



HuggingFace

Angie McMillan-Major and Emily M. Bender
DELPH-IN 2021 SIG
July 19, 2021



Questions to Explore

- What is [HuggingFace](#) 🤗?
 - What do they offer?
 - Who is their audience?
- What could be useful to DELPH-IN? What isn't so useful?
 - What are the overlaps?

What is HuggingFace 🤗?

Open source Python libraries for NLP (recently added speech and vision):

- Transformers
- Datasets
- Tokenizers
- Accelerate

Resources such as community forum, docs, and [introductory course](#)




The AI community building the future.


Build, train and deploy state of the art models powered by the reference open source in natural language processing.

Star 48,656

More than 5,000 organizations are using Hugging Face

 **Allen Institute for AI**
Non-profit · 53 models

 **Facebook AI**
Company · 121 models

 **asteroid-team**
Non-profit

 **Google AI**
Company · 141 models

 **Amazon Web Services**
Company · 1 model

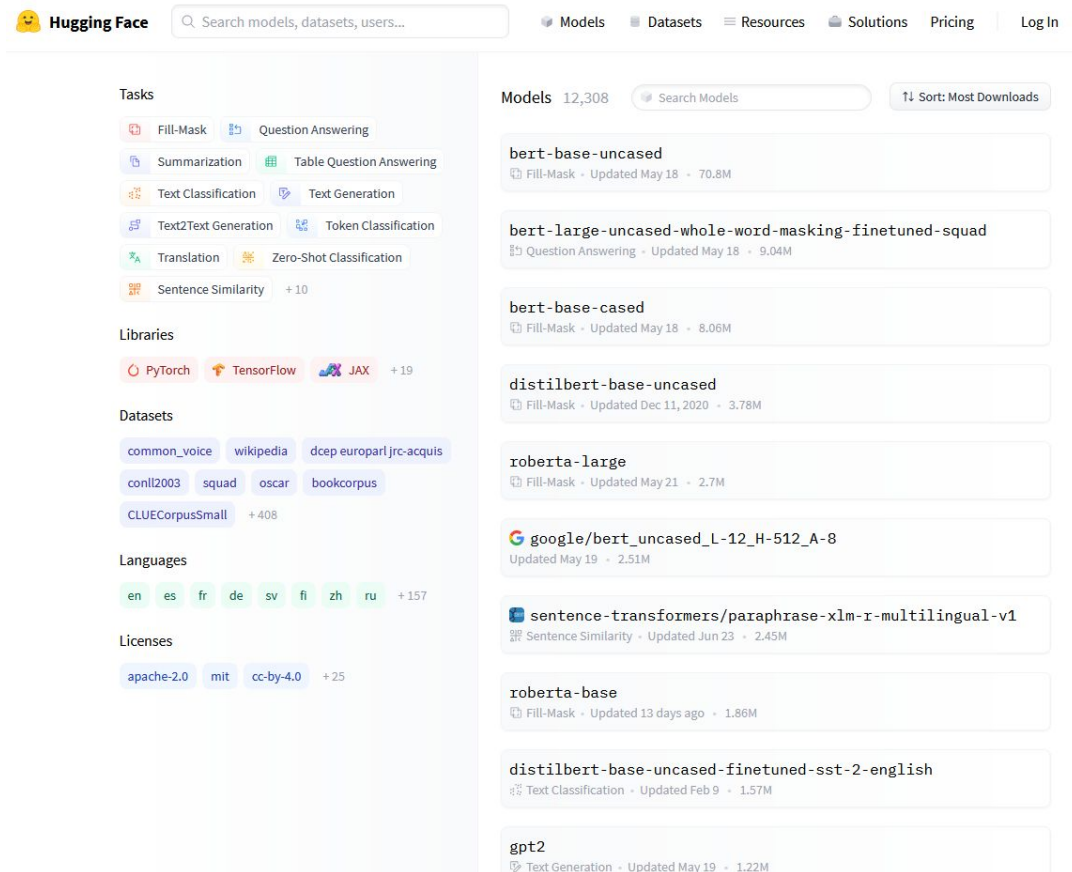
 **SpeechBrain**
Non-profit · 26 models

 **Microsoft**
Company · 51 models

 **Grammarly**
Company

HF Hub

- Web accessible catalog of models uploaded by organizations and individuals
- Searchable by task, libraries, datasets, languages, licenses



The screenshot shows the Hugging Face Hub homepage. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Resources, Solutions, Pricing, and Log In. Below the navigation bar, the page is divided into two main sections. The left section, titled 'Tasks', lists various natural language processing tasks such as Fill-Mask, Question Answering, Summarization, Table Question Answering, Text Classification, Text Generation, Text2Text Generation, Token Classification, Translation, Zero-Shot Classification, and Sentence Similarity. Below the tasks, there are sections for 'Libraries' (PyTorch, TensorFlow, JAX), 'Datasets' (common_voice, wikipedia, dcep europarl jrc-acquis, conll2003, squad, oscar, bookcorpus, CLUECorpusSmall), 'Languages' (en, es, fr, de, sv, fi, zh, ru), and 'Licenses' (apache-2.0, mit, cc-by-4.0). The right section, titled 'Models', displays a list of popular models, including bert-base-uncased, bert-large-uncased-whole-word-masking-finetuned-squad, bert-base-cased, distilbert-base-uncased, roberta-large, google/bert_uncased_L-12_H-512_A-8, sentence-transformers/paraphrase-xlm-r-multilingual-v1, roberta-base, distilbert-base-uncased-finetuned-sst-2-english, and gpt2. Each model entry shows its name, a brief description, the update date, and the number of downloads.

Hugging Face Search models, datasets, users...

Models **Datasets** **Resources** **Solutions** **Pricing** **Log In**

Tasks

- Fill-Mask Question Answering
- Summarization Table Question Answering
- Text Classification Text Generation
- Text2Text Generation Token Classification
- Translation Zero-Shot Classification
- Sentence Similarity +10

Libraries

- PyTorch TensorFlow JAX +19

Datasets

- common_voice wikipedia dcep europarl jrc-acquis
- conll2003 squad oscar bookcorpus
- CLUECorpusSmall +408

Languages

- en es fr de sv fi zh ru +157

Licenses


- apache-2.0 mit cc-by-4.0 +25

Models 12,308 Search Models Sort: Most Downloads

- bert-base-uncased**
Fill-Mask · Updated May 18 · 70.8M
- bert-large-uncased-whole-word-masking-finetuned-squad**
Question Answering · Updated May 18 · 9.04M
- bert-base-cased**
Fill-Mask · Updated May 18 · 8.06M
- distilbert-base-uncased**
Fill-Mask · Updated Dec 11, 2020 · 3.78M
- roberta-large**
Fill-Mask · Updated May 21 · 2.7M
- google/bert_uncased_L-12_H-512_A-8**
Updated May 19 · 2.51M
- sentence-transformers/paraphrase-xlm-r-multilingual-v1**
Sentence Similarity · Updated Jun 23 · 2.45M
- roberta-base**
Fill-Mask · Updated 13 days ago · 1.86M
- distilbert-base-uncased-finetuned-sst-2-english**
Text Classification · Updated Feb 9 · 1.57M
- gpt2**
Text Generation · Updated May 19 · 1.22M

HF Models

- Model card with information about the model
- Inference API for testing functionality
- Tags used for searching

 **Hugging Face**

[Models](#) [Datasets](#) [Resources](#) [Solutions](#) [Pricing](#) [Log In](#)

gpt2

[Text Generation](#) [PyTorch](#) [TensorFlow](#) [JAX](#) [TF Lite](#) [Rust](#) [Transformers](#) [en](#) [mit](#) [gpt2](#) [lm-head](#) [causal-lm](#) [exbert](#)

[Model card](#) [Files](#) [Train](#) [Deploy](#) [Use in Transformers](#)

GPT-2

Test the whole generation capabilities here:
<https://transformer.huggingface.co/doc/gpt2-large>


Pretrained model on English language using a causal language modeling (CLM) objective. It was introduced in [this paper](#) and first released at [this page](#).

Disclaimer: The team releasing GPT-2 also wrote a [model card](#) for their model. Content from this model card has been written by the Hugging Face team to complete the information they provided and give specific examples of bias.

Model description

GPT-2 is a transformers model pretrained on a very large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

Downloads last month
1,222,320



Hosted inference API

Text Generation

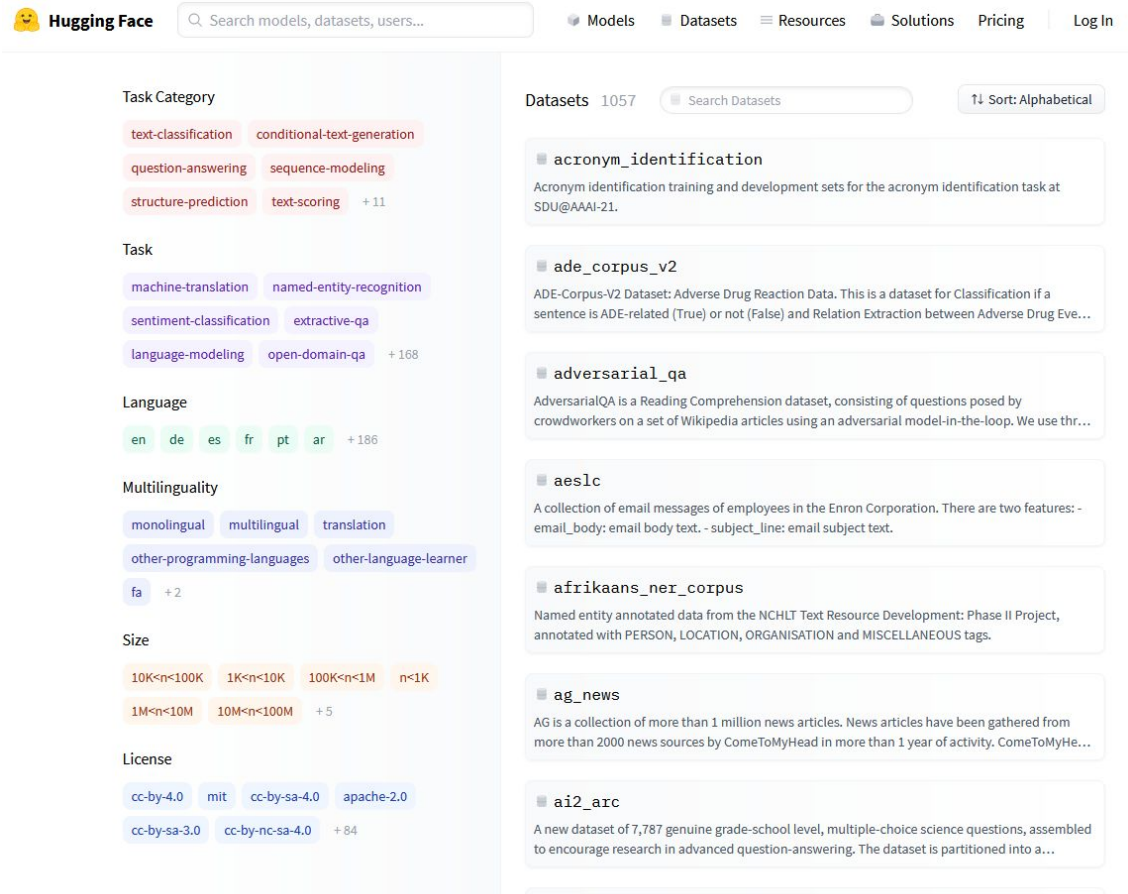
[Compute](#)

This model is currently loaded and running on the Inference API.

JSON Output [Maximize](#)

HF Datasets

- Web accessible catalog of datasets uploaded by organizations and individuals
- Searchable by task category, task, language, multilinguality, size, and license



The screenshot shows the Hugging Face Datasets interface. At the top, there's a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Resources, Solutions, Pricing, and Log In. The main content area is divided into two columns. The left column contains filters for Task Category, Task, Language, Multilinguality, Size, and License. The right column displays a list of datasets, including 'acronym_identification', 'ade_corpus_v2', 'adversarial_qa', 'aeslc', 'afrikaans_ner_corpus', 'ag_news', and 'ai2_arc'. Each dataset entry includes its name, a brief description, and a 'Search Datasets' button. The 'Sort: Alphabetical' button is also visible.

Hugging Face Search models, datasets, users...

Models **Datasets** Resources Solutions Pricing Log In

Task Category

- text-classification conditional-text-generation
- question-answering sequence-modeling
- structure-prediction text-scoring + 11

Task

- machine-translation named-entity-recognition
- sentiment-classification extractive-qa
- language-modeling open-domain-qa + 168

Language

- en de es fr pt ar + 186

Multilinguality

- monolingual multilingual translation
- other-programming-languages other-language-learner
- fa + 2

Size

- 10K<=n<=100K 1K<=n<=10K 100K<=n<=1M n<=1K
- 1M<=n<=10M 10M<=n<=100M + 5

License


- cc-by-4.0 mit cc-by-sa-4.0 apache-2.0
- cc-by-sa-3.0 cc-by-nc-sa-4.0 + 84

Datasets 1057 Search Datasets 14 Sort: Alphabetical

- acronym_identification**
Acronym identification training and development sets for the acronym identification task at SDU@AAAI-21.
- ade_corpus_v2**
ADE-Corpus-V2 Dataset: Adverse Drug Reaction Data. This is a dataset for Classification if a sentence is ADE-related (True) or not (False) and Relation Extraction between Adverse Drug Eve...
- adversarial_qa**
AdversarialQA is a Reading Comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles using an adversarial model-in-the-loop. We use thr...
- aeslc**
A collection of email messages of employees in the Enron Corporation. There are two features: - email_body: email body text. - subject_line: email subject text.
- afrikaans_ner_corpus**
Named entity annotated data from the NCHLT Text Resource Development: Phase II Project, annotated with PERSON, LOCATION, ORGANISATION and MISCELLANEOUS tags.
- ag_news**
AG is a collection of more than 1 million news articles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity. ComeToMyHe...
- ai2_arc**
A new dataset of 7,787 genuine grade-school level, multiple-choice science questions, assembled to encourage research in advanced question-answering. The dataset is partitioned into a...

HF Dataset Cards

- Dataset card with information about the dataset
- Links to related sites and models that use the dataset
- Tags used for searching

 **Hugging Face**

[Models](#) [Datasets](#) [Resources](#) [Solutions](#) [Pricing](#) [Log In](#)

Dataset: **snli**

Tasks: **natural-language-inference** Task Categories: **text-classification** Languages: **en** Multilinguality: **monolingual** Size Categories: **100K<n<1M**

Licenses: **cc-by-4.0** Language Creators: **crowdsourced** Annotations Creators: **crowdsourced**

Source Datasets: **extended|other-flicker-30k** **extended|other-visual-genome**

Dataset Structure

- Data Instances
- Data Fields
- Data Splits

Dataset Creation

- Curation Rationale
- Source Data
- Annotations
- Personal and Sen...

Considerations fo...

- Social Impact of ...
- Discussion of Bia...
- Other Known Lim...

Additional Inform...

- Dataset Curators
- Licensing Inform...
- Citation Informat...
- Contributions

Dataset Card for SNLI

Dataset Summary

The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE).

Supported Tasks and Leaderboards

[SembERT](#) (Zhousheng Zhang et al, 2019b) is currently listed as SOTA, achieving 91.9% accuracy on the test set. See the [corpus webpage](#) for a list of published results.

Languages

The language in the dataset is English as spoken by users of the website Flickr and as spoken by crowdworkers from Amazon Mechanical Turk. The BCP-47 code for English is en.

Dataset Structure

[Update on GitHub](#)
[Use in dataset library](#)
[Explore dataset](#)
[Edit Dataset Tags](#)
[Leaderboards on Papers with Code](#)

Homepage:
SNLI homepage



Repository:
Repository:

Paper:
A large annotated corpus for learning natural la...

Leaderboard:
(located on the homepage)


Point of Contact:
Gabor Angeli

Models trained or fine-tuned on snli

-  **cross-encoder/nli-MiniLM2-L6-H...**
Zero-Shot Classification · Updated Jun 21 ...
-  **cross-encoder/nli-deberta-base**

HF Dataset Viewer

- Inspect instances in dataset, features, and different splits before downloading locally
- Produces code for loading the dataset by command line

 **Datasets**

[github/huggingface/datasets](#)

[Docs](#) | [Overview](#) | [Add Dataset](#)

Filter by Tags
Choose an option

Dataset (Size: 683)
snli

Split
train

Offset (Size: 550152)
0 - +

☐ Show Citations

☐ Show List View

☒ Show Features

Code

Tags

- language : en
- task : text-classification
- purpose : NLI
- size : >100k
- language producers : crowdsourced
- annotation : crowdsourced
- tags : extended-from-other-datasets
- license : C, C, , B, Y, -, S, A, , 4, ,, 0

Features
premise × hypothesis × label ×

```
{
  "premise" : "string"
  "hypothesis" : "string"
  "label" : [
    0 : "entailment"
    1 : "neutral"
    2 : "contradiction"
  ]
}
```

| | premise | hypothesis | lat |
|---|--|---|-----|
| 0 | A person on a horse jumps over a broken down airplane. | A person is training his horse for a competition. | 1 |
| 1 | A person on a horse jumps over a broken down airplane. | A person is at a diner, ordering an omelette. | 2 |
| 2 | A person on a horse jumps over a broken down airplane. | A person is outdoors, on a horse. | 0 |
| 3 | Children smiling and waving at camera | They are smiling at their parents | 1 |
| 4 | Children smiling and waving at camera | There are children present | 0 |

















HF Audience

- Individual community members on the forum who use and contribute models and datasets
- Companies, universities and non-profits who also contribute

Organizations

Companies, universities and non-profits are an essential part of the Hugging Face community!

Also check out our awesome list of [contributors](#).

| | |
|--|---|
|  Facebook AI Company · 121 models |  Microsoft Company · 51 models |
|  Hugging Face Company · 7 models |  Google AI Company · 141 models |
|  Bayerische Staatsbibliothek Non-profit · 44 models |  Musixmatch Company · 2 models |
|  CLUE benchmark Non-profit · 11 models |  Allen Institute for AI Non-profit · 53 models |
|  AI Student Society Non-profit |  nboost Company · 4 models |
|  NLP for Healthcare 1 model |  Urduhack Non-profit · 1 model |
|  DeepChem Non-profit · 1 model |  RAPIDS Open GPU Data Science Company · 2 models |
|  Illuin Technology Company · 1 model |  SparkBeyond Company · 1 model |



Fit to DELPH-IN?

On the one hand:

- Greater visibility for datasets and models
- Opportunity to interact with the community

On the other hand:

- Focus is on NLP/ML rather than linguistics
- Audience is unlikely to be familiar with HSPG and MRS
- API and widget don't support parsing(?)



Your Thoughts?