Detecting Sockpuppetry on Wikipedia Using Meta-learning

Luc Raszewski and Christine de Kock

THE UNIVERSITY OF MELBOURNE

The Idea

Identify groups of sockpuppets (fake accounts) on Wikipedia by their writing style in article contributions.

The Task

Each **task** corresponds to a single sockpuppet **investigation** on Wikipedia, containing a single group of related sockpuppets. The **goal** of each task is to train a classifier that identifies the writing style of the sockpuppet group.

The Problem

Each sockpuppet group has very limited data to train an authorship classifier with, which is why previous approaches struggle.

The Solution

Use **Meta-learning** to train over multiple sockpuppet detection **tasks**, rather than one, to learn a general model that quickly adapts to new writing styles.

The Result

We find that **meta-learning** over a distribution of tasks **significantly improves** prediction precision over pre-trained approaches. This outcome is valuable for sockpuppet detection, where confidence in positive identifications is paramount. Our approach is applicable to other online communities as well.

Other Contributions...

We construct and publicly release a dataset of sockpuppet investigations on Wikipedia. Our dataset improves upon existing datasets which are either outdated, unreleased or do not preserve investigation structure.

We formulate a more realistic task definition. Previous approaches train on data from any number of accounts within a sockpuppet-group, preemptively revealing any deceptive efforts made by a puppetmaster to the model. Our model is fine-tuned on just one accused user, as it would be when deployed.

Sockpuppetry Example

Sockpuppets are accounts operated by a single user. They can be used to manipulate articles to align with a political agenda.

Here, two Sockpuppets are used to vandalise and dispute the neutrality of the “Boxer Rebellion” Wikipedia Article.

Editor 1

Who carries out this examination? That is, if an article is “nominated,” who is it nominated to? I found that TruthorDuty had no Talkpage or edits other than adding this nomination. I can see his or her point, but it would have been more helpful to have specifics.

Editor 2

Undo massive POV edits by KeepitImpartial- if you are trying to spread a pro christian and anti china agenda, do it more subtely, by removing the source before trying to tag information with ``citation needed".

Reverted Paragraph Removal.

TruthOrDuty (Sockpuppet)

This article contains numerous unsourced assertions of a subtly hyperbolic nature. I do not believe that the article, as it now stands, meets the encyclopedic neutrality requirements of Wikipedia. I understand that this article is one that is likely to bring out the more extreme elements in several societies. Therefore, it is extremely important, that the standards of Wikipedia are upheld.

Added a “Neutrality Disputed” Banner.

KeepItImpartial (Sockpuppet)

Need for citations and removal of dubious quotes/sources.

Removed Several Paragraphs and citations. Added citation requests.

The Dataset

A dataset of 23,610 sockpuppet investigations on Wikipedia, where each is a discrete sockpuppet detection task.

Train Set

Positive Samples are contibution messages of one accused Sockpuppet user, called the **Puppetmaster.**

Negative Samples are uniformly distributed contribution messages from non-sockpuppets drawn from the same time and article distributions.

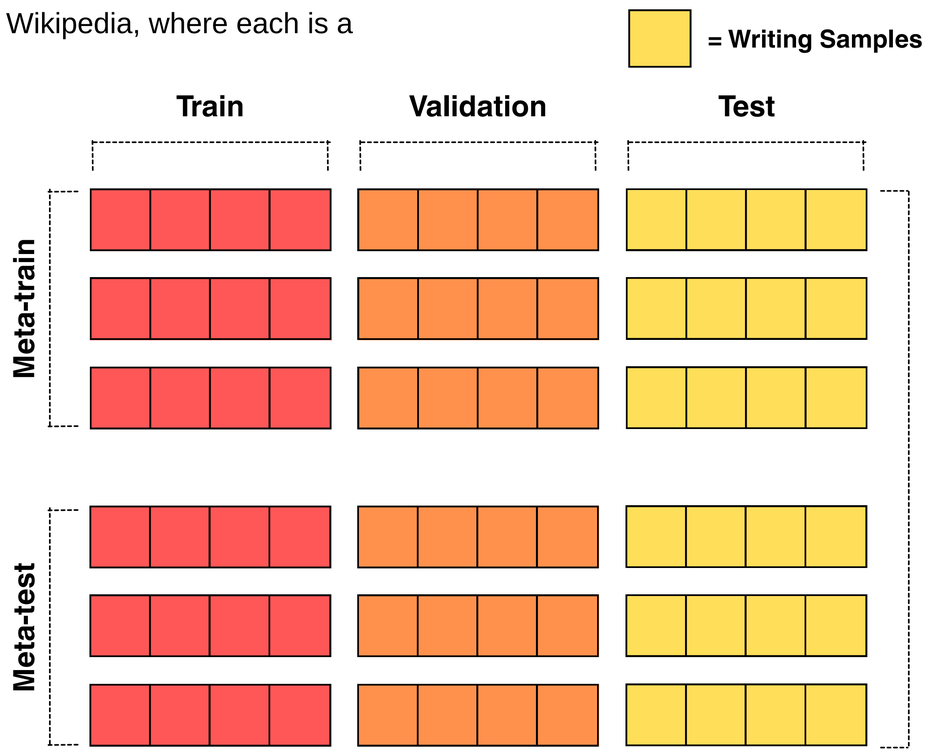
Validation Set

Subset of the training set, 20%.

Test Set

Positive Samples are contribution messages of all remaining accused users (sockpuppets).

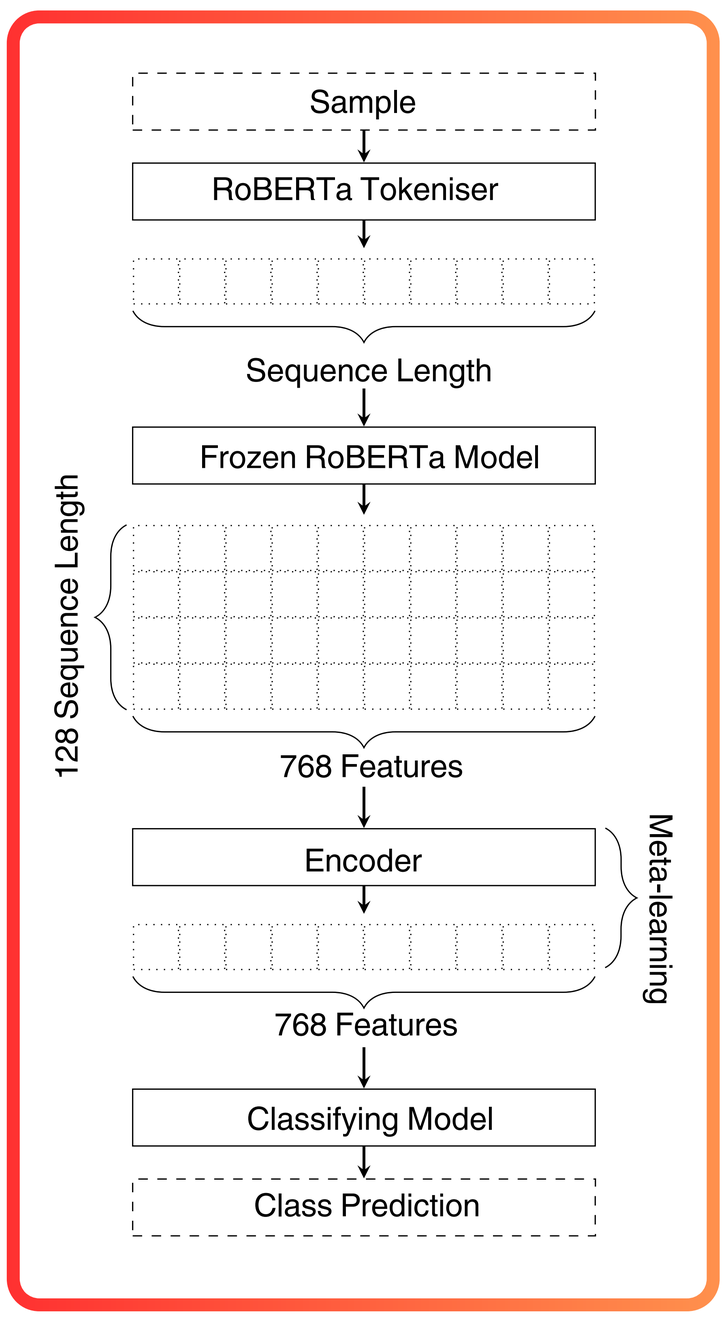
Negative Samples are the same as the training set.



The Model

Writing samples are first tokenized and encoded into a matrix of word-level contextual embeddings by a frozen RoBERTa model. A transformer encoder, meta-learned across tasks in the meta-train set, then interprets the matrix to produce a single authorship embedding.

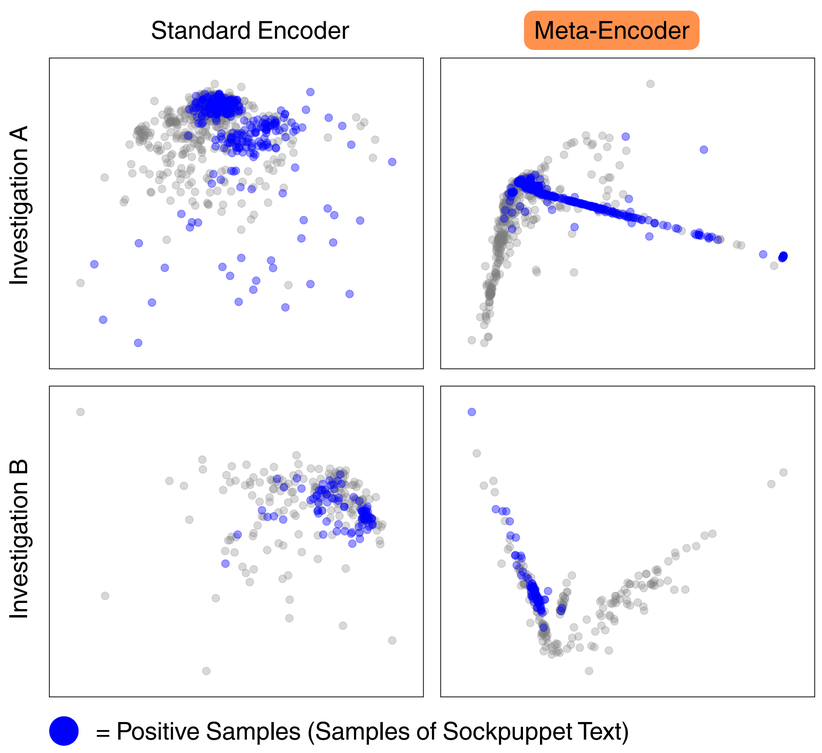
A fully connected, feed-forward neural network interprets the authorship embedding to produce the final class prediction (sockpuppet or not).



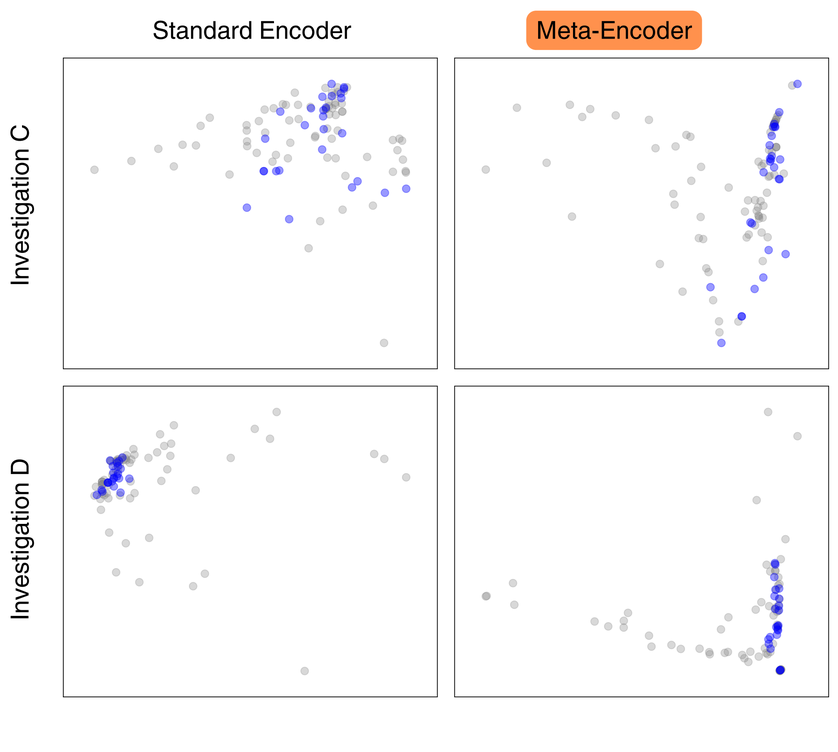
The Results

Approach	AUROC	AUPRC	F1-Score	F0.5-Score	Accuracy	Precision	Recall
Random	50.10 ± 0.14	50.85 ± 0.09	40.34 ± 0.11	36.52 ± 0.12	50.10 ± 0.16	34.46 ± 0.12	50.05 ± 0.15
Majority	-	-	-	-	65.60 ± 0.00	-	-
RoBERTa	65.70 ± 0.00	50.45 ± 0.03	57.97 ± 0.01	57.52 ± 0.06	66.98 ± 0.06	59.54 ± 0.13	67.63 ± 0.17
Standard Enc.	68.33 ± 0.09	50.67 ± 0.33	60.05 ± 0.18	58.73 ± 0.16	68.90 ± 0.07	59.72 ± 0.32	69.88 ± 0.34
Pre-trained Enc.	62.74 ± 0.02	44.80 ± 0.19	57.49 ± 0.13	52.90 ± 0.12	62.79 ± 0.05	51.45 ± 0.25	74.76 ± 0.28
Reptile Enc.	$\mathbf{78.98 \pm 0.12}$*	$\mathbf{62.21 \pm 0.08}$*	$\mathbf{67.46 \pm 0.53}$*	$\mathbf{67.89 \pm 0.17}$*	$\mathbf{77.51 \pm 0.19}$*	$\mathbf{69.43 \pm 0.26}$*	$\mathbf{70.81 \pm 0.82}$
Upper Limit	96.73 ± 0.00	93.56 ± 0.00	86.48 ± 0.00	91.11 ± 0.00	92.01 ± 0.00	95.38 ± 0.00	81.66 ± 0.00

PCA Of High Performing Tasks



PCA Of Low Performing Tasks



= Positive Samples (Samples of Sockpuppet Text)