identrics

Building Resilience to Foreign Authoritarian Disinformation in Southeast Europe

Vasil Shivachev, COO @ Identrics

February 2023

about us

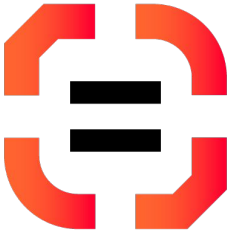
identrics

Vasil Shivachev

Vasil has 20 years' experience in information technologies and services, having held diverse roles, ranging from DevOps Engineer to managing a IT retail company. He's an avid audio and technophile, who collects turntables and rare and historic components for music and home theatre systems.

identrics

The company in a nutshell

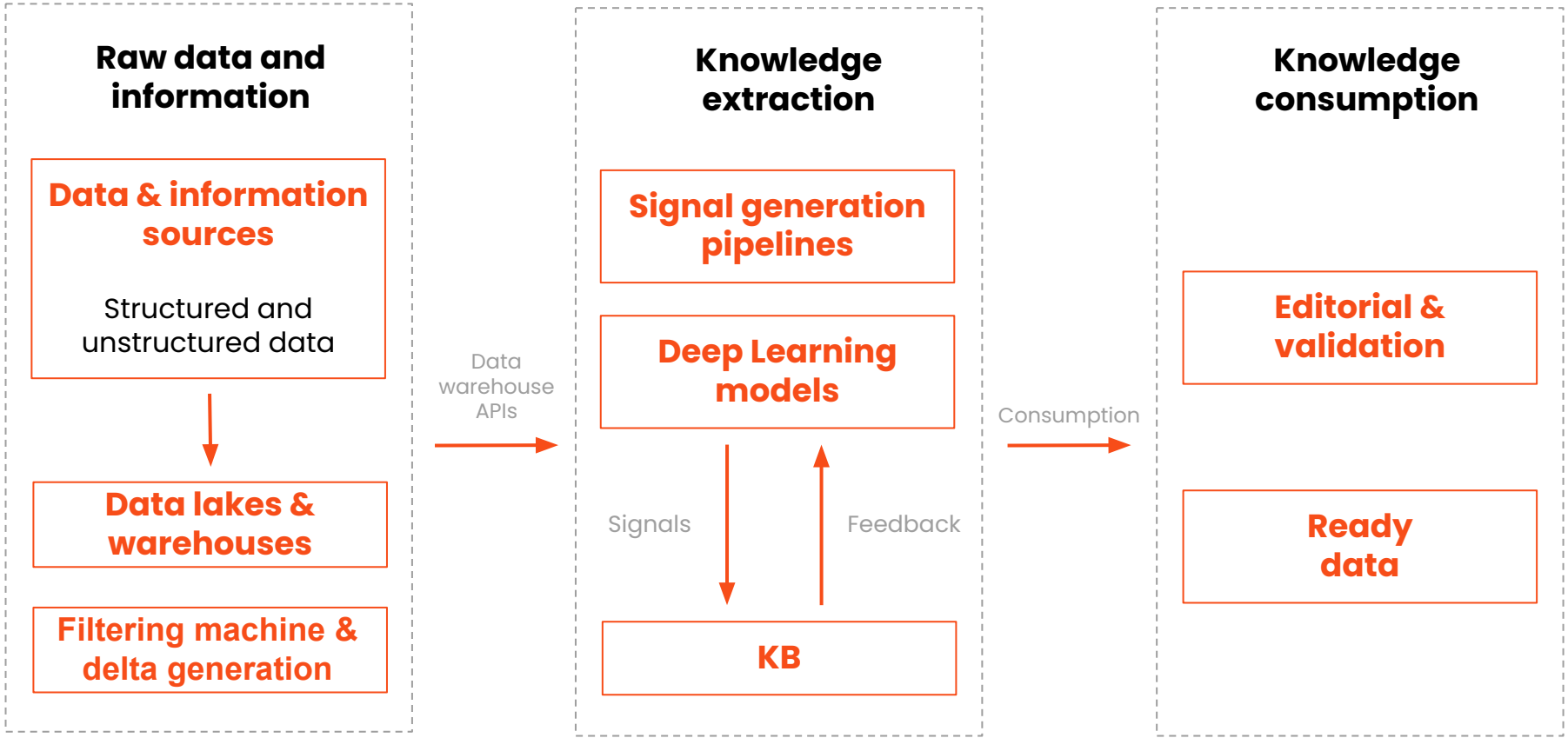


- **20+ years** in MI(Media Intelligence) and RI(Risk Intelligence) industry.
- One of the biggest global collectors of open data - **170K+ monitored sources** (traditional and social media, blogs, forums, telegram, comments and etc.) **200+ languages**
- Tailored **knowledge extraction** from unstructured data
- Design and maintenance of **knowledge bases**
- **Text generation** and automated abstracting

identrics

open sources intelligence

identrics


Raw data and information
Data & information sources
Structured and unstructured data
Data lakes & warehouses
Filtering machine & delta generation
Data warehouse APIs
Knowledge extraction
Signal generation pipelines
Deep Learning models
Signals
Feedback
KB
Consumption
Knowledge consumption
Editorial & validation
Ready data

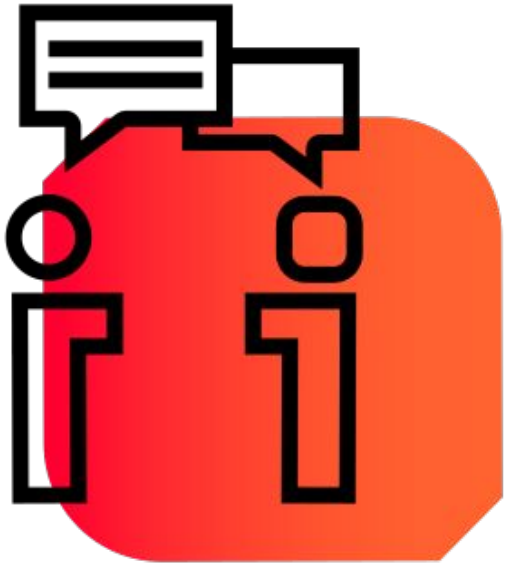
identrics

our experience fighting hate speech and propaganda

identrics

dis
mis
mal

information

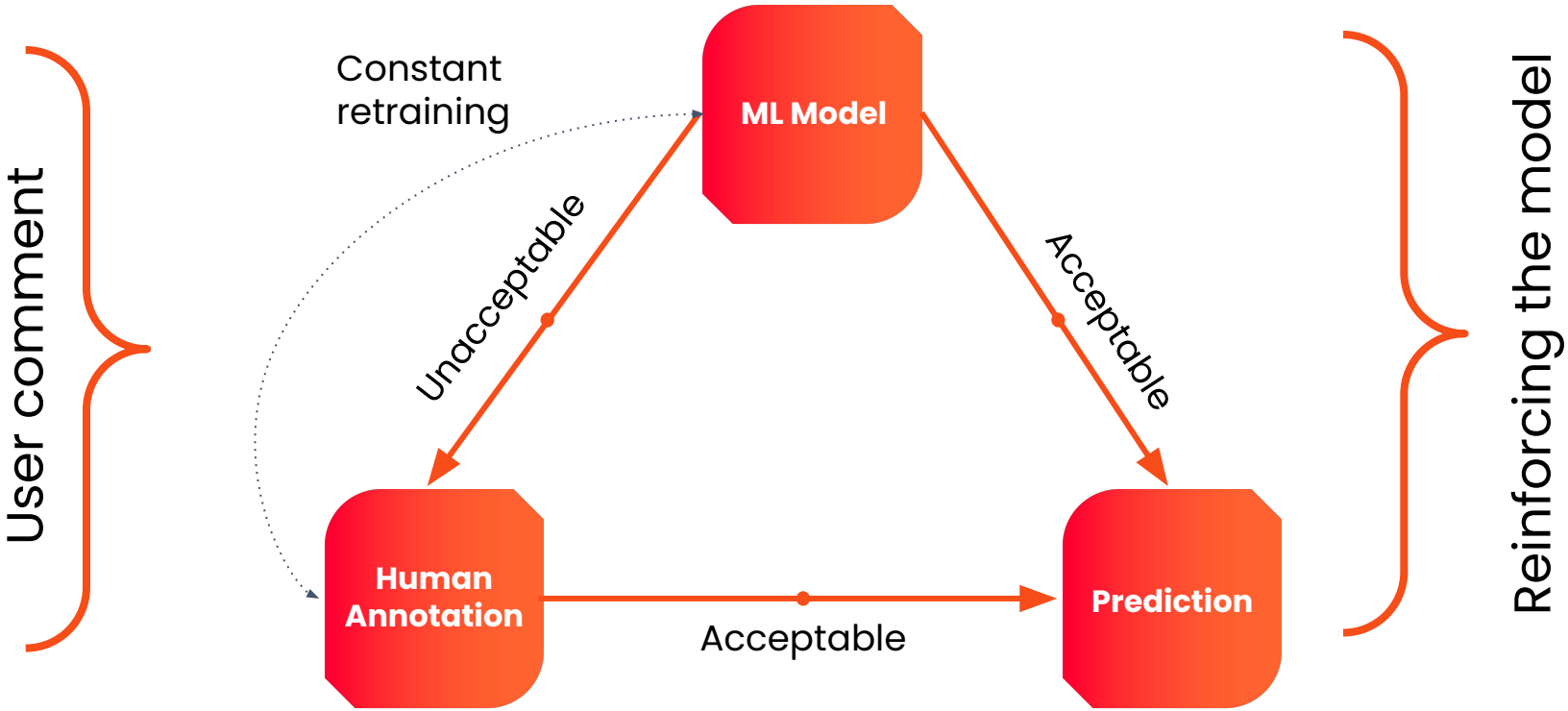


identrics

hate speech detection models

identrics

Hate speech detection models



context

77.9% readers have read the comments at some point

18.8% readers spent more time with comments

32.0% readers spent same amount of time with both

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

identrics

context

46.2% to learn about the opinions of others

40.1% to be entertained or amused by others' comments

33.9% to see how your opinion of the story or topic compares to others' view

29.9% to get more information on a story

27.9% to get additional reporting/updates to a story

26.4% to gauge the pulse of the community

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

identrics

issue / need

narrative can be hijack through storm of controversy hate comments

some articles generate engagement and large amount of **comments**

moderators have **limited resource** to check all comments for hate speech

identrics

approach

data science team supported by media' moderators

dataset created with the media' moderators

as a result **model** with result verified and accepted by media's moderators

identrics

result

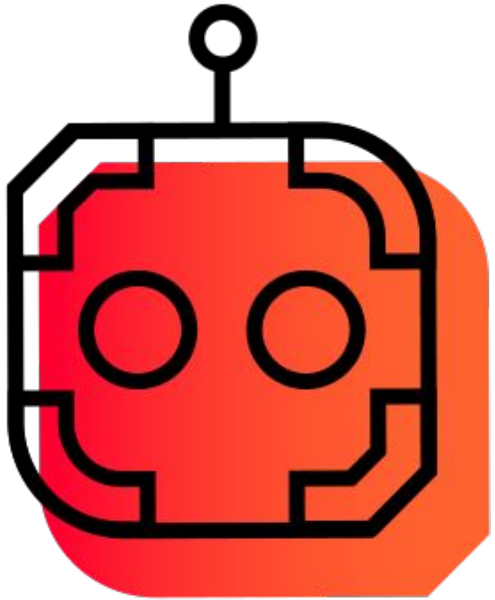
no hate comments - 0.85 precision

hate comments - 0.78 precision

accept / reject (**active learning**) strategy (Human-in-the-loop)

Supervised approach - generate alarms / signals for possible hate speech

Custom model per media



identrics

Next opportunity

Continuous improvement - HITL strategy

Reusable approach - for style based propaganda and disinformation detection

Custom model for advertisement detection

identrics

THANK YOU

vasil.shivachev@identrics.ai

www.identrics.ai

identrics