Navigating the Landscape of Automated Information Extraction for Financial Fund Prospectuses: Survey and Challenges

Yuyao Xu, Mohamad Farhan Mohamad Mohsin
School of Computing (SoC), University Utara Malaysia (UUM), 06010 Sintok, Kedah, Malaysia

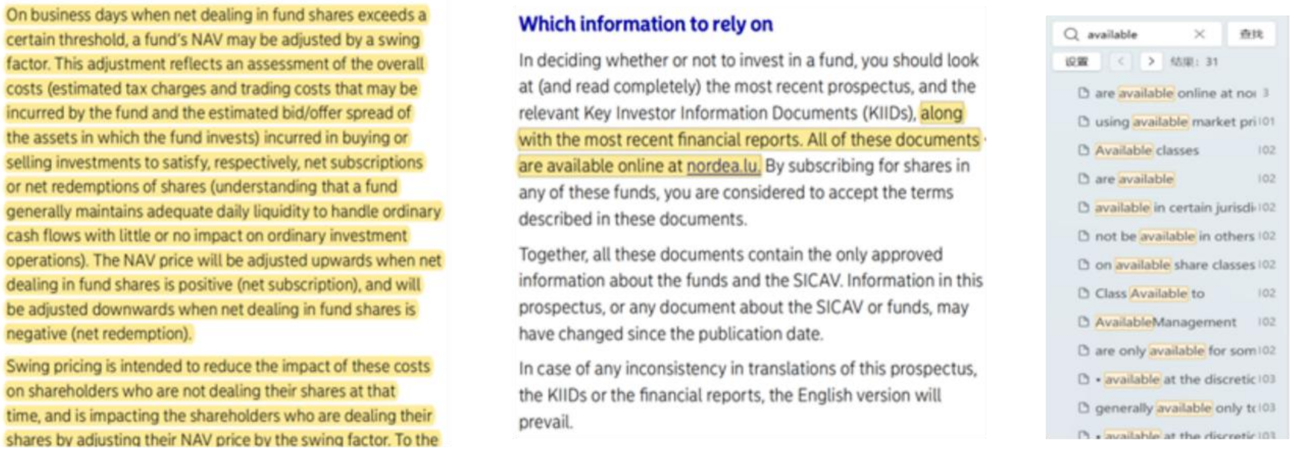
***Abstract*—In the financial sector, a fund prospectus is a critical document mandated by the Securities and Exchange Commission (SEC) that provides vital information about investments to the public. These documents encompass a range of financial concepts that define the fund's operations, including its name and disclaimers associated with periodic reports. Traditionally, the identification of these concepts has been a manual, labour-intensive, and costly task for financial regulators, aimed at ensuring the completeness of information. Automating this process is fraught with challenges, including the lengthy nature of prospectuses, the nuances of financial language, and the scarcity of labelled data for effective model training. This study explores state-of-the-art methods for information extraction, specifically within the context of financial documents. It begins with an overview of information extraction, detailing its definition and various types, such as Named Entity Recognition (NER) and event extraction. The discussion highlights the increasing significance of information extraction in the financial domain and reviews typical application areas. Ultimately, this research seeks to highlight the challenges within existing methods through a comprehensive literature review, emphasizing the need for more effective techniques tailored to the extraction of financial concepts in fund prospectuses. By enhancing and streamlining the extraction process, it aspires to improve efficiency and reduce costs for financial regulators, thereby ensuring more accurate and comprehensive information dissemination.**

***Keywords*—Machine learning; information automation; financial documentation**

I. Introduction

Financial documents (also known as financial statements) are formal records that provide a comprehensive and accurate representation of an enterprise's financial status, performance, and cash flow over a specific period. There are many types of financial documents, such as financial news, discussion boards, annual financial reports, and fund prospectuses. In the financial sector, every investment fund is mandated to provide detailed and informative financial documentation. These documents enable regulatory authorities and investors to comprehensively evaluate all pertinent and material information regarding the investment. For example, the prospectus offers an in-depth account of the public investment opportunity. Such documentation, primarily conveyed through natural language, supplemented with tables and mathematical formulations, serves as critical instruments in safeguarding regulatory compliance and ensuring transparency and oversight of the investment. These financial documents are highly regulated, which are important to every company. The regulations that apply depend on the type of fund to which the documents refer and on the countries in which the fund is distributed. Specifically, the regulations clearly outline the specific information that must be included in each type of document. Based on these requirements, legal experts can develop 'suggested wording' to ensure that fund documents are both compliant with regulatory standards and accurately convey the necessary details. The suggested wording is by no means universal, since asset and fund managers customize the narrative of fund documents to differentiate their products from those of the competitors. Furthermore, both funds and regulations change over time; as a consequence, it is important to update the corresponding fund documentation accordingly.

When financial documents are submitted to the financial regulator, the regulator needs to manually read the financial document and check the correctness of it. One of the most information factors that affects the correctness is that the financial document should clearly specify all the information needed in each type of document. The current practice for finding the sentences related to each type of information in the financial documents is still largely a manual activity [1, 2]. The financial regulator has to search for some “keywords” for each type of information, and check if the returned results are complete. This searching process is time-consuming and costly [1]. On one hand, there could be hundreds of types of information. The regulator has to manually type keywords related to all these types of information to completely check the financial documents, which is time-consuming. Besides that, many sentences can match the same keyword [3]. The regulator must select the right sentences describing each type of information to understand whether the sentences correctly explain the corresponding type of information. This process is costly. Doing so often turns out to be a complex, time-consuming and expensive activity, noting that the relevant information may not be in a structured form but rather provided through textual statements. For example, Fig. 1 shows two snippets of financial documents (i.e., prospectuses). The two snippets show a part of text churn related to the financial concepts “swing pricing” and “disclaimer on periodical reports". The text related to each financial concept is quite different. For swing pricing, a very long text churn is labelled as related to it [see Fig. 1(a)]; while there is only one sentence related to ‘disclaimer on periodical reports’ [see Fig. 1(b)], which shows that the report is available online. Such text churn is not easy to locate. If the regulator searches some keywords (e.g., available), 31 results are returned, as in Fig. 1(c). The regulator has to examine 31 paragraphs to find the text related to ‘disclaimer on periodical reports. An alternative is to locate the text based on the title of paragraph; however, there are still hundreds of titles in the document. The regulator may not easily know that the text related to ‘disclaimer on periodical reports’ is under the paragraph with the title “which information to rely on”.



On business days when net dealing in fund shares exceeds a certain threshold, a fund's NAV may be adjusted by a swing factor. This adjustment reflects an assessment of the overall costs (estimated tax charges and trading costs that may be incurred by the fund and the estimated bid/offer spread of the assets in which the fund invests) incurred in buying or selling investments to satisfy, respectively, net subscriptions or net redemptions of shares (understanding that a fund generally maintains adequate daily liquidity to handle ordinary cash flows with little or no impact on ordinary investment operations). The NAV price will be adjusted upwards when net dealing in fund shares is positive (net subscription), and will be adjusted downwards when net dealing in fund shares is negative (net redemption).

Swing pricing is intended to reduce the impact of these costs on shareholders who are not dealing their shares at that time, and is impacting the shareholders who are dealing their shares by adjusting their NAV price by the swing factor. To the

(a) swing pricing

Which information to rely on

In deciding whether or not to invest in a fund, you should look at (and read completely) the most recent prospectus, and the relevant Key Investor Information Documents (KIIDs), along with the most recent financial reports. All of these documents are available online at nordea.lu. By subscribing for shares in any of these funds, you are considered to accept the terms described in these documents.

Together, all these documents contain the only approved information about the funds and the SICAV. Information in this prospectus, or any document about the SICAV or funds, may have changed since the publication date.

In case of any inconsistency in translations of this prospectus, the KIIDs or the financial reports, the English version will prevail.

(b) Disclaimer on periodical reports

are available online at no: 3
using available market pri 101
Available classes 102
are available 102
available in certain jurisdi 102
not be available in others 102
on available share classes 102
Class Available to 102
AvailableManagement 102
are only available for som 102
• available at the discretic 103
generally available only tc 103
• available at the discretic 103

(c) Result of keyword search

Fig. 1. Examples of financial documents and financial concepts.

II. Motivation

A fund prospectus is a formal legal document required by regulatory authorities, such as the Securities and Exchange Commission (SEC), that provides detailed information about an investment fund [4]. It is designed to inform potential investors about the fund's objectives, strategies, risks, fees, performance, and other essential details to help them make informed investment decisions. The significant manual effort required frequently results in increased fund setup costs, extended time-to-market for fund products, and heightened compliance risks. There is therefore a need for automating finding the sentences related to each type of information. By doing so, the financial regulator can directly find the necessary sentences without manual checking. If the automated identification is accurate, the regulator also does not need to select the right sentences from all the returned results. In this study, the type of information to be extracted is referred to as financial concepts, which explain the concept, behaviour, definition, and metadata of financial activities, such as the detailed explanation of the calculation methodology for determining the issue price, along with the specific conditions and procedures governing the issuance process. All these concepts need to be clearly explained in certain types of financial documents.

Existing studies focus on mining financial data from web media (e.g., financial news and discussion boards [5]) and traditional financial documents (e.g., annual financial reports and 10-K) [6]. Besides, tracing information in documents has been widely studied in the areas of information extraction and software engineering [7, 8]. In these areas, they extract named entities from news and block [9]; based on different entities, the events related to each entity are also extracted [10, 11]. However, all these methods are not designed for the need of financial concept extraction in financial documents, which are more complex documents compared to discussion boards in the financial domain.

This study is proposed by analysing financial concepts in financial documents, taking fund prospectus documents in stock markets as a case study, and developing a tool-supported solution that can analyse a financial document (i.e., fund prospectuses) to automatically extract key types of information (i.e., financial concepts). The selection of the fund prospectus for the case study is due to its special characteristics. First, compared to other financial documents like financial news and discussion boards, fund prospectuses are lengthy with hundreds of pages, which increases the time for users to read the documents. Second, fund prospectuses are also difficult to read compared to financial news and discussion boards, since fund prospectuses may use similar sentences to refer to different types of financial concepts. Therefore, it could be helpful for users if the financial concepts could be automatically extracted from fund prospectuses.

III. Contribution

This study explores the significant challenges involved in the automated identification of financial concepts within fund prospectuses, a task that has traditionally required manual intervention by financial regulators. Fund prospectuses, which are essential documents regulated by the SEC, include crucial information about investment funds. However, their complexity and length pose considerable obstacles to automation.

The first challenge arises from the extensive nature of prospectuses, which often results in much longer texts than those typically addressed in existing research. This length complicates the extraction of relevant financial concepts. Additionally, the nuanced language specific to finance means that even minor variations in wording can lead to significant changes in meaning, making accurate identification even more difficult.

Furthermore, the scarcity of labelled datasets for training machine learning models in this domain presents another critical barrier, limiting the effectiveness of current automated approaches. While existing information extraction methods have proven successful with other document types, such as scientific articles and web pages, they fall short when applied to fund prospectuses.

By tackling these specific challenges, this research seeks to offer valuable insights that will lead to the development of more effective strategies for extracting financial concepts from fund prospectuses. This endeavour aims to enhance regulatory compliance and improve transparency within the financial sector, ultimately establishing a foundation for better information accessibility for all stakeholders involved.

IV. Organization

This study is structured into several key sections to provide a comprehensive exploration of information extraction in financial contexts, as shown in Fig. 2. The Introduction sets the stage by discussing the significance of fund prospectuses within the financial sector and underscores the critical role of information extraction in enhancing regulatory processes. In the Motivation section, the study elaborates on the rationale behind the study, emphasizing the challenges currently faced in automating the extraction of financial concepts from these complex documents. The Organization section outlines the study's structure, guiding readers through the subsequent content. The Contribution section identifies the unique contributions of the research to the field of information extraction, highlighting innovative approaches or insights that advance understanding. Following this, the Information Extraction section provides a general overview of various extraction techniques and their practical applications, laying the groundwork for more specific discussions. The study then shifts focus to Named Entity Recognition (NER), detailing various methods such as rule-based, feature-based, and deep learning-based approaches, and includes a critical discussion on the limitations of existing NER methods.

The Event Extraction section defines the essential concepts and tasks involved in extracting events from financial texts, followed by an examination of both pattern-matching and machine learning-based methods for event extraction. Finally, the study concludes with a section on Research Challenges and Future Directions, which identifies significant challenges in the field and offers recommendations for future research, aiming to inspire further advancements in the extraction of financial information.

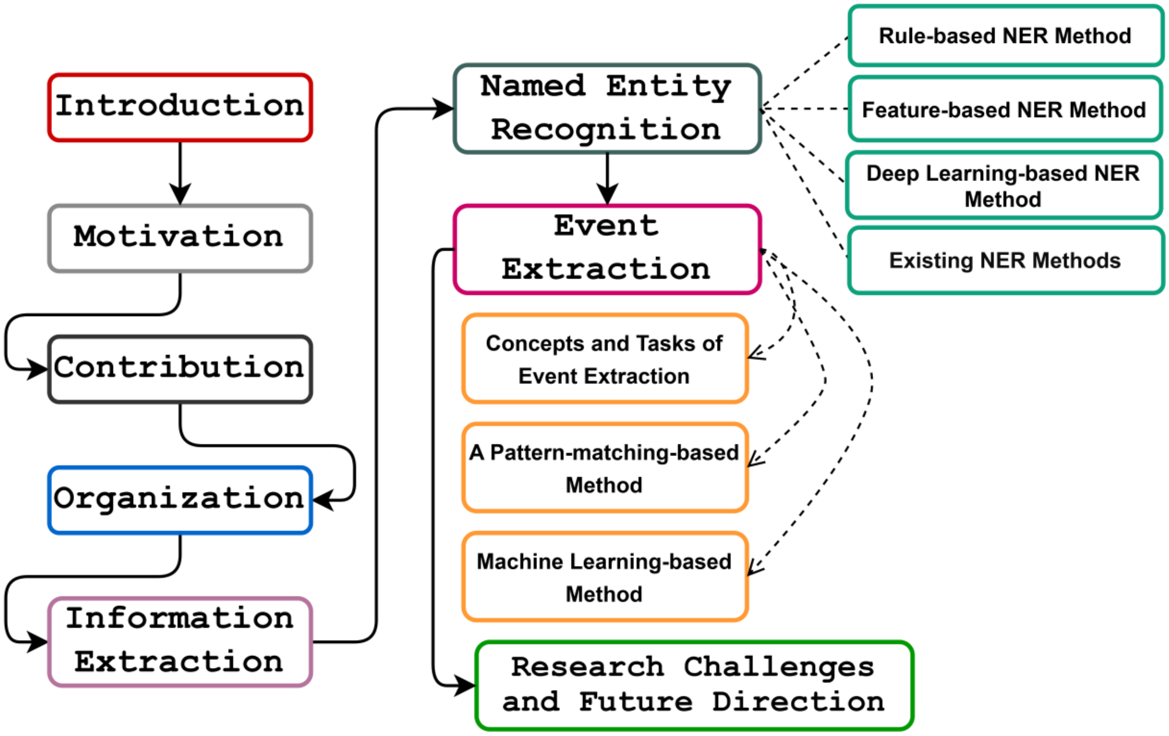


Fig. 2. Composition of the document.

V. Information Extraction

With the swift advancement of information technology, the volume of online data continues to grow daily. Therefore, it becomes the focus of people’s attention on how to obtain valuable information from massive data. To achieve this goal, information extraction is the technology for it. Information extraction refers to the structured processing of information contained in text into a specific structure of organization [12]. The typical information extraction tasks include NER and event extraction [13].

- NER refers to the extraction of entities in text; the entities are usually small pieces of phrases describing certain concepts. The example of NER in fund prospectus is the valuation day. The phrase ‘Weekly, each Tuesday’ explains the day for valuation. However, there are many dates in fund prospectuses; only this date indicates the valuation day, making this entity hard to identify. Another example is ‘disclaimer on periodical reports’, which should be a link to download the periodical reports. However, there are also many links in the fund prospectuses. Existing NER techniques cannot distinguish them.

- Event extraction is equivalent to the extraction of multiple sentences to explain 5W (i.e., what, who, when, where, why). The example of an event in fund prospectus is a disclaimer on periodical reports. The sentence ‘along with the most recent financial reports, all of these documents are available online at nordea.lu’ should be extracted as an event, which shows where the periodical reports can be obtained.

For instance, information extraction tasks can retrieve details such as time, location, and key individuals from news articles, as well as product names, development timelines, and performance metrics from technical documents. Additionally, it can extract factual information relevant to users from natural language texts. The applications of information extraction span various fields, including finance, journalism, and healthcare, highlighting its versatility and importance across multiple domains. This subsection defines the tasks of information extraction as follows.

First, for a document with N sentences, D = (S1, S2, ..., SN). Information extraction aims to extract entities, events, and relationships in D. The entity in D is represented as (E1, E2, …, EN) = H (S1, S2, ..., SN), where E1, E2, …, EN are the entities, which are the words in sentences S. NER is to find a function H, what can identify the target E in S. Regarding event extraction, it defines a function P, which can select a set of sentences in S. It has Eventm = Pm (S1, S2, ..., SN), which uses a subset of S to form different events. According to the above formal definition, the task of information extraction is:

“A process of automatically extracting structured information from unstructured and semi-structured data sources. It involves the identification, classification, and extraction of relevant information from textual, numerical, or multimedia data sources” [14].

In this study, the task is to extract the financial concept in fund prospectuses. Financial concepts are the phrases, sentences, and paragraphs to explain the related concepts in financial documents, such as ‘fund names, calculation method for issue prices, and disclaimer on periodical reports’. The identification of these concepts is important for regulators to check fund prospectuses. It is also possible to organize fund prospectuses into structured documents based on the extraction results, whose purpose is the same as the purpose of information extraction. Information extraction can be applied to a wide range of domains, including news articles, social media posts, clinical records, financial reports, and scientific papers. The extracted information may include entities (e.g., people, organizations, locations), events (e.g., births, deaths, mergers), and attributes (e.g., product features, customer data).

Information extraction methods are generally categorized according to the type of information being extracted, primarily into NER and event extraction techniques. In this study, this dimension is used for the literature review. The two classes of models are surveyed since this proposal also plans to extract entities (e.g., fund name, website) and events (e.g., the procedure to calculate issue prices) from fund prospectuses.

VI. Deep Learning Network Architecture

The NER (Named Entity Recognition) mainly completes the identification of predefined semantic type named entities from unstructured data. It is currently widely used in search engines, intelligent recommendations, and machine translation. For example, in machine translation, the named entities need to be identified, since each named entity can also have a specified translation.

NER methods can be divided into two methods: rule-based methods and machine learning-based methods [15], as shown in Fig. 3. Rule-based NER methods rely on predefined patterns, dictionaries, and linguistic rules to identify entities in text, making them particularly useful for domain-specific tasks like financial document processing. These methods often use a combination of regular expressions, lexical resources (e.g., lists of financial terms like "management fees" or "dividend yield"), and linguistic rules (e.g., POS tags or dependency parsing) to detect entities. For example, a rule might specify that a sequence of words following the phrase "expense ratio" and preceded by a dollar sign or percentage symbol should be tagged as a financial metric. While rule-based NER is highly interpretable and effective for structured or predictable text, it requires domain expertise to design rules and may struggle with variability in language.

Machine learning-based methods are divided into two categories according to the degree of dependence on the corpus: supervised NER and unsupervised NER. Machine learning-based NER methods use statistical models to automatically learn patterns and features from annotated data for identifying entities in text. These methods typically rely on sequence labelling models, such as Conditional Random Fields (CRFs) or deep learning architectures like BiLSTM-CRF and transformer-based models (e.g., BERT), to predict entity tags (e.g., "ORG," "PER," or "FINANCIAL_TERM") for each token in a sentence. Unlike rule-based methods, ML-based NER does not require manually crafted rules; instead, it learns from labelled datasets, capturing complex linguistic patterns and contextual dependencies. For example, a model trained on financial documents can learn to recognize entities like "management fees" or "dividend yield" based on their context.

This subsection introduces the relevant methods of NER. It also surveys the hot deep learning methods currently studied.

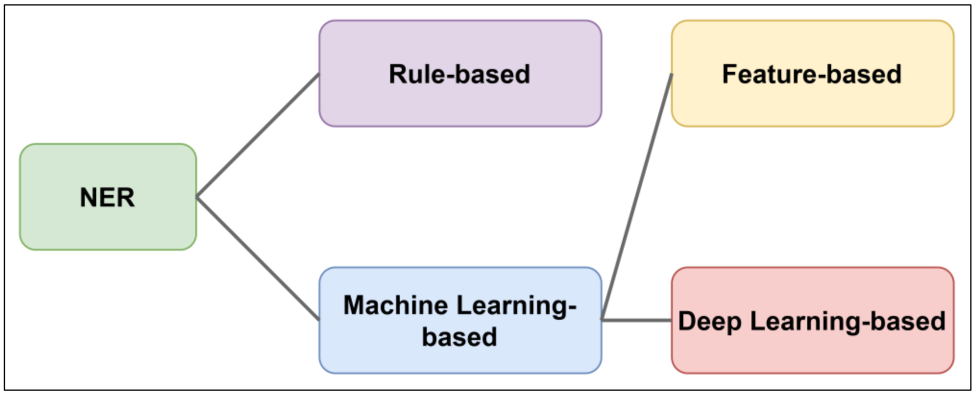


Fig. 3. Classification of NER methods.

A. *Rule-Based NER Methods*

In the early research stage of NER, the rule-based method was mainly used. The rules are mainly written and formulated by domain experts. The method first writes some simple rules by domain experts, and then experiments are carried out in the corpus. According to [16, 17] the experiment results, the rules are continuously improved after analysing the erroneous results until the effect of NER is satisfactory. Among the rule-based methods, the most commonly used is the dictionary-based matching method, which completes the NER through complete or partial matching of strings. This method is relatively simple and efficient. The implementation of dictionary-based matching methods can usually be implemented by various algorithms, such as forward matching, inverse matching, bidirectional longest matching, and dictionary trees.

The advantage of the rule-based NER method is that there is no need to label the corpus in advance, which has a good effect on a small-scale corpus. The NER system also runs quickly. The disadvantage is that the writing rules have high requirements for the individual level of personnel and poor system portability. Some well-known rule-based NER systems include LaSIE-II [16] and NetOwl [17].

B. *Feature-Based NER Method*

Feature-based NER methods are supervised methods using traditional machine learning. This method treats NER as a sequence labelling task. The specific algorithm model needs to be trained by using the labelled corpus. In the process of identifying named entities, the feature-based method usually includes the following steps:

a) Corpus labelling: The corpus is typically annotated using either the IOB (Inside-Outside-Beginning) or IO (Inside-Outside) annotation systems, which involve manually labelling the text to identify relevant entities.

b) Feature definition: Key features are defined by selecting the current word, as well as the preceding and following words, along with their parts of speech. These elements significantly influence the performance of NER tasks.

Model Training: Commonly employed models for training. Further, Table I shows several studies on feature-based NER recognition methods, which highlight the implemented methods, target entities, and identify the inadequacies in methods.

TABLE I SUMMARY OF FEATURE-BASED NER METHODS

Method	Description	Target	Type of Documents	Shortcomings
HMM [18]	Analyse word sequence with hidden Markov model	CoNLL2002, CoNLL2003	News and story	These methods demonstrate a notably low level of accuracy.
Traditional CRF [19]	Apply CRF for NER in biomedical domain	CoNLL2002, CoNLL2003	News and story	
CRF-based hybrid model [20]	Combine CRF with rules	CoNLL2002, CoNLL2003	News and story	
Traditional SVM [18]	Apply SVM for NER in biomedical domain	CoNLL2002, CoNLL2003	News and story	
Language independent SVM [21]	A language independent SVM model	CoNLL2002, CoNLL2003	News and story	
Traditional SVM [22]	Apply SVM for NER in real-time text	GENIA	Text in video	

C. *Deep Learning-Based NER Method*

Compared with the feature-based NER method, the deep learning-based NER method does not require manual rules or complex features. It is easy to extract hidden features from the input corpus. In NER tasks, commonly used neural networks mainly include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and neural networks based on attention mechanisms. Among them, Long Short-Term Memory Neural Network (LSTM) in RNNs has been widely used in NER tasks. Li et al., proposed a typical architecture for deep learning NER methods in 2020 [23], as shown in Fig. 4. The architecture is divided into three main parts:

1) Input distributed representations, which consider embeddings at the word and character levels, and combine additional features that are already effective in feature-based methods, such as part-of-speech tags and gazetteers.

2) Context encoder, which uses CNNs, RNNs, or other networks to capture context dependencies.

Tag decoder, which is mainly used to predict the label of the input sequence. Numerous studies on deep learning-based NER methods are presented in Table II, which outlines the employed techniques, target entities, and identifies the shortcomings of each method.

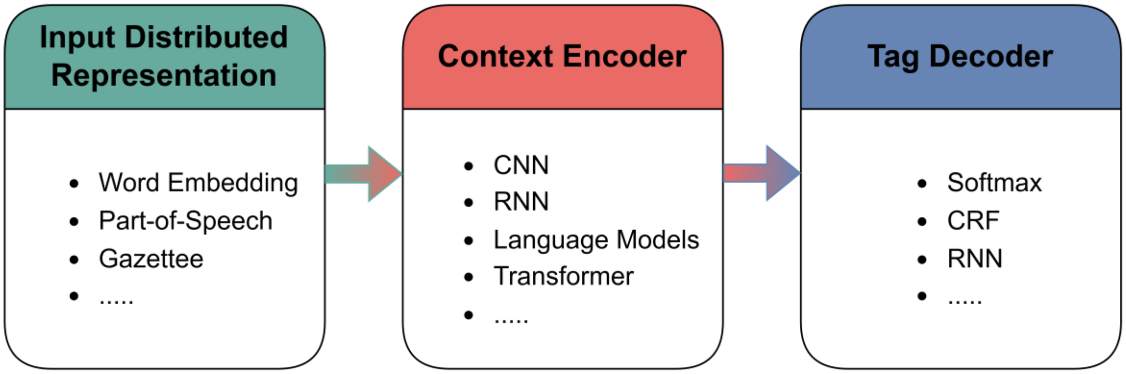


Fig. 4. A typical architecture for deep learning NER methods.

TABLE II SUMMARY OF DEEP LEARNING-BASED NER METHODS

Method	Description	Target	Type of Documents	Shortcomings
BI-LSTM-CRF [24]	A series of LSTM-based sequence labelling models	CoNLL2002, CoNLL2003 (Sentence)	News and story	These methods are intended to extract concise phrases from lengthy sentences (e.g., names, dates, and locations). However, they are built on a sentence-level model, and the target corpora do not pertain to the financial domain.
BiLSTM-CNNs-CRF [25]	An end-to-end sequence labeling model that combines bidirectional LSTM, CNN, and CRF	WSJ, CoNLL2003 (Sentence)	Newspaper, news, and story	
Attention-RNN-CRF [26]	A combination method	CoNLL2000, CoNLL2003, PTB-POS (Sentence)	News and story	
LSTM-CRF [27]	A method that uses LSTM and CRF	CoNLL2000, CoNLL2003 (Sentence)	News and story	
MRC-NER [28]	A method based on machine reading comprehension	WSJ, CoNLL2003 (Sentence)	Newspaper, news, and story	
TENER [29]	An improved Transformer encoder to model character-level and word-level features	CoNLL2003, OntoNotes 5.0 (Sentence)	News, story, blogs, conversations	
SDI-NER [30]	Combining syntactic dependency graphs for Chinese NER	Weibo	Blogs, short, messages	
WNER [31]	Using word-word relations for NER	CoNLL2003, Onto Notes 5.0, Weibo (Sentence)	News, story, blogs, short messages, conversations	
Survey [32]	A survey of NER using deep learning	NaN	NaN	
MINER [33]	Improving Out-of-Vocabulary Named Entity Recognition from an information-theoretic perspective	WNUT2017, TwitterNER, BioNER, Conll03-Typos, Conll03-OOV (Sentence)	Short messages， News, technical literature， Blogs	
kNN-NER [34]	NER with nearest neighbour search	CoNLL2003, OntoNotes,5.0OntoNotes, 4.0, MSRA, Weibo (Sentence)	Short messages, News, story	

E-NER [35]	Evidential deep learning for trustworthy NER	CoNLL2003, OntoNotes 5.0, WikiGold, TwitterNER, CoNLL2003-Typos, CoNLL2003-OOV (Sentence)	Short messages, News, story, Encyclopaedia	
GPT-NER [36]	NER via large language models	CoNLL2003, OntoNotes5.0 (Sentence)	News, short messages, story	
PromptNER [9]	Prompting for NER	CoNLL2003, CrossNER, GENIA (Sentence)	News, story, Biomedical terminology	
UniversalNER [37]	Targeted distillation from large language models for open NER	ACEO5, AnatEM, bc2gm, bc4chemd, bc5cdr, Broad Twitter, CoNLLO3, FabNER, Findvehice, GENIA, HarveyNER	News, short messages, story, Blogs etc.	

Huang et al. proposed a series of LSTM-based sequence labelling models, including LSTM, Bidirectional LSTM (BI-LSTM), LSTM (LSTM-CRF) with Conditional Random Field (CRF), and Bidirectional LSTM (BI-LSTM-CRF) with CRF layer, and compared the performance of the above models on NLP-labelled datasets [24]. In their work, the BI-LSTMCRF model is applied to the NLP benchmark sequence labelling dataset for the first time, and it is proven that the model can effectively utilize past and future input features and sentence-level labelling information. Ma et al. introduced an end-to-end sequence labelling model that integrates bidirectional LSTM, CNN, and CRF [25]. This model operates independently of task-specific resources, feature engineering, and data pre-processing, making it highly versatile.

Rei et al. (2016) proposed a combination method of word vector and character-level vector based on an attention mechanism for sequence labelling tasks [26]. This method argues that NER requires not only word vectors, but also character-level feature vectors in words. Based on the RNN-CRF model, the attention mechanism is used to splice word vectors and character-level feature vectors. Lample et al. (2016) proposed the LSTM-CRF NER model, which relies on two sources of information about words: character-based word representations are mainly learned from supervised corpora; Unsupervised word representation is learned primarily from an unannotated corpus [27].

Li et al. (2019) proposed a framework based on Machine Reading Comprehension (MRC) [28]. This method works with both non-nested and nested types of NER. Compared with the sequence annotation method, this method is simple and intuitive, and has strong portability. Experiments show that the MRC-based method can encode some prior semantic knowledge of the problem, so that it can perform better under small data sets and transfer learning. Yan, Deng, Li, & Qiu (2019) introduced the TENER model, which enhances the original Transformer architecture for NER tasks [29]. This model utilizes an improved Transformer encoder to effectively capture both character-level and word-level features.

Recent advancements in NER have seen diverse methodological innovations. Li et al. (2022) conducted a comprehensive survey highlighting the evolution of deep learning techniques in NER, providing a foundation for contemporary research [32]. Wang et al. (2022) introduced MINER, addressing out-of-vocabulary entity recognition through information-theoretic principles, while their kNN-NER leveraged nearest neighbour search to enhance contextual similarity modelling [33]. Zhang et al. (2023) proposed E-NER, integrating evidential deep learning to quantify uncertainty and improve trustworthiness in predictions [35]. The rise of large language models (LLMs) spurred novel approaches: Wang et al. (2023) developed GPT-NER [36], demonstrating LLMs' potential for direct entity recognition, while Ashok and Lipton (2023) introduced PromptNER [9], optimizing task-specific prompting strategies. Zhou et al. (2023) advanced UniversalNER [37], employing targeted distillation from LLMs to achieve robust performance in open-domain settings. Together, these works reflect a dynamic shift toward leveraging LLMs, probabilistic frameworks, and innovative training paradigms to tackle NER challenges across diverse contexts.

VII. Discussion on Existing NER Methods

As depicted in the last column of Table I, the existing information extraction techniques for NER are not suitable for financial concept extraction in fund prospectuses. The first reason is that all these methods are designed to extract short phrases from sentences, such as person names, URLs, times, and places. For example, a commonly used dataset for training and testing these methods is a dataset such as the CoNLL2003 dataset for NER tasks. CoNLL2003 covers English and German, including around 14,000 documents for English and 12,000 for German, with an average of 12 to 18 sentences per document. However, the financial concepts in fund prospectuses are usually associated with long sentences or paragraphs. These methods cannot model such sentence-level information to extract. The second reason is that, for machine learning based methods, they usually require a large dataset to train or fine-tune the model (e.g., the deep learning methods). As discussed previously, it is not easy to manually label such datasets for financial concept extraction; hence, the effectiveness of these methods could be limited [2].

This study takes the work by Wang et al. (2022) as a baseline [33, 34]. The work proposes MINER to address out-of-vocabulary entity recognition through information-theoretic principles. Since financial documents are different from many existing NER datasets (e.g., news), many out-of-vocabulary entities could exist. MINER can be a good baseline to address this issue. On top of MIER, this proposal represents domain expert rules as additional information to improve the ability of MINER in identifying long entities (i.e., financial concepts).

VIII. Event Extraction

An event is an objective fact consisting of a specific person, object, and action at a specific time and place. As an important part of information extraction, the key to event extraction is to extract information about events of interest from unstructured text and store it in a structured way [38]. This subsection first introduces the concept and tasks of event extraction. Then, the related works of event extraction based on pattern matching, machine learning, and neural networks are summarized.

A. Concepts and Tasks of Event Extraction

Gathering Event extraction techniques involve concepts such as triggers, event arguments, and argument roles, as listed in Table III.

TABLE III Basic Concepts and Terms for Event Extra

Concepts	Definition
Trigger	The rigger is a word. It is usually a verb or nouns and phrases that represent an action
Event arguments	The entity and entity property that are related to the event, including time, place, and persons
Argument roles	The role of the entity in the event, namely, the relationship between the entity and the event

The aim of event extraction is to identify the type of event and automatically extract arguments with varying roles from unstructured text using machine learning techniques [39]. The event extraction task is divided into two parts: Event Detection (ED) and Argument Extraction (EAE). Among them, event detection aims to identify specific types of event trigger words from a given text; it is a key step in event extraction [40]. Event Element Extraction identifies the elements of a particular event and marks its role [41]. For example, in the sentence “Biden swears inauguration at Capitol on January 20”, the event extraction task is specifically to detect the trigger word “inauguration”, which is the incident type; The words "Biden", "January 20" and "Capitol" are as the event elements, which can be determined as "person", "time" and "place", respectively.

The implementation of the above event extraction task is based on the technical basis of word segmentation, part-of-speech, and propositional entity recognition [42, 43]. The extracted structured information can further support deeper mining and understanding of unstructured text.

Table IV summarizes the main methods for event extraction. The methods can be classified into pattern-matching-based methods and machine learning-based methods. Early event extraction techniques typically use a pattern-matching-based method. Pattern-matching-based event extraction involves identifying specific events in text by matching predefined patterns or templates that describe how events are typically expressed.

TABLE IV Summary of Event Extraction Methods

Methods		Characteristics
Pattern matching-based		Works better in specific areas, but the process is cumbersome, prone to cascading errors, and depends on the specific text form of the specific domain
Machine learning-based	Feature-based	Relies on complex feature engineering and natural language processing tools
	Neural network-based	Works better in many EE tasks. It requires large corpus to train or fine-tune the model. It takes a long time to train the model and conduct prediction

These patterns often combine lexical cues (e.g., keywords like "acquire" or "merge"), syntactic structures (e.g., verb-argument relationships), and semantic constraints (e.g., entity types involved in the event). Pattern-matching methods are highly interpretable and effective for structured or predictable text, such as news articles or financial reports, where events follow consistent linguistic patterns. However, they require domain expertise to design patterns and may struggle with linguistic variability or complex sentence structures.

In recent years, due to the promotion of the construction of large-scale corpus and the rise of machine learning, researchers have begun to explore methods based on machine learning. Machine learning-based event extraction methods use statistical models to automatically identify and classify events (e.g., mergers, earnings announcements) in text by learning patterns from annotated data. These approaches often frame event extraction as a structured prediction task, where models like sequence labelling (e.g., BiLSTM-CRF) or transformers (e.g., BERT, GPT) are trained to detect event triggers (e.g., verbs like "acquire" or "announce") and their associated arguments (e.g., entities, time, location). For instance, a model might learn that the phrase "Company X announced a merger with Company Y" contains a "merger" event involving two organizations. Unlike rule-based methods, ML approaches generalize better to linguistic variability and implicit event mentions (e.g., "the acquisition was finalized"). They leverage contextual embeddings to capture semantic relationships and dependencies, but require large, labelled datasets for training.

B. A Pattern-Matching-Based Method

The pattern-matching-based method consists of two steps: pattern acquisition and event extraction. Firstly, the event pattern extraction is obtained through local text analysis, such as lexical analysis and syntactic analysis. Then, under the guidance of the event pattern, the event sentence to be extracted is matched to the corresponding pattern to detect and extract an event type, as shown in Fig. 5.

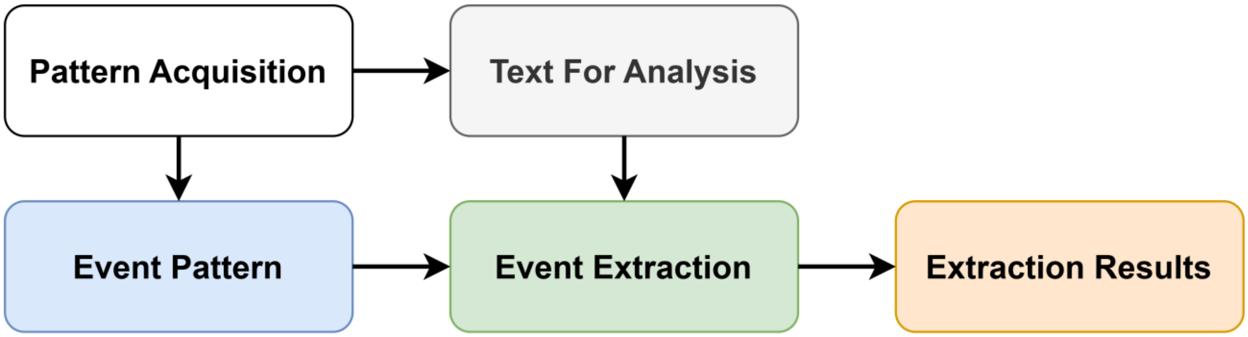


Fig. 5. Event extraction method based on pattern matching.

The supervised pattern matching method relies on a manually labelled corpus. The quality of the human-labelled corpus affects the effect of pattern learning. Therefore, Riloff (1993) mentions that statements near event elements often contain descriptions of the role of event elements in events [44]. This work constructs a domain-specific pattern for event extraction through supervised learning combined with human review.

Supervised pattern matching methods rely heavily on labelled corpus. Therefore, some researchers propose methods based on weak supervision. This type of method only requires manual pre-classification of the corpus or the definition of a small number of artificial seed patterns. Then, the method automatically learns the event mode according to the pre-classified corpus or seed mode. Based on the literature, the process only needs to label the event type but does not need to label all event elements in the corpus, which reduces the workload required to create a corpus. Table V summarizes the main works of pattern-based event extraction. A previous study (Roman, 2000) constructs a seed pattern set, divides the unlabelled corpus into related text sets and unrelated text sets [45]. It then explores the relevant patterns of related text sets to learn new models incrementally. Although this method (Roman, 2000) does not require artificial corpus labelling, the pre-classification labelling and seed pattern set they introduce are still relatively heavy. In order to solve this problem, the paper (Jiang, 2005) proposes a domain-independent pattern representation method, which is integrated into the WordNet conceptual knowledge base [46]. When performing pattern acquisition, this method only needs to define the event extraction task. The corresponding event extraction pattern can be automatically learned from the original corpus, which minimizes the intervention of manual annotation on pattern-matching event extraction. Other early studies, such as [47], demonstrated the utility of domain-specific patterns in biomedical event extraction, leveraging tailored rules to extract events from scientific literature. Similarly, Liao and Grishman (2011) combined semantic role labelling (SRL) with pattern-based rules to refine event extraction, showcasing the synergy between linguistic structures and rule-based methods. In recent years, pattern-based approaches have been enhanced by integrating them with deep learning techniques. For instance, Zhang et al. (2020) proposed a rule-augmented deep learning framework for event extraction in low-resource settings, emphasizing the value of pattern-based methods when annotated data is scarce [41]. Chen et al. (2021) further advanced this hybrid approach by incorporating dynamic multi-pooling convolutional neural networks (CNNs) to improve contextual understanding and event detection accuracy [48]. Additionally, Li et al. (2022) explored the use of transformer models like BERT alongside pattern-based rules for clinical event extraction, demonstrating their effectiveness in specialized domains [49]. These studies collectively illustrate the evolution of pattern-based methods, from standalone rule-based systems to hybrid frameworks that leverage the strengths of both traditional and modern NLP techniques. Learning models utilized for UATR often exhibit high complexity, which can lead to significant challenges in training effectiveness. When the complexity of a model exceeds the available quantity and quality of training data, achieving desired outcomes becomes increasingly difficult. While improving model accuracy is a primary objective, relying solely on accuracy as the sole metric for evaluating model performance is insufficient.

TABLE V SUMMARY OF PATTERN-BASED METHODS

Reference	Contribution	Usage scenarios
[45]	Build a seed pattern set for event extraction	Pattern-based methods require constructing patterns to extract events; however, the pattern is not easy to build, which limits their usage scenarios.
[46]	Propose domain-independent patterns	
[47]	Propose domain-specific patterns	
[50]	Combined semantic role labelling with pattern-based rules	
[41]	Propose a rule-augmented deep learning framework for event extraction	
[48]	A hybrid approach by incorporating dynamic multi-pooling CNNs	
[49]	Explored the use of transformer	

Although the pattern-based matching method is more effective in specific domains, the method is cumbersome in the process, prone to cascading errors, and depends on the specific form of the domain-specific text and requires participants to have strong professionalism in the process of obtaining patterns. Hence, its cost is high, and the formulated pattern is difficult to cover all event types. When the corpus changes, the pattern also needs to be updated, resulting in poor system generalization and low recall.

C. Machine Learning-Based Method

Compared with pattern matching methods, machine learning methods are more adaptable and portable to different domains. The machine learning-based method transforms the event extraction task into a classification problem on the basis of statistical learning and selects the appropriate feature input to the classifier to complete the extraction task. Typically, these classifiers include event trigger word classifiers, argument classifiers, and argument role classifiers. These classifiers can employ classification models such as maximum entropy and support vector machines in machine learning.

Chieu et al. first time applies the maximum entropy model for argument extraction and constructs a classifier to realize the event extraction of lecture announcements by defining simple features such as named entities, first words in sentences, lexical lowercase forms, and time expressions [51]. In order to extract more features in the feature engineering stage and improve performance, Llorens et al. (2010) also introduce semantic corner colour features; it use conditional random field model to realize event extraction and achieve good results [52]. Li et al. (2013) [53] further introduce global features to improve performance on the basis of Llorens et al. (2010) [52], and use trigger words and dependencies between multiple arguments for event extraction.

Different machine learning algorithms often have different advantages and disadvantages. It is possible to combine multiple machine learning methods to improve the event extraction effect. Ahn (2006) combines two machine learning models, which are the memory-based nearest neighbour learner and maximum entropy learner, to propose a simple modular event extraction method [54]. This method constructs multiple classifiers for each module separately; it uses lexical features, dictionary features, context features, dependency features, and entity features to complete the event extraction.

Machine learning-based event extraction often relies on complex feature engineering and natural language processing tools, which often require deep exploratory data analysis on the corpus. The cumbersome and tedious nature of this process, as well as the accumulation and propagation of errors caused by various types of feature extraction processes, will more or less affect the extraction results.

In recent years, deep learning has been a popular method to represent data in machine learning. Compared with traditional machine learning methods (feature-based), deep learning avoids the tedious work of manual feature engineering; it does not require rich domain expert knowledge and only needs to pass data directly to the built network. Its portability and flexibility have prompted more and more researchers to focus on deep learning-based methods.

The event extraction method based on deep learning can be divided into the method of pipeline model and the method of a document-level model, according to the different learning methods of the model. Pipeline model is a model that recognizes named entities and extracts events as a pipeline, while the method of document-level models is to learn the characteristics of events for extracting events directly without considering the named entities. Depending on the range of features used, it can be divided into sentence-level methods and chapter-level methods. Table VI displays recent works that utilize a combination of pipeline and document-level models, with the majority employing deep learning-based approaches.

TABLE VI SUMMARY OF MACHINE LEARNING-BASED METHODS

Ref	Contribution	Usage scenarios
[11]	Using regression forest for event extraction	Extract events in two steps. First it extracts named entities, and then events are extracted based on the named entities. As explain in existing methods for NER may not work on fund prospectus, which can affect the effectiveness of applying pipe models on financial concept extraction.
[55]	Ensemble learning for event extraction	
[56]	SVM based event extraction methods	
[57]	Event time extraction with decision tree	
[58]	Relation extraction for imbalanced data	
[48]	proposes a dynamic multi-pool convolutional network	
[59]	propose a framework based on the pre-trained model BERT	
[60]	uses LSTM (LongShort-Term Memory) for event detection	
[61]	use Bi-LSTM model	
[62]	uses the pre-trained model BERT to fine-tune Bi-LSTM	
[63]	adds an attention mechanism after the Bi-LSTM layer	
[64]	Add a gate mechanism in each graph convolutional layer	
[65]	Propose the edge-enhanced graph convolutional network	
[66]	Combine the syntactic and semantic structures through the application of the graph converter network	
[67]	Combine knowledge from hierarchical knowledge graphs and graphs neural networks on the basis of pre-training models	
[68]	Combine coreference resolution to better identify the event	
[49]	A survey for event extraction using deep learning	
[69]	Event Detection as Multi-task Text Generation	
[10]	A Data-Efficient Generation-Based Event Extraction Model	
[70]	Boosting Low-Resource Information Extraction by Structure-to-Text Data Generation with Large Language Models	

The event extraction method based on the pipeline model executes subtasks sequentially. Subsequent tasks depend on the results of the former. Chen at el. (2015) present event extraction as a two-stage multi-classification process, introducing a dynamic multi-pool convolutional network [48]. In the first stage, each word is classified to detect a trigger word. If a trigger is identified, the second stage classifies the trigger and assigns its meta-role. Yang at el. (2019) propose a framework based on the pre-trained model BERT (Bidirectional Encoder Representations Transformers), which includes a trigger word classifier and a metaclassifier that obtains results by inference and referencing the former [59].

Different classifiers can also be used for each stage. For example, Ding at el. (2019) uses LSTM (LongShort-Term Memory) in the event detection stage to construct a trigger perceptual feature extractor to dynamically learn character-level, part-of-speech, and semantic-level features to achieve event detection on the Chinese corpus [60]. Compared with the LSTM model, Bi-LSTM (Bi-directional Long Short-Term Memory) can learn long-term dependent information more effectively in sequence labelling tasks. Zeng et al. (2016) use a Bi-LSTM model to capture the semantic and structural features of sentences and achieve good results in the event detection stage [61]. Satyapanich et al. (2020) use the pre-trained model BERT to fine-tune Bi-LSTM in the event detection stage for trigger word classification [62], while the method of Zheng et al. (2018) in the argument extraction stage adds an attention mechanism after the Bi-LSTM layer for meta-recognition and finally assigns the meta role through the neural network [63].

In the deep neural network, hidden vectors in graph convolutional networks may contain noise information that is not related to candidate trigger words. Dac Lai et al. (2020) add a gate mechanism in each graph convolutional layer to filter the noise information in the hidden vector, but this method ignores the dependent label information [64]. The edge-enhanced graph convolutional network proposed by Cui et al. (2020) has achieved good results on the ACE2005 dataset by combining syntactic structure and type-dependent labels to improve event detection [65]. In order to capture the intermediate structure of the sentence, Veyseh et al. (2020) combine the syntactic and semantic structures through the application of the graph converter network; it captures the intermediate structure of the sentence from different input perspectives and then multiplies the intermediate structure to generate the final structure [66]. In addition, in order to enhance the inference ability of complex events, Li et al. (2019) combine knowledge from hierarchical knowledge graphs and graph neural networks on the basis of pre-training models [67].

Recent advancements in event extraction have introduced innovative approaches to address its inherent challenges. Lu et al. (2022) enhanced event identification by integrating coreference resolution, improving the ability to link event mentions across text [68]. Li et al. (2022) provided a comprehensive survey of deep learning techniques in event extraction, offering a roadmap for ongoing research [32]. Anantheswaran at el. (2023) reimagined event detection as a multi-task text generation problem, leveraging generative models to capture complex event structures [69]. Hsu et al. (2023) proposed a data-efficient generation-based model, addressing the scarcity of labelled data by optimizing generative frameworks for event extraction [10]. Ma et al. (2023) tackled low-resource scenarios by employing large language models for structure-to-text data generation, significantly boosting performance in information extraction tasks [70]. Together, these works highlight a shift toward leveraging generative models, coreference resolution, and data-efficient strategies to advance event extraction capabilities across diverse domains.

IX. Research Challenges and Future Directions

Event extraction extracts events from documents. The event is usually longer than named entities; hence, this task is more similar than financial concept extraction. Currently, many methods have been proposed. They extract events in two steps. First, it extracts named entities, and then events are extracted based on the named entities. As discussed previously, the existing methods for NER may not work on a fund prospectus, which can affect the effectiveness of applying pipe models on financial concept extraction. Second, there are ambiguity problems in fund prospectuses. Similar sentences may have different meanings when the context is different. The same entity can also have different meanings in different contexts, as discussed previously. Therefore, when extracting information, it is important to consider the context of the event (i.e., sentences) and conduct a multi-grained comparison on sentences (e.g., on word-level, sentence-level, keyword-level). Third, all these methods require a large, labelled data set to train the model, which is not available in financial concept extraction.

A. Small Training Set Issue in Financial Document Extraction Domain

Due to the scarcity of labelled training data, financial documents are often highly specialized, containing domain-specific terminology, complex structures, and sensitive information, making it difficult to obtain large, high-quality annotated datasets. There are a few studies to address the small training set issue in financial document information extraction.

B. Data Augmentation

The first solution is data augmentation. Anantheswaran et al. introduced a multi-task text generation framework for event detection, which can be adapted to financial information extraction [69]. By generating diverse textual representations of financial events, their method effectively expanded the training set without requiring additional manual annotations.

C. Transferring Learning and Pre-Trained Language Models

The sconed solution is transferring learning and pre-trained language models. Hsu et al. developed a data-efficient generation-based model for event extraction, leveraging pre-trained language models (PLMs) to transfer knowledge from general-domain corpora to financial documents [10]. Their approach minimized the need for large, annotated datasets by fine-tuning PLMs on small, domain-specific data.

D. Few-Shot and Zero-Shot Learning

The third solution is few-shot and zero-shot learning. Ashok and Lipton introduced PromptNER, a prompting-based approach for NER that enables few-shot and zero-shot learning [9]. By framing information extraction tasks as natural language prompts, their method reduced reliance on large, annotated datasets, making it particularly suitable for financial documents with limited labelled examples. Zhou et al. proposed UniversalNER, which uses targeted distillation from LLMs to perform open-domain NER with minimal supervision [37]. Their approach demonstrated robustness in low-resource settings, including financial document information extraction.

As shown in Table IV and Table V, there are many publicly available datasets for NER such as CoNLL2003. However, these datasets do not have financial concepts to extract. For example, the CoNLL2003 dataset is widely used benchmarks for NER tasks. CoNLL2003 covers English and German. CoNLL2003, published in 2003, includes around 14,000 documents for English and 12,000 for German, with an average of 12 to 18 sentences per document. The dataset primarily consists of news articles, making them representative of the news domain. Therefore, these datasets cannot solve the small training set issue in financial concept extraction.

X. Conclusion

In conclusion, the comprehensive analysis of information extraction techniques in this study highlights the critical importance of accurately capturing financial concepts within fund prospectuses. As established, the completeness of content in these documents is vital for investment funds, necessitating rigorous oversight from financial regulators. Despite the availability of various information extraction methods, significant challenges remain when applying these techniques to fund prospectuses, particularly due to their inherent complexity, linguistic ambiguity, and the scarcity of labelled data for effective training. The identified gaps in existing research emphasize the need for tailored frameworks specifically designed for financial concept extraction in this context. This research addresses these specific challenges to provide valuable insights that will facilitate the creation of more effective strategies for extracting financial concepts from fund prospectuses. The goal is to enhance regulatory compliance and increase transparency in the financial sector, thereby laying the groundwork for improved information accessibility for all stakeholders.

References

[1] H. Li, Q. Yang, Y. Cao, J. Yao, and P. Luo, "Cracking tabular presentation diversity for automatic cross-checking over numerical facts," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2599-2607.

[2] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," Financial Innovation, vol. 6, no. 1, p. 39, 2020.

[3] Y. Arslan et al., "A comparison of pre-trained language models for multi-class text classification in the financial domain," in Companion proceedings of the web conference 2021, 2021, pp. 260-268.

[4] M. Ceci, D. Bianculli, and L. C. Briand, "Defining a model for content requirements from the law: An experience report," in 2024 IEEE 32nd International Requirements Engineering Conference (RE), 2024, pp. 18-30: IEEE.

[5] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web media and stock markets: A survey and future directions from a big data perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 2, pp. 381-399, 2017.

[6] Q. Li, S. Shah, and R. Fang, "Table classification using both structure and content information: A case study of financial documents," in 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1778-1783: IEEE.

[7] J. Cleland-Huang, O. C. Gotel, J. Huffman Hayes, P. Mäder, and A. Zisman, "Software traceability: trends and future directions," in Future of software engineering proceedings, 2014, pp. 55-69.

[8] R. Grishman, "Twenty-five years of information extraction," Natural Language Engineering, vol. 25, no. 6, pp. 677-692, 2019.

[9] D. Ashok and Z. C. Lipton, "Promptner: Prompting for named entity recognition," arXiv preprint arXiv:2305.15444, 2023.

[10] I. Hsu et al., "DEGREE: A data-efficient generation-based event extraction model," arXiv preprint arXiv:2108.12724, 2021.

[11] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 20-31, 2014.

[12] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation," in Lrec, 2004, vol. 2, no. 1, pp. 837-840: Lisbon.

[13] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in COLING 1996 volume 1: The 16th international conference on computational linguistics, 1996.

[14] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," Journal of Big Data, vol. 6, no. 1, pp. 1-38, 2019.

[15] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, pp. 3-26, 2007.

[16] K. Humphreys et al., "University of Sheffield: Description of the LaSIE-II system as used for MUC-7," in Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998, 1998.

[17] G. Krupka and K. Hausman, "IsoQuest Inc.: description of the NetOwl™ extractor system as used for MUC-7," in Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998, 1998.

[18] S. Morwal, N. Jahan, and D. Chopra, "Named entity recognition using hidden Markov model (HMM)," International Journal on Natural Language Computing (IJNLC) Vol, vol. 1, 2012.

[19] U. Kanimozhi and D. Manjula, "A CRF based machine learning approach for biomedical named entity recognition," in 2017 second international conference on recent trends and challenges in computational models (ICRTCCM), 2017, pp. 335-342: IEEE.

[20] Z. Xu, X. Qian, Y. Zhang, and Y. Zhou, "CRF-based hybrid model for word segmentation, NER and even POS tagging," in Proceedings of the sixth SIGHAN workshop on Chinese language processing, 2008.

[21] A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," International Journal of Electrical, Computer, and Systems Engineering, vol. 4, no. 2, pp. 155-170, 2010.

[22] A. T. Imam, A. Alhroob, and W. Alzyadat, "SVM machine learning classifier to automate the extraction of SRS elements," International Journal of Advanced Computer Science and Applications (IJACSA), 2021.

[23] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE transactions on knowledge and data engineering, vol. 34, no. 1, pp. 50-70, 2020.

[24] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[25] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," arXiv preprint arXiv:1603.01354, 2016.

[26] M. Rei, G. K. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," arXiv preprint arXiv:1611.04361, 2016.

[27] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[28] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified MRC framework for named entity recognition," arXiv preprint arXiv:1910.11476, 2019.

[29] H. Yan, B. Deng, X. Li, and X. Qiu, "TENER: adapting transformer encoder for named entity recognition," arXiv preprint arXiv:1911.04474, 2019.

[30] P. Zhu et al., "Improving Chinese named entity recognition by large-scale syntactic dependency graph," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 979-991, 2022.

[31] S. Ali, K. Masood, A. Riaz, and A. Saud, "Named entity recognition using deep learning: A review," in 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022, pp. 1-7: IEEE.

[32] J. Li et al., "Unified named entity recognition as word-word relation classification," in proceedings of the AAAI conference on artificial intelligence, 2022, vol. 36, no. 10, pp. 10965-10973.

[33] X. Wang et al., "MINER: Improving out-of-vocabulary named entity recognition from an information theoretic perspective," arXiv preprint arXiv:2204.04391, 2022.

[34] S. Wang et al., "$ k $ NN-NER: Named Entity Recognition with Nearest Neighbor Search," arXiv preprint arXiv:2203.17103, 2022.

[35] Z. Zhang et al., "E-ner: evidential deep learning for trustworthy named entity recognition," arXiv preprint arXiv:2305.17854, 2023.

[36] S. Wang et al., "Gpt-ner: Named entity recognition via large language models," arXiv preprint arXiv:2304.10428, 2023.

[37] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "Universalner: Targeted distillation from large language models for open named entity recognition," arXiv preprint arXiv:2308.03279, 2023.

[38] G. Xiyue and H. Tingting, "Survey about research on information extraction," Computer science, vol. 42, no. 2, pp. 14-17, 2015.

[39] Z. Kan, L. Qiao, S. Yang, F. Liu, and F. Huang, "Event arguments extraction via dilate gated convolutional neural network with enhanced local features," IEEE Access, vol. 8, pp. 123483-123491, 2020.

[40] S. Liu, Y. Chen, S. He, K. Liu, and J. Zhao, "Leveraging framenet to improve automatic event detection," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2134-2143.

[41] N. Zhang et al., "Document-level relation extraction as semantic segmentation," arXiv preprint arXiv:2106.03618, 2021.

[42] Y. Tian et al., "Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8286-8296.

[43] R. Aly, A. Vlachos, and R. McDonald, "Leveraging type descriptions for zero-shot named entity recognition and classification," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1516-1528.

[44] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in AAAI, 1993, vol. 1, no. 1, p. 2.1.

[45] R. Yangarber, Scenario customization for information extraction. New York University, 2000.

[46] J. Jiang, "An event IE pattern acquisition method," Computer Engineering, vol. 31, no. 15, pp. 96-98, 2005.

[47] J. Björne and T. Salakoski, "Biomedical event extraction using convolutional neural networks and dependency parsing," in Proceedings of the BioNLP 2018 workshop, 2018, pp. 98-108.

[48] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 167-176.

[49] J. Li et al., "Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (BERT) and in-domain pre-training (IDPT)," BMC Medical Informatics and Decision Making, vol. 22, no. 1, p. 200, 2022.

[50] S. Liao and R. Grishman, "Acquiring topic features to improve event extraction: in pre-selected and balanced collections," in Proceedings of the international conference recent advances in natural language processing 2011, 2011, pp. 9-16.

[51] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," Aaai/iaai, vol. 2002, pp. 786-791, 2002.

[52] H. Llorens, E. Saquete, and B. Navarro, "TimeML events recognition and classification: learning CRF models with semantic roles," in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 725-733.

[53] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 73-82.

[54] D. Ahn, "The stages of event extraction," in Proceedings of the Workshop on Annotating and Reasoning about Time and Events, 2006, pp. 1-8.

[55] R. Xu, J. Hu, Q. Lu, D. Wu, and L. Gui, "An ensemble approach for emotion cause detection with event extraction and multi-kernel svms," Tsinghua Science and Technology, vol. 22, no. 6, pp. 646-659, 2017.

[56] Y. Li, K. Bontcheva, and H. Cunningham, "SVM based learning system for information extraction," in International Workshop on Deterministic and Statistical Methods in Machine Learning, 2004, pp. 319-339: Springer.

[57] N. Reimers, N. Dehghani, and I. Gurevych, "Event time extraction with a decision tree of neural classifiers," Transactions of the Association for Computational Linguistics, vol. 6, pp. 77-89, 2018.

[58] D. Song, J. Xu, J. Pang, and H. Huang, "Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data," Information Sciences, vol. 573, pp. 222-238, 2021.

[59] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 5284-5294.

[60] N. Ding, Z. Li, Z. Liu, H. Zheng, and Z. Lin, "Event detection with trigger-aware lattice neural network," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 347-356.

[61] Y. Zeng, H. Yang, Y. Feng, Z. Wang, and D. Zhao, "A convolution BiLSTM neural network model for Chinese event extraction," in Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24, 2016, pp. 275-287: Springer.

[62] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, no. 05, pp. 8749-8757.

[63] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 1049-1058.

[64] V. Dac Lai, T. Ngo Nguyen, and T. H. Nguyen, "Event Detection: Gate Diversity and Syntactic Importance Scoresfor Graph Convolution Neural Networks," arXiv e-prints, p. arXiv: 2010.14123, 2020.

[65] S. Cui, B. Yu, T. Liu, Z. Zhang, X. Wang, and J. Shi, "Edge-enhanced graph convolution networks for event detection with syntactic relation," arXiv preprint arXiv:2002.10757, 2020.

[66] A. P. B. Veyseh, T. N. Nguyen, and T. H. Nguyen, "Graph transformer networks with syntactic and semantic structures for event argument extraction," arXiv preprint arXiv:2010.13391, 2020.

[67] D. Li, L. Huang, H. Ji, and J. Han, "Biomedical event extraction based on knowledge-driven tree-LSTM," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1421-1430.

[68] Y. Lu, H. Lin, J. Tang, X. Han, and L. Sun, "End-to-end neural event coreference resolution," Artificial Intelligence, vol. 303, p. 103632, 2022.

[69] U. Anantheswaran, H. Gupta, M. Parmar, K. K. Pal, and C. Baral, "Edm3: Event detection as multi-task text generation," arXiv preprint arXiv:2305.16357, 2023.

[70] M. D. Ma, X. Wang, P.-N. Kung, P. J. Brantingham, N. Peng, and W. Wang, "STAR: boosting low-resource information extraction by structure-to-text data generation with large language models," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, vol. 38, no. 17, pp. 18751-18759.