative sentences, and 600 neutral sentences.
Updated a month ago

Key takeaways

- NLP brings us closer to people’s thoughts, emotions and intentions => It has the potential to impact billions of people;
- We’re not doing a great job at modeling the processes of which language is an output => We face the problem of too many <task, domain, language>s => Infinite room for research in this area;
- NLP(/ML) is not a one-size-fits-all problem => We need to rely on good engineering and products to fill the gap;

Thank you!

aylien.com

We’re hiring!

aylien.com/jobs