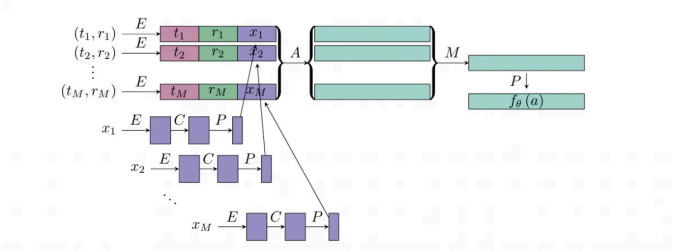
Re-Identifying Banned Social Media Commenters With Machine Learning

Originally published at https://www.unite.ai/re-identifying-banned-social-media-commenters-with-machine-learning/

Researchers from John Hopkins University have developed a Deep Metric approach to identifying online commenters who may have had previous accounts suspended, or may be using multiple accounts to astroturf or otherwise manipulate the good faith of online communities such as Reddit and Twitter.

The approach, presented in a new paper led by NLP Researcher Aleem Khan, doesn't require that the input data be automatically or manually annotated, and improves on the results of previous attempts even where only small samples of text are available, and where the text was not present in the dataset at training time.

The system offers a simple data augmentation schema, with embeddings of different sizes trained on a high-volume dataset containing over 300 million comments covering a million different user accounts.



The model architecture of the John Hopkins re-identification system, where the essential components are 1) text content, 2) a sub-Reddit feature and 3) publication time/date. Source: https://arxiv.org/pdf/2105.07263.pdf

The framework, based on Reddit usage data, considers text content, sub-Reddit placement and time published. The three factors are combined with diverse embedding methods including one-dimensional convolutions and linear projections, and are assisted by an attention mechanism and a max pooling layer.

Though the system concentrates on the text domain, the researchers contend that its approach can be translated to analysis of video or images, since the derived algorithm operates on frequency occurrences at a high level, despite a variety of input lengths for the training data points.

Avoiding 'Topic-Drift'

One trap that research of this nature can fall into, and which the authors have expressly addressed in the design of the system, is to place excessive emphasis on the re-occurrence of particular topics or themes across posts from different accounts.

Though a user may indeed write repetitively or iteratively in a particular strand of thought, the topic is likely to evolve and 'drift' over time, devaluing its use as a key to identity. The authors characterize this potential trap as 'being right for the wrong reasons' – a pitfall previously studied at John Hopkins.

Training Methodology

The system uses mixed precision training, an innovation presented in 2018 by Baidu and NVIDIA , which cuts memory requirements in half by using half-precision floats: 16-bit floating point values instead of 32-bit values. The data was trained on two V100 GPUs, with average training time coming in at 72 hours.

The schema employs simplified text encoding, with convolutional encoders limited to 2-4 subwords. Though the average length for frameworks of this nature is a maximum of five subwords, the researchers found that this economy not only had no impact on ranking performance, but that increasing the subwords to a maximum of five actually *degraded* ranking accuracy.

The Dataset

The researchers derived a dataset of 300 million Reddit posts from the 2020 Pushshift Reddit Corpus dataset, called the Million User Dataset (MUD).

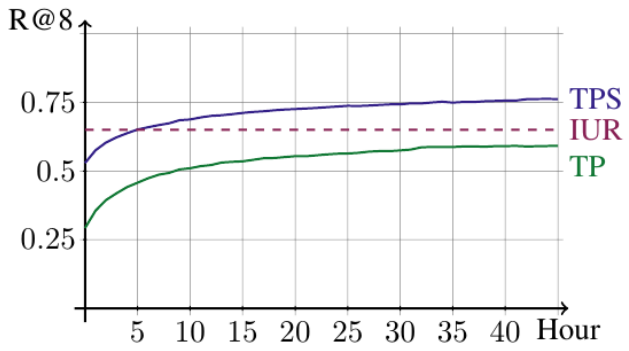
The dataset comprises all posts by Reddit authors that published 100-1000 posts between July 2015 and June 2016. Sampling over time in this way provides an adequate history length for the study, and lowers the impact of sporadic spam posts that are not within the scope of the research's objectives.

Number of users contributing	1,071,477
Number of posts	321,659,421
Mean post length	42.5 tokens
Mean number of posts contributed by a user	300.2
Mean number of subreddits accessed by a user	22.1
Mean number of months a user was active	9.9
Percentage of posts containing more than 64 tokens	17.37%

Statistics on the derived dataset for the John Hopkins re-identifier project.

Results

The image below shows cumulative improvement of results as ranking accuracy is tested at one-hour intervals in training. After six hours, the system outperforms the baseline achievements of related prior initiatives.



In an ablation study, the researchers found that removing the sub-Reddit feature from the workflow had surprisingly little impact on ranking accuracy, suggesting that the system generalizes very effectively, with robust feature tooling.

Posting Frequency As A Re-Identification Signature

This also indicates that the framework is highly transferable to other commenting or publishing systems where only the text content and date/time of publication is available – and, essentially, that the temporal frequency of posting is in itself a valuable collateral indicator to the actual text content.

The researchers note that attempting to perform the same estimation within the content of a single sub-Reddit poses a greater challenge, since the sub-Reddit itself serves as a topic proxy, and an additional schema would arguably be needed to fill this role.

The study was nonetheless able to achieve promising results within these constrictions, with the only caveat that the system works better at high volumes, and may have increased difficulty in re-identifying users where post volume is low.

Developing The Work

In contrast to a great deal of supervised learning initiatives, the features in the Hopkins re-identification schema are discrete and robust enough that the performance of the system improves notably as the volume of data scales up.

The researchers express interest in developing the system by adopting a more granular approach to analysis of publication times, since the often predictable schedules of rote spammers (automated or otherwise) are susceptible to identification by such an approach, and this would make it possible to either more effectively eliminate robot content from a study primarily aimed at vexatious users, or to aid in identifying automated content.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More