

# A keyphrase suggestion engine for semi-automated document characterization

**Nicholas Taylor** on behalf of James Powell, Dylan Johnson, Tim Mandzyuk, Daniel Waybright, and Alex Shocklee

[Research Library](#)

16 November 2023

LA-UR-23-32817





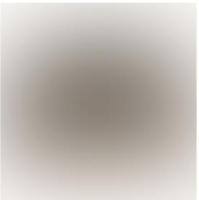

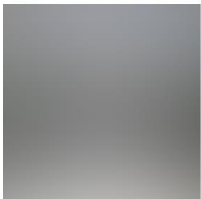

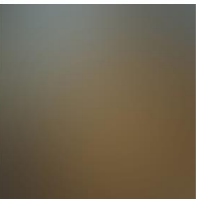
# problem statement



- self-deposit users are predominantly task-driven
- users like more and higher-quality metadata, for discovery
- users are more ambivalent about providing supplemental metadata, when depositing

**how can we streamline the generation of authoritative, supplemental metadata while leveraging the author's expertise?**

[“Reverse-CAPTCHA.png”](#)  
by Brogue Lessor Jig  
under [CC BY-SA 4.0](#)

Select all squares with **cars**

|   |   |   |
|---|---|---|
|  |  |  |
|  |  |  |
|  |  |  |



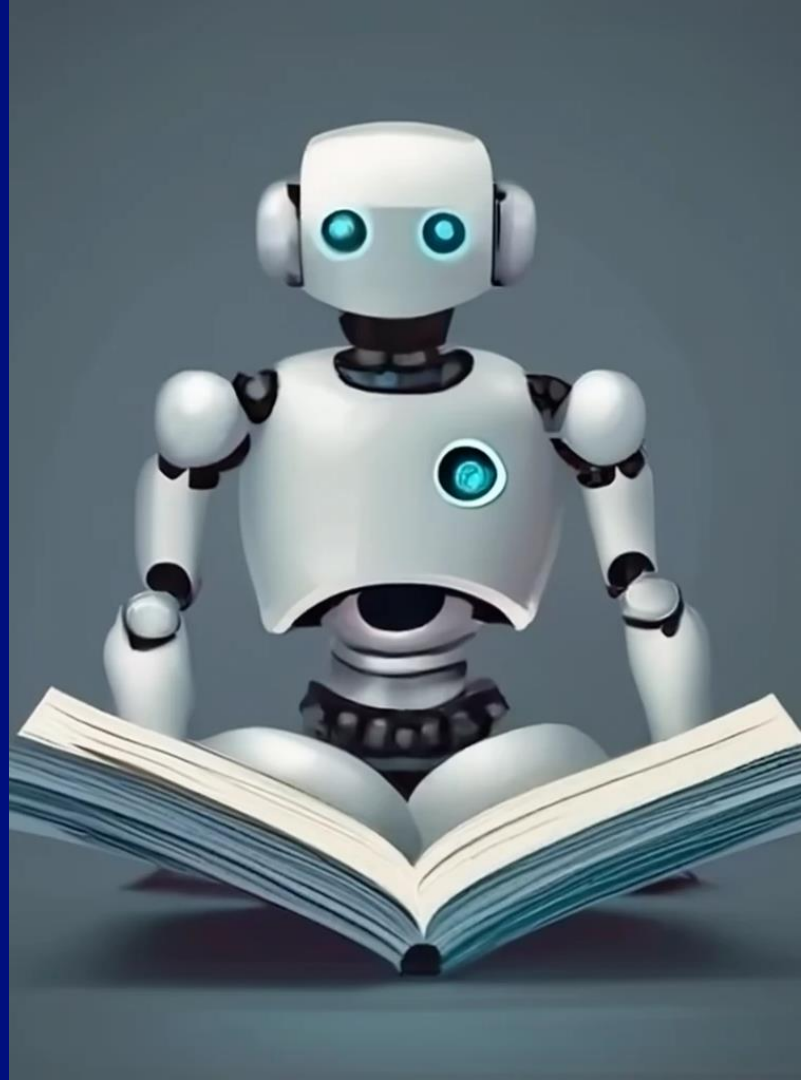
# solution?

- hire (many) more staff
- train them on breadth of LANL science
- train them on domain vocabularies
- have them read every submission
- have them supply keyphrases
- have authors validate quality



# solution!

- apply ML and NLP tools
- have those “read” submissions
- then have them suggest keyphrases
- validate quality of metadata output



# scope

- standalone web service
- input abstract full text
- output scored keyphrase suggestions
- model retraining optional
- performance on commodity hardware
- detailed logging

“checklist” by [cylonfingers](#)  
under [CC BY-SA 2.0](#)

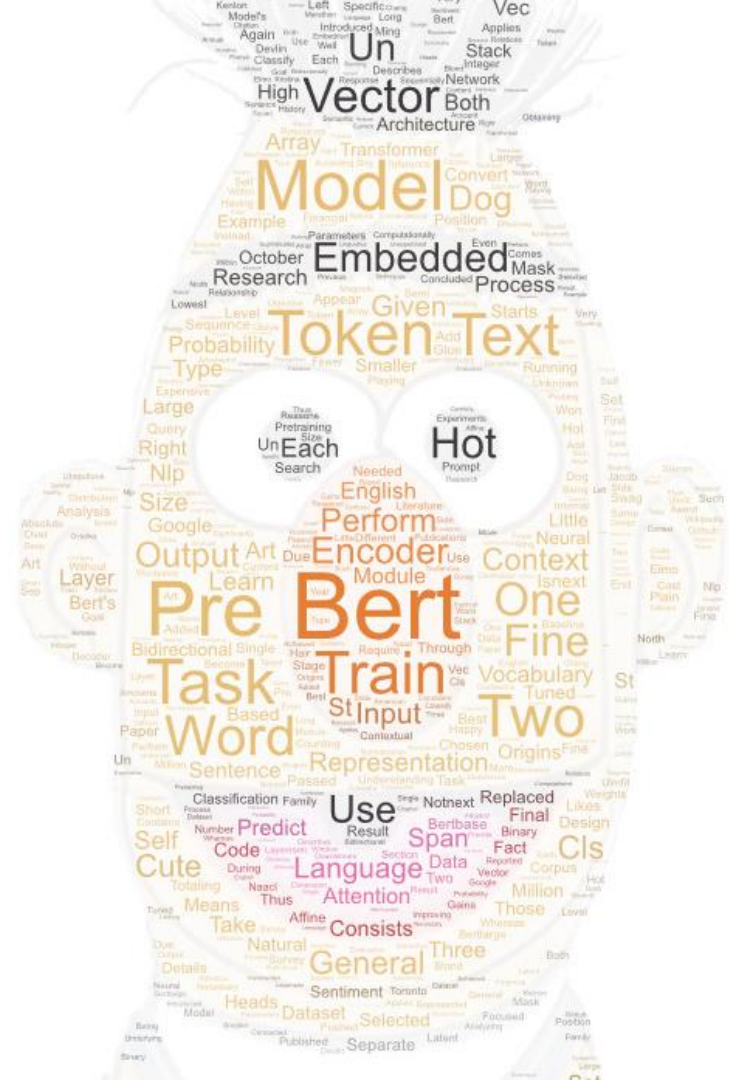




## solution


- **Bidirectional Encoder Representations from Transformers (BERT)**
- **general-purpose NLP tool**
- **superior to dumb statistical techniques as it understands semantic similarity**
- **lighter-weight, performant models: KeyBERT, SciBERT, DistilBERT**
- **runs on standard VM (for now)**
- **JSON API**

generated image for prompt “a jumble of words that together resemble bert from sesame street” by [Craiyon](#), processed through [Word Art](#) with text from the “[BERT \(language model\)](#)” [Wikipedia](#) page



# keyphrase app test interface

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

**SIGNED IN AS**  
[ntay@lanl.gov](#)

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ SciBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ DistilBERT  
Similarity:   
Confidence:   
Keyphrase #:

**Keybert Input** (required)


Text to be keybert-ed...

{1000 character limit}

**CRUNCH**

choose model(s)

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

SIGNED IN AS  
ntay@lanl.gov

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT

Similarity:

Confidence:

Keyphrase #:

☒ SciBERT

Similarity:

Confidence:

Keyphrase #:

☒ DistilBERT

Similarity:

Confidence:

Keyphrase #:

Keybert Input (required)

{1000 character limit}


CRUNCH



# specify similarity

- floating point number, 0-1
- maximum allowed similarity between generated keyphrases
- if similarity of any given pair exceeds threshold, lower-confidence keyphrase discarded

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

SIGNED IN AS  
ntay@lanl.gov

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT

Similarity:

Confidence:

Keyphrase #:

☒ SciBERT

Similarity:

Confidence:

Keyphrase #:

☒ DistilBERT

Similarity:

Confidence:

Keyphrase #:

**Keybert Input** (required)

Text to be keybert-ed...


{1000 character limit}

CRUNCH

# specify confidence

- floating point number, 0-1
- minimum confidence for a suggested keyphrase
- keyphrase not suggested if confidence subceeds threshold

LANL Inside   LANL Phonebook   LANL Research Library

 **Research Library**

SIGNED IN AS  
ntay@lanl.gov

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ SciBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ DistilBERT  
Similarity:   
Confidence:   
Keyphrase #:

**Keybert Input** (required)

Text to be keybert-ed...


{1000 character limit}

**CRUNCH**

# specify number of keyphrases

- whole number
- maximum number of keyphrases to generate

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

SIGNED IN AS  
ntay@lanl.gov

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ SciBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ DistilBERT  
Similarity:   
Confidence:   
Keyphrase #:

**Keybert Input** (required)

Text to be keybert-ed...


{1000 character limit}

CRUNCH

# supply full text

- 1,000 character limit
- designed to accommodate typical journal article abstract length

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

SIGNED IN AS  
ntay@lanl.gov

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ SciBERT  
Similarity:   
Confidence:   
Keyphrase #:

☒ DistilBERT  
Similarity:   
Confidence:   
Keyphrase #:

**Keybert Input** (required)  
  
{1000 character limit}

CRUNCH

# example values

- all models selected
- similarity: .9
- confidence: .35
- keyphrases: 5
- input: (abstract full text from [first COVID-19 pre-print](#) posted on [arXiv](#))

[LANL Inside](#) [LANL Phonebook](#) [LANL Research Library](#)

 **Research Library**

**SIGNED IN AS**  
[ntay@lanl.gov](#)

## BERT API Test Page

Select BERTs to use (required)

☒ KeyBERT

Similarity:

Confidence:

Keyphrase #:

☒ SciBERT

Similarity:

Confidence:

Keyphrase #:

☒ DistilBERT

Similarity:

Confidence:

Keyphrase #:

**Keybert Input** (required)

The 2019 novel coronavirus (2019-nCoV) is currently causing a widespread outbreak centered on Hubei province, China and is a major public health concern. Taxonomically 2019-nCoV is closely related to SARS-CoV and SARS-related bat coronaviruses, and it appears to share a common receptor with SARS-CoV (ACE-2). Here, we perform structural modeling of the 2019-nCoV spike glycoprotein. Our data provide support for the

(1000 character limit)

**CRUNCH**

# example outputs

Mark the accepted keywords. (required)

| KeyBert   | SciBert  | DistilBert   |
|---|--|--|
| <input type="checkbox"/> novel coronavirus (0.5918)       | <input type="checkbox"/> novel coronavirus (0.5555)      | <input type="checkbox"/> structural loop (0.9999413)         |
| <input type="checkbox"/> other coronaviruses (0.5677)     | <input type="checkbox"/> bat coronaviruses (0.4973)      | <input type="checkbox"/> structural modeling (0.99992114)    |
| <input type="checkbox"/> bat coronaviruses (0.5453)       | <input type="checkbox"/> other coronaviruses (0.493)     | <input type="checkbox"/> receptor binding module (0.9998897) |
| <input type="checkbox"/> ncov spike glycoprotein (0.5177) | <input type="checkbox"/> ncov spike glycoprotein (0.456) | <input type="checkbox"/> coronavirus (0.99986994)            |
| <input type="checkbox"/> common receptor (0.4064)         | <input type="checkbox"/> china (0.3574)                  | <input type="checkbox"/> fusion (0.9943869)                  |

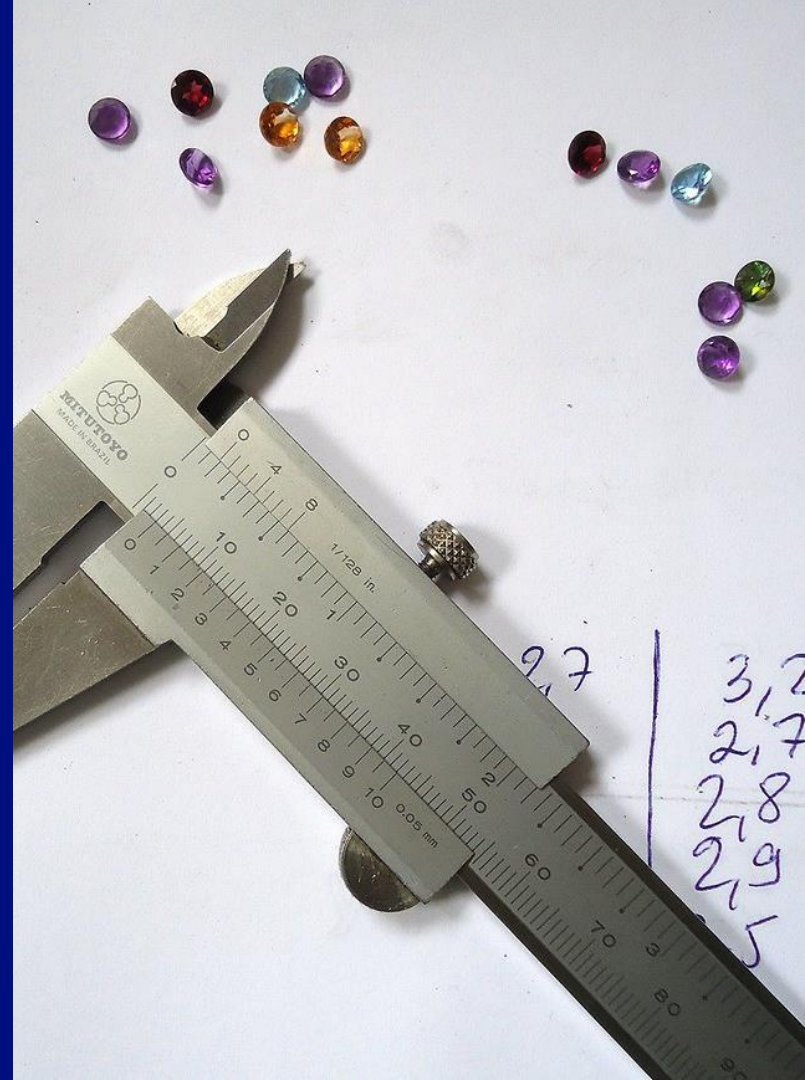
Add additional keywords here (separated by commas).

Transaction ID: 1b0e8b61-eac8-45b2-8727-d791c4cedac1

**SUBMIT**

# evaluating quality

- only informally, internally so far
- start with close partners
- (if sufficiently fast) deploy in production, leverage analytics to iterate





# how we'll use it

- integrate into self-deposit workflow to suggest candidate keyphrases
- separate work underway for automated extraction of document elements (including abstract) using GROBID
- potentially apply for digitization post-processing or previous submissions?
- prototype LLM-based fielded text extraction?



thank you!

