applied sciences

MDPI

Review

Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey

Wajid Ali [1,*], Wanli Zuo [1,*], Rahman Ali [2], Xianglin Zuo [1] and Gohar Rahman [3]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China; zuoxl17@mails.jlu.edu.cn
[2] Quaid-e-Azam College of Commerce, University of Peshawar, Peshawar 25000, Pakistan; rehmanali@uop.edu.pk
[3] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Parit Raja, Batu Pahat 86400, Malaysia; goharsa@gmail.com
* Correspondence: greatforyou86@gmail.com (W.A.); zuowl@jlu.edu.cn (W.Z.)

Abstract: The era of big textual corpora and machine learning technologies have paved the way for researchers in numerous data mining fields. Among them, causality mining (CM) from textual data has become a significant area of concern and has more attention from researchers. Causality (cause-effect relations) serves as an essential category of relationships, which plays a significant role in question answering, future events predication, discourse comprehension, decision making, future scenario generation, medical text mining, behavior prediction, and textual prediction entailment. While, decades of development techniques for CM are still prone to performance enhancement, especially for ambiguous and implicitly expressed causalities. The ineffectiveness of the early attempts is mainly due to small, ambiguous, heterogeneous, and domain-specific datasets constructed by manually linguistic and syntactic rules. Many researchers have deployed shallow machine learning (ML) and deep learning (DL) techniques to deal with such datasets, and they achieved satisfactory performance. In this survey, an effort has been made to address a comprehensive review of some state-of-the-art shallow ML and DL approaches in CM. We present a detailed taxonomy of CM and discuss popular ML and DL approaches with their comparative weaknesses and strengths, applications, popular datasets, and frameworks. Lastly, the future research challenges are discussed with illustrations of how to transform them into productive future research directions.

Keywords: cause-effect relation; causality survey; causality mining; deep learning; causality extraction; relation classification

Citation: Ali, W.; Zuo, W.; Ali, R.; Zuo, X.; Rahman, G. Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey. *Appl. Sci.* **2021**, *11*, 10064. https://doi.org/10.3390/app112110064

Academic Editor: Manuel Armada

Received: 23 August 2021
Accepted: 9 October 2021
Published: 27 October 2021

1. Introduction

Natural Language Processing (NLP) areas are also termed computational linguistics, which includes designing computational systems and procedures to handle natural language problems in informative software platforms. NLP works can be categorized into two comprehensive sub-fields, (a) Core fields, and (b) Applications. Whereas it is often hard for the researchers to differentiate exactly to which fields the problem belongs. The core fields report some issues including Language Modeling, Semantic Processing, Morphological Processing, and Syntactic Processing/Parsing. Application fields focus on mining valuable relational information including Cause-Effect relation, Part-Whole relation, Product-Produce relation, Content-Container relation, If-Then relations, Translation of text among and between languages, Sentiment analysis, Summarization, Automatic question answering, document classification, and Clustering. That relational information usually exists in images, graphics, video, text, audio and multimedia data domains. Among all domains, textual data preserves much human intelligence and conveys more contextual information. For a few decades, automated knowledge extraction from text has been a challenging task because it deals with the relationship of syntax, semantics, vocabulary,

metaphors, sarcasm, and ambiguous constructs like figurative expressions. In these cases, copying the human brain's knowledge is an important task for understanding written texts that require developing a complicated model using ML and DL approaches. However, computational linguistics and computer research societies have been made remarkable developments in the field over a few decades, especially for textual data mining tasks. The application area included different types of relationships. The basic concept, principles, extraction, representation, properties of Cause-Effect relation are related but different from other relationships, which plays a fundamental role based on their ability [1]. Causality defines relations among regularly correlated phenomena (p_1 and p_2) or two events (e_1 and e_2), such that the existence of p_1 or e_1 results in the occurrence of p_2 or e_2. A phenomena or event is expressed as a phrase, nominal, and small span of text in a sentence or different sentences [2]. However, more concisely, the idea of causality is tricky to describe [3]. Philosophers have copied the concept for periods. In modern the dictionary of sociology [4], a traditional definition of Cause-Effect relation can be found as:

An Event or Events that come first and results in the existence of another Event. Whenever the first event (the Cause) happens, the second event (the Effect) essentially or certainly follows. As well as the same is possible whenever the first event (the Effect) happens, the second event (the Cause) essentially or certainly follows.

By the idea of much causation, numerous possible causes may be seen for a specified event, any one of which may be enough but not essential condition for the existence of the effect, or an essential but not enough condition. Similarly, numerous possible effects may be seen for a specified event. Any one of which may be enough but not essential condition for the existence of the cause, or an essential but not enough condition.

Causality can be grouped into causality understanding tasks [5] and causal discovery tasks [6] for event pair in the text. Understanding the causal relationship among daily events is a fundamental task for common-sense understanding language, e.g., "Ali lost the match; the crude got angry." and causal discovery. Understanding the possible causality among events pair can play a significant role in Question Answering [7,8], Event-Prediction [9,10], Generating Future Scenarios, and Medical Text Mining [11,12], Decision Processing [13], Adverse effects of drugs [14], Machine Reading and Comprehension [15,16], and Decision-support jobs [17], Information retrieval [18], and introduces another fact of information extraction through its inherent ability to discover new knowledge in a wide range of disciplines. Major study field of causality are Medicine [19], Computer Science, Biology [20], Environmental Sciences [21], Psychology [22], Linguistics [23,24], Philosophy [25], and Process Extraction [26].

Causal discovery [6] often described as the detection of the Cause-Effect relation between events, which is a highly trending topic in different fields, which is targeted in different time period throughout the world. Causal discovery is performed by using Google's search keyword-based survey and Google trend (GT) based survey approaches. In Google's search keyword-based survey approach, researchers used key terms related to a specific topic for analyzing the problem, while the Google trend-based survey approaches can be used for frequently searched and top topic trends. GT search can be beneficial for reflecting community and public interests throughout diverse periods [27]. Studying such trends by data mining techniques might deliver valuable intuitions and remarks regarding causality mining. GT does not deliver several queries on daily basis. Instead, it gives a standardized figure between 0 and 100, where 0 denotes a low volume of data for the query and while 100 denotes a maximum approval for the terms [28]. We have cited some impotent paper using GT approach related to causality in the ML and DL section. The motivation of this survey is to deliver an extensive intuition of Cause-Effect relation by using shallow ML and DL approaches with comprehensive coverage of causality problem including Basic Concept and Types of causality, Representation of causality in text, Applications of causality, Data types, Extraction/Mining techniques, Comparison among different techniques, Future challenges, and Collective properties of others relations.

1.1. Concept and Representation of Causality

In many domains, including all disciplines, knowledge growth could be valuable to discover previously unknown relationships between entities and events. Moreover, the fundamental property of CM is how to represent causality in sentences. Hence, the simplest ways of representing causality are using propositions of the form, 'A causes B', 'B causes A', 'A is caused by B', and 'B is caused by A'. Many experts belonging to this field is often disagreeing with those representations for causally linked events. Similarly, it is necessary and understandable to express causality using different types. Causality can occur in numerous forms. The best common differences are i) Marked and Unmarked causality and ii) Implicit and Explicit causality [29–31]. In Marked causality, the linguistic signal of causation exists, e.g., "I won the prize because I was lucky", here, because is the marked causality. In the unmarked causality, there is no linguistics signal of causality, e.g., "Run gradually. There are slops", in this example there is no linguistics signal of causality. Also, in implicit causality, both cause and effect events are not explicitly stated, e.g., "The bullet is hit on his head", in this example, the cause and effect are not explicitly stated. In explicit causality, both cause and effect are explicitly stated, e.g., "The accident has been caused by heavy rain". In this example, both cause and effect events are explicitly stated. Further, explicit causality is represented by propositions (e.g., active, passive, subject-object, and nominal or verbal) and uses various syntactic representations. Linguistic literature identifies the following ways to express Cause-Effect relations [8] explicitly. Including Causative Adjectives and Adverbs [32], Resultative Constructions [33], Causal Links [34], If-Then Constructions [29,35–37], and Causative Verbs [38]. In Figure 1 the more concise representation of causality in the natural language text is presented.

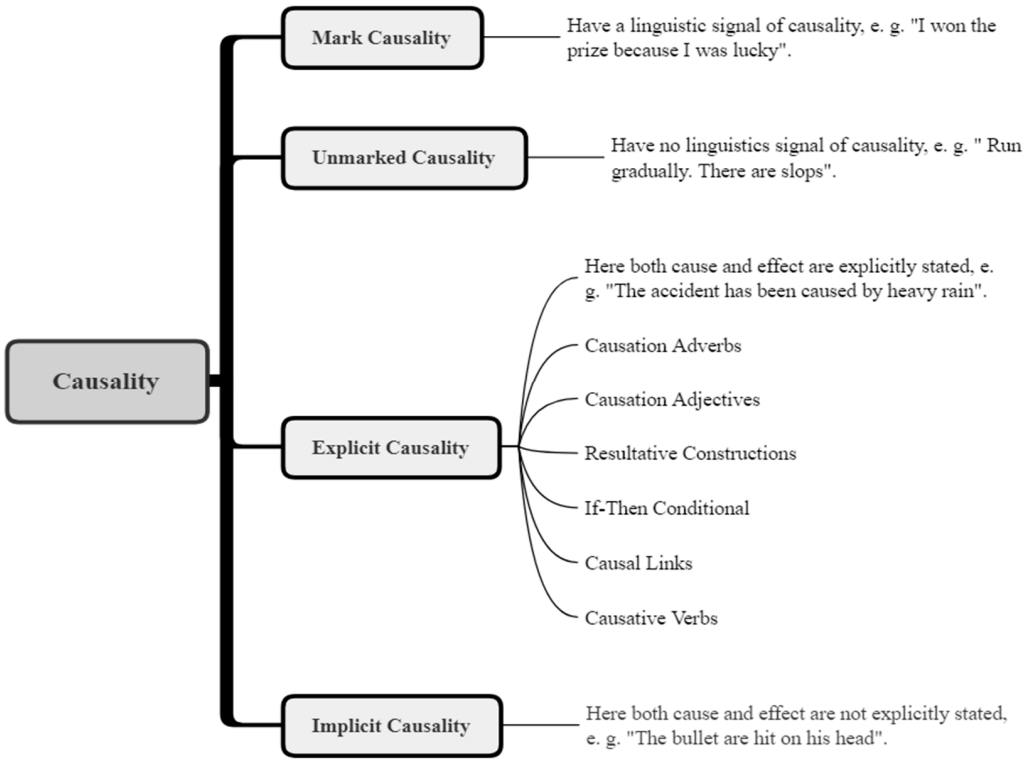

Figure 1. Representation of causality in the natural languages.

1.2. Research Contributions

Relation mining in NLP tasks is a vast research field in which causality plays an important role. The objective of this paper is to address a substantial research experience in the field of CM. However, around a number of prior surveys works in the field of CM were published. In [6], a review of basic theory of cause-effect relationships through structural causal models/networks are presented. In [39] a high-level views are presented about the

current formal complications and frameworks for causality learning. Ref. [40] described an Ensemble and Decision Trees (DT) ML networks for causality learning. Another notable survey motivated on mining causality for bivariate data [41]. In [42], different approaches for causality mining are summarized in time series data, on the other hand they targeted numerous semi-parametric score grounded techniques. In [38], a limited number of rule-based and statistical-based approaches are reviewed and overlooked the DL approaches. Most recently, in [43], a review of problems and techniques are presented, which mostly targeted the same traditional and statistical approaches. Different from prior review studies, this review addressing the shallow ML and DL techniques used in the latest research-oriented papers, various deep learning frameworks, various data types, researcher's experience, and models in research applications. The earlier surveys fail to deliver a complete discussion and a comparative analysis of shallow machine and deep learning-based approaches for CM, which specifically this survey aims to report. The current challenges in CM are to train massive implicit, ambiguous, and domain-independent datasets available at hand, which lead to causality as a critical task. In the light of ML and DL approaches, this article presents current perspectives and challenges in the field of causality that require more concentration such as optimization, scalability, power, and time to guide and educate practitioners and researchers in the area for the future development. To provide detailed and comprehensive attention to the issues above, different outlines are presented in this article including,

Sketched taxonomy for CM (Figure 2), which includes approaches using shallow ML and DL approaches.

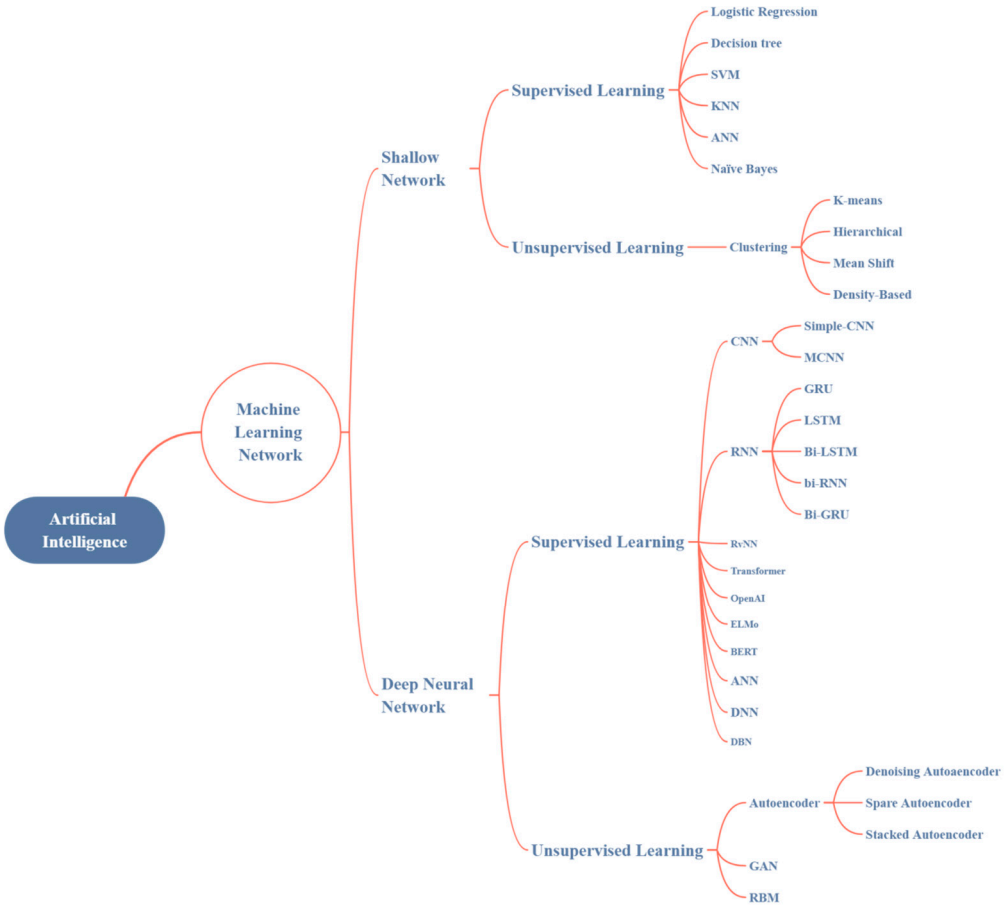

Figure 2. Taxonomy of most commonly used machine learning network for causality mining.

Brief historical literature of both shallow ML and DL approaches with their popular categories and summarized them.

Taxonomy of shallow ML and DL approaches are presented by evaluating present solutions using proposed classification.

Describes some common comparisons among all approaches, which lead us to mention some key research challenges and future direction in the field.

To get the objective done we follow the research methodology as follows, (1) we identified the most promising areas focusing on the issues of causality mining such as Shallow ML and DL approaches. (2) We designed our search criteria for extracting the articles of interest from the selected libraries. (3) We critically analyzed each article according to our designed review criteria. (4) Finally, we evaluate and summarize both competing paradigms.

The rest of this survey is organized in Figure 3. Section 2 provides brief literature and discussion about Machine Learning Techniques for CM. In Section 3, Deep Learning models, Frameworks, and techniques are discussed. Section 4 comparing both Machine and Deep Learning paradigms. Section 5 summarizes the key challenges and future guidelines. Finally, we conclude from Section 6.

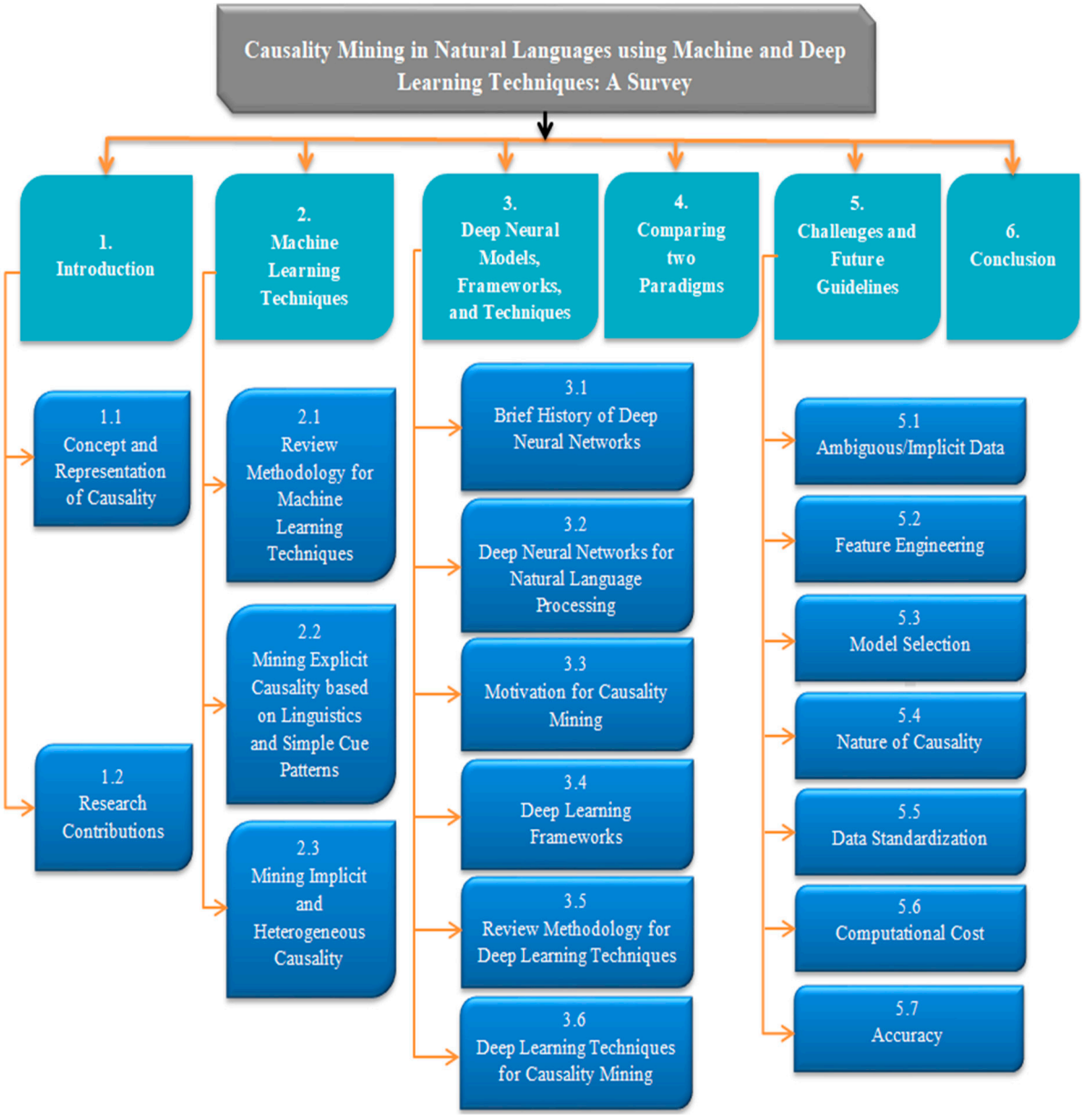

Figure 3. Organization of survey.

2. Machine Learning Techniques

Machine learning approaches are commonly divided into supervised learning (SL) and unsupervised learning (USL). The SL is based on labeled data, which contains some valuable information for the model. Text classification/mining is a common task in SL and is most frequently used in CM. Though, manually labeling the data is more cost-effective and time-consuming. Therefore, the absence of enough labeled data forms the major bottleneck to SL. On the contrary, USL mines the key feature knowledge from unlabeled data, making it much easier to gain training data. Though, the discovery performance of USL techniques is usually lower as compared to SL. Taxonomy of the most common shallow ML and DL algorithms used in CM is presented in Figure 2.

ML algorithms help progress the learning performance, simplify the learning process, and increase the possibilities of diverse applications. However, since the 80s, most studies relied on finding explicitly marked causality and cause-effect event pairs in domain-specific corpora, annotated manually by ruled-based/non-statistical techniques. Hence, at the beginning of the 2000s, there was sudden wholesale transformation. The paradigm of manual CM is shifted to ML approaches, which replaced the non-statistical approaches or enhanced with automatic feature engineering. Over time, researchers progressively began to account for implicit, heterogeneous, and ambiguous constructs through careful feature extraction by using a massive amount of textual labeled and domain-independent datasets to automatically extract implicit patterns in the text, demonstrating that ML approaches could potentially perform superior to purely linguistics-based approaches. In this section, a few distinctive ML techniques and frameworks are discussed. Table 1 summarizes the primary approaches, contributions, and limitations of linguistics-based and simple cue-pattern based ML approaches for CM. Similarly, Figure 4 represent the processing levels of shallow ML techniques, which consist of different parts, including Target Data Sources, Pre-Processing, Manual Features Designing and Extraction ((Principal Component Analysis (PCA), PPCA (Probabilistic Principle Component Analyzers), ICA (Independent Component Analysis), and GAM (generalized additive models)), Shallow Machine Learning Networks (Supervised Learning, Unsupervised Learning), Training and Testing, Prediction, and Performance Evaluation.

Table 1. Summary of linguistics and Cue phrase-based ML approaches.

SNo	Reference	Description	Pattern/Structure	Applications	Data Corpus	Languages	Limitations
1.	[7]	Improved version of C4.5 decision tree is used [44].	A pattern of causative verbs NP_1-verb-NP_2 are used.	Question Answering.	Domain-independent text. LATIMES section of the TREC 9 text group.	English	Not mentioned.
2.	[29]	A supervised approach for explicit causations.	Syntactic patterns (Phrase-relator-Cause).		SemCor 2.1 corpus for training.	✓	Only considered marked and explicit causations.
3.	[30]	Decision trees are used over POS-tagged data, and WordNet is used for mining semantic relations.	WordNet and POS-tagging features based features	Knowledge acquisition for decision making.	SemEval 2010 Task # 8 datasets (7954 instances for training and 2707 for testing) [30].	✓	Cost much more time in feature extraction.
4.	[45]	Syntactic parser for NP_1-Verb-NP_2 relation and WordNet knowledge base are used.	Used NP_1-Verb-NP_2 relation.		Penn Treebank dataset.	✓	Lack of ambiguity resolution and use of small dataset.

Table 1. *Cont.*

SNo	Reference	Description	Pattern/Structure	Applications	Data Corpus	Languages	Limitations
5.	[46]	Identifying relations among two-word noun compounds	Nouns pair patterns	Information retrieval, Information extraction, Text summarization	Bio-medical Text	✓	Only for nominal compound relations
6.	[47,48]	Use of Connexor dependency parser to extract NP_1-CuePhrase-NP_2 for inter-sentence relation.	NP_1-CuePhrase-NP_2 Pattern, cue phrase, and lexical pair probability.	✓	Five million articles from LA TIMES and WSJ for training set, two manually annotated test sets, including, WSJ article and Medline medical encyclopedia of A.D.A.M.	✓	System recall or F-score are ignored and no explanation of the use of NBC is provided.
7.	[49]	SemEvel2007 task-4 is applied for finding 7 frequently occurring semantic relations.	Events pair patterns of 7 relation types.	✓	Benchmark dataset to let the evaluation of diverse semantic relation classification algorithms		Only restricted to nominal based classification
8.	[50]	'PRE POST' model, extracted common-sense knowledge for the problem of CM.	Use Pre- and Post-condition pattern and SVM classifier	Knowledge acquisition for AI tasks.	Web text.		Based on a small set of labeled data.
9.	[51]	The similar SemEval-2010 task-8 used separate rule-based features for every type of relation.	Prepositions and verbs present among every nominal pair in combination with WordNet.	For information retrieval between nominal.	Training data of 8000 sentences, and test data of 2717 sentences	✓	Not specific to implicit causalities.
10.	[52]	Conditional text generation model.	Causal patterns and Cause-Effect graph.	Cause-effect event pairs generation.	Causal Bank corpus.	✓	Targeted only cause-effect event pairs

2.1. Review Methodology for Machine Learning Techniques

The following journal libraries have been exposed for this survey:

- IEEE Xplore Digital Library
- Google Scholar
- ACM Digital Library
- Wiley Online Library
- Springer Link
- Science Direct

We have cited over 100 popular papers from the above libraries and have shortlisted about 45 articles on CM, which focuses on shallow ML only. The search keywords used in these libraries include Causality Mining, Causality Classification/Detection, Cause-Effect relation classification with ML, Cause-Effect Event pair detection with ML. In this section, our goals are to study ML techniques focused on CM.

2.2. Mining Explicit Causality Based on Linguistics and Simple Cue Patterns

This section elaborates on some important work using Linguistics & simple Cue Pattern approaches for explicit CM. In such a direction, [45] was the first attempt using a syntactic parser of the NP_1-verb-NP_2 patterns. They used nine nouns hierarchies of semantic features for each value of NP_1 and NP_2 through WordNet knowledge base. The nouns hierarchies included entity, psychological feature and several constraints ranked based on frequency and accuracy. Ref. [46] Identified semantic relations at an intermediate

level of corpus for noun compounds description through multi-class mining approach that focused biomedical text because it preserves exciting challenges. In [48], a Naive Bayes classification at the lexical pair probabilities level is used to distinguish between various inter-sentence semantic relations. The purpose of this technique was a robust model for discourse-relation mining. They train a family of Naïve Bayes Classifiers (NBC) on an automatically generated set of samples in English sentences without annotations, and BLIPP English sentences, which is available at Linguistic Data Consortium (http://www.ldc.upenn.edu/).

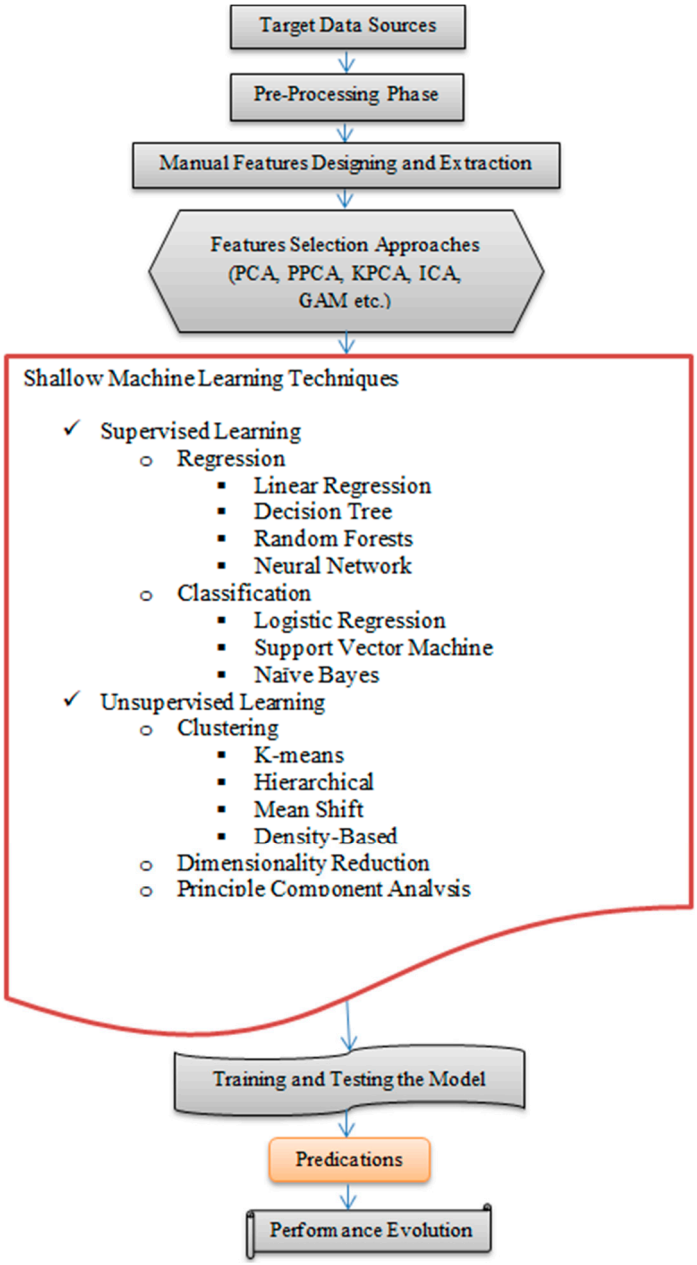

Figure 4. Processing levels of ML techniques.

Contrary to [53], Girju's [7], the ML model for the same problem is relatively a straightforward modification by using supervised method C4.5 decision tree, to change the semi-supervised pattern validation and ranking procedure, which consists of the identification of semantic constraints on each pattern ranking and causal pattern. For categorizing the noun-modifier relationship [54], they used a novel two-level hierarchy using common text in the base noun phrases. They targeted semantic similarity between

base noun phrases in clusters, determined by an extensive set of semantic relations. In [47], the followingnotable work is presented by using Connexor Dependency Parser (CDP) [55] to mine ternary expressions of the form NP_1-CuePhrase-NP_2, where lexical pair is created from NP_1 and NP_2 together using 60 causal verbs found in [7,53]. Such ternaries were filtered with a set of pre-defined cue phrases to get highly ranked ternaries. The initial classifier used cue phrase confidence scores treated as the initial causality annotated training set for NBC. The downside of this work is using a large unlabeled corpus to mine causality, but their evaluation results and error analysis left a little something to be desired.

They improved and focused only on the model precision, but they did not quote the model recall or F-score, which was not a good idea. The "Phrase-relator-Cause" patterns used the marked open-domain text and explicit causations [29]. The words such as due to, cause, after, because, as, due, and since are measured as the relators in the given patterns. They used semantically annotated corpus "SemCor 2.1" for training a C4.5 decision tree binary classifier. They targeted seven features for learning such as the type of relater, the tense of verbs, causal potentials of verbs, semantic classes of verbs, and its modifiers. With the beginning of benchmarked corpora for numerous NLP tasks, such as SemEval-2007 task-4 and SemEval-2010 task-8, the semantic relation mining, including causality, has been enhancing, and additional new approaches have come into existence. The winners of [56] for SemEval-2007 task-4 and [49] for SemEval-2010 task-8 used a combination of syntactic, semantic, and lexical features mined from numerous NLP toolkits, such as Syntactic Parser and knowledge-bases, and used SVM as the classifier. Though the mining outputs for causality are satisfactory, most of the samples in the corpus are simple causality, which is explicitly expressed with essential linguistic clues such as to cause, result in, due to, because, and lead to. The mining of implicitly express causality is still a big challenge where task-4 [57] was used for seven frequently occurring semantic relations mining including cause-effect, instrument-agency, theme-tool, part-whole, product-producer, origin-entity, and content-container. Moreover, [49] classified multi-way semantic relationships between nominal.

In [50], extracting common-sense knowledge for CM by post and pre-condition of actions and events from web corpus using 'PRE POST' tool with its post and pre-conditions. Ref. [51] Present WordNet-based semantic features in conjunction with separate role-based features for each type of relationship through verbs and prepositions existing among each nominal pair. For a broad analysis of this task, one can visit [30]. Training decision trees (DT) on SemEval 2010 corpus [58] with POS-tagging and WordNet-based dependency relations. They achieved an F-score of 0.858, and to train the CRF. The obtained F-score is 0.52 recorded. In [52], a conditional text generation network is suggested that make sentential terms of possible causes and effects for any free-form textual event. This is based on two resources; an extensive pool of English sentences that denote causal patterns (CausalBank) and large lexical causal knowledge graphs (Cause-Effect graph). They focused on explicit relations within a single sentence by linking one part of a sentence to another and using generated patterns instead of sentence-level human annotation. This approach performed superior on diverse causes and effects events in new inputs by automatic and human assessment. Table 2 summarizes some linguistics and Cue phrase-based patterns ML approaches covering Description, Pattern/Structure, Application, Dataset, Languages, and their Limitations. In [49,57], semantic classifications are performed, 'not more specific to the causality problem. Further, in [50], extracting common-sense knowledge by mining post- and pre-condition of actions and events in the web text by using their Pre-conditions and Post-conditions (PRE POST) for CM. They mine causality with their Pre-conditions and Post-conditions. Moreover, to save out extremely correlated words that are neither pre- or post-conditions, the model determines the set of feature words that describe the association among the candidate and action words by getting their three-way PMI with every feature word. They used SVMs by considering five action words for training and 35 for testing. A small set of the labeled corpus is considered for learning and generalize

well to unseen actions and events and also captures Preconditions relations, which are not directly captured by any of the prior approaches.

Table 2. Reviews ML approaches for implicit and ambiguous causalities.

SNo	Reference	Description	Pattern/Structure	Application Domain	Data Corpus	Language	Limitation
1.	[2]	Network (CausalNet) of cause-effect terms in a large web corpus.	Linguistic pattern, 'A ($event_1$) causes B ($event_2$)'.	Predication in short text.	10TB corpus from Bing.	English	Over-fitting issues.
2.	[9]	Pundit algorithm for future events prediction.	Handcrafted rules.	Predictions	News corpus last from 150 years news reports.	✓	It only applicable to textually denoted environment.
3.	[14]	Proposed ADRs.	Lexical patterns.	Healthcare field to decreases drug-related diseases.	Twitter and Facebook data.	✓	Worked only for explanatory messages related to drug and diseases.
4.	[31]	Applied pattern matching by phrasal and causative verbs that links ML and traditional methods.	Syntactic patterns	Used for large scale AI problems of events prediction.	News articles over 150 years old.	✓	Use of unrelated data, which result irrelevant causality predication.
5.	[59]	Extracting parallel and temporal causal relations, and differentiate among them	Feature based on WordNet and the Google N-gram corpus.	Decision making.	Their own corpus of temporal and causal relations .	✓	Hard to perform well on domain-independent data
6.	[60]	Discovered parallel temporal and causal relations.	PDTB, Prop Bank, and Time Bank data patterns.	Decision making.	Their own annotated corpus	✓	Overlooked in-depth analysis of both corpus and relations.
7.	[61]	A graphical framework for implicit causalities.	Semantic, lexical, and syntactic features.	Information retrieval in NLP.	Same corpus used [59].	English	Some vague verbs cause most of the errors.
8.	[62]	A distributional and connectives probability approach for event causality detection.	Follow features described Ruby-based discourse System [63].	Decision making.	Using news articles collected from CNN (http://www.cnn.com).	✓	More focused on explicit connective, and overlook implicit connective.
9.	[64]	Classifying causality among the verb and noun pairs.	Grammatically linked verb-noun pairs pattern based on extra knowledge with Linguistic features.	Prediction	Acquired 2 158 causal and 65, 777 non-causal from FrameNet.	✓	Bound to limited feature.
10.	[65]	MLR (The source code for relation mining is available in https://github.com/YangXuefeng/MLRE), mine all probable causality with any preposition or verb based.	Constituent and linguistic knowledge of the dependency grammar.	Extract causality in all language expression levels.	Prop bank [66],	✓	Small manually annotated dataset, typically lead to over fitting problem.

Table 2. *Cont.*

SNo	Reference	Description	Pattern/Structure	Application Domain	Data Corpus	Language	Limitation
11.	[67]	Mine causal and temporal relations, and propose guidelines to annotate casualty.	Used <CLINK> tag to indicate a causal link, and presented the idea of causal signals through the <C-SIGNAL> tag.	Prediction, risk analysis, and decision making.	Annotated dataset followed [68] guidelines.	✓	Complex annotation scheme.
12.	[69]	RHNB algorithm manages interactions among diverse features.	RHNB model based patterns	Prediction	SemEval-2010-Task8 dataset.	✓	Work on large set of feature vector, which usually slow the model processing.
13.	[70]	Explicit discourse connectives for mining alternative lexicalizations (AltLexes) of causal discourse relations.	Two kind of features: Parallel corpus derived feature and lexical semantic features.	Question Answering and text summa-rization.	Wikipedia from 11 Sept 2015.	✓	//
14.	[71]	CRF based model for CM.	Time-based sequence labeling, Lexical and syntactic features.	Emergency management.	Emergency cases corpus about typhoon disasters.	✓	Based on raw corpus which leads to low performance.
15.	[72]	First effort toward German causal language.	Annotated training suite and lexicon.	Identify new causal triggers.	English-German part of Europarl corpus [73].	German causal language.	Only focused English-German parallel corpus.
16.	[74]	BECauSE 2.0 corpus with broadly annotated expressions of causal language.	Annotated expression of causality.	Annotating causal relation.	BECauSE 2.0 corpus (https://github.com/duncanka/BECauSE).	English	Missing semantically fuzzy relations.
17.	[75]	CausalTriad, to mine causalities	Traid structures.	Medical related predication.	Health Boards dataset (https://www.healthboards.com/) and Traditional Chinese Medicine dataset.	✓	Only used for medical domain, and not useful in other domains.
18.	[76]	TCR, a joint inference model for understanding temporal and causal reasoning.	Using CCMs and ILP in the extraction of temporal and Causal relations.	Decision making in defense department.	Causal and temporal relations from the text (http://cogcomp.org/page/publication_view/835).	✓	Omitted the concept of jointly learning of temporal and Causal relations.
19	[77]	Extracting causality		Investigation of Malaria Epidemics	HAQUE-data and HANF-data	✓	Only targetd malaria related problems

2.3. Mining Implicit and Heterogeneous Causality

From a few decades, heterogeneous, implicit, and ambiguous causalities became a concentrated area for researches. If we move back toward past research in the field, most of the work has been targeted at explicit causalities through statistical techniques, and implicit causalities are ignored. Implicit causality was first tried by [59] to deal with implicit causalities in a well-organized way. Their model answered such questions by considering a sentence and taking two events occurring in that sentence, in which one event could be considered the cause of the other. They create parallel temporal and causality corpus and distinction among temporal and causal relations. Semantic and syntactic features are used for temporal relations, which encoded complementary indications that augmented

the knowledge of temporal relations to progress CM. Such an attempt achieved 0.49 and 0.524 F-score for both temporal and causal relations. Another work [60] explored parallel temporal and causal relations. They designed a corpus of causal and parallel temporal relations to fill a gap in the relation configuration annotated by present resources including, Penn Discourse Treebank (PDTB) [78], PropBank [66], and TimeBank [79]. This work defines the annotation of a corpus for both relations types, with an initial effort on the conjoined event creation. Such creation is often used to presents both causal and temporal relations. This was an opening idea to explore connections between causal and temporal relations. In the past, both causal and temporal relations keep the same conclusion. On the other hand, it was difficult to find causal and temporal relations in the arbitrary corpus, but finding these relations are more manageable in a wisely nominated subset of corpora.

Continuing toward implicit causalities, [61] suggested a graphical framework by catching contextual information pair of entities and discovering dependencies among different syntactic, semantic, and lexical features. Further, encode such features for molding the CM tasks in a graphical representation. They find related graph patterns that capture two events for a given pair in the same sentence for contextual information. This approach focused on setting up an ML model to learn every pair of events for causal or non-causal predication. Similarly, patterns for causal information are descriptive between two events. In the same way, the distributional and connectives probability approach is presented [62] for implicit CM. A pundit approach for future events prediction using handcrafted rules at newspaper headlines that last from 150 years of news reports collection [80]. Contrary to past approaches, the novelty of this work is considering a general-purpose mining algorithm, combines diverse web sources, and concentrates on future event predictions generation to improve and generalize historical events. For advanced research, one can visit a link (http://www.technion.ac.il/~kirar/Datasets.html). In [31], lexico-syntactic features and self-constructed rules are applied for CM. Such rules are kept precise by dependency structure and lexico-syntactic patterns to mine possible cause-effect pairs, and further, Laplace smoothing classifier is used to reject incorrect event pairs. In [81], implicit causal and non-causal relationship mining among verb-verb pairs is performed and produced a training corpus of causality between verbs and trained by a supervised system. The former approach is extended by [64], which classifies causality between the pattern of the verb-noun pair. Firstly, they recognize all nouns and verbs in the target sentence and then use a classifier to classify implicit causalities among grammatically linked noun-verb pair patterns. Similarly, a multi-level relation mining algorithm (MLRE) is presented to mine possible causalities with any verb or preposition-based linguistic pattern [65]. They used some lexical knowledge bases features, and feature selection approaches for learning.

In [67], a motivating work is presented to mine causal and temporal relationships among events pairs and proposed guidelines to annotate casualty among different events. This approach used an annotated corpus based on the suggested guidelines. In [69] causal connectives are applied, which is obtained by computing the similarity of sentences syntactic dependency structure through Restricted Hidden Naïve Bayes (RHNB) classifier to manage the interactions among lexico-syntactic patterns and causal connectives. Contrary to [31,82], this approach accounts for more significant features. In [2], a causality network of terms is produced from a group of web corpus by a linguistic approach, such as 'A causes B'. In such a graph, each node designates a term and each edge possesses a causal co-occurrence score. Finally, in the causal graph, the co-occurrence scores between terms compute the causal strength and using a co-occurrence score for CM. The same year [70] considered comparable monolingual corpora of simple and English Wikipedia of PDTB [78]. They used explicit discourse connectives for mining alternative lexicalizations (Altlexes) of causal discourse relations.

Next, a conditional redundancy field (CRF) based model is suggested by [71], which redefined the time-based sequence labeling process for CM. The proposed algorithm used LTP (Load, Transform, and Processing) technique by mining raw corpus related to emergency cases. Then, the mined corpus is used for causality candidates by using the

feature templates. The experimental result shows the practical impact of mined causality in the sentences. Among all earlier approaches, [83] is an innovative approach that employs graph LSTMs to classify relationships across sentences by building a document graph through dependency link and syntactic features among the root nodes of the parse tree. The first effort toward describing German causal language [72], creates resources, which contained annotated lexicon and training suite. Such an approach mine new causal triggers for automatic CM in English-German parallel corpus with negligible human management. The proposed approach is similar to identify transitive causal verbs, where the English verb has been taken as a seed source of causality.

In [74], the "BECauSE 2.0 corpus (Link for source document, at https://github.com/duncanka/BECauSE) is considered an extended version of the BECauSE 1.0 corpus with broadly annotated expressions of causal language. It comprises Penn Treebank [84], the New York Times corpus [85] that contains 59 randomly selected articles from Congress, "Dodd-Frank 679 sentences" transcribed [86], and manually annotated Sub-Corpus [87]. Contrary to BECauSE 1.0, the overall performance of this work is significantly enhanced by an F1-score 0.77 for causal connectives. The subsequent notable work using medical corpus through the Causal-Triad approach for CM [75]. The earlier works employed diverse mining techniques to discover pseudo causality in a single sentence, but causation transitivity knowledge often lies between sentences that were not considered. Furthermore, the rules of causation transitivity are followed by much pseudo causality, to yield new causal hypotheses and mine some hidden causality. They used Health Boards (HB) and Traditional Chinese Medicine (TCM) datasets. In the same year, [14] presented a causality reaction for learning adverse drug reactions (ADRs), in Twitter and Facebook platforms to automatically extract lexical patterns that denote the relationship between events and drugs. This work aims to notice a contrary response caused by a drug instead of a correlated sign based on causality measures. ADR mining has many applications for the usage of a direct implication of drug. Using past ADRs detection evidence and provided it to pharmaceutical companies, regulators, and health care departments. By the way, we can decrease drug-related diseases. Similarly, [76] presented an imperative natural language task for understanding temporal and causal relations between events using Constrained Conditional Models (CCMs). From a literature perspective, the effect must occur after the cause of the closely related temporal and causal relations, in which often one relation dictates the value of the other. However, to study these two relations, limited attention has been given in the past. However, the problem is formulated as an Integer Linear Programming (ILP) problem. The joint inference framework got significant improvements in extracting both causality and temporal relations in the text (The source code and the dataset at http://cogcomp.org/page/publication_view/835). In the joint system, the score of temporal and causal is added up for all event pairs. The temporal performance got strictly better in a recall, precision, and F1. The causal performance also improved by a large margin, representing that causal and temporal signals are helpful in each other.

In this section, we are including some GT-based approaches, which are beneficial for this causality problem including, [88], which examined the behavioral effect of the Internet search capacity for the financial crisis and oil prices on food price instability. They used GT to derive the subject variables that can be used to assist and describe food prices. This can help the market contributors tend to respond fast to information on the Internet to adjust to the new market situation. They used keywords in the Google search to discover the relationship between the overall agricultural price level and query volume. Where the market prices are designed by behavioral trends, subject to geographic areas and the interior dynamics of the country. In [77], five technique are used for causality extraction for the purpose to investigate malaria epidemics in dynamic processes using information encapsulated in time series by statistical techniques. This paper investigated transfer entropy [89], recurrence plot kernel [90], Granger causality [91,92], and causal decomposition and complex models [93,94]. They used (HAQUE-data) and (HANF-data) [95,96]. Lastly,

Table 2 summarizes major ML approaches for implicit and ambiguous causalities that cover Description, Pattern/Structure, Application, Data coups, Languages, and Their Limitations.

GT can be used to infer the popularity of dengue, which summarizes Google searches of associated subjects. Together the infection and its GTs have the same source of causation in the dengue virus, leading people to hypothesize that dengue occurrence and GTs outcomes results have a long-run equilibrium (LQ). This work considers the principle of LQ by using GT information for the primary discovery of future dengue outbreaks. They used the cointegration technique to evaluate LQ between dengue occurrence and GTs outcomes. The LQ is categorized by its linear arrangement that produces a stationary process. The final models are used to determine Granger causality among two processes. The model presented the direction of causality of the two processes, representing that GT outcomes will Granger-cause dengue occurrence (not in the reverse order). This study concluded that GTs outcomes can be used as the first indicator of future dengue outbreaks [97].

3. Deep Neural Models, Frameworks, and Techniques

Deep learning models for CM can enhance the performance of learning algorithms, improve the processing time, and increase the range of mining applications. However, the highly extended training time of the DL models leftovers a significant challenge for researchers. DL models combine the optimization, distribution, modularized techniques, and support to setups. These concepts are developed to streamline the implementation process and improve system-level progress and research. This section described the history of Deep Neural Networks, Deep Neural Frameworks (DNFs), and some effective Deep Learning Techniques for CM.

3.1. Neural Networks and Deep Learning

In the era of information processing tasks, ML has been merged in many disciplines, including information mining, relations classification, image processing, video classifications, recommendation, and analysis of different social networks. Including all ML algorithms, Neural Network (NN) and DL are identified as representation learning [98] extensively used. NN computes a result/predication/output, which generally states forward propagation (FF). During FF, the NN receives inputs vector X and result in a prediction vector Y. More generally, NN is based on interconnected layers (input, hidden, and output layer). Each layer is linked via a so-called weight matrix (W) to the next layer. Further, each layer consists of different combinations of neurons/nodes, where each node gets a particular number of inputs and computes a prediction/output. Every node in the output layers makes weighted addition based on received values from the input neurons. Further, the weighted addition is passed to some nonlinear activation functions (Sigmoid, Tan Hyperbolic (Tanh), Rectified Linear Unit (ReLU), Leaky ReLU, and Softmax Activation Function) to compute outputs. Figure 5 represents a simple NN with one input layer, three hidden layers (H_1, H_2, and H_3), one output layer, and four weight matrices (W_1, W_2, W_3 and W_4).

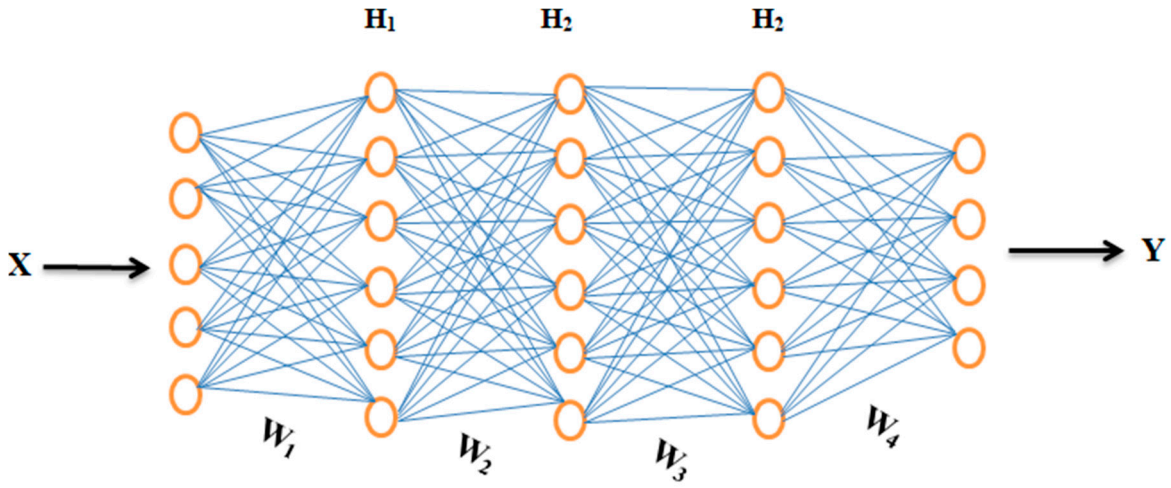

Figure 5. Represents a simple NN of one input layer, three hidden layers, and one output layer.

We set an input vector X to calculate dot-product by the first weight matrix (W_1) and used the nonlinear activation function to the result of this dot-product, which output a new vector h_1 that denotes values of the nodes in the first layer. Further, h_1 is used as a new input vector to the next layer, where similar operations are executed again. This process is repeated until the final output vector Y is produced, known as the NN prediction. While Equations (1)–(4) represent the whole set of operations in NN, where "σ" denotes an arbitrary activation function.

$$\vec{h}_1 = \sigma\left(\vec{x}.\ W_1\right) \tag{1}$$

$$\vec{h}_2 = \sigma\left(\vec{h}_1.\ W_2\right) \tag{2}$$

$$\vec{h}_3 = \sigma\left(\vec{h}_1.\ W_3\right) \tag{3}$$

$$\vec{y} = \sigma\left(\vec{h}_3.\ W_4\right) \tag{4}$$

More generally, we consider NN as a function instead of using a combination of interconnected neurons. With this function we combine all operations in a chained format in Equation (5) that we have seen in the above four equations.

$$\vec{y} = NN\left(\vec{x}\right) = \sigma\left(\sigma\left(\sigma\left(\vec{x}.W_1\right).W_2\right).W_3\right).W_4\right) \tag{5}$$

3.2. Loss Functions and Optimization Algorithms

The selection of loss function and optimization algorithms for DL networks can significantly generate optimal and quicker results. Every input in the feature vector is allocated its particular weight, which chooses the impact that the specific input desires in the summation function (Y). In simple word, certain inputs are made more significant than others by assigning them more weight, which have a superior effect in Y. Furthermore, a bias (b) is added to summation shown in Equation (6).

$$Y = bw_0 + w_1x_1 + w_2x_2 + w_3x_3 \tag{6}$$

The outcome Y is a weighted sum is converted to performed output using a non-linear activation function (f_{NL}). In this case the preferred result is the probability of an event, which is represented by Equation (7).

$$\hat{p} = f_{NL}(Y) \tag{7}$$

In many learning models, error (e) is calculated as the gap between the actual and predicted results in Equation (8).

$$J(w) = p - \hat{p} \tag{8}$$

The function for error calculation is called Loss Function J (.), which significantly effects on the model predication. Distinct J (.) will provide diverse errors for a similar prediction. Different J (.) deals with various problems, including classification, detection, extraction, and regression. Furthermore, error J (w) is a function of the network/model's inner parameters (weights and bias). Precise likelihoods, it requires minimizing the calculated error. In NN, this is achieved by Back Propagation (BP) [99], in which the existing error is commonly propagated backwards toward the preceding layer, where optimization function (OF), including Stochastic Gradient Decent (SGD), Adagrad, and Adam is used to modify the parameters in an efficient way to the minimize error.

The OF calculate the gradient (partial derivative) of J (.) concerning parameter (weights), and weights are improved in the reverse direction of the calculated gradient. This process is repetitive until it reaches the minimum J (.). Equation (9) represents the optimization process.

$$w^{(k+1)} = w^{(k)} = \frac{\partial}{\partial w^{(k)}} J(w) \tag{9}$$

The basic differences between different models are based on the number of layers and the architecture of the interconnected nodes. In those models neurons are structured into sequential layers, where each neuron receiving inputs only from previous layers neurons, called Feed forward Neural Networks (FFNNs). Though, there is no clear consensus on precisely what explains a Deep Neural Network (DNN), networks with several hidden layers are known as deep and those with several layers are known as very deep [100]. Contrary to traditional and ML techniques, DL techniques have enhanced performance in computer vision (Image Processing, Video Processing, Audio Processing, and Speech Processing) [101–103], and NLP tasks (Text Classification, Information Retrieval, Event Prediction, Sentiment Analysis, and Language Translation) [104–108].

Usually, the effectiveness of shallow ML algorithms is based on the goodness of input data representation. Compared to precise data representation, the performance of depraved data representation is usually lower. Hence, for shallow ML tasks, feature engineering is an effective research direction in raw datasets and will lead to various research studies. Usually, most of the features are domain-dependent which need much human effort e.g., in computer vision tasks, diverse features are compared and proposed including Bag of Words (BoW), Scale Invariant Feature Transform (SIFT) [109], and Histogram of Oriented Gradients (HOG) [110]. Similarly, in NLP tasks, diverse features sets are used including BoW, Linguistics Patterns (LP), and Clue Terms (CT), Syntactic, and Semantic context. Contrary, DL techniques work on automatic feature engineering, which lets researchers get more discriminative features with minimal human effort and domain knowledge [111]. As discussed above that DL techniques are based on a low-level, middle-level, and high-level layered structure for data representation, where the low-level layers are used for low-level features, the middle-level/hidden layers are used to extract hidden/middle-level features. Finally, the high-level features are extracted by high-level layers.

3.3. Brief History of Deep Neural Network

Since 300 B.C, the beginning of DL was a dream of experts by making machines that mimic the human brain. At that time, Aristotle recommended 'associationism', which led to the history of human motivation by realizing the human brain. His idea needs researchers who focus on understanding the recognition system of the human brain. Though, the current history of DL has been ongoing since 1993, when the McCulloch-Pitts (MCP) model was proposed as a prototype of the Artificial Neural Model (ANM) [112]. They developed a Neural Network (NN) system, which mimics the human brain neocortex [100] by using a threshold logic system', which combines mathematics and algorithms that mimic the human's thought process but not learn. From that, gradually the era of DL is grown. After

the MCP system, Hebbian theory [113] is applied for biological systems in the natural environment, leading to the first electronic device known as 'perceptron' in the cognition system. Hence, at the end of the first journey of AI, the advents of 'back propagandists' led to another baseline. Similarly, Verbose presented 'back propagation' for errors analysis in DL networks, which lead to a novel direction in modern NN.

In 1980, 'neocogitron' was inspired by CNN [114], and introduced its first milestone of an RNNs system for NLP tasks [115]. Furthermore, LeNet' made it possible for deep neural networks (DNN) effort practically well [116]. Due to the limitation of hardware resources, the 'LetNet' did not perform efficiently on a larger dataset. In 2006, a layer-wise pre-training framework was developed [117], named Deep Belief Networks (DBNs). The basic idea of this framework was an unsupervised two-layer network including Restricted Boltzmann Machines (RBMs), which freeze all parameters, and the other is to put a new layer on the top to train the parameters only for the new layer.

Hereafter, gradually diverse DL networks came into existence, enhancing the performance of models in every field. After developing artificial neural networks (ANNs), we have seen many DL techniques come into existence. Recently, in every field DL is a leading approach compared to traditional and shallow ML approaches. At the same time, there is big revolution came into hardware technologies, called graphical processing units (GPUs) [118,119] and Tensor Processing Unit (TPU) [120], which uses ANNs with billions of trainable parameters [121] for increasing the computational power and parallelization of DL techniques. Compared to the CPU, the GPU and TPU have significant computing power on a single machine using many distributed DL networks [122–124]. Moreover, most of the corpora come in raw format without or with raw labels. Since most of the practices are based on semi-supervised and unsupervised tools for training DL networks for enhancing the raw nature of data. Similarly, most of the prior DL techniques emphasized only a single modality that leads to a partial illustration of public data. In such cases, many scholars concentrated on cross-modality structure for keeping DL in an advanced option [125].

3.4. Deep Neural Network for Natural Language Processing

This section discusses some widespread DL networks including Convolutional Neural Network (CNN) [101,103,106,126], Recursive Neural Network (RvNN) [127,128], Recurrent Neural Network (RNN) [129,130], Gated Recurrent Units (GRU) [131], Long Short-term Memory (LSTM) [132], Bidirectional Long Short-term Memory (bi-LSTM) [133], Transformer [134], Embedding from Language Models (ELMo) [135], OpenAI [136], and Bidirectional Encoder Representations from Transformers (BERT-base) [137]. In the field of NLP, all of them contributed much more novel development. Table 3 summarizes the most fundamental and common DL models by their name, applications, and references.

Table 3. Common Deep learning models for NLP tasks.

SNo	References	Deep Neural Networks	Applications and Structure
1.	[101,103,106,126]	CNN	CNN's are made upon Fukashima's neurocognition [138,139], where the name originates from the convolution operation in signal processing and mathematics. CNN's use some specific type of function called filters, which lets simultaneous analysis of diverse features in the source data [101,140]. Though, CNN is considered as the foundation and inspiration of DL approaches, which beats its predecessors. It is based on a mash structure of neurons/nodes for information exchange, leading to various many-layered learning networks. In the beginning, it was applicable for computer vision. Further, enhanced to NLP.
2.	[127,128]	RvNN	Like CNNs, RvNN uses a method of weight sharing to decrease training. Though CNN's share their weights within a layer (horizontally), RvNN share weights between layers (vertically). This is interesting because it lets easy modeling of parse trees structures. In RvNN, a single tensor of parameters can be applied at a low level in the tree and further recursively used sequentially at higher levels [141]. It is applicable for sequential NLP tasks by using a tree-like architecture.
3.	[142,143]	DBNs	Applicable for unsupervised learning-based directed connections.
4.	[144,145]	Deep Boltzmann Machine (DBM)	Applicable for unsupervised learning based on undirected connections.

Table 3. *Cont.*

SNo	References	Deep Neural Networks	Applications and Structure
5.	[129,130,146–149]	RNN	RNN is a type of RvNN, comprehensively used in many NLP tasks. Since NLP is dependent on the sequence of words such as sentences /phonemes, it is beneficial to have a memory of the preceding elements when processing new ones. Sometimes, backward dependencies exist that correct processing of certain words/tokens may depend on words that follow it. Hence, it is crucial for RNN to look at the sentences in the forward and backward direction and integrate their outputs. This organization of RNN's is known as a bidirectional RNN. This design may allow the effect of input to longer than a single RNN layer and letting for longer-term effects. This sequential design of RNN cells is known RNN stack [150]. RNN is applicable for sequential NLP tasks, and as well as for speech processing.
6.	[151,152]	Generative Adversarial Network (GAN)	Applicable for unsupervised learning and using game-theoretical context.
7.	[153]	Variational Autoencoder (VAE)	Applicable for unsupervised learning and based on the Probabilistic Graphical model.
8.	[131,154]	GRU	GRU is an extended version of RNN and a simpler variant of the LSTM, usually perform better than standard LSTMs in several NLP sequential tasks.
9.	[130,132,155]	LSTM	LSTM is one of the prominently enhanced forms of RNN. In LSTMs, the recursive neurons are consist of many different neurons linked in a sequential structure to preserve, expose, or forget some precise information. While standard RNN's of the single node serving back to them and have some memory of long passed outcomes, these outcomes are merged in each consecutive iteration. Usually, it is significant to remember data from the distant past, however and at the same time, other very latest data may not be vital. LSTM can remember important data much longer, while inappropriate data can be forgotten. It plays a very important in sequential computation.
10.	[133]	bi-LSTM	Bi-LSTM is an enhanced form of LSTM that works in both left and right directions to deal with the problem. It is applicable for sequential NLP tasks and uses derived features from lexical resources such as NLP and WordNet systems.
11.	[134]	Transformer	Encoder-Decoder pair is typically used for text summarization, machine translation, or captioning, results is in textual form. An encoding ANN is used to yield a vector of a specific length and a decoding ANN is used to return variable size text based on the vector. Issue in this system: RNN is enforced to encode the whole sequence to a finite length vector without affections to whether or not any of the inputs are more significant than others. A strong solution to this issue: Using the attention mechanism. The first prominent use of an attention mechanism is the condensed layer for an annotated parameter of RNN hidden state, letting the network obtain what to pay attention in accordance with the annotation and current hidden state [156]. It is applicable for supervised learning with multi-head attention.
12.	[135]	ELMo	They used a feature-based approach and task-specific architectures that contain pre-trained representations as additional features.
13.	[136]	Generative Pre-trained Transformer (OpenAI GPT)	Applicable for unsupervised learning by Improving language understanding.
14.	[137]	BERT-base	BERT-base is the enhance form of Transformer, which deal the source sentence in both direction. It uses a bi-directional encoder-decoder along with attention mechanism. It is conceptually very simple and empirically influential.

3.5. Motivation for Causality Mining

DL applications are resulted based on feature representation and algorithms together with the design. These are related to data illustration/representation and learning structure. For data illustration, there is typically a disjunction among what information is said to be essential for the task, against what illustration produces good outcomes. For instance, Syntactic Structure, Sentiment Analysis, Lexicon Semantics, and Context are supposed by some linguists to be of fundamental importance. However, prior works are based on bag-of-words (BoW) system proven satisfactory performance [157]. The BoW [158], frequently seen as vector space models, includes an illustration that accounts only for the words/tokens and their frequency of existence. BoW overlooks the order and relations of words and treats every token as a distinctive feature. BoW neglects syntactic format, still delivers effective results for what some could consider syntax-oriented applications. This judgment recommends that simple illustrations, when combined with a big data set, may work superior to difficult representations. These outcomes verify the argument courtesy of the significance of DL architectures and algorithms. Often the effective language modeling guarantees the advancement of NLP. The aim of statistical language designing is the probabilistic illustration of word sequences, which is a complex job because of dimensionality curse. In [159], a breakthrough for language designing with NN aimed

to overcome the dimensionality cures by learning a distributed illustration of tokens and giving a likelihood function for structures.

A significant challenge in NLP study, related to other areas including computer vision, looks complicated to reach an in-depth illustration of language using statistical/ML networks. A core task in NLP is to illustrate of texts (documents), which comprises feature learning, i.e., mining expressive information to allow additional analysis and processing of raw data. Non-statistical approaches are based on handcrafted features engineering, which is time-consuming. Through, the development of algorithms needs careful human analysis to mine and exploit instances of such features. While, deep supervised approaches are more data-driven and can be used in extra general efforts, which directed a robust data illustration. In the presence of huge amounts of unlabeled dataset, unsupervised learning techniques are known to be critical tasks. With the beginning of DL and sufficiency of an unlabeled datasets, unsupervised techniques become a critical job for representation learning. At present, many NLP tasks depend on annotated data, while most unannotated data encourages study in employing deep data-driven unsupervised techniques. Given the possible power of DL techniques in NLP tasks, it looks critical to analyses numerous DL techniques extensively.

DL models have a hierarchical structure of layers that learn from data representation by input layer, then pass them through multiple intermediate layers (hidden layers) for further processing [121]. Finally, the last layer computes the output predation. ANN is a representative network using FP and backward propagation (BP). FP is used for processing weighted sum (WX) of input from the prior layer along with bias (b) term, and further passes it to a sequence of Convolutional, Non-linear, Pooling, and Fully connected layers to produce the required output (final prediction). Equation (10) represents the fundamental matrices of the neural networks.

$$Z = A(WX + b) \tag{10}$$

where 'W' represents the weight (number matrix), also known as parameters, X represents the input feature vector, 'b' represents the bias term, 'A' represents the activation function, and Z represents the final prediction. Similarly, the BP computes the derivative/slope/gradient of an objective function by chain rule of the gradient to the weights of a multilayer stack of modules via the chain rule of derivatives. DL play a role by deeply analyzing input and capturing all related features from low to high levels. The semantic configuration and representation learning are strengthened by neural processing and vector representation, making machines capable of feeding raw data to automatically determine hidden illustrations for final prediction [121] automatically. DL techniques have some fundamental strengths for CM, including, (1) By DL techniques, CM takes advantage of non-linear processing, which creates non-linear conversion from source to target output. They have the power to learn all related features from input data by a layered structure with different parameters and hyperparameters. (2) Compared to traditional and shallow ML techniques, DL can automatically capture important features without much human effort. (3) In DL network, the optimization function plays an important role for the end-to-end paradigm to train a more complex task for CM. (4) With DL techniques, both data-driven and program-driven techniques are easily structured for CM tasks.

3.6. Deep Learning Frameworks

Currently, some well-known DL frameworks are available at hand for diverse model designing. Such frameworks are either the library or interface tools that help ML developers and research scientists to develop and design DL networks more efficiently. Table 4 represent some well-known frameworks including Torch [160], TensorFlow [161], DeepLearning4j (DL4j) [162], Caffe [163], MXNet [164], Theano [165], Microsoft Cognitive Toolkit (CNTK) [166], Neon [167], Keras [168], and Gluon [169]. They all play a very significant role in DL architectures. Due to space limitations, it is advised for readers to visit [170] for detailed information about the mentioned Frameworks.

Table 4. Summary of Deep Learning Framework.

Frameworks	References	Primary Language	Interface Provision	RNN and CNN Provision	Key Note to Know About
Torch	[160]	C and Lua	Python, C/C++, and Lua	Yes	✓ Allow standard IDE for debugging, such as, PyCharm or PDA ✓ It works with dynamically updated graph ✓ It is mostly used for DL applications, such NLP and Computer Vision.
TensorFlow (TF)	[161]	Python and C++	Python, Java, JavaScript, C/C++, Julia, C#, and Go	Yes	✓ TF is a best choice for DL networks deployments ✓ Used for Data Integration (DI), such as, sql tables, input graphs, and images ✓ Along with deploying network on influential computing clusters, TF can run networks on mobile system (Android and iOS) as well.
DL4j	[162]	Java, JVM	Python, Java, and Scala	Yes	✓ It integrates the employment of the GloVe, Deep Autoencoder, Recursive Neural Tensor Network, Word2Vec, and Doc2Vec. ✓ It uses both Hadoop and Spark, this helps to accelerate network training. ✓ It trains neural network in parallel through repeated reduce through clusters.
Caffe	[163]	C++	MATLAB and Python	Yes	✓ It is open source DL framework ✓ Works fine in computer vision ✓ It support industrial and researchers applications
MXNet	[164]	//	Python, C++, Perl, R, Go, Matlab, Scala, and Julia.	Yes	✓ It can support several GPUs with optimized calculation and fast context switching. ✓ It is scalable and lean DL framework with provision of previous networks including, CNNs, GRU, and LSTM. ✓ It supports symbolic and imperative programming.
Theano	[165]	Python	Python	Yes	✓ It lets to process mathematical operations such as, multi-dimensional arrays ✓ It is used to handle computation for large algorithms used in DL ✓ It works well with GPU as compared to CPU
CNTK	[166]	C++/C#	C++, Python, and BrainScript	Yes	✓ It is the open source app for commercial DL. ✓ It easily combines feed-forward deep neural network, CNN, RNN, and LSTM. ✓ It describes the NN as a chain of computational stages through directed graph
Neon	[167]	Python	Python	Yes	✓ It is an open source DL framework ✓ It use its own GPU and CPU backend ✓ It perform well on large batches
Keras	[168]	Python	Python	Yes	✓ User friendly, easy, and modular ✓ It offers the advantages of comprehensive adoption, provision for a wide range of incorporation with at least five back-end engines including, Theano, TensorFlow, PlaidML, CNTK, and MXNet ✓ Support several GPUs and distributed training
Gluon	[169]	Python	Python	Yes	✓ Gluon provides a friendly API, for defining easy, clear, simple, and brief code ✓ It is easier for developers to understand and learn ✓ The model's definition is dynamic, it is easier to maintain because of its flexible structure.

3.7. Review Methodology for Deep Learning Techniques

The following journal libraries have been exposed for this survey:

- IEEE Xplore Digital Library
- Google Scholar
- ACM Digital Library
- Wiley Online Library
- Springer Link

- Science Direct

We have cited over 150 popular papers from the above libraries and have shortlisted about 107 articles on CM, which focuses on DL only. The search keywords used in these libraries include Causality Mining, Causality Classification/Detection, Cause-Effect relation classification with DL, Cause-Effect Event pair detection with DL. In this section, our goals are to study DL techniques focused on CM.

3.8. Deep Learning Techniques for Causality Mining

Recently several works have been published, and most of the attention has been given to supervised systems such as shallow ML and DL approaches. The basic distinction among these systems is that advanced features engineering is essential for ML techniques, wherein DL techniques; features are learned automatically by training. However, previous approaches were largely automated, only focused on extracting explicit and simple implicit causality, and did not address complex implicit and ambiguous causalities. Furthermore, most of the early works have focused on identifying whether a relation or sentence is causal or not, and little attention is given to determine the direction of causality that which entity is the effect, and which one is the cause. The challenges mentioned above are critical for NLP researchers. Recently, DL techniques have been applied to various NLP tasks such as sentiment analysis, sentence classification, topic categorization [81], POS tagging, named entity recognition (NER), semantic role labeling (SRL), relation classification, and causality mining. In this section, we are focusing on DL models that achieved extensive success in CM. The aim of building deep models is to permit the model to learn and extract suitable features automatically.

The two most widely used classifiers among various deep neural classifiers for relation classification are CNNs and RNN. In NLP, those classifiers are based on a discrete representation of words in vector space, known as word embedding that captures syntactic and semantic information of words [171,172]. The two most widely used classifiers among various deep neural classifiers for relation classification are CNNs and RNN. To the best of our knowledge, very few DL techniques are used for CM; some are discussed in this section. Similarly, Figure 6 represent the processing levels of DL techniques, which consist of different phases of processing till to the final prediction. In this figure, the model is provided the raw input data, passed it to pre-processing steps for cleaning it for further processing. Further, the pre-processed data is passed to the input layer of the model and followed by multiple hidden layers for deep analysis of hidden features by using different hyperparameter settings. Finally, the output prediction is achieved at the output layer. If the prediction is correct, then the model is finalized. Otherwise, the model is trained repeatedly by applying loss function to reduce the error until the final prediction is based on the model's performance evaluation metrics (precision, accuracy, and recall score).

In [173], two networks are presented, a Knowledge-based features mining network and Deep CNN, to train a model for implicit and explicit causalities and their direction. They used sentence context for designing the problem into a three-class classification of entity pairs, including class-1 that specifies the annotated pair with causal direction e1 -> e2 (cause, effect), class-2 entity pairs with causal direction e2 -> e1 (effect, cause), and class-3 entity pairs are non-causal. A list of hypernyms in WordNet is prepared for each of the two annotated entities in a source sentence They used two labeled datasets including, SemEval-2007 Task-4 (http://docs.google.com/View?docID=w.df735kg3_8gt4b4c) and SemEval-2010 Task-8" dataset (http://docs.google.com/View?docid=dfvxd49s_36c28v9pmw), in which total 479 samples are used for class-1, 927 for class-2, and 982 for class-3. The SemEval-2007 dataset has seven labeled relations and the SemEval-2010 has nine relations, including cause-effect relation. They extract causality from each dataset as positive labeled data and extract a random mix of other relations as negative data.

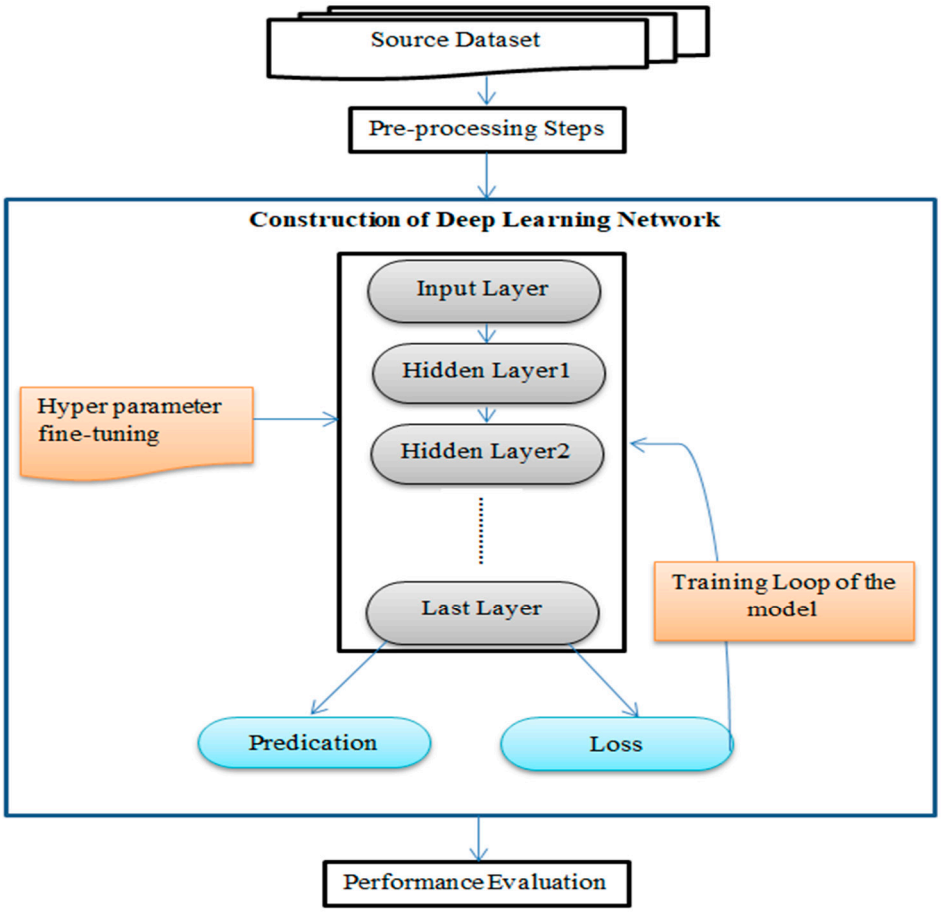

Figure 6. Processing level of DL techniques.

Ref. [174] propose a novel technique using multi-column convolutional neural networks (MCNNs) and source background knowledge (BK) for CM. It is a variant of CNN [175] with several independent columns. The inspiration for this work was [12]. They used short binary patterns to connect pairs of nouns like "A causes B" and "A prevents B" to increase the performance of event causality recognition. They focused on such event causalities, "smoke cigarettes" → "die of lung cancer" by taking an original sentence from which the candidate of causalities is extracted with the addition of related BK taken from the web texts. Three distinct methods are used to get related texts for a given causality candidate from 4 billion web pages as a source of BK, including (1) Why-question answering, (2) Using Binary Pattern (BP), and (3) Clues Terms. These techniques identify useful BK scattered in the web archives and feed into MCNNs for CM. In Figure 7, the architecture of MCNN is presented, which consists of 8 columns, where five columns are used to process event causality candidates and their nearby contexts in the original sentence. The other three columns deal with web archives. Then the output of all columns based on their layers combination is combined into the last layer for final prediction. Using all types of BK (Base + BP + WH + CL), the top achieved average precision is 55.13%, which is 7.6% higher than the best of [12] methods (47.52%). Note that by extending single CNN's to multi-column CNN's (CNN-SENT vs. Base), the proposed work obtained a 5.6% improvement, and further gave 5.8% improvement by adding with external BK.

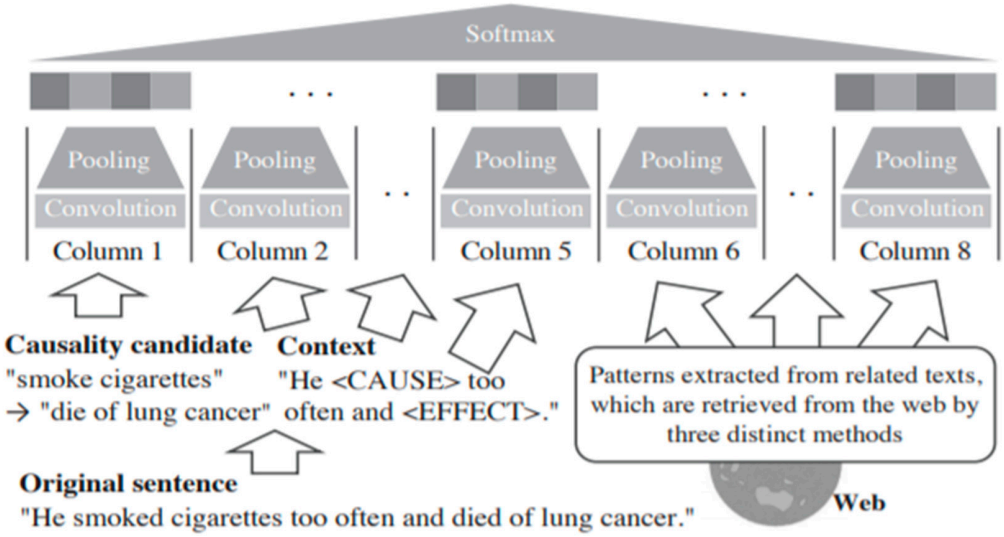

Figure 7. MCNNs Architecture for improving causality mining [174].

Ref. [176] enhanced MCNN by adding causality attention (CA), which results in the CA-MCNN model. This model is based on two notions that enhanced why-QA, which includes expressing implicitly expressed causality in one text by explicit cues from other text and describing the causes of similar events by using a set of similar words.

In [177], a novel set of event semantics and position features are used to train a Feed-Forward Network (FFN) for implicit causality. This work aims to improve ANN with features that take assistance from linguistic and associated works. It captures knowledge about the position and content of events contained in the relation. They used Penn Discourse Treebank (PDTB) and CST News (CST-NC) corpus. The whole objective function of the proposed algorithm is shown in Equation (11).

$$J = -\sum_{n=1}^{\infty} \sigma\left(\begin{array}{c} W_2||\max(\tanh(W_1.(Xi.E \oplus Xe + b_1) \\ \oplus Xp)||+b_2)\log P(y) + \ell||\theta||2 \end{array} \right) \quad (11)$$

where the set of parameters is θ = {E, W_1, b_1, W_2, b_2}, cross-entropy function is used for the loss function, which is regularized by the squared norm of parameters and scaled by hyperparameter (ℓ), positional features (Xp), input indices array (Xi), the true class label (y), and event-related features (Xe). Table 4 lists the most popular DL approaches for CM based on their targets, architecture, datasets, and references. A neural encoder-decoder approach predicts causally related events in stories through standard evaluation framework choice of plausible alternatives (COPA) [178]. This was the first approach to evaluate a neural-based model for such kinds of tasks, which learns to predict relations between adjacent sequences in stories as a means of modeling causality.

The bi-LSTM [179] is a linguistically informed architecture for automatic CM using word linguistics features and word-level embedding. It contains three modules: linguistic preprocessor and feature extractor, resource creation, and prediction background for cause/effect. A causal graph is created after grouping and proper generalization of the extracted events and their relations. They used the BBC News Article dataset, a portion of SemEval2010 task-8 related to "Cause-Effect", and adverse drug effect (ADE) dataset for training. In [180], a Temporal Causal Discovery Framework (TCDF), a DL model that learns temporal causal graph design by mining causality in continuous observational time series data. It applied multiple attention-based CNN along with a causal support step. It can also mine time interruption among cause and the existence of its effect. They used two benchmarks with multiple datasets including, simulated financial market and simulated functional magnetic resonance imaging (FMRI) data. Both contain a ground

truth comprising the underlying causal graph. The experimental analysis shows that this mechanism is precise in mining time-series data.

Ref. [181] propose a novel deep CNN using grammar tags for cause-effect pair identification from nominal words in natural language corpus knowledge reasoning. Though, the prior works mainly were based on predefined syntactic and linguistic rules. The modern approaches use shallow ML primarily Deep NN on top of linguistic and semantic knowledge to classify nominal word relations in a corpus. They used the SemEval-2010 Task 8 corpus for enhancing the performance of CM. In [182], a novel idea of Knowledge-Oriented CNN (K-CNN) for causality identification is presented. This model combined two channels: Data-Oriented Channel (DOC), which acquires important features of causality from the target data, and Knowledge-Oriented Channel (KOC), which integrates former human knowledge to capture the linguistic clues of causality. In KOC, the convolutional filters are automatically created from available knowledge bases (FrameNet and WordNet) without training the classifier by a huge amount of data. Such filters are the embedding of causation words. Additionally, it uses clustering, filters selection, and additional semantic features to increase the performance of K-CNN. They used three datasets including Causal-Time Bank4 (CTB), SemEval-2010 task-86, and Event StoryLine datasets7. More specifically, the KOC is used to integrate existing linguistic information from knowledge bases. Where DOC is used to learn important features from data by using a pre-defined convolutional filter. These two channels complement each other and extract valuable features of CM.

In the same year, a novel feed-forward neural network (FFNN) was used with a context word extension mechanism for CM in tweets [183]. For event context word extension, they used BK, extracted from news articles in the form of a causal network to identify event causality. They have used 2018 commonwealth game-related tweets held in Australia. This was a challenging job because tweets are mostly composed of unstructured nature, highly informal, and lacking contextual information. This approach is closely related to [177] for detecting causality between events using FFNN by enhancing the feature set by computing distances among events trigger word and related words in the phrase. Though, such positional knowledge for tweets might not show the causal direction more easily because tweets are mostly composed of noisy words and characters e.g., # (hashtags), @ sign, question marks (?), URLs, and emojis. Hence, such data is not appropriate for the detection of causality in tweets. Inspired by [183], the automatic mining of causality in a short corpus is a useful and challenging task [184], because it contains many informal characters, emojis, and questions marks. This technique was applied a deep causal event detection and context word extension approach for CM in tweets. They used more than 207k tweets using Twitter API (https://developer.twitter.com/en/docs/tweets/search/overview). They prepare to collect those tweets that were associated with the "Commonwealth Games-2018 held in Australia". This study [185] presents a BERT-based approach using multiple classifiers for CM inside a web corpus, which used independent labels given by multiple annotators in the corpus. By training multiple classifiers, hold all annotators procedure, where every classifier predicts the labels provided by a particular annotator, and integrate the result of all classifiers to predict the final labels found by the majority vote. BERT is a pre-trained network with a huge amount of corpus that learned some sort of BK for event-causal relations during pre-training. They used (Hashimoto et al., 2014) in the construction of source datasets. The experimentations prove that the performance is improved when BERT is pre-trained with a web corpus that covering a huge amount of event causalities instead of using Wikipedia texts. Though this effect was inadequate, hence, they further enhanced the performance by simply adding corpus associated with an input causality candidate as a BK to the input of the BERTs, which significantly beaten the state-of-the-art approach [174] by around 0.5 in average precision.

Ref. [186] explored the causality effect of search queries associated with bars and restaurants on every day new cases in the United State (US) areas with low and high everyday cases. GT searches for bars and restaurants presented a major effect on everyday new cases for areas with higher numbers of every day new cases in the US. They used

deep LSTM model for training, which is a typical problem in ML tasks. In [187], the Event Causality identification (ECI) model are proposed by targeting the limitations of past approaches by leveraging outside knowledge for reasoning, which can significantly improve the illustration of events and also mine event-agnostic, context-specific patterns, by a mechanism named "event mention masking generalization", which can significantly improve the capability of the model to handle new and previous unnoticed cases. Significantly, the important element of this model is "Knowledge-aware causal reasoned", which can exploit BK in external CONCEPTNET knowledge bases [188] to improve the cognitive process. They used 3 benchmark datasets including, Causal-TimeBank, Event Story Line, and Event Causality for experimentations, which show the model achieves state-of-the-art performance. In [189], the problem of causal impact is considered for numerous 'COVID-19' associated policies on the outbreak dynamics in diverse US states at different time intervals in 2020. The core issue in this work is the presence of time-varying and overlooked confounders. To address this issue, they integrated data from several COVID-19 related databases comprising diverse types of information, which help as substitutions for confounders. They used a neural network-based approach, which learns the illustrations of the confounders using time-varying observational and relational data and then guesses the causal effect of such policies on the outbreak dynamics with the learned confounder representations. The outcomes of this study confirming the proficiency of the model in controlling confounders for causal valuation of COVID-19 associated policies.

In [190], a self-attentive Bi-LSTM-CRF based approach is presented, named Self-attentive BiLSTM-CRF wIth Transferred Embedding (SCITE). This technique formulates CM as a sequence tagging problem. This is useful for directly mining cause and effect events without considering cause-effect pairs and their relationship separately. Moreover, to progress the performance of CM, a multi-head self-attention procedure is presented into the model to acquire the dependencies among causal words. To solve two issues, first, they included Flair embedding due to prior information deficiency in the [191]. Second, in terms of positions in the text, cause and effect are rarely far from each other's. For this, a multi-head self-attention [134] is applied. The SemEval 2010 task 8 is used with extended annotation, in which Flair-BiLSTM-CRF achieved progress of about 6.32% over the Bi-LSTM-CRF compared with BERT and ELMo (rises of 4.55% and 6.28%). Moreover, the causality tagging approach produced enhanced results compared to the general tagging approach under the SCITE model. This study [192] developed three network-architectures (Masked Event C-BERT, Event aware C-BERT, C-BERT) on the top of language models (pre-trained BERT) that influence the complete sentence context, events context, and events masked context for CM among expressed events in natural language text (NLT). They simply focus to recognize possible causality among marked events in a given sequence of text, but it doesn't find the validity of such relations.

This approach achieved state-ofthe-art performance in the proposed data distributions and can be used for mining causal diagrams and/or constructing a chain of events from an unstructured corpus. For experimentation, they generated their dataset from three benchmarks including, Semeval 2010 task 8 [30], Semeval 2007 task 4 [57], and ADE [193] corpus. This approach achieved state-ofthe-art performance in the proposed data distributions and can be used for mining causal diagrams and/or constructing a chain of events from an unstructured corpus. Table 5, represents the most common and well-used DL models by their Targets, Architecture, Datasets, References, and Drawbacks.

Table 5. Summary of deep learning networks for CM.

SNo	Architecture	References	Targets	Datasets	Language	Drawbacks
1.	Deep CNN with Knowledge-based features	[173]	This model mine both implicit and explicit causality, and direction of causality.	SemEval-2007 Task-4 and SemEval-2010 Task 8 datasets in English language.	English	Work on simple knowledge-based features
2.	MCNNs + BK	[174]	This work targeted implicit and ambiguous causality.	Four billion web pages in Japanese corpus.	Japanese	Only concentrated on Japanese corpus
3.	CA-MCNN	[176]	Target implicitly expressed cause-effect relations.	600 million Japanese web pages.	✓	✓
4.	FFNN	[177]	This architecture targeted implicit and ambiguous causalities.	The Penn Discourse Treebank and CST News Corpus in English language.	English	Over-fitting problem
5.	COPA Encoder-decoder models	[178]	They targeted causally related entities.	The Visual Storytelling (VIST), CNN/Daily Mail corpus, and CMU Book/Movie Plot Summaries in English language.	✓	Complex network design
6.	bi-LSTM	[179]	They focused causal events and their effects inside a sentence.	The BBC News Article, SemEval2010 task-8, and ADE (Adverse drug effect) datasets in English language.	✓	Time complexity
7.	Temporal Causal Discovery Framework (TCDF)	[180]	They learned temporal causal graph design by mining causality in a continuous observational time series data.	The simulated financial market (SFM) and simulated functional magnetic resonance imaging (SFMRI) dataset in English language.	✓	They executes rather worse on short time series.
8.	Deep CNN with grammar tags	[181]	Identifying cause-effect pair from nominal words.	SemEval-2010 Task 8 corpus.	✓	Over-fitting problem.
9.	Knowledge-Oriented CNN (K-CNN)	[182]	They targeted implicit causalities.	The Causal-Time Bank (CTB), SemEval-2010 task-8, and Event Story Line datasets in English language.	✓	Model over-fitting issue
10.	FFNN + BK	[183]	They targeted implicit causalities social media tweets.	Tweets associated to commonwealth Games, held in 2018 in Australia, in English language.	✓	This results in info loss. Due to opinionated posts.
11.	This technique applying a deep causal event detection and context word extension approach	[184]	They targeted implicit causalities in tweets.	More than 207k tweets related to Commonwealth Games-2018 held in Australia, in English language.	✓	Have knowledge or Information loss
12.	BERT-based approach using multiple classifiers	[185]	Mining Implicit Causality inside web corpus.	180 million news article snippets and titles corpus.	Japanese	Awareness and risk Management.
13.	BiLSTM-CRF-based model	[190]	They focused on implicit CM.	SemEval 2010 task 8 dataset with extended annotation in English language.	✓	Over-fitting issues
14.	Masked Event C-BERT, Event aware C-BERT, and C-BERT.	[192]	Influence the complete sentence context, events context, and events masked context for CM.	Semeval 2010 task 8 [30], Semeval 2007 task 4 [57], and ADE corpus.	✓	They simply focus to recognize possible causality among marked events in a given sequence of text, but it doesn't find the validity of such relations.

4. Comparing the Two Paradigms

Table 6 lists a comparison among two paradigms including Statistical/ML and DL approaches. Such comparison is made based on datasets preparation, domain types, applications, processing time, and their limitations.

Table 6. Comparison among mentioned techniques.

SNo	Statistical /ML Techniques	Deep Learning Techniques
1.	ML approaches used automatic tools for annotations, coding, and labeling e.g., crowdsourcing platforms like Amazon mechanical trunk (AMT).	DL approaches utilize deep neural architecture for analyzing data more deeply for automatic feature engineering.
2.	ML techniques focus on finding patterns automatically through small seed patterns.	They focus on finding patterns automatically by deep analysis without using seed patterns.
3.	They are trained and tested on huge textual corpora as compared to manual approaches.	They are trained and tested on unlimited text corpora.
4.	They work well using domain-independent corpus.	They combine both domain-dependency and independency into one framework.
5.	Such approaches are capable of catching those generalizations by appropriate feature sets.	Such approaches work well for both specific and other generalizes corpora.
6.	By using class-specific probabilities, the ambiguities can be captured automatically with ML algorithms.	Those approaches use their deep architecture by targeting implicit and ambiguous relations more efficiently.
7.	Such approaches focusing on both explicit and simple implicit causality.	They combine both implicit and explicit causalities into one model.
8.	Those approaches use lexical Knowledge bases and some other broad-based corpora like Wikipedia and DBpedia by creating knowledge bases and ontologies for training.	Those approaches combine all semantic lexicons and use web archives as a source of world knowledge.
9.	They are not working well for highly specialized domains. Besides, such annotated data may not be available in plenty, which results in good training and generalization.	Such approaches do not work well for highly specialized domains.
10.	Such approaches lacking standardized corpora, yet, no work provided empirical comparisons with existing approaches. This makes it a surprising and relatively fruitless exercise to compare the recall, precision, and accuracy of one approach with others.	Similarly, those approaches lack of standardized corpora, and yet, no work provided empirical comparisons with existing models. This makes it a surprising and comparatively fruitless exercise to compare the precision, recall, and accuracy of different approaches with each other.

5. Challenges and Future Guidelines

This section addresses significant research challenges based on distinct and deep literature on shallow ML and DL approaches. Based on several state-of-the-art works, we recognized some key research challenges faced during CM along with their future strategies and directions.

5.1. Ambiguous/Implicit Data

With the development of any tool, the fundamental and first step is how to arrange the data. Generally, the source datasets consist of images, graphics, sound, text, videos, and multimedia data. Moreover, each data instance contains diverse features, domain types, dimensions, data sizes, and characteristics. This varied nature of data keeps causality a challenging problem. Such problems can be handle through DL methods, which work based on their deep analysis and structure to deal with all critical features of the data.

5.2. Features Engineering

Features engineering is the second most fundamental step for any approach after data preparation. Most of the early works focused on hand-crafted approaches for features engineering, which was ineffective incapturing all necessary features in the source data. Hence, these issues are handled and explored through DL techniques because DL approaches work on automatic feature engineering, requiring little human effort and attention.

5.3. Model Selection

Model selection is the third basic step for any task modeling. Most of the non-statistical and ML approaches have been focused on supervised and unsupervised algorithms that produced unsatisfactory results because of the poor design of the model. These models need much attention from a human operator because of their diverse nature of parameters and design. Contrary, the DL models are more effective and work well based on their deep architecture and automatic feature engineering.

5.4. Nature of Causality

Causalities are usually present in explicit, ambiguous, and implicit nature where explicit causalities are the most occurring type of causalities in source datasets, which is simple to handle by traditional and ML techniques. While implicit and ambiguous causalities are very hard tasks to handle by such techniques. Hence, DL techniques are the best choice by their strong inference ability to deal with implicit and ambiguous causalities.

5.5. Data Standardization

For any model or algorithm, data standardization is the key source of accurate implementation. Due to the general lack of standardized datasets, no work delivers an observed comparison with existing approaches, which makes it a surprising and comparatively fruitless exercise to compare the precision, recall, and accuracy of one algorithm with the others. Hence, it is a challenging task in the field, which needs more attention from researchers to develop standardized datasets and data-driven models in the field.

5.6. Computational Cost

Most traditional practices incorporating diverse techniques into a single tool for performance enhancement, which leads to increase computation costs. Whereas, using DL techniques with shallow ML algorithms will mark the computational cost by combing parallel and distributed processing to make a matrix of multiple vectors, which minimizes the computational cost.

5.7. Accuracy

Prior algorithms are usually mine causalities more efficiently, though the reliance on the outcome was unsatisfactory because of their low accuracies. Although, in some situations, the accuracy is satisfactory by using explicit and domain-specific corpora, while insufficient for implicit, ambiguous, and domain-independent corpora. Merging shallow ML with DL approaches will tickle such issues, which emerge as an alternative tool at a certain level of accuracy. Besides, Table 7 lists some imperative research challenges with future guidelines that recognize the upcoming research directions to develop open and adaptable tools for CM, which helps to accumulate ambiguous, implicit, and domain-independent corpora. Similarly, future tools should cover hybrid, fast, and incremental learning algorithms for innovative challenges.

Table 7. Summary of diverse challenges and their future direction.

S_No	Challenges	Future Research Guidelines
1.	Ambiguous Data	The deep model can memorize a huge amount of information and data, but due to the heterogeneous nature of data makes it a black-box solution for many applications. The existence of such datasets is a key challenge, which needs the interpretability of data- driven DL techniques that produce more satisfactory results.
2.	Features Engineering	Using Deep CNN, RNN, GRU, LSTM, bi-LSTM, DCNN, BERT, and MCNN with their powerful feature abstraction capabilities to capture implicit and ambiguous features contribute most of the errors in the existing systems. Hence, new paradigms are required that can boost the learning ability of DL by integrating informative features-maps learned by supporting learners at the intermediate phases of DL models [70].

Table 7. *Cont.*

S_No	Challenges	Future Research Guidelines
3.	Model Sel	DL approaches still facing trouble by modeling complex data modalities. To achieve the best performance at various datasets, the combination of diverse and multiple DL architectures (DeepCNN, DeepRNN, Transformer, BERT, TinyBERT, ELECTRA, and attention-based bi-LSTM) can benefit the model robustness and generalization on various relations by mining diverse levels of semantic representations. Ideas of dropout, batch normalization, and novel activation functions are also important.
4.	Nature of Causality	For mining techniques implicit and ambiguous causality across the sentences is still a big challenge, which needs the ideas of single sentence rule and procedures that help us to develop a model for cross sentence CM.
5.	Data Standardization	By the general lack of standardized datasets, this is a surprising and relatively fruitless exercise to compare the precision, recall, and accuracy of different techniques. This needs attention in the preparation of a standardized dataset. And an experimental comparison of the existing systems is required on standardized data sets, and for now, CM is still full of challenges, included counterfactual causality and credibility of causality in text.
6.	Computational Cost	Review the applications of deep CNN on other associated tasks such as computer vision and NLP tasks will lead us to observe those models for CM.
7.	Accuracy	Combining a general semantic relations classifier e.g., SemEval-Tasks with any existing causality extraction system would be a valuable attempt toward accuracy improvement.
8.	Hypothesis generation	There is a need to use some techniques for event causality hypothesis generation and Scenario generation.
9.	Area of Interest	There should need to use some techniques for event causality hypothesis and Scenario generation.
10.	Attention	Attention is a fundamental visual organism in the human body, which automatically catches information from text and images in the surrounding. The attention system not simply mines the essential information from text and image but also stores its contextual relation with additional elements. In the future, research may be conceded in the track that reserves the whole semantics, syntactic features along with their discriminating features at the learning stages.

6. Conclusions

To the best of our knowledge, this is the first survey paper, which focuses on widespread state-of-the-art ML and DL research techniques, algorithms, and frameworks spanning a few decades for CM. Compared with other reviews, our paper is devoted to considering both ML and DL techniques, covering the most updated highly cited papers. We explored the causality problem and provided the researcher with the essential background knowledge of shallow ML and DL for the causality mining task. It begins with the history of shallow ML and DL algorithms, highly cited paper, related challenges, limitations, and developments in diverse applications. We notice that causality mining is a challenging NLP task mainly due to implicit, heterogeneous, and ambiguous linguistic concepts, which could or could not be causal. Data is another challenge for focused domains where much human expert annotation is needed, making it inflexible to use minimally supervised methods. Furthermore, model selection and data standardization is also the key challenges. At the present time, shallow ML and DL techniques with their automatic feature engineering approach have succeeded in reasonable results. Analysis of these use cases helped us to identify the upcoming challenges and suggest many existing solutions. Additionally, we discuss a flow of figurative representation of all those approaches that help to understand the CM process efficiently, which will lead to the suggestion of novel tools for implicit and ambiguous causalities in the future. Furthermore, there is a long way to go and get the required goals and objectives. We listed many findings and some possible future guidelines in Table 7.

Author Contributions: Conceptualization, W.A. and W.Z.; methodology, W.A.; software, W.A.; validation, W.A., R.A. and X.Z.; formal analysis, W.A.; investigation, W.Z.; resources, W.A. and G.R.; data curation, W.A.; writing—original draft preparation, W.A.; writing—review and editing, W.A., G.R. and R.A.; visualization, W.A. and G.R.; supervision, W.Z.; project administration, W.Z.; funding acquisition, No. All authors have read and agreed to the published version of the manuscript.

Funding: This work is sponsored by the National Natural Science Foundation of China (61976103, 61872161), the Scientific and Technological Development Program of Jilin Province (20190302029GX, 20180101330JC, 20180101328JC), Tianjin Synthetic Biotechnology Innovation Capability Improvement Program (no. TSBICIP-CXRC-018), and the Development and Reform Commission Program of Jilin Province (2019C053-8).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study did not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chan, K.; Lam, W. Extracting causation knowledge from natural language texts. *Int. J. Intell. Syst.* **2005**, *20*, 327–358. [CrossRef]
2. Luo, Z.; Sha, Y.; Zhu, K.Q.; Wang, Z. Commonsense Causal Reasoning between Short Texts. In Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'16, Cape Town, South Africa, 25–29 April 2016; pp. 421–430.
3. Khoo, C.; Chan, S.; Niu, Y. The Many Facets of the Cause-Effect Relation. In *The Semantics of Relationships*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 51–70.
4. Theodorson, G.; Theodorson, A. *A Modern Dictionary of Sociology*; Crowell: New York, NY, USA, 1969; p. 469.
5. Hassanzadeh, O.; Bhattacharjya, D.; Feblowitz, M. Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 5003–5009.
6. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **2009**, *3*, 96–146. [CrossRef]
7. Girju, R. Automatic detection of causal relations for question answering. In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, Sapporo, Japan, July 2003; Volume 12, pp. 76–83.
8. Khoo, C.; Kornfilt, J. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Lit. Linguist. Comput.* **1998**, *13*, 177–186. [CrossRef]
9. Radinsky, K.; Davidovich, S.; Markovitch, S. Learning causality for news events prediction. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 909–918.
10. Silverstein, C.; Brin, S.; Motwani, R.; Ullman, J. Scalable techniques for mining causal structures. *Data Min. Knowl. Discov.* **2000**, *4*, 163–192. [CrossRef]
11. Riaz, M.; Girju, R. Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision. In Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing, Pittsburgh, PA, USA, 20–22 September 2010; pp. 361–368.
12. Hashimoto, C.; Torisawa, K.; Kloetzer, J.; Sano, M. Toward future scenario generation: Extracting event causality exploiting semantic relation, coantext, and association features. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MA, USA, 22–27 June 2014; pp. 987–997.
13. Ackerman, E. Extracting a causal network of news topics. *Move Mean. Internet Syst.* **2012**, *7567*, 33–42.
14. Bollegala, D.; Maskell, S. Causality patterns for detecting adverse drug reactions from social media: Text mining approach. *JMIR Public Health Surveill.* **2018**, *4*, e8214. [CrossRef] [PubMed]
15. Richardson, M.; Burges, C. Mctest: A challenge dataset for the open-domain machine comprehension of text. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 193–203.
16. Berant, J.; Srikumar, V. Modeling biological processes for reading comprehension. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1499–1510.
17. Hassanzadeh, O.; Bhattacharjya, D.; Feblowitz, M.; Srinivas, K.; Perrone, M.; Sohrabi, S.; Katz, M. Causal Knowledge Extraction through Large-Scale Text Mining. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13610–13611.
18. Khoo, C.S.; Myaeng, S.H.; Oddy, R.N. Using cause-effect relations in text to improve information retrieval precision. *Inf. Process. Manag.* **2001**, *37*, 119–145. [CrossRef]
19. Khoo, C.; Chan, S. Extracting causal knowledge from a medical database using graphical patterns. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, 3–6 October 2000; pp. 336–343.

20. Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D.A.; Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **2005**, *308*, 523–529. [CrossRef]
21. Araúz, P.L.; Faber, P. Causality in the Specialized Domain of the Environment. In Proceedings of the Semantic Relations-II, Enhancing Resources and Applications Workshop Programme Lütfi Kirdar, Istanbul Exhibition and Congress Centre, Istanbul, Turkey, 22 May 2012; p. 10.
22. General, P.W. Representing causation. *J. Exp. Psychol.* **2007**, *136*, 1–82.
23. Talmy, L. *Toward a Cognitive Semantics*; Volume I: Concept Structuring Systems; MIT Press: Cambridge, MA, USA, 2000; pp. 1–565.
24. Semantics, J.H. Toward a useful concept of causality for lexical semantics. *J. Semant.* **2005**, *22*, 181–209.
25. White Peter, A. Ideas about causation in philosophy and psychology. *Psychol. Bull.* **1990**, *108*, 1–3.
26. Scaria, A.; Berant, J.; Wang, M.; Clark, P.; Lewis, J. Learning biological processes with global constraints. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1710–1720.
27. Ayyoubzadeh, S.; Ayyoubzadeh, S. Predicting COVID-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study. *MIR Public Health Surveill.* **2020**, *6*, e18828. [CrossRef] [PubMed]
28. FAQ about Google Trends Data—Trends Help. Available online: https://support.google.com/trends/answer/4365533?hl=en# (accessed on 3 October 2021).
29. Blanco, E.; Castell, N.; Moldovan, D. Causal Relation Extraction. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May 2008; pp. 310–313.
30. Hendrickx, I.; Kim, S.N.; Kozareva, Z.; Nakov, P.; Diarmuidó, D.; Diarmuidó, S.; Padó, S.; Pennacchiotti, M.; Romano, L.; Szpakowicz, S. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. *arXiv* **2019**, arXiv:1911.10422.
31. Sorgente, A.; Vettigli, G.; Mele, F. Automatic Extraction of Cause-Effect Relations in Natural Language Text. In Proceedings of the 7th International Workshop on Information Filtering and Retrieval Co-Located with the 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, 6 December 2013; pp. 37–48.
32. Cresswell, M. Adverbs of causation. In *Words, Worlds, and Contexts: New Approaches in Word Semantics*; De Gruyter: Berlin, Germany, 1981; pp. 21–37.
33. Simpson, J. Resultatives. In *Papers in Lexical-Functional Grammar*; Indiana University Linguistics Club: Bloomington, IN, USA, 1983; pp. 1–17.
34. Altenberg, B. Causal linking in spoken and written English. *Stud. Linguist.* **1984**, *38*, 20–69. [CrossRef]
35. Nastase, V. *Semantic Relations across Syntactic Levels*; University of Ottawa: Ottawa, ON, USA, 2004; pp. 1910–2010.
36. Sadek, J. Automatic detection of arabic causal relations. In *International Conference on Application of Natural Language to Information Systems (NLDB'13)*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 400–403.
37. Garcia, D. COATIS, an NLP system to locate expressions of actions connected by causality links. In Proceedings of the Knowledge Acquisition, Modeling and Management: 10th European Workshop, EKAW'97, Sant Feliu de Guixols, Catalonia, Spain, 15–18 October 1997; pp. 347–352.
38. Asghar, N. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *arXiv* **2016**, arXiv:1605.07895.
39. Gelman, A. Causality and statistical learning. *Am. J. Sociol.* **2011**, *117*, 955–966. [CrossRef]
40. Athey, S.; Stat, G.I. Machine learning methods for estimating heterogeneous causal effects. *Stat* **2015**, *1050*, 1–26.
41. Mooij, J.; Peters, J.; Janzing, D.; Zscheischler, J. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.* **2016**, *17*, 1103–1204.
42. Spirtes, P.; Zhang, K. Causal discovery and inference: Concepts and recent methodological advances. *Appl. Inform.* **2016**, *3*, 1–28. [CrossRef] [PubMed]
43. Guo, R.; Cheng, L.; Li, J.; Hahn, P.R.; Liu, H. A Survey of Learning Causality with Data: Problems and Methods. *ACM Comput. Surv.* **2020**, *53*, 1–37.
44. Quinlan, J. *C4.5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
45. Charniak, E. A maximum-entropy-inspired parser. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, 29 April–4 May 2000; pp. 132–139.
46. Rosario, B.; On, M.H. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01), Pittsburgh, PA, USA, 3–4 June 2001; pp. 82–90.
47. Chang, D.; KS, C. Causal relation extraction using cue phrase and lexical pair probabilities. In Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP'04), Hainan, China, 22–24 March 2004; pp. 61–70.
48. Marcu, D.; Echihabi, A. An unsupervised approach to recognizing discourse relations. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 368–375.
49. Rink, B.; On, S.H. Utd: Classifying semantic relations by combining lexical and semantic resources. In Proceedings of the 5th International Workshop on Semantic Evaluation; Association for Computational Linguistics, Uppsala, Sweden, 15–16 July 2010; pp. 256–259.
50. Sil, A.; Huang, F.; Series, A.Y. Extracting action and event semantics from web text. In Proceedings of the 2010 AAAI Fall Symposium Series, Westin Arlington Gateway, Arlington, Virginia, 11–13 November 2010; pp. 108–113.

51. Pal, S.; Pakray, P.; Das, D. JU: A supervised approach to identify semantic relations from paired nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, 15–16 July 2010; pp. 206–209.
52. Li, Z.; Ding, X.; Liu, T.; Hu, J.E.; Durme, B. Van Guided Generation of Cause and Effect. *arXiv* **2020**, arXiv:2107.09846, 1–8.
53. Schank, R.C. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*; Cambridge University Press: Cambridge, UK, 1983; p. 234.
54. Szpakowicz, S.; Nastase, V. Exploring noun-modifier semantic relations. In Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5), Tilburg University, Tilburg, The Netherlands, 15–17 January 2003; pp. 285–301.
55. Tapanainen, P.; Natural, T.J. A non-projective dependency parser. In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, USA, 31 March–4 April 1997; pp. 64–71.
56. Girju, R.; Beamer, B.; Rozovskaya, A. A knowledge-rich approach to identifying semantic relations between nominals. *Inf. Process. Manag.* **2010**, *46*, 589–610. [CrossRef]
57. Girju, R.; Nakov, P.; Nastase, V.; Szpakowicz, S.; Turney, P.; Yuret, D. Semeval-2007 task 04: Classification of semantic relations between nominals. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23 June 2007; pp. 13–18.
58. Pakray, P.; Gelbukh, A. An open-domain cause-effect relation detection from paired nominals. In *Mexican International Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2014; pp. 263–271.
59. Bethard, S.; HLT, J.M. Learning semantic links from a corpus of parallel temporal and causal relations. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, OH, USA, 16–17 June 2008; pp. 177–180.
60. Bethard, S.; Corvey, W.; Klingenstein, S.; Martin, J.H. Building a Corpus of Temporal-Causal Structure. In Proceedings of the European Language Resources Association (ELRA), Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May 2008; pp. 1–8.
61. Rink, B.; Bejan, C. Learning textual graph patterns to detect causal event relations. In Proceedings of the Twenty-Third International FLAIRS Conference, Datona Beach, FL, USA, 19–21 May 2010; pp. 265–270.
62. Do, Q.; Chan, Y.S.; Roth, D. Minimally Supervised Event Causality Identification. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing(EMNLP 2011), Edinburgh, Scotland, UK, 27–31 July 2011; pp. 294–303.
63. Lin, Z.; Ng, H.T.; Kan, M.Y. A pdtb-styled end-to-end discourse parser. *Nat. Lang. Eng.* **2014**, *20*, 151–184. [CrossRef]
64. Riaz, M.; Girju, R. Recognizing Causality in Verb-Noun Pairs via Noun and Verb Semantics. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), Gothenburg, Sweden, 26 April 2014; pp. 48–57.
65. Yang, X.; Mao, K. Multi level causal relation identification using extended features. *Expert Syst. Appl.* **2014**, *41*, 7171–7181. [CrossRef]
66. Kingsbury, P.; Palmer, M. From TreeBank to PropBank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands, Spain, 29–31 May 2002; pp. 1989–1993.
67. Mirza, P.; Kessler, F.B. Extracting Temporal and Causal Relations between Events. In Proceedings of the ACL 2014 Student Research Workshop, Baltimore, MA, USA, 22–27 June 2014; pp. 10–17.
68. Mirza, P.; Sprugnoli, R.; Tonelli, S.; Speranza, M. *Annotating Causality in the TempEval-3 Corpus*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2015; pp. 10–19.
69. Zhao, S.; Liu, T.; Zhao, S.; Chen, Y.; Nie, J.-Y. Event causality extraction based on connectives analysis. *Neurocomputing* **2016**, *173*, 1943–1950. [CrossRef]
70. Hidey, C.; Mckeown, K. Identifying Causal Relations Using Parallel Wikipedia Articles. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1424–1433.
71. Qiu, J.; Xu, L.; Zhai, J.; Luo, L. Extracting Causal Relations from Emergency Cases Based on Conditional Random Fields. *Procedia Comput. Sci.* **2017**, *112*, 1623–1632. [CrossRef]
72. Rehbein, I.; Ruppenhofer, J. Catching the Common Cause: Extraction and Annotation of Causal Relations and their Participants. In Proceedings of the 11th Linguistic Annotation Workshop, Valencia, Spain, 3 April 2017; pp. 105–114.
73. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the MT Summit, Phuket, Thailand, 12–16 September 2005; Volume 5, pp. 79–86.
74. Dunietz, J.; Levin, L.; Carbonell, J. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In Proceedings of the 11th Linguistic Annotation Workshop, Valencia, Spain, 3 April 2017; pp. 95–104.
75. Zhao, S.; Jiang, M.; Liu, M.; Qin, B.; Liu, T. CausalTriad: Toward Pseudo Causal Relation Discovery and Hypotheses Generation from Medical Text Data. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health, Washington, DC, USA, 29 August–1 September 2018; pp. 184–193.
76. Ning, Q.; Feng, Z.; Wu, H.; Roth, D. Joint reasoning for temporal and causal relations. In Proceedings of the ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2278–2288.
77. Craciunescu, T.; Murari, A.; Gelfusa, M. Causality detection methods applied to the investigation of malaria epidemics. *Entropy* **2019**, *21*, 784. [CrossRef] [PubMed]

78. Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.K.; Webber, B.L. The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May 2008; pp. 1–8.
79. Pustejovsky, J.; Hanks, P.; Saurí, R.; See, A.; Gaizauskas, R.; Setzer, A.; Radev, D.; Sundheim, B.; Day, D.; Ferro, L.; et al. The TIMEBANK Corpus. *Corpus Linguist.* **2003**, *2003*, 40.
80. Radinsky, K.; Davidovich, S.; Markovitch, S. Learning to Predict from Textual Data. *J. Artif. Intell. Res.* **2012**, *45*, 641–684. [CrossRef]
81. Riaz, M.; Girju, R. Toward a Better Understanding of Causality between Verbal Events: Extraction and Analysis of the Causal Power of Verb-Verb Associations. In Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial), Metz, France, 22–24 August 2013; pp. 21–30.
82. Ishii, H.; Ma, Q.; Yoshikawa, M. Incremental Construction of Causal Network from News Articles. *J. Inf. Process.* **2012**, *20*, 207–215. [CrossRef]
83. Peng, N.; Poon, H.; Quirk, C.; Toutanova, K.; Yih, W. Cross-Sentence N -ary Relation Extraction with Graph LSTMs. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 101–115. [CrossRef]
84. Marcus, M.; Kim, G.; Marcinkiewicz, M.A.; Macintyre, R.; Bies, A.; Ferguson, M.; Katz, K.; Schasberger, B. The Penn TreeBank: Annotating Predicate Argument Structure. In Proceedings of the Human Language Technology: Proceedings of a Workshop, Plainsboro, NJ, USA, 8–11 March 1994; pp. 110–115.
85. Sandhaus, E. The new york times annotated corpus. In Proceedings of the Linguistic Data Consortium, University of Philadelphia, Philadelphia, PA, USA, 17 October 2008; p. e26752.
86. Smith, N.A.; Cardie, C.; Washington, A.L.; Wilkerson, J.D. Overview of the 2014 NLP Unshared Task in PoliInformatics. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 26 June 2014; pp. 5–7.
87. Ide, N.; Baker, C.; Fellbaum, C.; Passonneau, R. The Manually Annotated Sub-Corpus: A Community Resource for and By the People. In Proceedings of the ACL 2010 Conference Short Papers, Stroudsburg, PA, USA, 11–16 July 2010; pp. 68–73.
88. UZMAY, G.; Gokce, K. The Causality Effect of Interest in the Financial Crisis and Oil Market on Food Prices: A Case Study of Internet Search Engine Behavior. In Proceedings of the IX. IBANESS Congress Series, Edirne, Turkey, 29–30 September 2018; pp. 1–10.
89. Faes, L.; Nollo, G.; Porta, A. Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Phys. Rev.* **2011**, *83*, 51112. [CrossRef] [PubMed]
90. Eckmann, J.; Kamphorst, S. Recurrence plots of dynamical systems. *World Sci. Ser. Nonlinear Sci. Ser. A* **1995**, *16*, 441–446.
91. Society, C.G. Investigating causal relations by econometric models and cross-spectral methods. *Econom. J. Econom. Soc.* **1969**, *37*, 424–438.
92. Marinazzo, D.; Pellicoro, M. Kernel method for nonlinear Granger causality. *Phys. Rev. Lett.* **2008**, *100*, 144103. [CrossRef] [PubMed]
93. Yang, A.; Peng, C. Causal decomposition in the mutual causation system. *Nat. Commun.* **2018**, *9*, 1–10. [CrossRef] [PubMed]
94. Craciunescu, T.; Murari, A.; Gelfusa, M. Improving entropy estimates of complex network topology for the characterization of coupling in dynamical systems. *Entropy* **2018**, *20*, 891. [CrossRef] [PubMed]
95. Haque, U.; Hashizume, M.; Glass, G.E.; Dewan, A.M.; Overgaard, H.J.; Yamamoto, T. The role of climate variability in the spread of malaria in bangladeshi highlands. *PLoS ONE* **2010**, *5*, e14341. [CrossRef] [PubMed]
96. Hanf, M.; Adenis, A.; Nacher, M.; Journal, B.C. The role of El Niño southern oscillation (ENSO) on variations of monthly Plasmodium falciparum malaria cases at the cayenne general hospital, 1996–2009. *Malar. J.* **2011**, *10*, 1–4. [CrossRef] [PubMed]
97. Syamsuddin, M.; Fakhruddin, M. Causality analysis of Google Trends and dengue incidence in Bandung, Indonesia with linkage of digital data modeling: Longitudinal observational study. *J. Med. Internet Res.* **2020**, *22*, e17633. [CrossRef]
98. Deng, L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, 1–29. [CrossRef]
99. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Foundations*; MIT Press: Cambridge, CA, USA, 1986; pp. 399–421.
100. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
101. LeCun, Y.; Neural, Y.B. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
102. Burney, A.; Syed, T.Q. Crowd Video Classification Using Convolutional Neural Networks. In Proceedings of the 2016 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 19–21 December 2016; pp. 247–251.
103. Abdel-Hamid, O.; Mohamed, A.R.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1533–1545. [CrossRef]
104. Yan, Y.; Chen, M.; Saad Sadiq, M.; Shyu, M.-L. Efficient Imbalanced Multimedia Concept Retrieval by Deep Learning on Spark Clusters. *Int. J. Multimed. Data Eng. Manag. IJMDEM* **2017**, *8*, 1–20. [CrossRef]
105. Yan, Y.; Chen, M.; Shyu, M. Deep learning for imbalanced multimedia data classification. In Proceedings of the 2015 IEEE international symposium on multimedia (ISM), Miami, FL, USA, 14–16 December 2015; pp. 483–488.
106. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2016**, arXiv:1408.5882.

107. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MA, USA, 22–27 June 2014; pp. 655–665.
108. Dos Santos, C.; Gatti, M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 69–78.
109. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Eventh IEEE International Conference on Computer Vision, TKerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
110. Dalal, N.; Histograms, B.T.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
111. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1–21. [CrossRef]
112. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
113. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386. [CrossRef] [PubMed]
114. Fukushima, K.; Miyake, S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In Proceedings of the Competition and cooperation in neural nets, Berlin, Heidelberg, Kyoto, Japan, 15–19 February 1982; pp. 267–285.
115. Jordan, M. Serial order: A parallel distributed processing approach. *Adv. Psychol.* **1997**, *121*, 471–495.
116. LeCun, Y.; Bottou, L.; Bengio, Y. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
117. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
118. Misic, M.; Đurđević, Đ.; Tomasevic, M. (PDF) Evolution and Trends in GPU Computing. Available online: https://www.researchgate.net/publication/261424611_Evolution_and_trends_in_GPU_computing (accessed on 19 August 2021).
119. Raina, R.; Madhavan, A.; Ng, A.Y. Large-scale deep unsupervised learning using graphics processors. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 873–880.
120. Osborne, J. Google's Tensor Processing Unit Explained: This is Google Scholar. Available online: https://scholar.google.com/scholar?q=Google%27s+Tensor+Processing+Unit+explained%3A+this+is+what+the+future+of+computing+looks+like (accessed on 21 August 2021).
121. Ian, G.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; pp. 436–444.
122. Azarkhish, E.; Rossi, D.; Loi, I. Neurostream: Scalable and energy efficient deep learning with smart memory cubes. *Proc. IEEE Trans. Parallel Distrib. Syst.* **2017**, *29*, 420–434. [CrossRef]
123. McMahan, H.; Moore, E.; Ramage, D.; y Arcas, B. Federated learning of deep networks using model averaging. *arXiv* **2016**, arXiv:1602.05629.
124. Yan, Y.; Zhu, Q.; Shyu, M.-L.; Chen, S.-C. A Classifier Ensemble Framework for Multimedia Big Data Classification. In Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA, USA, 28–30 July 2016; pp. 615–622.
125. Kaiser, Ł.; Brain, G.; Gomez, A.N.; Shazeer, N.; Vaswani, A.; Parmar, N.; Research, G.; Jones, L.; Uszkoreit, J. One Model to Learn Them All. *arXiv* **2017**, arXiv:1706.05137.
126. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012): 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
127. Goller, C. *A Connectionist Approach for Learning Search Control Heuristics for Automated Deduction Systems*; Akademische Verlagsgesellschaft AKA: Berlin, Germany, 1999; pp. 1–8.
128. Socher, R.; Chiung, C.; Lin, Y.; Ng, A.Y.; Manning, C.D. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In Proceedings of the 28th International Conference on Machine Learning, ICML, Bellevue, WA, USA, 28 June 28–2 July 2011; pp. 129–136.
129. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
130. Li, X.; Wu, X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In Proceedings of the ICASSP 2015—2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4520–4524.
131. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
132. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

133. Zhang, S.; Zheng, D.; Hu, X.; Yang, M. Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation, Shanghai, China, 30 October–1 November 2015; pp. 73–78.
134. Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
135. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
136. Alec, R.; Karthik, N.; Tim, S.; Ilya, S. Improving language understanding with unsupervised learning. *Citado* **2018**, *17*, 1–12.
137. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.; Language, G.A.I. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
138. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202. [CrossRef]
139. Fukushima, K.; Miyake, S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.* **1982**, *15*, 455–469. [CrossRef]
140. Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *arXiv* **2014**, arXiv:1404.5997, 1–7.
141. Socher, R.; Huang, E.; Pennin, J. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1–9.
142. Hinton, G. Deep belief networks. *Scholarpedia* **2009**, *4*, 5947. [CrossRef]
143. Hinton, G.; Deng, L.; Yu, D. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
144. Salakhutdinov, R.; Statistics, G.H. Deep boltzmann machines. In Proceedings of the Artificial Intelligence and Statistics, Hilton Clearwater Beach Resort, Clearwater Beach, FL, USA, 16–18 April 2009; pp. 448–455.
145. Salakhutdinov, R.; Hinton, G. An efficient learning procedure for deep Boltzmann machines. *Neural Comput.* **2012**, *24*, 1967–2006. [CrossRef]
146. Fausett, L. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*; Pearson Education: London, UK, 1994; p. 480.
147. Mikolov, T.; Karafiát, M.; Burget, L.; Honza, J.; Cernocky, J.H.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
148. Mikolov, T.; Kombrink, S.; Burget, L.; Černocký, J.; Khudanpur, S. Extensions of recurrent neural network language model. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 5528–5531.
149. Mikolov, T.; Deoras, A.; Povey, D.; Burget, L. Strategies for training large scale neural network language models. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village Resort, Big Island, HI, USA, 11–15 December 2011; pp. 196–201.
150. El Hihi, S.; Bengio, Y. Hierarchical recurrent neural networks for long-term dependencies. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1996; pp. 493–499.
151. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
152. Radford, A.; Metz, L. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
153. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114, 1–14.
154. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
155. Greff, K.; Srivastava, R.; Koutník, J. LSTM: A search space odyssey. *IEEE Trans. Neural Networks Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef] [PubMed]
156. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
157. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, PA, USA, 6–9 July 2002; pp. 79–86.
158. Harris, Z.S. Distributional Structure. *Word* **1954**, *10*, 146–162. [CrossRef]
159. Popov, M.; Kulnitskiy, B.; Perezhogin, I.; Mordkovich, V.; Ovsyannikov, D.; Perfilov, S.; Borisova, L.; Blank, V. A Neural Probabilistic Language Model. *Fuller. Nanotub. Carbon Nanostruct.* **2003**, *3*, 1137–1155.
160. Collobert, R.; Bengio, S.; Mariethoz, J. Torch: A Modular Machine Learning Software Library. Available online: http://publications.idiap.ch/downloads/reports/2002/rr02-46.pdf (accessed on 15 October 2020).
161. Abadi, M.; Agarwal, A.; Barham, E.B. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.

162. Skymind Skymind. Deeplearning4j Deep Learning Framework. 2017. Available online: https://deeplearning4j.org/ (accessed on 16 October 2020).
163. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
164. Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; Zhang, Z. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv* **2015**, arXiv:1512.01274.
165. Al-Rfou, R. Theano: A Python Framework for Fast Computation of Mathematical Expressions. *arXiv* **2016**, arXiv:1605.02688.
166. Agarwal, A.; Akchurin, E.; Basoglu, C.; Chen, G.; Cyphers, S.; Droppo, J.; Eversole, A.; Guenter, B.; Hillebrand, M.; Huang, X.; et al. An Introduction to Computational Networks and the Computational Network Toolkit. MSR-TR-2014-112 (DRAFT Vol.1.0). 2016, pp. 1–50. Available online: https://www.microsoft.com/en-us/research/wp-content/uploads/2014/08/CNTKBook-20160217.pdf (accessed on 1 October 2021).
167. NervanaSystems. The Neon Deep Learning Framework. Available online: https://github.com/NervanaSystems/neon (accessed on 11 May 2017).
168. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
169. Wood, M. Introducing Gluon: A New Library for Machine Learning from AWS and Microsoft: Introducing Gluon. 2017. Available online: https://aws.amazon.com/blogs/aws/introducing-gluon-a-new-library-for-machine-learning-from-aws-and-microsoft/ (accessed on 1 October 2021).
170. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.L.; Chen, S.C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. CSUR* **2018**, *51*, 1–36. [CrossRef]
171. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
172. Turian, J.; Ratinov, L.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 384–394.
173. De Silva, T.N.; Zhibo, X.; Rui, Z.; Kezhi, M. Causal relation identification using convolutional neural networks and knowledge based features. *Int. J. Comput. Syst. Eng.* **2017**, *11*, 696–701.
174. Kruengkrai, C.; Torisawa, K.; Hashimoto, C.; Kloetzer, J.; Oh, J.-H.; Tanaka, M. Improving Event Causality Recognition with Multiple Background Knowledge Sources Using Multi-Column Convolutional Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3466–3473.
175. Ciresan, D.; Meier, U.; Schmidhuber, J. Multi-column Deep Neural Networks for Image Classification. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 16–21 June 2012; pp. 3642–3649.
176. Oh, J.; Torisawa, K.; Kruengkrai, C.; Iida, R.; Kloetzer, J. Multi-column convolutional neural networks with causality-attention for why-question answering. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, CA, USA, 6–10 February 2017; pp. 415–424.
177. Ponti, E.M.; Korhonen, A. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Valencia, Spain, 3 April 2017; pp. 25–30.
178. Roemmele, M.; Gordon, A.S. An Encoder-decoder Approach to Predicting Causal Relations in Stories. In Proceedings of the First Workshop on Storytelling, New Orleans, Louisiana, 5 June 2018; pp. 50–59.
179. Dasgupta, T.; Saha, R.; Dey, L.; Naskar, A. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, 12–14 July 2018; pp. 306–316.
180. Nauta, M.; Bucur, D.; Seifert, C. Causal Discovery with Attention-Based Convolutional Neural Networks. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 312–340. [CrossRef]
181. Ayyanar, R.; Koomullil, G.; Ramasangu, H. Causal Relation Classification using Convolutional Neural Networks and Grammar Tags. In Proceedings of the 2019 IEEE 16th India Council International Conference (INDICON), Marwadi University, Rajkot, India, 13–15 December 2019; pp. 1–3.
182. Li, P.; Mao, K. Knowledge-oriented Convolutional Neural Network for Causal Relation Extraction from Natural Language Texts. *Expert Syst. Appl.* **2019**, *115*, 512–523. [CrossRef]
183. Kayesh, H.; Islam, M.S.; Wang, J. On Event Causality Detection in Tweets. *arXiv* **2019**, arXiv:1901.03526.
184. Kayesh, H.; Islam, M.S.; Wang, J.; Kayes, A.S.M.; Watters, P.A. A deep learning model for mining and detecting causally related events in tweets. *Concurr. Comput. Pract. Exp.* **2020**, e5938. [CrossRef]
185. Kadowaki, K.; Iida, R.; Torisawa, K.; Oh, J.H.; Kloetzer, J. Event causality recognition exploiting multiple annotators' judgments and background knowledge. 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), Hong Kong, China, 3–7 November 2019; pp. 5816–5822.
186. Mehrabadi, M.A.; Dutt, N.; Rahmani, A.M. The Causality Inference of Public Interest in Restaurants and Bars on COVID-19 Daily Cases in the US: A Google Trends Analysis. *JMIR Public Health Surveill.* **2020**, *7*, 1–6.

187. Liu, J.; Chen, Y.; Zhao, J. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), Yokohama, Japan, 7–15 July 2020; pp. 3608–3614.
188. Speer, R.; Lowry-Duda, J. ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge. *arXiv* **2017**, arXiv:1704.03560.
189. Ma, J.; Dong, Y.; Huang, Z.; Mietchen, D.; Li, J. Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US. *arXiv* **2021**, arXiv:2106.01315.
190. Li, Z.; Li, Q.; Zou, X.; Ren, J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing* **2021**, *423*, 207–219. [CrossRef]
191. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
192. Khetan, V.; Ramnani, R.; Anand, M.; Sengupta, S.; Fano, A.E. Causal-BERT: Language models for causality detection between events expressed in text. *arXiv* **2021**, arXiv:2012.05453v2, 965–980.
193. Gurulingappa, H.; Rajput, A.; Roberts, A.; Fluck, J. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed.* **2012**, *45*, 885–892. [CrossRef] [PubMed]