Investor protection

54 000 PRIIPs KIDs – how to read them (all)

Contact: adrien.amzallag@esma.europa.eu

Summary

This article presents the results of an ESMA pilot exercise to apply natural language processing techniques on a unique dataset of c. 54 000 Key Information Documents that describe structured retail products produced under the Packaged Retail Investment and Insurance-Based Products Regulation. The techniques involved include measuring linguistic richness and semantic uncertainty, as well as sentiment analysis. This work – an application of SupTech – aims to illustrate how these techniques can produce useful measures for European supervisors, policymakers and risk analysts. Information extracted from text opens up new possibilities for supervisory assessments, for example with respect to information completeness and to legal requirements that a document be comprehensible to investors. In addition, text-based information is uncorrelated with (i.e. complementary to) numerical information, which can help policymakers determine if the legislation is working as intended. Lastly, text-based information can identify new sources of financial risks to investors.

Introduction

European retail investors now receive more information than ever, as transparency and disclosure requirements enacted following the 2007–2008 global financial crisis are implemented. The majority of this increased information is in the form of text, located for example in prospectuses and KIDs for funds or structured retail products.

It can be challenging for investors to make sense of so much information. It can also be challenging for supervisors, who are legally tasked with verifying these documents' compliance with a multitude of detailed requirements that span highly technical (and often lengthy) texts, produced by thousands of financial entities, across numerous languages and styles. It is, however, crucial – for investor protection, for orderly financial markets and for financial stability – that supervisors be able to effectively supervise this exponentially increasing amount of regulatory text.

This article summarises recent ESMA efforts to extract information of interest from a specific set of regulatory documents. The aim is to illustrate how natural language processing can assist both supervision and supervisory convergence, as well as evidence-based policymaking and risk-monitoring efforts by the public sector in Europe.

The article applies these perspectives to information extracted from a data set of KIDs for PRIIPs, most of which are structured retail products. Although the total number of KIDs is unknown, there are indications that tens of thousands are available, and that the market is worth at least several hundred billion euro. This market size, coupled with ongoing Joint Committee work to review the PRIIPs KID Regulation (Joint Committee of the European Supervisory Authorities, 2019), makes PRIIPs a worthwhile area for investigation and application of these techniques[135].

By law, a KID must be provided to retail investors when they consider purchasing a PRIIP. The structure, content, presentation and length of the KID are tightly controlled by the PRIIPs

[135] Structured retail products have also attracted some prior interest from both regulators and academics. See for example Demartini and Mosson (2020) and Célérier and Vallée (2017).

Regulation and PRIIPs KID Delegated Regulation[136]. For example, the PRIIPs KID Regulation specifies dozens of phrases that must be mentioned in specific locations in the KID. The KID must also include a variety of numbers calculated under specific formulae, such as performance under several scenarios. All of this information can be extracted and assessed.

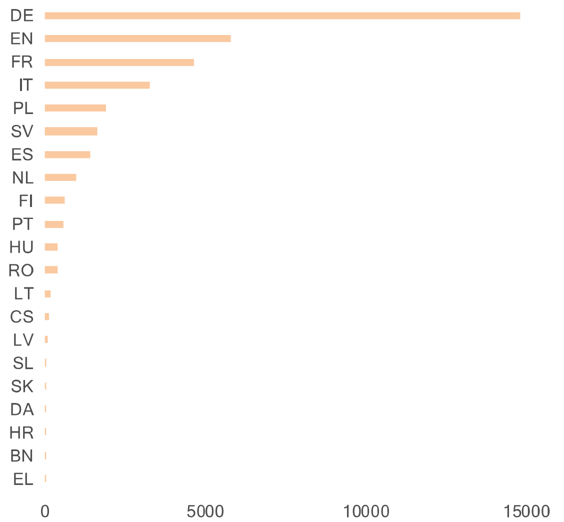
The next section describes the data set and methodology. Subsequent sections illustrate how this information can be used for supervision and supervisory convergence, policy development and risk monitoring. The conclusions connect these results with wider policy discussions.

Data set and methodology

The article uses a unique data set of 54 384 KIDs retrieved from public websites and a specialised commercial data provider, manufactured and sold in the EU by 333 unique issuers. These KIDs describe PRIIPs issued between 1 January 2018 (when the requirement to produce KIDs began) and 31 December 2020. The sample includes KIDs written in nearly all official EU languages – shown in Chart RA.1 below[137].

German-language KIDs are by far the most prevalent, followed by English-, French- and Italian-language KIDs. However, it is difficult to assess the extent to which this data sample is representative of the overall PRIIPs universe. KIDs are not centralised; there is no single location where they can be found. As a result, the total number of KIDs is unknown.

RA.1
Number of languages included in database
Much variation in available KIDs per language



Note: Number of KIDs grouped by language of document. Duplicate documents have been removed prior to graphing.
Sources: ESMA

In any case, a number of items can be extracted from a KID, such as the presence of certain words or phrases, various cost-related figures, simulated returns under different performance scenarios, the Summary Risk Indicator (SRI, discussed below), and descriptive information such as the product ISIN, issuance date and recommended holding period.

However, there are a number of technical challenges before this can be done. First, KIDs are nearly always provided in PDF format, which implies that text is 'frozen' and needs to be unpacked before it can be read and analysed by a computer. The conversion process means that the text loses its intended structure: tables are split, word order is reversed, and words can be duplicated. This inevitable (for PDF documents) step in natural language processing is time-consuming and prone to error, and destroys content. This leads to a recommendation for future policymaking: when a law requires the widespread production of documents, it is essential that these be made available in a flexible format such as open document format, even if in addition to PDF.

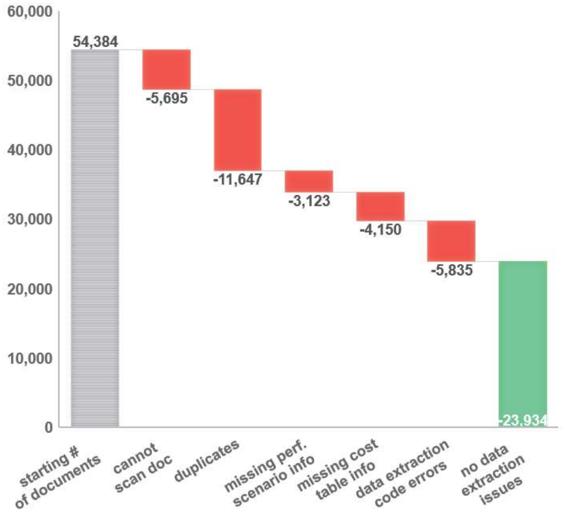
[136] Regulation (EU) No 1286/2014 (PRIIPs Regulation) and Commission Delegated Regulation (EU) 2017/653 (PRIIPs KID regulation).

[137] Country tables are not shown, because the same product may be sold in multiple countries. Duplicate products (i.e. the same product but with multiple KIDs across European languages) have been reduced to a single KID. Where multiple KIDs are available for the same product in the same language; the earliest (i.e. oldest) KID is used as a basis for these assessments. The aim is to focus on primary market issuance as much as possible.

A second technical challenge is in tailoring an algorithm to handle the inevitable idiosyncratic cases that arise when documents are written in multiple languages and styles. Thus, the exercise described in this article could never be a substitute for human review. The techniques outlined below aim to support public-sector agents in observing patterns and in conducting inspections effectively across tens of thousands of documents. It is in no way a recommendation that comprehensive reviews, decisions and/or sanctions be outsourced to machines.

Chart RA.2 below illustrates how these and other technical challenges reduce the data set by more than half. In the end, there remain about 24 000 KIDs that are entirely free from data extraction issues[138]. In addition, 81 % of KIDs in the sample refer to structured retail products, while about 19 % of the sample refer to funds (including, but not limited to, Category 2 products in the PRIIPs KID Regulation)[139]. For the remainder of the analysis, we exclude funds, in order to have as homogeneous a data sample as possible and given that in the PRIIPs KID Regulation, funds use different calculation methodologies to produce some of the metrics discussed below (e.g. performance scenarios). Lastly, insurance-based investment products and multi-option products are not included in the analysis.

RA.2
Reduction in data sample
Types of text extraction issues



Note: The vertical axis is the number of PRIIPs KIDs. 'Cannot scan doc' refers to technical issues when a PDF file cannot be converted to a text document (and instead a series of numbers and symbols appears). 'Data extraction code errors' refers to situations in which a computer code leads to inconsistencies in numerical information being extracted (i.e. numbers from some parts of the KID, e.g. on the stressful performance scenario, can be extracted, while information from elsewhere, e.g. the moderate performance scenario, cannot be obtained); this represents areas where the computer code can be further refined.
Sources: ESMA, SRP, individual financial entities' websites.

Natural language processing and supervision

This section considers how natural language-processing techniques can provide additional metrics to assist supervisors in enforcing compliance with disclosure requirements. We present below some first findings from the analysis of PRIIPs KIDs. This work will be further refined in cooperation with the NCAs going forward, including the consistency of KID phrasing.

Measuring information completeness

One key application is to compare the extent to which each PRIIPs KID includes the specific phrases it is required to mention. In total, there are approximately 65 distinct items that must be

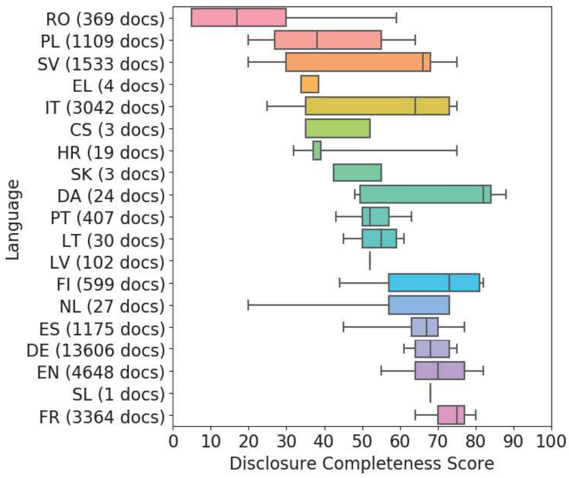
138 The sample size used in this analysis will vary depending on the topic. For example, assessing the completeness of information disclosures uses a sample closer to 30 000 KIDs, insofar as we seek to examine KIDs that also contain missing information. In addition, linguistic complexity measures do not require performance scenario information, so it is not necessary to focus only on KIDs that include performance scenarios. Where possible, the number of KIDs included in the specific analysis is mentioned.

139 UCITS and AIFs (the most common fund types) are currently out of scope of the PRIIPs Regulation. These funds must prepare a key investor information document (KIID). Although the KIID is an EU-wide information document, based on EU law, there is national discretion regarding the scope, for example on whether to apply the KIID regime to an AIF.

mentioned verbatim in the KID, regardless of the PRIIP subcategory. As a result, a 'completeness score' can be produced, which measures the number of phrases observed relative to the total number required to be mentioned[140].

Chart RA.3 below presents the range in completeness scores across EU languages, and shows several interesting facts. First, few KIDs in any jurisdiction achieve 100 % completeness of the required disclosures[141]. For example, for the 5 788 English-language KIDs, most mention between 64 % and 77 % of the required phrases, but in extreme cases only *c.* 55 % or up to *c.* 82 %. Second, the completeness score varies substantially across language groups.

RA.3
Range in disclosure completeness score by language
Many required KID phrases are not mentioned



Note: The vertical line in each box shows the median KID completeness score for that language group. Box edges are the 25th and 75th percentile scores, and additional lines ('whiskers') illustrate the 10th and 90th percentiles for that language group.
Sources: ESMA, SRP, individual financial entities' websites.

Chart RA.3 above raises interesting questions related to supervisory convergence across jurisdictions. For example, to what extent can 'similar meaning' be tolerated, if a required phrase is not included[142]? In addition, what is an acceptable threshold for a less-than-perfect completeness score, before further supervisory assessment should occur?

Table RA.4 below displays the phrases that are most often missing in KIDs[143]. It appears that descriptions of the cost tables, performance scenarios and SRI are the most challenging for PRIIPs KID manufacturers to comply with, compared with other required phrases such as standardised table and section headings, or elements to mention at the beginning of the KID. This information can help indicate focus areas for supervision, and/or areas where the legislation is misunderstood and guidance (such as questions and answers and guidelines) may be needed. It may also signal a need to adjust the legislation (see next section)[144].

RA.4
Top 10 required phrases not found in KIDs
Descriptive phrases appear most problematic

Asset type	Number of KIDs missing this item	% of KIDs missing this item
AVII (Descr. of costs, sent. 3)	18 244	61
AVII (Descr. of costs, sent. 4)	17 895	59
AV (Perf. Scen., Element C, sent. 1)	10 720	36
AIII.7 (SRI, Element A, sent. 2)	10 195	34
AVII (Descr. of costs, sent. 1)	9 462	31
AVII (Descr. of costs, T2, sent. 1)	9 084	30
AV (Perf. Scen., Element D)	8 843	29.4
AVII (Cost Table 1, row 2 text)	8 744	29.1
PRIIPS Regulation Art. 8(2) (sent. 3)	8 719	29.0
AVII (Descr. of costs, sent. 2)	8 518	28

Note: Table rows refer to regulatory requirements; the top 10 missing phrases in the KIDs data sample (after removal of duplicates) are shown. All rows denoted with 'A###' indicate an annex to the PRIIPs KID Regulation. Descr. of costs = presentation of cost information in the KID; Perf. Scen. = performance scenarios; SRI = summary risk indicator; T1 and T2 = Tables 1 and 2; sent. = sentence. See the abovementioned regulatory text for further details on the specific phrases in question.
Sources: ESMA, SRP, individual financial entities' websites.

It can be useful to combine the disclosure completeness score with other information sources. For example, Chart RA.5 below

[140] Phrases that are optional to mention are not included in this analysis; only mandatory disclosures are considered.

[141] A small number of KIDs in the data set may still achieve 100 % compliance (outliers are not shown in Chart RA.3).

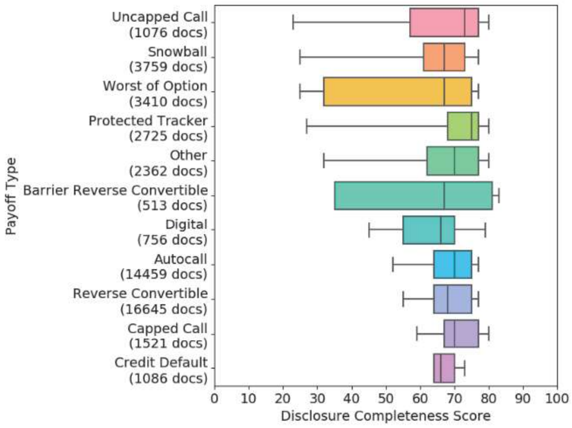
[142] Some tolerance is provided in the search function, e.g. for punctuation differences and capitalisation. However, the use of similar words, or word order being reversed in the same phrases, is not permitted, as it is assumed that the legislature had a clear intention in mind (i.e. standardisation) when going to the trouble of specifying, directly in the legislation, the phrase to be mentioned.

[143] 'Missing' here denotes cases where the phrase is entirely absent in the KID or is incorrectly copied from the legislation. See also footnote **Error! Bookmark not defined.**.

[144] When searching across multiple languages and document formats (i.e. templates from issuers), it is nearly impossible to eliminate false positives (i.e. indicating that a phrase is missing when in fact it is not). Results like those in Table RA.4 can also help indicate if the search algorithms are sufficiently precise.

illustrates the range in the completeness score for the most common PRIIP pay-off types in the data set. As can be seen, most difficulties in complying with the required disclosure phrases appear to be clustered among PRIIPs that include worst of option, autocallable (also known as knock-out) and/or barrier reverse convertible pay-off types. Another application area could be to group KIDs by PRIIP manufacturer, and thus identify, at the level of a supervised entity, manufacturers whose KIDs tend to have particularly low scores.

RA.5
Disclosure completeness score by pay-off type
Certain pay-off types may be worth focusing on



Note: The vertical line in each box shows, within each pay-off type, the range in the disclosure completeness score. Box edges are the 25th and 75th percentiles, and additional lines ('whiskers') illustrate the 10th and 90th percentiles for that pay-off type. One product can contain multiple pay-off types. 'Other' collects all PRIIPs in the data sample for which there are 400 or fewer observations for that pay-off type.
Sources: ESMA, SRP, individual financial entities' websites.

Measuring information complexity

KIDs are required to be written 'in language that is clear, succinct and comprehensible'[145]. These notions are also found in many pieces of EU law that involve disclosure requirements. For example, the Prospectus Regulation requires that the summary be written 'in language that is clear, non-technical, concise and comprehensible for investors'[146]. Elsewhere, MiFID II stipulates that 'All information, including marketing communications, addressed by the investment firm to clients or potential clients shall be fair, clear and not misleading' and that best execution policies must 'explain clearly, in sufficient detail and in a way that can be easily understood by clients, how orders will be executed by the investment firm for the client'[147].

'Clarity', 'comprehensibility', 'succinctness' and similar words are subjective concepts (which we refer to collectively as reflecting 'complexity'). Therefore, it can be challenging for supervisors to, first, assess a document according to these concepts and, second, develop an appropriate benchmark with which to compare documents[148].

At the same time, these requirements are not trivial. For example, the very first recital of the PRIIPs Regulation makes it clear that the main purpose of the KID is to facilitate investor understanding of products that 'can be complex and difficult to understand'[149]. If retail investors are unable to understand the information being provided to them, the investor protection motive mentioned immediately afterwards in the PRIIPs Regulation (recital 2) cannot be fulfilled.

The field of linguistics has developed a number of ways to assess the complexity of a text[150]. These range from basic metrics, such as sentence length, to more complicated econometric-based methods. We now apply several of these to the data set. Importantly, each metric chosen is language-independent, which means that we can safely compare KIDs across the data set,

145 Article 6(4)(c) of the PRIIPs Regulation. See also recital 13: 'To meet the needs of retail investors, it is necessary to ensure that information on PRIIPs is accurate, fair, clear and not misleading for those retail investors. This Regulation should therefore lay down common standards for the drafting of the key information document, in order to ensure that it is comprehensible to retail investors. Given the difficulties many retail investors have in understanding specialist financial terminology, particular attention should be paid to the vocabulary and style of writing used in the document. Rules should also be laid down on the language in which the key information document should be drawn up. Furthermore, retail investors should be able to understand the key information document on its own without referring to other non-marketing information.'

146 Article 7(3)(b) of Regulation (EU) 2017/1129. See also recital 27: 'In order to enable investors to make an informed investment decision, that information [information contained in the prospectus] should be sufficient and objective and should be written and presented in an easily analysable, concise and comprehensible form.'

147 Article 24(3) and Article 27(5) of Directive 2014/65/EU, respectively.

148 Demartini and Mosson (2020) assess complexity by counting the number of product features, and number of pay-off scenarios.

149 Recital 1 of Regulation (EU) No 1286/2014.

150 These measures can be applied to all types of language, including whether the document relates to a financial product or not. No external benchmark is necessary; the purpose is to identify outliers from within the PRIIPs KID universe, so as to guide supervisors for where to focus any human review efforts. Indeed, the approaches discussed in this section can only be additional to human review, which is necessary to conclusively determine whether a KID is truly clear, succinct and comprehensible

regardless of the language in which they were written[151].

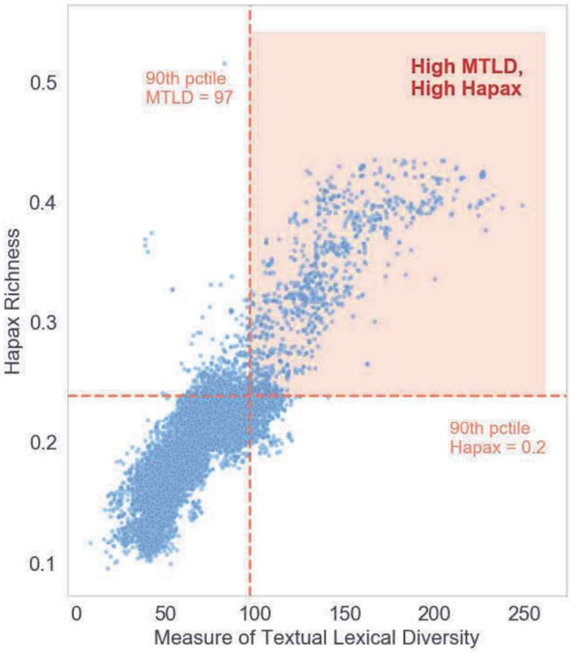
The subjectivity associated with these metrics cannot be eliminated; there is no unambiguous threshold beyond which a text can be said to be 'complex'. Nevertheless, these measures can facilitate supervisors' prioritisation of cases for further inspection, by identifying outliers. These can also be combined with further information (e.g. if several outliers are from the same issuer).

To begin with, Chart RA.6 below compares PRIIPs KIDs using two related scores: the measure of textual lexical diversity (MTLD) and the hapax richness. The MTLD is derived from the ratio of the number of unique words to the total number of words in the KID (the type–token ratio), corrected for differences in length[152]. Hapax richness measures the number of words that appear only once in the document relative to the total number of words. Both MTLD and hapax richness indicate the linguistic diversity of the text: greater diversity can indicate more precision, but can also indicate the presence of less common words (i.e. jargon) in the KID[153].

RA.6

Assessing KIDs using measures of linguistic diversity

Identifying extreme KIDs for further inspection



Note: The chart displays the hapax richness and MTLD for each KID in the sample used for this analysis (18 565 documents). Hapax richness is the number of words that appear only once in the KID, relative to the total number of words. The MTLD is derived from the ratio of the number of unique words to the total number of words in the document, subsequently corrected for differences in document length.
Sources: ESMA, SRP, individual financial entities' websites.

Chart RA.6 above demonstrates how visualisations can identify outliers. The MTLD and hapax richness are clearly positively correlated, and 90 % of KIDs are clustered in the blue cloud in the bottom left of the chart. However, there is less clustering and correlation in the top 10 % of the sample (i.e. above the respective 90th percentiles). This can provide an indication for prioritisation: KIDs in this upper-right region could be assessed first to determine if they are truly written in 'language that is clear, succinct and comprehensible'.

Chart RA.7 below assesses KIDs using two measures that examine language from an uncertainty perspective. The first is Yule's I metric, which measures the probability that two randomly selected words in a text are identical[154].

151 This rules out some popular metrics, such as average word and sentence length, the Flesch–Kincaid readability test (Kincaid et al., 1975), the Automated Readability Index (Senter and Smith, 1967) and the fog Index (Gunning, 1952).

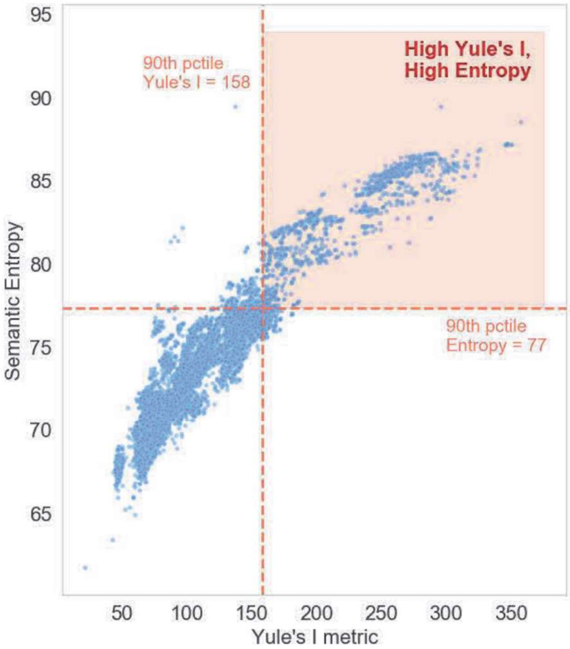
152 The standard threshold of 0.72 is used. See McCarthy and Jarvis (2010) and Tolochko and Boomgaarden (2019).

153 An additional approach could be to compare the frequency of words in PRIIP KIDs with the frequency of those words in general. However, the appropriate benchmark for these highly specific products is not clear (i.e. literature and the news, which are the most common types of corpus available for natural language processing, are not satisfactory). Moreover, we are working with 19 or 20 languages, so benchmarks would need to be language-specific.

154 Yule's I metric is calculated as $\frac{M_1 \times M_2}{M_1 - M_2}$, where M_1 is the total number of words in the document, and M_2 is the sum, across all unique words in the document, of the squared frequency of each unique word. See Yule (1944).

The second is semantic entropy, which measures how likely it is that a reader can predict the next word after the word they have just read in the text[155].

RA.7
Assessing KIDs according to linguistic uncertainty
Identifying extreme KIDs for further inspection



Note: The chart displays the Yule's I metric and semantic entropy for each KID in the sample used for this analysis (18 614 documents). Yule's I measures the probability that two randomly selected words from a text are identical. Semantic entropy measures how likely it is that a reader can accurately predict the next word after a given word in the KID.
Sources: ESMA, SRP, individual financial entities' websites.

Like the previous chart, Chart RA.7 above also identifies substantial clustering of KIDs in the bottom-left quadrant, which denotes 90 % of the data set yet covers only one third of the chart area. A much smaller share of KIDs (10 % of the sample) exists in the upper-right quadrant, which identifies KIDs with both high Yule's I and high semantic entropy. Extreme values for these linguistic uncertainty metrics may indicate KIDs that are particularly difficult for readers to follow, despite the PRIIPs Regulation requirement (Article 6(4)(c)) that KIDs '*be written in language and a style that communicate* in a way that facilitates the understanding of the information'.

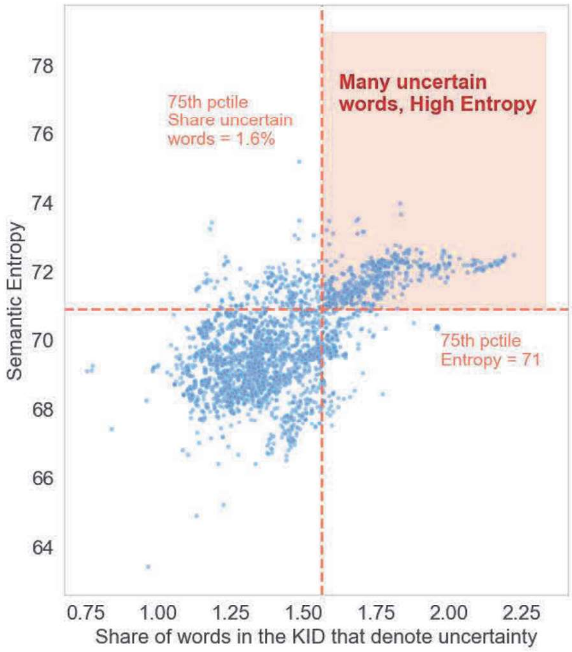
Another area where natural language processing has made a significant contribution is sentiment analysis, which assesses the overall 'feeling' associated with a given text. At first glance, financial documents may seem like a strange area to assess for emotive connotations. However, sentiment analysis can be useful in assessing uncertainty and possibility ('modality'), as well as positive or negative feeling. This has been assessed, for example, by Loughran and McDonald (2011), who also provide, and regularly update, a set of word lists associated with the above, and other, emotions.

To do this, we count the number of occurrences of words associated with 'uncertainty' and 'modality' in each KID (using only English-language KIDs in the sample). This number is then divided by the total number of words in each KID to form a normalised measure of uncertainty. It seems reasonable to assume that, the more words in a document are associated with a particular emotion, the more likely it is that investors reading that document will enter that emotional state.

This sentiment analysis-derived measure of uncertainty (and modality) is displayed in Chart RA.8 below, alongside the semantic entropy measure discussed above. It is clear from this chart that there is a positive correlation between the two metrics. As in the preceding charts, supervisors could potentially use these metrics to indicate which KIDs to first focus their limited resources on. KIDs with both high numbers of word denoting uncertainty and high semantic entropy may be worthwhile and primary candidates for further inspection, for example.

155 Calculated as $-100\frac{\sum p_i \times log\, p_i}{\log N}$, where p_i is the probability of observing a specific word in the document, and N is the total number of words (Shannon et al., 1963; Dale et al., 2000; Tolochko and Boomgaarden, 2019).

RA.8
Identifying KIDs with especially unclear language
Identifying extreme KIDs for further inspection



Note: The chart displays the share of each KID containing a set of words deemed to increase the uncertainty associated with understanding the KID according to the dictionary first presented by Loughran and McDonald (2011). 'Uncertain words' in the graph refers to the combination of the 'uncertainty' and 'modal' word lists provided by the above academic paper. The vertical axis provides the semantic entropy for each KID in the sample used for this analysis (3 546 documents). Semantic entropy measures how likely it is that a reader can accurately predict the next word that follows a given word in the KID.
Sources: ESMA, SRP, individual financial entities' websites.

These information complexity metrics can be combined with other information extracted from the KID. For example, the PRIIPs KID Regulation requires an SRI to be produced. The SRI aggregates the estimated credit risk (i.e. issuer default risk) and adverse market price risk associated with the PRIIP, and ranges from 1 (lowest risk) to 7 (highest risk). The necessary simulations and formulae used to produce the SRI are also set out in the PRIIPs KID Regulation.

An investigation was conducted into whether the SRI already, somehow, reflects the fact that a KID is written in more complex language (for example because products with greater risk require more complicated drafting to describe them). If so, the use of these information complexity measures is trivial. However, little co-movement was found between information complexity measures and the SRI. This supports the idea that information complexity metrics can provide supervisors with complementary insights[156].

Evidence-based policymaking

In line with the EU's Better Regulation principles, EU law is often reviewed and evaluated, for example to ascertain the effectiveness of certain provisions. However, it can be challenging for policymakers to gather a sufficiently large database to make such assessments, particularly for qualitative provisions. This section illustrates how natural language-processing techniques can support these efforts.

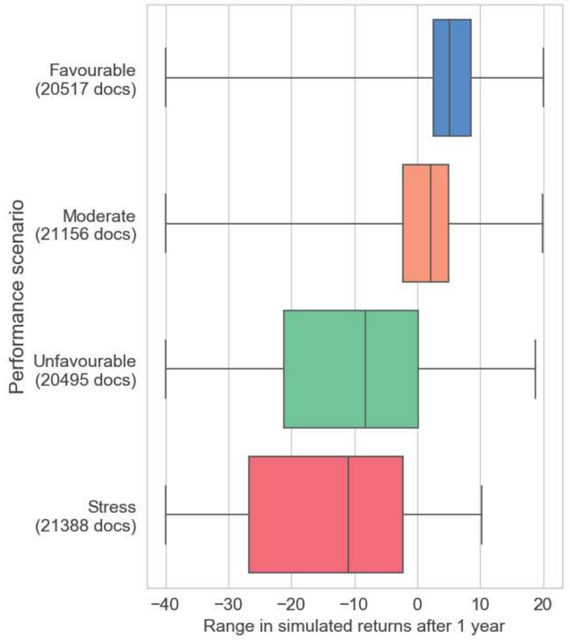
For example, PRIIPs KIDs must include simulated after-cost returns under at least four different performance scenarios. The calculation methodology is specified in detail within the PRIIPs KID Regulation. In particular, the simulations reflect performance under favourable (90th percentile of returns), moderate (50th percentile, i.e. the median), unfavourable (10th percentile) and stress (1st or 5th percentile, depending on the type of product) conditions.

Chart RA.9 below displays the variation in returns across these different scenarios. The simulated returns in both the stress and unfavourable scenarios are, as expected, usually below the moderate scenario returns. However, the simulated moderate and favourable scenario returns (blue and orange boxes, respectively) are both highly similar and clustered tightly (i.e. the boxes are not very wide). This raises the question of whether these scenarios sufficiently distinguish PRIIPs. In doing so, the chart provides evidence in support of the efforts of the Joint Committee of the European Supervisory Authorities in late 2018 / early 2019 to consult on revising the PRIIPs KID Regulation scenario calculation methodologies[157].

[156] Results, available upon request, are identical using hapax richness, MTLD and Yule's I metric.

[157] See Joint Committee of the European Supervisory Authorities (2019). One might think that product-specific differences could be driving such divergences across scenarios. However, the very large sample size suggests that the divergences go beyond product-specific features and are more related to the scenario calculation methodologies. Moreover, the results (available upon request) are unchanged if the difference between the favourable and moderate scenarios in each individual KID is first taken and the range for that difference is plotted (i.e. take the difference between the two scenarios within each product and then plot that difference).

RA.9
Added value of each performance scenario
Similar favourable and moderate scenarios

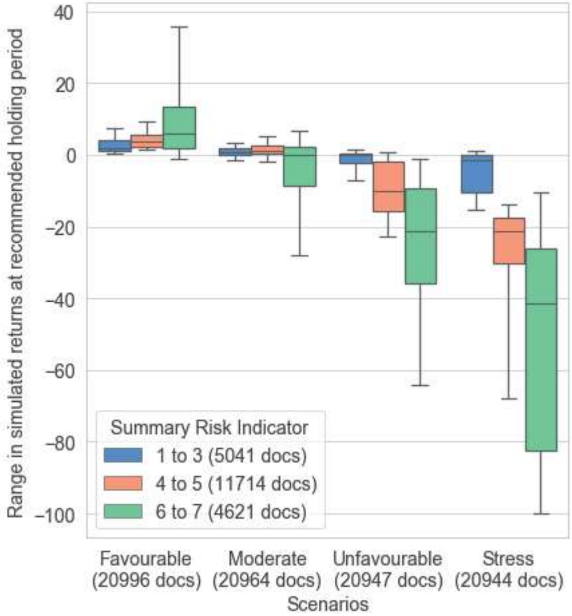


Note: The chart presents the range in performance returns of PRIIPs in each performance scenario category (favourable, moderate, etc.), using only scenarios that may occur after 1 year of holding the PRIIP. The methodology for calculating each scenario is set out in the PRIIPs KID Regulation. Similar results are obtained when comparing scenario returns at product maturity (or recommended holding period), rather than 1 year. The vertical line in each box shows the median simulated return in that performance scenario category. Box edges are the 25th and 75th percentiles, and additional lines ('whiskers') illustrate the 10th and 90th percentiles for that performance scenario category.
Sources: ESMA, SRP, individual financial entities' websites.

Next, Chart RA.10 below examines the extent to which the SRI varies with each performance scenario across PRIIPs. This can help policymakers assess whether the SRI complies with recital 5 of the PRIIPs KID Regulation, i.e. that 'information on the risks should be aggregated as far as possible and numerically presented as a single summary risk indicator…in order for retail investors to fully understand those risks'.

As can be seen in Chart RA.10 below, in the favourable and moderate scenarios there is little variation in simulated returns across SRI categories within the same scenario. This is sensible, because these scenarios reflect 'upside' or 'moderate' risk for an investor. However, in the more pessimistic unfavourable and stress scenarios (which are likely to more closely reflect the 'risk' situation that the legislature had in mind in the above recital), the SRI is associated with clear differences in simulated returns: the higher the SRI for a PRIIP, the lower the returns within the same scenario. This provides evidence for policymakers that the SRI calculation methodology in the PRIIPs KID Regulation is functioning as intended.

RA.10
Evaluating the summary risk indicator
SRI seems to differentiate PRIIPs effectively



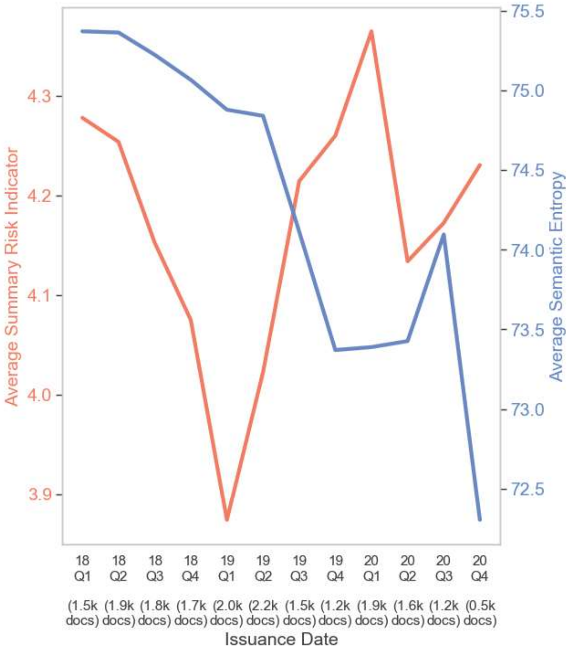
Note: The boxes and vertical lines indicate the range of returns (at the recommended holding period) across PRIIPs grouped by the SRI. The SRI aggregates the estimated credit risk (default risk) and market risk (adverse market price risk) associated with the PRIIP. The necessary simulations and formulae used to produce the SRI are set out in the PRIIPs KID 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') illustrate the 10th and 90th percentiles for that same group.
Sources: ESMA, SRP, individual financial entities' websites.

From words to risks

ESMA and many national authorities are tasked with assessing risks to financial markets, and in particular risks to retail investors. The texts of PRIIPs KIDs also contain insights useful for satisfying these mandates.

For example, following on from the previous section, Chart RA.11 below tracks developments in the average SRI together with those in semantic entropy in PRIIPs issued in each quarter since early 2018. Doing so allows one to observe how estimated product risks to investors (i.e. the SRI) are evolving over time and, in parallel, if the complexity of information provided to investors has moved in the same direction.

RA.11
Evolution in financial risks and linguistic complexity
Increasing product risk, but less complex text



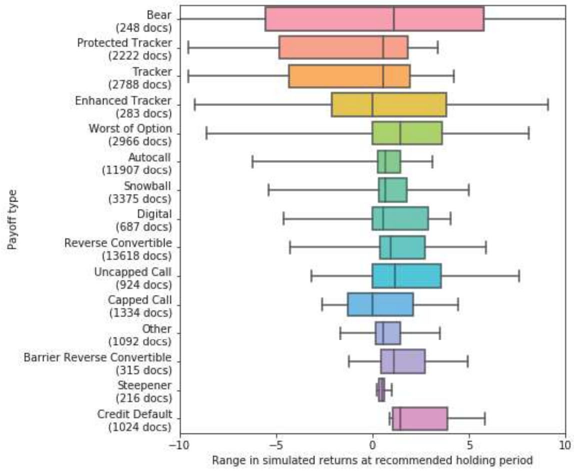
Note: The chart presents the average, for each quarterly issuance period, of the SRI (left-hand side, in red) and the semantic entropy (right-hand side, in blue) of PRIIPs in the data set since the beginning of the legislative requirement to produce a PRIIPs KID. The SRI ranges from 1 (lowest risk) to 7 (highest risk). Semantic entropy measures how likely a reader is able to accurately predict the next word that follows a given word in the KID.
Sources: ESMA, SRP, individual financial entities' websites.

As can be seen from Chart RA.11 above, it appears that the average SRI has, after falling during 2018, returned to and remained at the levels of the start of 2018. This suggests that there is a steady state of PRIIP risk for investors, at around the 'medium' risk level (using the description associated with the numerical SRI categories set out in the PRIIPs KID Regulation).

At the same time, the average uncertainty in KID texts has tended to fall since its peak at the beginning of 2018. This is interesting for several reasons. First, it confirms that the SRI and semantic complexity measures are complementary metrics rather than overlapping ones (as discussed in the previous section). Second, the mostly steady decline in semantic entropy could indicate that PRIIPs manufacturers are improving their compliance with the PRIIPs KID 'clear language' requirements (although human review would be needed to ultimately confirm this). Third, and following on from the previous point, although PRIIPs sold to retail investors are often around a 'medium' risk level, the clarity of presentation of those risks may be improving in parallel.

Risk to investors will also depend on the pay-off type of the PRIIP, and here as well text-based extraction methods may provide useful information for risk-monitoring efforts. To demonstrate this, Chart RA.12 below presents the variation in simulated moderate scenario returns across the data set, grouped by PRIIP pay-off type. Interestingly, a non-negligible share of PRIIPs in many pay-off type categories appear to offer negative returns were the moderate scenario to materialise, despite this being the 'middle' scenario (i.e. neither favourable nor unfavourable). It is unlikely that many issuers would voluntarily present such figures to potential retail investors, which demonstrates the wisdom of requiring, in the PRIIPs KID Regulation, that performance returns be expressed **net** of costs. However, there may be other reasons why simulated returns under the moderate scenario are negative (i.e. even without removing costs from the returns), such as the PRIIP pay-off type. In any case, this approach could help authorities identify the PRIIP types on which they should focus their efforts to make sure that investors are aware of the risks when making an investment.

RA.12
Moderate scenario returns across pay-off types
Many cases of low or negative scenario values



Note: The chart presents the range in moderate scenario returns (after costs) at the product maturity / recommended holding period for PRIIPs grouped by pay-off type. The vertical line in each box shows, within each pay-off type, the median moderate scenario returns (after costs) at the recommended holding period. Box edges are the 25th and 75th percentiles, and additional lines ('whiskers') illustrate the 10th and 90th percentiles for that pay-off type. One product can contain multiple pay-off types. 'Other' collects all PRIIPs containing pay-off types that have 150 or fewer observations in the data sample.
Sources: ESMA, SRP, individual financial entities' websites.

Conclusion

This article has presented the results of a recent ESMA pilot exercise to apply natural language-processing techniques on a unique data set of *c.* 54 000 PRIIPs KIDs produced between 1 January 2018 and 31 December 2020. These tools – a form of SupTech – can help supervisors, policymakers and risk assessors within the European public sector meet their respective mandates in areas that have seen a sizeable increase in regulatory documentation.

Natural language-processing techniques can help identify the extent to which regulatory documents mention required words and phrases. These techniques can also help in an area that can be challenging to assess, but is crucial for investor protection: the widespread legal requirement that documents be written in language that is clear and comprehensible. Linguistic complexity metrics, as well as sentiment analysis, can help supervisors to identify which documents, in preference to others, should be subjected to comprehensive supervisory scrutiny. Moreover, language-independent linguistic complexity measures can be useful in developing common benchmarks across the EU, which is useful for supervisory convergence.

Policymakers can also benefit from these techniques, which uncover additional areas in which to assess key legislative provisions. For example, data extracted from the KID help illustrate the effectiveness of the PRIIPs KID performance scenario calculation methodology (assuming, of course, that issuers comply with the calculation requirements). It also demonstrates that the SRI calculation methodology successfully distinguishes (*ex ante*) PRIIPs that carry greater risks for investors.

Risk-monitoring departments can also use natural language-processing techniques to refine their risk assessment activities. For example, the joint EU-wide joint evolution in the SRI and linguistic complexity over time suggests that the tendency of PRIIPs to remain around the 'medium' risk level may be tempered by less complexity in the language used to describe these products. Pending further human review of individual documents, this may help mitigate concerns about a return to the situation feared in recital 1 of the PRIIPs Regulation, namely that 'Existing disclosures to retail investors … often do not help retail investors to compare different products, or understand their features. Consequently, retail investors have often made investments without understanding the associated risks and costs and have, on occasion, suffered unforeseen losses.'

Information extracted from PRIIPs KIDs can also be combined with information from other databases, for example to identify PRIIP pay-off types in which simulated returns under the moderate performance scenario are negative for investors. This can help identify PRIIP types for which authorities may wish to particularly ensure that investors are aware of the risks when making an investment.

Natural language processing opens up powerful new possibilities for public entities to better meet their mandates and, ultimately, for more effective investor protection. European policymakers can continue to support the development of these activities by ensuring that, when a law requires the widespread production of documents, these are made available in a flexible format such as open document format, even if in addition to PDF. Centralisation of document provision is also crucial for supervisors, policymakers and risk analysis departments to have an overview (and thus sufficiently large sample sizes) of the universe of text available. European efforts such as the Commission's digital finance strategy and European strategy for data are likely to prove highly beneficial in this regard.

ESMA will continue to explore and apply these techniques where relevant, in conjunction with the European System of Financial Supervision.

References

Célérier, C. and Vallée, B. (2017), 'Catering to investors through security design: Headline rate and complexity', *Quarterly Journal of Economics*, Vol. 132, No 3, pp. 1469–1508.

Dale, R., Moisl, H. and Somers, H. (2000), *Handbook of Natural Language Processing*, CRC Press, New York.

Demartini, A. and Mosson, N. (2020), 'The complexity of structured products marketed in France: What impact has the AMF's action had?', Autorité des Marchés Financiers, January.

European Commission, 'Evaluating laws, policies and funding programmes'.

Gunning, R. (1952), *The Technique of Clear Writing*, McGraw-Hill, New York.

Joint Committee of the European Supervisory Authorities (2019), 'Joint Consultation Paper concerning amendments to the PRIIPs KID', 16 October.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom, B. S. (1975), 'Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel', *Research Branch*

Reports, No 8-75, Chief of Naval Technical Training, Naval Air Station Memphis.

Loughran, T. and McDonald, B. (2011), 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks', *Journal of Finance*, Vol. 6, No 1, pp. 35–65.

McCarthy, P. M. and Jarvis, S. (2010), 'MTLD, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment', *Behavior Research Methods*, Vol. 42, No 2, pp. 381–392.

Senter, R. J. and Smith, E. A. (1967), 'Automated Readability Index', Wright-Patterson Air Force Base, iii. AMRL-TR-6620, November.

Shannon, C. E., Weaver, W. and Burks, A. W. (1963), *The Mathematical Theory of Communication*, University of Illinois Press, Chicago.

Tolochko, P. and Boomgaarden, H. (2019), 'Determining political text complexity: Conceptualizations, measurements, and application', *International Journal of Communication*, Vol. 13, 21.

Yule, C. U. (1944), *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.