Trends, challenges, and collaborations in Suptech: ESMA's perspective

Artificial Intelligence and Machine Learning for SupTech

EU Supervisory Digital Finance Academy

Florence, 15 March 2023

Giulio Bagattini
Risk Analysis Officer
Consumer, Sustainability and Innovation Analysis Unit
Economics, Financial Stability and Risk Department

ESMA
European Securities and Markets Authority

Table of contents

1. SupTech at ESMA: an overview
2. Data analytics for anomaly detection
3. Deep dive in NLP applications
4. Looking forward

1. **SupTech at ESMA: an overview**
2. Data analytics for anomaly detection
3. Deep dive in NLP applications
4. Looking forward

SupTech’s role at ESMA

- Relevant to multiple departments (potentially all those that use data)
- Enables combining traditional and innovative data sources
- Mainly using Python
- Only a few staff per department applying these techniques, but growing interest and training possibilities
- Staff are encouraged to propose, develop and present tools
- Ad-hoc workshops with NCAs to disseminate knowledge
- Ad-hoc focused activities with groups of NCAs
- Recurring item for discussion at the Financial Innovation Standing Committee

SupTech's role at ESMA

SupTech supports various ESMA activities:

- Market monitoring and risk assessment
  - — Collect and analyse alternative and unstructured data
- Direct supervision
  - — Detect anomalies and flag instances for human review
- Supervisory convergence
  - — Compare compliance with regulations at national level, support peer reviews (e.g. prospectus)
- Policy development
  - — Provide quantitative evidence for review of regulations, ESMA's advice to EC (e.g. PRIIPs KIDs)

SupTech's role at ESMA

Main tools and methods:

Natural language processing (NLP)

- Transforming text into data, it enables a quantitative analysis of unstructured information (e.g. compliance documents)

Data analytics for anomaly detection

- To enable compliance monitoring and anomaly detection in the supervision of credit rating agencies and transaction data reporting infrastructures

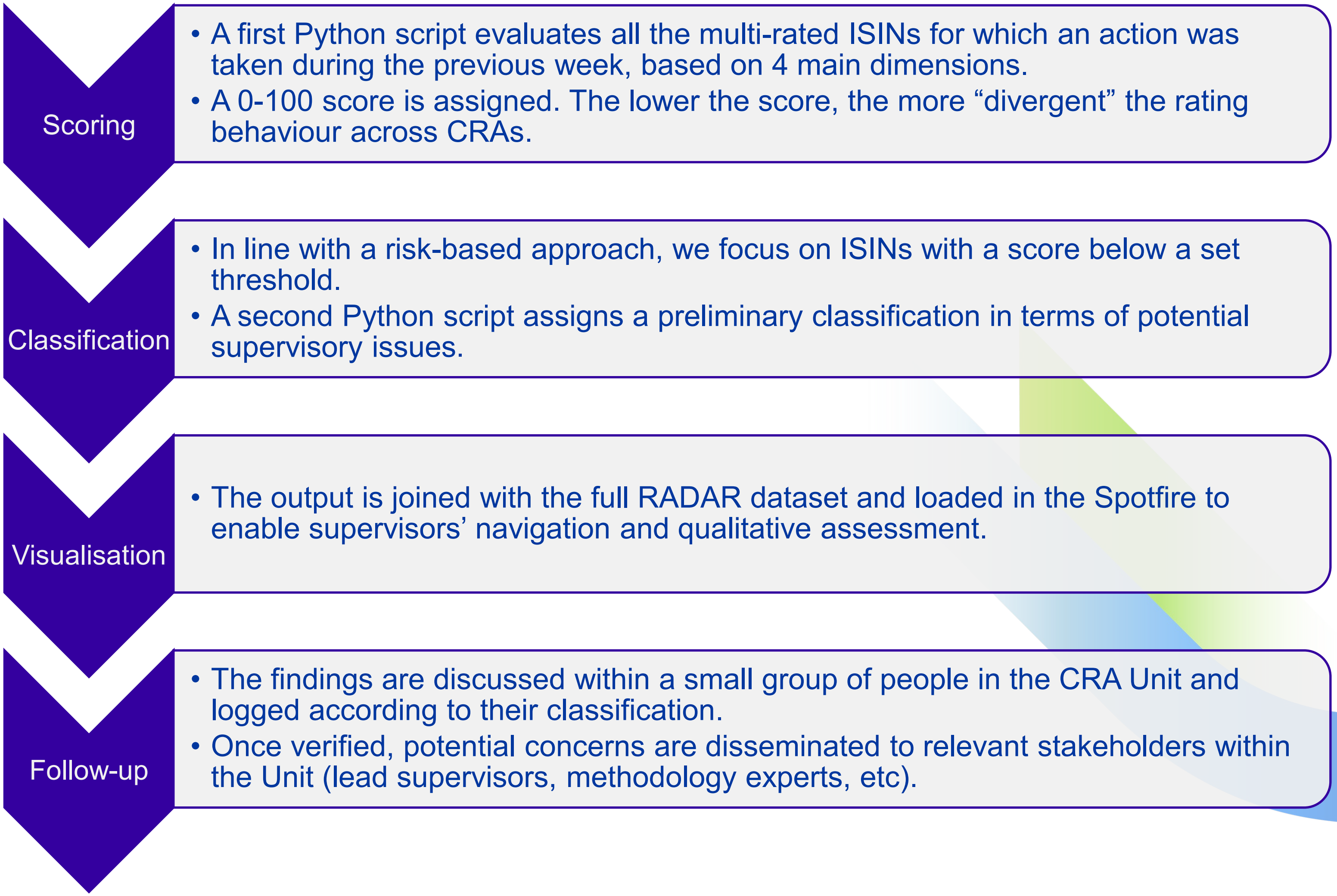
Web-scraping

- To retrieve information from the web and structure it into datasets (compatibly with applicable policies)

1. SupTech at ESMA: an overview
2. **Data analytics for anomaly detection**
3. Deep dive in NLP applications
4. Looking forward

Anomaly detection in credit ratings

Rule-based tool to identify abnormal ratings



Scoring

- A first Python script evaluates all the multi-rated ISINs for which an action was taken during the previous week, based on 4 main dimensions.
- A 0-100 score is assigned. The lower the score, the more “divergent” the rating behaviour across CRAs.

Classification

- In line with a risk-based approach, we focus on ISINs with a score below a set threshold.
- A second Python script assigns a preliminary classification in terms of potential supervisory issues.

Visualisation

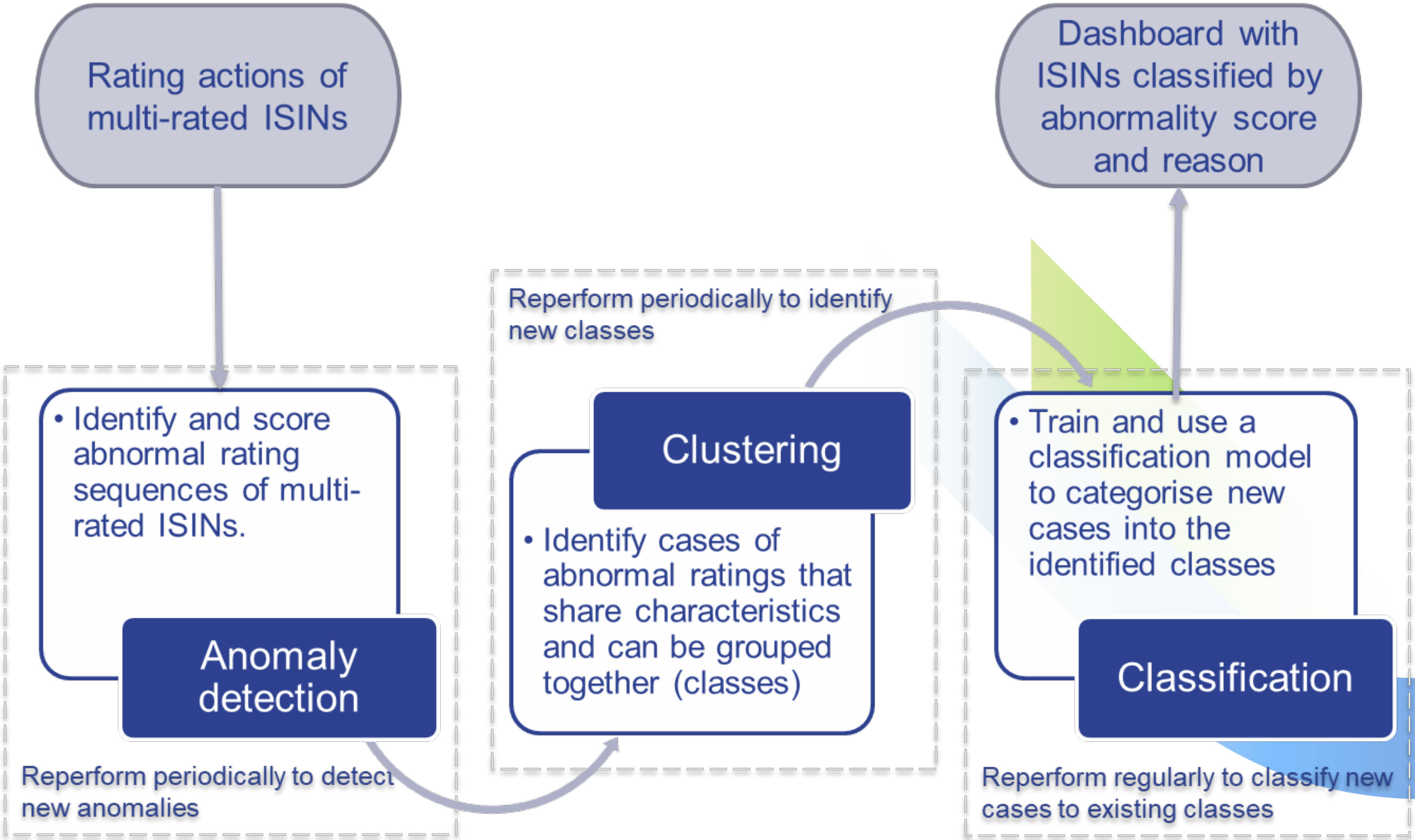
- The output is joined with the full RADAR dataset and loaded in the Spotfire to enable supervisors’ navigation and qualitative assessment.

Follow-up

- The findings are discussed within a small group of people in the CRA Unit and logged according to their classification.
- Once verified, potential concerns are disseminated to relevant stakeholders within the Unit (lead supervisors, methodology experts, etc).

Anomaly detection in credit ratings

ML-based tool to identify, group and classify abnormal ratings



Rating actions of multi-rated ISINs

Anomaly detection

- Identify and score abnormal rating sequences of multi-rated ISINs.

Reperform periodically to detect new anomalies

Clustering

- Identify cases of abnormal ratings that share characteristics and can be grouped together (classes)

Reperform periodically to identify new classes

Classification

- Train and use a classification model to categorise new cases into the identified classes

Reperform regularly to classify new cases to existing classes

Dashboard with ISINs classified by abnormality score and reason

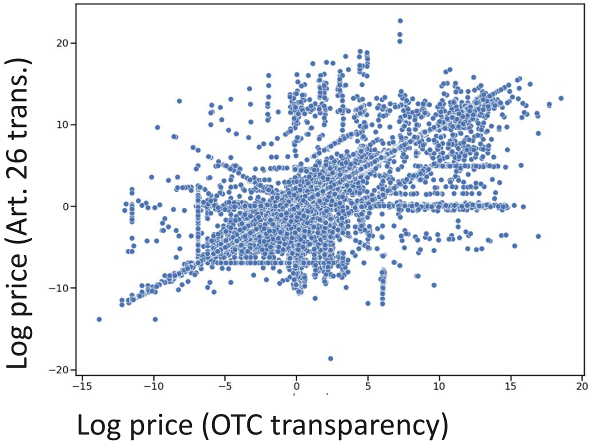
Anomaly detection in transaction data

- ESMA is responsible for supervising key data reporting infrastructures under EMIR, SFTR, MiFIR and SECR
- ESMA monitors consistency between transaction data reported directly to regulators (Art. 26 of MiFIR) and OTC transaction data published for transparency purposes
- ESMA uses data analytics – including ML – to detect inconsistencies of price/quantity information for a given instrument or trading venue, over time and at an aggregate level

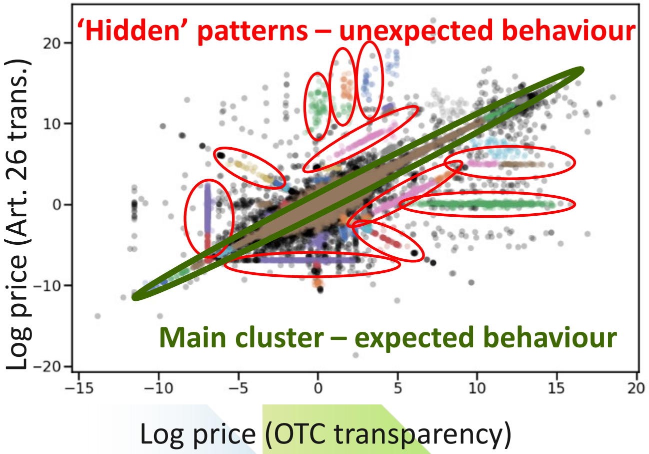
Anomaly detection in transaction data

I. Consistency on an aggregate level

ML techniques such as clustering can reveal ‘hidden’ patterns that link data to behaviours driven by specific instruments/reporting firms/trading venues etc.



Detection of ‘hidden’ patterns via clustering

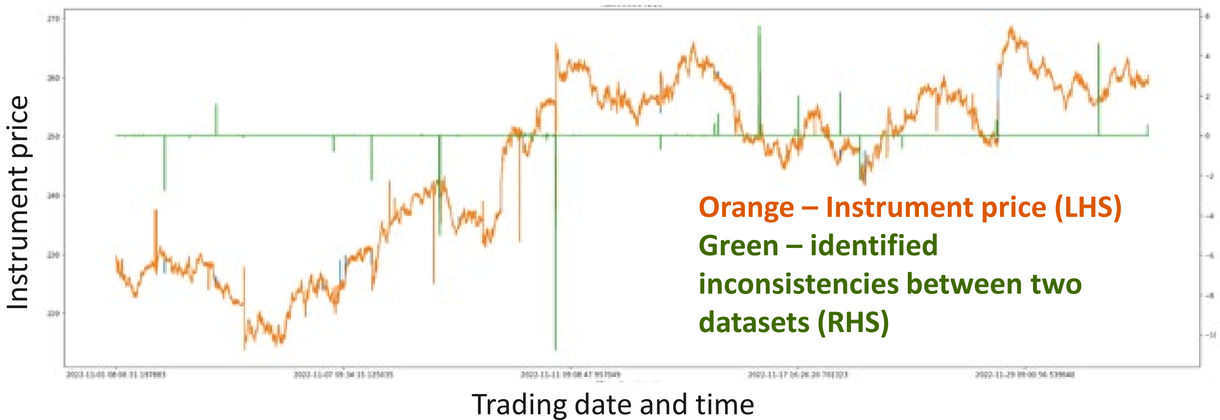


‘Hidden’ patterns – unexpected behaviour

Main cluster – expected behaviour

II. Consistency over time

Consistency is monitored also at the level of individual transactions at each point in time



Orange – Instrument price (LHS)
Green – identified inconsistencies between two datasets (RHS)

1. SupTech at ESMA: an overview
2. Data analytics for anomaly detection
3. **Deep dive in NLP applications**
4. Looking forward

Deep dive in NLP

- Increasing amount of information available in the form of text only

  - From supervised entities: compliance and disclosure documents, communications to investors, marketing material…

  - On the web: social media, news outlets, etc

- Impossible for humans alone to process all information → Computer processing extends information set

- Main focus of work has been on documents; growing interest on other sources on the web (relevant to trends and risks monitoring)

Deep dive in NLP

NLP has a cross-cutting role:

Applied to documents produced by investment funds, securities issuers, structured retail products, CRAs…

Supports all mentioned ESMA mandates:

- Market monitoring and risk assessment: transform key information from text into data for statistical analysis
- Direct supervision: assist supervisors' ability to systematically monitor compliance with regulations and detect anomalies (e.g. required phrases and disclosures, requirements like 'clear' and 'comprehensible' language, description of risks)
- Supervisory convergence: produce quantitative measures to assess implementation of regulations and guidelines
- Policymaking: produce data-based indicators for evaluation and review of EU regulations

Deep dive in NLP

Our approach:

- Leverage off-the-shelf methods/packages when possible
- Start from simple methods, increasing complexity when warranted
- Develop solutions specific to context and objectives, based on expert judgment
- This means that often subjective decisions are involved→ NLP-based analyses also come with limitations. E.g.:
  - Phrases can be "missing" in different ways (entirely absent vs. rephrased)
  - Criteria to determine whether language is "clear and comprehensible" depend on the linguistic metrics used, which are subject to prior selection
  - Text length, repetitiveness, and complexity may be viewed differently

    Long → unfocused *or* comprehensive?

    Repetitive → Low information content *or* consistency across a document?
- Transparency about analytical criteria and feedback from subject matter experts (e.g. policy depts) are essential

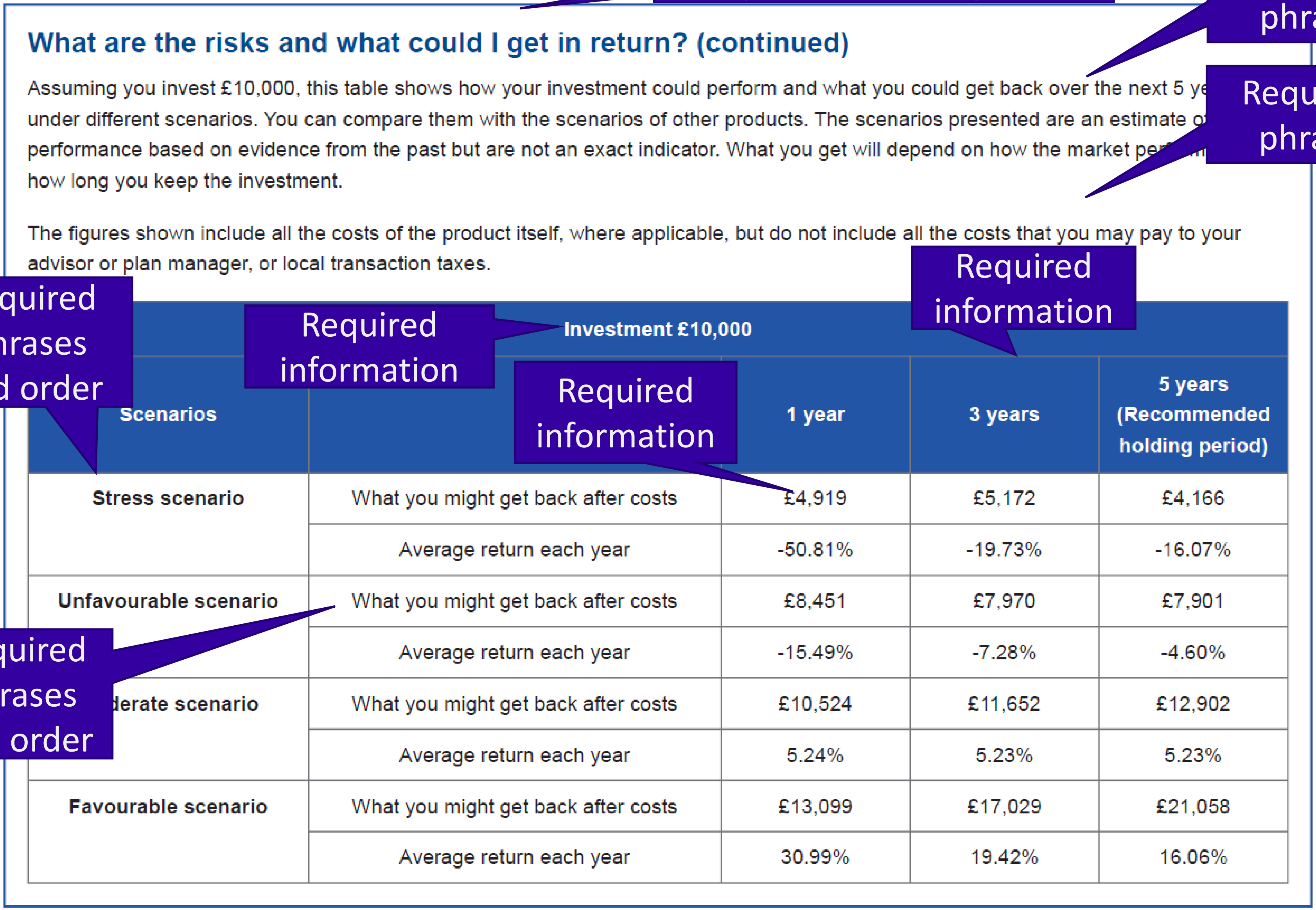
Deep dive in NLP

Example: text mining of PRIIPs KIDs

1. Load pdf and convert to text
   a) Scanning issues: ‘the quick brown fox’ → ‘the the quick quick brown brown…’
   b) Content of tables parsed in different orders
2. Confirm that required phrases and data are present
3. Extract information of interest (e.g. costs and performance scenarios)
4. Clean information
   a) standardise numbers: from “€ 12 400,89” to “12,400.89”
   b) identify numbers of interest compared with useless information (i.e. is 448190 a EUR number or just a document ID?)
5. Analyse the data

Deep dive in NLP

From pdf…



What are the risks and what could I get in return? (continued)

Assuming you invest £10,000, this table shows how your investment could perform and what you could get back over the next 5 ye under different scenarios. You can compare them with the scenarios of other products. The scenarios presented are an estimate o performance based on evidence from the past but are not an exact indicator. What you get will depend on how the market per how long you keep the investment.

The figures shown include all the costs of the product itself, where applicable, but do not include all the costs that you may pay to your advisor or plan manager, or local transaction taxes.

	Investment £10,000			
Scenarios		1 year	3 years	5 years (Recommended holding period)
Stress scenario	What you might get back after costs	£4,919	£5,172	£4,166
	Average return each year	-50.81%	-19.73%	-16.07%
Unfavourable scenario	What you might get back after costs	£8,451	£7,970	£7,901
	Average return each year	-15.49%	-7.28%	-4.60%
derate scenario	What you might get back after costs	£10,524	£11,652	£12,902
	Average return each year	5.24%	5.23%	5.23%
Favourable scenario	What you might get back after costs	£13,099	£17,029	£21,058
	Average return each year	30.99%	19.42%	16.06%

Required section/ phrase

Required phrase

Required phrase

Required phrases and order

Required information

Required information

Required information

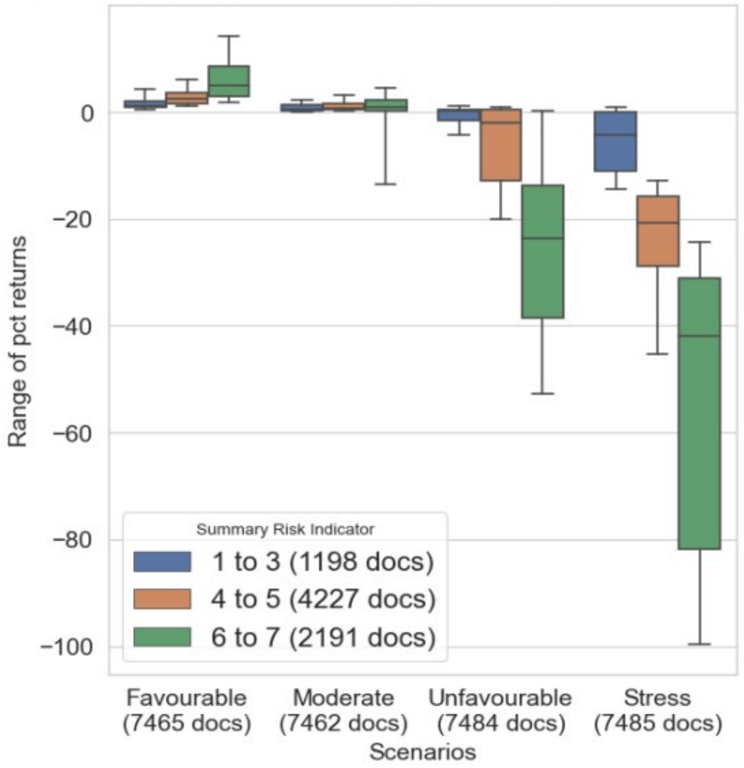
Required phrases and order

Deep dive in NLP

…to data analysis

SRI and simulated returns

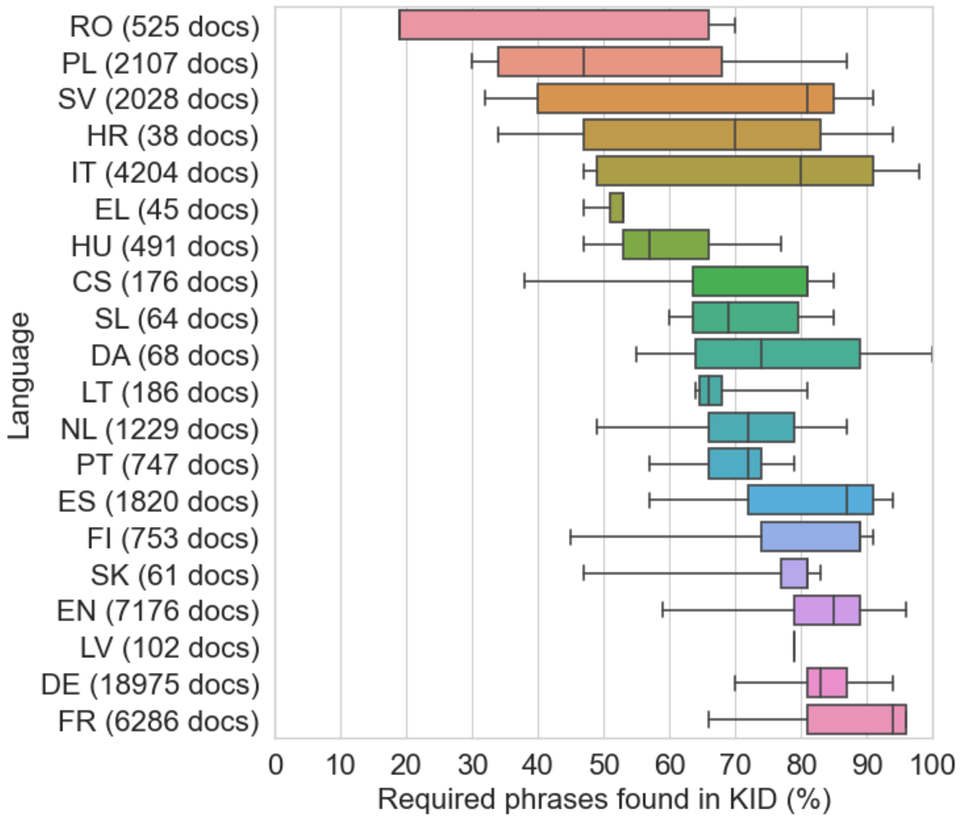
SRI consistent with volatility of product's performance



Note: The boxes and vertical lines indicate the range of returns (at the recommended holding period) across SRPs grouped by the SRI (the number of products in each sample varies slightly as information for some scenarios could not be retrieved from some documents). The SRI aggregates the estimated Credit Risk (default risk) and Market Risk (adverse market price risk) associated with the SRP. The necessary simulations and formulae used to produce the SRI are set out in the PRIIPs KIDs Regulation. The SRI ranges from 1 (lowest risk) to 7 (highest risk). The horizontal line in each box shows the median KID simulated return rate for that specific performance scenario and SRI grouping. Box edges are the 25th and 75th percentile simulated returns across the group, and additional lines ('whiskers') represent the 10th and 90th percentiles for that same group.

Required phrases

Phrases frequently not reported correctly or entirely



Note: Distribution of search success rates obtained as the number of required phrases found in a KID over the total number of required phrases which are mandated in the PRIIPs regulation, delegated regulation, and Q&A on PRIIPs KIDs. The box edges correspond to the 25th and 75th percentiles of the distribution, the outer segments ("whiskers") represent the 5th and 95th percentiles, and the line inside the box is the median.

Sources: ESMA, Structuredretailproducts.com, financial entities' websites.

Deep dive in NLP

Examples of current and recent projects (1)

- Analysis of key information documents of structured retail products
  - To assess aspects relevant to investor protection such as information completeness and linguistic complexity
  - To extract cost, performance and risk figures
- Analysis of ESG language and words
  - In credit rating agencies' press releases to assess the impact of ESMA guidelines on including ESG topics
  - In investment funds' names to support the development of ESMA guidelines on using ESG-related terms

Deep dive in NLP

Examples of current and recent projects (2)

- Analysis of prospectuses
  - For financial instruments to assess certain Prospectus Regulation provisions (e.g. risk factors, required statements)
  - For investment funds to assess Sustainable Finance Disclosure Regulation provisions, use of artificial intelligence in investment strategies

1. SupTech at ESMA: an overview
2. Data analytics for anomaly detection
3. Deep dive in NLP applications
4. **Looking forward**

Looking forward: ESMA’s vision

SupTech integral to ESMA’s “supervisory and facilitation” role (ESMA strategy 2023-2028):

- *“ESMA will continue to **facilitate NCAs’ supervision – strengthening convergence on use of the digital technologies and the use of SupTech tools**, including sharing best practices and undertaking joint projects.”*
- *“In close cooperation with NCAs and other EU authorities, ESMA will explore the centralisation of some supervisory technologies, in order to **pool resources and achieve efficiencies**.”*
- *“The objective will be to **share expertise and benefit from available opportunities**, and where possible making tools available to NCAs.”*

Looking forward: concrete prospects

- Exploit centralised repositories (e.g. European Single Access Point)
  - Facilitate collaboration between ESMA and NCAs on supervisory convergence topics (e.g. develop common thresholds for further action)
  - (later) develop common code repositories
- Leverage machine-readable documents
  - Reading PDF format leads to substantial loss of information + efforts to recreate the structure of the document
  - Machine-readable formats (e.g. XHTML) make extraction of information much easier and more comprehensive → Would unlock additional benefits of NLP

Any questions?

ESMA
European Securities and Markets Authority

Giulio Bagattini
Risk Analysis Officer
Economics, Financial Stability and Risk Department

201-203 rue de Bercy, 75012 Paris - France

Tel: +33 (0)1 58 36 64 88
Email: giulio.bagattini@esma.europa.eu

www.esma.europa.eu

@ESMAComms

European Securities and Markets Authority (ESMA)