Introduction to Natural Language Processing (NLP) with Twitter and HuggingFace

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

Hollie Johnson

National Innovation Centre for Data

Plan for today

- The National Innovation Centre for Data
- What makes good data science
- Introduction to Natural Language Processing
- Discussion
- Practical session

About me

- Maths undergrad → (👩🎪) → Comp sci conversion → Stats PhD (Topological event history analysis for global wind data)
- Previously worked as software developer, RA, now senior data scientist at NICD
- Collector of hobbies 🏋️🚴🎻👩

National Innovation Centre for Data

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data

The situation

- National data skills shortage
- Recruitment challenges
- Data quality issues
- Cost of data - **data is not inherently an asset!**
- Low level of data readiness in organisations

Who we are

Our goal is to be a hub for data innovation and data skills. We are not academics, nor are we a traditional consultancy.

We wear many hats, but our main activity is running **collaborative data skills projects** with external organisations.

WiFi: Catalyst | 👩‍⚕️ oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data

Who we work with

- Public sector
- SMEs
- Voluntary sector
- Large corporations
- Start-ups
- University departments

Data skills projects

- Owned by the organisation
- A mixture of collaborative skills transfer sessions and offline work
- Solving a real business problem
- Their problem, their people, their data → their solution

Organisations finish with valuable project output, but more importantly their people have gained new data skills to take forward to future projects.

Data skills projects

- Prepare and deliver conventional teaching materials
- Pair programming with client
- Read and understand cutting edge research to find what is relevant
- Help organisations understand the options available
- Oversee the ‘data science process’

…and many other things!

What makes a good data science project?

A successful data science project is one that delivers *value* to an organisation.

Value can look like:

- increased sales
- reduced churn
- better resourcing (human or otherwise)
- reduced waste/increased efficiency

Research shows most project failures are due to poor project management and scoping

Asking the right questions is as important as having the required technical capabilities

NICD data science workflow

- business understanding
  - *goal*
  - *objectives*
  - *deliverables*
  - *resources*
- data preparation
- modelling
- deployment + monitoring

Defining the goal, objectives, deliverables and resources forms part of our popular Data Science Kick-off workshop.

Getting started with data science

As an organisation:

- consider current use of data:
  - is it effective
  - is the data strategy supported by the organisational culture?
  - who/what drives decisions on data projects?
- consider current business challenges:
  - what challenges are you facing that can be addressed with data?
  - pick the lowest hanging fruit first
- aim for small wins, often
- iterate!

Getting started with data science

As an individual:

- data science needs varied skill sets
  - how do your existing skills can support data science projects
  - where can you add value to your organisation?
- development of technical skills - science or engineering
- development of data literacy and communication

Natural Language Processing

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data

Natural Language Processing

- What is NLP?
- A brief history
- Word embeddings
- Modern architectures

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

What is NLP?

NLP (Natural Language Processing) is a field covering many language-based tasks, including

- text segmentation
- named entity recognition
- **sentiment analysis**
- text summarisation
- machine translation

and is used for chatbots/virtual assistants, spam filtering, web search and text analysis.

A brief history *

- 1940s: NLP emerged following WWII and the desire for automated language translation
- 1960s: Researchers split between symbolic and stochastic NLP
- 1970s: Further research into areas including logic-based paradigms, natural language understanding, discourse modelling
- 1980s: Move towards probabilistic models, information generation and extraction
- 1990s: Introduction of the LSTM model
- 2010s: Popularisation of deep learning methods including techniques such as word embeddings
- 2020s: Dominance of the transformer architectures

Text processing

'Traditional' NLP requires text preprocessing, since most machine learning models require numeric data as input.

The main stages of text preprocessing are

- cleaning
- tokenisation
- stop word removal
- stemming

Cleaning

Data cleaning will vary depending on the context but may include tasks such as

- removal of punctuation
- removal of digits
- conversion to lower case
- removal or replacement of special cases (e.g. @hollie → @user)

Tokenisation

Tokenisation is how we divide text into individual tokens. Commonly, a token corresponds to a word, however a tokens could also be

- characters
- subwords
- words
- n-grams

Stop words

Stop words can be

- 🌎 Global: always have low information
- Subject-specific: may have low information in a given context (e.g. ‘coffee’ if considering reviews of a cafe)
- Document-specific: may have low information in given documents (e.g. words appearing in a document header)

Stemming

Stemming is the replacement of words with a stem word. For instance

- walking, walked → walk
- story, stories → stori

This can help to reduce the number of unique tokens. But be careful - meaning can be lost!

Lemmatisation is an alternative method that uses a morphological analysis of words to preserve meaning.

How do we develop meaningful representations of language?

Bag of Words

A unique token is assigned to each word in the text

- allows development of identity between words
- all ordering is discarded
- no notion of similarity

How do we develop meaningful representations of language?

Bag of Words

```
"it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
"it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
"it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
```

Common words can dominate, potentially without much information, so frequency can be penalised using **Term Frequency - Inverse Document Frequency** (TF-IDF), allowing distinct words in a given document to have more weight.

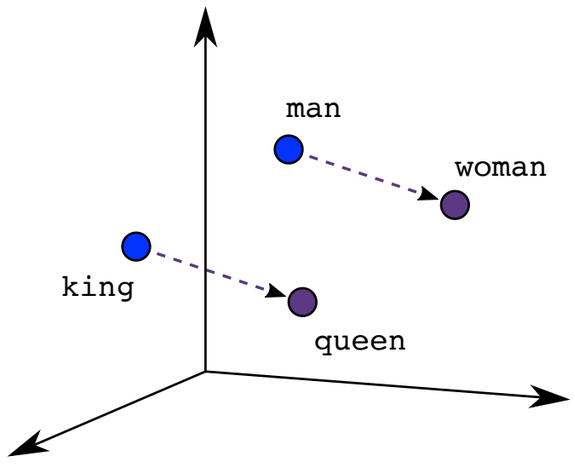
How do we develop meaningful representations of language?

Word embeddings

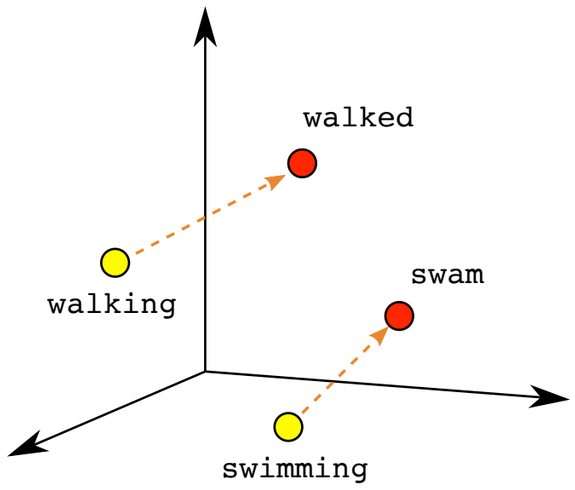
An unsupervised learning approach that learns continuous multidimensional vector representation for each word, by learning to predict a ‘centre word’ given a fixed side window around it

- goal is to capture meaning within the embedding
- the surrounding words are used to capture meaning
- common methods are Word2Vec (Google) and GloVe (Stanford)

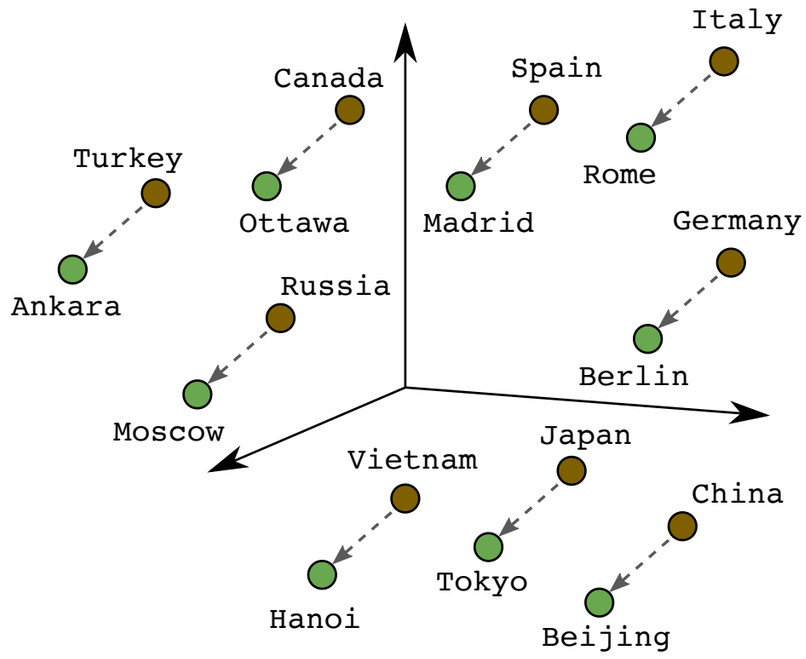
How do we develop meaningful representations of language?



Male-Female



Verb Tense



Country-Capital

developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

Modern architectures

Modern NLP methods work very differently

- do not require the same preprocessing steps
- current SoTA
- use self-attention mechanisms and positional encodings
- Large Language Models are trained in a self-supervised way, then customised for specific tasks
- transformer based architectures

BERT

Bidirectional Encoder Representations from Transformers

Can be used for

- token classification (e.g. named entity recognition or question answering)
- text classification (e.g. sentiment analysis)
- text-pair classification (e.g. sentence similarity)

Not suitable for

- text generation
- machine translation
- text summarisation

BERT

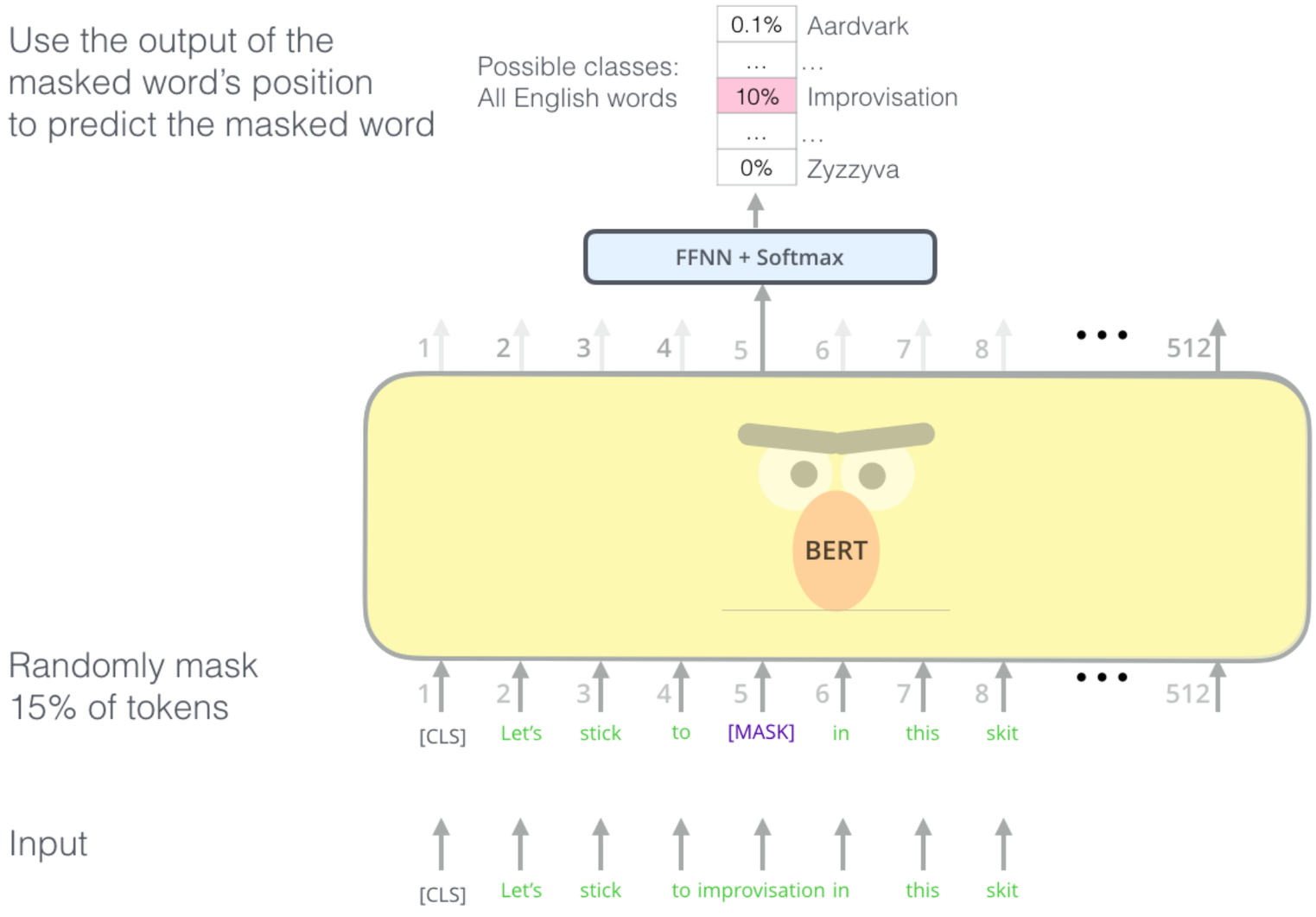
BERT's novelty lies in the way it was pre-trained:

- masked language model (MLM) – randomly mask some of the tokens from the input and predict the original vocabulary id of the masked token
- next sentence prediction (NSP) – predict if the two sentences were following each other or not

These two tasks are trained concurrently.

BERT

Use the output of the masked word's position to predict the masked word

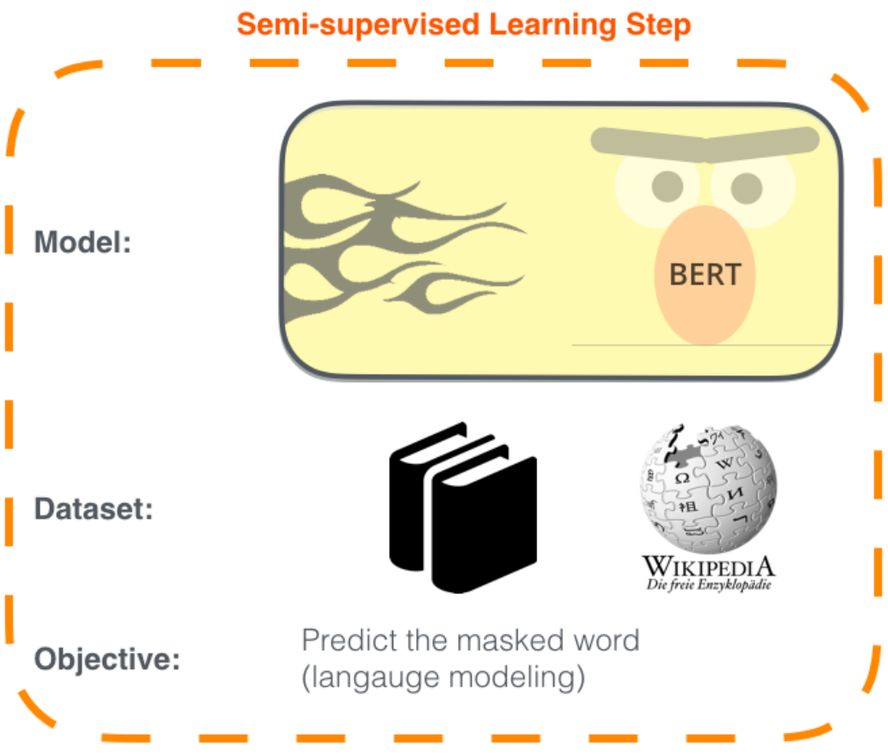


jalammar.github.io/illustrated-bert

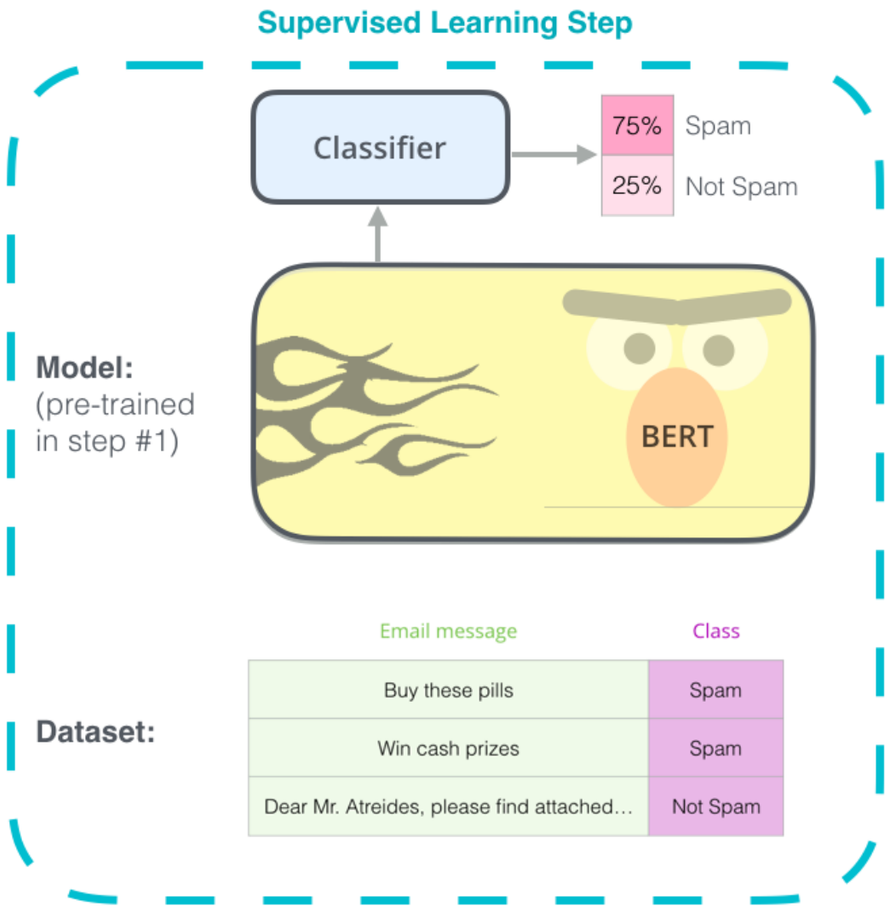
BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



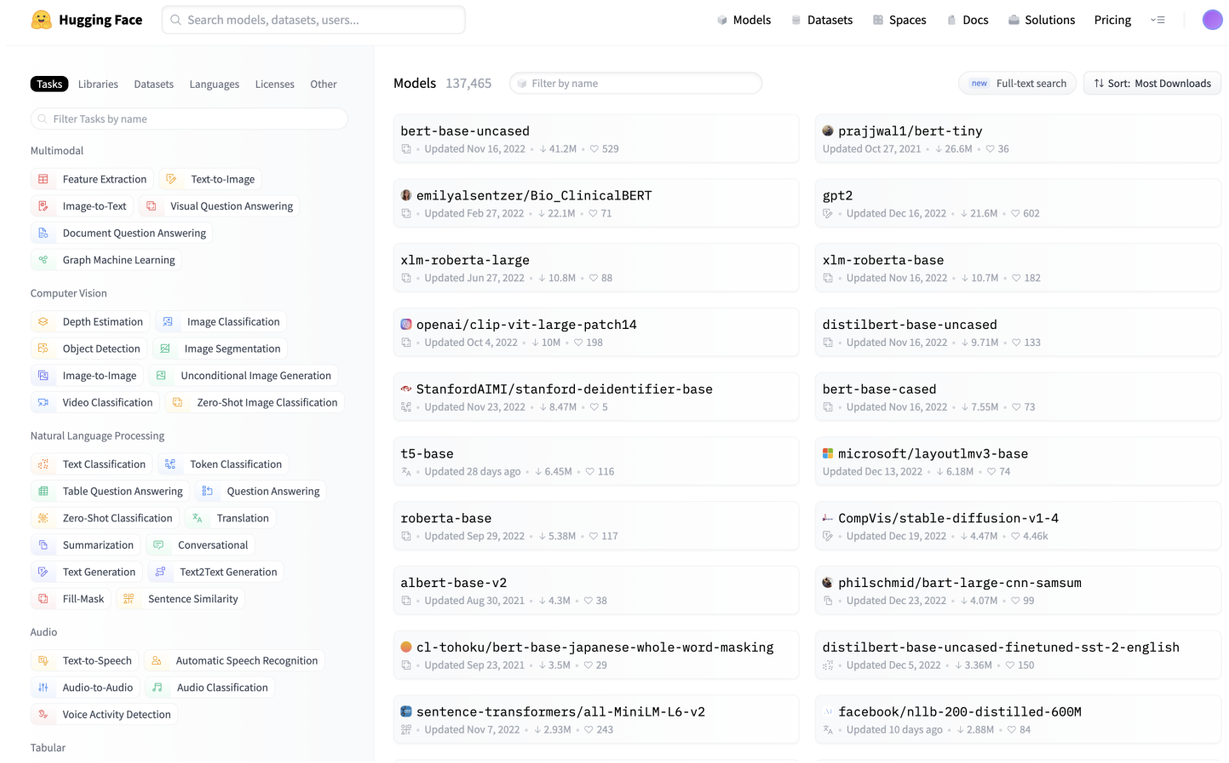
2 - Supervised training on a specific task with a labeled dataset.



jalammar.github.io/illustrated-bert

In practice

We do not need to implement (or train) any of these ourselves.



huggingface.co/models

Many pre-trained models are available, some of which are also fine tuned for specific tasks, so we will take advantage of this and use a model from HuggingFace 🤗

National Innovation Centre Data

Discussion

Over to you

- What sort of language data do you encounter in your roles and what problems/challenges do you face with this?
- What regulations do you think should be placed on the use of language models (if any)?
- What does the future look like?

There are post-its on the tables, please add your thoughts to the large sheets of paper on the walls

10:00

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data

Workshop

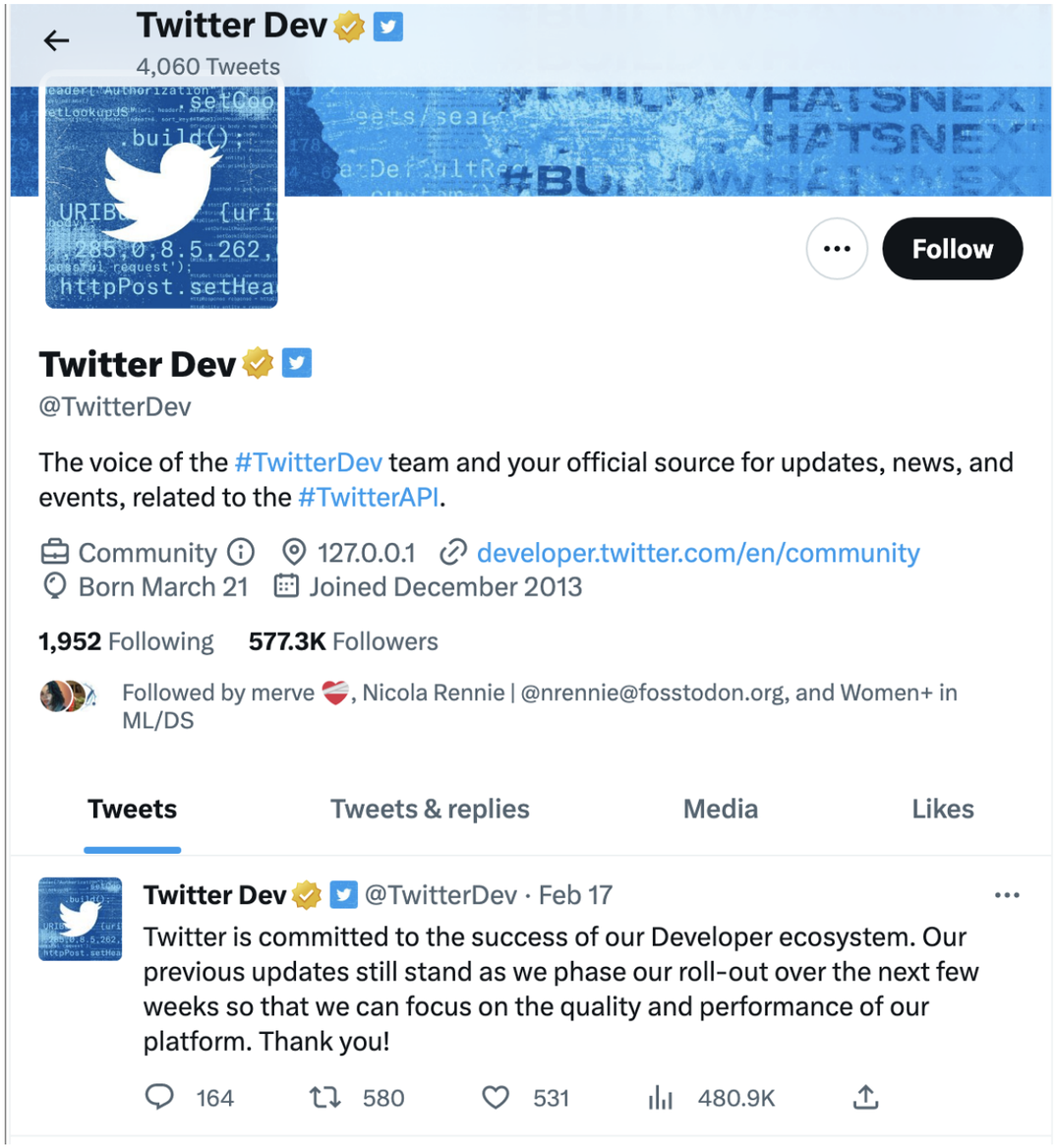
WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

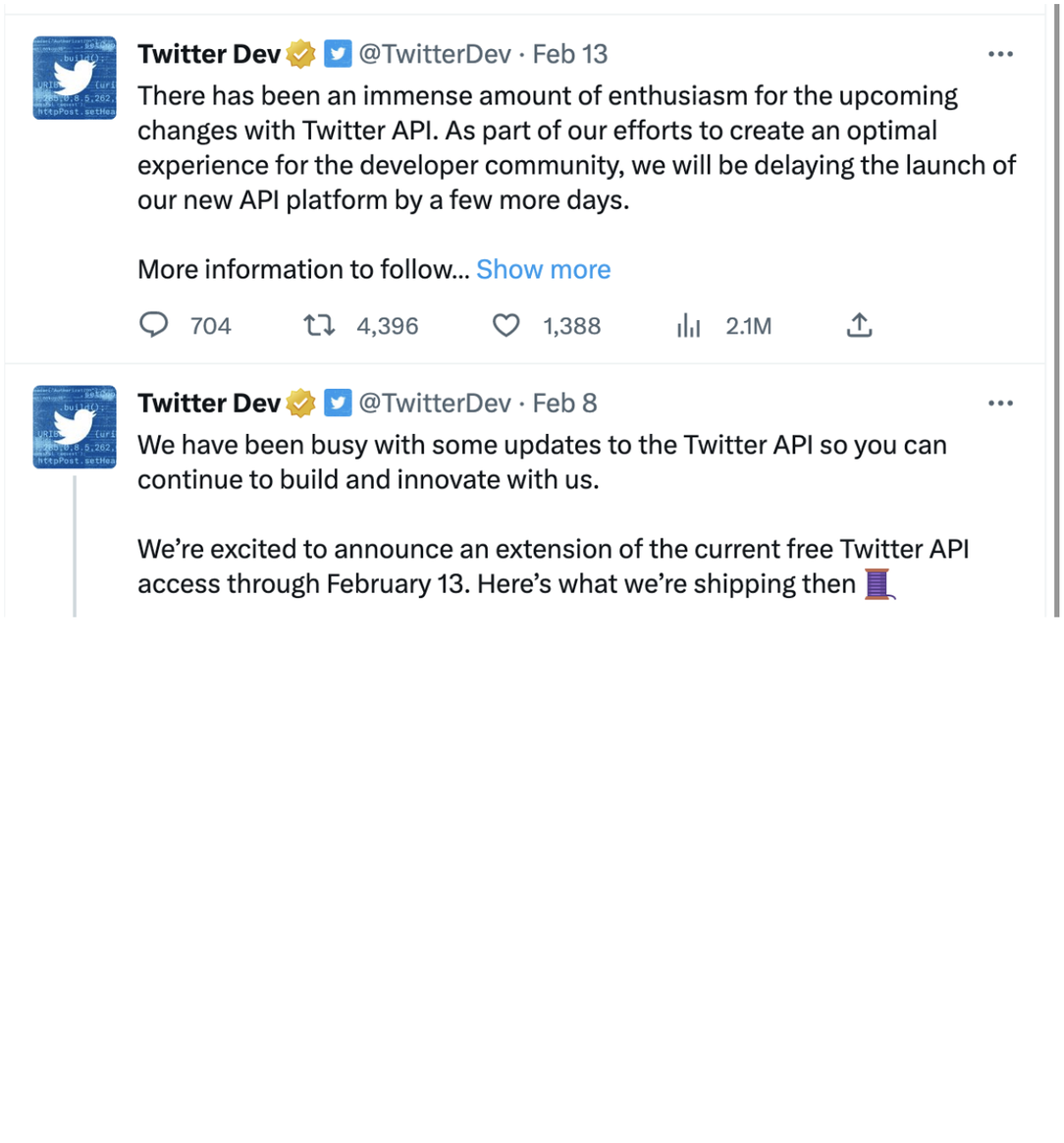
National Innovation Centre Data

The Twitter API saga…

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data


Twitter Dev
4,060 Tweets
Follow
Twitter Dev
@TwitterDev
The voice of the #TwitterDev team and your official source for updates, news, and events, related to the #TwitterAPI.
Community
127.0.0.1
developer.twitter.com/en/community
Born March 21
Joined December 2013
1,952 Following
577.3K Followers
Followed by merve, Nicola Rennie | @nrennie@fosstodon.org, and Women+ in ML/DS
Tweets
Tweets & replies
Media
Likes
Twitter Dev @TwitterDev · Feb 17
Twitter is committed to the success of our Developer ecosystem. Our previous updates still stand as we phase our roll-out over the next few weeks so that we can focus on the quality and performance of our platform. Thank you!
164
580
531
480.9K


Twitter Dev @TwitterDev · Feb 13
There has been an immense amount of enthusiasm for the upcoming changes with Twitter API. As part of our efforts to create an optimal experience for the developer community, we will be delaying the launch of our new API platform by a few more days.
More information to follow... Show more
704
4,396
1,388
2.1M
Twitter Dev @TwitterDev · Feb 8
We have been busy with some updates to the Twitter API so you can continue to build and innovate with us.
We're excited to announce an extension of the current free Twitter API access through February 13. Here's what we're shipping then

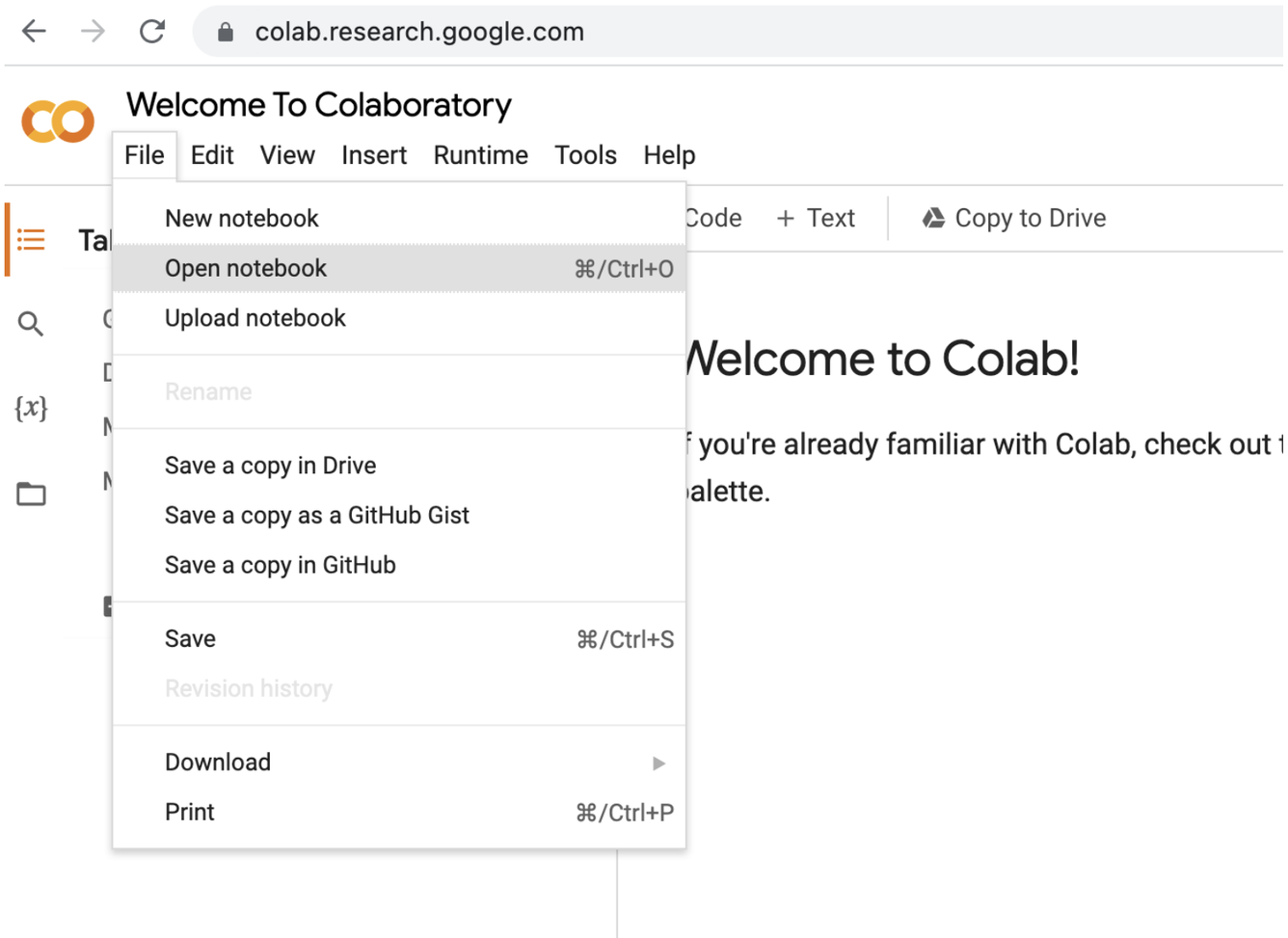
Getting set up

- Connect to WiFi - details at the bottom of the slide
- Open the repository for this session on GitHub https://github.com/NICD-UK/IWD-twitterxhuggingface
- Open GoogleColab at colab.research.google.com

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

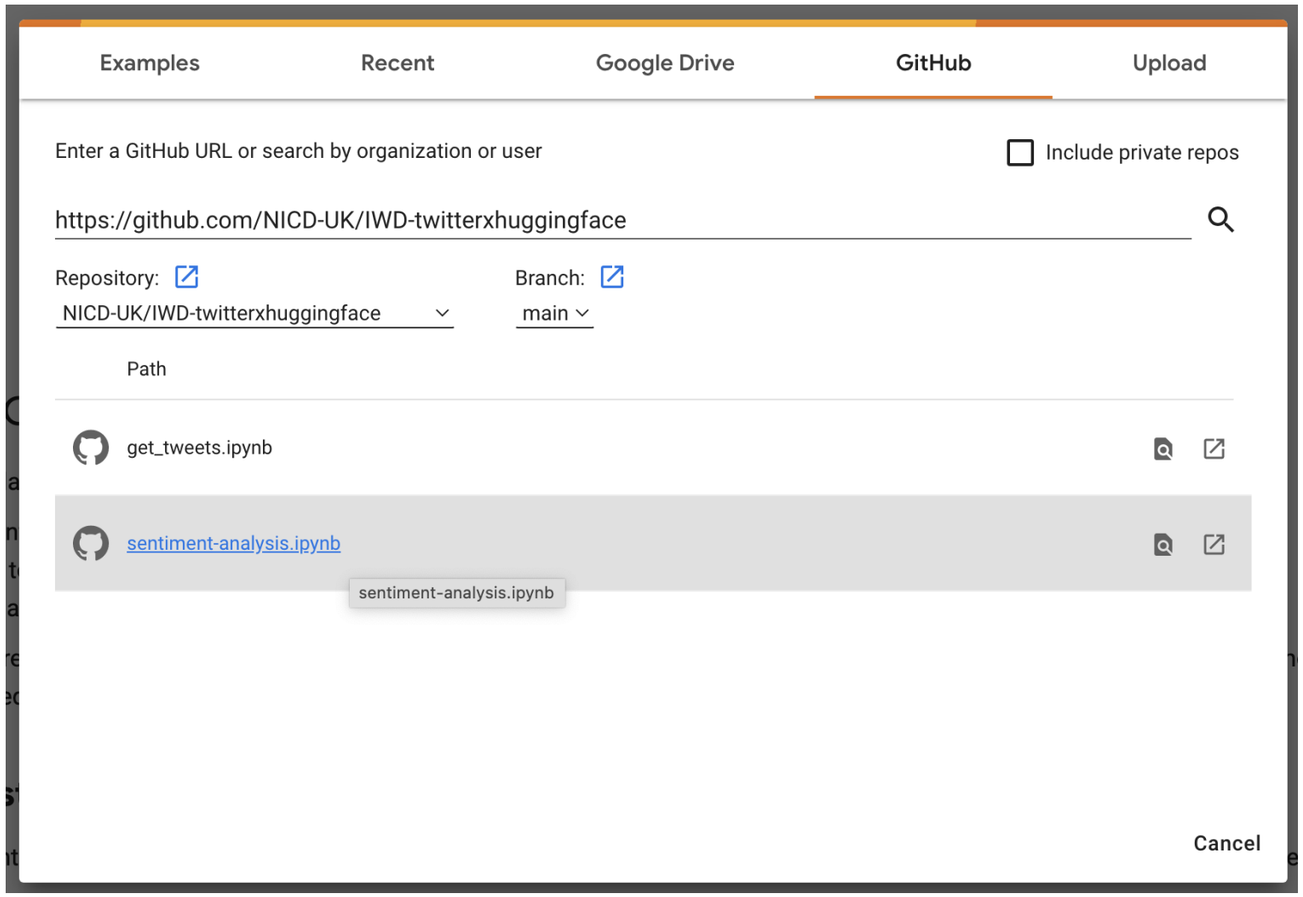
Getting set up

In GoogleColab, select **File**, then **Open notebook**



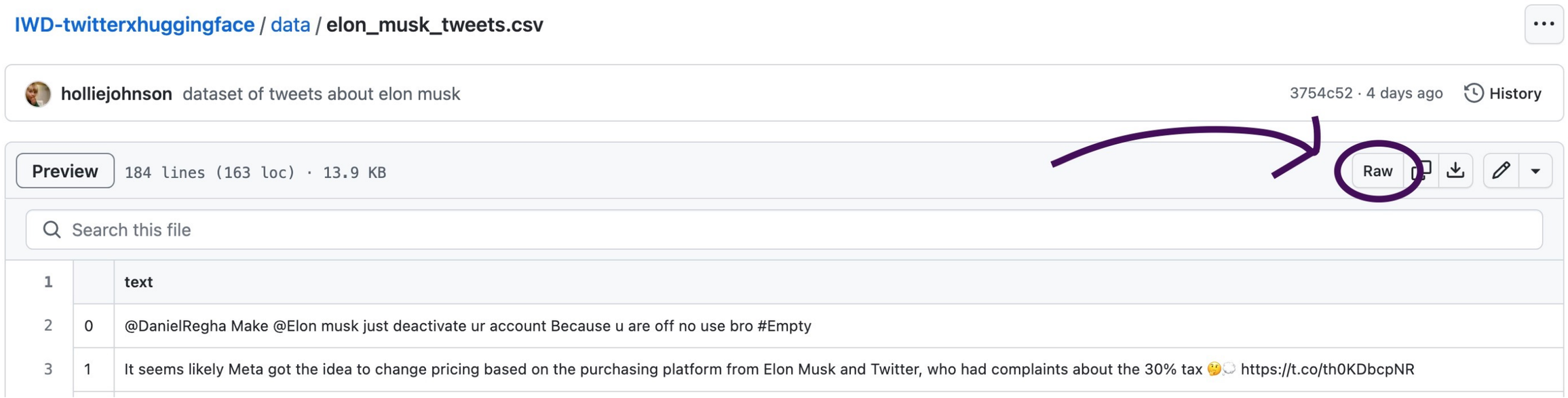
Getting set up

Paste in the url to the GitHub repository and select the notebook **sentiment-analysis.ipynb**



Data

There are pre-saved data sets available at the same GitHub repository. To access the url for the raw data, navigate to the data set you wish to use and click `raw`.



Data

> ⚠ **Warning**
>
> The data set comprises actual tweets obtained using the free API. These have NOT been filtered for toxicity, profanity etc.

HuggingFace model

For our model, we will be using **Twitter-roBERTa-base for Sentiment Analysis**

- roBERTa base model trained on ~124M tweets from January 2018 to December 2021
- Fine-tuned for sentiment analysis with the TweetEval benchmark
- English language
- Encoder transforms tweets into tokens that can be used by the model

Thanks for attending!

- We're always eager to hear about real-life use cases of the content we share. If you end up using any of these materials in your organisation, please let us know how.
- Feel free to contact me hollie.johnson@ncl.ac.uk

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data

WiFi: Catalyst | 👰 oxinn-event | 🔑 event-6YJh!

National Innovation Centre Data