Quaderni FinTech

Sperimentazione dell'IA nell'azione di contrasto della CONSOB agli abusivismi finanziari

M. Trerotola, A. Boghi, P. Deriu, G. Frega

19

luglio 2026

CONSOB
COMMISSIONE NAZIONALE
PER LE SOCIETÀ E LA BORSA

Nella collana dei Quaderni ***FinTech***

sono raccolti lavori di ricerca relativi

al fenomeno «FinTech» nei suoi molteplici aspetti

al fine di promuovere la riflessione e

stimolare il dibattito su temi attinenti

all'economia e alla regolamentazione

del sistema finanziario.

CONSOB

00198 Roma – Via G.B. Martini, 3
t +39.06.84771 centralino
f +39.06.8477612

20121 Milano - Via Broletto, 7
t +39.02.724201 centralino
f +39.02.89010696

h www.consob.it
e studi_analisi@consob.it

Sperimentazione dell'IA nell'azione di contrasto della CONSOB agli abusivismi finanziari

*M. Trerotola, A. Boghi, P. Deriu, G. Frega**

Sintesi del lavoro

L'Intelligenza Artificiale (IA) si configura come una delle più rilevanti innovazioni tecnologiche contemporanee, con un impatto significativo e trasversale su tutti i settori, incluso quello finanziario. Questo studio esplora l'impiego dell'IA nell'azione di contrasto ai fenomeni di abusivismo finanziario tramite l'utilizzo di un *software* sviluppato dalla CONSOB in collaborazione con il Politecnico di Torino e l'Università Politecnica delle Marche. Il lavoro si concentra sull'utilizzo di tecnologie IA specifiche, quali i *Large Language Model* (LLM) e il *Natural Language Processing* (NLP), evidenziando il loro ruolo nell'analisi, estrazione e trattamento delle informazioni contenute negli esposti presentati dai risparmiatori, così come nell'analisi dei siti *web* dei soggetti segnalati.

Il seguente articolo descrive lo sviluppo del prototipo in questione che integra tecniche basate su intelligenza artificiale ibrida, con un *focus* particolare sull'accuratezza e la robustezza dell'estrazione di informazioni da contenuti a disposizione.

JEL Classifications: O32, O33, C55, G21.

Parole chiave: intelligenza artificiale, abusivismo finanziario, software per il trading, piattaforme di trading, CONSOB's effort aginst fraud.

(*) Mario Trerotola - PhD in Intelligenza Artificiale;
Alessio Boghi - CONSOB, Divisione Ispettorato;
Paola Deriu - CONSOB, Responsabile della Divisione Studi e Regolamentazione;
Giuseppe Frega - Responsabile Ufficio Vigilanza su Fenomeni Abusivi, Divisione Ispettorato.

Si ringraziano la Prof.ssa Caterina Lucarelli dell'Università Politecnica delle Marche, Giampiero Longobardi (CONSOB, Vice Direttore Generale), Nicola De Caro e Brizio Leonardo Tommasi (Divisione Informatica e Intelligenza Artificiale) per gli utili commenti. Errori e imprecisioni sono imputabili esclusivamente agli autori. Le opinioni espresse nel lavoro sono attribuibili esclusivamente agli autori e non impegnano in alcun modo la responsabilità dell'Istituto. Nel citare il presente lavoro, non è, pertanto, corretto attribuire le argomentazioni ivi espresse alla CONSOB o ai suoi Vertici.

La valutazione dei risultati cui il sistema perviene è stata condotta attraverso il confronto tra i risultati ottenuti dal sistema e quelli derivanti dall'analisi di un operatore, con particolare attenzione all'accuratezza, coerenza e consistenza delle risposte generate. In particolare, i risultati evidenziano il potenziale dell'IA nell'automazione di processi complessi di vigilanza finanziaria, pur riconoscendone le attuali limitazioni. Nell'articolo, viene, infatti, fornita la descrizione delle limitazioni intrinseche delle attuali tecnologie IA, come la gestione della complessità semantica e la spiegabilità degli LLM, evidenziando i principali spunti di ottimizzazione per miglioramenti futuri.

Muovendo dal prototipo esaminato, il lavoro sviluppa, infine, una riflessione più ampia sulle implicazioni tecnico-organizzative e sul quadro giuridico di riferimento per l'impiego dell'IA nei processi di vigilanza, con particolare attenzione ai profili di supervisione umana, tracciabilità, *accountability* e compatibilità con l'AI Act e con la disciplina nazionale.

Con specifico riferimento al tema della supervisione umana, è bene evidenziare che il sistema sperimentato non assume in alcun caso decisioni in via autonoma: l'intervento del *software* resta circoscritto alle fasi preliminari e preparatorie, secondo una logica di *human in the loop* nella quale la valutazione e la decisione finale, e la relativa responsabilità, restano in capo all'operatore. Il sistema prevede, infatti, che attraverso un sistema di *feedback* l'operatore possa valutare ed eventualmente modificare gli esiti cui è pervenuto il *software* sia in relazione al risultato finale sia relativamente all'estrazione di specifiche informazioni acquisite durante le differenti fasi di analisi dell'esame degli esposti e dei siti *web*.

Experimenting with AI in CONSOB's Efforts to Combat Financial Fraud

*M. Trerotola, A. Boghi, P. Deriu, G. Frega**

Abstract

Artificial Intelligence (AI) is one of the most significant contemporary technological innovations, with a substantial and far-reaching impact across all sectors, including the financial sector. This study explores the use of AI in combating financial fraud through the use of software developed by CONSOB in collaboration with the Polytechnic University of Turin and the Polytechnic University of the Marche. The study focuses on the use of specific AI technologies, such as Large Language Models (LLMs) and Natural Language Processing (NLP), highlighting their role in analyzing, extracting, and processing the information contained in complaints filed by investors, as well as in analyzing the websites of the entities reported.

The following article describes the development of the prototype in question, which integrates techniques based on hybrid artificial intelligence, with a particular focus on the accuracy and robustness of information extraction from available content.

The evaluation of the system's results was conducted by comparing the system's outputs with those derived from an analyst's review, with particular attention to the accuracy, coherence, and consistency of the generated responses. In particular, the results highlight the potential of AI in automating complex financial supervision

(*) Mario Trerotola - PhD in Artificial Intelligence;
Alessio Boghi - CONSOB, Inspections Division;
Paola Deriu - CONSOB, Head of Research and Regulation Division:
Giuseppe Frega - CONSOB, Head of Unauthorised Financial Business Office, Inspections Division.

The authors would like to thank Caterina Lucarelli Polytechnic University of Marche, Giampiero Longobardi (CONSOB, Deputy General Manager), Nicola De Caro and Brizio Leonardo Tommasi (IT and Artificial Intelligence Division) for useful comments to an earlier version of the paper. Of course, the authors are the only responsible for errors and imprecisions. The opinions expressed here are those of the authors and do not necessarily reflect those of CONSOB.

processes, while acknowledging its current limitations. In fact, the article describes the inherent limitations of current AI technologies, such as handling semantic complexity and the explainability of LLMs, highlighting key areas for optimization to facilitate future improvements.

Building on the prototype examined, the study ultimately offers a broader reflection on the technical and organizational implications and the legal framework governing the use of AI in supervisory processes, with particular attention to human oversight, traceability, accountability, and compliance with the AI Act and national regulations.

With specific reference to the issue of human supervision, it is important to note that the tested system does not, under any circumstances, make decisions autonomously: the software's intervention is limited to the preliminary and preparatory phases, following a human-in-the-loop approach in which the evaluation, the final decision, and the associated responsibility remain with the operator. In fact, the system provides that, through a feedback mechanism, the operator can evaluate and, if necessary, modify the results produced by the software, both in terms of the final outcome and with regard to the extraction of specific information gathered during the various stages of analyzing the complaints and websites.

Indice

Introduzione

PARTE PRIMA

1 Evoluzione dell'intelligenza artificiale: brevi cenni

L'intelligenza artificiale (IA) può essere definita come «*la capacità di un sistema di interpretare correttamente dati esterni, di apprendere da tali dati e di utilizzare gli apprendimenti per raggiungere obiettivi e compiti specifici attraverso un adattamento flessibile*»[1]. Grazie all'impiego di algoritmi di apprendimento automatico, il sistema è in grado, dunque, di inferire regole e modelli dai dati, adattarsi a nuove situazioni e generalizzare la propria operatività oltre le istruzioni predefinite. Tale capacità ha trovato applicazioni in molteplici settori, tra cui quello finanziario, dove viene impiegata per analisi predittive, gestione del rischio, trading algoritmico, ottimizzazione delle strategie di investimento e in vigilanza.

L'evoluzione dell'IA ha visto diverse fasi a partire dai primi algoritmi simbolici sviluppati negli anni '50 e '60 che cercavano di modellare il ragionamento umano attraverso regole logiche e sistemi di inferenza. Questi primi approcci erano limitati dalla capacità computazionale e dalla mancanza di dati sufficienti per creare modelli complessi. Negli anni '80 e '90 l'aumento della potenza di calcolo e la crescente disponibilità di dati, hanno favorito l'affermazione del *machine learning* (ML), una branca dell'intelligenza artificiale che si concentra sull'uso di algoritmi capaci di apprendere dai dati e di migliorare progressivamente le loro prestazioni. Negli anni successivi, il *deep learning*, evoluzione del *machine learning*, basata su reti neurali profonde ispirate al funzionamento dei neuroni biologici, ha poi ulteriormente rivoluzionato il campo dell'IA. Grazie alla capacità di apprendere e modellare relazioni complesse e non lineari nei dati, ha reso possibili progressi straordinari in ambiti come il riconoscimento delle immagini, la traduzione automatica e la comprensione del linguaggio naturale, aprendo nuove prospettive applicative e ampliando costantemente i confini dell'intelligenza artificiale.

Negli ultimi anni, l'IA ha compiuto ulteriori progressi con lo sviluppo di Modelli di Linguaggio di Grandi Dimensioni (LLM), quali quelli della famiglia

1 Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62* (1), 15–25.

GPT-5 di OpenAI e Gemini di Google. Basati sull'architettura Transformer[2] e su paradigmi di *pre-training* bidirezionale[3], e caratterizzati da capacità di generalizzazione *few-shot*[4], questi hanno trasformato il modo in cui le macchine comprendono e generano il linguaggio, grazie alla capacità di cogliere relazioni complesse e sfumature nei testi. Una loro evoluzione sono i modelli multimodali, che integrano diverse tipologie di dati – come testo, immagini e audio – offrendo una comprensione più ricca e articolata delle informazioni e aprendo la strada a nuove applicazioni in campi molto diversi tra loro.

Gli LLM operano elaborando istruzioni fornite sotto forma di *prompt*, che guidano il modello nella produzione di testi o altri contenuti. Attraverso il processo di inferenza, essi generano risposte coerenti con la richiesta, sfruttando le conoscenze acquisite durante l'addestramento e adattandosi a diversi formati, come spiegazioni, sintesi o traduzioni. Inoltre, possono essere integrati con altri componenti *software*, assumendo così il ruolo di veri e propri agenti capaci di interagire con l'ambiente, eseguire compiti complessi e supportare processi decisionali.

Anche nel settore finanziario, l'intelligenza artificiale sta assumendo un ruolo sempre più centrale, favorendo una profonda trasformazione attraverso applicazioni innovative nelle tecnologie finanziarie (fintech). L'IA riveste una funzione determinante in diverse aree del settore finanziario, tra cui l'analisi del rischio di credito[5], la previsione dei prezzi degli *asset*[6], l'antirici-

2 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Kaiser, Ł., Polosukhin, I. (2017), Attention is All You Need, Advances in Neural Information Processing Systems (NeurIPS), 30, 5998-6008. L'architettura *Transformer* è il modello di rete neurale alla base della maggior parte degli LLM attuali. La sua innovazione principale è il meccanismo di attenzione (*attention*), che consente al modello di valutare, per ogni parola, quali altre parole del testo siano più rilevanti per coglierne il significato, considerando l'intera sequenza nel suo insieme anziché elaborarla rigidamente parola dopo parola. Ciò permette di cogliere in modo efficiente le relazioni anche tra termini distanti nel testo.

3 Devlin, J., Chang, MW., Lee, K., Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT 2019, 4171-4186. Il *pre-training* bidirezionale consiste in una fase di addestramento preliminare su grandi quantità di testo non etichettato, durante la quale il modello impara a ricostruire parole mancanti tenendo conto sia del contesto che le precede sia, contemporaneamente, di quello che le segue (da cui "bidirezionale"). Rispetto agli approcci che leggono il testo in una sola direzione, questo consente una comprensione più ricca del significato delle parole in funzione del contesto. Il modello così pre-addestrato può poi essere specializzato (*fine-tuning*) su compiti specifici.

4 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, JD., Dhariwal, P., *et al.* (2020), Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-1901. La capacità di generalizzazione *few-shot* indica l'abilità di un LLM di svolgere un compito nuovo a partire da pochi esempi forniti direttamente nelle istruzioni (il *prompt*), senza necessità di un nuovo addestramento. Quando il modello opera sulla sola descrizione del compito, senza esempi, si parla di *zero-shot*. È questa proprietà a rendere gli LLM flessibili e adattabili a compiti diversi senza interventi tecnici dedicati.

5 Studi recenti hanno evidenziato l'applicazione di modelli AI per l'estrazione automatizzata di informazioni rilevanti da documenti finanziari, come i *report* annuali, al fine di migliorare i processi di valutazione del rischio e le decisioni relative alla concessione del credito [*Cfr.* Acheampong, A. and Elshandidy, T., 2021. Does soft information determine credit risk? Text-based evidence from European banks. *Journal of international financial markets, institutions and money*, *75*, p.101303].

6 Modelli di intelligenza artificiale, quali Random Forest e Reti Neurali Profonde, hanno trovato ampio utilizzo in letteratura, dimostrando una maggiore accuratezza nel prevedere le dinamiche di prezzo, in particolare per quanto riguarda l'individuazione di *pattern* non lineari nei mercati [*Cfr.* Parente, M., Rizzuti, L. and Trerotola, M., 2024. A

claggio[7] e la finanza comportamentale[8]. Indicazioni utili sullo stato dell'adozione dell'intelligenza artificiale nel settore finanziario italiano si ricavano dal *report* OECD (2026), *Artificial Intelligence in Italian Financial Markets*, che segnala la progressiva diffusione di tali tecnologie e sottolinea l'esigenza di accompagnarne lo sviluppo con appropriati presidi di *governance*, qualità del dato, *accountability* e supervisione[9].

L'introduzione degli LLM nel settore finanziario ha ampliato il ventaglio dei compiti di analisi dei contenuti economico-finanziari automatizzabili, dalla classificazione e dall'analisi del *sentiment* al riconoscimento di entità, dalla sintesi documentale alla risposta automatica a domande[10]. Come evidenziato, tali modelli mostrano una particolare efficacia nell'estrazione di informazioni da dati non strutturati e complessi, quali documenti testuali e immagini, configurandosi come un sostanziale avanzamento rispetto alle tradizionali tecniche di *text mining*[11].

In tal senso, risulta di particolare interesse, per un'Autorità di vigilanza, l'applicazione dell'intelligenza artificiale per l'elaborazione di documenti, che veicolano informazioni eterogenee per formato, struttura e grado di affidabilità. L'impiego di algoritmi basati su IA consente non solo di estrarre informazioni utili per l'attività di vigilanza, ma anche di individuare schemi ricorrenti e correlazioni latenti all'interno dei dati. In particolare, gli LLM possono essere impiegati per attività quali la sintesi automatica dei contenuti documentali e il riconoscimento di entità nominate (ad esempio aziende, individui o località), migliorando significativamente le capacità di analisi e supportando processi decisionali fondati su evidenze testuali complesse.

Tale approccio favorisce l'automazione dei processi di vigilanza che si caratterizzano per l'elevato livello di standardizzazione. In concreto, l'estrazione delle informazioni presenti negli esposti e nei siti *web* oggetto di indagine – e la connessa valutazione, tramite il *software,* finalizzata a verificare l'eventuale sussistenza dei profili di competenza – consente di velocizzare i

profitable trading algorithm for cryptocurrencies using a Neural Network model. *Expert Systems with Applications*, *238*, p.121806].

7 Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G. and Lorentzen, J., 2020. Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, *23* (1), pp.173-186.

8 Le tecniche basate su AI supportano una comprensione più approfondita dell'influenza dei comportamenti e delle emozioni degli investitori sulle dinamiche di mercato, migliorando la capacità di previsione delle reazioni a eventi economici e politici [*Cfr.* Ahmed, S., Alshater, M.M., El Ammari, A. and Hammami, H., 2022. Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, *61*, p.101646].

9 OECD (2026), Artificial Intelligence in Italian Financial Markets, OECD Publishing, Paris, https://doi.org/10.1787/6f42c977-en.

10 Li, Y., Wang, S., Ding, H. and Chen, H., 2023, November. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance* (pp. 374-382).

11 Li, J., Sun, A., Han, J., Li, C. (2022), A Survey on Deep Learning for Named Entity Recognition, IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70, https://doi.org/10.1109/TKDE.2020.2981314

tempi istruttori e, conseguentemente, processare un maggior numero di segnalazioni.

2 L'intelligenza artificiale in CONSOB

L'intelligenza artificiale – elemento di rilievo nella strategia CONSOB sin dal 2022 – assume carattere primario nel Piano strategico 2025-2027, nel quale la transizione digitale è assunta come fattore trasversale del cambiamento organizzativo e dell'evoluzione dell'azione di vigilanza. In tale quadro, l'innovazione tecnologica è considerata funzionale sia al rafforzamento dell'efficacia e dell'efficienza dell'attività dell'Istituto, sia al potenziamento della tutela degli investitori, anche attraverso il presidio dei rischi derivanti dall'evoluzione tecnologica e l'impiego di strumenti avanzati a supporto dei processi di analisi e controllo. In continuità con il precedente ciclo di pianificazione, ma in una prospettiva più ampia di consolidamento del cambiamento organizzativo e operativo, il Piano valorizza dunque l'uso di soluzioni digitali e di approcci *data-driven* come leva per una vigilanza più tempestiva, proporzionata e coerente con la crescente complessità dei mercati finanziari.

In tale contesto, sono state avviate diverse sperimentazioni che mirano a rafforzare l'approccio di vigilanza *data-driven* e *risk-based*, con particolare attenzione all'uso di sistemi di IA a supporto delle analisi preliminari e alla loro compatibilità con il quadro normativo vigente[12].

A tali iniziative si affiancano le attività tecnico-progettuali volte alla progressiva industrializzazione, progettazione e integrazione delle componenti applicative *software*, incluse quelle dedicate alla acquisizione, normalizzazione e selezione dei contenuti rilevanti per i casi d'uso sottoposti ad analisi tramite IA. In tale ambito, si valorizzano le diverse esperienze maturate in materia di IA, che comprendono, tra l'altro, lo sviluppo di modelli di classificazione e analisi semantica dei contenuti, l'utilizzo di tecniche di *machine learning* per il supporto alle attività di *screening* e rating dei casi, l'integrazione di soluzioni di *Natural Language Processing* per l'estrazione automatica di informazioni da fonti eterogenee, nonché l'evoluzione di *pipeline* di gestione del dato e di soluzioni tecnologiche orientate a garantire scalabilità, tracciabilità e conformità ai requisiti regolamentari.

La vigilanza sull'*insider trading* è una delle aree in cui la CONSOB ha concentrato i suoi sforzi nell'applicazione delle tecniche di IA. La CONSOB, in

12 Deriu P., Racioppi S., *Riflessioni in tema di intelligenza artificiale e attività di vigilanza*, con presentazione a cura di Lalli A., Quaderni FinTech CONSOB, n. 15, agosto 2025

collaborazione con la Scuola Normale Superiore di Pisa[13], ha sviluppato un prototipo sperimentale a supporto delle analisi preliminari nell'ambito della *detection* di condotte sospette di costituire abusi di mercato. In particolare sono stati sviluppati tre *proof of concept*: il primo, basato su *clustering k-means* dinamico, consente di individuare comportamenti anomali dei singoli investitori; il secondo, fondato sulle *Statistically Validated Networks* (SVN), è volto a rilevare condotte coordinate di gruppi di operatori; il terzo adotta tecniche di riduzione dimensionale per isolare anomalie contestuali nell'operatività di mercato in prossimità di eventi *price sensitive*, offrendo un supporto preliminare all'identificazione di possibili abusi di informazione privilegiata.

L'esperienza ha evidenziato come i diversi approcci siano complementari, rafforzando l'efficacia complessiva del modello vigilanza sul quale si basano le analisi preliminari, a supporto della decisione di avviare un'istruttoria.

Per quanto riguarda l'automazione del monitoraggio dei KID (*Key Information Documents*) relativi ai PRIIPs (*Packaged Retail and Insurance-based Investment Products*), la CONSOB, in collaborazione con l'Università La Sapienza[14], ha adottato tecniche di intelligenza artificiale per automatizzarne l'analisi. L'approccio combina metodologie di *information extraction* basate sia su *pattern* che su modelli di apprendimento, con l'obiettivo di costruire automaticamente un *dataset* strutturato, utilizzabile tanto da sistemi automatizzati quanto da operatori umani. Ciò consente di ottenere significativi vantaggi analitici nella prioritizzazione delle attività di vigilanza, grazie a uno *screening* automatico dei KID rapido ed efficace.

Infine, CONSOB, in collaborazione con l'Università di Trento[15], ha avviato una sperimentazione per la realizzazione di un sistema di supporto alla vigilanza sui *green bond* europei, basato sull'impiego di tecniche di intelligenza artificiale e di elaborazione del linguaggio naturale (NLP). Il prototipo analizza automaticamente la documentazione testuale associata alle emissioni, individuando le frasi rilevanti sotto il profilo ambientale, etichettandole in base al contenuto, attribuendo un giudizio di *sentiment* e verificando la presenza di riferimenti agli Obiettivi di Sviluppo Sostenibile (SDG). Il sistema misura, inoltre, lo scostamento tra gli SDG dichiarati formalmente dall'emittente e quelli effettivamente rilevabili nei documenti, considerato un potenziale indicatore di *greenwashing*. I risultati vengono restituiti in forma strutturata,

13 CONSOB – Scuola Normale Superiore di Pisa, *Dimensionality reduction techniques to support insider trading detection*, Quaderni FinTech CONSOB, n. 12, febbraio 2024.

14 Lembo D. *et al.*, "Information Extraction through AI techniques: The KIDs use case at CONSOB," *arXiv*, 2022, pp. 1–4. doi: 10.48550/arXiv.2202.01178

15 Paterlini S., Nicolodi A., Gentile M., Foglia Manzillo V., Sancilio M.R., Deriu P., *Greenwashing alert system for EU green bonds: The CONSOB–University of Trento prototype*, Quaderni FinTech CONSOB, n. 14, luglio 2025

consentendo un'analisi documentale scalabile e sistematica e generando "*alert*" su documenti e soggetti che presentano caratteristiche potenzialmente anomale.

Tra i vari *use case* è possibile rintracciare un filo comune: tutti condividono una logica di supporto conoscitivo all'attività di vigilanza, con centralità della validazione umana e dell'integrazione nei processi esistenti. Su questi aspetti si tornerà diffusamente nella Parte IV.

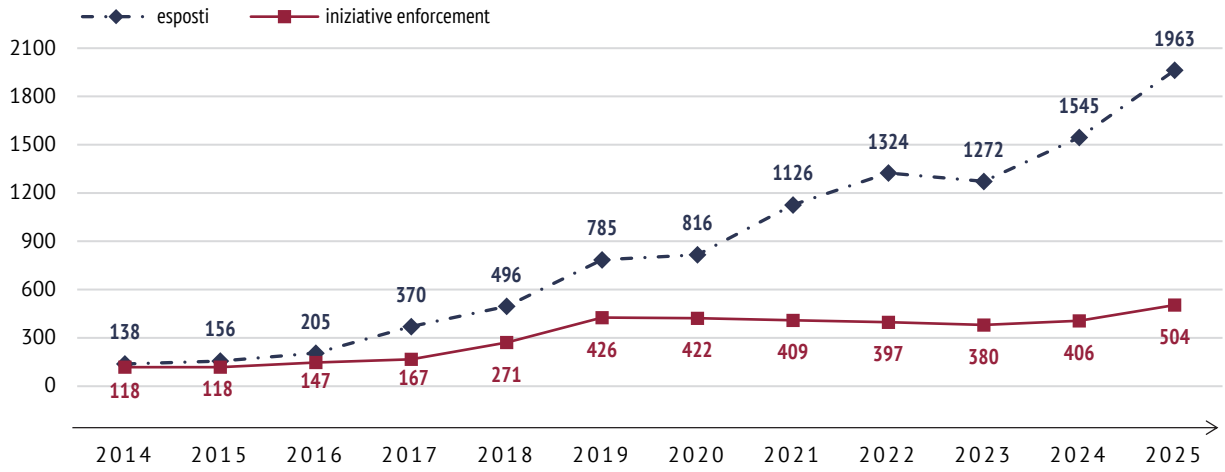
Attualmente la CONSOB sta lavorando alla possibile introduzione di nuovi strumenti di intelligenza artificiale per innovare altri processi di vigilanza, rendendoli più efficienti e meno onerosi. In particolare, la transizione in atto riguarda l'adozione di modelli di vigilanza basati sul rischio, volti a consentire di indirizzare risorse umane esperte verso situazioni particolarmente critiche e meritevoli di attenzione, nonché lo sviluppo di soluzioni di automazione per alcune fasi dell'attività di controllo. In questo quadro si colloca il prototipo descritto nel presente studio, sviluppato per rendere più efficiente l'azione di contrasto agli abusivismi finanziari attraverso un *software* che integra anche sistemi di intelligenza artificiale.

Il prototipo presentato costituisce, tuttavia, non solo un caso applicativo di impiego dell'IA a supporto della vigilanza, ma anche un'occasione per riflettere sui limiti tecnico-operativi di tali strumenti e sulle condizioni giuridiche e organizzative del loro utilizzo nei processi istruttori di un'Autorità di vigilanza.

1 Il contesto di riferimento

Negli ultimi anni il fenomeno degli abusivismi finanziari è evoluto significativamente, alimentato dall'uso crescente delle nuove tecnologie e di *internet*. Queste pratiche illecite compromettono la sicurezza degli investitori e l'integrità del mercato finanziario. A partire dal 2014 il fenomeno ha mostrato una continua espansione e con essa vi è stata una progressiva crescita del numero delle segnalazioni che vengono trasmesse alla CONSOB, indice di come sia sempre più elevata l'aspettativa del pubblico dei risparmiatori rispetto all'azione di contrasto ai casi di abusivismo, come indicato nella Figura 1.

Figura 1 – Evoluzione nel tempo del numero di segnalazioni/esposti dei risparmiatori per ipotesi di abusivismo e del relativo numero di iniziative di contrasto assunte dalla CONSOB



Fonte: CONSOB, menù dinamico e database dell'Ufficio Vigilanza Fenomeni Abusivi.

L'esigenza di protezione è ancora più avvertita quando si è in presenza di fenomeni di volta in volta nuovi utilizzati con finalità illecite. I fenomeni di abusivismo si caratterizzano, infatti, per la continua mutevolezza sia delle concrete modalità operative, in ciò favorite anche dall'innovazione tecnologica, sia degli schemi negoziali. Le piattaforme di *trading* e di *crowdfunding* non autorizzate e l'uso improprio della tecnologia a registro distribuito (DLT) rappresentano alcune delle modalità operative oggetto di indagine.

Si è, pertanto, cercato di realizzare un *software* che fosse, soprattutto, capace di adattarsi ai mutamenti esogeni, *in primis* derivanti da modifiche normative delle disposizioni delle discipline di settore. In tal senso, sono state create delle liste di parole sensibili, sia lato servizi sia lato prodotti, essenziali per rilevare la sussistenza di eventuali profili di competenza. Tali liste possono essere agevolmente integrate con nuovi termini e aggiornate in totale autonomia senza intervento degli sviluppatori e senza che ciò imponga una revisione delle operazioni di analisi dei testi (esposti e contenuti presenti nei siti *web*), né un ripensamento dell'intero approccio ogni volta che tali mutamenti si presentino. Ciò, rende possibile soddisfare, a risorse date, l'aumento della domanda (di protezione) da parte dei risparmiatori utilizzando la stessa configurazione di base apportando modifiche e perfezionamenti a fronte dei mutamenti causati da fattori esterni.

In sostanza, il *software* presenta un'elevata flessibilità e capacità di scalabilità in quanto *(i)* consente di processare un maggior numero di esposti attraverso l'automazione di attività seriali relative a processi che presentano un elevato livello di standardizzazione ed *(ii)* è capace di adattarsi ai mutamenti esterni senza la necessità di dover intervenire sulla configurazione di base.

L'attività abusiva posta in essere tramite il canale *internet* rappresenta una sfida crescente per l'Autorità di vigilanza. Per tale ragione è stato realizzato il prototipo di un sistema che consente di efficientare il processo di vigilanza in materia di contrasto agli abusivismi finanziari posti in essere tramite il canale *internet*, con specifico riferimento alle fattispecie di abusiva prestazione di servizi e attività di investimento e servizi su *crypto-asset*.

Nel dettaglio, il progetto si sviluppa su due direttrici:

a) l'analisi degli esposti trasmessi alla CONSOB, finalizzata all'individuazione di specifiche informazioni di interesse negli stessi contenute che poi confluiscono in un *report*. In particolare, tale attività mira a:
   - verificare la sussistenza di profili di competenza;
   - identificare eventuali URL di siti *internet* riconducibili ai soggetti segnalati;

b) l'analisi dei siti *web* dei presunti operatori abusivi, con l'estrazione di informazioni rilevanti che, anch'esse, confluiscono in un (secondo) *report* e vengono successivamente trasferite in modo automatico negli appositi documenti.

Proprio la combinazione di queste due direttrici rende necessario un impianto metodologico capace di integrare fonti eterogenee, tecniche diverse di estrazione e criteri di validazione coerenti con le esigenze dell'attività istruttoria.

In tale cornice si colloca anche il lavoro tecnico di industrializzazione sulla componente di *crawling*, finalizzato a rendere robusta l'acquisizione di contenuti specifici, di siti multi-pagina, di contenuti dinamici e pagine effettivamente rilevanti ai fini dell'analisi.

2 Fonti di dati per l'analisi di segnalazioni e attività *online*

Le segnalazioni dei risparmiatori, i siti *web* dei presunti operatori abusivi e i portali delle Autorità di vigilanza offrono informazioni specifiche e complementari che, se opportunamente integrate, permettono di svolgere analisi approfondite e di verificare la legittimità delle operazioni e delle attività oggetto di esame.

Il processo, unitariamente inteso, poggia su più fasi sequenziali e coniuga logiche finalizzate a estrarre informazioni dalle fonti esaminate (esposti e siti *web*) con logiche di controllo delle informazioni estratte tramite verifica su fonti esterne. L'elaborazione delle fonti, l'analisi dei dati e l'estrazione delle informazioni restituiscono all'utente un *output* che, al verificarsi di determinate condizioni, conduce all'avvio della fase successiva.

In sintesi, partendo dall'analisi dei fatti narrati nella segnalazione trasmessa dal risparmiatore, il *software* estrae le informazioni atte, *in primis,* a stabilire la sussistenza di eventuali profili di competenza, individua il sito *internet* da ispezionare e genera un *report*. L'eventuale rilevanza dell'esposto per i profili di competenza conduce alla seconda fase focalizzata sull'estrazione delle informazioni contenute nel sito *web* denunciato, a valle della quale, viene generato un secondo *report* nel quale confluiscono, tra le altre, informazioni in ordine al dichiarato possesso - all'interno del sito esaminato - di eventuali autorizzazioni rilasciate da Autorità di vigilanza omologhe alla CONSOB.

Il *software* consente, altresì, al funzionario istruttore di verificare l'eventuale presenza della società cui è riconducibile l'operatività oggetto di indagine negli elenchi dei soggetti autorizzati tenuti dalla CONSOB. Di seguito, verranno esaminate in dettaglio le caratteristiche e l'importanza di ciascuna di queste fonti, evidenziando il loro contributo all'efficacia delle analisi e delle valutazioni di conformità.

La natura eterogenea di queste fonti spiega anche la scelta di una *pipeline* ibrida, nella quale moduli diversi intervengono in momenti distinti dell'analisi per trattare contenuti che differiscono per struttura, formato e grado di affidabilità.

2.1 Esposti

L'esposto costituisce un documento formale tramite il quale un soggetto segnala alla CONSOB un comportamento scorretto, un'irregolarità o una presunta violazione di leggi o regolamenti. Gli esposti rappresentano una fonte primaria di informazioni, cruciali per l'analisi delle attività sospette o irregolari. Nella maggior parte dei casi, tali documenti forniscono dettagli specifici sui soggetti segnalati, comprese le modalità di contatto e una descrizione degli eventi o delle condotte oggetto della segnalazione. Gli esposti possono essere presentati attraverso una procedura guidata *online* o inviati tramite altri canali e redatti in forma libera offrendo flessibilità nella modalità di segnalazione. Nel caso in cui si usi la procedura guidata *online* gli esposti sono generati secondo un modello preformato. L'esponente può, inoltre, allegare documentazione a supporto, come contratti, comunicazioni via *e-mail*, che necessitano di elaborazione per estrarre tutte le informazioni complementari ai fini dell'analisi. La qualità e la completezza degli esposti sono fattori determinanti per il successo delle analisi successive, poiché forniscono un contesto ricco e dettagliato che può orientare efficacemente le indagini preliminari, guidando verso una valutazione più accurata e mirata dei fatti segnalati. Gli esposti costituiscono dunque una fonte ricca ma anche potenzialmente disomogenea, incompleta o "rumorosa".

2.2 Siti *web* operativi

I siti *web* dei soggetti segnalati costituiscono una fonte essenziale per acquisire informazioni sulle loro attività *online*. Essi contengono spesso una vasta gamma di contenuti utili a ottenere una comprensione approfondita delle operazioni dichiarate e delle modalità con cui queste vengono presentate al pubblico. I siti *web* costituiscono una fonte auto-rappresentativa preziosa dal punto di vista informativo, ma anche da verificare criticamente.

Un aspetto particolarmente rilevante riguarda l'estrazione di eventuali autorizzazioni dichiarate; la loro presenza consente di verificare la conformità legale del soggetto rispetto alla normativa vigente. L'analisi approfondita dei contenuti consente non solo di valutare i servizi proposti e le dichiarazioni contrattuali, ma anche di segnalare, nel caso di mancanza di riferimenti ad autorizzazioni, potenziali attività non autorizzate, offrendo un quadro complessivo più accurato e completo.

2.3 Portali delle Autorità di vigilanza

I portali delle Autorità di vigilanza rappresentano una risorsa autorevole e indispensabile per la verifica del soggetto a cui è riferibile il sito *internet* e per la validazione delle autorizzazioni dichiarate dagli stessi. Ad esempio, la CONSOB offre accesso a registri pubblici contenenti le società autorizzate a operare in Italia e le licenze che sono state loro concesse. Queste informazioni sono fondamentali per confermare la legittimità delle operazioni dei soggetti rispetto alle informazioni dichiarate sui loro siti *web*.

L'aspetto legato ai controlli, tramite accesso a fonti esterne, è di centrale importanza considerate le implicazioni che deriverebbero come conseguenza dell'adozione di eventuali provvedimenti non pertinenti. Per tale motivo, il sistema è basato su più livelli (macchina/essere umano) e ambiti di controllo (siti *internet* delle singole Autorità di vigilanza/sito *internet* della IOSCO[16]). In concreto, prima la macchina e poi l'essere umano verificano se il soggetto indagato:

- sia in possesso di eventuali autorizzazioni a operare nei confronti dei soggetti residenti in Italia mediante verifica negli elenchi tenuti dalla CONSOB;
- sia stato oggetto di un *alert/warning* adottato da un'Autorità di vigilanza dei mercati finanziari aderente alla IOSCO mediante accesso al portale I-SCAN all'interno del quale vengono pubblicate tali comunicazioni nei confronti di soggetti che operano in assenza di autorizzazione mettendo in atto *scam/fraud*.

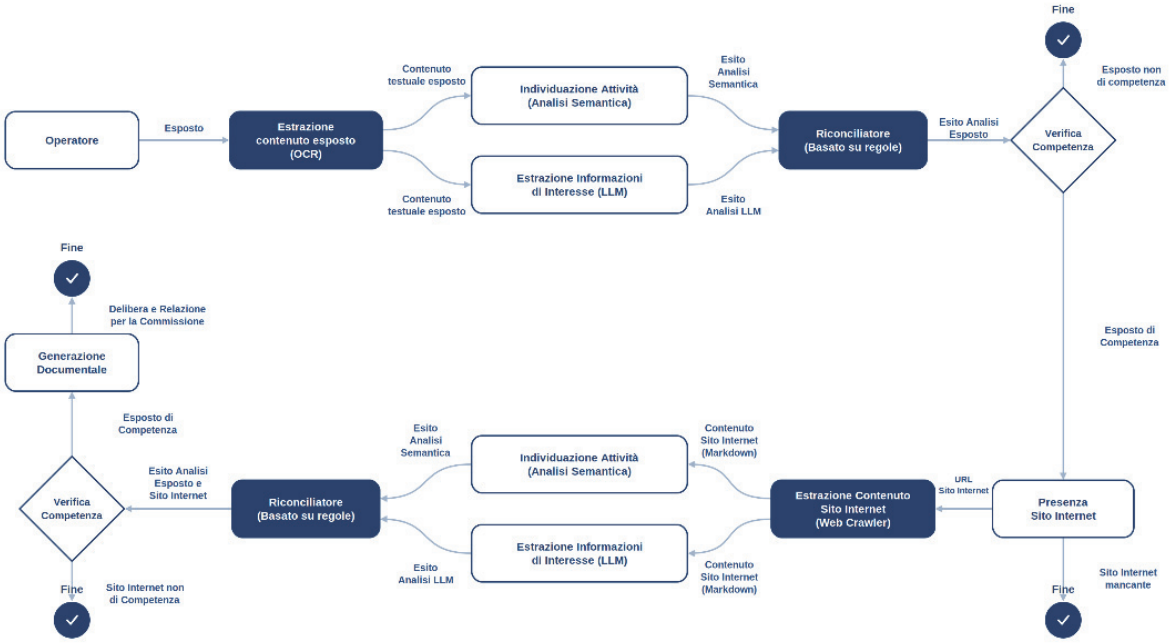
Per quanto riguarda i soggetti non presenti negli elenchi tenuti dalla CONSOB la verifica viene eseguita esclusivamente dall'essere umano mediante accesso ai siti *internet* delle singole Autorità che, stando a quanto riportato nel sito *internet* del soggetto indagato, avrebbero rilasciato l'autorizzazione.

3 Metodologia

La metodologia adottata in questo studio è stata sviluppata con l'obiettivo di fornire alla CONSOB uno strumento automatizzato per l'analisi degli esposti e dei contenuti dei siti/pagine *web* dei soggetti segnalati, con lo scopo di supportare l'attività di contrasto agli abusivismi finanziari.

16 Il sistema di allerta per gli investitori dell'International Organization of Securities Commissions (IOSCO) costituisce una risorsa per l'identificazione di soggetti segnalati da autorità di vigilanza internazionali per attività fraudolente o irregolari. L'accesso al portale I-SCAN (International Securities & Commodities Alerts Network della IOSCO raggiungibile all'indirizzo *web* https://www.iosco.org/i-scan/) consente non solo di confrontare le informazioni raccolte con segnalazioni e avvisi emessi da altre autorità mondiali omologhe alla CONSOB, ma anche di rafforzare l'efficacia delle analisi e delle valutazioni di conformità, permettendo di individuare con maggiore precisione e rapidità i soggetti già noti per comportamenti illeciti, facilitando così l'adozione di misure correttive adeguate.

Figura 2 – Diagramma del flusso operativo del prototipo sviluppato per l'analisi automatizzata degli esposti e dei siti *web* segnalati



L'approccio seguito combina tecniche di riconoscimento ottico di caratteri (OCR), *web crawling* in combinazione con analisi semantica e LLM, integrandole in una *pipeline* scalabile[17]. Dal punto di vista metodologico, le attività che il *software* svolge possono essere idealmente schematizzate in quattro aree di intervento.

La prima riguarda **l'acquisizione dei dati**, comprendente: *(i)*la raccolta degli esposti trasmessi dai risparmiatori attraverso i diversi canali disponibili, sia in forma libera sia secondo modelli preformati, corredati da eventuali allegati in formato eterogeneo; *(ii)* il contenuto dei siti *web* segnalati indicati negli esposti ritenuti di competenza. L'operatore può inoltre interrogare gli elenchi CONSOB e il portale I-SCAN della IOSCO per riscontrare la presenza del soggetto a cui è riferibile il sito *internet*.

La seconda consiste nel *pre-processing* e nella **normalizzazione dei contenuti acquisiti dagli esposti e dai siti *web***. Prima di applicare l'OCR ai

17 Nell'ambito di tale metodologia, le attività tecniche di industrializzazione hanno riguardato, in particolare, l'affinamento delle logiche di navigazione automatizzata, la gestione di contenuti generati dinamicamente e la trasformazione dei contenuti acquisiti in formati più idonei all'elaborazione automatica. Sul piano architetturale, il sistema è stato ulteriormente migliorato rendendo l'esecuzione delle analisi asincrona e introducendo la possibilità di avviare più analisi in parallelo. Tale intervento, che ha richiesto una revisione dell'architettura applicativa, consente di aumentare significativamente la capacità di elaborazione complessiva, riducendo i tempi di attesa per l'operatore e permettendo di gestire in modo più efficiente i picchi di carico legati all'afflusso di nuove segnalazioni.

documenti degli esposti, sono applicate tecniche per favorirne la leggibilità come la binarizzazione e la verifica dell'orientamento. Per i siti *web*, invece, una volta estratto il contenuto, questo viene normalizzato, ripulito dagli elementi non informativi e rappresentato in formato *Markdown*[18], così da preservarne la struttura semantica e facilitarne l'elaborazione da parte degli LLM.

La terza riguarda **l'estrazione delle informazioni**. In questo passaggio i contenuti acquisiti presenti negli esposti e nei siti *web* analizzati vengono elaborati lungo due direttrici. Da un lato, l'analisi semantica, basata su *set* di *keyword* predefinite, consente l'identificazione di attività e sottostanti, di competenza CONSOB, presenti nel contenuto. Dall'altro, l'impiego di LLM permette di estrarre le informazioni necessarie. L'integrazione dei due approcci garantisce sia la robustezza dell'estrazione sia la flessibilità necessaria per gestire testi complessi e non standardizzati.

La quarta prevede la **validazione e riconciliazione dei dati** estratti e classificazione dei contenuti. Le informazioni estratte dall'analisi semantica e dall'LLM vengono classificate secondo regole condizionali che riflettono il processo di vigilanza. Le regole condizionali assolvono dunque a un duplice compito: oltre a classificare il contenuto, ne riconciliano gli *output*, dirimendo eventuali sovrapposizioni tra risultati prodotti dall'analisi semantica e dagli LLM[19]. Il riscontro su fonti esterne autorevoli, quali gli elenchi CONSOB, il portale I-SCAN di IOSCO e le ricerche *web*, è invece condotto dall'operatore, che valida l'esito complessivo dell'analisi.

In questa architettura ciascun modulo presidia un diverso livello di analisi e l'affidabilità del risultato non dipende dal componente singolarmente più accurato, bensì dalla robustezza dell'intera *pipeline*.

L'accuratezza del sistema di analisi di esposti e siti *internet* è stata verificata attraverso il confronto dell'*output* del sistema automatico con l'analisi manuale condotta da operatori esperti, valutando la capacità di estrazione dei contenuti, valorizzazione dei campi estratti, classificazione e consistenza dell'*output* in selezionati scenari applicativi. L'analisi condotta ha evidenziato risultati positivi. I test effettuati, come appresso meglio descritti[20], hanno

18 Il *Markdown* è un formato di testo leggero che rappresenta la struttura logica essenziale di un documento (titoli, elenchi, gerarchie) mediante pochi simboli, senza l'apparato di *tag* proprio dell'HTML. Ciò consente di mantenere l'organizzazione dei contenuti in una forma compatta e facilmente elaborabile.

19 Le regole condizionali classificano il contenuto e ricompongono in un unico *output* i risultati dei due percorsi, che operano su dimensioni in larga parte complementari. La verifica dell'operatività è affidata esclusivamente all'analisi per parole chiave; l'estrazione delle informazioni strutturate (soggetto, tipologia di servizio, *claim* di autorizzazione) all'LLM. L'unica sovrapposizione riguarda siti ed *e-mail*, gestita tramite meccanismi che eliminano i doppioni: se un dominio è rilevato da entrambi gli approcci, compare una sola volta nell'*output*. La classificazione dell'esposto resta determinata dall'analisi per parole chiave, arricchita dai campi estratti dall'LLM. Le regole svolgono quindi una funzione di combinazione, non di arbitraggio tra esiti contrastanti.

20 Vedi *infra* § 2 "Risultati", Parte Terza.

dimostrato un livello di accuratezza elevato, con *output* che sono stati consistenti con le interpretazioni fornite dagli operatori umani.

Questa metodologia garantisce un approccio modulare, in cui ciascun componente tecnologico contribuisce a un diverso livello di analisi: l'OCR e il *web crawling* assicurano l'estrazione dei contenuti, l'analisi semantica consente l'identificazione dell'operatività, mentre i LLM permettono l'estrazione di informazioni da contenuti non strutturati e da contesti semanticamente complessi. Nei paragrafi successivi vengono descritti nel dettaglio i tre moduli principali della *pipeline*: analisi semantica (3.1), modelli di linguaggio di grandi dimensioni (3.2) e OCR (3.3).

La descrizione di tali moduli consente, inoltre, di mettere in luce non solo il contributo operativo del sistema, ma anche alcuni profili critici — in termini di spiegabilità, affidabilità, qualità del dato e supervisione umana — che assumono rilievo nella successiva riflessione sul quadro regolatorio e sui profili di *governance*.

3.1 Analisi semantica

L'individuazione delle attività segnalate dagli esponenti avviene attraverso un algoritmo di analisi semantica che si basa su un insieme di parole chiave definite a partire dall'esperienza degli operatori.

L'analisi semantica è un ramo della linguistica computazionale e dell'intelligenza artificiale che si concentra sull'interpretazione e la rappresentazione del significato delle parole, delle frasi e dei testi in un linguaggio naturale. Tra i principali componenti figurano la *tokenizzazione*[21], il *Part-of-Speech* (PoS) *Tagging*[22], il riconoscimento di entità nominate (NER)[23] e l'analisi sintattica[24].

21 La *tokenizzazione* è il processo di suddividere un testo in unità minime chiamate *token*, che possono essere parole, frasi o simboli. Questa è la fase preliminare dell'elaborazione del linguaggio naturale (NLP) e prepara il testo per l'elaborazione successiva. *Esempio:* Consideriamo il testo "Il mercato è rialzista oggi." La *tokenizzazione* produce: ['Il', 'mercato', 'è', 'rialzista', 'oggi', '.'].

22 Il *PoS Tagging* consiste nell'assegnare etichette grammaticali a ciascun *token*, indicando la funzione grammaticale della parola, come sostantivo, verbo, aggettivo, ecc. Questo è essenziale per comprendere la struttura sintattica delle frasi. Esempio: I *token* ['Il', 'mercato', 'è', 'rialzista', 'oggi', '.'] diventano: [('Il', 'DT'), ('mercato', 'NN'), ('è', 'VBZ'), ('rialzista', 'JJ'), ('oggi', 'NN'), ('.', '.')].

23 Il NER è il processo di identificazione e classificazione delle entità nominate nel testo, come persone, organizzazioni, luoghi, date, ecc. Questo passaggio è cruciale per estrarre informazioni specifiche e strutturate da testi non strutturati. Esempio: Consideriamo il testo "Apple Inc. ha riportato utili più alti nel Q2." Il NER produce: [('Apple Inc.', 'ORGANIZATION'), ('Q2', 'DATE')]. Per una trattazione generale del riconoscimento di entità nominate, si veda Nadeau, D., Sekine, S. (2007), *A survey of named entity recognition and classification*, Lingvisticae Investigationes, 30(1), 3-26, https://doi.org/10.1075/li.30.1.03nad; per un approccio basato su architetture neurali, si veda Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016), *Neural Architectures for Named Entity Recognition*, Proceedings of NAACL-HLT 2016, 260-270.

24 L'analisi sintattica è il processo di analisi della struttura grammaticale delle frasi. Costruisce un albero sintattico che rappresenta la struttura gerarchica della frase, mostrando come le parole si combinano in unità più complesse,

Una delle caratteristiche avanzate dell'analisi semantica è la capacità di riconoscere schemi complessi, utile, nel caso in esame, per identificare le relazioni tra attività e i loro sottostanti. Questa capacità permette di associare attività di *trading* con strumenti finanziari anche se i termini non sono vicini nel testo[25].

Esempio: consideriamo il seguente testo da una piattaforma di *trading*:

> *"L'utente può svolgere varie attività di trading sulla nostra piattaforma. Le nostre offerte includono il trading di CFD su indici azionari principali, commodity, e coppie forex."*

In questo esempio, i termini "*trading*" e "*commodity*" non sono adiacenti come bigramma all'interno della frase, essendo separati da una sequenza di parole intermedie. Tuttavia, l'analisi semantica riconosce che "*trading*" è associato a "*commodity*" come "CFD". Questa capacità di riconoscere schemi complessi consente una comprensione più profonda delle relazioni tra azioni (come il *trading*) e gli strumenti finanziari coinvolti (come i CFD su *commodities*), fornendo così informazioni più accurate e dettagliate estratte dal testo.

Una volta individuata la presenza e il ruolo delle *keyword* di interesse nei contenuti del sito *web* e dell'esposto, si procede alla verifica di eventuali negazioni. In alcuni casi, infatti, le frasi possono specificare che un operatore non svolge l'attività rilevata. Pertanto, nelle frasi contenenti combinazioni di *keyword* vengono individuate le negazioni e le corrispondenti rilevazioni vengono escluse dall'analisi.

3.2 Modelli di linguaggio di grandi dimensioni (LLM)

Nel contesto del progetto in analisi, i modelli linguistici di grandi dimensioni (LLM) vengono impiegati per estrarre informazioni specifiche in formato strutturato dagli esposti e siti *web*. Per l'estrazione dei dati, è stata definita in linguaggio naturale una dettagliata descrizione del compito da svolgere e sono stati individuati i campi richiesti in *output*, ciascuno corredato

denominate sintagmi (ad esempio sintagma nominale o sintagma verbale), e le relazioni tra di essi. Esempio: nella frase "I prezzi delle azioni stanno aumentando rapidamente", l'analisi sintattica individua il sintagma nominale "I prezzi delle azioni", con funzione di soggetto, e il sintagma verbale "stanno aumentando rapidamente", con funzione di predicato.
La frase verbale interna (VP) è composta da un verbo al gerundio o participio presente (VBG) "aumentando" e una frase avverbiale (ADVP).
La frase avverbiale (ADVP) è composta da un avverbio (RB) "rapidamente".

25 Per una descrizione di una *pipeline* NLP comprendente *tokenizzazione*, *PoS tagging* e *parsing* sintattico, si veda Manning, CD., Surdeanu, M., Bauer, J., Finkel, J., Bethard, SJ., McClosky, D. (2014), *The Stanford CoreNLP Natural Language Processing Toolkit*, Proceedings of ACL 2014: System Demonstrations.

da una dettagliata descrizione in linguaggio naturale, basata sull'esperienza dell'operatore. La spiegazione del compito, insieme ai campi e al contenuto da processare, costituisce l'*input* per il processo di inferenza del modello, che genera i risultati nel formato di risposta definito.

Esempio: supponiamo di voler identificare la presenza di riferimenti a pubblicità all'interno di un testo. Sarà necessario fornire al modello un formato di risposta specifico, includendo la descrizione dettagliata per ciascun campo, e il contenuto da cui estrarre le informazioni rilevanti.

```
output_format = {
       presenza_pubblicita : bool = Field(
              default = false,
              description = "Campo booleano che rileva se
              l'esponente fa riferimento a pubblicità utilizzata
              dai soggetti segnalati. Impostare su true se il
              testo evidenzia, in modo esplicito o implicito,
              l'impiego di materiale promozionale; in assenza di
              tali riferimenti, il valore rimane false."
              )
       }

input_testuale = "Ho completato la registrazione alla
piattaforma XYZ dopo aver visualizzato un annuncio pubblicitario
su Facebook, il quale prometteva un rendimento garantito di €200
mensili."
```

A partire dal formato di risposta fornito, il modello restituirà un *output* conforme alla struttura specificata, dove il campo "*presenza_ pubblicita"* sarà valorizzato in funzione del contenuto dell'*input*. In questo esempio, l'*output* sarà il seguente:

```
{presenza_pubblicita = true}
```

Il sistema ha rilevato correttamente che si fa riferimento a pubblicità nell'*input* fornito, valorizzando il campo "*presenza_pubblicita*" come *true*.

Oltre alla capacità di processare testo, i LLM Multimodali[26], tra cui i modelli della famiglia GPT-5 di OpenAI, consentono di gestire *input* di molteplici formati, come testo, immagini e audio. In caso di utilizzo di questi modelli è possibile processare direttamente il file, senza la necessità di estrarne il contenuto in formato testuale.

26 Un LLM multimodale è un modello di intelligenza artificiale capace di elaborare e comprendere dati provenienti da diverse modalità, come testo e immagini. Integra simultaneamente informazioni visive e linguistiche, consentendo una comprensione più completa rispetto ai modelli tradizionali che analizzano solo il testo. Esempi includono modelli come CLIP e GPT-4o Vision, utilizzati per estrarre significati sia dai contenuti testuali che visivi.

L'adozione degli LLM per l'estrazione di contenuti da documenti complessi si è dimostrata particolarmente efficace, specialmente in presenza di *layout* articolati o di immagini a bassa risoluzione. Gli LLM sono in grado di processare tali contenuti, estraendo le informazioni in modo strutturato, garantendo risultati di elevata precisione[27]. L'efficacia dell'estrazione di contenuti testuali attraverso LLM dipende dall'applicazione specifica e dal modello impiegato. In tale ambito, i modelli multimodali che integrano contenuto testuale e informazione di *layout* - come LayoutLM[28]- rappresentano una direzione di sviluppo particolarmente promettente.

Nel caso di *input* documentali di grandi dimensioni, l'elevato volume di informazioni da elaborare può rendere necessarie strategie di segmentazione dell'*input*, con un conseguente aumento del numero di chiamate al modello, del costo computazionale dell'inferenza e della latenza complessiva del processo di analisi.

3.3 OCR

L'*Optical Character Recognition* (OCR) consente di estrarre testo da immagini o scansioni, trasformandolo in contenuto testuale elaborabile[29]. Nel contesto di esposti inviati dai risparmiatori alla CONSOB e siti *internet* segnalati, l'OCR è impiegata per l'estrazione di contenuti testuali dagli esposti e allegati.

L'OCR opera attraverso diverse fasi mirate all'estrazione del testo contenuto in immagini. Questa sequenza di fasi pre-elabora il contenuto dell'immagine, ottimizzandola per l'estrazione del contenuto, garantendo che questo sia estratto correttamente per ulteriori elaborazioni come l'utilizzo di tecniche basate su *Natural Language Processing* (NLP). Gli esposti dei risparmiatori sono in formati eterogenei, tra cui *screenshot*, documenti, ed

27 Un esempio di tale capacità estrattiva applicata in ambito RegTech è offerto da Trerotola M., Parente M., Calvaresi D., A Hybrid Multi-Agent System for Early Scam Detection in Crypto-Assets, in Applied Sciences, vol. 16, n. 7, 2026, 3122, ove i modelli linguistici di grandi dimensioni sono impiegati, in combinazione con logiche deterministiche basate su regole, per la classificazione tassonomica degli emittenti di cripto-attività, la verifica degli obblighi di disclosure previsti dal Regolamento (UE) 2023/1114 (MiCAR) e l'individuazione precoce di schemi fraudolenti mediante l'analisi congiunta della documentazione off-chain e dei segnali on-chain

28 Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. (2020), LayoutLM: Pre-training of Text and Layout for Document Image Understanding, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1192-1200.

29 Il Riconoscimento Ottico dei Caratteri (OCR) è una disciplina della visione artificiale e dell'intelligenza artificiale che si occupa della conversione automatica di immagini contenenti testo stampato, manoscritto o dattiloscritto in dati testuali codificati digitalmente. Utilizzando tecniche avanzate di elaborazione delle immagini e algoritmi di apprendimento automatico, l'OCR analizza le caratteristiche visive dei caratteri per segmentare, riconoscere e interpretare le singole lettere e numeri. Questo processo coinvolge diverse fasi, tra cui la pre-elaborazione dell'immagine, la segmentazione dei caratteri, l'estrazione delle caratteristiche e il riconoscimento mediante modelli statistici o reti neurali profonde. L'OCR trova applicazione in numerosi ambiti, tra cui la digitalizzazione di documenti cartacei, l'archiviazione elettronica, l'estrazione automatica di informazioni e la facilitazione dell'accessibilità digitale. Per una revisione sistematica delle tecniche OCR, con particolare riferimento al riconoscimento di testo manoscritto, si veda Memon, J., Sami, M., Khan, RA., Uddin, M. (2020), *Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)*, IEEE Access, 8, 142642-142668,

e-mail. La qualità di queste immagini può variare considerevolmente. È pertanto necessaria una prima fase di pre-elaborazione per migliorare l'accuratezza del riconoscimento del testo, attraverso: (i) l'applicazione di un filtro per rimuovere il "rumore visivo" causato da bassa qualità della scansione o da artefatti dovuti alla compressione dell'immagine; (ii) sistemi di correzione dell'orientamento e allineamento del testo; e (iii) binarizzazione dell'immagine attraverso un processo di conversione in bianco e nero aumentando il contrasto tra il testo e lo sfondo, rendendo più semplice il riconoscimento dei caratteri.

Segue poi una seconda fase di segmentazione del testo che permette il riconoscimento di righe, parole e caratteri. Dopo la pre-elaborazione, il testo viene suddiviso in segmenti logici, come righe, parole e infine singoli caratteri. L'OCR utilizza algoritmi di *pattern matching* per confrontare i caratteri riconosciuti con un *set* di modelli noti. Nel sistema oggetto di questo lavoro, tale fase è supportata da Tesseract OCR[30], che utilizza tecniche di riconoscimento basate su reti neurali ricorrenti, in particolare LSTM (*Long Short-Term Memory*)[31], per migliorare l'accuratezza dell'estrazione testuale anche in presenza di font complessi o qualità visiva non ottimale.

Questo passaggio è cruciale per migliorare l'accuratezza del risultato finale. In tale fase, i LLM possono essere utilizzati per verificare la coerenza delle frasi e correggere eventuali incongruenze.

Nel progetto in esame, l'OCR è essenziale per digitalizzare del contenuto presente in esposti. Gli esposti inviati dai risparmiatori spesso, infatti, includono allegati in formati non strutturati, da cui non è possibile estrarne direttamente il contenuto, come *screenshot* o scansioni di documenti. L'OCR consente di estrarre il contenuto di questi documenti, utilizzandolo per le successive analisi.

È importante sottolineare che l'accuratezza del processo di OCR dipende fortemente dalla qualità dell'immagine in *input* che se distorte o a bassa risoluzione possono compromettere la capacità del sistema di riconoscere correttamente il testo.

30 Smith, R. (2013), History of the Tesseract OCR Engine: What Worked and What Didn't, Proceedings of SPIE 8658, Document Recognition and Retrieval XX, 865802.

31 LSTM *(Long Short-Term Memory*) è un tipo di rete neurale ricorrente (RNN) progettata per apprendere e ricordare relazioni di lungo termine nelle sequenze di dati. A differenza delle RNN tradizionali, le LSTM hanno celle di memoria speciali che consentono loro di mantenere e aggiornare informazioni rilevanti per periodi di tempo più lunghi, superando il problema della *vanishing gradient*. Queste caratteristiche rendono le LSTM particolarmente efficaci per compiti sequenziali come il riconoscimento vocale, la traduzione automatica e il riconoscimento ottico nei caratteri (OCR).

1 Elaborazione e analisi delle segnalazioni

Il processo si articola in due fasi strettamente connesse e complementari. La prima riguarda l'elaborazione dei documenti inviati dall'esponente, con l'obiettivo di estrarne in modo strutturato le informazioni rilevanti. La seconda si concentra invece sui siti *internet* riconducibili al soggetto segnalato, al fine di estrarre contenuti e identificare attività e informazioni rilevanti. La *ratio* della scelta architetturale, di integrare i risultati attraverso una *pipeline* ibrida, consente di ricostruire un quadro informativo completo e coerente, utile a supportare le attività di vigilanza e contrasto agli abusivismi finanziari.

Figura 3 – *Sequence Diagram* del funzionamento del sistema di analisi esposti

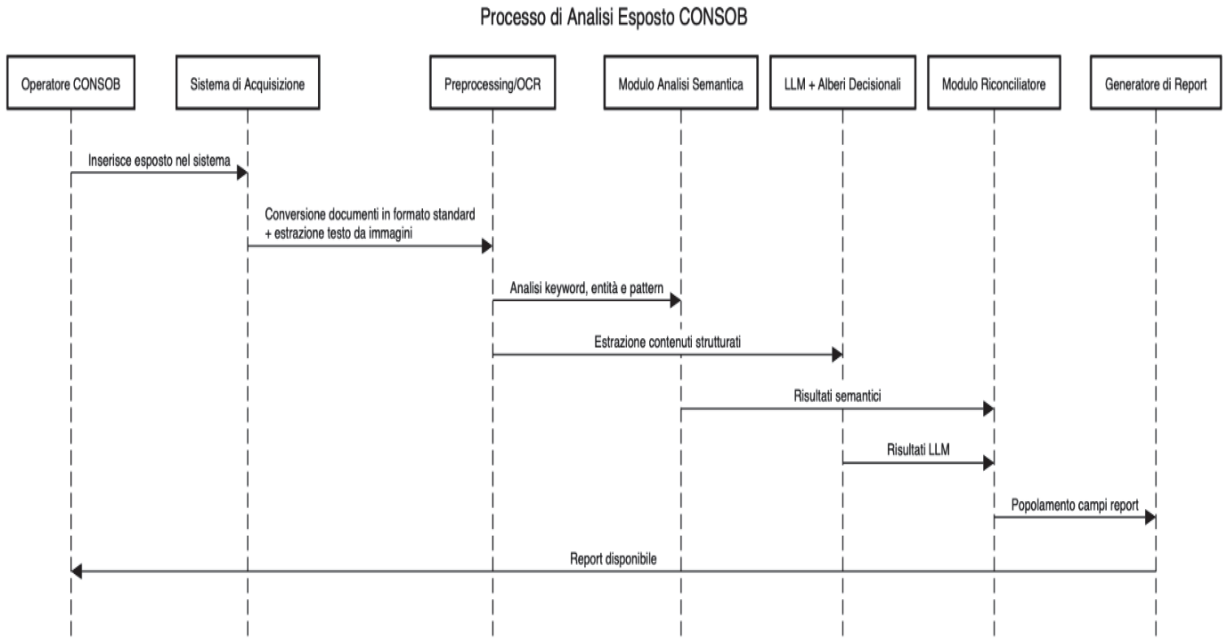
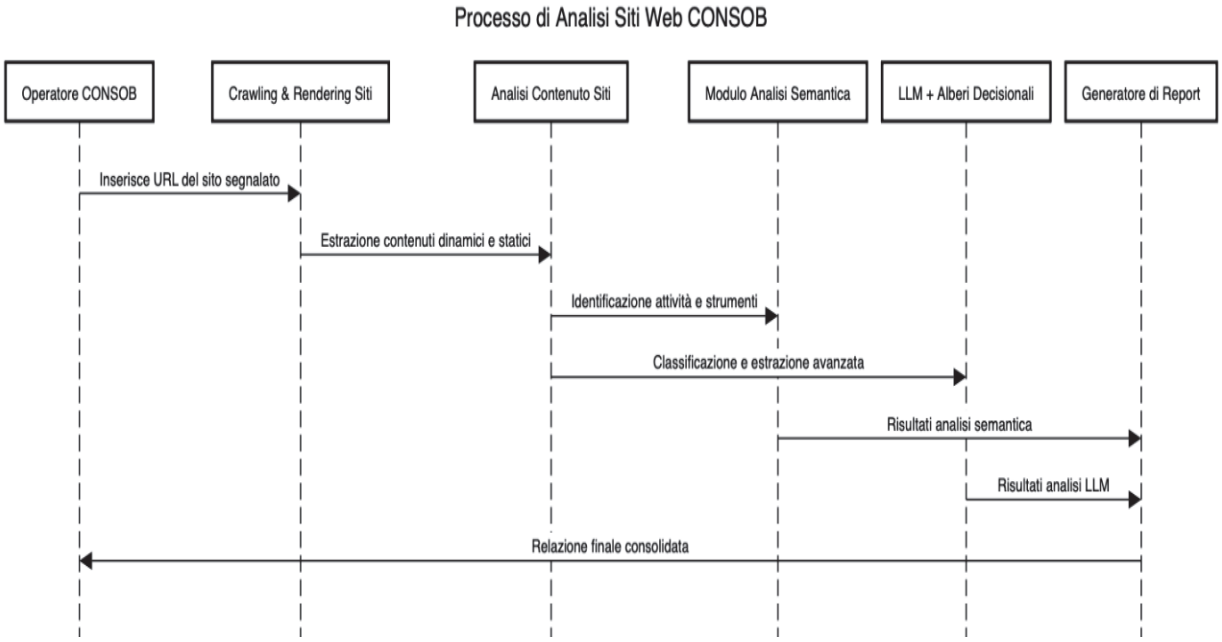


Figura 4 – *Sequence Diagram* del funzionamento del sistema di analisi siti internet



Il primo passo per la trattazione e l'analisi di un esposto trasmesso da un risparmiatore consiste nella standardizzazione del formato dei documenti ricevuti. Gli *input*, spesso eterogenei (PDF, immagini scannerizzate o fotografie), vengono convertiti in immagini uniformi sulle quali applicare procedure di *pre-processing* prima dell'esecuzione dell'OCR. In questo contesto è stato utilizzato *Tesseract*, uno degli strumenti più consolidati per il riconoscimento ottico dei caratteri.

Il *pre-processing* dell'immagine segue una *pipeline* mirata a migliorarne leggibilità e qualità. L'immagine a colori viene innanzitutto convertita in scala di grigi, riducendone la complessità visiva. Successivamente viene applicata l'equalizzazione dell'istogramma, tecnica che consente di aumentare il contrasto e rendere più netta la distinzione tra i caratteri e lo sfondo, anche in presenza di documenti sbiaditi o acquisiti in condizioni di luce non ottimali. Infine, l'immagine viene binarizzata attraverso l'algoritmo di Otsu[32], che individua in modo automatico la soglia ottimale per separare le aree testuali da quelle di sfondo. Questa sequenza di trasformazioni garantisce un *input* pulito e standardizzato, migliorando la separazione tra testo e sfondo e l'accuratezza del riconoscimento da parte dell'OCR. Trattandosi del primo anello della *pipeline*, la qualità dell'estrazione operata dall'OCR condiziona l'affidabilità di tutte le fasi successive: un riconoscimento impreciso può compromettere

32 Otsu, N. (1979), A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62-66.

l'individuazione di informazioni decisive per l'istruttoria, come gli URL di siti *internet*, i nominativi di persone o società e gli indirizzi *e-mail*; il rischio è che l'errore si propaghi lungo l'intera catena di analisi e ne degradi l'esito complessivo.

L'esecuzione di *Tesseract* sul documento così pre-processato produce un'unica stringa di testo, che costituisce il contenuto dei documenti inviati dall'esponente e la base per le successive fasi di analisi automatizzata. Tale stringa viene elaborata nei moduli di analisi semantica ed estrazione tramite LLM, con l'obiettivo di identificare degli elementi necessari per l'analisi e la classificazione del contenuto.

Nel modulo di analisi semantica vengono identificate coppie di parole sia lato attività sia lato prodotti (es. strumenti finanziari), utili a rilevare la presenza di attività di interesse per la CONSOB all'interno del testo. Il sistema utilizza un *set* di parole chiave, definito sulla base dell'esperienza degli operatori e distinto per esposti e siti *internet*, per identificare frasi in cui tali parole compaiano congiuntamente. Una volta identificata la frase, questa viene validata escludendo i casi in cui siano presenti negazioni[33] nel contesto e coppie attività/prodotti confermate vengono quindi aggiunte alle attività rilevate. Questo approccio facilita la verifica da parte dell'operatore e aumenta la trasparenza del sistema, migliorandone l'*explainability*.

Ogni contenuto viene elaborato sistematicamente, in parallelo all'analisi basata su parole chiave ed espressioni regolari, da un modello di linguaggio di grandi dimensioni (LLM). L'LLM non opera come *fallback*, ma è applicato a tutti gli esposti; il suo contributo è particolarmente rilevante per le informazioni che dipendono dal contesto e che l'esponente può esprimere con un'ampia varietà di termini (ad esempio la modalità di contatto o la presenza di un contatto attivo), non riconducibili a un dizionario chiuso di parole chiave. Nel prototipo in esame è stato utilizzato il modello GPT-5[34] di OpenAI[35].

Per impiegare un LLM è necessaria una chiara definizione del compito, dell'*input* fornito e dell'*output* atteso. Nel caso dell'estrazione di informazioni strutturate in formato chiave-valore, al modello vengono fornite specifiche

33 Si è scelto di escludere dall'analisi in parola le frasi contenenti negazioni (es. non forniamo servizi d'investimento), normalmente collocate all'interno dei *disclaimer* presenti nei *footer* delle pagine *web* indagate, al fine di ridurre i falsi positivi, ossia evitare che il *software,* nel considerare attendibili tali informazioni, escluda dal perimetro d'intervento il sito oggetto di verifica. Nella prassi operativa è, infatti, emerso che sovente nei siti *internet* degli operatori abusivi fosse presente tale avvertenza.

34 OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., *et al.* (2023), GPT-4 Technical Report, arXiv:2303.08774, https://doi.org/10.48550/arXiv.2303.08774

35 GPT-4o è stato inizialmente adottato in quanto rappresentava la versione più recente disponibile al momento della sperimentazione del prototipo; la successiva famiglia GPT-5, rilasciata da OpenAI nell'agosto 2025, offre miglioramenti in termini di accuratezza, capacità di ragionamento, riduzione delle allucinazioni e ampiezza della finestra di contesto. L'architettura del sistema è parametrica rispetto al modello: il passaggio a versioni successive richiede aggiornamenti minimi alla *pipeline*.

coppie campo-descrizione, in cui la descrizione illustra in linguaggio naturale il significato del campo. Attraverso un processo iterativo di ottimizzazione dei *prompt*, è possibile incrementare l'accuratezza dell'estrazione. L'*output* prodotto dal modello viene poi validato e trasformato in un formato compatibile con gli altri componenti dell'analisi.

Il prototipo è strutturato in due componenti principali: un'interfaccia utente (*frontend*), tramite cui l'operatore interagisce con il sistema, e una componente lato server (*backend*), che gestisce l'elaborazione dei contenuti e l'interazione con il modello GPT-4o. Il *backend* è realizzato in Python sulla base dei *framework* FastAPI e LangChain[36]. Sfruttando la funzionalità di *function calling*[37], è possibile definire modelli di risposta che specificano per ogni campo: la tipologia, il valore di *default* e le informazioni relative alla valorizzazione in linguaggio naturale, in modo da ottenere in *output* direttamente un oggetto validato e popolato secondo precise istruzioni. Pur non trattandosi di una forma piena di *explainable* AI, al modello viene richiesto di fornire una giustificazione in linguaggio naturale per ciascun campo estratto, offrendo all'operatore elementi utili a verificare la correttezza del risultato; va peraltro ricordato che le giustificazioni generate dal modello stesso costituiscono una razionalizzazione successiva e non riflettono fedelmente il processo inferenziale interno, e vanno pertanto trattate come ausilio operativo, non come spiegazione formale. Approcci formali alla spiegazione delle predizioni dei modelli di *machine learning*, come SHAP[38] e LIME[39] costituiscono un riferimento metodologico in materia.

36 L'interfaccia utente è sviluppata in Next.js, un *framework* per la realizzazione di applicazioni *web*. Python è un linguaggio di programmazione ampiamente utilizzato in ambito IA; FastAPI è un *framework* per la realizzazione di interfacce di programmazione (API), che consentono la comunicazione tra i diversi componenti *software*, mentre LangChain è una libreria concepita per lo sviluppo di applicazioni basate su modelli linguistici di grandi dimensioni (gestione dei *prompt*, strutturazione degli *output*, orchestrazione delle chiamate al modello).

37 Il *function calling* (o *tool calling*) è una funzionalità dei modelli linguistici più recenti che consente di vincolare l'*output* a uno schema predefinito dallo sviluppatore: anziché restituire testo libero, il modello produce un oggetto strutturato (tipicamente in formato JSON) conforme ai campi e ai tipi di dato specificati. Ciò rende l'*output* direttamente elaborabile dagli altri componenti *software* e ne consente la validazione automatica, riducendo il rischio di risposte in formati non previsti.

38 Lundberg, SM., Lee, SI. (2017), A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (NeurIPS), 30, 4765-4774. SHAP (*SHapley Additive exPlanations*) è un metodo di spiegazione a posteriori (*post-hoc*) e indipendente dal modello, che attribuisce a ciascuna variabile di *input* un contributo quantitativo alla singola predizione. Tali contributi si fondano sui valori di Shapley, concetto mutuato dalla teoria dei giochi cooperativi che ripartisce in modo equo l'esito tra i fattori che vi concorrono. SHAP consente così di spiegare la singola decisione e, per aggregazione, di stimare l'importanza complessiva delle variabili.

39 Ribeiro, MT., Singh, S., Guestrin, C. (2016), "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. LIME (*Local Interpretable Model-agnostic Explanations*) è un metodo di spiegazione a posteriori e indipendente dal modello che approssima localmente il comportamento di un classificatore, in prossimità della specifica istanza da spiegare, mediante un modello surrogato semplice e interpretabile (tipicamente lineare). L'approssimazione si ottiene perturbando l'*input* e osservando come variano le predizioni, individuando in tal modo le variabili che hanno maggiormente inciso su quel singolo esito.

Un esempio rilevante riguarda l'identificazione degli URL presenti negli esposti, necessari per l'avvio dell'analisi dei siti *internet*. In questo caso, l'impiego dell'LLM consente di distinguere tra URL pertinenti e non pertinenti.

Una volta completata l'analisi semantica e l'elaborazione tramite LLM, le informazioni estratte vengono riconciliate e utilizzate per classificare l'esposto. La classificazione avviene tramite un approccio ibrido basato su regole e intelligenza artificiale, che riproduce la logica decisionale dei funzionari: gli esposti vengono così suddivisi tra rilevanti e non rilevanti.

Conclusa questa fase e identificato il sito *internet* segnalato, si procede alla sua analisi. Le pagine da esaminare possono essere estratte dalla *sitemap* del sito, oppure, in assenza di questa, individuate a partire dalla *homepage* tramite un processo iterativo di *crawling*.

Alcuni siti generano parte dei propri contenuti direttamente nel *browser* dell'utente, anziché trasmetterli già pronti dal *server*: è il caso, in particolare, di quelli realizzati con tecnologie come *React*[40]. Per gestire la navigazione e l'estrazione dei contenuti è stato quindi implementato un processo di *web crawling* con *Python* e *Selenium*[41], utilizzando *Chrome* come *browser*. Questo approccio simula la navigazione di un utente reale, garantendo l'accesso e il caricamento completo delle pagine dinamiche. Poiché il codice HTML contiene *markup*[42] e metadati non sempre rilevanti ai fini dell'elaborazione, il contenuto viene estratto e convertito in formato *Markdown*, che conserva la gerarchia logica dei contenuti (titoli, sottosezioni, elenchi) eliminando al contempo l'apparato di *tag* superfluo. Questa scelta non rappresenta un mero dettaglio implementativo, ma risponde a problemi concreti: riduce il rumore informativo, presentando al modello il contenuto sostanziale e non il codice di formattazione; contiene costo e latenza dell'inferenza, limitando i *token* superflui; e, offrendo un *input* più pulito e strutturato, riduce il rischio di allucinazioni nell'analisi di pagine lunghe o articolate. Questa conversione consente un'elaborazione più efficiente da parte dei modelli di linguaggio, evitando spreco di *token* dovuto a *tag* e strutture ridondanti.

40 React è una libreria JavaScript per la creazione di interfacce *web*, utilizzabile sia generando i contenuti sul server sia direttamente nel browser dell'utente. In quest'ultimo caso (esecuzione *client-side*) i contenuti non sono presenti nel codice HTML inviato dal server: il semplice prelievo del codice sorgente non ne restituisce quindi il contenuto effettivo.

41 Selenium è uno strumento di automazione del browser, qui impiegato per caricare ed esplorare le pagine *web* come farebbe un utente reale.

42 Il *markup* è l'insieme di etichette (*tag*) e metadati con cui una pagina *web* viene strutturata e formattata: nel codice HTML indica, ad esempio, dove iniziano titoli, paragrafi, *link* o tabelle. Si tratta di informazioni utili alla visualizzazione nel *browser*, ma in larga parte irrilevanti ai fini dell'analisi del contenuto.

Dai contenuti estratti vengono ricavate le informazioni rilevanti tramite una *pipeline* analoga a quella utilizzata per l'analisi degli esposti, combinando analisi semantica ed estrazione basata su LLM. Nei casi in cui la lunghezza del contenuto superi la finestra del contesto del modello[43], espressa in *token*, il contenuto viene suddiviso in *chunk* e processato separatamente, con una fase successiva di riconciliazione. I modelli di ultima generazione, come GPT-5, consentono di processare anche siti *web* di dimensioni significative grazie a finestre di contesto estese (sino a vari milioni di *token* nei modelli di frontiera), sebbene *benchmark* recenti evidenzino fenomeni di degrado dell'attenzione sui contesti molto lunghi[44] che suggeriscono cautela nell'interpretazione dell'*output*.

Successivamente, è possibile verificare la presenza del soggetto nei registri CONSOB, il riscontro del sito o del nominativo nel portale IOSCO I-SCAN, e infine condurre ricerche sul *web* (ad esempio tramite il motore di ricerca Bing) per individuare eventuali riferimenti ad attività fraudolente.

Al termine dell'analisi, l'operatore valida l'esito, corregge eventuali errori di estrazione e aggiunge note utili al miglioramento del sistema. Solo a valle di questa validazione, l'operatore può richiedere la generazione di *report* strutturati relativi al soggetto segnalato, a supporto dei passaggi preliminari di estrazione, selezione, classificazione e predisposizione documentale: il sistema popola con i dati estratti i campi di modelli predefiniti, producendo un supporto documentale che resta interamente soggetto alla revisione dell'analista. Il sistema non assume in alcun caso decisioni in via autonoma: il suo intervento resta circoscritto a tali fasi preliminari e preparatorie, secondo una logica di *human in the loop* nella quale la valutazione e la decisione finale, e la relativa responsabilità, restano in capo all'operatore.

È proprio questa prossimità alle attività istruttorie — pur in una logica di supporto e non di sostituzione dell'analista — a rendere particolarmente rilevanti i temi dell'affidabilità, della verificabilità degli *output* e della tracciabilità dei passaggi di analisi.

43 La finestra del contesto di un modello di linguaggio di grandi dimensioni (LLM) è definita come il numero massimo di *token* (T) che il modello è in grado di accettare e processare come *input* durante una singola operazione di inferenza. Questo parametro, denotato come T, è intrinseco all'architettura del modello e può variare significativamente a seconda della configurazione specifica e delle capacità del modello. In pratica, T rappresenta la capacità del modello di mantenere un'attenzione simultanea su una porzione limitata di dati di *input*, determinando così la portata del contesto che il modello può utilizzare per generare risposte o fare previsioni.

44 Liu, NF., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P. (2024), Lost in the Middle: How Language Models Use Long Contexts, Transactions of the Association for Computational Linguistics, 12, 157-173.

2 Risultati

La valutazione delle prestazioni dello strumento sviluppato è stata condotta attraverso una metodologia rigorosa che ha previsto il confronto tra i risultati prodotti dal *software* e quelli ottenuti dall'analisi da un operatore esperto.

La progettazione del sistema si è articolata lungo tre direttrici:

- lo sviluppo del modulo di analisi degli esposti;
- lo sviluppo del modulo di analisi dei siti *internet*; e
- l'integrazione e il miglioramento incrementale del sistema, comprensivo di entrambe le *pipeline*[45].

Quest'ultima direttrice ha riguardato sia l'esperienza d'uso dell'operatore e l'affinamento delle procedure di estrazione, sia il trasferimento delle informazioni estratte in appositi *report*. L'obiettivo di tale processo è stato quello di verificare l'accuratezza, la coerenza, l'affidabilità e la stabilità dello strumento in differenti scenari applicativi.

La procedura di valutazione è stata articolata nei seguenti passaggi:

1. Selezione del campione di test: è stato predisposto un campione rappresentativo di esposti e siti *internet*, selezionato in modo da coprire una vasta gamma di casi d'uso. Il campione comprendeva documenti con vari livelli di complessità, inclusi testi con linguaggio figurato, ambiguità semantica e *layout* complessi.
2. Applicazione dello strumento: lo strumento sviluppato è stato applicato all'intero campione di test e i risultati generati dallo strumento sono stati registrati e documentati per consentire successivi confronti.
3. Confronto manuale: parallelamente, un operatore umano esperto ha condotto un'analisi manuale sugli stessi casi, seguendo i medesimi criteri adottati dal *software*. I risultati ottenuti sono stati quindi confrontati con quelli dello strumento, al fine di individuare eventuali discrepanze o incoerenze.

Le differenze emerse tra i risultati dello strumento e quelli dell'analisi manuale sono state esaminate in dettaglio per individuarne le cause. Tale analisi ha permesso di evidenziare i limiti dello strumento e di definire le situazioni in cui l'intervento umano risulta ancora preferibile o necessario,

45 I primi test sulle *pipeline* separate sono stati condotti nel mese di aprile 2024 e hanno riguardato 20 esposti e 58 siti *internet*. I successivi test sulla *pipeline* completa sono stati condotti tra il mese di maggio del 2024 e settembre dello stesso anno e hanno riguardato 89 diversi esposti in esito ai quali sono stati esaminati i contenuti di 85 siti *internet* (in 4 esposti non era menzionato alcun sito *internet*).

offrendo al contempo indicazioni utili per futuri sviluppi e miglioramenti del sistema.

2.1 Verifiche sul modulo di analisi degli esposti

Il lavoro ha preso avvio da un *subset* iniziale di 20 esposti (debitamente anonimizzati), selezionati appositamente per la loro eterogeneità, al fine di individuare le diverse tipologie e i diversi formati delle segnalazioni trasmesse dai risparmiatori e di definire una *pipeline* di estrazione dei contenuti adeguata. Le segnalazioni si presentavano in formati eterogenei: documenti PDF, *screenshot*, modelli precompilati generati dal *form* del sito CONSOB e scansioni di documenti. Nello specifico, nel campione inizialmente esaminato il 60% delle segnalazioni (12 su 20) era redatto secondo il *template* CONSOB, il 30% (6 su 20) era costituito da scansioni in formato PDF e il restante 10% (2 su 20) da *screenshot*[46]. La verifica dell'accuratezza del sistema è stata condotta mediante confronto con un operatore umano; l'intervento ha riguardato principalmente il miglioramento del processo di estrazione del testo dai documenti trasmessi in forma libera o rappresentati da foto. Laddove l'esposto è stato trasmesso utilizzando il *form* della CONSOB l'estrazione delle informazioni ha dato immediatamente risultati soddisfacenti nella totalità dei casi.

L'estrazione delle informazioni dagli esposti si è rivelata meno complessa rispetto, invece, all'analisi dei siti *internet* stanti le rilevate difficoltà connesse, principalmente, alla limitata persistenza temporale dei siti fraudolenti, la cui durata in rete è circoscritta.

2.2 Verifiche sul modulo di analisi dei siti *internet*

Il modulo è stato, inizialmente, sviluppato su un insieme di 58 siti *internet*. In una fase iniziale è stato individuato un set di *keyword* da ricercare all'interno del sito per accertarne l'operatività; elemento parimenti essenziale è la presenza di riferimenti ad autorizzazioni e delle informazioni identificative del soggetto. Poiché tali informazioni possono trovarsi anche in sottopagine, si è reso necessario predisporre un'estrazione multipagina. È stato infatti osservato che in 6 casi su 58 le autorizzazioni non comparivano nel *footer* del sito, bensì in apposite sottopagine che richiedevano un'estrazione separata. Per

46 Si è scelto di seguire un approccio a difficoltà incrementale partendo, quindi, dagli esposti trasmessi dai risparmiatori tramite il *format* presente nel sito della CONSOB, il quale si caratterizza per l'ordinata allocazione delle informazioni rilevanti secondo una procedura guidata. Questo ha permesso di isolare più agevolmente le incongruenze e di effettuare un confronto con gli esiti cui è pervenuto il *software* in esito all'esame dei documenti trasmessi in forma libera. Ad esempio, era inizialmente emersa la difficolta del *software* di leggere gli indirizzi *e-mail* e gli URL dei siti *web* che apparivano di colore celeste o arancione perché il redattore aveva attivato il *link.* Hanno, altresì, formato oggetto di specifica verifica l'esame congiunto di plurimi esposti trasmessi da risparmiatori in relazione alla medesima vicenda. Il *software* permette, infatti, il caricamento di più documenti e restituisce un unico esito che tiene conto delle informazioni complessivamente estratte.

questo motivo è stato sviluppato un sistema di *crawling* che non si limita all'estrazione della *homepage*, ma analizza tutte le pagine di primo livello del sito, attribuendo particolare priorità a quelle contenenti informazioni legali e riferimenti ad autorizzazioni. In 14 casi, inoltre, i siti presentavano sistemi di verifica *anti-bot*, il cui superamento ha richiesto il ricorso a soluzioni dedicate per garantire l'accesso ai contenuti.

2.3 Integrazioni e controlli

Il perfezionamento delle *pipeline* è proceduto in parallelo per i due moduli — siti *internet* ed esposti — secondo un approccio di sviluppo incrementale. In tale ambito è stata svolta una specifica attività di *prompt tuning*: una volta sviluppata la *pipeline* di estrazione dei contenuti, è stato definito un *prompt* di base, successivamente affinato nel tempo sulla base dei *feedback* forniti dall'operatore, sia per i siti *internet* sia per gli esposti. Tale attività ha dovuto tenere conto anche del rilascio di nuove versioni dei *large language model*, che hanno richiesto specifici aggiustamenti. Parallelamente, a partire dal *subset* originario, il set di *keyword* impiegato per la verifica dell'operatività dei siti è stato rivisto e ampliato con cadenza settimanale, sulla base dei medesimi riscontri[47].

Il sistema è stato distribuito secondo un ciclo di rilasci settimanali, finalizzati alla verifica delle nuove funzionalità, all'individuazione di eventuali *bug* e al controllo delle mitigazioni relative ai *bug* riscontrati nelle precedenti sessioni di test. I test sono stati condotti da un operatore di CONSOB. A supporto del processo di test era stato predisposto un sistema di segnalazione dell'esito, mediante il quale l'operatore indicava se i contenuti fossero stati processati correttamente; in caso negativo, un apposito campo note consentiva di descrivere nel dettaglio gli errori riscontrati.

2.4 Risultati *pipeline* unificata

Raggiunta una versione (sperimentale) stabile di entrambi i moduli, le *pipeline* sono state unificate, rendendo possibile, a partire da un esposto, l'analisi dei siti *internet* segnalati dagli utenti.

47 In questo percorso incrementale, le attività tecniche di industrializzazione hanno contribuito anche al progressivo irrobustimento delle componenti tecnologiche, affinamento dei criteri di selezione delle pagine e miglioramento della qualità dei contenuti acquisiti a supporto dell'analisi automatizzata. Sul piano architetturale, il sistema è stato ulteriormente migliorato rendendo l'esecuzione delle analisi asincrona e introducendo la possibilità di avviare più analisi in parallelo. Tale intervento, che ha richiesto una revisione dell'architettura applicativa, consente di aumentare significativamente la capacità di elaborazione complessiva, riducendo i tempi di attesa per l'operatore e permettendo di gestire in modo più efficiente i picchi di carico legati all'afflusso di nuove segnalazioni.

La *pipeline* completa è stata testata su un set di 89 esposti, di cui 4 privi di siti *internet* associati. Gli esiti dei test hanno evidenziato che in questa fase il sistema ha rilevato la presenza di sistemi *anti-bot* in 16 casi, per il cui superamento è stato impiegato *BrightData*, che ha consentito di processare la totalità dei siti. L'estrazione del testo è avvenuta correttamente nella totalità degli esposti, consentendo la generazione di 85 *report* completi di analisi a supporto dell'attività dell'operatore. Inoltre, in tutti i casi è stato possibile rilevare la presenza della lingua italiana e acquisire gli *screenshot* delle pagine rilevanti del sito.

In particolare, lo strumento ha mostrato notevole affidabilità e stabilità nei diversi test condotti, generando risposte coerenti e precise anche in presenza di *input* identici ripetuti. Inoltre, nei casi più complessi, sebbene siano state rilevate minime discrepanze che non hanno compromesso l'efficacia complessiva della soluzione, il confronto tra i risultati dello strumento e quelli dell'analisi manuale ha *dimostrato* che lo strumento è in grado di replicare con precisione le decisioni dell'operatore umano, confermando la sua utilità in applicazioni pratiche.

2.5 Metriche di valutazione del prototipo su analisi esposti

La valutazione delle prestazioni del prototipo si fonda sul confronto, per ciascun caso, tra la classificazione prodotta automaticamente e quella di riferimento effettuata da un operatore.

Per il confronto, sono stati definiti i seguenti criteri di valutazione:

- Accuratezza: misura la precisione con cui lo strumento ha valorizzato i campi dell'analisi, confrontando i risultati con quelli ottenuti manualmente dall'operatore.
- Coerenza: valuta la capacità del sistema di mantenere risposte pertinenti, in particolare nei casi caratterizzati da *prompt* complessi e articolati.
- Affidabilità: verifica la consistenza dei risultati ottenuti su documenti differenti, valutando la replicabilità e uniformità delle prestazioni.
- Stabilità: analizza la capacità dello strumento di restituire lo stesso *output* a fronte di *input* identici, garantendo prevedibilità e uniformità delle prestazioni.

Da tale confronto si distinguono quattro categorie di esiti: i veri positivi, ossia i casi positivi correttamente individuati; i veri negativi, ossia i casi negativi correttamente esclusi; i falsi positivi, ossia i casi negativi erroneamente classificati come positivi; e i falsi negativi, ossia i casi positivi non individuati.

I risultati di seguito riportati riflettono gli esiti delle analisi svolte su infrastruttura locale di test, non proprietaria, condotte su un campione di 78 esposti.

a) Efficienza

metrica	valore
tempo medio di lavorazione assistita / esposto	5,63 s
– di cui estrazione testo / OCR	0,82 s
– di cui estrazione LLM	4,81 s
tempo mediano / esposto	4,39 s

Una volta che sarà completata la fase di industrializzazione del *software* nell'infrastruttura della CONSOB (vedi *infra* § 2.6) sarà possibile svolgere compiute e puntuali valutazioni in merito alle *performance* del *software* e al concreto contributo dallo stesso apportato all'attività di vigilanza. In particolare, verrà verificato: *(i)* il tempo medio di lavorazione dell'esposto e del collegato sito *web*; *(ii)* il tempo medio di durata dell'istruttoria, inteso come lasso temporale che intercorre tra *l'upload* dell'esposto nel *software* e il *download* dei documenti finali (durata *input/output*); *(iii)* il numero di esposti che sarà possibile processare - e che a valle dell'analisi conducono all'adozione dei provvedimenti di competenza - in un determinato periodo di riferimento (settimana/mese) rispetto a quanto viene attualmente fatto nel medesimo arco temporale senza l'ausilio del *software*.

b) Efficacia - estrazione dei campi (LLM)

Precision / *recall* / F1 per tipo di campo (*matching strict*, normalizzato).

Al fine di valutare l'efficacia di estrazione delle informazioni è stata verificata la *precision*[48], la *recall*[49] e calcolata la media armonica di *precision* e *recall*[50].

48 La *precision* esprime l'affidabilità delle attribuzioni positive, tanto più elevata quanto minori sono i falsi positivi. *Precision* risponde alla domanda "quante delle entità estratte dal sistema sono corrette". Esempio: su 10 siti *web* estratti dal testo dell'esposto, 9 corretti → *precision* 0,90. Misura la qualità di ciò che il sistema produce.

49 La *recall* esprime la capacità di copertura, tanto più elevata quanto minori sono i falsi negativi. *Recall* risponde alla domanda "quante delle entità realmente presenti nel documento il sistema ha trovato". Esempio: il documento conteneva 10 siti, il sistema ne ha trovati 8 → *recall* 0,80. Misura la completezza.

50 F1 – media armonica di *precision* e *recall*: un numero solo che bilancia entrambe. Penalizza i casi sbilanciati (un sistema che trova tutto ma sbaglia tutto, o preciso ma incompleto, ha F1 bassa). Valore massimo: 1,00. Esempio: trova 10 siti *internet* nel testo ma li estrae con URL errato.

campo	estratti	*precision*	*recall*	F1
soggetti segnalati	135	0,81	0,83	0,82
siti *web*	153	0,87	0,95	0,91
e-mail	70	0,99	0,97	0,98
nomi contatti	55	1,00	0,92	0,96
metodi primo contatto	30	0,80	0,89	0,84
totale	***443***	***0,88***	***0,91***	***0,89***

E-mail e nomi quasi perfetti; il limite è sulla *precision* dei siti *web*.

c) Efficacia - classificazione e competenza CONSOB

Rilevazione nel testo delle parole presenti nelle liste "attività" e "prodotti". Decisione binaria di competenza (definizioni per regolamento: MiFID = strumento + attività; MiCAR = cripto-attività + attività).

voce	n.	accuratezza	*precision*	*recall*	falsi +	falsi -
rilevazione strumenti finanziari	78	0,86	0,92	0,90	5	6
rilevazione cripto-attività	78	0,99	1,00	0,97	0	1
competenza MiFID	78	0,83	0,92	0,84	4	9
competenza MiCAR	78	0,90	1,00	0,73	0	8
competenza complessiva	78	0,86	1,00	0,85	0	11

Profilo conservativo: *precision* alta, zero falsi positivi; il limite è la sotto-rilevazione (*recall*).

Nel caso della competenza complessiva, per ciascun esposto il prototipo determina se la fattispecie rientri nell'ambito di competenza della CONSOB, ossia se attenga a uno strumento finanziario o a una cripto-attività riconducibili a un'attività rilevante ai sensi della disciplina MiFID o MiCAR, e tale determinazione viene confrontata con la classificazione effettuata dall'operatore.

Sui 78 esposti esaminati l'accuratezza è pari a 0,86. Ciò significa che il *software* ha classificato correttamente (competenza o non competenza) 67 esposti. L'accuratezza esprime la correttezza complessiva dell'estrazione e corrisponde alla quota di elementi corretti sul totale degli elementi considerati nel confronto con il valutatore umano, tenendo conto sia degli elementi estratti in modo errato sia di quelli omessi.

La *precision* è pari a 1,00, la classificazione operata dal *software* è risultata corretta nel 100% degli esposti di competenza. Vuol dire che il *software* non ha mai sbagliato quando ha classificato come di competenza l'esposto esaminato. Un limite è invece emerso nella valutazione di non competenza. Infatti, gli errori sono emersi colo con riferimento alla classificazione degli esposti come non rilevanti (11 falsi negativi).

d) Robustezza / affidabilità

metrica	valore
esecuzioni totali	78
tasso di successo *pipeline*	87,9%
esposti non classificati (insufficienza informativa)	12,1%
stabilità classificazione su *input* identici (5×3 run)	100%
variabilità estrazione su *input* identici (jaccard liste)	0,37

Il 12,1% degli esposti non ha superato la validazione dello schema per insufficienza informativa del documento sorgente, non per errore della *pipeline*: si tratta di esposti che non contenevano elementi sufficienti per procedere alla classificazione. La decisione di rilevanza è 100% stabile su *input* identici; gli elementi estratti variano (temperatura ≠ 0) — da impostare a temperatura = 0 in produzione. Il *crawling* dei siti è modulo distinto, con valutazione di robustezza separata.

e) Accuratezza OCR vs estrazione LLM (qualificazione distinta)

sottoinsieme	n.	fedeltà testo	tempo OCR
PDF nativi (66,7%)	52	98,9% (*recall*-parole oggettivo)	0,29 s
PDF scansionati (33,3%)	26	~80,7% (stima)	1,82 s

Due terzi degli esposti processati sono PDF nativi con testo preservato al 99%. Vale a dire che una minima parte (1%) del testo è risultata talvolta scritta con poco contrasto di colore con lo sfondo (es. colore arancione o celeste dei siti *web* con *link* attivo).

In tutti i casi l'OCR ha estratto informazioni dalle immagini sottoposte all'esame. LLM ha estratto informazioni dall'87,9% degli esposti, in quanto, come detto, il 12,1 degli esposti non ha superato la validazione dello schema per insufficienza informativa del documento sorgente.

f) Tipologia di errori

tipologia	conteggio	%
informazione corretta	1260	54,3%
informazione non estratta	524	22,6%
errata / attribuzione sbagliata	444	19,1%
parziale / troncato	63	2,7%
allucinazione LLM (no OCR)	22	0,9%
formato / normalizzazione	6	0,3%

2.6 Industrializzazione nell'architettura CONSOB

I risultati confermano dunque il valore applicativo dello strumento come supporto valido per automatizzare compiti complessi nell'ambito dell'attività istruttoria, consentendo di ridurre l'intervento umano senza intaccare la qualità dell'analisi. Proprio tale potenzialità, tuttavia, rende necessario interrogarsi sulle condizioni tecniche, organizzative e regolatorie entro le quali strumenti di questo tipo possano essere impiegati in modo affidabile e coerente con la funzione di vigilanza.

Allo stato, il *software* è in fase di industrializzazione nell'architettura infrastrutturale della CONSOB e prima del pieno rilascio in esercizio è prevista l'implementazione di ulteriori funzionalità finalizzate a migliorare l'attività di estrazione di alcune informazioni presenti nei siti *internet* (es. individuazione della lingua italiana). Parallelamente, è stato inserito nel Piano informatico 2026 un progetto che prevede lo studio e la realizzazione di un sistema di valutazione dei risultati cui, allo stato, perviene il *software*. La soluzione evolutiva ipotizzata si basa su un sistema di *feedback* fornito dagli utenti durante le differenti fasi di analisi sugli esiti dell'esame degli esposti e dei siti *web,* così da creare un *dataset* di informazioni atte a favorire la costruzione di modelli che consentano al *software* di migliorare, attraverso meccanismi di apprendimento supervisionato, gli esiti delle analisi.

3 Analisi delle limitazioni attuali delle tecnologie IA

Gli algoritmi di intelligenza artificiale implementati nello strumento in esame si fondano su quattro tecnologie cardine: OCR, *web crawling*, analisi semantica e LLM. L'utilizzo di analisi semantica rispetto all'utilizzo di LLM offre un vantaggio significativo in termini di *explainability*[51], in quanto il controllo è basato sulla presenza e sul ruolo nella frase di un *set* di *keyword* preimpostate dall'operatore, che da un lato necessita di aggiornamenti, ma dall'altro permette la verifica puntuale delle attività indicate.

In netto contrasto, gli LLM presentano sfide considerevoli in termini di *explainability*[52]. Questi modelli, caratterizzati da una complessità intrinseca derivante dall'enorme numero di parametri, nell'ordine di miliardi, operano in modo di difficile interpretabilità per un operatore. Il funzionamento di tali modelli si basa su trasformazioni non lineari e complesse, che complicano ulteriormente la comprensione del processo che ha portato all'*output* e creando una significativa barriera alla spiegazione del risultato dell'inferenza. A tale opacità si aggiunge la forte dipendenza dai dati di addestramento: le prestazioni e la capacità di generalizzare di questi modelli sono strettamente legate alla qualità, quantità e diversità dei dati su cui sono stati addestrati. A tale opacità si lega inoltre il fenomeno delle "allucinazioni".

Un'ulteriore limitazione dei LLM riguarda la gestione di finestre di contesto molto estese: con l'aumentare della lunghezza del *prompt*, la loro efficacia tende infatti a degradare, riducendo la coerenza dell'*output* e con conseguente diminuzione della capacità di mantenere rilevanza per il contesto a lungo termine. Questo fenomeno, legato alla difficoltà dei modelli nel mantenere un'attenzione efficace sull'intero contenuto, si combina con un'altra sfida tipica dei LLM non specializzati per compiti specifici: le prestazioni e la capacità di generalizzazione dipendono fortemente dal numero di parametri, oltre che dalla qualità e quantità dei dati di addestramento. Tali modelli richiedono infatti risorse computazionali considerevoli, in particolare memoria, generalmente disponibili solo su infrastrutture specializzate. Ne consegue che l'utilizzo *on-premise* richiede infrastrutture specializzate, favorendo il ricorso a servizi *cloud* gestiti, i quali introducono ulteriori complessità legate a latenza, sicurezza dei dati e costi operativi. L'emergere di modelli *foundation* eseguibili su *hardware* proprio — come gpt-oss di OpenAI[53] eseguibile su singola *GPU*

51 L'*explainability* si riferisce alla capacità di un modello di intelligenza artificiale (IA) di fornire interpretazioni comprensibili del suo funzionamento e delle sue decisioni. In altre parole, è la qualità che consente agli esseri umani di comprendere come e perché un modello di IA arriva a una determinata conclusione, rendendo trasparenti i processi decisionali e permettendo di valutare la correttezza e l'affidabilità del modello.

52 Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018), A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys, 51(5), Art. 93, 1-42.

53 OpenAI (2025b), gpt-oss-120b & gpt-oss-20b Model Card, arXiv:2508.10925, https://doi.org/10.48550/arXiv.2508.10925

consumer — unitamente all'evoluzione di tecniche di quantizzazione[54] dei modelli, sta progressivamente ridimensionando questo *trade-off*, consentendo di eseguire modelli di dimensioni significative anche su *hardware* non specialistico. Resta aperto, sul piano critico, il tema dei rischi connessi all'impiego di LLM di grandi dimensioni, evidenziati in letteratura: gli ingenti costi computazionali e ambientali dell'addestramento, l'opacità e la scarsa documentabilità dei dati su cui i modelli sono addestrati (con il connesso rischio di incorporare e amplificare *bias* e contenuti discriminatori) e la capacità di generare testi fluenti ma privi di ancoraggio fattuale, potenzialmente sfruttabili per la produzione di disinformazione[55].

Accanto a tali problematiche, anche l'OCR presenta limitazioni che possono incidere sulla qualità complessiva dello strumento. Le *performance* degli algoritmi di riconoscimento ottico dei caratteri dipendono fortemente dalla qualità del contenuto da processare: immagini con bassa risoluzione, documenti scansionati con rumore visivo, distorsioni grafiche o testi scritti in caratteri non standard possono compromettere la precisione del riconoscimento. Inoltre, la presenza di formati eterogenei, *layout* complessi o combinazioni di lingue diverse introduce ulteriori difficoltà nell'estrazione corretta del testo. Queste criticità possono tradursi in errori sistematici che, se non adeguatamente rilevati e corretti, rischiano di propagarsi nelle fasi successive dell'analisi, compromettendo l'accuratezza complessiva del processo e riducendo l'affidabilità delle informazioni estratte. Le criticità non riguardano soltanto i singoli moduli, ma anche la qualità e la natura delle fonti trattate. Gli esposti possono essere incompleti, disomogenei o rumorosi, mentre i siti *web* costituiscono una fonte auto-rappresentativa: le informazioni dichiarate dal soggetto, comprese le eventuali autorizzazioni vantate, non sono di per sé attendibili e richiedono un riscontro esterno. Il sistema estrae fedelmente quanto presente nel contenuto, ma la distinzione tra dichiarazioni veritiere e ingannevoli resta affidata alla verifica successiva e al giudizio dell'operatore. A ciò si aggiunge la variabilità dei contenuti *web*, particolarmente accentuata nel contrasto agli abusivismi: i siti analizzati hanno spesso natura effimera o deliberatamente elusiva (contenuti caricati dinamicamente, pagine che mutano rapidamente, testo veicolato tramite immagini per ostacolarne l'estrazione, accorgimenti volti a mostrare contenuti diversi a un utente e a un sistema automatico), con conseguente riduzione della completezza e dell'affidabilità del contenuto acquisito.

54 La quantizzazione è una tecnica di compressione che riduce la precisione numerica con cui sono memorizzati i parametri di un modello di intelligenza artificiale (ad esempio, da numeri in virgola mobile a 16 bit a interi a 4 bit), riducendone significativamente l'occupazione di memoria a fronte di una perdita di accuratezza generalmente contenuta. Tale tecnica consente di eseguire modelli di grandi dimensioni anche su hardware con risorse limitate.

55 Bender, EM., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610-623.

Infine, trattandosi di una *pipeline* che concatena moduli eterogenei, un'imprecisione a monte (nell'OCR, nell'acquisizione del contenuto *web* o nell'estrazione) può propagarsi alle fasi successive e condizionarne l'esito: l'affidabilità complessiva dipende dunque dalla tenuta dell'intera catena e dalla capacità di intercettare e correggere gli errori prima che si consolidino.

I limiti qui richiamati — in termini di spiegabilità, rischio di allucinazioni, gestione di contesti estesi, modalità di deployment e costi infrastrutturali — non rilevano, tuttavia, soltanto sul piano tecnico. In un contesto istituzionale come quello della vigilanza, essi incidono direttamente sulle condizioni di impiego del sistema, sulla configurazione dei controlli umani, sulla tracciabilità delle operazioni e, più in generale, sulla compatibilità dell'uso dell'IA con i principi di legalità, trasparenza, proporzionalità e *accountability*. È da questa prospettiva che prende avvio la riflessione sviluppata nella Parte Quarta.

Uno sguardo d'insieme. IA e vigilanza: dal caso applicativo ai presidi di utilizzo

PARTE QUARTA

1 Osservazioni a partire dal prototipo

Il prototipo sviluppato nelle Parti Seconda e Terza consente di osservare, su un caso concreto, il contributo che tecniche diverse – OCR, *web crawling*, analisi semantica e LLM – possono offrire per trattare in modo più sistematico materiali istruttori eterogenei. La sperimentazione conferma che la tecnologia, da sola, non basta. Conta il modo in cui i diversi moduli si innestano nel processo di vigilanza.

L'impiego di strumenti di intelligenza artificiale nei processi di vigilanza solleva questioni che non si possono ricondurre alla mera efficienza tecnica. Il tema riguarda il modo in cui l'Autorità costruisce la propria base conoscitiva – raccoglie, organizza e valuta gli elementi informativi necessari ai fini dell'istruttoria – in un contesto segnato da una crescente complessità dei fenomeni osservati e dall'esigenza di preservare presidi essenziali di legalità, proporzionalità, trasparenza, responsabilità e controllo umano[56].

In questa prospettiva, tali presidi non dipendono soltanto dalle capacità del modello impiegato, ma anche dal lavoro tecnico di progettazione applicativa, integrazione tra moduli, gestione delle dipendenze tecnologiche e governo della *pipeline* svolte nell'ambito delle attività tecniche di industrializzazione chiamate a rendere il sistema utilizzabile nel contesto infrastrutturale dell'Istituto.

Le sperimentazioni *SupTech* avviate in CONSOB possono essere lette in questa ottica. Nei casi finora sviluppati l'intelligenza artificiale non è concepita come un sostituto dell'analista umano, né come uno strumento destinato ad automatizzare integralmente la decisione amministrativa. Più realisticamente, l'IA viene impiegata per alleggerire alcune fasi del lavoro istruttorio, trattare grandi volumi di dati eterogenei, far emergere ricorrenze e rendere più ordinata la ricostruzione di elementi di fatto che, altrimenti, richiederebbero tempi molto più lunghi[57].

56 Kaplan A., Haenlein M., *Op. cit.* pp. 15-25.

57 Sul punto, v. anche Artificial Intelligence in Italian Financial Markets, pubblicato da OECD, che insiste sulla necessità di accompagnare l'adozione dell'IA con presidi di *governance*, qualità del dato, *accountability* e attenzione alle dipendenze da terzi.

Questa impostazione trova riscontro anche in alcuni contributi recenti della collana Quaderni FinTech. In particolare, il Quaderno n. 15 ha sottolineato il carattere di supporto di questi strumenti alle analisi condotte dagli Uffici e la centralità della valutazione umana; il Quaderno n. 14, dedicato al prototipo CONSOB–Università di Trento per l'individuazione di possibili fenomeni di *greenwashing* nei *green bond*, ha mostrato come il valore dell'innovazione tecnologica dipenda soprattutto dalla sua capacità di integrarsi in processi di vigilanza già strutturati e verificabili[58] [59].

Il prototipo sviluppato per supportare l'azione di contrasto agli abusivismi finanziari si inserisce nello stesso percorso. La combinazione di OCR, *web crawling*, analisi semantica, modelli linguistici di grandi dimensioni e regole condizionali consente di affrontare in modo più sistematico materiali molto eterogenei, come esposti, allegati, contenuti presenti sui siti *internet* e altre evidenze digitali riconducibili ai soggetti segnalati. Proprio per questa natura ibrida il modello è interessante poiché consente di misurare, nello stesso tempo, il potenziale applicativo di queste tecnologie e i limiti che ancora ne condizionano l'impiego in un contesto istituzionale. Tra questi rientrano l'opacità dei modelli, il rischio di *hallucinations*, la qualità dei dati, la difficoltà di trattare contesti molto estesi e, più in generale, la necessità di rendere verificabili gli *output* prodotti dal sistema.

Su questi aspetti incide oggi anche il mutato contesto regolatorio. L'entrata in vigore del Regolamento (UE) 2024/1689 e l'adozione, sul piano nazionale, della Legge 23 settembre 2025, n. 132 inducono ad andare oltre alla domanda se il sistema funzioni; deve essere chiarito entro quali limiti e con quali presidi esso possa essere utilizzato in modo coerente con le attività di vigilanza.

Questa sezione del lavoro sviluppa quindi tre passaggi. Il primo riguarda i limiti tecnico-operativi delle tecnologie impiegate nei sistemi ibridi di analisi documentale e informativa. Il secondo riguarda la compatibilità di tali strumenti con il nuovo quadro europeo e nazionale, con specifico riferimento all'AI Act e alla Legge italiana sull'intelligenza artificiale. Il terzo richiama i presidi che stanno progressivamente delineando, anche in CONSOB, un *framework* interno di adozione responsabile dell'IA, fondato su sperimentazione controllata, centralità del fattore umano, qualità del dato, tracciabilità e integrazione con i processi di vigilanza esistenti.

58 Deriu P., Racioppi S., *Op. Cit.*

59 Paterlini S., Nicolodi A., Gentile M., Foglia Manzillo V., Sancilio M.R., Deriu P., *Op. cit.*

2 L'IA come infrastruttura conoscitiva della vigilanza

Le applicazioni dell'intelligenza artificiale in ambito finanziario sono ormai numerose e riguardano ambiti anche molto diversi tra loro, dall'analisi del rischio al monitoraggio delle transazioni, dall'automazione documentale all'individuazione di anomalie.

Quando queste tecnologie vengono impiegate nell'esercizio di funzioni pubbliche, la questione cambia natura. Non si tratta più soltanto di misurarne l'efficienza, ma di capirne l'incidenza sul modo in cui si forma, si organizza e si giustifica la conoscenza amministrativa.

Le esperienze più recenti maturate in CONSOB offrono un punto di osservazione interessante. I prototipi sviluppati per l'individuazione di *pattern* anomali nell'*insider trading*, per l'estrazione di informazioni dai KID, per l'analisi automatizzata della documentazione relativa ai *green bond* o, nel caso illustrato in questo lavoro, per il contrasto agli abusivismi finanziari, hanno finalità diverse ma una medesima impostazione di fondo. L'IA si rivela utile quando rafforza la capacità di analisi dell'Autorità, non quando pretende di sostituire il giudizio dell'operatore. Il suo contributo è più promettente quanto più viene impiegata per trattare passaggi specifici del processo – acquisizione, normalizzazione, estrazione, classificazione preliminare, generazione di *alert* – lasciando all'intervento umano la verifica, l'interpretazione e la decisione finale[60].

Anche il dibattito internazionale va in questa direzione. Il rapporto *Artificial Intelligence in Italian Financial Markets*, pubblicato dall'OECD nel 2026, evidenzia che l'adozione dell'IA nel settore finanziario richiede non solo sperimentazione tecnologica, ma anche adeguati presidi di *governance*, qualità del dato, *accountability*, auditabilità e attenzione alle dipendenze da terzi. Nella stessa direzione si colloca il *paper Regulatory approaches to Artificial Intelligence in finance*, che sottolinea come l'uso dell'IA in finanza debba essere valutato alla luce dei quadri di *policy*, dei presidi di rischio e delle responsabilità istituzionali che ne accompagnano l'impiego.

Anche il prototipo descritto nelle Parti 2 e 3 del presente lavoro, dedicato al contrasto degli abusivismi finanziari, può essere letto come un'evoluzione dell'infrastruttura conoscitiva dell'Autorità. Non automatizza la vigilanza in senso pieno, ma automatizza alcune operazioni conoscitive che ne costituiscono il presupposto. Ciò si traduce in operazioni molto concrete, quali la lettura di testi non strutturati, l'individuazione di URL pertinenti, l'acquisizione di contenuti *web*, la classificazione preliminare, la verifica

60 https://fa000000125.resources.office.net/033f92d3-bc6d-439a-858a-a17acf70360a/1.0.2605.18012/en-us_web/taskpane.html?_host_Info=Word$Win32$16.01$it-IT$$$$19#user-content-fn-5

rispetto a registri e fonti esterne, la predisposizione di bozze e *report* a supporto dell'istruttoria. Il valore dello strumento risiede nella capacità di rendere più scalabile e tempestiva una parte del lavoro istruttorio, senza alterarne la natura essenziale. In questa capacità si riflettono le attività tecniche di industrializzazione con cui sono progettate, integrate e manutenute le componenti applicative, le *pipeline* di trattamento del dato e gli ambienti di esecuzione necessari a trasformare un prototipo in uno strumento operabile, controllabile e sostenibile nel tempo.

3 Limiti tecnico-operativi

3.1 Spiegabilità, opacità e razionalizzazione *ex post*

I limiti emersi nell'analisi del prototipo non riguardano solo la *pipeline* tecnica. In un contesto di vigilanza incidono direttamente sulle condizioni di impiego istituzionale dello strumento. In altre parole, l'affidabilità tecnica non può essere considerata come una questione isolata, perché si intreccia inevitabilmente con profili di responsabilità, verificabilità e tenuta organizzativa.

Uno dei nodi più delicati riguarda la spiegabilità dei sistemi impiegati. Nei moduli basati su regole, *keyword* o logiche condizionali, il percorso che conduce al risultato è in larga misura ricostruibile dato che è possibile risalire alla regola attivata, al termine rilevato, al passaggio logico che ha determinato l'esito. Il livello di trasparenza è quindi relativamente elevato, anche se richiede manutenzione continua di dizionari, regole e criteri di classificazione.

Diverso è il caso dei modelli linguistici di grandi dimensioni. Qui l'*output* deriva da processi interni complessi, non facilmente traducibili in una sequenza inferenziale immediatamente intelligibile per l'analista umano. Anche quando il modello accompagna la risposta con una giustificazione testuale, quella giustificazione non coincide con la spiegazione del percorso seguito. È piuttosto una razionalizzazione *ex post*, utile per orientare la verifica dell'operatore, ma insufficiente a fondare da sola la comprensibilità dell'esito[61]. In un contesto istituzionale, questo limite non è secondario, perché la comprensibilità dell'esito contribuisce alla legittimazione stessa dell'uso della tecnologia.

61 Guidotti R. *et al.*, *Op. cit.,* 51(5), art. 93; Ribeiro M.T., Singh S., Guestrin C., *Op. cit.*, pp. 1135-1144; Lundberg S.M., Lee S.I., *Op. cit.* pp. 4765-4774.

3.2 Allucinazioni e affidabilità dell'estrazione informativa

Un secondo profilo di criticità riguarda il fenomeno delle allucinazioni. Nei sistemi generativi, l'errore più insidioso non è sempre quello macroscopico. Più spesso il problema nasce nella generazione di risultati plausibili, ben costruiti e apparentemente coerenti, che però non trovano un adeguato riscontro nelle evidenze disponibili [62].

Nel contrasto agli abusivismi finanziari, questo rischio è particolarmente sensibile. Un'estrazione errata può riguardare elementi essenziali (autorizzazioni dichiarate, attività svolte, URL pertinenti, riferimenti a prodotti o servizi, contatti, nomi di soggetti o collegamenti tra fonti diverse). Per questo l'affidabilità del sistema non dipende solo dal modello utilizzato; contano la struttura e la qualità del *prompt*, il formato dell'*output*, le procedure di validazione *post-inferenza* e, soprattutto, la possibilità di verificare le evidenze a monte del risultato prodotto.

3.3 Contesti lunghi, *chunking* e perdita di informazione

Un ulteriore profilo critico emerge quando l'*input* documentale è molto ampio o particolarmente frammentato. Sebbene i modelli di frontiera dispongano oggi di finestre di contesto molto più estese rispetto al passato, la possibilità tecnica di processare testi lunghi non equivale alla garanzia che tutte le informazioni rilevanti siano considerate con la stessa accuratezza. L'aumento della lunghezza del contesto non elimina del tutto il rischio di perdita di attenzione su passaggi intermedi, marginali o meno salienti del documento[63].

La gestione di siti *web* complessi o di esposti corredati da numerosi allegati richiede allora strategie di segmentazione del contenuto, seguite da una fase di riconciliazione dei risultati ottenuti sui singoli blocchi. Questa scelta, tuttavia, introduce un *trade-off* non trascurabile tra accuratezza, costo, latenza e rischio di perdita o duplicazione di informazione. Da qui l'esigenza di *pipeline* progettate in modo selettivo, nelle quali *crawling*, *pre-filtering*, tecniche di *retrieval* e moduli LLM siano combinati in funzione del tipo di contenuto da trattare.

62 Ji Z. *et al.*, Op. *cit.* 55 (12), art. 248.

63 Liu N.F. *et al.*, Lost in the Middle: How Language Models Use Long Contexts, in Transactions of the Association for Computational Linguistics, (2024), 12, pp. 157-173.

3.4 Propagazione dell'errore e robustezza della *pipeline*

Nel prototipo in esame, il risultato finale dipende dall'interazione tra moduli eterogenei. Questo comporta un tipico rischio di propagazione dell'errore: un'imprecisione nella fase di OCR può compromettere l'estrazione semantica successiva; un contenuto *web* rumoroso o mal normalizzato può falsare l'*input* fornito al modello; un'estrazione iniziale non corretta può condizionare le classificazioni o le bozze documentali prodotte a valle. L'errore, quindi, non resta necessariamente confinato al punto in cui si genera.

La qualità del sistema va quindi valutata sull'intera catena, non sul singolo componente. Diventano essenziali la standardizzazione degli *input*, il controllo delle trasformazioni intermedie, il *logging* e la gestione delle anomalie e la possibilità di risalire, quando necessario, al passaggio che ha originato un determinato esito.

Si tratta di profili che rinviano a un lavoro tecnico continuativo di *feedback*, monitoraggio, *versioning*, gestione delle eccezioni e affinamento dei singoli moduli, senza il quale l'affidabilità del sistema resterebbe esposta a fragilità difficilmente governabili in sede operativa.

3.5 *Deployment*, dipendenze tecnologiche e costi di *governance*

Anche le modalità di *deployment* incidono sulla valutazione complessiva dello strumento. I servizi *cloud* gestiti consentono oggi, in molti casi, di accedere a modelli più avanzati, a finestre di contesto più ampie e ad aggiornamenti più rapidi; al tempo stesso introducono dipendenze da fornitori terzi e sollevano questioni di sicurezza, *governance* dei dati, continuità operativa e *auditability*.

Le soluzioni *on-premise* o basate su modelli *open-weight* offrono un maggiore controllo sull'ambiente di esecuzione e possono ridurre alcuni rischi di esternalizzazione, ma richiedono investimenti infrastrutturali, competenze tecniche e valutazioni costi-benefici adeguate. La scelta del sistema non può essere formulata in astratto ma deve dipendere dalla funzione del sistema, dalla sensibilità dei dati trattati e dal livello di affidabilità richiesto dal processo di vigilanza.

Tale valutazione comparativa tra architetture, ambienti di esecuzione, modelli di servizio e requisiti di sicurezza costituisce uno degli ambiti in cui il contributo tecnico all'industrializzazione assume rilievo, in quanto condiziona la possibilità stessa di impiegare l'IA in modo compatibile con le esigenze di continuità operativa, auditabilità e controllo istituzionale.

A questa valutazione iniziale deve affiancarsi un presidio del sistema lungo l'intero ciclo di vita. Nel caso di un'applicazione come quella qui

esaminata, il passaggio dal prototipo all'uso operativo non coincide con il solo rilascio tecnico. Occorre sapere quali modelli, *prompt*, regole condizionali, dizionari e componenti di *pipeline* sono effettivamente in uso, quale versione sia stata impiegata in un dato momento e chi sia responsabile del relativo aggiornamento. Lo stesso vale per le modifiche successive, perché l'affinamento di un *prompt*, l'aggiornamento di un dizionario, la sostituzione di un modello o la correzione di una regola possono incidere sugli *output* prodotti e sulla loro confrontabilità nel tempo.

Prima del rilascio in esercizio, e poi in occasione degli aggiornamenti più significativi, il sistema dovrebbe essere testato su casi rappresentativi e documentati. Anche dopo il rilascio, il presidio non si esaurisce. Diventa necessario monitorare la stabilità delle prestazioni, intercettare eventuali fenomeni di *drift* – intesi come scostamenti progressivi delle prestazioni rispetto ai risultati attesi o alle versioni precedenti –, conservare evidenza delle configurazioni utilizzate, degli *output* generati e degli interventi di correzione o validazione effettuati dall'operatore, gestire anomalie e incidenti, e prevedere momenti periodici di revisione. Se una modifica riduce l'affidabilità del sistema o produce risultati non coerenti con le attese, devono essere disponibili criteri per sospenderne l'uso, tornare a una versione precedente o dismettere la componente interessata.

I costi di *governance*, dunque, non si concentrano nella fase iniziale di sviluppo, ma accompagnano l'intera vita dello strumento. Questa prospettiva trova riscontro anche nel rapporto OECD *Artificial Intelligence in Italian Financial Markets*, che collega l'adozione dell'IA nel settore finanziario alla qualità del dato, alla tracciabilità, al controllo delle dipendenze tecnologiche e alla responsabilità organizzativa.

4 Compatibilità con l'AI Act e con la Legge italiana

Nel caso del prototipo qui esaminato, la questione della compatibilità con il quadro regolatorio non riguarda un sistema destinato a sostituire la decisione amministrativa, ma uno strumento che interviene in fasi istruttorie, di selezione, classificazione preliminare ed elaborazione documentale. È questa collocazione funzionale, più che la tecnologia in sé, a orientare la valutazione del regime giuridico applicabile.

Una qualificazione astratta rischierebbe di essere poco utile. La compatibilità di strumenti come quello in esame con la disciplina europea e nazionale in materia di intelligenza artificiale dipende dalle concrete modalità di progettazione, impiego, supervisione e integrazione organizzativa del sistema nel processo di vigilanza.

La valutazione dovrebbe concentrarsi sul ruolo concreto del sistema, quanto incide sull'attività dell'analista, quali controlli sono previsti, quali dati tratta, quali passaggi restano tracciabili e chi ne presidia il ciclo di vita.

4.1 Il quadro dell'AI Act

Il Regolamento (UE) 2024/1689 prevede un'applicazione progressiva delle proprie disposizioni. Dopo l'entrata in vigore del 2024, dal 2 febbraio 2025 trovano applicazione i divieti e gli obblighi di alfabetizzazione in materia di IA; dal 2 agosto 2025 si applicano le regole sui modelli di uso generale e sulla *governance*; dal 2 agosto 2026 entra poi in vigore una parte rilevante della disciplina relativa ai sistemi *high-risk* e agli obblighi di trasparenza[64] [65].

Nel caso di un prototipo come quello qui considerato, non sembra opportuno anticipare una qualificazione rigida in termini di *high-risk* senza una valutazione puntuale del *deployment* e della funzione effettivamente svolta. Utile, sul piano analitico, è osservare che il sistema appare orientato principalmente a una funzione di supporto istruttorio, selezione e prioritizzazione, e non all'adozione di decisioni interamente automatizzate. La sua qualificazione regolatoria va quindi verificata alla luce del ruolo concreto che esso assume nel processo.

Anche ove non ricada formalmente in una categoria ad alto rischio, il sistema dovrebbe comunque essere progettato seguendo i principi di gestione del rischio, di qualità e *governance* dei dati, di documentazione, di tracciabilità, supervisione umana, di robustezza e sicurezza, definiti dal regolamento. L'adozione di presidi ispirati a questi principi non rappresenta solo un adempimento eventuale, ma costituisce un criterio di buona amministrazione tecnologica.

Pur ritenendo poco opportuno in questa sede operare una netta classificazione in termini di rischio del prototipo in esame si ritiene utile richiamare il considerando n. 58) del Regolamento (UE) 2024/1689, secondo il quale «***i sistemi di IA previsti dal diritto dell'Unione al fine di individuare frodi nell'offerta di servizi finanziari*** *e a fini prudenziali per calcolare i requisiti patrimoniali degli enti creditizi e delle imprese assicurative* ***non dovrebbero essere considerati ad alto rischio»***. Questo riferimento non esaurisce la valutazione del

64 Per il quadro temporale di applicazione del regolamento, v. Regolamento (UE) 2024/1689 e la documentazione normativa europea collegata.

65 Molto recentemente, nell'ambito del cosiddetto "Digital Omnibus" in materia di IA, il Consiglio e il Parlamento europeo hanno raggiunto il 7 maggio 2026 un accordo politico provvisorio volto, tra l'altro, a differire l'applicazione di parte degli obblighi relativi ai sistemi di IA ad alto rischio: la nuova data sarebbe il 2 dicembre 2027 per i sistemi *high-risk stand-alone* e il 2 agosto 2028 per i sistemi *high-risk* incorporati in prodotti, fermo restando che, sino al completamento dell'iter formale di adozione e pubblicazione, il testo vigente dell'AI Act continua a costituire il riferimento giuridico applicabile.

prototipo, ma offre un elemento utile per inquadrarne la funzione e la possibile portata degli obblighi applicabili.

4.2 La Legge italiana n. 132 del 2025

Sul piano nazionale, la Legge 23 settembre 2025, n. 132, recante *Disposizioni e deleghe al Governo in materia di intelligenza artificiale*, promuove un impiego corretto, trasparente e responsabile dell'IA in una prospettiva antropocentrica e stabilisce espressamente che le proprie disposizioni si interpretano e si applicano conformemente al regolamento europeo. Gli artt. 1 e 3 richiamano, tra gli altri, i principi di trasparenza, proporzionalità, sicurezza, protezione dei dati personali, riservatezza, accuratezza, non discriminazione e sostenibilità, e sottolineano la necessità che i sistemi e i modelli di IA siano sviluppati e applicati nel rispetto dell'autonomia e del potere decisionale dell'uomo, assicurando sorveglianza e intervento umano.

Applicata al prototipo qui analizzato, questa impostazione suggerisce tre implicazioni. In primo luogo, lo strumento resta compatibile con il quadro nella misura in cui mantiene un ruolo di assistenza, e non di sostituzione, rispetto all'analista umano. In secondo luogo, deve restare ferma la riconducibilità della decisione rilevante sul piano amministrativo o istruttorio a una persona fisica, dotata del potere e della responsabilità di valutare criticamente gli *output* del sistema. In terzo luogo, devono essere tracciabili le fonti, i *prompt*, le versioni dei modelli, le correzioni umane e i passaggi che hanno condotto alla classificazione o all'estrazione di una data informazione. Infine, il sistema deve essere collocato in un assetto organizzativo che consenta verifiche *ex ante* ed *ex post*, auditabilità e controllo interno.

La Legge, peraltro, non si limita a introdurre cautele. Essa riconosce anche il ruolo dell'IA come leva per migliorare l'efficienza dell'azione pubblica, ridurre i tempi dei procedimenti e accrescere la qualità dei servizi resi a cittadini e imprese. Per un'Autorità di vigilanza, questo significa che l'adozione di strumenti di IA non va letta solo in chiave difensiva, ma come parte di un percorso di innovazione organizzativa da governare con adeguati presidi.

5 Un *framework* interno di adozione responsabile dell'IA

Le esperienze già maturate in CONSOB, insieme alla riflessione sviluppata nei più recenti Quaderni FinTech, consentono di intravedere il progressivo consolidarsi di un *framework* interno di adozione responsabile dell'IA. Non si tratta ancora di uno schema chiuso, ma di un insieme di criteri ricorrenti, quali la sperimentazione controllata, l'impiego dell'IA in funzione di supporto e non di sostituzione dell'analista, l'attenzione alla qualità e

all'affidabilità del dato, la combinazione tra tecniche deterministiche e modelli più complessi, la verifica della compatibilità con il quadro normativo vigente e lo sviluppo di competenze interdisciplinari. In tale quadro, il *framework* non si esaurisce in principi generali o in presidi giuridico-organizzativi, ma comprende anche componenti tecnologiche, ambienti di sviluppo e collaudo, criteri di rilascio, controlli sugli *input* e sugli *output*, nonché scelte architetturali e infrastrutturali coerenti con il livello di rischio e con la funzione svolta dal sistema.

Questa impostazione è coerente con i principi di legalità e di imparzialità e buon andamento dell'attività amministrativa; l'obiettivo è evitare che l'impiego dell'IA si traduca in processi decisionali automatizzati o poco controllabili che potrebbero porre rischi di "arbitrarietà" dell'attività amministrativa.

In tal senso, il ricorso a sistemi di IA va inteso come strumento organizzativo e ausiliario. Le soluzioni adottate non comportano una delega della funzione decisoria dell'amministrazione ma supportano le strutture responsabili nello svolgimento di attività preliminari.

Coerentemente con tale impostazione, gli output dei sistemi di IA non sostituiscono la valutazione dell'ufficio. Forniscono elementi informativi che il funzionario può utilizzare, insieme ad altre evidenze, nell'ambito del procedimento.

In termini più generali, anche le criticità emerse nel caso del prototipo illustrato in questo lavoro – dall'eterogeneità delle fonti alla necessità di validazione umana, dalla dipendenza dalla qualità del dato all'esigenza di tracciabilità – permettono di individuare alcuni presidi essenziali per un'adozione responsabile dell'IA nei processi di vigilanza.

In termini operativi, questi presidi possono essere ricondotti a quattro dimensioni.

Sul piano funzionale, il sistema deve avere una finalità chiaramente delimitata e proporzionata, orientata al supporto dell'istruttoria, della selezione dei casi o della definizione delle priorità di analisi, senza assumere in via autonoma decisioni amministrative o valutazioni para-decisionali suscettibili di incidere direttamente sulla posizione dei destinatari.

Sul piano informativo, occorre assicurare la qualità delle fonti, il controllo del rumore, la gestione degli errori di OCR, la corretta acquisizione dei contenuti dinamici, la documentazione delle trasformazioni del dato e la verificabilità delle evidenze utilizzate dal sistema. In contesti come quello qui esaminato, la solidità dell'*output* dipende infatti, in misura rilevante, dalla qualità della base informativa su cui si innesta l'elaborazione automatizzata.

A livello organizzativo, ruoli, responsabilità, livelli di autorizzazione, procedure di validazione, *logging*, modalità di *escalation* dei casi dubbi e monitoraggio continuo delle prestazioni devono essere definiti in modo esplicito e integrati nei processi esistenti. L'introduzione di strumenti di IA, infatti, non richiede soltanto una scelta tecnologica, ma anche un adattamento consapevole dell'organizzazione che ne governa l'impiego.

Resta infine il presidio giuridico-regolatorio, che impone di coordinare l'adozione della tecnologia con l'AI Act, con la Legge italiana n. 132 del 2025, con la disciplina in materia di protezione dei dati, con i requisiti di resilienza operativa digitale e, più in generale, con le specificità del settore finanziario e dell'azione amministrativa di vigilanza.

Questi presidi spostano il baricentro della riflessione. La domanda non è se un sistema di IA sia, in astratto, utilizzabile, ma a quali condizioni possa esserlo in modo affidabile, proporzionato e coerente con la missione dell'Autorità. È su questo passaggio – dalla semplice possibilità tecnica alle condizioni concrete di impiego – che si misura il grado di maturità di un effettivo *framework* interno di *governance*.

6 Considerazioni conclusive

L'impiego dell'intelligenza artificiale nell'azione di contrasto agli abusivismi finanziari rappresenta un banco di prova particolarmente significativo per comprendere in che modo un'autorità di vigilanza possa integrare tecnologie avanzate nei propri processi senza comprometterne le garanzie fondamentali. Il prototipo mostra che l'IA può offrire un contributo per trattare contenuti eterogenei, per ridurre i tempi di analisi, per classificare preliminarmente i casi e predisporre supporti documentali per l'istruttoria. Ma questo contributo non è automatico, dipende dalle scelte architetturali, organizzative e normative che accompagnano lo sviluppo e l'utilizzo dello strumento.

Considerato nel suo insieme, il percorso sviluppato in questo lavoro – dal contesto degli abusivismi finanziari alla costruzione del prototipo, dalla verifica empirica dei risultati all'analisi dei limiti e delle condizioni di impiego – mostra come il tema dell'IA nella vigilanza non possa essere affrontato separando il piano tecnico da quello organizzativo e regolatorio.

La lezione principale è che, nella vigilanza, la qualità della *governance* conta almeno quanto la capacità computazionale. La sua compatibilità con il Regolamento (UE) 2024/1689, con la Legge 23 settembre 2025, n. 132 e con il *framework* interno in via di consolidamento non va ricercata in classificazioni astratte del sistema, ma nella capacità di progettare e utilizzare questi strumenti in modo coerente con i principi di legalità, proporzionalità,

trasparenza, *accountability* e supervisione umana. In questa prospettiva, il contributo effettivo dell'IA alla vigilanza dipende anche dalla capacità dell'amministrazione di disporre di strutture tecniche interne per l'industrializzazione in grado di progettare, integrare, verificare, manutenere ed evolvere nel tempo tali soluzioni, assicurandone la coerenza con i processi istituzionali e con i vincoli regolatori che ne disciplinano l'impiego.

La prospettiva più convincente è quella di un'amministrazione assistita dalla tecnologia, non automatizzata. Un'amministrazione più capace di leggere la complessità, ma ancora pienamente responsabile delle scelte compiute; più veloce nel trattare le informazioni, ma non meno attenta alla verifica degli *output*; più aperta all'innovazione, ma consapevole dei limiti degli strumenti impiegati.

Questo richiede competenze multidisciplinari. Non basta conoscere il procedimento amministrativo, né basta conoscere la tecnologia. Occorre saper leggere le capacità e i limiti del sistema, monitorarne il funzionamento, riconoscere anomalie e prestazioni inattese, e intervenire tempestivamente quando l'*output* non appare coerente con le evidenze disponibili.

Bibliografia

Acheampong, A., Elshandidy, T. (2021), Does soft information determine credit risk? Text-based evidence from European banks, Journal of International Financial Markets, Institutions and Money, 75, 101303

Ahmed, S., Alshater, MM., El Ammari, A., Hammami, H. (2022), Artificial intelligence and machine learning in finance: A bibliometric review, Research in International Business and Finance, 61, 101646

Bender, EM., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610-623

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, JD., Dhariwal, P., et al. (2020), Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-1901

Calvaresi M., Parente M., Trerotola M., D. (2026), A Hybrid Multi-Agent System for Early Scam Detection in Crypto-Assets, *Applied Sciences*, 16 (7), 3122.

CONSOB (2024), Relazione per l'anno 2023 - L'adozione di tecnologie innovative nella vigilanza, pp. 52-53, https://www.consob.it/web/area-pubblica/relazione-annuale

CONSOB (2024), Relazione per l'anno 2023 - L'innovazione di processi, sistemi e strumenti, p. 85, https://www.consob.it/web/area-pubblica/relazione-annuale

CONSOB - Scuola Normale Superiore di Pisa (2024), Dimensionality reduction techniques to support insider trading detection, Quaderni FinTech CONSOB, n. 12, https://www.consob.it/web/area-pubblica/fintech

Deriu, P., Racioppi, S. (2025), Riflessioni in tema di intelligenza artificiale e attività di vigilanza, con presentazione a cura di A. Lalli, Quaderni FinTech CONSOB, n. 15

Devlin, J., Chang, MW., Lee, K., Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT 2019, 4171-4186

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018), A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys, 51(5), Art. 93, 1-42

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, YJ., Madotto, A., Fung, P. (2023), Survey of Hallucination in Natural Language Generation, ACM Computing Surveys, 55(12), Art. 248, 1-38

Jullum, M., Løland, A., Huseby, RB., Ånonsen, G., Lorentzen, J. (2020), Detecting money laundering transactions with machine learning, Journal of Money Laundering Control, 23(1), 173-186

Kaplan, A., Haenlein, M. (2019), Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence, Business Horizons, 62(1), 15-25

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016), Neural Architectures for Named Entity Recognition, Proceedings of NAACL-HLT 2016, 260-270

Lembo, D., Limosani, A., Medda, F., Monaco, A., Scafoglieri, FM. (2022), Information Extraction through AI techniques: The KIDs use case at CONSOB, arXiv:2202.01178

Li, J., Sun, A., Han, J., Li, C. (2022), A Survey on Deep Learning for Named Entity Recognition, IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70

Li, Y., Wang, S., Ding, H., Chen, H. (2023), Large Language Models in Finance: A Survey, Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF '23), 374-382

Liu, NF., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P. (2024), Lost in the Middle: How Language Models Use Long Contexts, Transactions of the Association for Computational Linguistics, 12, 157-173

Lundberg, SM., Lee, SI. (2017), A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (NeurIPS), 30, 4765-4774

Manning, CD., Surdeanu, M., Bauer, J., Finkel, J., Bethard, SJ., McClosky, D. (2014), The Stanford CoreNLP Natural Language Processing Toolkit, Proceedings of ACL 2014: System Demonstrations, 55-60

Memon, J., Sami, M., Khan, RA., Uddin, M. (2020), Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR), IEEE Access, 8, 142642-142668

OECD (2024), Regulatory approaches to Artificial Intelligence in finance, OECD Artificial Intelligence Papers, No. 24, OECD Publishing, Paris, DOI: 10.1787/f1498c02-en.

OECD (2026), Artificial Intelligence in Italian Financial Markets, OECD Publishing, Paris, DOI: 10.1787/6f42c977-en.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2023), GPT-4 Technical Report, arXiv:2303.08774

OpenAI (2025b), gpt-oss-120b & gpt-oss-20b Model Card, arXiv:2508.10925, https://doi.org/10.48550/arXiv.2508.10925

Otsu, N. (1979), A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62-66

Parente, M., Rizzuti, L., Trerotola, M. (2024), A profitable trading algorithm for cryptocurrencies using a Neural Network model, Expert Systems with Applications, 238, 121806,

Paterlini, S., Nicolodi, A., Gentile, M., Foglia Manzillo, V., Sancilio, MR., Deriu, P. (2025), Greenwashing alert system for EU green bonds: The CONSOB-University of Trento prototype, Quaderni FinTech CONSOB, n. 14

Ribeiro, MT., Singh, S., Guestrin, C. (2016), "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144

Smith, R. (2013), History of the Tesseract OCR Engine: What Worked and What Didn't, Proceedings of SPIE 8658, Document Recognition and Retrieval XX, 865802

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Kaiser, Ł., Polosukhin, I. (2017), Attention is All You Need, Advances in Neural Information Processing Systems (NeurIPS), 30, 5998-6008

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. (2020), LayoutLM: Pre-training of Text and Layout for Document Image Understanding, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1192-1200

Quaderni FinTech

19 – luglio 2026 **Sperimentazione dell'IA nell'azione di contrasto della CONSOB agli abusivismi finanziari**
M. Trerotola, A. Boghi, P. Deriu, G. Frega

18 – giugno 2026 **La comunicazione finanziaria tramite il canale digitale**
Profili di rischio per l'investitore retail
A. Canepa, M. Caratelli, D. Colonnello, A. Ottavianelli, P. Soccorso

17 – dicembre 2025 **L'aggregazione di dati finanziari nella prospettiva dell'*open finance***
Profili di mercato e regolatori
A. Burchi, S. Mezzacapo, P. Musile Tanzi, T.N. Poli, G. Quaresima, V. Troiano

16 – dicembre 2025 **Sustainable Development Goals omission and environmental sentiment metric for greenwashing and ESG controversies alert in green bonds**
A. Nicolodi, S. Paterlini, M. Gentile, V. Foglia Manzillo, G. Vittorioso

15 – agosto 2025 **Riflessioni in tema di intelligenza artificiale e attività di vigilanza**
P. Deriu, S. Racioppi; con presentazione a cura di A. Lalli

14 – luglio 2025 **Greenwashing alert system for EU green bonds**
The CONSOB–University of Trento prototype
S. Paterlini, A. Nicolodi, M. Gentile, V. Foglia Manzillo, M.R. Sancilio, P. Deriu

13 – luglio 2024 **Il crowdinvesting made in Italy**
Un'indagine conoscitiva
V. Caivano, C. Lucarelli, F.J. Mazzocchini, P. Soccorso

12 – febbraio 2024 **Dimensionality reduction techniques to support insider trading detection**
CONSOB - Scuola Normale Superiore di Pisa

11 – dicembre 2022 **A machine learning approach to support decision in insider trading detection**
CONSOB - Scuola Normale Superiore di Pisa

10 – luglio 2022 **How Covid mobility restrictions modified the population of investors in Italian stock markets**
CONSOB - Scuola Normale Superiore di Pisa

9 – giugno 2022 **L'intelligenza artificiale nell'*asset* e nel *wealth management***
N. Linciano, V. Caivano, D. Costa, P. Soccorso,
T.N. Poli, G. Trovatore; in collaborazione con Assogestioni

8 – aprile 2021 **La portabilità dei dati in ambito finanziario**
A cura di
A. Genovese e V. Falce

7 – settembre 2020 **Do investors rely on robots?**
Evidence from an experimental study
B. Alemanni, A. Angelovski, D. Di Cagno, A. Galliera, N. Linciano, F. Marazzi, P. Soccorso

6 – dicembre 2019 **Valore della consulenza finanziaria e *robo advice* nella percezione degli investitori**
Evidenze da un'analisi qualitativa
M. Caratelli, C. Giannotti, N. Linciano, P. Soccorso

5 – luglio 2019 **Marketplace lending**
Verso nuove forme di intermediazione finanziaria?
A. *Sciarrone Alibrandi, G. Borello, R. Ferretti, F. Lenoci, E. Macchiavello, F. Mattassoglio, F. Panisi*

4 – marzo 2019 **Financial Data Aggregation e Account Information Services**
Questioni regolamentari e profili di business
A. Burchi, S. Mezzacapo, P. Musile Tanzi, V. Troiano

3 – gennaio 2019 **La digitalizzazione della consulenza in materia di investimenti finanziari**
Gruppo di lavoro CONSOB, Scuola Superiore Sant'Anna di Pisa, Università Bocconi, Università di Pavia, Università di Roma 'Tor Vergata', Università di Verona

2 – dicembre 2018 **Il FinTech e l'economia dei dati**
Considerazioni su alcuni profili civilistici e penalistici
Le soluzioni del diritto vigente ai rischi per la clientela e gli operatori
E. Palmerini, G. Aiello, V. Cappelli G. Morgante, N. Amore, G. Di Vetta, G. Fiorinelli, M. Galli

1 – marzo 2018 **Lo sviluppo del FinTech**
Opportunità e rischi per l'industria finanziaria nell'era digitale
C. Schena, A. Tanda, C. Arlotta, G. Potenza