SUSTAINALYTICS

SEEING THE FOREST FOR THE TREES

Machine Learning Powered Open Source Intelligence

Arik Brutian
Stefan Dumitrescu

June 2019

Agenda

01 The open source intelligence challenge: information overload and relevancy

02 Machine learning and NLP as a solution

03 Lessons learnt and the way forward

The open source intelligence challenge: Information overload and relevancy

SUSTAINALYTICS

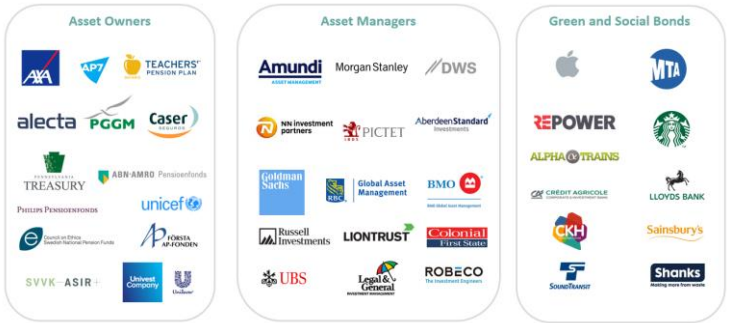
The Context

Who We Are

Vision: We believe that it is imperative for the global economy to become more just and sustainable;

Mission: To provide the insights required for investors and companies to make more informed decisions that lead to a more just and sustainable global economy.

- The largest pure-play investment research and ratings firm dedicated to responsible investment and Environmental, Social and Governance (ESG) research and ratings
- Over 500 professional staff with more than half engaged in research
- In excess of 550 clients (asset owners and managers)
- Over 25 years experience in the fields of ESG and corporate governance research



The Context

What We Do

» Sustainability considerations have become one of primary factors for assessing the viability of an enterprise in broader context.

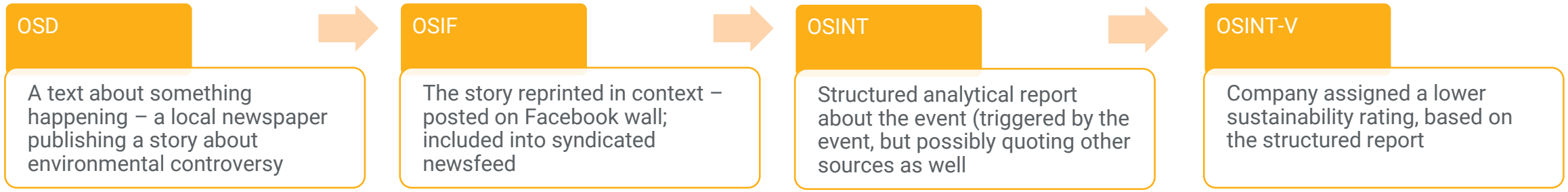
The Open Source Intelligence Challenge

What is OSINT?*

Open Source Intelligence: the process of discovery, discrimination, distillation, and dissemination of unclassified information to a select audience to support decision-making.

» **Phases of OSINT**

- OSD: Open source data (data in the primary source)
- OSIF: Open source information (data that can be put together after some filtering and validation – news digest, Facebook news wall, Twitter)
- OSINT: generation of intelligence out of OSIF
- OSINT/V: Validated OSINT (intelligence coupled with expert opinion, acting as an insight for decision-making)



*This slide conceptually relies on *Steele (2007)* "Open Source Intelligence" in *Johnson (editor), Handbook of Intelligence Studies,* New York: Routledge.

The Open Source Intelligence Challenge

OSINT in Sustainability Context

» **Research and ratings as highest level of OSINT**

- Watching all publicly traded companies in all key financial centres (20 000 + companies)
- Identifying any story in the online, print, or (recently) social media, mentioning any of these companies (90 000 sources covered in all key languages)
- Investigating the story and using the findings as key inputs into ESG research/ratings
- Example: Volkswagen controversy

The Open Source Intelligence Challenge

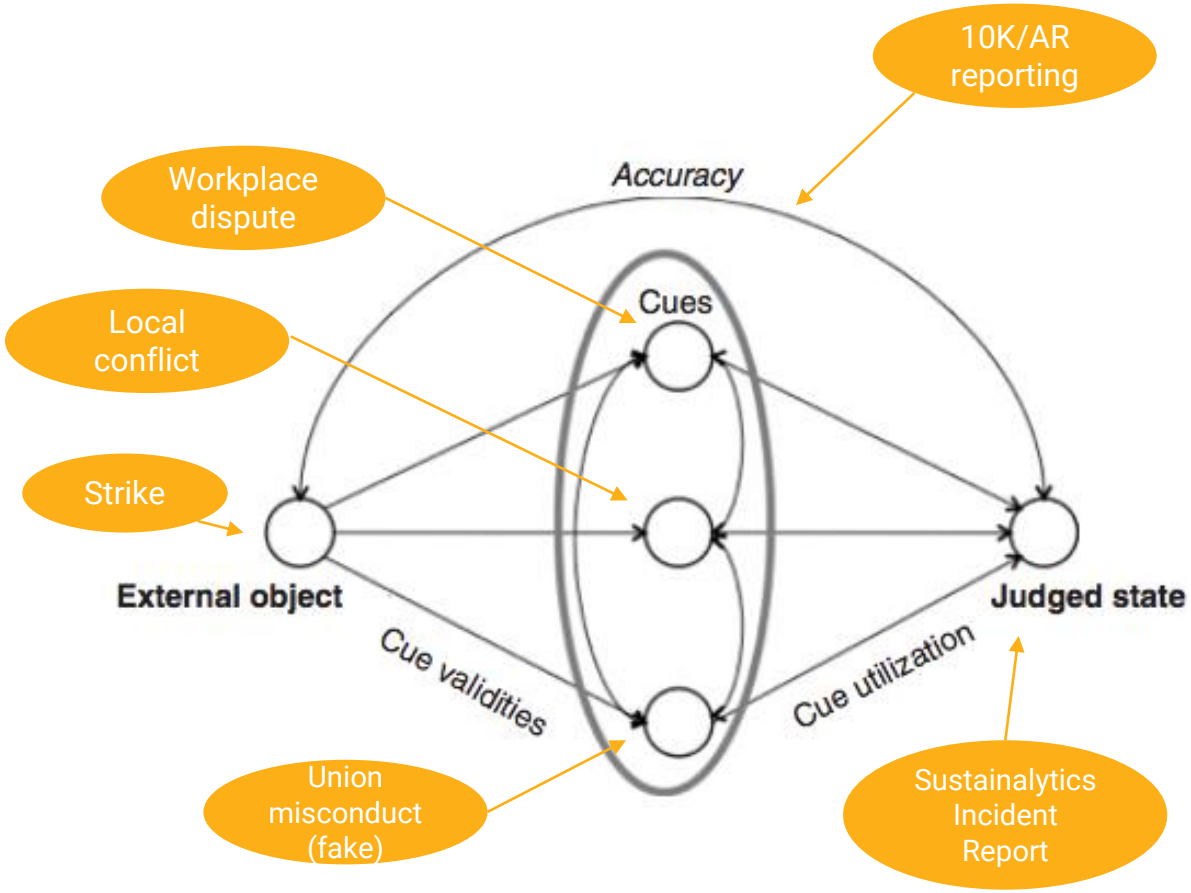
Recall – Comprehensive Monitoring

» **OSINT Recall**

- Real-life events are represented in the media in distorted manner;
- Framing / agenda setting affect how the event will be covered;
- Some events produce overlapping stories, and some events are insufficiently covered.

"the belief that more data or information automatically leads to better decisions is probably one of the most unfortunate mistakes of the information society"

Woods and Hollnagel 2006



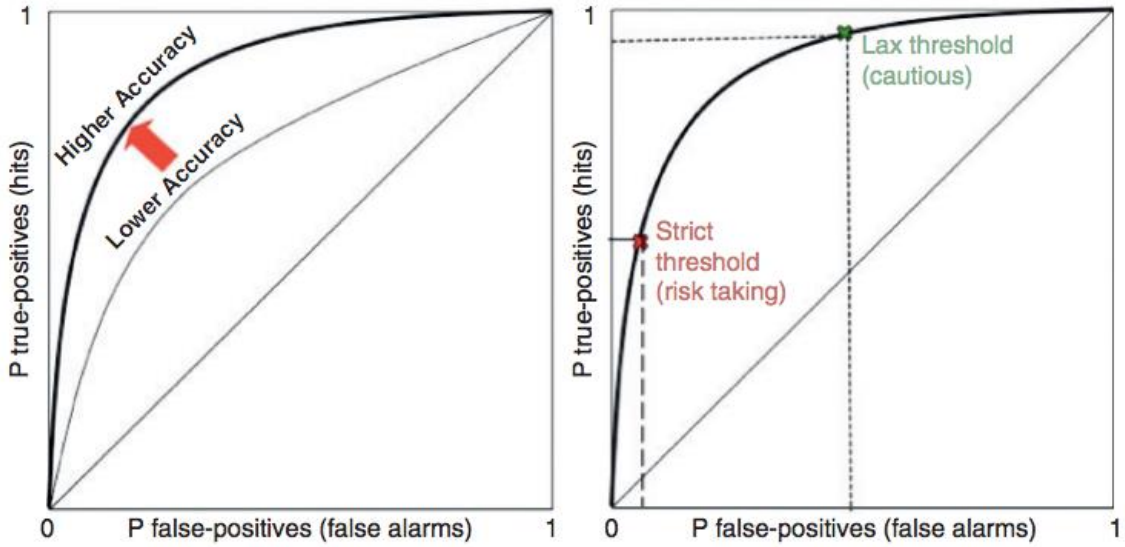
Adapted from Choo (2009) "Information Use and Early Warning Effectiveness: Perspectives and Prospects", *in* Journal of the American Society for Information Science and Technology, 60(5): 1071-1082.

The Open Source Intelligence Challenge

Precision: Making a Decision

» **OSINT Accuracy**

- Accuracy of the system is determined by the ratio of its true positives to false positives
- The time/effort spent on false positives can be prohibitively high, so high precision is necessary



Adapted from Choo (2009) "Information Use and Early Warning Effectiveness: Perspectives and Prospects", *in* Journal of the American Society for Information Science and Technology, 60(5): 1071-1082.

» **OSINT Sensitivity**

- Lax threshold: "weak evidence sufficient to say yes" – high probability of hits on the expense of false positives
- Strict threshold: "strong evidence needed to say yes" – reduced probability of false alarms, but also reduces hits

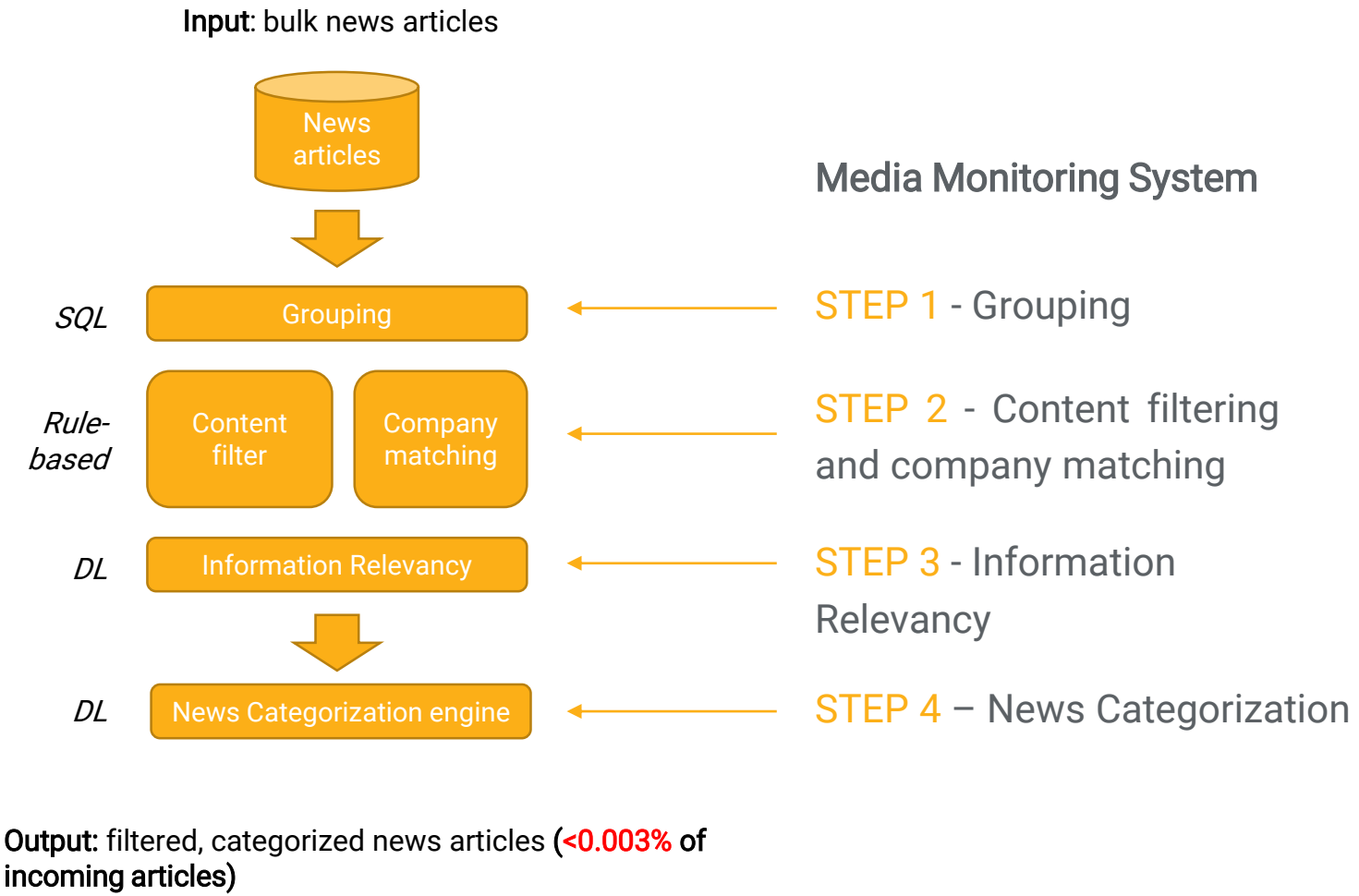
Machine learning and NLP as a solution

Solution

ML and NLP enabling Comprehensive Media Monitoring

» **Business Process:** Media Monitoring

- Processed around millions of news articles/day
- Received through an API connection
- Each article is received as an xml file enriched with metadata information related to sentiment or identified entities



Solution

ML and NLP enabling Comprehensive Media Monitoring

Incident/Relevant

July 26, 2016

9-Year-Old Child Worker Dies in Bangladeshi Textile Mill

The father of Sagar Barman accused the supervisors at Zobeda Textile Mill in Dhaka of killing his son by pumping air into his rectum.

By Julfikar Ali Manik and Geeta Anand

June 04, 2016

In Turkey, a Syrian Child 'Has to Work to Survive'

Over one million Syrian children live in Turkey, and thousands work in factories or sweatshops to provide for their families, rather than attend school.

By Ceylan Yeginsu

May 25, 2016

Indonesian Children Face Hazards on Tobacco Farms, Report Says

Children as young as 8 working on tobacco farms are exposed to harmful nicotine and pesticides, according to Human Rights Watch researchers.

By Joe Cochrane

April 10, 2016

JAKARTA JOURNAL

Gridlocked Jakarta Becomes Even Worse, at Least for a Week

The governor of Jakarta temporarily suspended its "three-in-one" car-pool system in part because it exploits children

Noise

May 30, 2019

We Don't Need to Be Saved From Making Smoothies

Food delivery services have tried to convince us that cooking is a difficult, boring task. It's actually a life skill.

By David Tamarkin

May 28, 2019

FRONT BURNER

In Brooklyn, a New Food Hall With Breathtaking Views

Time Out Market New York, from the publishers of the magazine, features 21 vendors in a two-story space in Dumbo.

By Florence Fabricant

May 28, 2019

FRONT BURNER

Greek Goods for Your Pantry

The cookbook author Diane Kochilas has opened an online marketplace for unusual Greek food stuffs.

By Florence Fabricant

May 28, 2019

Kawi, From Momofuku, Takes Korean Food Head On

The cuisine was always crucial to David Chang's empire, but at Kawi, the chef Eunjo Park brings it front and center.

Solution

ML and NLP enabling Comprehensive Media Monitoring

July 26, 2016

9-Year-Old Child Worker Dies in Bangladeshi Textile Mill

The father of Sagar Barman accused the supervisors at Zobeda Textile Mill in Dhaka of killing his son by pumping air into his rectum.

By Julfikar Ali Manik and Geeta Anand

June 04, 2016

In Turkey, a Syrian Child 'Has to Work to Survive'

Over one million Syrian children live in Turkey, and thousands work in factories or sweatshops to provide for their families, rather than attend school.

By Ceylan Yeginsu

May 25, 2016

Indonesian Children Face Hazards on Tobacco Farms, Report Says

Children as young as 8 working on tobacco farms are exposed to harmful nicotine and pesticides, according to Human Rights Watch researchers.

By Joe Cochrane

April 10, 2016

JAKARTA JOURNAL

Gridlocked Jakarta Becomes Even Worse, at Least for a Week

The governor of Jakarta temporarily suspended its "three-in-one" car pool system in part because it exploits children

July 25, 2016

DHAKA, Bangladesh — A supervisor at a textile mill was arrested after a 9-year-old worker died over the weekend, and the boy's father accused the supervisor and others of killing him because he had protested against abuse.

Ismail Hossain, the officer in charge of the Rupganj police station in Narayanganj District in central Bangladesh, said that Nazmul Huda, an assistant administrative officer at the Zobeda Textile Mill, had been taken into custody for questioning, and that others would also be detained as the inquiry continued.

The father of the boy, Ratan Barman, 70, filed a complaint on Sunday with the Rupganj police, accusing supervisors at the mill of killing his son by pumping air from a compressor machine into his rectum.

The boy, Sagar Barman, had been working at the mill for seven months, along with his parents, his father said in a telephone interview on Monday.

"I thought, as we are poor, it will be helpful to run our family if my son Sagar can do some work in this factory," Mr. Barman said. "I used to gather empty bobbins," putting them into a trolley, he added. "My son also used to do the same work."

Last year, a 12-year-old boy died in a similar manner at the motorcycle repair shop where he had worked. Though the official minimum working age is 14, child labor has long been widespread in Bangladesh, and the government does not keep records of workplace deaths or injuries involving children. But cases like Sagar's capture the public's attention.

Solution

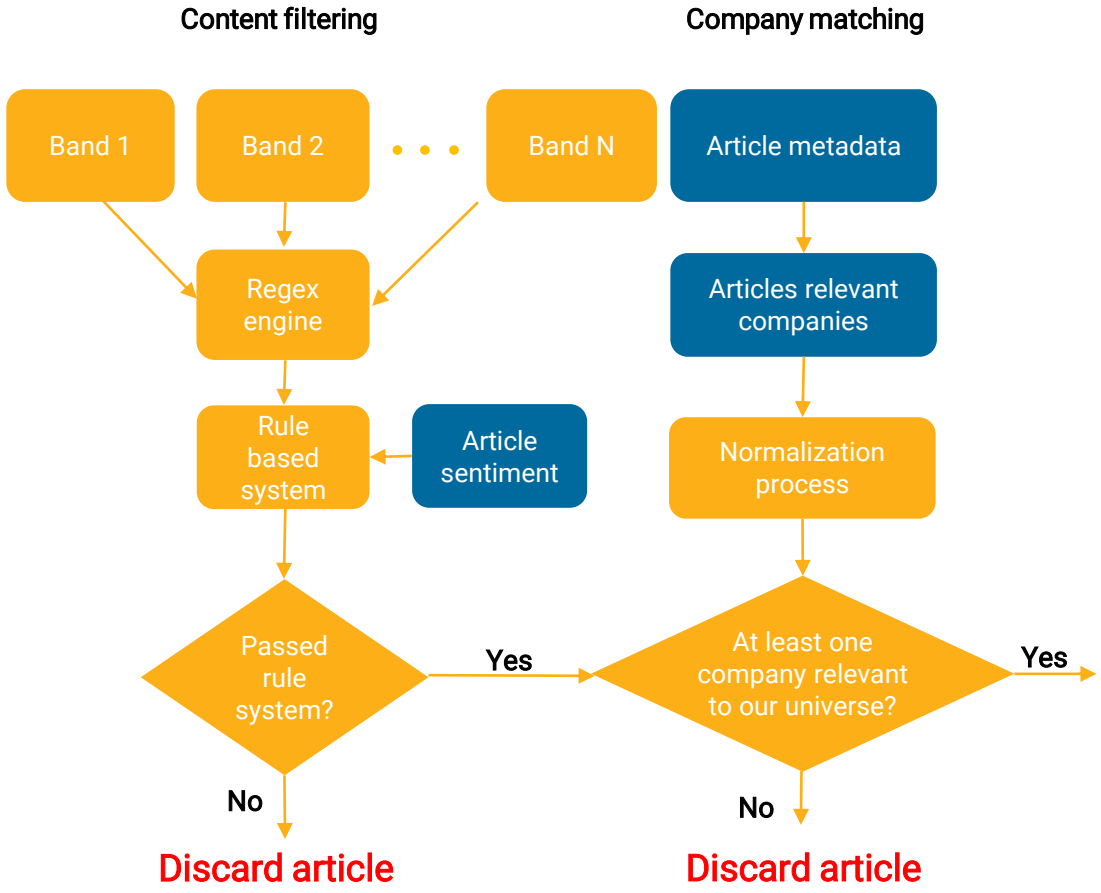
Content Filtering and Company Matching

» **Problem:** Discard irrelevant articles

» **Task:** reject those articles that do not contain certain key-words and that do not refer to at least one of our target companies

» **Solution**

1. The Content filter is a rule-based system that uses as input several bands of words ranked based on their relevancy and the article sentiment (found in the metadata). Regular expressions are used to calculate the frequency of each word within an article and the output is combined with the sentiment in a proprietary set of rules. An article is considered relevant and moves forward if it passes at least one of these rules.
2. Every article that passes the content filter is checked for companies that are relevant to our universe. Company matching filter uses Levenshtein distance for string similarity coupled again with a rule-based system. Each company mention, before entering the system, is normalized by lowercasing, removing/replacing common words.



Solution

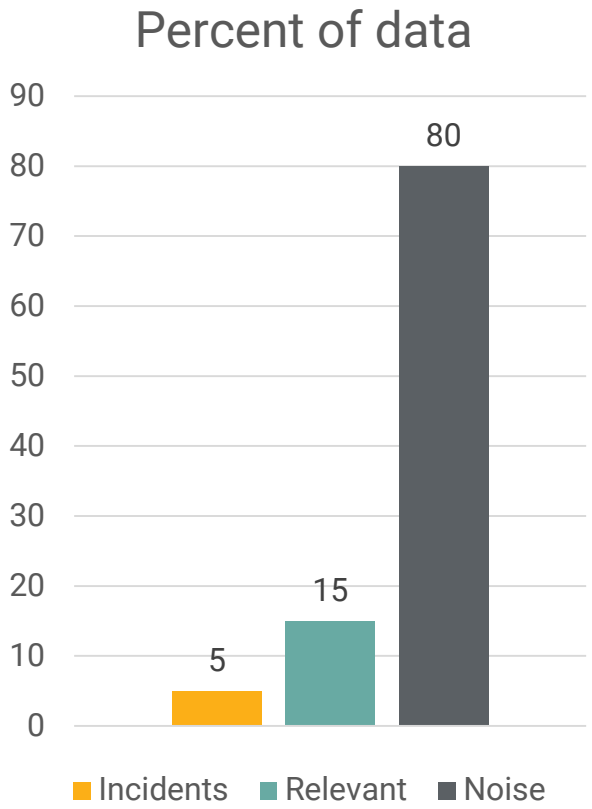
Informational Relevancy

- **Problem:** further cut down on irrelevant news articles. Articles that still trickle down to this level: > 5K/day, with >10K+ after weekends
- **Task:** assign a label (Incident, Relevant and Noise) to a news article
- **Constraint:** false negative rate should be less than a specified amount (3-5%)

Confusion Matrix

Predicted→	Incident	Relevant	Noise
Incident	TP	TP	FN
Relevant	..	..	..
Noise	..	..	..

- **For incidents**
  1. TP includes Incidents classified as Relevant
  2. FN rate < 3%

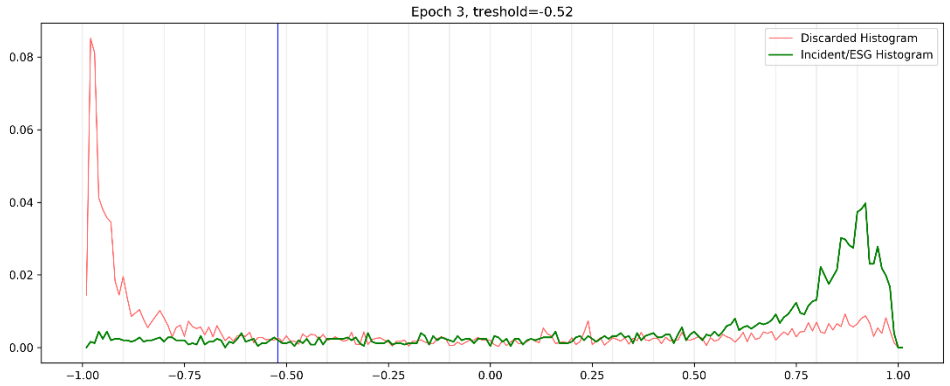


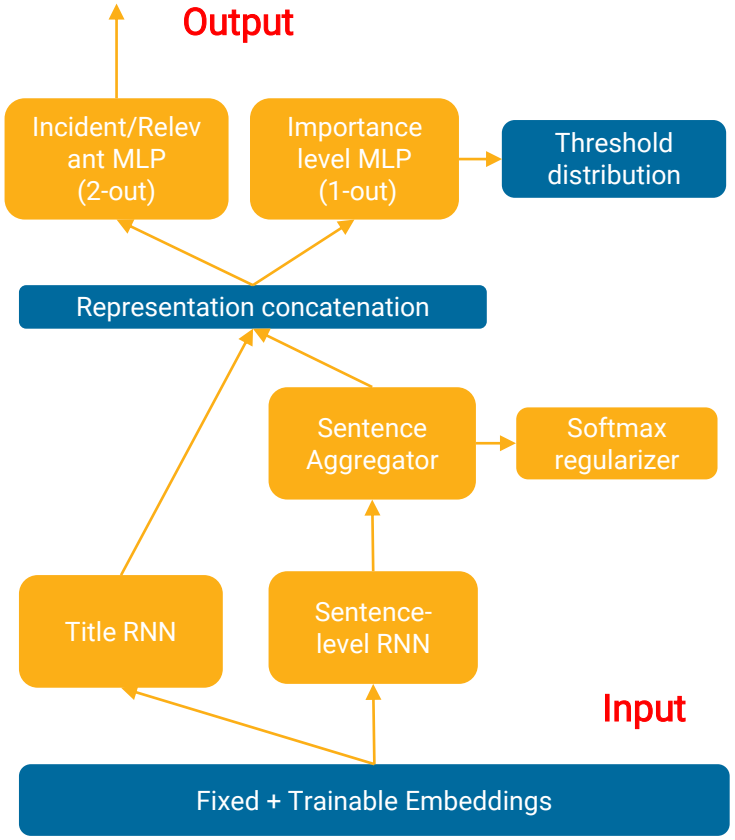
Solution

Content Filtering and Company Matching

» **Solution**

- RNNs
- Dynamically adjustable parameter that controls false negatives for the Incidents class
- Weighted loss
- Additional regularization (3-way softmax)
- Inference: first pass the Importance level MLP (1*tanh), then evaluate a 2*logistic MLP to decide Incident vs Relevant class





Solution

Content Filtering and Company Matching

» Solution

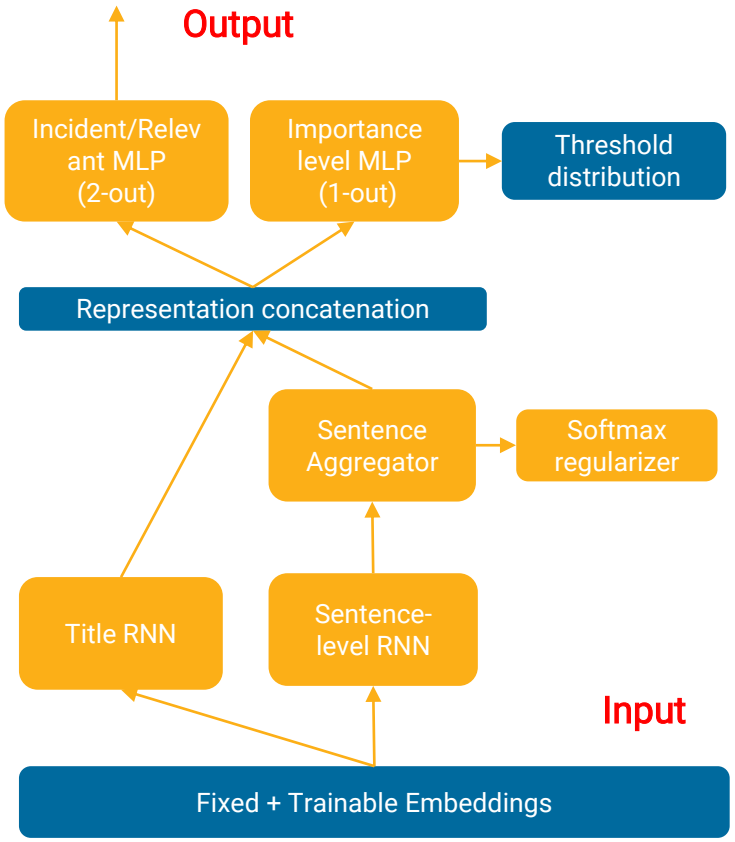
- RNNs
- Dynamically adjustable parameter that controls false negatives for the Incidents class
- Weighted loss
- Additional regularization (3-way softmax)
- Inference: first pass the Importance level MLP (1*tanh), then evaluate a 2*logistic MLP to decide Incident vs Relevant class

» What we tried

- Classic approaches (SVM/RF/etc.) as well as classic RNN approach with softmax on top – works well but can't control FN rate
- Decide only title / only on text

» Lessons learned

- Mid-level weighted softmax works well as a regularizer Decide only title / only on text
- Top 400 words are enough
- Text preprocessing is very important
- Trainable embeddings work better than pretrained ones (Glove/FastText) in this case …
- … though we did not try ELMO or other context dependent embeddings



Solution

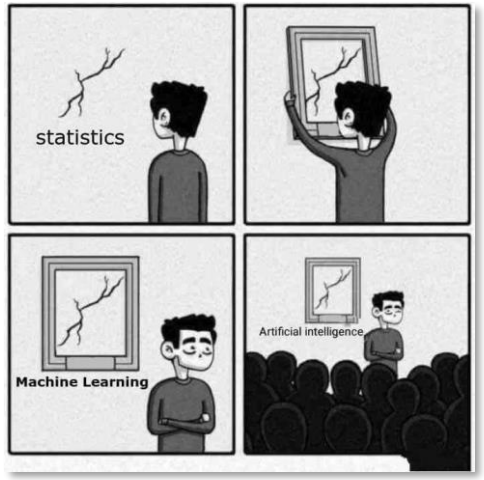
Content Filtering and Company Matching

» Problem

- automatically categorize news articles for further processing and research

» Task

- correctly choose one of 40+ mutually overlapping categories



» Constraint

- categories are rather fuzzy (subjective interpretation), sometimes seem to overlap or include one another, thus ITA is rather low. Dataset is highly imbalanced (long tail: biggest 2 classes have over 60% of data with smallest classes having <0.01%)

» Solution

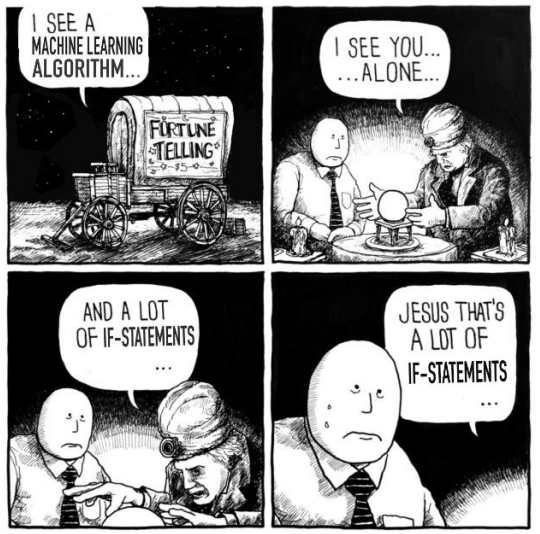
- simple TFIDF + SVM, **works ~3-5% better than anything else (score on the test set)**

Solution

Content Filtering and Company Matching

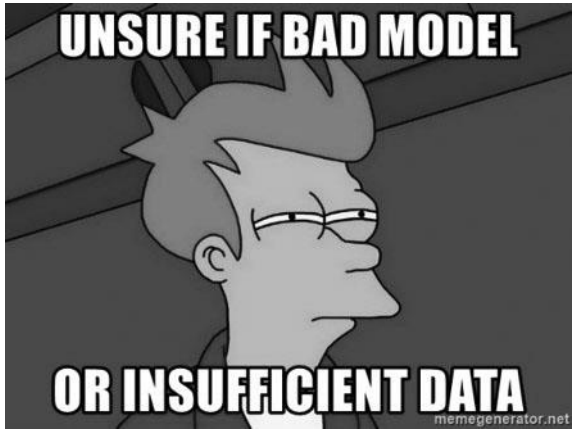
» What we tried

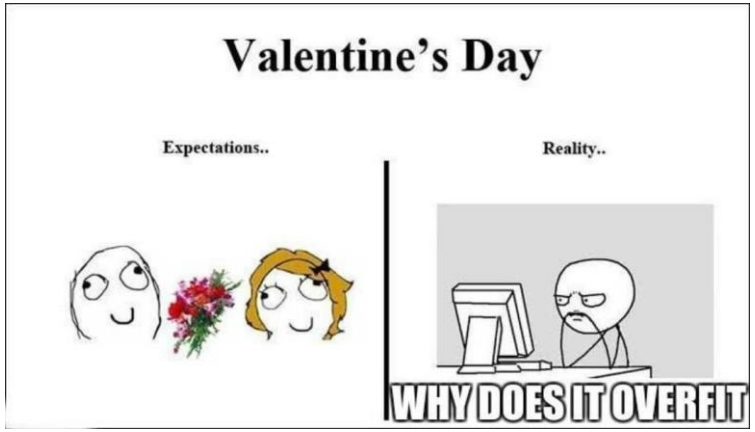
- All the “classic” algos, with and without LSA
- Standard recipe: embeddings + RNNs + softmax
- Partially retagging the dataset
- Unbalanced learning approaches (including oversampling methods like SMOTE)



» Lessons learned

- We live in a fuzzy world
- LSA doesn’t always help
- Networks overfit, and quickly
- Oversampling: sounds good, doesn’t (always) work
- Sometimes classic NLP works better





Lessons learnt and the way forward

Outcomes and Outgrowths

» **Results**

- Comprehensiveness: Dynamically monitor 20k+ entities
- Accuracy: Enabling the analysts to deal with 0.003% of daily incoming articles

» **Business Value**

- Increased the production of monitoring-based products by 3 times, while controlling for the human resources
- Decreased curator time spent on screening (due to increased accuracy) by half, thus freeing resources for high-value jobs Oversampling: sounds good, doesn't (always) work

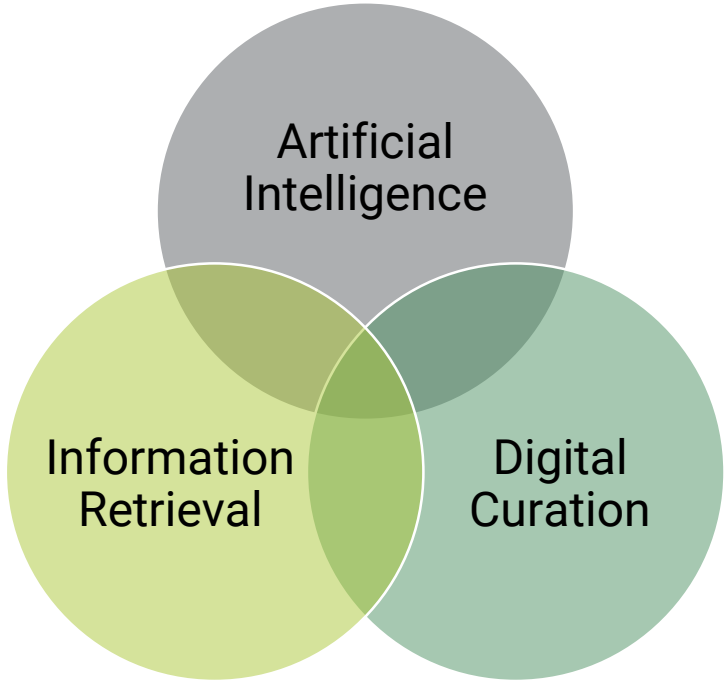
The Context

The Role of Digital Innovation

- **WHAT:** Create digital solutions that add intelligence to the ***information retrieval and analytics*** process.

- **HOW:** ***Identify***, ***research***, ***design***, ***develop***, ***test***, and ***integrate*** smart technologies into Sustainalytics' information processing value chain in order to support our mission.

SMART: INTERSECTION OF TECHNOLOGIES



Sustainalytics AI/ML Program

- Move **from** classification **to** generation
- Move **from** descriptive **to** predictive analytics
- Move **from** unstructured data **to** multimodal data (imagery, mines, water basins)

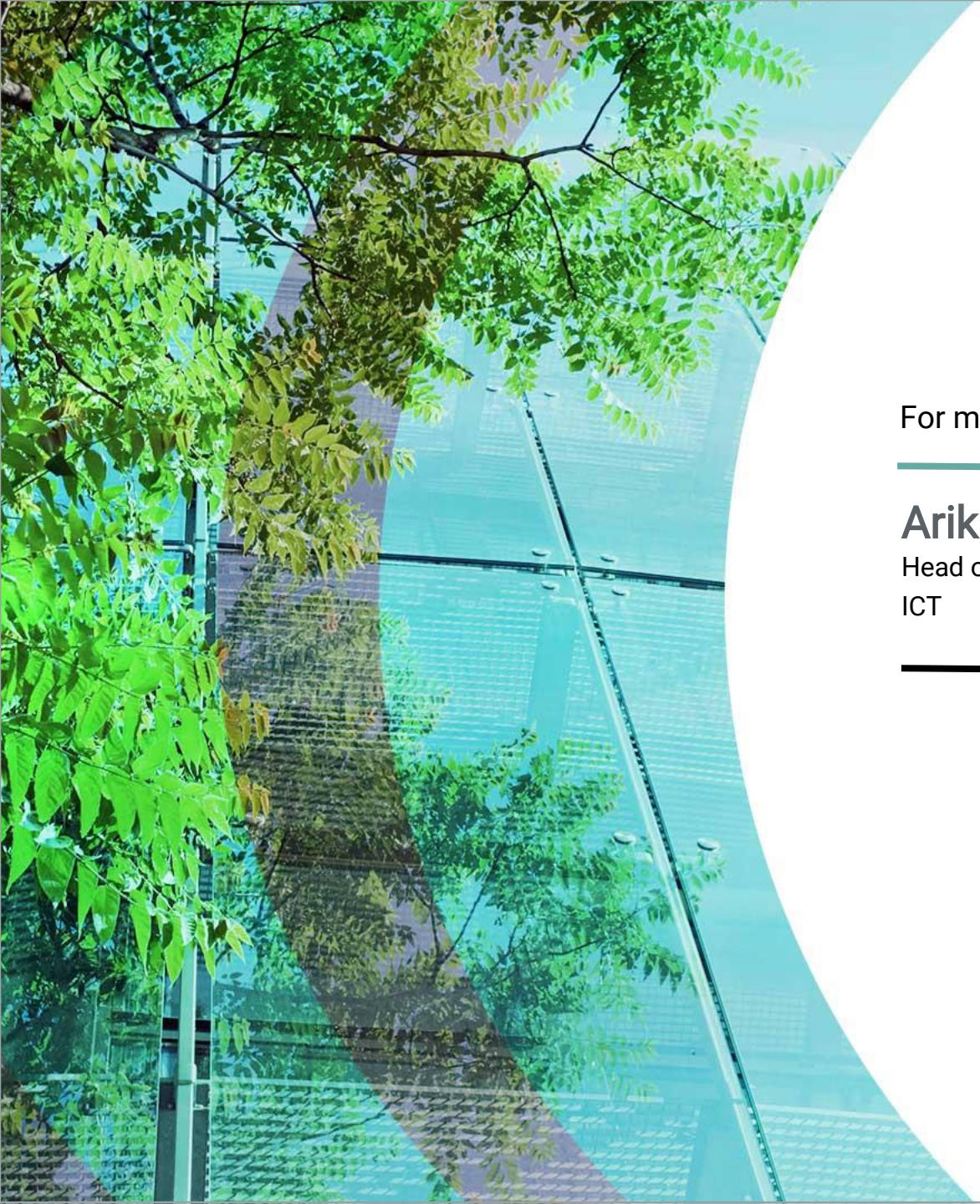
Gentlemen, we have run out of money; now we have to think.

Churchill

We are always out of resources, so we need the machines to think.

Thank You!

SUSTAINALYTICS



SUSTAINALYTICS

For more information, please contact:

Arik Brutian

Head of Digital Innovation

ICT

Stefan Dumitrescu

A.I. Solutions Architect

ICT

www.sustainalytics.com