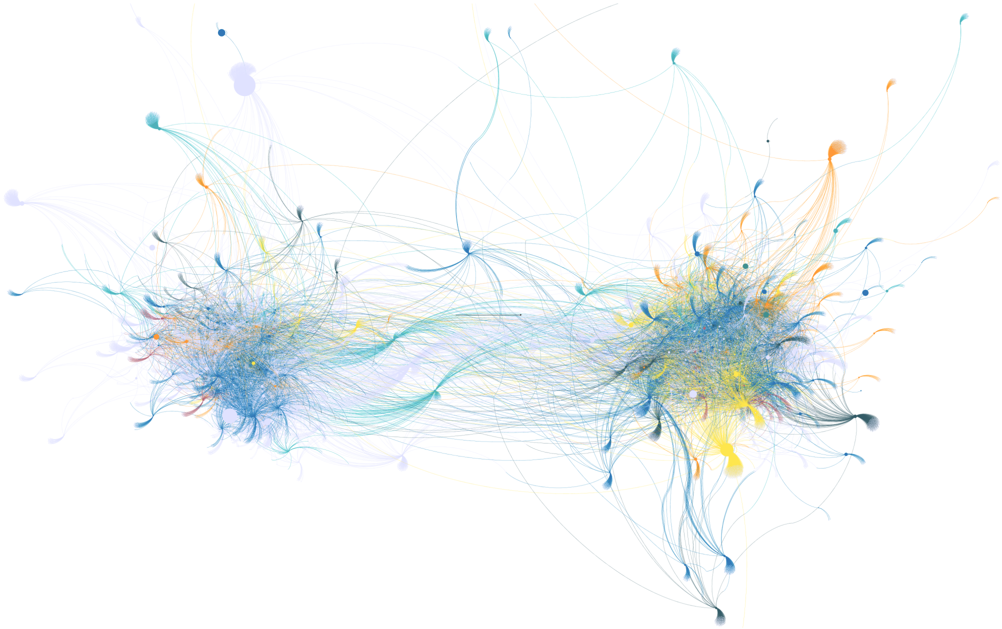
June 2021



Twincler Spark for faster detection of harmful content and platform abuse with large-scale content moderation and recommendation systems

Demo Abstract

June 2021

ABSTRACT

We would like to introduce Twincler Spark, a new technology that generates high-accuracy output and actionable leads for large-scale systems that run on noisy or weak supervision deep learning models such as content moderation or recommendation systems. The goal is to enable those systems to be ahead of disinformation, previously unknown conspiracy theories and platform abuses before they impact users, the platform and off-platform environments. The new approach of Twincler Spark is to understand harmful content and platform abuses as data anomalies in the datastream. Thus, Twincler Sparks allows unbiased flagging of harmful content and untrustworthy accounts related to data anomalies in real-time and to scale across languages. Our demonstration will introduce the new approach as well as showcase two case studies: (1) Detecting harmful content and conspiracy theories timely related to YouTube data in order to prevent multi-task recommenders from jumping on harmful content and untrustworthy accounts and (2) reducing the noise of leads for human content evaluators to review related to Twitter data.

INTRODUCTION

Conventional content moderation systems use either (1) content-based approaches or approaches that are based on (2) virality and reach. Content-based approaches use machine learning, deep learning models in particular. Deep learning models work quite well for harmful content that does not change a lot over time such as some forms of sexual suggestive content, content related to child abuse or violence. However, there are three challenges to deep learning models in the fields of harmful content and platform abuse.

First, creating training sets for deep learning models is expensive and time consuming. It is common that companies hire entire teams to label the data manually. When the model is ready to go, many use cases of harmful content and platform abuse have already changed. Second, deep learning models have a hard time identifying harmful content and platform abuses that are previously unknown or change rapidly such as conspiracy theories. They have a hard time identifying cases they have never been trained on. Third, hand-labeled data is prone to biases. When human annotators are too divided during training over whether content is deceptive and misleading or not, the model receives ambiguous or conflicting feedback. Annotating disagreement is common among training sets that require a high level of domain expertise. Disinformation in particular requires domain training for content evaluators and inhabits a significant number of edge cases such as borderline content. Borderline content allows different perspectives on a political or societal issue. Consequently, the trained model finds itself in a rather confused situation in certain cases of disinformation or may not recognize such content at all.

Approaches that try to identify harmful content and platform abuse through virality and reach classify content based on its amplification. Those correlation-based models detect harmful content when it has already been amplified, ignore nuances of amplification causes and are blind to harmful content that is not boosted. One tactic in the playbooks of information operations is to have harmful content sit on the platform for some time, ready to use in an operation years later . This backdoor content has a minimum reach when it's being posted, oftentimes under 50 views. Approaches based on virality and reach have a hard time identifying such content.

Unbiased data evidence

A new and faster approach is to understand harmful content and platform abuse as data anomalies. Data anomalies are related to both structured and unstructured data. The important difference to conventional approaches is to be able to identify signals of harmful content and scaled abuses more timely and more accurately in the context of the data as it streams on multiple platforms.

There are 3 cornerstones of Twincler design:

- Ability to identify previously unknown abuses and harmful content timely
- Providing unbiased, data-based evidence of abusive behavior
- Output of high-accuracy and actionable leads to support large-scale systems that run on noisy or weak supervision models

The core process of Twincler is to provide early data-driven evidence of harmful content and abusive behavior via API fully automated and in real-time. Twincler's signals are directly sent to internal systems that process and integrate them into the content moderation process. The important detail is that Twincler is built without access to the ground truth of the platforms' policies defining harmful content or platform abuse. By not pre-defining abuse factors, Twincler is particularly effective in identifying previously unknown abuses and providing additional data-driven evidence for edge cases such as borderline content. That said, later versions of Twincler additionally provide certain types of abuse factors in real-time such as coordinated behavior or irregular account activity.

Multi-platform approach

Another important feature of Twincler is the ability to identify, process and weigh signals from multiple data pipelines from different platforms in real-time. Multiple implicit and explicit signals of multiple objectives that sit on multiple platforms and on multiple surfaces are being processed, related and weighed to one another in real-time. This optimizes the results and the accuracy rather than analyzing an isolated activity on just one platform. Twincler is enabled by our Knowledge Base. Our Knowledge Base is a library of anomaly patterns we've evaluated worldwide since 2016 related to (1) events such as elections, terrorist attacks or mass shootings, (2) operational behavior of accounts and certain actors in the past and (3) signature patterns of several platforms about content amplification and account behavior.

Context in real-time

One observation with the early versions of Twincler was that real-world users had difficulties in applying data-based evidence of abusive behavior to internal enforcement policies. For this reason, we implemented additional layers of contextual information such as automated narrative detection or account attribution to the alerts in real-time. This feature enables users to develop their own reasoning within their policy framework.

Evaluation

Twincler has been tested with real-world users and enables teams in three areas:

1. Trust & safety teams such as intelligence desks, content evaluators or policy enforcement. Twincler enables teams to faster detect harmful content and previously unknown abuses by days. Due to high-accuracy leads, Twincler significantly reduces the number of flagged cases to review. Twincler can also help teams to prioritize leads and provide data-driven evidence for reasonable policy enforcement.

2. Engineering teams who work on recommendations and ranking systems. Twincler provides additional layers of unbiased and data-driven evidence of anomalies related to content or accounts. Thus, Twincler prevents recommenders from jumping on content that shows signals of anomalies and enables ranking systems to downrank untrustworthy accounts.

3. Research and development of deep learning models. Twincler provides unbiased and data-driven evidence for labeling content, accounts, abusive behavior and edge cases such as borderline content. With Twincler, the process of hand labeling can be significantly reduced. Creating training sets is cheaper and faster and more accurate.

The key takeaways from evaluating Twincler's performance are:

- Twincler outperformed conventional systems in identifying policy violations by days
- Twincler provides actionable leads
- Twincler significantly reduced the number of leads to review
- Twincler has the ability to automatically identify previously unknown conspiracy theories or scaled abuses
- Twincler offers a faster way to detect harmful content and platform abuse at scale

DEMONSTRATION

Our demonstration consists of two main parts: First, the introduction of our new approach, the underlying paradigm and the evaluation of Twincler Spark with real-world users. Second, we will walk through two case studies that show Twincler Spark in real-world environments. One case study allows insights into the performance of the system detecting harmful content and conspiracy theories timely, relevant for YouTube and recommendation systems. Another case study will show how Twincler Sparks reduces the number of leads to review for human content evaluators, relevant for Twitter and content evaluation. The data of both case studies is related to the US 2020 elections.

A Appendix

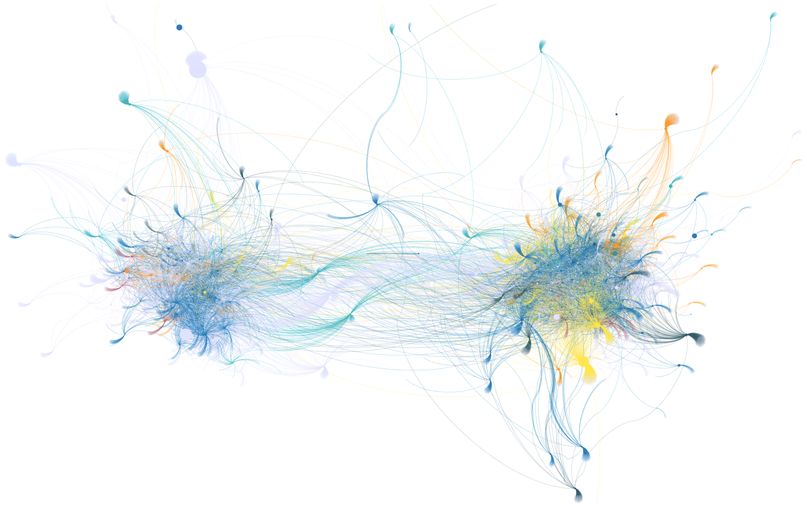


Figure 1: Network mapping of the 500 most active Twitter accounts with signals of synthetic engagement or untrustworthy behavior related to YouTube resources and the YouTube videos they have shared while referring to the US 2020 elections in the time between October 1, 2020 and October 20, 2020. Nodes represent Twitter accounts and YouTube videos. They are colored according to their operational behavior: Promotes content inorganically - orange, misrepresents identity - blue, advanced capabilities to amplify resources at scale - moss green, advanced capabilities to simulate popularity - dusted green, advanced capabilities to simulate organic engagement - dusted rose, extraordinary capabilities to amplify resources at scale - clear green, extraordinary capabilities to simulate organic engagement - yellow, other - night green. The connections between the accounts are determined by the engagement to one another, measured by retweets, favs, quotes and replies and the content they have shared. The sizes of the nodes reflect the number of referrals to YouTube videos.

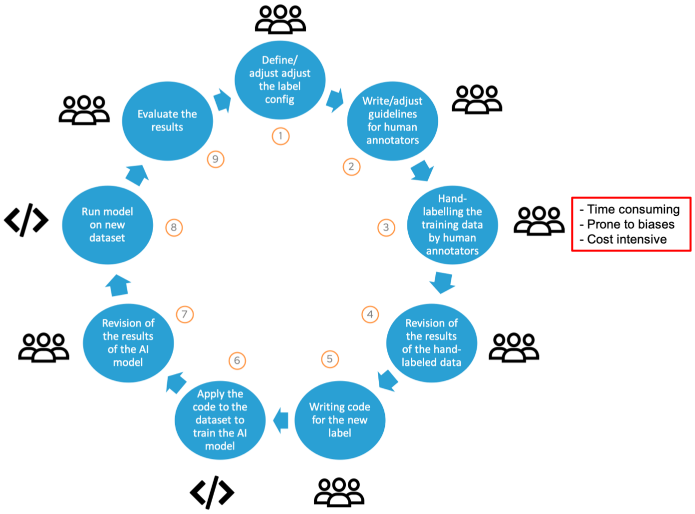

Define/ adjust adjust the label config
1
Write/adjust guidelines for human annotators
2
Hand-labelling the training data by human annotators
3
- Time consuming
- Prone to biases
- Cost intensive
Revision of the results of the hand-labeled data
4
Writing code for the new label
5
Apply the code to the dataset to train the AI model
6
Revision of the results of the AI model
7
Run model on new dataset
8
Evaluate the results
9

Figure 2: Training of AI-enabled systems for content moderation. Source: "Instagram's Adam Mosseri on the future of Reels, moderation, and the responsibility of social media platforms", in: The Verge, Decoder with Nilay Patel, January 2021. Image ©Twincler 2021

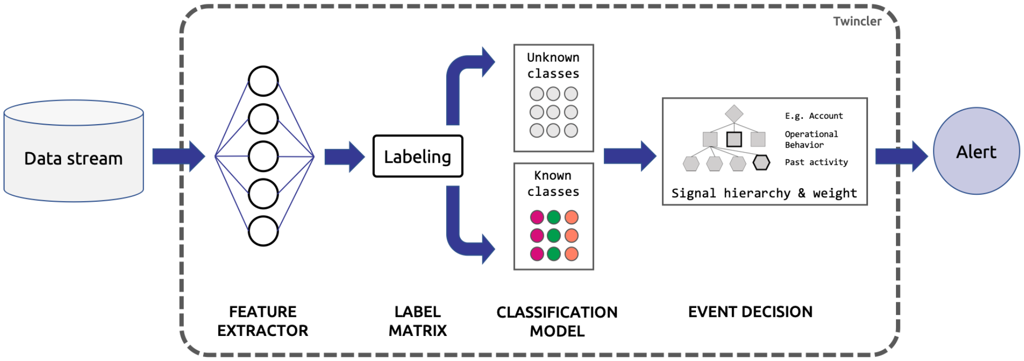


Figure 3: An overview of the Twincler process.

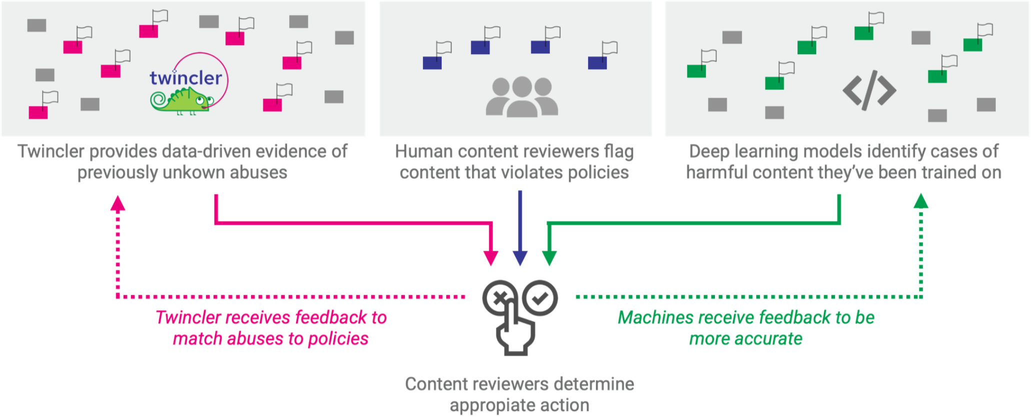

twincler
Twincler provides data-driven evidence of previously unkown abuses
Human content reviewers flag content that violates policies
Deep learning models identify cases of harmful content they've been trained on
Twincler receives feedback to match abuses to policies
Machines receive feedback to be more accurate
Content reviewers determine appropiate action

Figure 4: Twincler integration into a pre-existing content evaluation process. Twincler can also be integrated into multi-task recommendation or ranking systems.