future internet

MDPI

Article

Analysis of Digital Information in Storage Devices Using Supervised and Unsupervised Natural Language Processing Techniques

Luis Alberto Martínez Hernández, Ana Lucila Sandoval Orozco and Luis Javier García Villalba *

Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain

* Correspondence: javiergv@fdi.ucm.es; Tel.: +34-91-394-7638

Abstract: Due to the advancement of technology, cybercrime has increased considerably, making digital forensics essential for any organisation. One of the most critical challenges is to analyse and classify the information on devices, identifying the relevant and valuable data for a specific purpose. This phase of the forensic process is one of the most complex and time-consuming, and requires expert analysts to avoid overlooking data relevant to the investigation. Although tools exist today that can automate this process, they will depend on how tightly their parameters are tuned to the case study, and many lack support for complex scenarios where language barriers play an important role. Recent advances in machine learning allow the creation of new architectures to significantly increase the performance of information analysis and perform the intelligent search process automatically, reducing analysis time and identifying relationships between files based on initial parameters. In this paper, we present a bibliographic review of artificial intelligence algorithms that allow an exhaustive analysis of multimedia information contained in removable devices in a forensic process, using natural language processing and natural language understanding techniques for the automatic classification of documents in seized devices. Finally, some of the open challenges technology developers face when generating tools that use artificial intelligence techniques to analyse the information contained in documents on seized devices are reviewed.

Keywords: artificial intelligence; digital information; computer forensic; entity extraction; entity recognition; storage devices; text processing

Citation: Martínez Hernández, L.A.; Sandoval Orozco, A.L.; García Villalba, L.J. Analysis of Digital Information in Storage Devices Using Supervised and Unsupervised Natural Language Processing Techniques. *Future Internet* **2023**, *15*, 155. https://doi.org/10.3390/fi15050155

Academic Editor: Ivan Serina

Received: 12 February 2023
Revised: 17 April 2023
Accepted: 18 April 2023
Published: 23 April 2023

1. Introduction

Due to the rise of technology, we live in an increasingly connected world. It is now common to have more than one device connected to the internet, constantly sharing information. Because of this, cybercriminals are looking for ways to commit illegal actions, protected by the privacy of the internet. For this reason, law enforcement agencies need to pay more attention to this type of problem, in order to prevent crimes that could harm people from being committed. According to [1], the number of cybercrime suspects arrested between 2011 and 2019 has increased considerably, from 4800 in 2011 to almost 9000 in 2019. When a suspect is arrested, it is very common for material assets to be seized, including technological devices that can store evidence of a crime. One of the activities that agents must carry out is a content analysis, using forensic methodologies on the seized electronic devices in search of evidence, however this task can be one of the most complex due to the large amount of data that may be stored on a device, and relationships between documents must be sought in order to generate evidence that can be presented in a trial. Although there are tools that allow the analysis process to be automated, they must be correctly adjusted to the purpose of the process. In addition, constant supervision by the analyst will be necessary to ensure that details are not overlooked that could be a precursor to a clue that

could lead to crucial evidence for a legal case. Advances in different areas of machine learning allow the creation of architectures to increase the performance of information extraction and analysis in devices, by performing an intelligent search process trying to find relationships between files based on the search for a given term. Automatic document classification is a technique that makes use of artificial intelligence to sort documents into classes or categories. Many document classification tools make use of natural language processing techniques for document analysis and relationship extraction.

This paper presents a literature review of the artificial intelligence algorithms and techniques that can be used to analyse the context of documents contained in removable devices retrieved during a digital evidence collection process in order to find relationships between them automatically, providing relevant information to investigations and reducing human error. It also provides a comparison of the main algorithms and methods analysed. The rest of the work is organised as follows: In Section 2, a description of each of the phases of the digital forensic process is given. Section 3 describes the text processing techniques currently available and reviews the literature, to verify the areas of application of the methods. A detailed description of natural language understanding is given in Section 4. In Section 5, the challenges faced by AI models for processing, understanding, and classifying text documents are analysed. Finally, the conclusions of the work are included in Section 6.

2. Computer Forensics

Computer forensics, also defined as computer forensic science [2], is a speciality of digital forensics that focuses on finding evidence in computers, servers, smartphones or digital storage media based on the scientific method. Computer forensics is the process of preserving, identifying, acquiring, and analysing electronic evidence that can be used in a judicial process, by using methodological resources to search for evidence on digital devices such as computers, smartphones, servers, or the internet. The process of computer forensics consists of several stages (see Figure 1), the process starts with the identification of digital material on a storage device, under the assumption that it may be potential evidence in criminal proceedings, and ends when an expert presents the final report with the findings of the analysis [3]. All of the following steps, executed sequentially, constitute the digital forensic investigation development lifecycle [4,5].

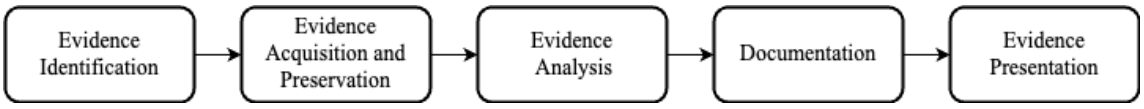


Figure 1. Digital Forensics Process.

- Identification: Identifying digital evidence is the first step in the digital forensic process. This step involves identifying one or more storage sources such as hard drives, USB sticks, SD memory sticks, mobile phones, remote storage services, IoT devices, and virtualised equipment. Typically, these devices are derived from a legal process and must be properly seized, following the chain of custody, and isolated to prevent any tampering with potential evidence. When the investigation is conducted on commonly used equipment in an organisation, such as servers, network equipment, or cloud-hosted services, the investigation team and the organisation must ensure that no one other than the investigation team's analysts have access to them.
- Evidence examination: This step is performed using forensic tools and methodologies for the extraction of data that may be useful in an investigation in a legal process, storing all evidence securely. The evidence must be securely stored in various storage devices to leave the original information, also called the "forensic image", intact until they are needed for further research.
- Analysis: In the process of analysing the extracted data, the analyst thoroughly investigates the extracted information in order to identify, interpret, classify, and convert it into useful research information, using specialised tools and techniques.

This process is perhaps the most complex and can take the most time to analyse. To be successful, the analyst must be experienced in looking for patterns of information that can add value to the research.

- Documentation: Once the digital evidence has been obtained, the relevant documentation of the findings is carried out, providing a summary and conclusion of the research carried out [6].
- Presentation: Data obtained through proper forensic methodology, using the scientific method, may be admitted by a judge as evidence in litigation cases.

The analysis of files from a device that is part of a legal process is one of the most laborious tasks, due to two factors: (1) the fact that the data must be given a meaning in order to be relevant to an investigation, and (2) the variety of data collected from the approach of the media to be represented, which can be an image, video, audio, or text [7]. However, there are many challenges that analysts have to face, among which are: the large amount of storage space, in terabytes, that can currently be accessed by anyone and the different types of files and data included in the collection process. This paper presents a collection of different artificial intelligence techniques that can help to generate tools capable of automatically identifying and classifying a document by its context.

3. Natural Language Processing

In an increasingly connected world, and due to the sheer volume of textual data that exists, it is increasingly difficult for human beings to discover knowledge that can be useful to society, especially when there is a time limit to identify it [8]. Natural language processing (NLP) is an area of research and development that mainly performs written and spoken language analysis and generation, having its beginnings in tasks such as cryptanalysis and machine translation. NLP emphasises text processing and the applications of each of the developments, and as such can be considered as a strand of computational linguistics, which focuses on the analysis and formal modelling of language and its applications while maintaining a very close relationship with linguistics, computer science, and psychology [9].

Currently, natural language processing has attracted particular interest in the academic community. Linguistic technologies are gaining ground and are gradually spreading their use in professional sectors, with the aim of discovering, classifying, organising, or searching content automatically, which allows for a more efficient use of time, cost reduction, and agile decision making in organisations. One of the applications of NLP is in named-entity recognition (NER) detector tools. As the name suggests, NER detects entities such as people, locations, organisations, or brands. NER uses machine learning technology, rules, and linguistic corpora to identify entities based on words or phrases effectively and classify them from a set of items with similar characteristics [10].

Figure 2 shows a general outline of the relationship between named-entity recognition and relationship extraction techniques. A common way to implement NER is through the use of dictionaries and entity lists. Dictionaries are sets of words or phrases that are used to identify a specific entity [11]. For example, a dictionary of people's names might contain common names such as "John", "Mary", and "Charles". Entity lists, on the other hand, are lists of words or phrases that are used to identify a category of entity. In addition, like NER, dictionaries and relationship lists can be used to implement relationship extraction [12], but with the difference that dictionaries are sets of phrases or words that are used to find a specific relationship between two entities. For example, a dedicated dictionary for financial relationships may contain terms such as "acquire", "invest", "buy", and "sell". Relationship lists, on the other hand, are a set of phrases or words that are used to identify a category of relationship, for example, containing terms such as "financing" or "acquisition".

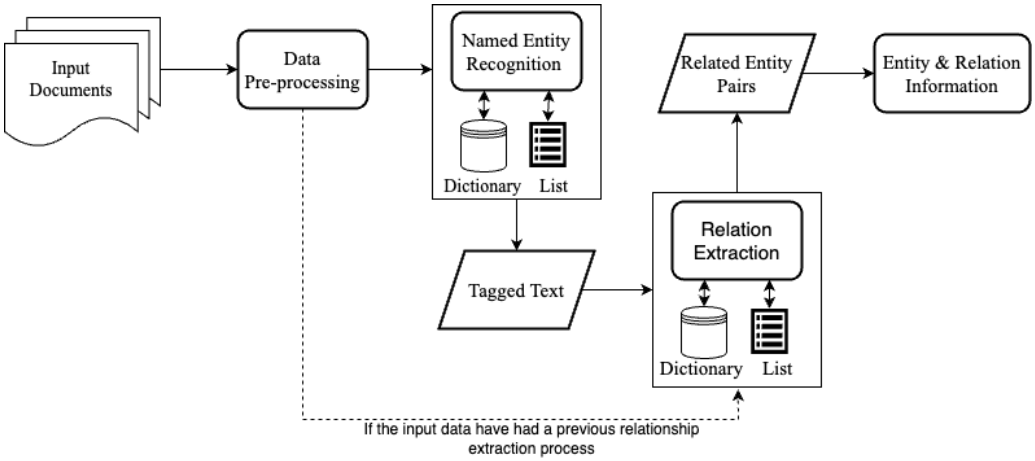


Figure 2. General Diagram of Operation of Named-Entity Extraction and Relationship Extraction.

3.1. Named-Entity Recognition

The first step is to detect a word or phrase that forms entities, where each word represents a token, e.g., "The bluebirds fly high" is an entity formed of four tokens. One form of labelling is inside–outside–outside–principle, which allows the indication of where words begin and end [13]. The second step is the generation of entity categories. These categories allow the algorithms to identify the different ways of writing a certain type of entity, e.g., dates could be "mm/dd/yyyy or yy/mm/dd". An NER tool allows text to be tagged according to its context, and this tagged text is assigned a label to differentiate it from other categories. NER models, in their most basic version, seek to identify only some types of entities such as people, organisations, or places, however, models can also be found that allow the identification of streets and dates, among others. Figure 3 shows an example of the recognition of entities in a given text carried out by an NER tool which is able to identify in the text certain patterns that refer to people (*PER*), locations (*LOC*), dates (*DATE*), and organisations (*ORG*). An important aspect of NER models is that previously they have only been able to focus on a single language such as Spanish, English, or Portuguese and a single subject such as sports, news, or justice.

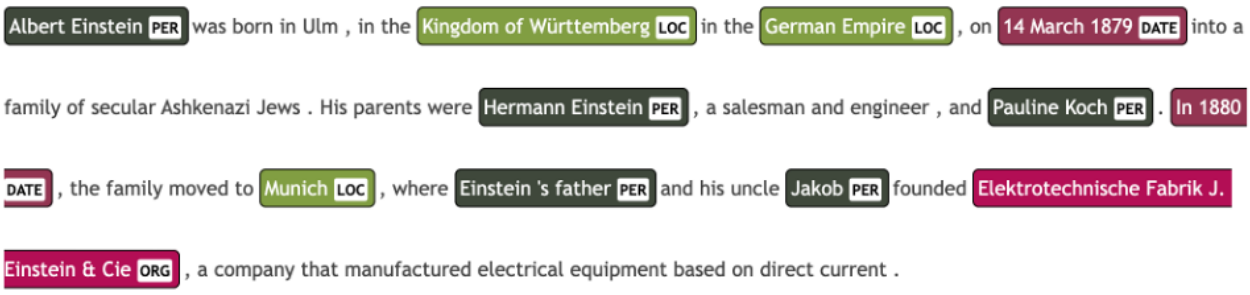


Figure 3. Example of Labelling Entities.

An important aspect to consider to be successful is to have quality labelling for learning models, such as evaluation. A labelled corpus is a theme-specific text containing annotations of one or more entities [14]. Table 1 shows a summary of the datasets and tools to perform the named-entity recognition task. Recently, the data sources used for corpus creation include text conversations such as Twitter comments, film reviews, or social media comments. In addition, Wikipedia articles allow the number of tags per corpus type to increase.

Table 1. NER Datasets.

Name	Language	Text Source	#Tags
CoNLL 2003 [15]	English, German, Dutch	Reuters news	4
Ontonotes V5 [16]	English, Chinese, Arabic	Mobile conversations, religious texts, newsgroups, broadcast news and conversation weblogs, newscast	18
NCBI disease [17]	English	PubMed	14
WNUT 2017 [18]	English	Emerging discussions	6
GENIA [19]	English	PubMed	36
WNUT 2020 [20]	English	Twitter	2
LeNER-Br [21]	Portuguese	Legal text	6
WikiFiger [22]	English	Wikipedia	112
WikiNEuRal [23]	English, Italian, German, Dutch, French, Portuguese, Russian, Polish	PubMed	5
Broad Twitter corpus [24]	English	Twitter	3

Thanks to the use of modern data sources such as digital encyclopaedias or Twitter, it is possible to increase both the number of tags and the size of corpora for training NLP models. An example of this is the WikiFiger dataset [22], which has 112 tags of different topics. This improvement has also been noticed in corpora with the same subject matter, for example in corpora dedicated to medicine, NCBI disease [17], with 14 tags compared to GENIA [19], which has 36 tags.

In addition to the datasets available to train your own models, there are now pre-trained online tools that allow one to perform automatic or semi-automatic named-entity recognition of any text on a language-dependent basis. Table 2 shows some of these tools. These make use of one or more of the datasets mentioned in Table 1 and usually have an API with which developers can create their own applications using relationship extraction models. However, these tools may inherit the language limitations of the datasets they were trained on.

Table 2. Pretrained NER tools.

Tool	Platform	Source URL
GATE	Java	[25]
NLTK	Python	[26]
Stanford	Java	[27]
Spacy	Python	[28]
Poliglot	Python	[29]
Flair	Python	[30]
DeepPavlov	Python	[31]
Allen	Python	[32]
Annie	Python	[33]

3.2. Relationship Extraction

Relation extraction (RE) allows the extraction of relations between two or more previously identified entities [34] (e.g., persons, places, ID), in order to classify them into a set of characteristics (e.g., "son of", "employee of", or "lives in") using semantic relations in the text. This process allows structured knowledge to be acquired from unstructured data. For example, in Figure 4 the phrase "Madrid is in Spain" is establishing an "is in" relationship between Madrid and Spain.

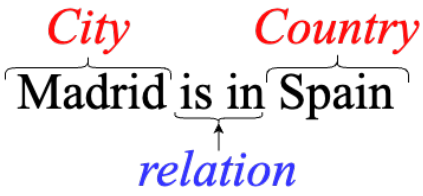


Figure 4. Relationship Example.

One way to represent the detected relationships and facilitate understanding is by means of a diagram. Figure 5 shows the graph of the identified relations in a text, in which some characteristics are added to facilitate the interpretation of the entities, for example "*born in*" , "*occupation*", "*developed*", "*part of*".

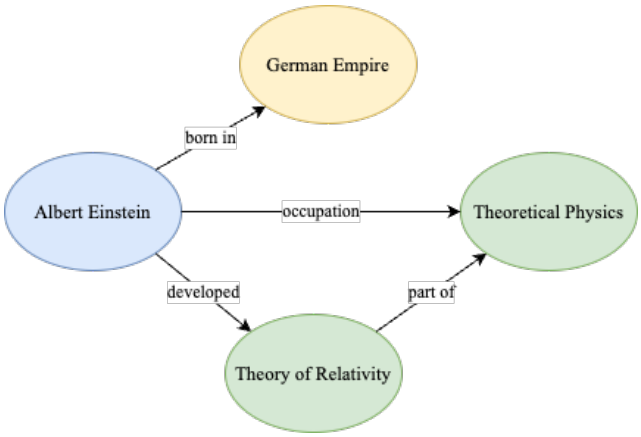


Figure 5. Graphical Representation of Relationships.

There are a few different methods to perform relationship extraction:

- RE rule-based: In rule-based methods, a manual analysis of a set of sentences is carried out in order to identify what sentences that include a relation look like [34]. This method seeks to identify patterns of the type 1, where *X_1* and *Y_1* are identified entities and alpha (α) is intermediate words. These types of patterns are called word sequence patterns, because they follow a coherent syntactic and semantic order in the text, however, the implemented rules are inefficient for sequences of larger scope and variety, e.g., "John and Mary got married". In this example, the rule specifies a pattern that follows the sequence of the text, which defines easily identifiable patterns in word sequences.

$$(X\_1\ \alpha\ Y\_1) \quad (1)$$

- Weakly supervised RE: This method starts with a set of manually created rules and, based on these, finds new ones from the unmarked text data. One way to start is to create a set of "seed tuples" that describe specific relationships between entities [35]. For example, seed = seed = {(PER: Bob, LOC: USA), (PER: Joe, LOC: Spain)} these seeds establish entities that have a relationship based on the "is in" relationship PER is in LOC), from this pattern new ones are generated from the text recursively, (PER:

Alice, LOC: Germany). This method is increasingly error-prone and new seeds will be needed when new types of relations are needed.

- Supervised RE: The most common way to extract existing relationships is to train a binary classifier that determines whether a relationship exists between two entities [36,37]. These binary classifiers take as input semantic and syntactic features of the text, which requires the text to be previously marked by other NLP methods.
- Distantly supervised RE: This method combines the techniques of using a classifier and seed data, with the difference that instead of using a set of tuples [38], all knowledge is taken from an existing knowledge base such as Freebase, DBpedia, Ontonotes, Wikipedia, WNUT, Yago. It allows for reduced manual effort, as well as the scalability to use a large number and variety of tags and relations. This method is limited to the knowledge base used and, in case it is needed for the research in question, an adjustment of the trained data will be necessary.
- Unsupervised RE: This method, unlike the previous ones, is based on a very general and heuristic set of constraints, so there is no need to use labelled data, seed sets or rules are used to capture the different relationships in the text [39]. In this method, more general rules of thumb are used to find the tuples. For some cases, even taking advantage of small labelled text datasets to design and modify systems. In general, these methods tend to require less supervision overall.

There are currently multiple works related to the extraction of relationships based on identified entities. This paper focuses its research on supervised and unsupervised relationship extraction methods, due to the large amounts of data derived from a judicial process that can be unstructured.

3.3. Supervised RE

Supervised relationship extraction models are used in a variety of applications where it is necessary to extract specific relationships between entities in unstructured data, such as text. Examples of the use of supervised relationship extraction models include medical information extraction, sentiment analysis, social network monitoring, and financial analysis. However, it is important to note that these models require a labelled dataset and significant effort in the training process. Among the most common characteristics among the supervised relationship extraction methods are word distance, word context, entity dependency, part-of-speech labels, tokens, NER, and labels. This method ensures that the relationships extracted from the text are the most relevant. However, it is a method that has difficulty in including new relationships (a new classifier has to be trained) and is only effective for a reduced set of types of relationships between entities.

In [40], a model based on multi-task learning for relation extraction is proposed. It uses several models, the first one helps the entity extraction model to obtain an abstract representation of multitasking, by adding other auxiliary tasks. This allows the extraction of additional semantic information from the relation extraction model in a single task. In addition, they include single and auxiliary task models, that learn from relation extraction by obtaining knowledge for each task, which, through distillation and knowledge extraction algorithms, improves their performance.

Kumar Sahu et al. [41] demonstrate a method for relation extraction that automatically learns features using convolutional neural networks, reducing the dependence on manual feature engineering. This proposal takes as input a complete sentence with the entities and generates a vector of probabilities corresponding to the total number of possible relationship types. For each feature, there is a randomly initialised vector representation, except for word embedding, for which a pretrained word vector model is used by learning from PubMed articles.

3.4. Unsupervised RE

Unsupervised relationship extraction models are used in a variety of applications where it is necessary to discover unknown patterns and relationships in unstructured data, such as text. Examples of use are text clustering, social network analysis, discovery of additional relationships, and anomaly detection. In this method, more general rules of thumb are used to find the tuples. In some cases, even small labelled text datasets are exploited to design and modify systems. However, these models can be more difficult to interpret and validate than supervised models, as there is no labelled dataset to compare the results with.

Genest et al. [42] propose a method, PromptORE, that adapts an embedding framework to work in an unsupervised statement-based environment, and is used to embed operations that express relationships and does not require hyperparameter tuning. The embeddings are then grouped into clusters to discover relationships and the appropriate number of clusters is automatically estimated. In addition, in [43] a method using sentence supervision for unsupervised relation extraction is presented, which uses the SBERT-based pretrained model for unsupervised relation extraction and sentence encoding. The model uses a clustering algorithm to classify identical patterns and the extraction of relationships between entities in a sentence, calculating a confidence value to avoid semantic drift between sentences. In [44], general domain knowledge is incorporated, which allows first-order rules to be encoded and automatically combined with the model developed for relation extraction. The paper proposes an unsupervised approach to relation extraction that does not require relation training data and allows the incorporation of global constraints expressing domain knowledge, encoding it as first-order logic rules and integrating it automatically with a thematic model, to produce clusters formed by the available data and constraints.

In [45], the authors present GraphRel, a graph convolutional network (GCN)-based model for extracting and learning entities and relationships. This model handles three key aspects of relation extraction. GraphRel automatically extracts the features of each sentence by stacking a Bi-LSTM sentence encoder and a dependency tree encoder, labels the words, and predicts triplets of relations connecting the mentions. The first phase of the proposed model extracts the hidden features of the nodes along the dependencies and a new connected graph is established with edges weighted by relations, and the second phase considers the interaction of entities and relations before the final classification.

In [46], a method for relationship extraction is created using the BERT model, which consists of predicting the relationship between two entities, given a sentence and two non-overlapping entity sections. First, the input sequence of type [[CLS] sentence [SEP] subject [SEP] object [SEP]] is constructed, to avoid overfitting, the entities are replaced by masks composed of argument type, entity argument type, and entity type.

For the extraction of relations at document level, there is the proposal given in [47], which makes use of different nodes and edges to generate a document-level graph. Inferences on the edges of the graph allow the learning of intra- and inter-entity relationships using multi-instance learning. The model extracts neural relationships based on a partially connected graph of entities, where entity mentions constitute the nodes, and directed edges correspond to ordered pairs of entity mentions. It uses multi-instance learning when mention-level annotations are available. An important aspect to consider, is that in order to achieve the extraction of document-level relationships, representative datasets related to the research topic are necessary, in this sense Yao et al. [48] propose DocRed, a dataset derived from Wikipedia and Wikidata, this dataset allows the extraction of entities and relationships between them. For its correct functioning, it is necessary to read multiple sentences from the document to extract relationships and thus predict their relationships. An important aspect is that it can be used in both supervised and weakly supervised environments. Table 3 summarises the analysed papers related to relation extraction techniques. Although most of the techniques were created to analyse short sentences, they can be

adjusted to process large volumes of text contained in documents with the same results and trying to find relations between two or more documents.

Table 3. Relationship Extraction Proposals.

Proposal	Technique	Approach	Scope
Wang et al. [40]	Extraction-Based Relationships	Supervised	Phrase
Genest et al. [42]	Instruction-Based Relationship Extraction	Unsupervised	Phrase
Ali et al. [43]	Phrase Supervision	Unsupervised	Phrase
De Lacalle et al. [44]	General Domain Knowledge	Unsupervised	Phrase
Fu et al. [45]	Convolutional Graph Networks	Unsupervised	Phrase
Sahu et al. [41]	Convolutional Neural Networks	Supervised	Phrase
Shi et al. [46]	BERT	Unsupervised	Phrase
Yao et al. [48]	Dataset from Wikipedia and Wikidata	Supervised/Unsupervised	Dataset
Christopoulou et al. [47]	Multi-Instance Learning Graphs	Unsupervised	Document

4. Natural Language Understanding

Natural language understanding (NLU) or natural language interpretation (NLI) [49] is an area of knowledge that studies the automatic understanding of text. This field has gained special interest nowadays, as it can be applied to different areas where society currently demands innovative solutions. For example, in [50] an identification of linguistic forms based on the frequencies of word usage in legal documents is made possible by using a statistical approach to explore the language and extract the linguistic forms it contains. The data extracted by the proposed method can be used to analyse links and references and search for information in the documents, as well as looking for links between correlated legal documents and establishing relevance between them. By using NLU tasks it is possible to perform neural machine translation tasks, for example, in [51] a BERT-based model is proposed, to extract input sequence representations and then fuse them with each layer of the NMT model using attentional models. In [52], a forgery detection model is proposed for the veracity of an article question, taking into account phrase matching based on key phrase retrieval. Moreover, ref. [53] proposes a hybrid model that overcomes the limitations of sequence models by making use of the RoBERTa model for word or sub-word tokenisation and word embedding generation, and also makes use of the LSTM model for encoding long-distance temporal dependencies in word embedding. In the voice command area, ref. [54] proposes a sequence-by-sequence neural architecture for training NLU models to learn intention prediction tasks, labels, and values slots, which is possible without the need for aligned data.

Further, for archiving and content analysis, in [55] the authors present a variational neural decoder (VND), that makes use of latent variables to model the semantics at each time step of the source and target texts. This is possible by introducing a variational autoencoder (VAE) in the decoding process and incorporating latent variables into the hidden state of the VND.

In recent years, with the transformer revolution, deep learning models for NLU tasks have gained prominence in the scientific community and industry. A transformer model [56] has multiple encoders and decoders stacked together, auto attendant care in each of the encoded and decoded units, and cross attendant care between the encoders and decoders [57]. In this section, some of the natural language comprehension models are described and a general description of some works that make proposals based on these algorithms is given. Figure 6 shows a diagram summarising the different themes in which the techniques investigated in this paper have been used.

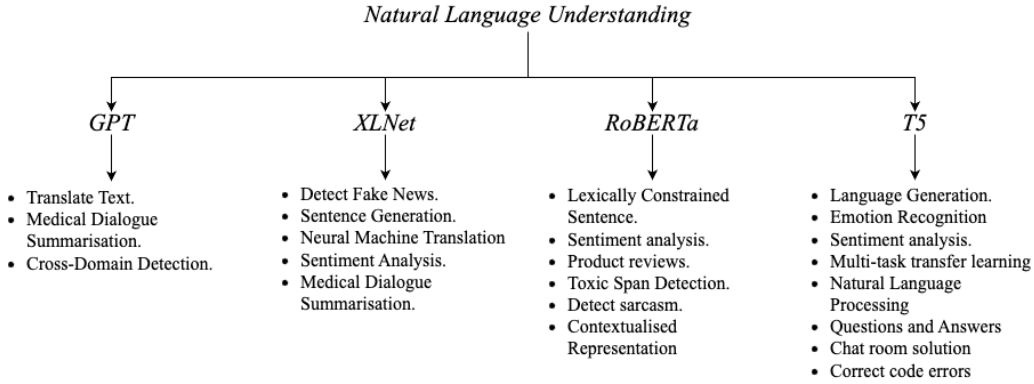


Figure 6. Uses of Natural Language Understanding Techniques.

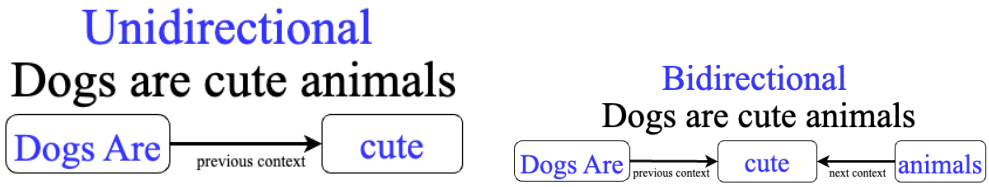
Table 4 shows a summary of the natural language understanding models studied, highlighting the characteristics of the models, the identification method used, and the number of parameters required for training; this last one is necessary to take into consideration when it is necessary to implement the algorithm in a productive environment, since having a low number of parameters means that the processing resources required to perform the inferences are low, however, on some occasions accuracy is sacrificed for performance.

Table 4. Natural Language Understanding Models.

Model	#Parameters	Features Layer/Hidden/Heads	Method
BERT [58]	Base: 110M Large: 340M	12/768/12 24/1024/16	Bidirectional transformer MLM, NSP
RoBERTa [59]	Base: 123M Large: 355M	12/768/12 24/1024/16	BERT without NSP using dynamic masking
XLNet [60]	Base: 110M Large: 340M	12/768/12 24/1024/16	Bidirectional transformer
GPT [61]	110M	12/768/12	Transformers
DistilBERT [62]	134M	6/768/12	BERT distillation
ALBERT [63]	Base: 11M Large: 17M	12/768/12 24/1024/16	BERT with reduced parameters, SOP (not NSP)
T5 [64]	Base: 220M Large: 770M	12/768/12 24/1024/16	Text-to-text

4.1. Bidirectional Encoder Representations From Transformers

Bidirectional encoder representations from transformers (BERT) is a technique based on neural networks, for NLP pretraining. The aim of Google's algorithm is to interpret our search language in a more natural way, using neuro-linguistic programming. To do this, BERT uses an open source neural network to process the natural language of entered searches, which is achieved through bidirectionality [58]. This consists of analysing the same phrase in two directions, from left to right and from right to left of the keyword. Figure 7 shows an example of unidirectional and bidirectional parsing, where in Figure 7a it only takes as reference the previous context of the sentence and in Figure 7b it takes both the previous and the next context of the sentence, this will allow the algorithm to better understand what the text is trying to convey and the topic of each sentence in depth [65].



(**a**) (**b**)

Figure 7. Unidirectional and Bidirectional sentence analysis. (**a**) *Unidirectional*. (**b**) *Bidirectional*.

4.2. Generative Pretrained Transformer

A generative pretrained transformer (GPT) is an NLP algorithm based on deep learning, that allows for the generation of human-like text from text input [61]. For the model to work, only one sentence is needed, the transformer creates meaningful information and text based on the context of the given sentence using publicly available datasets. The technology can process any type of text, including computer code. GPT models are trained in order to predict subsequent tokens based on all previously identified tokens. This is achieved through language autoregression. There are two additional versions of GPT: GPT-2 and GPT-3. In Table 5 the characteristics of each version of the model are presented.

- GPT-2: This version contains 1.5 billion parameters [66]. The model is initially trained using data collected from web pages and subsequently the model is fitted with a custom dataset that is oriented to a particular task. This technique is called two-step training: pretraining and tuning.
- GPT-3: This model uses a different learning strategy to its predecessor, using prompts which learn from examples of NLP or NLG tasks. This model is trained with 400 billion tokens and has a maximum of 175 billion parameters [67].

Table 5. GPT Models.

Model	Parameters (Billions)	Decoder Layers	Context Token Size	Hidden Layer	Batch Size	Training Data
GPT	0.117	12	512	768	64	BookCrawl
GPT-2	1.5	48	1024	1600	512	WebText
GPT-3	175	96	2048	12,288	3.2M	CommonCrawl

The GPT model is used in several areas of knowledge, for example in [68] the authors present CodexDB which, through the use of the GPT-3 model, customises SQL query processing by translating text into SQL code. MacNeil et al. [69] perform an analysis of the scope of using natural language to automatically generate explanations from a given code fragment using GPT-3. In the area of medicine, the model has also been used and has shown good results, for instance in [70] a model for capturing relevant medical information and using GPT-3 to generate synthesised training data is presented. This model allows synthetically generated data to be combined with manually labelled data by humans, obtaining better results than models trained only with manually labelled data. On the other hand, in [71] a proposal is made for the detection of technical research texts manipulated by means of GPT-2, which poses a risk to genuine research which is undermined by synthetically generated texts.

4.3. XLNet

XLNet [60] is a transfer learning model introduced by Google AI in 2019, this model leverages regressional pretraining (AR) language modelling and autoencoding (AE) to overcome BERT's limitations of context capture. It allows the analysis of sentence contexts in a bidirectional way, by maximising the permutation probabilities of the factorisation order.

In order to minimise the limitations of BERT for capturing bidirectional contexts, XLNet introduces permutation language modelling (PLM). RA-based pretraining reconstructs the original data from a corrupted input, instead of performing an explicit density calculation. This allows for better performance by addressing the limitations of bidirectional information in AR language modelling. However, there are differences between the pretraining and fitting process, because the artificial token symbols used by BERT during training do not appear in the data when fitting is performed. It is good at linguistic tasks involving long contexts. Thanks to its autoregressive formulation, the model performs better than BERT on 20 tasks, such as sentiment analysis, document classification, natural language inference, and question answering [60]. XLNet has been used in several knowledge domains, e.g., for detecting news where Kumar et al. [72] present a refined XLNet model for predicting fake news in binary and multi-layer problems. The proposals used the LIAR dataset to detect fake news in social networks, and used two and six classes to perform the classification. Achieving 44% accuracy for the 6-class and 72% accuracy for the 2-class.

In [73], a method is proposed that makes use of a classifier to indicate where and how to refine candidate sentences for Markov chain Monte Carlo (MCMC)-based models. Two methods are proposed to generate synthetic samples to refine the pretrained model and a two-step approach to generate sentences with constraints, calculating the percentage reliability of the candidate sentence, for machine translation and generation of automatic responses. In addition, to improve the neural machine translation (NMT) model in [74], contextual feature extraction is performed on Chinese to Mongolian, Uygur, and Tibetan translation tasks using an XLNet-based pretraining method. Furthermore, in [75] a study is conducted to analyse the results of combining the closed recurrent unit (GRU) with XLNet (XLNet-GRU) for the detection of news generating misinformation about the Bitcoin and Ethereum cryptocurrencies, focusing on the English and Malay languages, using manually labelled datasets to understand the purpose of the sentence.

In relation to health areas, XLNet has had very good results, for example in [76] an unsupervised development for the analysis of tweets with the XLNet model, to see the trend of two vaccines, used transfer learning to classify the tweets. This technique outperformed techniques such as VADER, TextBlob, Bi-LSTM, and BERT for sentiment analysis based on social media updates. In addition, a model proposed to analyse drugs prewritten in discharge documents in a non-standardised format (AB-XLNet), which used training based on the random insert technique and the pretrained BERT model, improved the accuracy of XLNet by 3% and extracted ADE and form, which previously could not be extracted from XLNet.

4.4. DistilBERT

DistilBERT is a smaller and more efficient BERT-derived model, trained in the same way as the base model but under a self-supervised scheme [62]. DistilBERT was pretrained on three targets:

- Distillation loss: The model was trained to return the same training rate as the baseline BERT.
- Masked language modelling (MLM): It bases its learning on a bidirectional analysis of the sentence.
- Loss of cosine embedding: The training of the model generates hidden states as close as possible to the base BERT model.

The DistilBERT model is a model that has been distilled from the base BERT model, with the distillation technique the model has 40% fewer parameters and 60% less runtime, while retaining over 95% of the performance of BERT (as measured by the GLUE language comprehension benchmark).

In [77], a logistic regression algorithm combined with the BERT and DistilBERT algorithms is used to predict the time required for error correction based on LiveCode bug reports, to measure the effectiveness of BERT and DistilBERT under the same conditions. The results of the experiments show that DistilBERT retains almost the same language

understanding capabilities as BERT and in some conditions is 63.28% faster than BERT. Furthermore, by performing some modifications on its parameters, the logistic regression model, DistilBERT obtains a better accuracy value than BERT.

The DistilBERT model is notable for its speed in making inferences. This feature has been used by many papers. For example, in [78] it collects data from texts published on social networks and classifies comments that denote harassment of children through online comments. This work demonstrates the risks faced by children on the internet. The proposal achieves an improvement of about 3% accuracy over other algorithms addressing the same problem, demonstrating the accuracy of DistilBERT. In addition, in [79] they test a DistilBERT-based model to detect comments that denote online aggression and it is claimed that the use of information from multiple sources increases the performance and accuracy of the model. In the area of medicine, DistilBERT has also had good results. In Jojoa et al. [80], a method based on DistilBERT is proposed to detect positive or negative tendencies on responses in surveys during the COVID-19 pandemic, the work obtained outstanding results (F1 score of 80.3%) considering that large volumes of data are necessary to obtain good results. In the area of banking, government, and global news, in [81], a study on sentiment analysis using the DistilBERT model applied to news was carried out. This model was fitted and fed to four different classifiers. The results obtained by Dogra et al. indicate that the trained model can transfer semantic understanding to other domains, achieving a higher accuracy than the reference TF-IDF. The authors conclude that the random forest model combined with DistilBERT leads to a higher accuracy than other classification models, i.e., a comparison of TF-IDF achieves a 7.5% higher accuracy, of 78%.

For the classification of documents, TopicBERT is presented in [82], a proposal that allows the optimisation of computational resources necessary for fine tuning. This proposal bases its operation on the complementary learning of unified subject and linguistic models. The proposal has a reduction in the number of self-service operations, reaching a speed 1.4 times higher than previous proposals, obtaining a performance of 99.9% in 5 datasets.

On the other hand, DC-BERT is proposed in [83], which is based on two BERT models, the first one encodes the question in one occasion and the second one performs a precoding of the documents, storing its cache offline, in addition to using a decoupled contextual encoding framework. For the tests carried out by the authors, two datasets, SQuAD Open and Natural Questions Open, are used, in which the proposed model obtains a performance of up to 10 times faster in document retrieval, while retaining almost the same performance of the previously proprietary approaches for open domain questions. Likewise, in [84] a novel architecture is proposed by applying the concept of "long" attention to a distilled BERT model, focusing the solution on recognising legal domain phrases and contexts. This allows the model to recognise the context for longer text sequences by combining a window of local attention with task-driven global attention. The proposed model is faster and outperforms other models, such as BERT, for document classification in a legal context.

In Varun Dogra, et al. [85], a model for the classification of news and banking events is proposed, using a hybrid architecture using the context-independent language representation TFIDF. These representations are introduced in five classifiers, with the result that DistilBERT conveys domain-general knowledge better to others compared to the TFIDF baseline. Furthermore, it is concluded that the use of a hybrid model improves accuracy with classifiers such as random forest.

Quevedo et al. [86] make a proposal, based on the performance of several linguistic models that base their operation on transformers, for the development of a specialised question answering solution in the legal field, comparing the results obtained with other developments that focus their operation on question answering without a specific field. The PolicyQA dataset was used, by conforming it with documents related to users' data processing policies, which fall within the legal scope of the software. The proposal used ALBERT, BERT, DistilBERT, LEGAL-BERT, and RoBERTa as base encoders and compared their performance on the SQuAD V2.0 and PolicyQA question answering reference dataset. Fur-

thermore, it is shown that domain-general BERT-based models, such as ALBERT, perform better than a model trained for a more specific domain, such as LEGAL-BERT.

4.5. Text-to-Text Transfer Transformer

The Text-to-Text Transfer Transformer (T5) model is a technique where the input and output are always text strings, using a transformer architecture that replaces tasks such as removing spaces with a combination of alternative tasks [87]. Each task used is approached with the input model input and trained by generating a target text including tasks such as question answering, translation, and text classification, and is also used for language generation in different languages [88,89]. Among the changes that can be identified with respect to models such as BERT, are causal decoders for bidirectional textual revision.

For multi-modal emotion recognition tasks, in [90] a method is proposed that allows a representation of emotions through speech, audio, and text to be found. This proposal uses pretrained models such as TRILL and the single-modal text-to-text transform, which fit the emotion dataset. In multi-task transfer learning tasks, work has been presented for the analysis of multi-domain and multi-task learning behaviour using T5 (MD-T5) models [91], specifically in two domains, Python code and chess. This work reaffirms the challenges of using negative knowledge transfer and model forgetting, with good results using joint pretraining plus domain austerity for multi-domain and multi-task learning. Nagoudi et al. [92] propose a transformer-based model for natural language processing focused on indigenous languages, using the T5 model. The *IndCorpus* dataset of ten indigenous languages and Spanish was used to fit the model. This work demonstrates the good results of machine translation using various approaches for translation between American Indian languages and Spanish.

In [93], a proposal is made that by using a T5 pretrained model, performance improvements are achieved in processing tasks such as error correction in code, injecting mutations into code, or generating comments, in addition to exploiting additional data for the self-controlled pretraining phrase with respect to other proposals. In addition, ref. [94] uses the model for neural generation of questions and Phakmongkol and Vateekul [95] use T5 to answer questions. On the other hand, T5 is used to create a chat room, as in [96], for the regulation of emotions in chat rooms as a framework for integral dialogue.

4.6. ALBERT

In the work presented in [63], a BERT-based algorithm, ALBERT, is presented that improves the performance of twelve BERT natural language processing tasks, including answering questions from the SQuAD v2.0 dataset and the RACE SAT-style reading comprehension benchmark. To improve the performance of the model, capacity allocation is achieved in an efficient way. ALBERT achieves an 80% reduction in projection block parameters without sacrificing performance on the benchmark datasets. This is made possible by context-independent learning of embeddings such as words or subtokens.

4.7. RoBERTa

Developed by Facebook, RoBERTa, the robust and optimised BERT method, is a model based on BERT using an improved methodology for the training process, using 10 times the data used in BERT and improved computational power [59]. It is an optimised method for pretraining NLP systems. RoBERTa is a model in which the system learns to predict sections of hidden text within previously unlabelled linguistic examples using BERT's masking strategy [97]. For RoBERTa training, the BERT hyperparameters are modified and the training target of the next phrase and minibatch training is lined and the learning rates are modified. Meta's RoBERTa model has been used in relevant proposals, to identify patterns in text.

RoBERTa uses large amounts of text data for training, approximately 160 GB, including data from books and Wikipedia used by BERT. Sources supplementing the training data include news data (CommonCrawl news, 76 GB), text from websites (38 GB), and stories

(31 GB). The training of this data is performed over the course of a day using the latest generation graphics card. This model outperforms its predecessor BERT and XLNet in the GLUE benchmark data. An important point to consider about the RoBERTa model is that if an attempt was made to reduce the model to a smaller model, using techniques such as pouring, quantisation, or distillation, the model would have lower prediction metrics.

Currently there are many jobs that use RoBERTa to perform tasks in different areas. For example, in [98] the model is used for the analysis of English language tweets for the COVID-19 classification of adverse pregnancy outcomes and contingency cases, resulting in F1 scores of 93% and 75%, respectively. Furthermore, in the area of sentiment analysis and classification, in [99] ASK-RoBERTa is presented, a pretraining model that allows the prediction of sentiment inclinations from sentences or documents based on adaptive pretraining of the model to sentiment knowledge. In the paper, some rules are developed to perform term and sentiment mining based on grammar and parts of speech. ASK-RoBERTa performs better than previously proposed sentiment analysis models based on BERT and current deep learning models. On the other hand, a model that makes use of RoBERTa and interactive attention networks (IAN) for sentiment analysis for product reviews is proposed in [100], called RoBERTa-IAN. This work proposes the use of low-dimensional vectors to serve as input to a model for the extraction of semantic features of the text and obtaining the hidden representation.

In [101], a solution is presented that, by using RoBERTa and conditional random fields (CRF), is used to detect offensive language on a social media post. The model takes into account the existing differences with the general language, in addition to the jargon used, the proposed solution obtained a 66.34% F1 score. Further, in [102] RoBERTa is used to detect sarcastic comments in tweets posted in English. The proposed model is based on a model pretrained with Twitter data and uses a three-layer forward connected neural network. The model obtains an F1 score of 52.6%.

Furthermore, in [103] a model based on RoBERTa trained only on Czech language data is presented. This proposal improves on existing multilingual contextualised linguistic representation models, such as multilingual XLM-RoBERTa [104] and SlavicBERT [105], in NLP tasks such as tagging and lemmatisation, dependency parsing, and semantics.

In addition, some proposals allow the categorisation of large numbers of digital documents by automatically assigning unlabelled text documents to predefined categories, providing a solution to the growing demand to organise, store, and retrieve these documents accurately and efficiently. For example, in [106] EVI-IBLMM (extended variational inference for inverted Beta-Liouville mixture model) is proposed for text categorisation. Two datasets are used in the proposal: WebKB and 20Newsgroups, which have four and twenty categories, respectively. For the reduction of words to their simplest form, they make use of Porter's stemming. Experiments on 20 runs show that the proposal of X et al. has the best categorisation accuracy, 90.36% with the WebKB dataset and 81.11% on 20Newsgroup, among all mixture-based approaches for the text categorisation task (EVI-GIDMM, EVI-IDMM, EVI-GaMM).

This paper focuses its research on the RoBERTa model, because it provides great flexibility and can be adapted to a wide range of tasks such as text classification, entity recognition, or sentiment analysis. Furthermore, compared to BERT, RoBERTa has shown superior performance in various natural language processing tasks [59]. Furthermore, the pretraining process is improved compared to BERT, using unrestricted pretraining, which helps improve its ability to identify semantic and syntactic features. It also offers great flexibility. This model can be useful for resource-constrained teams that want to take advantage of pretrained machine learning without having to invest in training custom models from scratch.

Table 6 shows a comparison of the proposals that make use of the RoBERTa model. The table shows that the model can be tuned for a wide variety of approaches, such as medical, lexical, sentiment analysis, or marketing while maintaining good results.

Table 6. Overview of RoBERTa Proposals.

Proposal	Extraction Approach	Model Scope	Datasets	Results F1 Score
Adverse Pregnancy Outcomes and Potential COVID-19 [98]	Lexically constrained sentence generation	Medical	Medication Abuse in Tweets	0.9305
ASK-RoBERTa [99]	Sentiment analysis	Lexical	Restaurant-14-16, Laptop-14	0.779, 0.821, 0.71, 0.792
RoBERTa-IAN [100]	Product reviews	Marketing	SEMVAL	0.90
RoBERTa-CRF [101]	Toxic span detection	Sentiment analysis	Civil Comments	0.66
Sarcastic RoBERTa [102]	Detect sarcasm in tweets	Sentiment analysis	iSarcasm	0.526
RobeCzech [103]	Contextualised representation	Lexical	Czech Facebook dataset	0.801

Table 7 shows a summary of the works studied in this paper, mentioning the name of the proposal, the approach given to the model based on the training data, the scope, and the NLU model used for the proposal.

Table 7. Overview of Natural Language Understanding Models Proposals.

Proposal	Extraction Approach	Model Scope	Model	Datasets
CodexDB [68]	Translate text	SQL query processing	GPT-3	WikiSQL, SPIDER
GPT-3-ENS [70]	Medical dialogue summarisation	Medical	GPT-3	Human labeled, GPT-3-ENS
Cross-Domain Detection [71]	Cross-domain detection	Technical text	GPT-3	SME-labeled, Proxy data
Fake News Detection [72]	Detect fake news	News	XLNet	LIAR
Show Me How To Revise [73]	Sentence generation	Lexical	XLNet	One-Billion-Word
Low-resource neural machine translation [74]	Neural machine translation	Translation	XLNet	CCMT2019
XLNet-GRU [75]	Detect fake news	News	XLNet-GRU	Cryptocurrency news
Sentiment Analysis [76]	Tweet-based sentiment analysis	Medical	XLNet	COCO val
AB-XLNet [107]	Drugs in discharge summaries	Medical	XLNet	
AraT5 [88]	Language generation	Lexical	T5	Arabic MT datasets
mT5 [89]	Language generation	Lexical	T5	
Multimodal Emotion Recognition [90]	Emotion recognition	Sentiment analysis	T5	IEMOCAP
Extreme Multi-Domain [91]	Multi-task transfer learning	Lexical	T5	
Indt5 [92]	Natural language processing for indigenous languages	Lexical	T5	IndCorpus
Ensemble-NQG-T5 [94]	Neural generation of questions	Questions and answers	T5	SQuAD 2.0
Generated Questions for Thai [95]	Answer questions	Questions and answers	T5	Wiki QA and iApp Wiki QA
ER-Chat [96]	Chat room	Chat	T5	
Code-Related Tasks [93]	Correct code errors	Code-related	T5	Bug-Fix Pairs
Adverse Pregnancy Outcomes and Potential COVID-19 [98]	Lexical constrained sentence generation	Medical	RoBERTa	Medication Abuse in Tweets
ASK-RoBERTa [99]	Sentiment analysis	Lexical	RoBERTa	Restaurant-14, Restaurant-16 , Laptop-14
RoBERTa-IAN [100]	Product reviews	Marketing	RoBERTa	SEMVAL
RoBERTa-CRF [101]	Toxic span detection	Sentiment analysis	RoBERTa	Civil Comments
Sarcastic RoBERTa [102]	Detect sarcasm in tweets	Sentiment analysis	RoBERTa	iSarcasm
RobeCzech [103]	Contextualised representation	Lexical	RoBERTa	Czech Facebook dataset

5. Challenges

One of the challenges facing technology developers is to develop a solution that understands natural language. This is because modern languages are long and complex, containing a large number of phrases that can mean different things depending on the region of the world in which they are used. One way of tackling the problem is through syntax, as it allows for the analysis of the combination of sentences that could have different meanings [8].

Solutions that use NLP models in their logic to understand language are growing, but as demand grows, so do the challenges of the technology. This section lists some of the challenges faced when analysing the context of files on storage devices.

- Abbreviations: One of the most important challenges faced when analysing a document, regardless of language, is recognising words that may have multiple meanings or words that may be part of different sentences, and classifying similar words. Several words or sentences can be written in different ways, these can be abbreviated to facilitate writing, reading, and understanding. The same words can be written in long forms, e.g., BTW means By the way, and in many cases abbreviations may coincide with an organisation, a place, a position, or a position title.
- Errors related to speed and text: Models that rely on semantics cannot be trained if the speech and text data are wrong. This problem is analogous to the implication of misused or even misspelled words, which allow the model to learn over time. Although evolved grammar correction tools are good enough to remove sentence-specific errors, the training data must be error-free to facilitate accurate development in the first place.
- Spelling Variation: The vowels in the English language are very important. These letters do not make a big difference when heard, but they do make a big difference in spelling. Everybody makes spelling mistakes, but for the majority of us, we can gauge what the word was actually meant to be. However, this is a major challenge for computers, as they do not have the same ability to infer what the word was actually meant to be.
- Foreign Words: Words that are currently not used or heard very often are an area of interest in this field. Such words include names of people, place names, or ancient organisations.
- Different types of text: Sometimes, when analysing a document, it is difficult to relate two texts of different subject matter, for example, it is difficult to relate a judicial text to a common text, given that the words used to refer to the same thing may vary.
- Synonyms: Some phrases or words can have exactly the same meanings at different grammatical levels and with the same grammatical category. In any language, people often use synonyms to denote slightly different meanings within their vocabulary without changing the meaning of the sentence. Small and little can be synonyms when automatically parsing a sentence to denote the same meaning, but they are not interchangeable in all contexts, as one could denote only size and the other could denote both size and emotion. For example, buy a small cup of coffee for the office and even the small change can make a difference, in this example the words are not interchangeable.
- Colloquialisms: In every culture, phrases, expressions, and idiomatic jargon are used that have specific meanings, posing a problem for NLP developments. These expressions may exist in more than one culture, yet have different meanings in different geographical regions or simply have no coherent meaning. Even if NLP services try to scale beyond ambiguities, errors, and homonyms, it is not easy to include specific words that have different connotations from one culture to another. There are words that lack a definition in the language, but may still be relevant to a specific audience. It is important to include relevant references so that the resource is sufficiently perceptive.
- Sarcasm: Generally used words and phrases that can be positive or negative, but in reality connote the opposite. When expressing a sentence, the intention with which it is intended to be transmitted, the emotions that were present when creating it, and the personality of the author or speaker can influence it. Some of them, such as irony and sarcasm, can make a sentence be taken as positive, however, the emotions of the author can go in an opposite sense to the literal one. Although sentiment analysis has now made advances, the correct extraction of context when confronted with sarcastic sentences remains a research challenge.

- Disambiguation of the meaning of words: To perform a correct sentence disambiguation process, it is necessary to understand the context in which it was written. For this, it is necessary to extract the meaning of the word by taking into account the adjacent words, because they have related meanings.

6. Conclusions

In this work we carry out a bibliographical review of some proposals that make use of machine learning methods and algorithms, specifically natural language processing models, for the processing, comprehension, and classification of texts contained in removable devices in a forensic process, in order to find evidence about a specific case.

The identification of named entities is a major problem for natural language processing, since valuable information must be extracted from texts. However, this process can have various difficulties in identifying and detecting the context of a phrase such as abbreviations, spelling and grammatical errors, the use of synonyms, colloquialisms used only in certain cultures, or the use of sarcastic phrases.

The algorithms that have shown great efficiency in entity detection are those based on transformers, specifically the BERT algorithm, surpassing previous models in precision and performance. BERT allows the processing of texts in several languages, which has made it possible to generate versions for specific languages or purposes to cover a need. In addition, by performing some distillation techniques, the BERT base model can be used in low-resource devices, because the size is reduced and performance is improved. In this case, multiple jobs sacrifice processing time to improve prediction accuracy. One of the important technical aspects of BERT is that it can check the context of a sentence in a bidirectional way, which allows a better understanding of the background of the text and correct entity recognition. Although most of the algorithms allow the processing of sentences, they can be tuned to allow the processing of documents and the search for relationships between them.

Therefore, the use of natural language processing algorithms for the classification and identification of relationships between documents in a forensic analysis process, can provide advantages such as greater precision in the extraction and identification of evidence, identifying patterns in the texts and data linkage, reduction in human errors, and automation of repetitive tasks, which would allow researchers to focus on more complex tasks.

In general, the use of machine learning models in a forensic process can improve the efficiency, accuracy, and reliability of evidence analysis, which can help investigators to reach stronger and more accurate conclusions.

Author Contributions: Conceptualization, L.A.M.H., A.L.S.O. and L.J.G.V.; methodology, L.A.M.H., A.L.S.O. and L.J.G.V.; validation, L.A.M.H., A.L.S.O. and L.J.G.V.; investigation, L.A.M.H., A.L.S.O. and L.J.G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work also was supported by the European Commission under the Horizon 2020 research and innovation programme, as part of the project HEROES (Grant Agreement no. 101021801). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission—EU. Neither the European Union nor the European Commission can be held responsible for them.

Data Availability Statement: Not applicable; this study does not report any data.

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

References

1. Statista. Annual Number of Suspected and Arrested Individuals for Cybercrimes in Spain from 2011 to 2019. Available online: https://www.statista.com/statistics/1173433/cybercrime-number-of-detained-and-investigated-spain/ (accessed on 6 January 2023).
2. Noblett, M.; Pollitt, M.; Presley, L. Recovering and Examining Computer Forensic Evidence. Available online: https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/oct2000/computer.htm (accessed on 6 January 2023).
3. Raghavan, S. Digital forensic research: Current state of the art. *CSI Trans. ICT* **2013**, *1*, 91–114. [CrossRef]

4. Patel, J. Forensic Investigation Life Cycle (FILC) using 6'R' Policy for Digital Evidence Collection and Legal Prosecution. *Int. J. Emerg. Trends Technol. Comput. Sci.* **2013**, *1*, 129–132.
5. Ćosić, J.; Cosic, J.; Cosic, Z. Chain of Custody and Life Cycle of Digital Evidence. *Comput. Technol. Appl.* **2012**, *3*, 126–129.
6. Agarwal, A.; Agarwal, A.; Gupta, S.; Gupta, S.C. Systematic Digital Forensic Investigation Model. *Int. J. Comput. Sci. Secur.* **2011**, *5*, 118–131.
7. Amato, F.; Cozzolino, G.; Giacalone, M.; Moscato, F.; Romeo, F.; Xhafa, F. A Hybrid Approach for Document Analysis in Digital Forensic Domain. In Proceedings of the International Conference on Emerging Internetworking, Data & Web Technologies, EIDWT 2019, Fujairah Campus, United Arab Emirates, 26–28 February 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 170–179.
8. Chowdhary, K.R. Natural Language Processing. In *Fundamentals of Artificial Intelligence*; Springer: New Delhi, India, 2020; pp. 603–649. [CrossRef]
9. Meurers, D. Natural language processing and language learning. *Encycl. Appl. Linguist.* **2012**, *9*, 4193–4205.
10. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
11. Wu, C.; Wu, F.; Qi, T.; Huang, Y. Named Entity Recognition with Context-Aware Dictionary Knowledge. In Proceedings of the Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, 30 October 30–1 November 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 129–143.
12. Mollá, D.; Van Zaanen, M.; Smith, D. Named entity recognition for question answering. *Proc. Australas. Lang. Technol. Workshop* **2006**, *2006*, 51–58.
13. Modrzejewski, M.; Exel, M.; Buschbeck, B.; Ha, T.L.; Waibel, A. Incorporating external annotation to improve named entity translation in NMT. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 3–5 November 2020; pp. 45–51.
14. Dill, S.; Eiron, N.; Gibson, D.; Gruhl, D.; Guha, R.; Jhingran, A.; Kanungo, T.; Rajagopalan, S.; Tomkins, A.; Tomlin, J.A.; et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*; Association for Computing Machinery: New York, NY, USA, 2003; pp. 178–186. [CrossRef]
15. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv* **2003**, arXiv:cs/0306050.
16. Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. OntoNotes Release 5.0. Available online: https://catalog.ldc.upenn.edu/LDC2013T19 (accessed on 6 January 2023).
17. Doğan, R.I.; Leaman, R.; Lu, Z. The NCBI Disease Corpus. Available online: https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/ (accessed on 6 January 2023).
18. Derczynski, L.; Nichols, E.; van Erp, M.; Limsopatham, N. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, 7 September 2017; pp. 140–147.
19. Tsujii, J. GENIA Corpus. Available online: http://www.geniaproject.org/home (accessed on 6 January 2023).
20. Reddy, S.; Biswal, P. IIITBH at WNUT-2020 Task 2: Exploiting the best of both worlds. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*; Association for Computational Linguistics: Online, 2020; pp. 342–346. [CrossRef]
21. Luz de Araujo, P.H.; de Campos, T.E.; de Oliveira, R.R.R.; Stauffer, M.; Couto, S.; Bermejo, P. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*; Lecture Notes on Computer Science (LNCS); Springer: Canela, Brazil, 2018; pp. 313–323. [CrossRef]
22. Ling, X.; Weld, D.S. Fine-grained entity recognition. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
23. Tedeschi, S.; Maiorca, V.; Campolungo, N.; Cecconi, F.; Navigli, R. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 2521–2533.
24. Derczynski, L.; Bontcheva, K.; Roberts, I. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 1169–1179.
25. Project, G. Gate. Available online: https://gate.ac.uk/ (accessed on 6 January 2023).
26. NLTK. Natural Language Toolkit. Available online: https://www.nltk.org/ (accessed on 6 January 2023).
27. Group, T.S.N.L.P. Stanford Named Entity Recognizer (NER). Available online: https://nlp.stanford.edu/software/CRF-NER.shtml (accessed on 6 January 2023).
28. Spacy. Entity Recognizer. Available online: https://spacy.io/api/entityrecognizer (accessed on 6 January 2023).
29. Poliglot. Named Entity Extraction. Available online: https://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html (accessed on 6 January 2023).
30. flairNLP. Flair. Available online: https://github.com/flairNLP/flair (accessed on 6 January 2023).
31. DeepPavlov. Named Entity Recognition (NER). Available online: https://docs.deeppavlov.ai/en/0.0.8/components/ner.html (accessed on 6 January 2023).

32. AllenNLP. Named Entity Recognition (NER). Available online: https://allenai.org/allennlp (accessed on 6 January 2023).
33. Project, G. ANNIE: A Nearly-New Information Extraction System. Available online: https://gate.ac.uk/sale/tao/splitch6.html#x9-1200006.1 (accessed on 6 January 2023).
34. Claro, D.B.; Souza, M.; Castellã Xavier, C.; Oliveira, L. Multilingual Open Information Extraction: Challenges and Opportunities. *Information* **2019**, *10*, 228. [CrossRef]
35. Huang, H.; Wong, R.K.; Du, B.; Han, H.J. Weakly-Supervised Relation Extraction in Legal Knowledge Bases. In *Digital Libraries at the Crossroads of Digital Information for the Future*; Jatowt, A., Maeda, A., Syn, S.Y., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 263–270.
36. Gao, T.; Han, X.; Qiu, K.; Bai, Y.; Xie, Z.; Lin, Y.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction. *arXiv* **2021**, arXiv:2105.09543.
37. Shao, C.; Li, M.; Li, G.; Zhou, M.; Han, D. CRSAtt: By Capturing Relational Span and Using Attention for Relation Classification. *Appl. Sci.* **2022**, *12*, 1068. [CrossRef]
38. Chen, X.; Huang, X. EANT: Distant Supervision for Relation Extraction with Entity Attributes via Negative Training. *Appl. Sci.* **2022**, *12*, 8821. [CrossRef]
39. Lange Di Cesare, K.; Zouaq, A.; Gagnon, M.; Jean-Louis, L. A Machine Learning Filter for the Slot Filling Task. *Information* **2018**, *9*, 133. [CrossRef]
40. Wang, W.; Hu, W. Improving Relation Extraction by Multi-Task Learning. In *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence (HPCCT & BDAI '20)*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 152–157. [CrossRef]
41. Sahu, S.K.; Anand, A.; Oruganty, K.; Gattu, M. Relation extraction from clinical texts using domain invariant convolutional neural network. *arXiv* **2016**, arXiv:1606.09370.
42. Genest, P.Y.; Portier, P.E.; Egyed-Zsigmond, E.; Goix, L.W. PromptORE—A Novel Approach Towards Fully Unsupervised Relation Extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 561–571. [CrossRef]
43. Ali, M.; Saleem, M.; Ngomo, A.C.N. Unsupervised Relation Extraction Using Sentence Encoding. In *The Semantic Web: ESWC 2021 Satellite Events*; Verborgh, R., Dimou, A., Hogan, A., d'Amato, C., Tiddi, I., Bröring, A., Mayer, S., Ongenae, F., Tommasini, R., Alam, M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 136–140.
44. De Lacalle, O.L.; Lapata, M. Unsupervised relation extraction with general domain knowledge. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing: Seattle, Washington, DC, USA, 18–21 October 2013; pp. 415–425.
45. Fu, T.J.; Li, P.H.; Ma, W.Y. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1409–1418. [CrossRef]
46. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv* **2019**, arXiv:1904.05255.
47. Christopoulou, F.; Miwa, M.; Ananiadou, S. Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs. *arXiv* **2019**, arXiv:1909.00228.
48. Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; Sun, M. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. *arXiv* **2019**, arXiv:1906.06127.
49. Semaan, P. Natural language generation: An overview. *J. Comput. Sci. Res.* **2012**, *1*, 50–57.
50. Petrović, D.; Stankovic, M. Use of linguistic forms mining in the link analysis of legal documents. *Comput. Sci. Inf. Syst.* **2018**, *15*, 5. [CrossRef]
51. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T. Incorporating BERT into Neural Machine Translation. *arXiv* **2020**, arXiv:2002.06823.
52. Kudande, D.; Dolai, P.; Hole, A. Fake News Detection & Sentiment Analysis on Twitter Data Using NLP. *Int. Res. J. Eng. Technol. (IRJET)*, **2021**, *8*, 1571–1574.
53. Tan, K.L.; Lee, C.P.; Anbananthen, K.S.M.; Lim, K.M. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access* **2022**, *10*, 21517–21525. [CrossRef]
54. Mishakova, A.; Portet, F.; Desot, T.; Vacher, M. Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes. In Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kyoto, Japan, 11–15 March 2019; pp. 832–837. [CrossRef]
55. Zhao, H.; Cao, J.; Xu, M.; Lu, J. Variational neural decoder for abstractive text summarization. *Comput. Sci. Inf. Syst.* **2020**, *17*, 537–552. [CrossRef]
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
57. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 604–624. [CrossRef] [PubMed]
58. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

59. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
60. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
61. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 6 January 2023).
62. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
63. Soricut, R.; Lan, Z. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. Available online: https://ai.googleblog.com/2019/12/albert-lite-bert-for-self-supervised.html (accessed on 6 January 2023).
64. Roberts, A.; Raffel, C.; Lee, K.; Matena, M.; Shazeer, N.; Liu, P.J.; Narang, S.; Li, W.; Zhou, Y. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*; Technical Report; Google. 2019. Available online: https://research.google/pubs/pub48643/ (accessed on 18 January 2023).
65. Jwa, H.; Oh, D.; Park, K.; Kang, J.M.; Lim, H. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Appl. Sci.* **2019**, *9*, 4062. [CrossRef]
66. Solaiman, I.; Clark, J.; Brundage, M. GPT-2: 1.5B Release. Available online: https://openai.com/blog/gpt-2-1-5b-release/ (accessed on 6 January 2023).
67. OPENAI. Models: Overview. Available online: https://beta.openai.com/docs/models/overview (accessed on 6 January 2023).
68. Trummer, I. CodexDB: Synthesizing Code for Query Processing from Natural Language Instructions Using GPT-3 Codex. *Proc. VLDB Endow.* **2022**, *15*, 2921–2928. [CrossRef]
69. MacNeil, S.; Tran, A.; Mogil, D.; Bernstein, S.; Ross, E.; Huang, Z. Generating Diverse Code Explanations Using the GPT-3 Large Language Model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research—Volume 2 (ICER '22)*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 37–39. [CrossRef]
70. Chintagunta, B.; Katariya, N.; Amatriain, X.; Kannan, A. Medically aware gpt-3 as a data generator for medical dialogue summarization. In Proceedings of the Machine Learning for Healthcare Conference (PMLR), Online, 6–7 August 2021; pp. 354–372.
71. Rodriguez, J.; Hay, T.; Gros, D.; Shamsi, Z.; Srinivasan, R. Cross-Domain Detection of GPT-2-Generated Technical Text. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics. Seattle, WD, USA, 10–15 July 2022; pp. 1213–1233. [CrossRef]
72. Kumar J, A.; Esther Trueman, T.; Cambria, E. Fake News Detection Using XLNet Fine-Tuning Model. In Proceedings of the 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 18–19 June 2021; pp. 1–4. [CrossRef]
73. He, X.; Li, V.O. Show Me How To Revise: Improving Lexically Constrained Sentence Generation with XLNet. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 12989–12997. [CrossRef]
74. Wu, N.; Hou, H.; Guo, Z.; Zheng, W. Low-Resource Neural Machine Translation Using XLNet Pre-training Model. In *Artificial Neural Networks and Machine Learning—ICANN 2021*; Farkaš, I., Masulli, P., Otte, S., Wermter, S., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 503–514.
75. Mohamad Zamani, N.A.; Liew, J.S.Y.; Yusof, A.M. XLNET-GRU Sentiment Regression Model for Cryptocurrency News in English and Malay. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*; European Language Resources Association: Marseille, France, 2022; pp. 36–42.
76. Bansal, A.; Susan, S.; Choudhry, A.; Sharma, A. Covid-19 Vaccine Sentiment Analysis During Second Wave in India by Transfer Learning Using XLNet. In *Pattern Recognition and Artificial Intelligence*; El Yacoubi, M., Granger, E., Yuen, P.C., Pal, U., Vincent, N., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 443–454.
77. Ardimento, P. Predicting Bug-Fixing Time: DistilBERT Versus Google BERT. In *Product-Focused Software Process Improvement*; Taibi, D., Kuhrmann, M., Mikkonen, T., Klünder, J., Abrahamsson, P., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 610–620.
78. Saha, U.; Mahmud, M.S.; Keya, M.; Lucky, E.A.E.; Khushbu, S.A.; Noori, S.R.H.; Syed, M.M. Exploring Public Attitude Towards Children by Leveraging Emoji to Track Out Sentiment Using Distil-BERT a Fine-Tuned Model. In *Third International Conference on Image Processing and Capsule Networks*; Chen, J.I.Z., Tavares, J.M.R.S., Shi, F., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 332–346.
79. Palliser-Sans, R.; Rial-Farràs, A. HLE-UPC at SemEval-2021 Task 5: Multi-Depth DistilBERT for Toxic Spans Detection. *arXiv* **2021**, arXiv:2104.00639.
80. Jojoa, M.; Eftekhar, P.; Nowrouzi-Kia, B.; Garcia-Zapirain, B. Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. *AI Soc.* **2022**, 1–8. [CrossRef]
81. Dogra, V.; Singh, A.; Verma, S.; Kavita; Jhanjhi, N.Z.; Talib, M.N. Analyzing DistilBERT for Sentiment Classification of Banking Financial News. In *Intelligent Computing and Innovation on Data Science*; Peng, S.L., Hsieh, S.Y., Gopalakrishnan, S., Duraisamy, B., Eds.; Springer Nature Singapore: Singapore, 2021; pp. 501–510.
82. Chaudhary, Y.; Gupta, P.; Saxena, K.; Kulkarni, V.; Runkler, T.A.; Schütze, H. TopicBERT for Energy Efficient Document Classification. *arXiv* **2020**, arXiv:2010.16407.

83. Nie, P.; Zhang, Y.; Geng, X.; Ramamurthy, A.; Song, L.; Jiang, D. DC-BERT: Decoupling Question and Document for Efficient Contextual Encoding. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1829–1832. [CrossRef]
84. Bambroo, P.; Awasthi, A. LegalDB: Long DistilBERT for Legal Document Classification. In Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19–20 February 2021; pp. 1–4. [CrossRef]
85. Dogra, V. Banking news-events representation and classification with a novel hybrid model using DistilBERT and rule-based features. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 3039–3054.
86. Caballero, E.Q.; Rahman, M.S.; Cerny, T.; Rivas, P.; Bejarano, G. Study of Question Answering on Legal Software Document using BERT based models. *Latinx Nat. Lang. Process. Res. Work.* **2022**.
87. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
88. Elmadany, A.; Abdul-Mageed, M. AraT5: Text-to-Text Transformers for Arabic Language Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 628–647.
89. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
90. Deng, J.J.; Leung, C.H.C.; Li, Y. Multimodal Emotion Recognition Using Transfer Learning on Audio and Text Data. In *Computational Science and Its Applications—ICCSA 2021*; Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 552–563.
91. Oshingbesan, A.; Ekoh, C.; Atakpa, G.; Byaruagaba, Y. Extreme Multi-Domain, Multi-Task Learning With Unified Text-to-Text Transfer Transformers. *arXiv* **2022**, arXiv:2209.10106.
92. Nagoudi, E.M.B.; Chen, W.R.; Abdul-Mageed, M.; Cavusogl, H. Indt5: A text-to-text transformer for 10 indigenous languages. *arXiv* **2021**, arXiv:2104.07483.
93. Mastropaolo, A.; Scalabrino, S.; Cooper, N.; Nader Palacio, D.; Poshyvanyk, D.; Oliveto, R.; Bavota, G. Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Madrid, Spain, 22–30 May 2021; pp. 336–347. [CrossRef]
94. Hwang, M.H.; Shin, J.; Seo, H.; Im, J.S.; Cho, H.; Lee, C.K. Ensemble-NQG-T5: Ensemble Neural Question Generation Model Based on Text-to-Text Transfer Transformer. *Appl. Sci.* **2023**, *13*, 903. [CrossRef]
95. Phakmongkol, P.; Vateekul, P. Enhance Text-to-Text Transfer Transformer with Generated Questions for Thai Question Answering. *Appl. Sci.* **2021**, *11*, 267. [CrossRef]
96. Katayama, S.; Aoki, S.; Yonezawa, T.; Okoshi, T.; Nakazawa, J.; Kawaguchi, N. ER-Chat: A Text-to-Text Open-Domain Dialogue Framework for Emotion Regulation. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2229–2237. [CrossRef]
97. AI, M. RoBERTa: An Optimized Method for Pretraining Self-Supervised NLP Systems. Available online: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/ (accessed on 6 January 2023).
98. Lee, L.H.; Hung, M.C.; Lu, C.H.; Chen, C.H.; Lee, P.L.; Shyu, K.K. Classification of Tweets Self-reporting Adverse Pregnancy Outcomes and Potential COVID-19 Cases Using RoBERTa Transformers. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*; Association for Computational Linguistics: Mexico City, Mexico, 2021; pp. 98–101. [CrossRef]
99. You, L.; Han, F.; Peng, J.; Jin, H.; Claramunt, C. ASK-RoBERTa: A pretraining model for aspect-based sentiment classification via sentiment knowledge mining. *Knowl.-Based Syst.* **2022**, *253*, 109511. [CrossRef]
100. Dai, J.; Pan, F.; Shou, Z.; Zhang, H. RoBERTa-IAN for aspect-level sentiment analysis of product reviews. *J. Phys. Conf. Ser.* **2021**, *1827*, 012079. [CrossRef]
101. Suman, T.A.; Jain, A. AStarTwice at SemEval-2021 Task 5: Toxic Span Detection Using RoBERTa-CRF, Domain Specific Pre-Training and Self-Training. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics. Online, Bangkok, Thailand, 5–6 August 2021; pp. 875–880. [CrossRef]
102. Hercog, M.; Jaroński, P.; Kolanowski, J.; Mieczyński, P.; Wiśniewski, D.; Potoniec, J. Sarcastic RoBERTa: A RoBERTa-Based Deep Neural Network Detecting Sarcasm on Twitter. In *Big Data Analytics and Knowledge Discovery*; Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 46–52.
103. Straka, M.; Náplava, J.; Straková, J.; Samuel, D. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*; Ekštein, K., Pártl, F., Konopík, M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 197–209.
104. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.
105. Arkhipov, M.; Trofimova, M.; Kuratov, Y.; Sorokin, A. Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy, 2 August 2019; pp. 89–93.

106. Ling, Y.; Guan, W.; Ruan, Q.; Song, H.; Lai, Y. Variational Learning for the Inverted Beta-Liouville Mixture Model and Its Application to Text Categorization. *arXiv* **2021**, arXiv:2112.14375.
107. Oh, K.; Kang, M.; Oh, S.; Kim, D.H.; Kang, S.H.; Lee, Y. AB-XLNet: Named Entity Recognition Tool for Health Information Technology Standardization. In Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 19–21 October 2022; pp. 742–744. [CrossRef]