Risk Sciences 1 (2025) 100018

Contents lists available at ScienceDirect

Risk Sciences

journal homepage: www.keaipublishing.com/en/journals/risk-sciences/

A news monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions☆

Kris Boudt[a,b,c], Olivier Delmarcelle[a,b], Pascal Ringoot[a,d,*]

[a] Faculty of Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium
[b] Department of Economics, Ghent University, Belgium
[c] School of Business and Economics, Vrije Universiteit Amsterdam, the Netherlands
[d] National Bank of Belgium, Belgium

ARTICLE INFO

Keywords:
Anti-money laundering
Micro-prudential supervision
News event monitoring
Offshore leaks
Risk assessment

ABSTRACT

Investigative journalism plays a pivotal role in uncovering financial misconduct, yet its integration into anti-money laundering (AML) supervision remains underexplored. This paper presents a data-driven media monitoring system designed to provide timely alerts on news articles relevant to AML oversight in financial institutions. The system produces early-warning risk signals, offering actionable insights for qualitative assessments by micro-prudential supervisors. Relevance is quantified through thematic and entity-specific keywords, with these scores aggregated into weekly risk indicators. Through event analysis, we demonstrate the system's effectiveness by examining how Belgian newspapers covered eight offshore leaks between 2013 and 2021. Furthermore, we assess the system's robustness to machine translation and explore the use of prompting as an alternative methodology. The findings highlight the potential of integrating news-based alerts into AML supervision, enhancing the monitoring and response capabilities of supervisory authorities.

1. Introduction

Money laundering and the financing of terrorism (ML/FT) have long posed significant threats to global economies and societies. Investigative efforts by journalists, financial institutions, regulators, and other stakeholders play a pivotal role in addressing these risks. Regulatory frameworks mandate that financial institutions implement robust anti-money laundering and counter-terrorism financing (AML/CFT) systems to monitor customers and transactions effectively (Baesens et al., 2021). National competent authorities (NCAs) oversee the adequacy of these systems through prudential supervision, which includes both onsite and offsite inspections to ensure compliance with legal requirements. Among the tools available for identifying non-compliance risks, news articles frequently provide timely and actionable insights. High-profile examples, such as the offshore leaks investigated by the International Consortium of Investigative Journalists (ICIJ), highlight the potential of journalistic work to uncover misconduct. However, manual news monitoring is both time-intensive and prone to human error. A systematic, data-driven approach to monitoring news articles can enhance timeliness and reliability by quantifying relevance based on thematic

☆ The opinions expressed are strictly those of the authors and do not necessarily reflect the views of the institutions with which they are associated.

* Corresponding author at: Faculty of Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium.
E-mail addresses: pascal.ringoot@nbb.be, pascal.ringoot@vub.be (P. Ringoot).

https://doi.org/10.1016/j.risk.2025.100018
Received 15 September 2024; Received in revised form 11 April 2025; Accepted 12 April 2025
Available online xxxx

and entity-specific rules. Such an approach ensures that AML supervision teams can efficiently identify articles discussing pertinent topics or financial institutions under their oversight.

In this paper, we study the integration of automated news monitoring into the existing process. Specifically, we investigate the use of a news aggregator's Application Programming Interface (API) to access all available news from news providers and to transform that news into signals on which the AML expert can act. Our integrated setup enables us to run sophisticated queries allowing to be more selective to highlight articles wherein the supervised institutions are mentioned. The signal alerting the analyst is not at the article level but at the level of monitored period allowing drilldown to the news articles contributing to the signal. We set up an interaction between the AML domain expert defining the relevancy keywords and validating the signals, and the NLP domain expert implementing the rules that select the AML-relevant news. A further post-processing improves the selection. We find that by taking the grammatical structure of each sentence into account by means of named entity recognition and dependency parsing (Constant & Nivre, 2016, Straka & Strakova, 2017), the error rate can be significantly reduced in an automated way.

We apply the framework to financial institutions licensed in Belgium. In recent years, the preventive framework for combating money laundering and terrorist financing has undergone significant developments at the international, European and Belgian Level. The European Union integrated the anti-money laundering and countering the financing of terrorism (AML/CFT) measures into the legal framework in 2015 (EU Directive 2015/849[1]). In 2019, the European Banking Authority (EBA, 2019) included amendments of the Capital Requirements Directive (CRD) to indicate prudential supervisors should act on AML/CFT information and notify the EBA hereof without delay. In Belgium, the AML legal framework was implemented in the Belgian AML/CFT law of September 2017.[2] On 21 November 2017, the National Bank of Belgium (NBB) published a Regulation on prevention which is applicable to all Belgian financial institutions falling under its supervisory competence. As AML/CFT competent authority in Belgium, the NBB provides regularly updated information on its website with regard to all legal and regulatory obligations. Supervised institutions need to be compliant with the AML/CFT requirements, among other things they need to provide a yearly Anti Money Laundering Compliance Officer (AMLCO) report to the NBB and complete a yearly survey containing quantitative and qualitative aspects. This information is then verified by the AML/CFT analysts. Based on detailed analysis and expert judgement of these reports, a rating indicating how well the institutions copes with AML/CFT is assigned to each of the players in the Belgian financial market. Periodical onsite and offsite inspections are organised for each individual institution. The final goal of this process is to associate a risk level (high /medium-high/ medium-low/low) with each financial institution. If deemed necessary, the NBB informs international bodies like the EBA or ECB on a timely basis of the latest situation and important events of the financial institutions.

This yearly process using questionnaires is complemented with an event-based analysis. A trigger of such an event can be a media article discussing the involvement of a financial institution in money laundering activities. The active expert-based monitoring of such news event is prone to a delay in reaction to the news, and subject to analyst biases and errors (Antonakis, 2017). This analysis is related with the requirement of financial institutions to monitor early warnings for credit risk, ESG controversies and "adverse media" or "special interest person" for their customers. One solution is to rely on external "know your customer" (KYC) databases made available by risk data providers, but they are often limited by their focus on international news data and do not allow fine-tuning of the algorithms used.

The proposed news monitoring system enhances the AML analysis. For every monitoring period, the system computes an AML signal per monitored entity. When the signal exceeds the threshold, the action of the AML analysts is triggered. At this stage the AML analyst receives an overview of news articles that have contributed to the AML signal. This information supports the analyst decision making about changing the AML risk score. Based on the global risk assessment, the analyst takes measures and notifies the European Banking Authority or the European Central Bank if deemed necessary.

We implement the proposed system to monitoring the news in Belgium about the financial institutions supervised by the AML team of the National Bank of Belgium. Our input news article database consists of the main newspapers and magazines in Belgium, which are available through the Belga news agency's API. To ensure comparability over time, we focus on print articles only for which the number of articles has remained stable (in contrast with online news articles). There are on average of 3500 printed articles per day. This news article database was used before for ESG monitoring (Borms et al., 2021) and thematic indicator calculation (Algaba et al. 2023).

We show the results in terms of detected relevant articles per financial institution for news about the offshore leaks published by the International Consortium of Investigative Journalists (ICIJ). Our event analysis considers eight offshore leaks over the period 2013–2021. Relevance screening allows to reduce that number to three leaks in which Belgian banks were prominently mentioned, namely the Offshore Leaks (April 2013), Panama Papers (April 2016) and FinCEN Files (September 2020). The event window plots illustrate the timing of the articles. We find that media attention has positive dependence: following the attention spike on the day of the release, additional news articles are subsequently published by the newspapers and magazines not directly involved in the investigation as well as articles covering new elements of information such as the actions taken by policymakers.

This paper is structured in the following way. Section 2 introduces the proposed AML news event monitoring system describing the rules used for a systematic analysis of news covering the financial institution's involvement in AML events. Section 3 describes post-processing steps to improve the signal. Section 4 gives an overview of the data used in our analysis. The results are shown in Section 5. Subsequently, in Section 6, we examine the robustness to machine translation and compare the proposed implementation to alternatives, such as prompting that utilizes large language models. The final section concludes and presents an outlook for future work.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015L0849&rid= 2

[2] https://www.ejustice.just.fgov.be/eli/wet/2017/09/18/2017013368/justel

2. Detecting relevant news articles for AML supervision

A growing number of articles have shown the incremental information of news compared to surveys. Thorsrud (2020) created a daily business cycle index based on the newspapers to reflect economic fluctuations. Algaba et al. (2023) used the news for nowcasting consumer confidence. In this paper, we study the design of an automated News Monitoring (NEMO) system that enhances the AML analysis by using smart algorithms to execute the repetitive task of screening the news for relevant information for AML oversight and improve the information available to the analysts.

2.1. NEMO augmented workflow for micro-prudential oversight

Fig. 1 shows the proposed business process diagram of our approach to enhance analysis of the Belgian press regarding AML content. There are four panels. The first panel describes the news data inflow as provided by the news aggregator.[3] Our implementation uses the BelgaPress API. This provider aggregates all the articles from all Belgian sources and enables us to focus on institution and keyword relevance. The second panel visualizes the news event monitoring process based on entity and thematic relevance leading to a signal per period per entity. In the third panel, we describe actions for post-processing of the selected news to improve further the signal. The fourth panel describes the interaction between the AML analyst and the news event monitoring system. A first interaction consists of calibrating the list of keywords and the thresholds used. The calibration uses both supervised and unsupervised methods. A second interaction is triggered by the AML signal exceeding the threshold. At this stage the AML analyst receives an overview of news articles that have contributed to the AML signal. This information supports the analyst decision making about changing the AML risk score. A third interaction consists of contextualising the current AML score and events by analysing historical signals. Based on the global risk assessment, the analyst takes measures and notifies the European Banking Authority or the European Central Bank if deemed necessary.

The feedback loops within the proposed system facilitate the implementation of consistency checks between quantitative and qualitative information and integrates both categories in the proposed risk rating. Moreover, it enables the detection of systematic tendencies like the way a sector deals with specific countries.

2.2. NEMO: entity filtering

When an indicator for a financial institution is based on the news media, it is key to classify the articles correctly so that only those where the financial institution was the subject are kept. First and foremost, we need to understand how the financial institutions are referenced in the news.

Zhang and Liu (2012) introduce entity categorisation to refer to the grouping of actual words or phrases, also called *entity expressions*, that refer to a unique entity. In line with Barrière (2016), we use the term *surface forms* for the identification of mentions of named entities in text in this paper to make the distinction between the named entity itself. Articles that refer to "Bank of New York Mellon" for example might also use "BNYM" or "BNY Mellon" as surface forms.

For each institution i, we define a list of n_i surface forms, where each of them is a sequence of words $w_{i,j}$:

$$k_i = (k_{i,1}, ..., k_{i,n_i}) \text{ where } k_{i,j} \equiv (w_{i,j,1}, ..., w_{i,j,v_{i,j}}). \tag{1}$$

As soon as an article contains at least one of the surface forms k_i, it is considered to refer to institution i. By convention, the first surface form $k_{i,1}$ corresponds to the formal entity name. Most of the time, this will be the official name as defined in the Crossroads Bank for Enterprises in Belgium or Moody's Analytics BankFocus for institutions from abroad. To simplify the representation, we omit the time representation.

Entity relevance is measured using the BM25 algorithm (Robertson & Zaragoza, 2009). Starting from a set of keywords (for example, the possible names of a company), the BM25 assigns a score to each document, measuring how relevant the document is to the input keywords. BM25 records how often keywords appear in each document, penalizing words that frequently appear across documents and rewarding documents containing many keywords.[4] Further details concerning the BM25 algorithm are provided in Appendix A.

Given the surface form of a company k_i, (or, in other words, a query) and a news article a, the BM25 algorithm returns a positive score:

[3] One may wonder if a Google search is adequate to handle adverse media monitoring for financial institutions. The first problem is to find the best search criteria. Herein, the name of the institution is clearly crucial, but whatever keywords will be taken, chances exist the articles just used other words or synonyms. Using only the institution's name brings us to another problem, namely that of engine optimization, where the order of the articles found is often based on site optimisation and leads us to more marketing related content. A final issue is that premium news content behind paywall is not accessible through Google search. To avoid all these problems and focus on institution and keyword relevance, we opt to monitor news articles received from print media news providers.

[4] BM25 is also known as an advanced Term Frequency - Inverse Document Frequency (TF-IDF) scheme. In the simple TF-IDF scheme, relevance is determined by the ratio between Term Frequency (TF) and Inverse Document Frequency (IDF). The TF part counts the number of times the searched keywords appear in a document. The IDF part counts the number of documents in which keywords appear, penalizing highly frequent keywords.

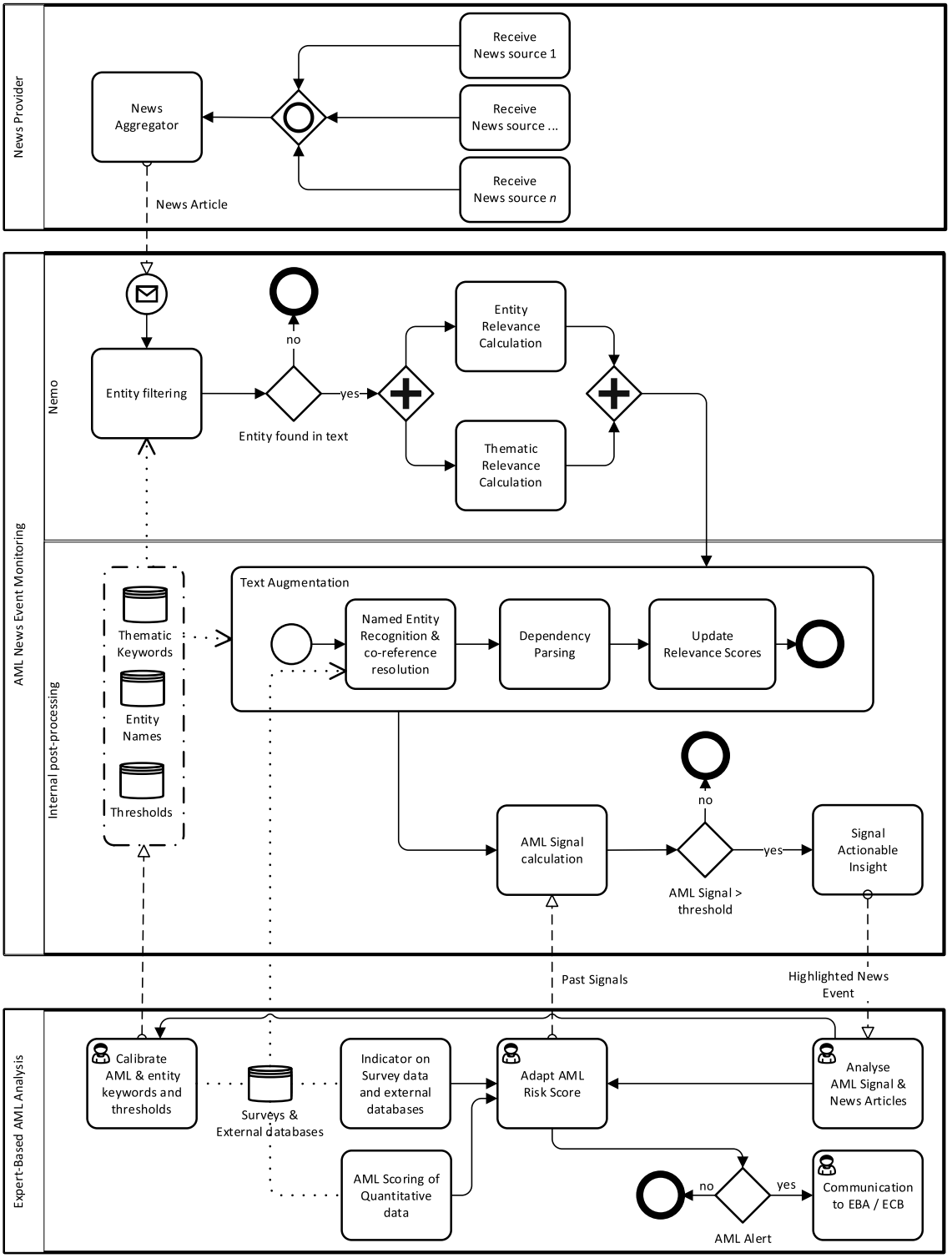

Fig. 1. Pipeline of the AML news event monitoring process to extract indicators and highlight news articles.

$$score_{a,i} = BM25(a, k_i). \tag{2}$$

This score is, however, not normalized and can only be used to rank documents with respect to a constant query k_i. To obtain interpretable scores, we apply a transformation from the value returned by BM25. Kanoulas et al. (2009) suggest modeling the

distribution of BM25 scores using a mixture of Gamma and Normal distributions. However, estimating this mixture in practice is only possible for large queries, comprising more than a dozen of keywords. Since the surface form of a company k_i generally contains only a couple of elements, the estimation of the Gamma-Normal mixture is not practical. Instead, we re-scale the BM25 scores to a bounded interval to enable comparability across companies.

We consider an entity relevance scale that is zero if the entity is not mentioned, and otherwise ranges from 25 to 100, with 25 being the minimum relevance for an article mentioning the company, and 100 the maximum achievable relevance. For the extreme situations where scores are positive but below a low threshold $l_{a,i}$ the value sticks at 25, while for the values above a high threshold $h_{a,i}$ the entity relevance is 100. In between, there is a linear transformation. Our rationale behind this choice is that BM25 scores do not possess upper bounds and can potentially reach very large values. We calibrate the low threshold to the historical minimum observed BM25 score, and the high threshold to the 90 % quantile. The latter safeguards the analysis against the effect of outliers in the BM25 scores.

The transformation to the new relevance scale can be formalized as follows:

$$U_{a,i} = \begin{cases} 0, & score_{a,i} = 0 \\ l_{a,i}, & 0 < score_{a,i} \leq l_{a,i} \\ score_{a,i}, & l_{a,i} < score_{a,i} < h_{a,i} \\ h_{a,i}, & score_{a,i} \geq h_{a,i} \end{cases}$$

$$Relevance_{a,i} = \left\{25 + 75\frac{U_{a,i}}{(h_{a,i} - l_{a,i})}\right\} * I(score_{a,i} > 0), \tag{3}$$

with $I(\cdot)$ the indicator function, $l_{a,i}$ and $h_{a,i}$ the low and high thresholds used for the BM25 score for entity i, and $\min(a, b)$ and $\max(a, b)$ being the minimum and maximum between a and b. Given the heterogeneity of media attention to firms, we set the thresholds to the minimum and 90 % quantile computed on the historical scores for every entity separately.

We illustrate the computation of the entity relevance score for two companies in Fig. 2. The top plot shows the distribution of the original score ($score_{a,i}$) while the bottom plot shows the effect of the transformation function (3). Specifically, when the score is 0, the transformed score is 0 as well. All strictly positive score values below the 10 % quantile are mapped to the low threshold value. Then the score linearly increases to reach its maximum of 100 for all scores above the high threshold (set to the 90 % quantile).

2.3. NEMO: thematic relevance calculation

For each article, the thematic AML relevance is assessed using a predefined list of language-specific keywords, denoted as k_{AML}. This list comprises essential terms identified by AML subject matter experts and serves as a standardized, institution-invariant reference for evaluating thematic relevance.

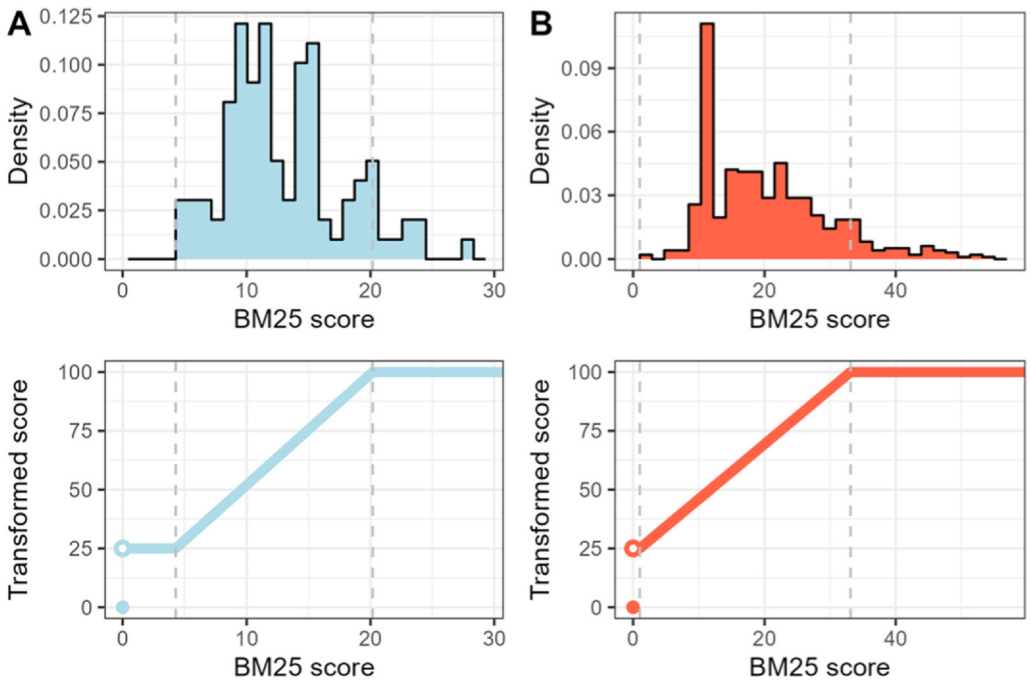

Fig. 2. Computation of entity relevance as a function of the entity BM25 scores. Each side (left and right) of the figure illustrates the relationship between the distribution of the BM25 scores and the transformation function for a specific company. The upper charts display the observed distribution of the BM25 scores. The bottom charts display the result of the transformation to the new scale for each company.

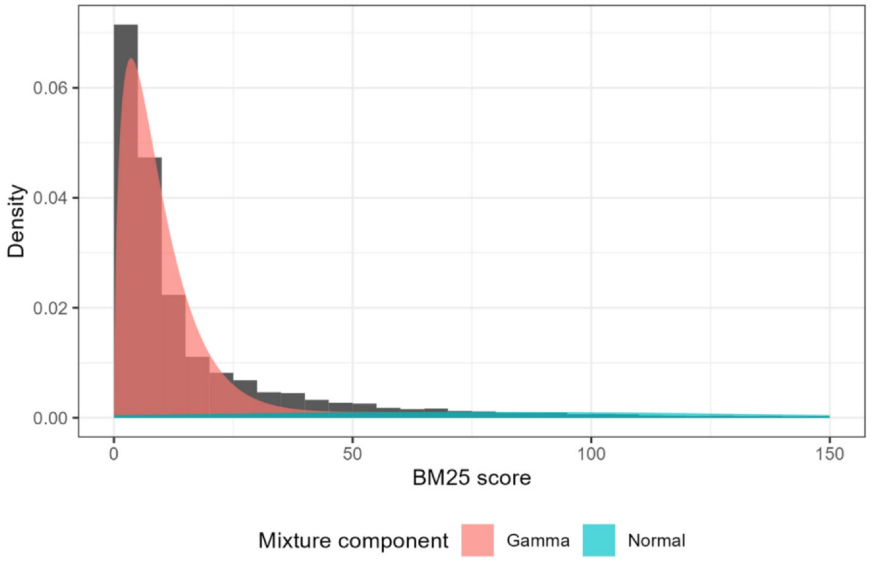

Fig. 3. Gamma-Normal mixture model on the thematic (AML) BM25 scores. The histogram describes the measured scores of documents matching at least one of the keywords. The left density (red) is the gamma component of the mixture, modelling the scores of the non-relevant documents. The right density (blue) is the normal component of the mixture, modeling the scores of the relevant documents.

As for entity relevance, thematic relevance calculation is based on the BM25 algorithm, computing the relevance to the set of keywords k_{AML}:

$$score_{a,AML} = BM25(a, k_{AML}). \tag{4}$$

As we compute the scores for all articles, the obtained scores have a mixture distribution, since some articles are relevant and others are irrelevant (Kanoulas et al., 2009). A gamma distribution can well approximate the scores of irrelevant documents, characterized by scores close to but above 0. In contrast, the score distribution of relevant articles can be assumed to be normal as the number of keywords increases. Fig. 3 displays the estimated Gamma-Normal mixture on the BM25 thematic scores. Since the Gaussian component models the relevant documents, it is particularly helpful in scaling the BM25 scores to a bounded interval. Using the normal cumulative distribution function, each document can be mapped to a number between 0 (not relevant) and 1 (highly relevant):

$$U_{a,AML} = \Phi\left(\frac{score_{a,AML} - \mu_{AML}}{\sigma_{AML}}\right),$$

$$Relevance_{a,AML} = \frac{U_{a,AML} - \Phi(-\mu_{AML}/\sigma_{AML})}{1 - \Phi(-\mu_{AML}/\sigma_{AML})} * 100. \tag{5}$$

The transformation in (5) thus maps the BM25 scores to a relevancy score between 0 and 100 using the Gaussian distribution function $\Phi(z)$. This monotone S-shaped transformation preserves the ranking. An additional min-max transformation is used to ensure that articles with a zero score effectively have a zero relevance. The minimum is reached when $score_{a,AML} = 0$ which corresponds to $\Phi(-\mu_{AML}/\sigma_{AML})$. The maximum of $\Phi(z)$ is of course 1. Hence the range used is $1 - \Phi(-\mu_{AML}/\sigma_{AML})$.

Consistent with the view that the BM25 scores have a mixture distribution, we recommend to estimate μ_{AML} and σ_{AML} using the maximum likelihood estimator under the Gamma-Gaussian mixture distribution.

2.4. NEMO: Overall relevance of an article

The overall relevance combines the entity relevance ($Relevance_{a,i}$) and the AML thematic relevance ($Relevance_{a,AML}$) into a single metric, representing the relevance of each article to both a financial institution and AML warnings.

The overall relevance of an article a in terms of discussing entity i in the context of AML is denoted by $Relevance_{a,i,AML}$. We compute it as the geometric mean of entity relevance and thematic relevance:

$$Relevance_{a,i,AML} = \sqrt{Relevance_{a,i} * Relevance_{a,AML}}. \tag{6}$$

In contrast with composite indicators using the arithmetic average, the geometric average does not allow compensation and is pulled towards the lowest of its terms. This property is desirable for determining the overall relevance, as it will heavily penalize news articles that are either 1) not focused on a given institution, or 2) irrelevant to the AML context.

The overall score can be used to effectively rank articles, helping analysts to target the most relevant news.

2.5. NEMO: aggregation to an AML warning signal

The analyst receives the warnings on a periodic basis. This requires aggregating the AML-entity relevance scores of the articles of the period considered. We do this using the sum of the underlying articles' score, weighted by a geometric series:

$$AMLSignal_{t,i} = Relevance_{t\,(1),i,AML}{*}weight_1 + Relevance_{t\,(2),i,AML}{*}weight_2 + Relevance_{t\,(3),i,AML}{*}weight_3 + ..., \tag{7}$$

where $Relevance_{t\,(j),i,AML}$ is the score of the j-th highest scoring article for entity i of period t, and $weight_j$ is the corresponding weight. To put more weight on high score articles, the weight function is a geometric series $weight_j = (ar^{j-1})$ where the parameters are set to $a = 0.8$ and $r = 0.2$. For instance, if we have three articles with relevancy scores equal to 90, 70, and 30, then the signal equals $90 * 0.8 + 70 * (0.8 * 0.2) + 30 * (0.8 * 0.04) = 84.16$. In case of an infinite number of articles, the total weight equals $a + ar + ar^2 + ...=a/(1 - r)$, which is 1 for $a = 0.8$ and $r = 0.2$. Hence if we have a large number of highly relevant articles, the signal will be close to 100. The decision to provide the AML analyst a warning is dichotomous and thus requires to use a cutoff value. As a rule of thumb, we use the threshold of 75. Extensive live testing validates this choice.

2.6. NEMO search relevance for AML enhanced news event searches

The scores generated from the product of entity and event relevance measure whether an article is of interest to an AML analyst. While it is appropriate when monitoring AML events as a broad theme, the analysis sometimes requires focusing on some affairs or investigating new risks. This then requires combining the entity and relevance scores with a specific *search relevance* score. Unlike entity and event relevance, search relevance is used solely to re-rank articles and does not need an interpretable scale. Thus, there is no need for a normalization step before using this relevance. This allows for very fast computation of search relevance, suitable for dealing with user requests.

For every article a previously processed and a user-defined query Q, the search relevance is the BM25 score for that article and query:

$$Relevance_{a,Q} = BM25(a, k_Q) \tag{8}$$

Then, the search relevance is combined with the existing relevance of that article for the *AML* event and an entity *i*. This determines the rank of each article in the results shown to the user:

$$SearchRank = \text{rank}(Relevance_{a,Q}{*}Relevance_{a,i,AML}), \tag{9}$$

where rank(.) is the rank function.

3. NEMO post-processing

Up till now, we have defined an independent relevance scoring of the entity and the AML thematic relevance. The resulting relevance signal enables us to do a first-pass filtering. The application of business rules with regular expressions to select articles containing all the known surface forms of the named entity k_i and thematic keywords k_{AML} is a good starting point and ensures a high-recall list of candidate articles at the expense of some false positives. Post-processing then allows to do further finetuning.

Now, we apply high-precision filters on the selected articles, by extending the selection process with more computationally intensive post-processing tasks. The goal is to reduce false positives where the institution was wrongly associated with a thematic keyword k_{AML} and depends on grammatical analysis of the text.

Suppose an article a is relevant for at least two financial institutions i and j and some thematic keywords k_{AML}. An in-depth analysis of the grammatical structure of the text is recommended to verify for which institution the news is relevant.

The analysis of co-occurrence of surface forms of the institution i and thematic keywords k_{AML} enables to detect a change of scope or subject throughout an article. By applying Named Entity Recognition (NER), places where an institution's surface form is not in the context of a proper noun and overlaps with a specific language construct are detected. Lemmatization allows us to detect all inflected forms of a word in a text and relate them with the word's lemma or dictionary form. Doing so enables enriching the list of thematic keywords k_{AML} so that it can be taken into consideration in a feedback loop.

To perform our text augmentation, we use UDPipe (Straka & Strakova, 2017) as a basis to perform the data preparation tasks. It provides support for tokenisation, Part of Speech tagging, lemmatisation and dependency parsing for several languages, including French and Dutch.

3.1. Named entity recognition

Named Entity Recognition (NER) is an important information extraction technique that was thoroughly evaluated within the Conference on Computational Natural Language Learning (CoNNL) framework. It involves the identification of proper nouns in text and a classification to a specific type (person, company, location).

Although UDPipe is capable to detect all proper nouns in the news articles, it does not perform NER. Since we are trying to identify the entities related to financial institutions, we can safely ignore entities referring to locations or dates and limit our scope to persons and companies.

3.1.1. Disambiguation and normalisation of named entities

Due to lack of uniformity in writing style and domain dependency, multiple surface forms of a single named entity are expected to occur inside an article (Barrière, 2016). Jijkoun et al. (2008) describe the impact of Named Entity Normalisation (NEN) for information retrieval in a question answering context. This goes beyond NER and addresses two phenomena "ambiguity" and "synonymy".

By means of example, let us have a look what might happen when selecting articles for the institution Corona Direct, a subsidiary company from Belfius insurance. Given the bunch of news articles that mention the disease 'Corona' and the increased risk in money laundering due to this, a lot of articles get highlighted as they combine the noun 'Corona' and some AML keyword, but these should be flagged as false positives.

The Corona Direct insurance company can be referred to by multiple surface forms in the news articles (e.g., "Corona Direct", "the Belfius subsidiary", its former name "Corona"). For some of the surface forms in a text, e.g., "Corona", it will not be clear it refers to the insurance company, the disease, or the brand of beer produced by a Mexican brewery.

Techniques like entity name disambiguation can answer such questions. These verify the co-occurrence of specific words that are oftentimes stored as features into a profile. Jijkoun et al. (2008) use Wikipedia data for the disambiguation of surface forms. Nguyen & Cao (2012) went one step further and combined the retrieval from Wikipedia with multiple iterations to create a semantic web. Using external databases like Wikipedia provides useful information about the location of the administrative branch or the members of the board for big financial institutions in our case, but it is of no use for smaller institutions that are hardly known and not mentioned on Wikipedia.

In our case, text from the reported AML compliance documents is available for all institutions and the surface forms entered herein can be used in the disambiguation process. Note that these surface forms are not necessarily identical to those used for the entity filtering, although most ones will be present as well. On top of this, specific keywords from the financial sector are included to ease the classification.

Initially, all named entities that are a superset of the institution's surface forms used for the entity filtering are accepted as surface forms to refer to the institution. As such, occurrences of the named entities "KBC CEO" or "KBC Group" are added to the list of surface forms for the institution whose initial surface forms contained "KBC". In contrast, the name of some institutions might be frequently used in person names. "Isabel" is such an example. To identify named entities that corresponds to locations, we verify if the entity's name contains the words 'street', 'avenue', 'square', etc. and query WikiData with each named entity to verify if it is a geographical entity.

On top of this, we integrate a feedback loop in our pipeline to store the profile for each entity. Each named entity found in a news article that is identified as a person or company is integrated in the institution's profile, like the names for the institution 'Isabel'. Analysts can then flag these, so they are considered as surface forms for subsequent NEN.

3.1.2. Coreference resolution

Suppose that, in article *a,* the financial institution is mentioned in sentence s_i and the money laundering activity in sentence s_{AML}. It is not trivial to conclude that the wrongdoing is done by the institution mentioned, especially if s_i and s_{AML} are not the same sentence. However, pronouns (e.g., it, he, her, whose, etc) are often referring to specific nouns preceding them. These can help identify a relationship between the financial institution and the wrongdoing.

Here, a technique called Coreference Resolution (CR) identifies linguistic expressions that refer to the same entity (Sukthanker et al., 2020). Important to note is that CR is far from being solved, let alone for Dutch and French. A commonly used heuristic to implement CR is to select the closest grammatically compatible mention in the subject position as its antecedent. In our pipeline, we apply CR that associates each pronoun to the named entity that was associated with the last surface form before it.

3.2. Dependency parsing

The aim of dependency parsing is to analyse the grammatical structure of each sentence and to establish relationships between the words. Such a relationship facilitates the identification of entities that are related to thematic keywords in the text and will detect additional false positives in the relevant articles. We take a simplified approach in our implementation and limit the dependency parsing to co-occurrence analysis. Specifically, we consider all thematic keywords in a sentence s_{AML} to be associated with the named entity in that sentence for as long as one is present. A sentence without named entity uses the entities that are mentioned in the latest sentences preceding s_{AML}. If none of these named entities has been identified as a surface form of the institution k_i, all thematic keywords k_{AML} in the sentence s_{AML} are ignored for the updated calculation of the relevance score.

3.3. Update relevance score

The relevance scores calculated by NEMO give a first approximation and can be adjusted thanks to the text augmentation tasks. Instead of using each occurrence of the name of the institution i in a text, we require the grammatically analysed text to contain a named entity that was accepted by the institution's profile. In addition, the relevance score used so far is based on a bag of words approach only considering the multiplicity of the words and thus ignoring the number of words between the thematic and entity keywords. We improve the relevance by using this distance to remove articles in which there is no paragraph that relates the two. A paragraph is hereby defined by a collection of at most κ consecutive sentences that are not separated by an empty line. In the implementation, we set κ to 5 and use the UDPipe sentence segmenter (Straka & Strakova, 2017)

$$Relevance^{*}_{a,i,AML} = Relevance_{a,i,AML} * J_{a,i}, \tag{10}$$

where $J_{a,i}$ is one when the article has at least one paragraph in which the entity and AML keyword co-occur and was not removed in the text augmentation phase. By plugging the improved relevancy scores into the signal definition of (7) we obtain the final definition of the AML signal

$$AMLSignal^{*}_{t,i} = Relevance^{*}_{t(1),i,AML} * weight_1 + Relevance^{*}_{t(2),i,AML} * weight_2 + \quad Relevance^{*}_{t(3),i,AML} * weight_3 + ... \tag{11}$$

This AML signal takes values between zero (no articles) and 100 (maximum score). The higher the score, the more evidence there is that relevant articles are present about that entity in an AML context for the period analysed. This AML signal needs to be compared with a threshold to trigger the attention of the AML analyst. Note that a high signal does not necessarily mean there is a severe issue, hence qualitative analysis by the analyst is still required. As such the news event monitoring enhances the work of the AML analyst but it can not replace the expert judgement.

4. Data and calibration

4.1. Universe of financial institutions monitored

The National Bank of Belgium is the competent authority with a supervisory role for many institutions that are grouped in four categories: Banks (100 in total); Insurance and reinsurance undertakings (70); Stockbroking firms (35); Payment institutions and Electronic money institutions (43).[5]

The AML supervision team of the National Bank of Belgium highlights the importance of AML/CFT prevention and raises awareness in the Belgian financial sector. For this process, specific guidelines have been released for each of the four mentioned sectors.

4.2. Role of questionnaires in AML event monitoring

Similarly, as for other competent authorities, the current AML micro-prudential performance monitoring process at the NBB is mainly manual and its execution is on a permanent basis.

All financial institutions in Belgium must report yearly an AML compliance document to the NBB. In this qualitative report, the AML compliance officer (AMLCO) of a bank, stockbroking firm, insurance or electronic money or payment institution provides six types of information:[6]

- The business model and a discussion on the changes regarding its products offered, target customer base or distribution strategy.
- If there was a change in the enterprise-wide risk assessment, its main conclusions, and impacts.
- Comments on the sufficiency of the human resources allocated to AML/CFT, application of the AML/CFT law, and possibly outsourcing of any AML related function.
- Description of any changes in the policies and procedures, training, trends and developments, or internal audit applied.
- Information on events on the control measures, results from the monitoring.
- Other points requiring attention.

As part of the implemented framework, once a year, the NBB organises surveys in which the financial institutions must provide information about business activities from the past year and to which extent the situation changed in comparison with the year before. Based on the sector and situation, some financial institutions are exempted from the quantitative questionnaire and only report the AML compliance report.

The timing varies slightly from one year to another but in general institutions have two months to complete the surveys. The domains covered in the survey are defined in such a way the supervisor gets a better idea of the risk taken by the financial institution and enables them to verify if the necessary measures were in response to proposed guidelines from the past. It provides a global overview of the customers' base, the distribution of the type of transactions and the concerned counterparties, as well as information about the people in charge to verify the institution's compliancy and the composition of the board in general.

The yearly NBB surveys have a quantitative and qualitative component. The quantitative information financial institutions must provide concerns[7]

- Figures about customers and their transactions split in several categories (per risk level, country, type of transaction, occasional, politically exposed persons (PEP), ultimate beneficial owners (UBO), etc.
- Information about their activity and if they have specific procedures for AML/CFT as well as figures and actions taken for atypical and suspect transactions for each of the type of clients.

[5] A complete list of supervised institutions is available on the NBB website https://www.nbb.be/en/financial-oversight/prudential-supervision/areas-responsibility for the two national languages French and Dutch.

[6] https://www.nbb.be/en/articles/circular-nbb202204-prudential-expectations-regarding-amlco-activity-report

[7] https://www.nbb.be/en/articles/circular-nbb202206-periodic-questionnaire-combating-money-laundering-and-terrorist

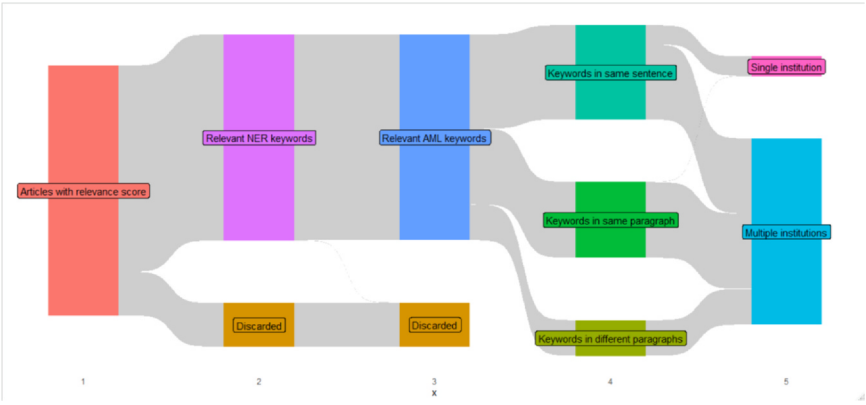

Fig. 4. The effect of applying named entity recognition and co-occurrence analysis in the pipeline.

- Figures about the staff located in Belgium, and assigned to specific divisions like internal audit, compliance, AML, etc. as well as the training for their domains.
- If specific AML/CFT services are outsourced and if specific measures are taken therefor.

Prudential analysts verify this information, compare it with the previous situation and put this in relationship with other public information. The final goal of this process it to associate a risk level (high / medium-high / medium-low / low) with each financial institution. Part of this evaluation is done automatically using the collected quantitative data, but the most significant conclusions are based on manual expert judgement.

In addition to responding to the survey, financial institutions need to communicate without delay important changes in their situation. The supervisor monitors other data, such as print media, to ensure that the relevant information is obtained on a timely basis. In the remainder of the paper, we describe an AML news event monitoring system that alerts the AML analysts of the publication of relevant news.

4.3. Universe of news articles

We use the same media archive in this paper as Algaba et al. (2023) of the Belgian News Agency (Belga) which contains around 27 million media news articles in Dutch and French over the period November 2001 until October 2022.

4.4. Effect of postprocessing

All signals obtained with the news event monitoring system must contain the institution's name and should have a thematic keyword. Fig. 4 shows the effect of the postprocessing on the articles of 293 signals between 2020 and 2021 having a relevance above 50.

In our sample, the named entity recognition filters 109 out of 623 articles (or 17 %) and labels them as false positives. Here, the feedback loop seems to play an important role.

- Some institutions have a name that was commonly used for persons (e.g. 'Isabelle', 'Alena'). In general, it does not harm to accept surface forms that have additional words before or after the institution's name. It is even preferred to catch occurrences that include a department, a group (Isabelle Group), or a specific function (Alena CEO) in the proper noun. To avoid the false positives, the feature is disabled for these cases so that only named entities with a perfect matching name are accepted. In the future, this will be solved using a feedback loop so that words like 'group' or 'CEO' will still be recognised as valid institution names.
- The name of some other institutions is also commonly used in Dutch or French and often not in the context of a proper noun (e.g. 'Integrale', 'FIL', 'Federale'). Here, the pipeline is able to detect each reference without modification.

The results strongly depend on the quality of the grammatical parsing and named entity recognition. For some sentences (with the institution name 'ABN AMRO' for example), only part of the name (i.e. 'AMRO') is detected as a proper noun, whereas the system associates 'ABN' as the verb. We fix this problem by taking into consideration the surrounding words as part of the proper nouns too.

We implement two extensions to the default filtering in the text augmentation process. The first is that thematic keywords are optionally reduced to the required list of nouns, auxiliaries, proper nouns, and verbs. Doing so enables the system to accept all other part-of-speech language constructs. In this regard, the French keyword "transactions louches" (shady dealings) matches "transactions très louches" (very shady dealings). A second optimisation is the use of lemmas. Lemmatization offers a solution as it can transform all inflected forms of a word in a text and relate them with the word's lemma or dictionary form. Applying the advanced thematic keyword filtering did not reveal new occurrences in the text for the given sample.

When analysing the co-occurrence of surface forms of the institution in combination with thematic keywords, we distinguish three situations

- "Keywords in the same sentence" means both the institution and the AML keywords are present in a single sentence. In these situations, there is clearly a link between the institution and the AML theme.
- Documents classified under "Keywords in the same paragraph" will have no co-occurrence of the institution and AML keywords in the same sentence as the institution. Coreferences (Sukthanker et al., 2020) can still put the link between the two sentences.
- Chapters are separated with blank lines, and in the third category the institution's keyword is in disjunct paragraphs than the thematic keywords. This is clearly the weakest possible link. Although not impossible there are coreferences, it is more likely another person or company is mentioned in the paragraph with the AML keywords.

We see that the number of pairs that pop-up in a single sentence (235 documents) is higher than the combination in a single paragraph (189). Given only articles with a relevance score above 50 are taken in our sample, chances are higher both keywords co-occur in the same sentences. A small test shows that if we do not filter on relevance score before, the number of pairs in single sentences was clearly lower than the pairs in a single paragraph.

Most documents have references to other organisations and persons on top of the supervised institutions (approximately 90 %). This means some articles do not necessarily relate all the AML keywords with the supervised institution or do also relate them to other entities and persons.

Only 89 documents (or 20 %) have the institution and the thematic keywords in different paragraphs. These are most likely false positives as all the documents in this category refer to multiple organisations and persons.

4.5. Overview of relevant articles

Fig. 5 gives an overview of the distribution of the frequency of relevant news articles of the four financial sectors supervised by the NBB. The font size reflects the number of news articles with an AML relevance above 75 in the Belgian press. Given the importance of financial transactions for credit institutions, it is not surprising the banking sector is the most important of all. The insurance sector is the second most important, but important to note is a single group is often active in both insurance and banking. The entity keywords used in our implementation do not disambiguate between them, which explains why the number of relevant articles for the bank is comparable to the insurance of the same group. This is clearly the case for KBC, Belfius and AXA as we see in the word cloud. The electronic money and payment institutions are rarely mentioned in the news articles detected as relevant for the AML supervision.

Given the importance of the banking sector, the risk of double counting when adding the insurance companies, and the small representation of the other two sectors, we'll concentrate solely on the banking sector in the next chapter.

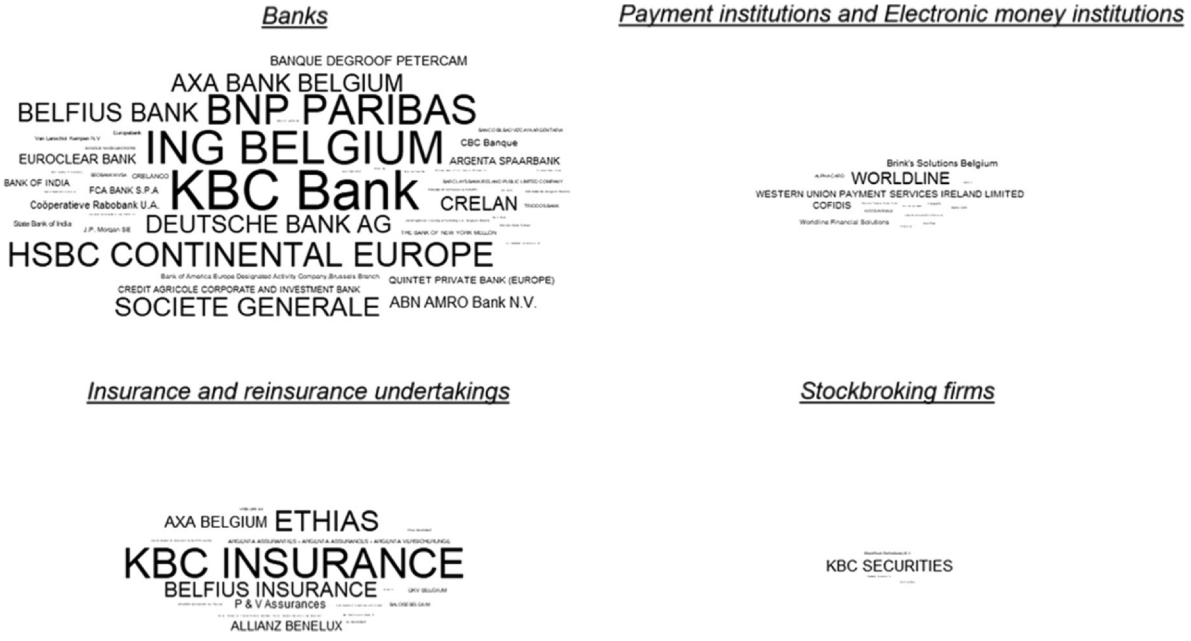

Fig. 5. Word cloud of mentions of the financial institutions in the articles with an AML and entity relevancy above 75. The word cloud shows only institutions supervised by National Bank of Belgium and grouped by sector: banks (top left), electronic money and payment institutions (top right), Stockbroking firms (bottom right) and Insurance sector (bottom left). The size represents the number of relevant articles found between 2000 and 2023.

5. Compliance signals for Belgian banks around the publication of offshore leaks

Since the offshore leaks publication in 2017 by the International Consortium of Investigative Journalists (ICIJ), there have been seven additional significant leaks publications. They received attention by the news media, but not all of them were relevant for the AML supervision of banks in Belgium. In this section, we use the proposed news monitoring system to identify which leaks have been most prominent.

The ICIJ started in 2017 as a project of the American Center for Public Integrity. It has since then received several extensive and complex data sets through data leaks from across the world (Berglez & Gearing, 2018). The results of the investigation are published in news articles. Reporters typically set a joint publication date month ahead in order to allow them to work through the data.

Table 1 shows the number of relevant articles per event for each of the supervised institutions. We count an article as relevant when both entity and thematic relevance is above the 75 threshold. Institutions for which the signal was below the threshold of 75 for the entity as well as the thematic relevance are not represented. Based on Table 1, we conclude that the relevance screening suggests that there are three leaks in which Belgian banks were prominently mentioned, namely the Offshore Leaks (April 2013), Panama Papers (April 2016) and FinCEN Files (September 2020). For each of these leaks, the relevancy analysis concludes that there are more than three articles discussing more than one financial institution. The accusation mentioned in the Offshore Leaks is that banks were involved in the creation of business subsidiaries in tax havens such as the Cayman Islands and the Channel Islands. The Panama Papers exposed the widespread use of offshore entities for illicit activities. They revealed how wealthy individuals, politicians, celebrities, and businesses worldwide used offshore companies and tax havens to hide assets, evade taxes, and, in some cases, launder money. Similarly, as for the Offshore Leaks, several Belgian banks are mentioned in the Panama Papers newspaper articles for their role in facilitating the creation of offshore companies for their clients. The FinCEN Files leak underscored weaknesses in the anti-money laundering systems of some big Belgian banks, prompting calls for reforms and stronger regulatory enforcement.

In Fig. 6 we zoom in on the counts of relevant newspaper articles detected around the release of these three major leaks. We focus on the twenty days following the release. The post-event window analysis is useful since not all newspapers and magazines are involved in the ICIJ, explaining a delay in reporting. Moreover, as also mentioned by Díaz-Struck and Cabra (2018), the sharing of the ICIJ data with the public has the advantage of giving a “second life to the investigation” leading to additional discoveries and discussions. For this reason, it is useful to not only study the news on the day of the ICIJ release, but also the presence of relevant news following the initial publication.

Note first that we only consider the analysis of print articles. When there is no relevant article on the day of the release, it means that the scoop was given to non-Belgian newspapers or magazines. While for the FinCEN Files there are publications on the day of the release, there is no relevant news detected on the day of the ICIJ announcement of the Offshore Leaks and Panama Papers. For instance, several Belgian newspaper articles discussing the Offshore Leaks refer to an earlier article published in the French newspaper ‘Le Monde’.

The spike in relevant articles on Belfius five days after the release of the FinCEN files strikes not because of their involvement, but their relationship with a gold trader, Tony Goetz. It suggests renewed interest in the regulatory and procedural aspects of a ‘Suspicious Activity Report’ for which Belfius, constrained by Belgian Banking Law at that moment, was unable to seek additional clarification. The event underscores the broader regulatory hurdles banks encountered when trying to address potential money laundering cases.

We further note that there is no regular pattern: each leak is different, highlighting the need of a performant news monitoring system alerting the AML supervision team about the publication of relevant articles. Thanks to the whistleblowers and investigative journalism, relevant news articles are published, that are signalled by the News Monitoring (NEMO) algorithm to the AML analysts who needs this information to assess the AML systems of the supervised financial institutions.

6. Robustness analysis

The proposed system analyses each article in the local language. This comes at the cost of maintaining a lexicon for each local language and possible inconsistencies across languages. To avoid this, Ashwin et al. (2024) and Barbaglia et al. (2024) use first machine translation and analyze all texts in English. While such machine translation increases the computational burden of the analysis, it reduces the complexity of maintaining up-to-date keywords lists and increases the consistency between the analysis of articles published in different languages. Ashwin et al. (2024) conclude that lexicon-based sentiment analysis in the context of nowcasting GDP growth is robust to machine translation. In this subsection, we study whether this is also the case in the context of news analytics for anti-money laundering supervision.

Another robustness analysis we undertake involves replacing the dictionary-based BM25 algorithm with prompt-driven approaches in a generative AI setting. In this method, a large language model (LLM) executes the task of identifying relevant articles based on carefully designed prompts. This approach eliminates the need to specify a lexicon and compute relevancy using BM25, instead relying on tailored instructions capable of addressing edge cases. Additionally, the prompt-based method can seamlessly process articles in multiple languages. However, this versatile approach has notable drawbacks, including higher computational costs and reduced transparency. In our context, it also encounters a significant limitation: most publishers do not provide licenses that permit the application of prompts to complete news articles.

We perform our robustness analysis on the 20-day period following the ICIJ leaks announcement for the two leaks with the highest number of financial institutions involved, namely the Offshore Leaks (2013–04–03, 6 financial institutions with at least three relevant articles) and the FinCEN files (2020–09–20, 7 financial institutions). In total we have 975 Dutch and 906 French articles with

Table 1
Signals with articles wherein ICIJ or a reference of the concerned leak is mentioned. Only articles whose entity and thematic relevance are above the threshold of 75 are shown.

	Offshore Leaks	Luxembourg Leaks	Swiss Leaks	Panama Papers	Paradise Papers	FinCEN Files	Pandora Papers	Congo Hold-up	Total
ABN AMRO Bank N.V.	2			1					**3**
BELFIUS BANK	4					12			**16**
BNP PARIBAS	11	1	1	3		6		2	**24**
CREDIT AGRICOLE	3	1							**4**
DEUTSCHE BANK AG	1			2		5			**8**
HSBC CONTINENTAL EUROPE	4	2	89	9	3	12	1		**120**
ING BELGIUM	14			1		52		10	**77**
KBC Bank	2			5		5			**12**
QUINTET PRIVATE BANK	3								**3**
AXA BANK BELGIUM		1							**1**
BANQUE DEGROOF PETERCAM		1		2		1			**4**
CBC Banque		2							**2**
SOCIETE GENERALE		1		3					**4**
BANQUE NAGELMACKERS				1					**1**
BARCLAYS BANK						2			**2**
J.P. Morgan SE						3			**3**
BANK OF NEW YORK MELLON						1			**1**
Total number of financial institutions with at least three relevant articles	6	0	1	4	1	7	0	1	
Total number of relevant articles	44	9	90	27	3	99	1	12	**285**

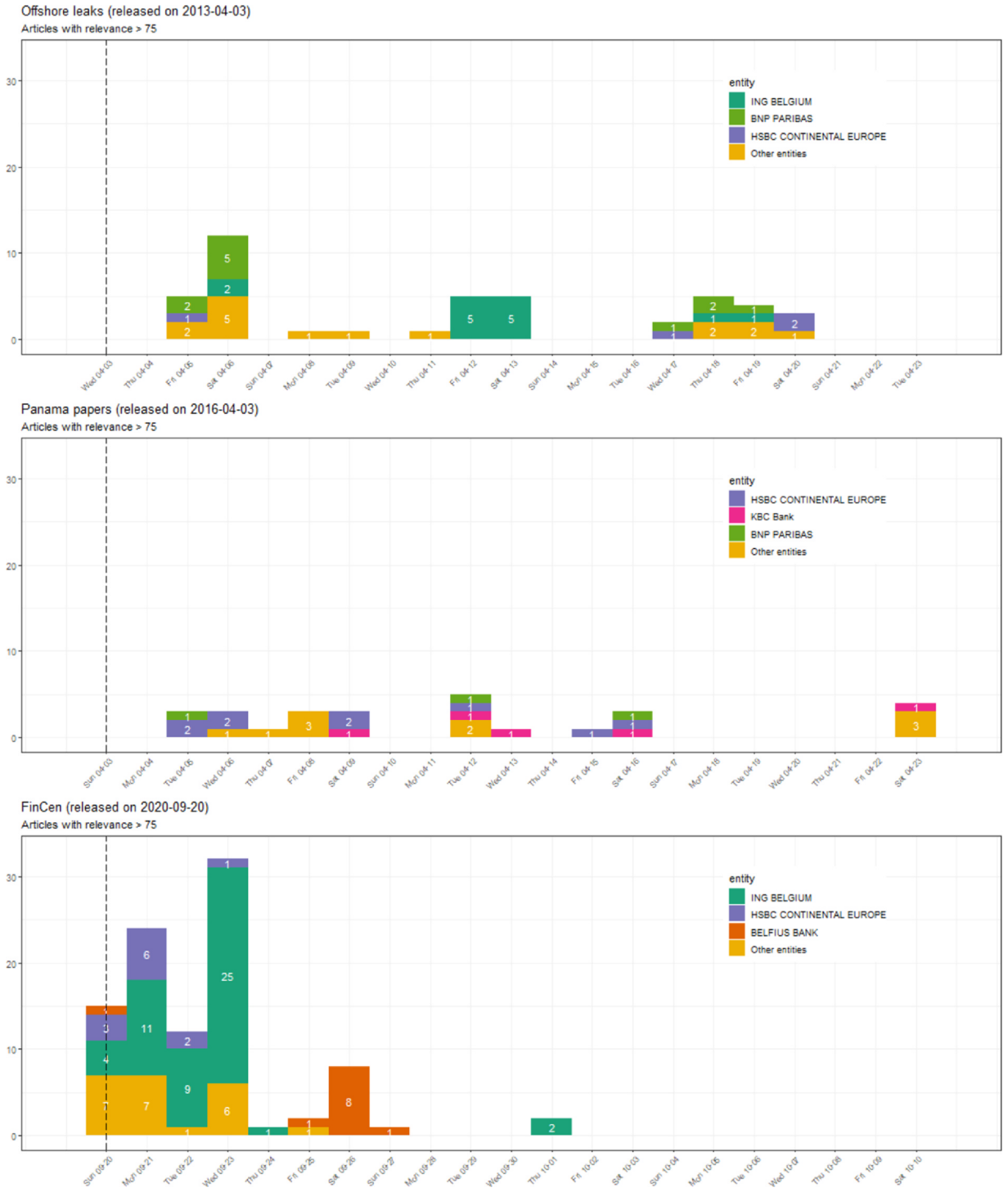

Fig. 6. Distribution of number of relevant articles around the release of the Offshore Leaks, Panama Papers and FinCEN Files.

entity relevance higher than 75 for these two periods. We first use Google's Compact Language Detector 2 to find the language in which the article was written. Although, the article's source provides already a good indication for the language used, some online sources provide text in different languages. Manual inspection of the articles for which an unexpected language was found referred to sports results or stock prices and are manually rectified. Finally, like Ashwin et al. (2024), we use the Google Translate API to translate each of these articles in French and English leading to a total corpus of 1881 articles in each of three languages. For each of these articles, we compute the thematic relevance using 4 different approaches:

1. **Baseline**: BM25 using the baseline lexicon in Dutch (resp. French) when the language of the original news article is Dutch (resp. French)
2. **NL2FR and FR2NL Articles**: BM25 using the baseline lexicon in Dutch (resp. French) when the language of the original news article is French (resp. Dutch) but the article itself has been translated to Dutch (resp. French)
3. **NL2EN Lexicon**: Using the to English translated baseline Dutch lexicon and the article (Dutch or French) is also translated to English
4. **FR2EN Lexicon**: Using the to English translated baseline French lexicon and the article (Dutch or French) is also translated to English.

Using the cutoff of 75 for the thematic relevance scores, this implies four binary classifications of the relevance of the news article for the AML team.

A key advantage of keyword-based approaches is their low cost and speed as compared to the increasingly popular approach of using a Large Language Model (LLM) to classify articles. To benchmark the outcomes, we created the prompt in Appendix B to analyze the news articles in Dutch, French and English. The prompt is implemented using the service Azure OpenAI which host a number of large language models. We used the prompt in combination with the model 'gpt-4o-mini' version '2024–07–18'. A new interaction with the model is created for each analyzed article, discarding previous answers.

Using the output from the prompt, we obtain three additional binary classifications of the relevance of the news article for the AML team:

5. **Prompting**: Using the prompt on the original article
6. **NL2FR and FR2NL Prompting:** Using the prompt on the to Dutch (resp. French) translated article when the original language is French (resp. Dutch)
7. **NL2EN and FR2EN Prompting:** Using the prompt on the English translated article

This leads to a total of seven binary classifications for each article. We study the coherence between the classifications using Cohen's kappa (Cohen, 1960). Unlike accuracy, Cohen's kappa does not require to know whether each article is truly relevant. The metric evaluates the agreement between two classification methods, yielding a value between -1 and 1. The result can be interpreted using the scale described by Landis and Koch (1977): [<0] Poor agreement. [0–0.2] Slight agreement. [0.21–0.4] Fair agreement. [0.41–0.6] Moderate agreement. [0.61–0.8] Substantial agreement. [0.81–1] Almost perfect agreement.

The Cohen's kappa between the seven classifications are shown in Fig. 7. The lower (resp. upper) diagonal corresponds to the results for articles in Dutch (resp. French). Overall, agreement between most methods ranges from "moderate" to "almost perfect", indicating that the AML information value is still captured on the translated text.

Prompt approaches have an almost perfect agreement ($0.81 \leq \kappa \leq 1.00$) among them. This highlights the capabilities of large language models in dealing with multilingual content. Whether the prompt is applied to English, native, or translated text, results are very similar.

In contrast, the BM25 block shows more sensitivity to translations, with substantial differences between Dutch and French. The agreement between Baseline BM25 and the "NL2FR/FR2NL" method, which simply involves a translation of the original text, amounts to 0.79 for Dutch texts and 0.61 for French texts. This suggests that the translation from French to Dutch is less likely to retain the keywords used in the BM25 lexicon. Other translations involving Dutch show a lower agreement than their French counterpart. This aligns with the accuracy analysis of Google Translate reported by Aiken (2019).

Comparing Baseline BM25 and the prompt on the original text, the agreement is substantial for French texts while only moderate for Dutch texts. This might be explained by differences between the Dutch and French lexicons, which were constructed independently. This gap disappears for the "NL2EN Lexicon" and "FR2EN Lexicon" that apply the same translated lexicon on Dutch and French texts, indicating that the translation is effective in uniformizing results across languages. The "FR2EN Lexicon" yields the highest agreement with the prompt methods, even higher than the Baseline BM25 on French texts.

While Cohen's kappa is useful to evaluate the robustness of the detection of relevant news across methods and languages, it does not explain the source of the disagreement or whether one method is performing better than others.

To understand this disagreement between Baseline BM25 and Prompting, we report in Table 2 the contingency matrix between the Baseline BM25 classification (in row) and the Prompting results (in column). The numbers on the diagonal of Table 2 show the cases of agreement between the BM25 and Prompting methods. There is $\frac{1270+295}{1881} = 83.2\%$ of agreement, which is substantially higher than the hypothetical probability of random agreement, $\frac{297}{1881} \times \frac{609}{1881} + \frac{1584}{1881} \times \frac{1272}{1881} = 62.06\%$, confirming a positive dependence between the two baseline methods. The corresponding Cohen's kappa is 0.56.

The elements outside of the diagonal show that the dependence is not perfect. Consider first the case of disagreement, where the BM25 approach identifies an article as AML-relevant while the prompting method does not. Such outcomes are rare: only two articles fall into this category. A manual review reveals that both articles discuss a sudden drop in gold prices and highlight a market opportunity to invest in gold. Given that gold is frequently used in money laundering schemes due to its portability, high value, and ease of conversion to cash, several keywords from the BM25 lexicons appeared in these articles. Consequently, the BM25 method erroneously classified them as AML-relevant. While such misclassifications could theoretically be mitigated by incorporating a classifier into the BM25 framework, the low frequency of these cases makes the additional effort unjustifiable.

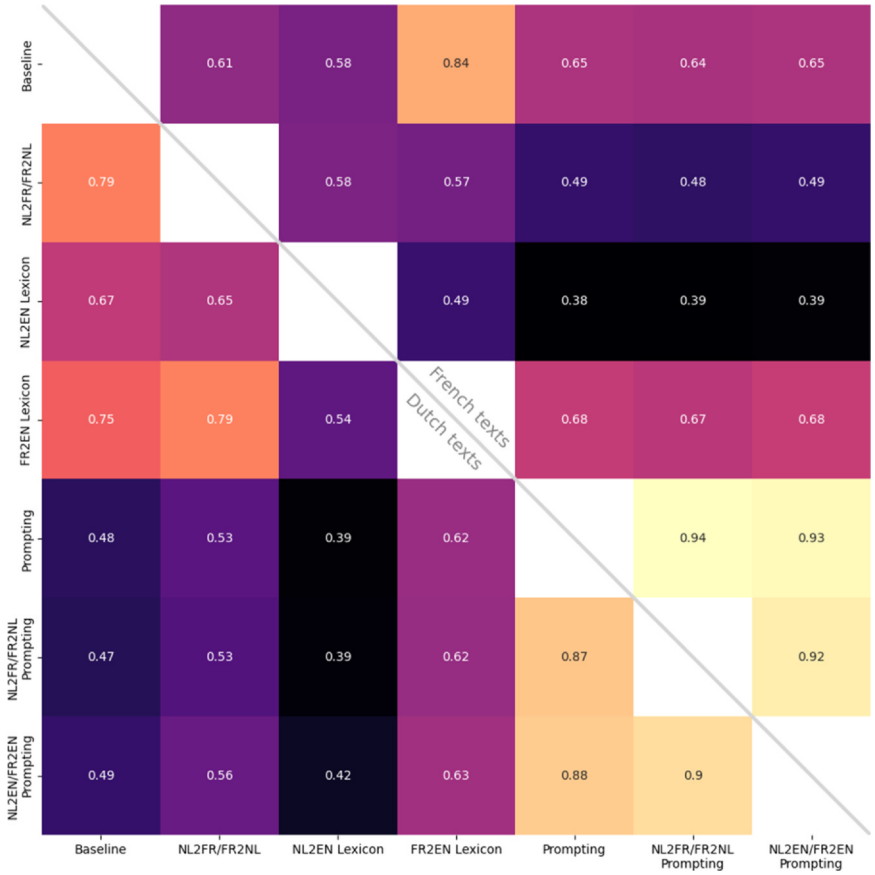

Fig. 7. Cohen's kappa heatmap of the seven binary classification methods. This heatmap illustrates the (dis)agreement between the approaches. The heatmap is subdivided into two parts, corresponding to the analysis of the Dutch (lower triangle) and French (upper triangle) subsets of newspaper articles.

Table 2
Contingency table between the AML news article alerts by the BM25 baseline, where articles are considered as being relevant if their score is over the 75 threshold, and the prompting of articles in their original language.

		Prompt		**Total**
		Not AML relevant	AML relevant	
BM25	Not AML relevant	1270	314	1584
	AML relevant	2	295	297
Total		1272	609	

The majority of disagreements in the contingency table (314 out of 1881 total articles) arise from articles identified exclusively by the prompting method. To examine this discrepancy, we sampled cases of disagreement and identified several instances of over-detection by the prompting method:

- Articles about scrutiny of financial institutions for reasons that are not related to AML.
- Articles in which AML is not the main topic. It is only mentioned anecdotally in a few sentences with no new information. Examples are articles covering a stock market analysis, a robbery, or cyberattack.
- Articles about harm to customers of financial institutions that are not related to AML
- Articles in which the AML keywords are not present but extrapolated to be present by the prompt, while mechanically the BM25 score is zero.

BM25 does not tend to detect these articles as relevant for AML because of the low frequency of AML relevant words. The prompt does not instruct to take into account the relative frequency of the AML content in the article. This may explain why it tends to detect more articles in which AML is not the main topic. We find indeed that lowering the threshold for BM25 results in more consistency with the results from prompting. This is at the expense of a higher number of alerts. A manual check indicates that the majority of them are false positives. We therefore conclude to not change the threshold of 75 as we overall estimate that the majority comes from

an overdetection by the prompt. Future work may look at improvement of the prompt and the BM25 or possibly use a combination of both with a more extensive test set.

7. Conclusion and future work

A well-functioning financial system requires financial institutions to implement a reliable AML/CFT control framework to monitor both customers and transactions. The exceptional occurrence of fraudulent transactions makes their detection difficult and is often compared to a blindfolded quest for a needle in a haystack. Fortunately, regulators can among others rely on investigative journalists seeking actively for those needles in a haystack. A prime example of this are the news article publications following the analysis of data leaks by the International Consortium of Investigative Journalists (ICIJ). The key practical question is then: how to integrate these articles in the workflow of regulators through the automatic detection of relevant news articles published in the media?

In this paper, we propose a data-driven system to compute early warning signals indicating per financial institutions that needs to be supervised whether relevant articles have been published. The proposed solution classifies a news article as relevant when it discusses for a non-negligible extent a relevant topic (thematic relevance) and when it relates to a financial institution that needs to be monitored (entity relevance). The obtained relevance score per article is then aggregated into a weekly early warning risk signal. We show how these signals fit in the workflow of AML supervision.

The proposed News Event Monitoring (NEMO) process is a modular and scalable system. We apply it to the supervision of financial institutions using Belgian newspaper and magazine articles as input data. The event analysis shows that, in terms of media coverage, there are three highly relevant leaks published by the ICIJ for the Belgian AML supervision, namely the Offshore Leaks (April 2013), Panama Papers (April 2016) and FinCEN Files (September 2020).

In future work, we aim to refine the post-processing stage by leveraging the prompting capabilities of Large Language Models (LLMs) to reduce false positives. We will also study the use of the proposed scores as features in a classification model and evaluate the incremental value compared to features obtained using other approaches such as the neural-network-based technique BERT. We also anticipate that advancements in coreference resolution will further enhance the precision of filtering relevant articles. Additionally, our research will focus on deriving AML performance scores for financial institutions by balancing positive and negative news coverage. This will involve examining the incremental informational value of news monitoring signals and time series relative to qualitative and quantitative indicators obtained from annual surveys. An exciting challenge lies in exploring the application of these insights during fit-and-proper assessments, which are conducted when new board members of financial institutions are nominated. Addressing this will require robust person name disambiguation techniques to effectively manage cases of homonymy.

CRediT authorship contribution statement

Ringoot Pascal: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Boudt Kris:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Delmarcelle Olivier:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Research Foundation Flanders (#G014420N, #W001021N) and the COST action (#CA21163) on Text, functional and other high-dimensional data in econometrics.

Appendix A. The BM25 algorithm to compute relevancy

Suppose we have a corpus of news articles, A, including $|A|$ news articles, and a query Q consisting of keywords $q, \ldots, q_{|Q|}$. For any given news article $a \in A$, the BM25 score aggregates the relevancy score of all query keywords for that article:

$$BM25(Q, a, A) = \sum_{i=1}^{|Q|} IDF(q_i, A)*TF(q_i, a, A), \tag{A.1}$$

where the relevancy of each keyword is the product between the *inverse document frequency* (penalizing words that appear frequently in the corpus) and the *term frequency* of that keyword.

The inverse document frequency is defined as follows:

$$IDF(q_i, A) = \log\left(\frac{|A| - df(q_i, A) + 0.5}{df(q_i, A) + 0.5} + 1\right), \tag{A.2}$$

where $df(q_i, A)$ is the number of news article containing the keyword q_i. The term frequency equals:

$$TF(q_i, a, A) = \frac{tf(q_i, a)}{tf(q_i, a) + k_1(1 - b + b\frac{dl_a}{avdl_A})}, \tag{A.3}$$

where dl_a is the document length (number of words) of news article a, $avdl_A$ the average document length in corpus A, $tf(q_i, a)$ the count of occurrences of keyword q_i in a, k_1 and b are constants. The factor $(1 - b + b\frac{dl_a}{avdl_A})$ is a normalization of the document length designed such that documents longer than the average will see their scores reduced. The parameter b controls the strength of the normalization and is generally set to 0.75. The function becomes saturated as $tf(q_i, a)$ increases. Hence, multiple occurrences of a single keyword contribute marginally less to the overall score. The parameter k_1 controls the strength of the saturation and is usually set to 1.2. Because of these normalizations, BM25 tends to favour shorter documents containing multiple keywords of the query.

Appendix B. System and user prompts used to detect the relevant news for the anti-money laundering supervision of financial institutions

System prompt

You are a news monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions. You are tasked with determining if a news article is relevant to a given theme. You are given step-by-step evaluation criteria and a news article. You always follow these instructions, evaluate critically and explain your reasoning.

User prompt

The news article provided to you relates to an event involving the < analyzed company > identified by the name "${name}". Your goal is to assess whether the part of the news article related to the < analyzed company > is HIGHLY relevant to the COMPLIANCE theme, using the criteria outlined below.

Steps:

1. **Understand the Context:** Carefully read the news article (< news_article_text >) to grasp the full context and nature of the event being discussed. Be wary of the causality of events described in the news.
2. **Evaluate the relevance to the Compliance theme** Assess whether the segment of the news article related to the < analyzed company > is HIGHLY relevant to the Compliance theme.
   i. **Context of Compliance:** Compliance is defined from the perspective of the regulator of financial institutions, such as banks, insurance companies, and investment firms. Consider general regulatory priorities, such as preventing misconduct, ensuring transparency, protecting consumers, maintaining ethical practices, and safeguarding financial system integrity.
   ii. **Evaluation Criteria:**
      - Does the text suggest or imply any compliance risks or regulatory concerns?
      - Are there references to laws, regulations, policies, or practices that might attract regulatory scrutiny?
      - Does the text indicate deviation from commonly accepted compliance standards or principles?
      - Are there elements in the text that suggest gaps in risk management, ethical practices, or operational controls?
      - Are there any mentions of regulatory investigations, fines, or penalties?
      - Does the text highlight any compliance failures, breaches, or violations?
      - Is there any indication of non-compliance with anti-money laundering (AML) or know your customer (KYC) regulations?
      - Does the text suggest any involvement in fraud, corruption, bribery, or other financial crimes?
      - Does the text mention any regulatory enforcement actions, or compliance challenges faced by the company?
      - Are there any indications of conflicts of interest, insider trading, market manipulation, or other unethical behavior?
      - Does the text reveal any compliance-related controversies, scandals, or reputational risks?

Article Details:

- **Company Name**: ${name}
- **News Article**: ${article_text}

Output Format (JSON):

Write your answer in valid JSON format according to the schema below:

```
{
"Context": "Your reasoning about the overall event and the company's presence.",
"Compliance relevance": {
"Explanation": "Explain the relevance of the news segment to the Compliance theme.",
"Value": true | false
}
}
```

References

Aiken, M. (2019). An updated evaluation of Google translate accuracy. *Studies in Linguistics and Literature, 3*(3), 253–260.

Algaba, A., Borms, S., Boudt, K., & Verbeken, B. (2023). Daily news sentiment and monthly surveys: A mixed-frequency dynamic factor model for nowcasting consumer confidence. *International Journal of Forecasting, 39*, 266–278.

Antonakis, J. (2017). On doing better science: From thrill of discovery policy implications. *The Leadership Quarterly, 28*, 5–21.

Ashwin, J., Kalamara, E., & Saiz, L. (2024). Nowcasting euro area GDP with news sentiment: a tale of two crises. *Journal of Applied Econometrics, 39*, 887–905.

Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems, 150*, Article 113492.

Barbaglia, L., Consoli, S., & Manzan, S. (2024). Forecasting GDP in Europe with textual data. *Journal of Applied Econometrics, 39*, 338–355.

Barrière, C. (2016). *Entities, Labels, and Surface Forms. In: Natural Language Understanding in a Semantic Web Context.* Cham: Springer.

Berglez, P., & Gearing, A. (2018). The Panama and Paradise Papers. The rise of a global fourth estate. *International Journal of Communication, 12*, 4573–4592.
Borms, S., Boudt, K., Van Holle, F., & Willems, J. (2021). *Semi-supervised text mining for monitoring the news about the ESG performance of companies. In Data science for economics and finance*. Cham: Springer217–239.
Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37–46.
Constant, M., & Nivre, J. (2016). *A Transition-Based System for Joint Lexical and Syntactic Analysis. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *1*, Berlin, Germany: Association for Computational Linguistics161–171.
Díaz-Struck, E., & Cabra, M. (2018). Uncovering international stories with data and collaboration. *Digital Investigative Journalism: Data, Visual Analytics and Innovative Methodologies in International Reporting,* 55–65.
EBA 2019, Opinion of the European Banking Authority on communications to supervised entities regarding money laundering and terrorist financing risks in prudential supervision (EBA-Op-2019-08).
Jijkoun, V., Khalid, M., Marx, M., & De Rijke, M. (2008). Named entity normalization in user generated content. *Information In Proceedings of the Second Workshop on Analytics for noisy unstructured text data,* 23–30.
Kanoulas, E., Pavlu, V., Dai, K., & Aslam, J. (2009). *Modeling the score distributions of relevant and non-relevant documents. In Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings*, *2*, Springer Berlin Heidelberg152–163.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Nguyen, H., & Cao, T. (2012). Named entity disambiguation: A hybrid approach. *International Journal of Computational Intelligence Systems, 5*, 1052–1067.
Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval, 3*, 333–389.
Straka, M. & Strakova, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. 88-99.
Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and Coreference Resolution: A Review. *Information Fusion, 59*, 139–162.
Thorsrud, L. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business Economic Statistics, 38*, 393–409.
Zhang, L., & Liu, B. (2012). Sentiment Analysis and Opinion Mining. Encyclopedia of Machine Learning and Data Mining.