mathematics

MDPI

Article

Finite State Automata on Multi-Word Units for Efficient Text-Mining [†]

Alberto Postiglione

Department of Business Science and Management & Innovation Systems, University of Salerno,
Via San Giovanni Paolo II, 84084 Fisciano, Italy; ap@unisa.it

[†] This paper is an extended version of our paper published in Advances in Internet, Data & Web Technologies—The 12th International Conference on Emerging Internet, Data & Web Technologies (EIDWT-2024), Naples, Italy, 21–23 February 2024.

Abstract: Text mining is crucial for analyzing unstructured and semi-structured textual documents. This paper introduces a fast and precise text mining method based on a finite automaton to extract knowledge domains. Unlike simple words, multi-word units (such as credit card) are emphasized for their efficiency in identifying specific semantic areas due to their predominantly monosemic nature, their limited number and their distinctiveness. The method focuses on identifying multi-word units within terminological ontologies, where each multi-word unit is associated with a sub-domain of ontology knowledge. The algorithm, designed to handle the challenges posed by very long multi-word units composed of a variable number of simple words, integrates user-selected ontologies into a single finite automaton during a fast pre-processing step. At runtime, the automaton reads input text character by character, efficiently locating multi-word units even if they overlap. This approach is efficient for both short and long documents, requiring no prior training. Ontologies can be updated without additional computational costs. An early system prototype, tested on 100 short and medium-length documents, recognized the knowledge domains for the vast majority of texts (over 90%) analyzed. The authors suggest that this method could be a valuable semantic-based knowledge domain extraction technique in unstructured documents.

Keywords: text mining; knowledge extraction; finite automata; ontology; multi-word units; natural language processing

MSC: 68T50

check for updates

Citation: Postiglione, A. Finite State Automata on Multi-Word Units for Efficient Text-Mining. *Mathematics* **2024**, *12*, 506. https://doi.org/10.3390/math12040506

Academic Editor: Chengjie Sun

Received: 29 December 2023
Revised: 31 January 2024
Accepted: 4 February 2024
Published: 6 February 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. The Growth of Natural Language Textual Documents

In this paper, we focus on natural language textual documents, whose format is typically unstructured (unstructured text data refer to text that is not organized in a predefined way, such as natural language free-form text documents, emails, social media posts, articles, etc.) and *semi-structured* (semi-structured textual data have some organization but lack sufficient structure to be stored in a relational database. Examples include XML, JSON, and HTML files). Until recently, the number of textual documents has remained relatively small, allowing effective management through classical algorithmic methods. However, in recent times, there has been a significant surge in the volume of textual data, encompassing books, news, research articles, web pages, emails, and social media posts (chat, blog, forum), among other sources. This surge can be attributed to several concurrent factors: the widespread diffusion of Big Data [1–6], the growth of corporate data, and the widespread adoption of the Internet of Things (IoT) [7–11] and IIoT (Industrial Internet of Things) [12–15]. This surge is further fueled by the rapid expansion of digital text documents, influenced by global government policies promoting the dematerialization of paper documents, the increasingly widespread use of interaction tools such as chatbots [16–20] and, of course, the growing pervasiveness of social media in the daily lives of billions of people.

As a consequence, a significant number of documents exist in unstructured or semi-structured formats, with only a small percentage being structured and/or numeric data. According to estimates in [21], at least 80% of real-world data comprises unstructured or semi-structured natural language textual documents. Similarly, in [22], it is stated that over 85% of current data are unstructured or semi-structured natural language textual data. Regardless of the exact percentage, the volume of natural language textual data generated has now taken on the consistency of Big Data: the data are enormous, heterogeneous, poorly typed, and continuously generated.

The sheer volume of data generated surpasses the capabilities of traditional databases and algorithmic methods, rendering these methods inadequate for efficient manipulation, management, and processing of such data. This challenge is further exacerbated by the necessity to process these data quickly, if not instantaneously. Consequently, there is a demand for robust data management systems capable of handling these large datasets and conducting real-time analytics.

1.2. Text Mining

Text mining [21–27], also referred to as text data mining or text analytics, is an interdisciplinary field that integrates various technologies, including natural language processing (NLP), machine learning, computer science, and statistical methods. The primary objective of text mining is to automatically extract knowledge and meaningful information from unstructured or semi-structured natural language text documents. This process may encompass the identification of entities (such as names and locations) and/or meaningful patterns, leading to the revelation of new insights, the discovery of hidden relationships between entities, as well as sentiment analysis and categorization. It is important to note that text mining represents a specialized branch of data mining [22,28,29]. While data mining encompasses a broad range of techniques applied to diverse types of data, text mining specifically focuses on the analysis of unstructured or semi-structured textual documents. Despite being easily manipulated by humans, unstructured text poses significant complexity for computer programs to handle effectively.

Text mining can be applied to a wide spectrum of domains, including industry, academia, medicine, the web, communication, and more [23,24,29,30]. Its utilization extends to crucial areas such as search engines, customer relationship management systems, email and document filtering, product suggestion analysis, fraud detection, and social media analytics. In these areas, text mining is employed for a range of tasks, such as opinion mining and sentiment analysis, feature extraction and document classification, trend analysis, topic modeling, summarization, and named entity recognition.

Text annotation in natural language processing (NLP) refers to the process of adding relevant information or labels to text data. The goal is to enhance the understanding of the text by providing additional context, structure, or meaning. Text annotation can be performed manually by human annotators, or it can be automated using certain tools or pre-existing datasets. The quality of text annotation is critical. The creation of annotated datasets often requires domain expertise to ensure accurate and meaningful labeling. One noteworthy application in the realm of research involves the classification of short scientific texts, typically comprising a title, a brief abstract, and keywords. Text annotation techniques are frequently employed for this purpose [31,32]. The medical field also benefits from text annotation, as evident in studies such as [33–35].

Beyond research and medicine, text mining plays a pivotal role in various fields, including risk management [36], predictive maintenance of industrial machines [37], financial domains [38], finance [39], service management [40], policy-making [41], healthcare [42], bio-medicine [43], COVID-19 literature [44], psychiatry [45], agriculture [46], social networks [47], social media [48], education and training [49], among others. The versatility of text mining makes it a valuable tool for extracting insights and patterns from textual data in diverse and evolving fields.

In [21], a text mining system is categorized into pre-processing, text representation, and four operational phases:

1. Pre-processing,
2. Text Representation,
3. Dimensionality Reduction,
4. Features Extraction,
5. Document Classification,
6. Evaluation.

During the initial phases (Phases 1 and 2), the primary objective is to convert unstructured text into structured features, with the aim of standardizing the input text to a uniform format. This process involves the cleaning and preparation of raw text data to make it suitable for analysis. Tasks within this phase may include the removal of irrelevant or unnecessary characters and words, handling special characters, converting text to lowercase, and tokenization—breaking the text into individual words or phrases. Figure 1 in [26] visually depicts the interactions among the four operational phases (Phases 3–6), offering a comprehensive illustration of the subsequent stages of the text mining process.

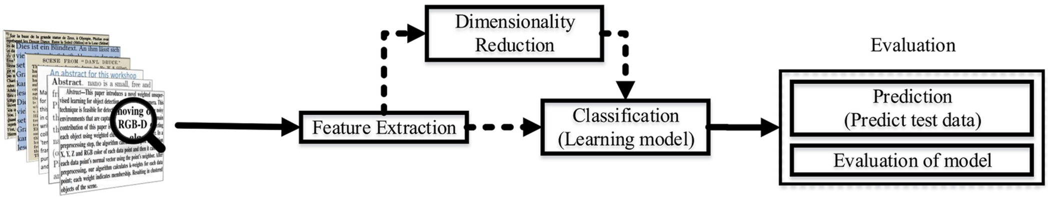


Figure 1. Text Mining system phases (from [26]).

1.3. Brief Description of the Proposal

This paper represents an extended version of [50], introducing AUTOMETA, a Linguistic-Based Text Mining system employing a finite automaton grounded in the concept of a "*unit of meaning*". AUTOMETA specializes in identifying "*multi-word units*", small portions of text with high semantic content. These units facilitate the automatic determination of the main knowledge domains within natural language text. Knowledge identification in text mining involves extracting relevant information from unstructured text data.

Our proposal has extensive applications, spanning diverse industries and providing valuable insights from large volumes of unstructured textual data. Key applications include Information Retrieval, Document Summarization, Question Answering Systems, Sentiment Analysis, Entity Recognition, Topic Modeling, Information Extraction, Clinical Text Mining in Healthcare, Business Intelligence, Legal Text Analysis, and more.

The algorithm is designed to address challenges presented by very long multi-word units composed of a variable number of simple words. During pre-processing, it integrates user-selected Multi-Word Unit ontologies into a single finite automaton. At runtime, the Finite Automaton reads the input text character by character, concurrently identifying all multi-word units from the selected ontologies, even if they partially or completely overlap. Ontologies can be updated without incurring additional computational costs.

This approach demonstrates efficiency for both short and long documents and does not require pre-training. A prototype of the system underwent testing on 100 input documents. It interacted with a union of one generic multi-word unit ontology and 22 terminological multi-word unit ontologies, totaling 21,396 entries. The system achieved knowledge domain identification in 91% of cases. It is configured as stand-alone software, making it embeddable in websites and portals for online use.

AUTOMETA is the product of decades of research in computer science and computational linguistics. The scientific basis can be traced back to studies on Lexicon–Grammar

(LG) [51–54] and on string matching through finite automata [55–58]. AUTOMETA represents a profound evolution compared to its predecessor, CATALOGA [59–62], as in 2021 we undertook a complete redefinition, redesign and rewriting of the system. It maintains the general approach of its predecessor and some ontologies (which we intend to review anyway), but every other aspect underwent a thorough redesign and rewriting process. This includes implementations of technology, algorithms and data structures, source code and interface.

1.4. Structure of the Paper

This paper is organized as follows: in Section 2, we provide an overview of the two primary Large Pre-trained Language Models, namely BERT and GPT; in Section 3, we describe the distinctive features of the system; in Section 4, we provide an overview of the algorithmic performances of the system; in Section 5, we illustrate the system performances on real data; in Section 6, we present a brief description of a case study designed to test the system on a set of documents; in Section 7, we delve into the primary challenges associated with the automatic processing of natural language texts; in Section 8, we compare the performance of the system with other existing methods; in Section 9, we provide suggestions for potential future research and draw overall conclusions.

2. Large Language Models: A Recent Milestone in NLP

In recent years, the field of Natural Language Processing has witnessed significant advancements, with Large Language Models (LLMs) [63], particularly Large Pre-trained Language Models (PLMs) [64], playing a pivotal role in reshaping the NLP landscape.

2.1. Large Pre-Trained Language Models (PLM)

PLMs are designed to comprehend and generate human-like text by learning intricate patterns and representations from extensive textual data. Their text generation capabilities have reached a level comparable to that of human writing [65], permeating various aspects of daily life and becoming essential in many professional workflows.

The core idea behind PLMs involves acquiring a generic, latent representation of language through universal task learning and applying it across diverse NLP tasks. PLMs play a crucial role in pushing the boundaries of NLP capabilities, constituting an active area of research and development. Key characteristics of PLMs include:

- Transformer Architectures: The Transformer architecture, introduced by Vaswani et al. [66], is a neural network architecture highly effective in capturing complex relationships in sequential data, such as language.
- Pre-training on Extensive Datasets and Transfer Learning: PLMs undergo pre-training on extensive datasets with diverse textual information, followed by fine-tuning for specific tasks [67–69].
- Versatility: PLMs find applications in various NLP tasks, including text classification, named entity recognition, question answering, sentiment analysis, language translation, and more.

Notable PLMs achieving cutting-edge performance include BERT (introduced by Google) and GPT (developed by OpenAI).

2.2. BERT (Bidirectional Encoder Representations from Transformers)

BERT [70–76], introduced by Google researchers in 2018 [70], has significantly advanced language-related tasks and was integrated into Google's search engine by 2019. An analysis published in 2020 highlighted BERT's rapid ascent in the research world, with over 150 scientific publications within just over a year [71]. As of December 2023, Scopus records 3716 scientific articles with the term "BERT" in the title.

Built on the transformer architecture, BERT employs bidirectional pre-training (on extensive corpora), enhancing its understanding of word context and relationships. During pre-training, it predicts missing words, enhancing its grasp of language semantics and context. After pre-training, the model can be fine-tuned for NLP tasks like language understanding, machine translation, question answering, sentiment analysis, and named entity recognition.

BERT uses the *WordPiece tokenizer* algorithm to tokenize input into smaller units, allowing for tokens to range from single characters to subword tokens; for example, the word "running" might be tokenized into subword tokens "run" and ′′###ing". Special tokens like [CLS] and [SEP] are added to delimit the text portion to be considered as a *unit of meaning*, so defining where that unit begins and ends within the text. Each token is assigned a unique ID from the BERT predefined vocabulary, facilitating the transformation of text into numeric sequences for neural network calculations. For example, the input "The quick brown fox jumps over the lazy dog" is tokenized as [101, 1996, 4248, 2829, 4419, 5086, 2058, 1996, 7887, 3899, 1012, 102], where 101 represents [CLS], 1996 corresponds to the ID for "The", 4248 to "quick", and so on.

2.3. GPT (Generative Pre-Trained Transformer)

GPT [77–79], developed by OpenAI, is a highly versatile language model known for generating coherent and contextually relevant text. Its evolution spans multiple iterations, including GPT-1 (2018), GPT-2 (2019), GPT-3 (2020), and GPT-4 (2023). GPT excels in various tasks such as text completion, translation, summarization, and question-answering.

Built on the transformer architecture, GPT undergoes pre-training a large neural network on diverse text corpora. This initial phase equips the model with a broad understanding of language. Following pre-training, GPT can be fine-tuned for specific tasks or employed as a general language model for various natural language understanding and generation tasks.

Tokenization is a fundamental process in GPT, allowing for the model to operate on discrete *units of meaning* and comprehend language structure. Tokens, the basic building blocks for GPT's processing, result from breaking down the text into smaller units, such as words or subwords. Special tokens like [CLS] (classification token) and [SEP] (separator token) are added to delimit the text portion to be considered as a *unit of meaning*. GPT's tokenization handles a wide vocabulary, incorporating rare or unseen words. The tokenized representation serves as input for both pre-training and fine-tuning computations. Tokenization is conceptually similar to that of BERT, except that GPT uses left-to-right tokenization, while BERT uses bidirectional tokenization.

Post tokenization, each token receives a unique numeric identifier (ID) using the GPT vocabulary. This vocabulary, derived from the pre-training corpus, includes complete words and subword units, enabling GPT to capture morphological variations and generate flexible, contextually rich text. The final numerically represented tokenized sequence is processed by the GPT model. For instance, the input text "The quick brown fox jumps over the lazy dog" is tokenized as [CLS], "The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog", ".", [SEP], with each token substituted by an appropriate numerical ID from the GPT vocabulary, including special tokens like [CLS] and [SEP].

In summary, we note that, in addition to the common characteristics already reported in Section 2.1 as both PLMs:

- While GPT and BERT have distinct training objectives and optimize for different tasks, their shared foundation highlights commonalities in handling contextualized language representations.
- Both GPT and BERT aim to capture contextual understanding of language. GPT achieves this by predicting the next word in a sequence (left-to-right), while BERT focuses on bidirectional representations, considering context from both the left and right sides of a word.
- Both GPT and BERT are designed to be versatile and adaptable to various natural language processing tasks. They can be fine-tuned for specific applications, such as sentiment analysis, named entity recognition, question answering, and more.

3. System Description

In this paper, we delve into AUTOMETA ([50]), a Linguistic-Based Text Mining system that employs a finite automaton to efficiently identify small portions of text with very high semantic content within an input document. The primary goal is to automatically determine the main knowledge domains of the text.

We must, first of all, define the linguistic constructs on which the system is based: the Multi-word units (MWUs) and the MWUs ontologies; then, we offer an idea of the algorithms and the related computational complexities on which the system is based.

3.1. Simple Words and Multi-Word Units

The words in a sentence can be broadly classified into two main categories: *simple words*, which are meaningful, uninterrupted sequences of letters delimited by spaces, and *multi-word units (MWUs)* [80–82], also known as multi-word expressions (MWEs). MWUs are composed of two or more simple words separated by spaces or other diacritical marks. These combinations of words function together as a single semantic unit, and the meaning of multi-word units is, to some extent, independent of that of their individual components.

A multi-word unit, can be exemplified by the following: *credit card* (*"It is a card that allows the holder to obtain, from businesses managed by the institution that issues the card itself, or from those affiliated with the issuer, goods or services without the simultaneous payment of the related prices, which are instead debited to an account the balance of which is subject to periodic collection."*) (Enciclopedia online Treccani, https://www.treccani.it/ (accessed on 15 September 2023)). It is typically *monosemic*, precisely identifying a semantic area or a well-defined object and nothing else. Multi-word units can encompass idioms, collocations, phrasal verbs, and other lexical combinations. Here are some examples:

- Idiomatic Expressions: *Kick the bucket* (to die) or *Break a leg* (good luck) or *Hit the hay* (go to sleep).
- Collocations: *Make a decision* or *Take a shower* or *Strong coffee*.
- Phrasal Verbs: *Turn on* (activate) or *Look up* (search for information).
- Proverbs: *Too many cooks spoil the broth* (when too many people are involved in a task, it may not turn out well) or *Don't count your chickens before they hatch* (do not make plans or celebrate success prematurely).
- Technical Terms: *Artificial intelligence* (the simulation of human intelligence in machines) or *Climate change* (long-term change in the average weather patterns).
- Compound Words: *Airplane* (Air+plane) or *Raincoat* (Rain+coat) or *Sunglasses* (Sun+glasses).
- Fixed Expressions (or Prepositional Phrases): *By the way* (incidentally) or *In the meantime* (meanwhile) or *All of a sudden* (unexpectedly).
- Named Entities: *New York City* or *United States* or *Mona Lisa* (the name of a picture).
- Acronyms and Initialisms: *NASA* (National Aeronautics and Space Administration) or *UNESCO* (United Nations Educational, Scientific and Cultural Organization).
- Technical Terminology: *Quantum mechanics* or *Genetic engineering* or *Nuclear fusion*.

An algorithmic challenge arises from the fact that a multi-word unit lacks explicit terminators to signal its start and end, unlike white space or punctuation that separates simple words. This poses a significant challenge in the automatic treatment of multi-word units since they can be very long and composed of any number of simple words. Consequently, when reading an input text, it remains uncertain how many consecutive elementary words must be read to form a multi-word unit. Furthermore, there is a close relationship between multi-word units and terminology. Specialized field lexicons are primarily composed of multi-word units, but this relationship extends beyond specialized lexical domains and encompasses all semantic areas. In a particular knowledge domain, the terminological multi-word units can be regarded as metadata for the corresponding ontology.

In contrast, a simple word (*credit* or *card*, taken individually) is usually *polysemic*, having various meanings and belonging to many different semantic areas, often distant from the specific context (for example, for the word "credit": *banking*, but also *psychology*, *mathematics*, *medicine*, *economics*, *industry*, etc.; similarly, for the word "card": *material*, *literature*, *medicine*, *commerce*, *economy*, *computer science*, etc.).

A single word almost always requires inclusion in a sentence to be disambiguated. In contrast, a multi-word unit retains its meaning regardless of the sentences in which it occurs and specifies its semantic content (e.g., Alice performed *market analysis*). Multi-word units are considered *units of fixed meaning* with specific formal and semantic characteristics: they are not extemporaneous combinations of simple words. Established linguistic studies indicate that, in natural language, the majority of cognitive information is conveyed through multi-word units.

3.2. Ontology System

AUTOMETA accesses terminological ontologies where each entry is assigned an ontological identification, comprising tags (or labels) referencing one or more knowledge domains or semantic fields (metadata) where the entry is commonly used and holds an unambiguous meaning. This tagging system is referred to as a lexical ontology, establishing a distinct and unequivocal relationship between a multi-word unit in a terminology dictionary and one or more tags denoting the semantic domain(s) to which it pertains. Essentially, it specifies the particular domain(s) of knowledge where that multi-word unit possesses a well-defined meaning. To illustrate, the multi-word unit *analisi cliniche* (*clinical analysis*) is tagged with MEDIC, signifying *Medicine*, while the multi-word unit *carta di credito* (*credit card*) is tagged with ECONO, signifying *Economy*. Each ontology undergoes creation and verification under the supervision of experts in the respective sector. It is important to note that our lexical ontologies do not serve encyclopedic purposes. For instance, while they may not encompass Subnuclear Physics, they provide all the ontology used by Subnuclear Physics to unequivocally reference a set of established, standardized, and universally shared knowledge.

The connection between multi-word units and domain field tags is ontological, establishing the unambiguous meaning that multi-word units hold within specific cognitive contexts. As previously mentioned, it has been emphasized that the information conveyed in any text is predominantly conveyed by terminological multi-word units. This implies that by extracting terminological multi-word units from a given text, we automatically extract its general meaning in the form of ontology strings, comprising a multi-word unit and an unambiguous domain field tag.

3.3. System Functionality

The system comprises a preliminary linguistic phase for the design and development of the IT data structure for the ontologies, followed by an algorithmic phase.

3.3.1. Preliminary Linguistic Step

For AUTOMETA to recognize a specific knowledge domain, the system requires the existence of an ontology of terminological multi-word units for that domain. If the ontology is not available, it must be created. This operation is a one-time task and should be conducted in collaboration with experts in the relevant sector. At runtime, the system allows the combination of multiple ontologies from different knowledge domains, integrating them into a unified structure.

The preliminary linguistic phase involves the following elementary steps:

Step 1 Multi-Word Unit Choice and Normalization: Experts in the sector identify all multi-word units of the knowledge domain (metadata) and normalize the data by transforming the set of multi-word units into a standard format, following the normalization choices determined in the previous step.

Step 2 Association of the Semantic Domain with Each Multi-Word Unit: The system builds a linguistic model that associates one or more semantic domains with each of the multi-word units identified in the previous step.

Step 3 Ontology Construction: A digital data structure is created, containing all possible multi-word units in the form of an ontology. Each line in the ontology includes a multi-word unit and its associated semantic domains.

3.3.2. Algorithmic Step

The algorithmic process comprises a one-time pre-processing phase and a two-stage processing phase for each input text: Matching and Analysis.

Pre-Processing: The system builds a finite automaton data structure based on user-selected ontologies.

Matching: The algorithm reads the input text character by character, navigating through the finite automaton, where each character of the input text corresponds to a state of the finite automaton. It identifies terminological multi-word units immediately and simultaneously, even if they are partially or completely overlapped, without revisiting previous characters.

Analysis: After reading the entire input text, the algorithm analyzes metadata and classifies the text based on its knowledge domains.

4. Algorithm Performances

The system processes input text by navigating a finite automaton, searching for Multi-Word Units from selected ontologies. When reaching a state corresponding to the end of a multi-word unit, it outputs the associated semantic domain(s). The algorithms are highly performant, fast, and precise, capable of identifying all occurrences of multi-word units, even if they partially or completely overlap.

Phase 1: Preprocessing

Initially, the user selects ontologies, and the system constructs a finite automaton in the computer's memory based on these choices. This occurs once per session and whenever reference ontologies change. The time complexity of the preprocessing phase (i.e., the total number of elementary algorithmic steps, hence the number of state transitions on the finite automaton) is $O(m)$ [55], where m is linearly proportional to the sum of the lengths of all entries in the selected ontologies (n elements), as given by

$$m = \sum_{i=1}^{n} \text{length}(a_i).$$

Phase 2: Matching

The automaton continuously reads lines of input text, character by character, never stepping back through the input text. The characters guide it through the finite automaton; when it encounters a state corresponding to the end of a multi-word unit, it emits its associated semantic domains. All terminological multi-word units are recognized simultaneously in a single pass (i.e., with only one reading of the input text) even if partially or totally overlapping, however long the multi-word units are. Thus, the algorithm works in linear time with respect to the number of characters in the input text [55]; therefore, its performance does not depend on the size of the ontologies (i.e., the set of all possible multi-word units). It also runs on a data structure entirely in the computer's main memory.

Phase 3: Analysis

At the end of the Matching phase, when the input text is completely read, the system counts the frequency of each semantic field that emerges and proposes a classification of the input text knowledge domain based on the most frequent ones.

5. System Performances

The system operates on a standard architecture without specific hardware or software requirements. The tests were conducted using the following hardware configuration:
Laptop: DELL INSPIRON 16;

CPU: 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz 2.30 GHz;
RAM: 32.0 GB;
OS: Windows 11 Home, 64-bit, x64, ver. 22H2.

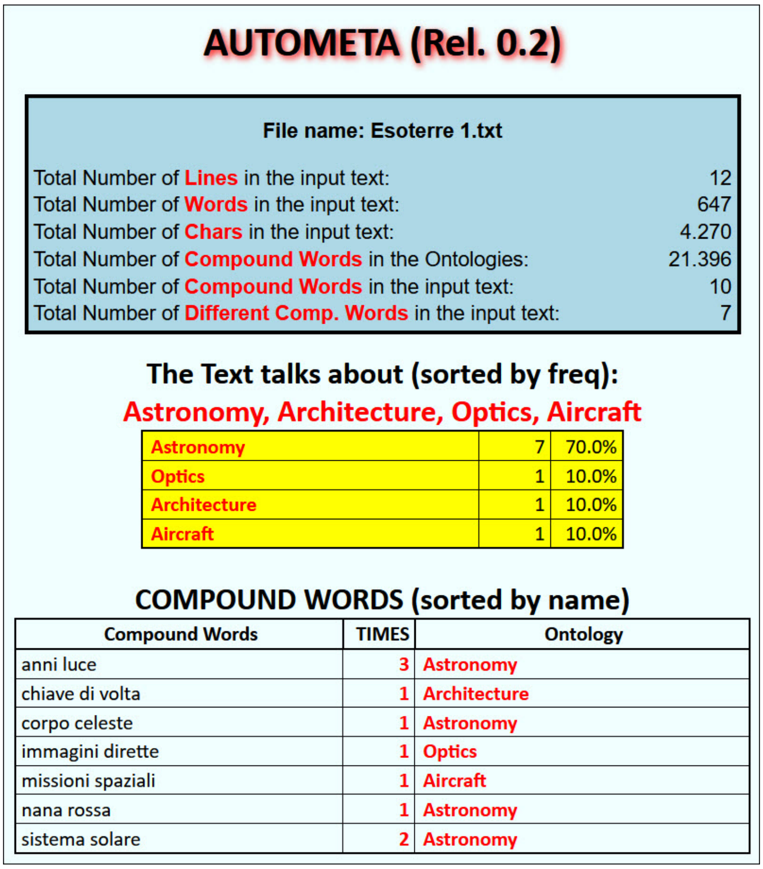
There is no need for a particular video card or mass memory beyond standard specifications. The software was developed using "Embarcadero Delphi 11.2", a robust RAD (Rapid Application Development) visual software development tool based on an Object Programming Language. The computer was running system applications during program execution.

5.1. AUTOMETA at Work

In this paper, we present a test conducted on an Astronomy document titled "Two potentially habitable exo-earths discovered 16 light years from us" from "Le Scienze" (the Italian edition of ″Scientific American″) dated 15 December 2022 (https://www.lescienze.it/news/). The document consists of 625 words, 4231 characters, and 12 paragraphs. We performed the test in five different scenarios:

1. On the entire document.
2. On the first two thirds of the document.
3. On the first third of the document.
4. On the central part of the document.
5. On the union of the first and last paragraphs of the document.

It is noteworthy that the system effectively identifies target information even with very few words, less than 200. The analysis results for the entire document are summarized in Figure 2, and the outcomes of the five tests are presented in Table 1.



AUTOMETA (Rel. 0.2)

```
File name: Esoterre 1.txt

Total Number of Lines in the input text:                          12
Total Number of Words in the input text:                         647
Total Number of Chars in the input text:                       4.270
Total Number of Compound Words in the Ontologies:             21.396
Total Number of Compound Words in the input text:                 10
Total Number of Different Comp. Words in the input text:           7
```

The Text talks about (sorted by freq):
Astronomy, Architecture, Optics, Aircraft

Astronomy	7	70.0%
Optics	1	10.0%
Architecture	1	10.0%
Aircraft	1	10.0%

COMPOUND WORDS (sorted by name)

Compound Words	TIMES	Ontology
anni luce	3	Astronomy
chiave di volta	1	Architecture
corpo celeste	1	Astronomy
immagini dirette	1	Optics
missioni spaziali	1	Aircraft
nana rossa	1	Astronomy
sistema solare	2	Astronomy

Figure 2. Example of AUTOMETA output.

Table 1. Summary of the 5 subtests for the example.

Words	Chars	Knowledge Domains
647	4270	Astronomy (7), Architecture (1), Aircraft (1), Optics (1)
440	2884	Astronomy (7), Aircraft (1), Optics (1)
168	1093	Astronomy (6)
186	1220	Astronomy (2)
182	1178	Astronomy (3), Architecture (1)

5.2. Running System Performances

During the startup phase, the software constructs a Finite Automaton in RAM, encompassing all chosen ontologies. The system includes one generic multi-word unit ontology and 22 terminological multi-word unit ontologies, totaling 21,396 entries. The time needed to build the finite automaton for the entire set of ontologies ranges from 8 to 12 s, depending on system overhead.

The pre-processing time for a text varies from 0.1 s for the smallest text (103 words and 644 characters) to 2.86 s for the longest text (556,907 words and 3,449,325 characters).

The processing time for a text ranges from 0.1 s for the smallest text to 1.35 s for the longest text.

The maximum space used by the entire system, including the Finite Automaton and 23 ontologies with 21,396 elements, is 709 MB for the longest text.

6. Case Study Description

An early prorotype of the system correctly recognized the knowledge domains for the vast majority of texts (over 90%) analyzed.

6.1. Corpus

We worked on a dataset of 100 short-length texts (a few hundred words) and medium-length texts (a few thousand words), selected from the websites of the main national and regional newspapers, as well as from specialized magazines. The average length of the corpus is 766 words and 5436 characters. The minimum-length text contains 252 words and 1777 characters, while the maximum-length text contains 8800 words and 56,715 characters.

The total number of Multi-Word Units is 21,396 which comes from the union of 1 generic multi-word unit ontology and 22 terminological multi-word unit ontologies.

6.2. Methodology

After conducting a personal (human) reading of each text, we subjectively classified the documents by associating each with a maximum of three knowledge domains, those which, in our opinion, exhibited the closest correspondence between text and content.

In the second operational phase, we subjected the same documents to the text mining system and then compared the results obtained with the previous "human" categorization, summing the percentages of all the words found that belong to one of the three domains associated with the text in the first step.

6.3. Results

When the sum of the percentages from the computer analysis, compared to the human analysis, was

- from 85% upwards, we considered the objective fully achieved;
- between 70% and 84%, we considered the objective well achieved;
- between 50% and 69%, we considered the objective achieved;
- between 30% and 49%, we considered the objective poorly achieved;
- from 29% down, we considered the objective not reached.

The results of the test are resumed in Table 2.

Table 2. Results of the test on 100 documents.

Goal	Documents	Avg. Words	Avg. Chars	Avg. Percentage
Fully Achieved	52	721.50	5279.87	93.56%
Well Achieved	31	956.93	6731.37	75.97%
Achieved	8	698.25	4598.13	52.50%
Poorly Achieved	2	581.50	3687.50	37.50%
Not Achieved	7	440.71	2716.14	2.86%

It should be noted that 6 out of 7 documents that did not reach the objective were passages taken from literary works (by Shakespeare, Sartre, Eliot, Plautus, Fruttero and Lucentini, De Carlo).

7. Natural Language Texts Main Issues

Text mining faces several intrinsic problems when dealing with unstructured or semi-structured natural language texts. Analyzing these types of documents poses difficulties for computers, in contrast to the ease with which humans can manipulate them. These challenges need to be addressed by all text mining systems, regardless of their approach and techniques.

These issues arise mainly because natural language textual documents originate from diverse sources, primarily created by human beings with different backgrounds and cultures. Human language is inherently complex, ambiguous, and variable, and these issues are mainly related to semantics, context, formatting, and standardization.

While some issues can be addressed automatically through effective text processing or neutralized by algorithmic and statistical–mathematical methods, others demand deep linguistic and semantic understanding of the text.

The field of text analysis is currently focused on processing, employing statistical and mathematical analyses that are linguistically superficial and fall short of achieving the deep semantic understanding accomplished by humans [21]. In essence, the more a text mining approach relies on language knowledge, the better it can overcome some of these problems. Until these challenges are overcome, semantic compromises may be necessary in text analysis. Here are some of the main issues in text mining natural language texts:

7.1. Ambiguity

7.1.1. Lessical Ambiguity

Polysemous Words (words with multiple inflections) require context for disambiguation (e.g., “park”: a public area OR the act of parking a vehicle).

7.1.2. Syntactical Ambiguity

Understanding *syntactic ambiguity* is particularly challenging. For example, the written sentence “Mario saw Franco in the park with binoculars” can be interpreted in four different ways, highlighting the complexity of language comprehension:

1. Mario is in the park and (Mario) has binoculars (with which he saw Franco);
2. Mario is in the park and saw Franco (who may not be in the park) who has binoculars;
3. Franco is in the park and Mario has binoculars (with which he saw Franco);
4. Franco is in the park and has binoculars and Mario saw him.

Incidentally, we note that this problem is so difficult that, simply in its written formulation, it cannot be solved even by human beings, unless other information derived from other senses (e.g., sight) or other knowledge is added.

7.2. Word Variability

7.2.1. Synonyms

Different words might have very similar, but not exact, meanings; sometimes the nuances could be very important for understanding the written text.

7.2.2. Variations in Spelling

Variations in spelling refer to different ways words can be written, leading to differences in their appearance. These variations can occur due to regional differences, language evolution, historical influences, individual preferences, and errors. It is important to recognize these variations, especially in natural language processing and text mining, where understanding different forms of words is crucial for accurate analysis and interpretation. Here are some examples:

- British vs. American English: Color (American) vs. Colour (British)
- Alternative Spellings: Traveling vs. Travelling
- Abbreviations and Contractions: Can't vs. Cannot
- Improper Word Usage due to dialectical habits (e.g., "the teacher learns me to speak" instead of ″teaches")
- Neologisms: Emergence of new words borrowed from online dialogues (e.g., "L8er″ for "later″)

7.2.3. Typos and Misspellings

In a text, there could be typos and/or misspelling that could be of two main types:

- non-existing word (e.g., banf instead of band);
- existing word (e.g., bank instead of band) that could change the meaning of a sentence ("the bank went bankrupt" is a possible thing, just like "the band went bankrupt").

7.3. Writing Style: Standardization and Formatting Problems

These problems are primarily due to the absence of universal rules for writing texts that a computer can consistently understand:

- Various uses of *accented characters* (only for vowel "a″ there are 10 variants: $\acute{a}$, $\grave{a}$, $\hat{a}$, $\check{a}$, $\breve{a}$, $\tilde{a}$, $\bar{a}$, $\vec{a}$, $\dot{a}$, $\ddot{a}$),
- Varied use of *apostrophes* (e.g., pò instead of po' in Italian) highlights dialectical nuances.
- Variations in *punctuation* and *capitalization*.
- Diverse *sentence structures* contribute to the uniqueness of writing styles.
- *Colloquial Expressions* are also widely present, ranging from online texts to more formal contexts like literature, politics, and journalism.

7.4. Hidden Information

These problems arise from the complexity of the problem and the cognitive limitations of computers. Instead, these issues require "linguistic–cognitive" methods closer to human comprehension to understand the information hidden at deeper levels than the simple frontal text, since humans use natural language to *express thoughts, emotions, and feelings*, employing

- *Euphemisms, Disguises, and Metaphors* to convey nuanced meanings.
- *Irony and Rhetorical Figures* to add layers of meaning.

These features often *hide information* beyond direct readability, posing a significant challenge for technology due to the need for deep understanding and reasoning. For example: ″Bob left us yesterday" (Where did Bob go? Did he move, change jobs, or die?), or "You are the apple of my eyes", or "Messi is a poor footballer″ (ironically said about Lionel Messi after just one bad match) or "Ronaldo is an ex-player″ (ironically said about Cristiano Ronaldo because he is an old player).

8. Comparisons

Let us now focus our attention mainly on the comparison between AUTOMETA and PLMs, both because the latter are the most scientifically advanced systems, and those with the highest level of diffusion and success, and because non-linguistic text mining approaches present, in principle, the same problems of PLMs.

8.1. PLM Family Systems Main Issues

In their exploration of Large Language Models (LLMs), Zhang et al. [83] state that *"Despite its powerful modeling and description capabilities, LLMs present significant problems and limitations"*.

An article by the editors of "MIT Technology Review" on 24 February 2021 (https://www.technologyreview.com/2021/02/24/1014369/10-breakthrough-technologies-2021) states that *"GPT-3 can mimic human-written text with uncanny—and at times bizarre—realism, making it the most impressive language model yet produced using machine learning. But GPT-3 doesn't understand what it's writing, so sometimes the results are garbled and nonsensical. It takes an enormous amount of computation power, data, and money to train, creating a large carbon footprint and restricting the development of similar models to those labs with extraordinary resources. And since it is trained on text from the internet, which is filled with misinformation and prejudice, it often produces similarly biased passages"*.

In a statement on 20 November 2022 (https://openai.com/blog/chatgpt#OpenAI), OpenAI acknowledges challenges with ChatGPT's accuracy. They note that *"ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows"*.

Now, let us delve into the primary issues associated with PLMs.

Pre-training: The process involves training a language model on a corpus of textual data. Schramowski et al. [84] explore the ethical and moral dimensions of LLMs. They emphasize that during pre-training, PLMs, exposed to vast amounts of unfiltered textual data, not only capture linguistic knowledge but also implicitly retain general knowledge present in the data.

- Data size: Pre-training requires a huge quantity of unfiltered data. The system's statistical engine improves with the quantity of data it processes. As of 28 May 2020, GPT-3 counts 570 GB (300 billion words) of data from CommonCrawl, WebText, English Wikipedia, and two books corpora (Books1 and Books2) [85].
- Data quality: There is no certainty that the pre-training data are correct, as online text is rife with errors, biases, malicious content, and misinformation. In some cases, pre-training data may violate moral and legal laws or contain potentially sensitive information, as discussed by Huang et al. [86] in their analysis of the personal information disclosure tendencies of individuals with PLMs. Consequently, PLMs inherit and perpetuate biases from the training data, resulting in uncontrollable texts with distorted and biased content.

Tokenization: PLMs break the text into tokens, then transform each token into a number and thereafter look for patterns in these tokens. PLMs implicitly capture information about words and their relationships based on the patterns and sequences they observe in the tokens from the data they were trained on. These models are capable of recognizing or generating coherent text, but they do not have explicit knowledge of semantics or the ability to understand concepts in the same way humans do. The model generates contextually appropriate text by predicting the next token in a sequence based on the preceding tokens.

Semantics: Tokenization is performed without considering meanings or semantics. Consequently, these models lack a deep understanding of written language. This implies that the models cannot justify or explain exactly what they emit, which is particularly critical in applications such as healthcare and finance. Furthermore, recognizing ambiguity, both lexical and syntactic, is not straightforward. In general, questions related to the semantic knowledge of a text, such as detecting hidden meanings, remain challenging to resolve.

Multi-Word Units: Multi-word units (MWUs) are "units of meaning" with high semantic value, but processing them is not straightforward.

- MWUs in PLMs: PLMs have no explicit knowledge of multiword units (MWUs) or phrases as atomic entities; they process an MWU as sequences of individual tokens. For example, the phrase "natural language processing" would be tokenized into three separate tokens: "natural", "language", and "processing". Each token is considered independently during the model's training and inference.
- MWU Size: An MWU can be very long and is composed of any number of simple words, as discussed in Section 3.1. An MWU has no "end unit" symbol (like a space or punctuation for a simple word). In real life, an MWU could theoretically extend to the end of the entire text. In the tokenization process, PLM determines where a "unit of meaning" stops and puts an "end of sentence" token.
- Occurrences of MWUs: In texts, MWUs occur with low frequency compared to single words; therefore, in an overall statistical analysis, they tend to be overlooked. To address this problem, it would be advisable to separate MWU statistics from simple word statistics.

Spelling:

- Text Flaws: A text, both in the pre-training phase and in the mining phase, could contain spelling errors, typos, misspellings, and features of writing styles. Every word and every little part of a text is important in both the pre-training and mining processes, so a text should be corrected for these issues.
- Pre-processing: Pre-processing cleaning of the input text is necessary, but this is an expensive, time-consuming operation, and often it is not even possible. A pre-processing routine might detect a spelling error, but it might not correct it. Sometimes it may not even detect the error; for example, if there is a word that is lexically but not semantically correct.

Behavior with texts of rare or unknown topics to the model: These models do not work reasonably well on rare or unknown events. LLMs may struggle to provide accurate responses or predictions for rare or unprecedented events due to their heavy reliance on learned patterns from historical data.

Resource: These models, mainly for the pre-training process, require a lot of computational power, resources, and capital investment.

Management: These models could be managed only by very large companies.

8.2. A Comparison between PLMs and AUTOMETA

Pre-Training

PL Typically, these techniques necessitate a significant initial learning process, often involving the analysis of millions of documents. The system's statistical engine improves with the quantity of data it processes.

AU In contrast, the proposed system does not require an extensive initial learning process and can recognize the semantic domain of a document without the need to read other documents from the same knowledge domain.

Pre-processing

PL Natural language texts lack a standard format (refer to Section 7), requiring an effort to make a text as "standard" as possible. Achieving this involves both a syntactic pre-processing phase and a deep understanding of the language, which statistical methods lack. This phase is delicate and time consuming, particularly because simple words, the focus of these methods, are numerous in a text. Probabilistically, they are often not very distinct from each other, making them easily confused. For instance, in a sentence discussing one's bank, the phrase "my band is far away" might remain unnoticed as a writing error where "band" should have been "bank". Context investigation is necessary to identify such errors.

AU In contrast, the proposed system involves a "small" pre-processing phase to normalize the formal spelling of words, as multi-word units are few and diverse, minimizing pre-processing and dimensionality reduction. The system recognizes that these tasks are time consuming, error prone, and language specific.

auto-correction

PL Due to the lexical proximity among simple words, implementing auto-correcting techniques for errors is challenging.

AU In contrast, the proposed system aims to auto-correct as many errors as possible in "useful" but ill-written text.

polysemic

PL In natural language, simple words are inherently polysemous, belonging to multiple semantic domains.

AU In contrast, the proposed system detects as few "pieces of text" as possible, each carrying a substantial amount of semantic information and having few synonymy relationships.

8.3. A Comparison between AUTOMETA and Other Methods

AUTOMETA is a Linguistic-Based Text Mining system that utilizes a finite automaton to efficiently identify small portions of text with very high semantic content (MWUs) within an input text. These MWUs are then compared with MWU ontologies to automatically determine the main knowledge domains of the text.

8.3.1. Different Strategies from AUTOMETA

Algorithmically, employing a different method for matching simple words or MWUs against a dictionary can be highly time consuming and depends on the size of the dictionary: the larger it is, the longer the wait for a response. Identifying a multi-word unit from an ontology with m elements in a text with n words requires $O(\log m)$ accesses to an ordered data structure for each single search and globally $O(n)$ accesses to the data structure for each single word in the input text. This is because the complete chain with all the successive words of the focused word has to be attempted. Thus, the comprehensive cost to analyze a single word in the input text requires at least $O(n \log m)$ accesses to a data structure if it is maintained sorted with respect to the alphabetical order of the items. Since the input text contains n words, there are necessary $O(n^2 \log m)$ searches in the ontology to analyze the entire input text. In addition, in the case of deletion, insertion, or modification of an entry in an ontology, such an operation must be performed while maintaining the order in the ontology. This costs at least $O(\log m)$ accesses to the ontology, possibly involving the physical movement of entries from one memory area to another.

Furthermore, simple word text-mining systems tend to perform better on long texts and less effectively on shorter documents. This is attributed to the presence of numerous simple words in a text, making it easier for statistical regularities to emerge in longer texts.

8.3.2. AUTOMETA

- Independent of the length and number of elementary words in a multi-word unit, it addresses challenges posed by very long units composed of a variable number of simple words.
- The processing time for a text of k characters requires $O(k)$ state transitions. The Finite Automaton reads the input text character by character, concurrently identifying all multi-word units from selected ontologies, performing the analysis in linear time.
- Can handle large dictionary sizes without issues.
- Recognizes partially or completely overlapping multi-word units without additional computational effort.
- The entire ontology, stored in a finite automaton in central memory, ensures significantly faster access times compared to secondary memory.
- Ontologies do not need to be sorted, making addition, modification, or deletion of entries virtually computationally cost-free; we simply insert a new word at the end of the ontology or modify or delete a word without touching the others.
- Recognizes multi-word units directly while reading the input text, without providing feedback.
- Efficient for both short and long documents.
- Does not require a pre-training process.

9. Conclusions and Future Research Perspectives

9.1. Conclusions

This paper introduces an efficient algorithmic–linguistic text mining approach that identifies knowledge domains within a text by detecting terminological multi-word units and comparing them with ontologies. The algorithm utilizes a finite state automaton for all ontologies, emitting knowledge domains upon recognizing multi-word units. Upon reading the entire input text, the algorithm analyzes the metadata issued to associate one or more knowledge domains with the document. This approach is applicable to both long and short texts, showcasing high efficiency, speed, and precision.

In conclusion, we assert that our semantics-based text mining and information retrieval approach is well-suited for achieving the goals of automatic recognition and the related semantic cataloging of (semi) unstructured natural language texts.

9.2. Future Research Perspectives

To achieve efficient semantic-based text mining, special attention must be paid to analyzing and promptly capturing linguistic changes, primarily driven by the evolving facets of civilization. Potential avenues for future research could include:

- Ontology Structure Update: Focus on updating existing ontologies and creation of new electronic terminology ontologies for various semantic domains, such as e-government, biomedicine, ecological transition, etc. The flexibility of our approach allows modifications without the need to reorder ontology items.
- Testing and Validation: Conduct of comprehensive testing and validation of existing ontologies by applying them to more extensive corpora.
- Integration into Websites and Portals: Exploration of the integration of ontologies and software tools into websites and portals, allowing for internet users to test the system and provide feedback on its quality and usability.
- Automatic Translation of Multi-Word Units: Investigation of the addition of a translation feature to our approach, enabling the translation of multi-word units from one language to another. This capability is crucial, as translating multi-word units is not always straightforward by translating their individual components, given that a multi-word unit is a whole. The rapid translation feature could address errors in technological and scientific translation caused by the lack of reliable terminological bilingual electronic glossaries/dictionaries.

Funding: Research funded by Università degli Studi di Salerno (300639FRB22POSTI) | Università degli Studi di Salerno.

Data Availability Statement: The data will be made available by the authors on request.

Acknowledgments: My main thanks go to Maurizio Talamo, a teacher and a friend who, as always, advised me and gave me fundamental ideas to be able to proceed with my work. I am very grateful also to Mario Monteleone for the long discussions on computational linguistics, always very useful and stimulating. Without his contribution "Cataloga", the ancestor of AUTOMETA, would not have existed.

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

References

1. Chen, M.; Mao, S.; Liu, Y. Big Data: A Survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
2. Philip Chen, C.; Zhang, C.Y. Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [CrossRef]
3. Tsai, C.W.; Lai, C.F.; Chao, H.C.; Vasilakos, A.V. Big Data Analytics: A Survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
4. Oussous, A.; Benjelloun, F.Z.; Ait Lahcen, A.; Belfkih, S. Big Data Technologies: A Survey. *J. King Saud Univ.-Comput. Inf. Sci.* **2018**, *30*, 431–448. [CrossRef]
5. Adadi, A. A Survey on Data-efficient Algorithms in Big Data Era. *J. Big Data* **2021**, *8*, 24. [CrossRef]
6. Zhang, H.; Lee, S.; Lu, Y.; Yu, X.; Lu, H. A Survey on Big Data Technologies and Their Applications to the Metaverse: Past, Current and Future. *Mathematics* **2023**, *11*, 96. [CrossRef]
7. Atzori, L.; Iera, A.; Morabito, G. The Internet of Things: A Survey. *Comput. Netw.* **2010**, *54*, 2787–2805. [CrossRef]
8. Tsai, C.W.; Lai, C.F.; Chiang, M.C.; Yang, L.T. Data Mining for Internet of Things: A Survey. *IEEE Commun. Surv. Tutorials.* **2014**, *16*, 77–97. [CrossRef]
9. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2347–2376. [CrossRef]
10. Qadri, Y.A.; Nauman, A.; Zikria, Y.B.; Vasilakos, A.V.; Kim, S.W. The Future of Healthcare Internet of Things: A Survey of Emerging Technologies. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1121–1167. [CrossRef]
11. Zhong, Y.; Chen, L.; Dan, C.; Rezaeipanah, A. A Systematic Survey of Data Mining and Big Data Analysis in Internet of Things. *J. Supercomput.* **2022**, *78*, 18405–18453. [CrossRef]
12. Xu, L.D.; He, W.; Li, S. Internet of Things in Industries: A Survey. *IEEE Trans. Ind. Inform.* **2014**, *10*, 2233–2243. [CrossRef]
13. Boyes, H.; Hallaq, B.; Cunningham, J.; Watson, T. The Industrial Internet of Things (IIoT): An Analysis Framework. *Comput. Ind.* **2018**, *101*, 1–12. [CrossRef]
14. Sisinni, E.; Saifullah, A.; Han, S.; Jennehag, U.; Gidlund, M. Industrial Internet of Things: Challenges, Opportunities, and Directions. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4724–4734. [CrossRef]
15. Paniagua, C.; Delsing, J. Industrial Frameworks for Internet of Things: A Survey. *IEEE Syst. J.* **2021**, *15*, 1149–1159. [CrossRef]
16. Akhtar, M.; Neidhardt, J.; Werthner, H. The Potential of Chatbots: Analysis of Chatbot Conversations. In Proceedings of the 21st IEEE Conference on Business Informatics, CBI 2019, Moscow, Russia, 15–17 July 2019; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019; Volume 1, pp. 397–404. [CrossRef]
17. Chaves, A.P.; Gerosa, M.A. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *Int. J. Hum. Comput. Interact.* **2021**, *37*, 729–758. [CrossRef]
18. Chao, M.H.; Trappey, A.J.C.; Wu, C.T. Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics. *Complexity* **2021**, *2021*, 5511866. [CrossRef]
19. Rapp, A.; Curti, L.; Boldi, A. The Human Side of Human-Chatbot Interaction: A Systematic Literature Review of Ten Years of Research on Text-Based Chatbots. *Int. J. Hum. Comput. Stud.* **2021**, *151*, 102630. [CrossRef]
20. Luo, B.; Lau, R.Y.K.; Li, C.; Si, Y.W. A Critical Review of State-of-the-Art Chatbot Designs and Applications. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1434. [CrossRef]
21. Zong, C.; Xia, R.; Zhang, J. *Text Data Mining*; Springer: Singapore, 2021; pp. 1–351. [CrossRef]
22. Tandel, S.S.; Jamadar, A.; Dudugu, S. A Survey on Text Mining Techniques. In Proceedings of the 2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019, Coimbatore, India, 15–16 March 2019; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019; pp. 1022–1026. [CrossRef]
23. Aggarwal, C.C.; Zhai, C. *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 1–526.
24. Aggarwal, C.C.; Zhai, C. A Survey of Text Classification Algorithms. In *Mining Text Data*; Aggarwal, C.C., Zhai, C., Eds.; Springer: Boston, MA, USA, 2012; pp. 163–222. [CrossRef]
25. Usai, A.; Pironti, M.; Mital, M.; Aouina Mejri, C. Knowledge Discovery out of Text Data: A Systematic Review via Text Mining. *J. Knowl. Manag.* **2018**, *22*, 1471–1488. [CrossRef]
26. Kowsari, K.; Meimandi, K.J.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]

27. Kumar, M.; Kumar, S.; Yadav, S.L. *Data Mining for the Internet of Things: A Survey*; Apple Academic Press: Waretown, NJ, USA, 2023; pp. 93–109.
28. Navathe, S.B.; Ramez, E. Data warehousing and data mining. In *Fundamentals of Database Systems*; Pearson Education: Singapore, 2000; pp. 841–872.
29. Gupta, V.; Lehal, G.S. A survey of text mining techniques and applications. *J. Emerg. Technol. Web Intell.* **2009**, *1*, 60–76. [CrossRef]
30. Liao, S.H.; Chu, P.H.; Hsiao, P.Y. Data Mining Techniques and Applications—A Decade Review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]
31. Kusakin, I.; Fedorets, O.; Romanov, A. Classification of Short Scientific Texts. *Sci. Tech. Inf. Process.* **2023**, *50*, 176–183. [CrossRef]
32. Danilov, G.; Ishankulov, T.; Kotik, K.; Orlov, Y.; Shifrin, M.; Potapov, A. *The Classification of Short Scientific Texts Using Pretrained BERT Model*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 83–87. [CrossRef]
33. Ongenaert, M.; Van Neste, L.; De Meyer, T.; Menschaert, G.; Bekaert, S.; Van Criekinge, W. PubMeth: A cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.* **2008**, *36*, D842–D846. [CrossRef] [PubMed]
34. Cejuela, J.M.; McQuilton, P.; Ponting, L.; Marygold, S.; Stefancsik, R.; Millburn, G.H.; Rost, B. Tagtog: Interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database* **2014**, *2014*, bau033. [CrossRef]
35. Baltoumas, F.A.; Zafeiropoulou, S.; Karatzas, E.; Paragkamian, S.; Thanati, F.; Iliopoulos, I.; Eliopoulos, A.G.; Schneider, R.; Jensen, L.J.; Pafilis, E.; et al. OnTheFly2.0: A text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. *NAR Genom. Bioinform.* **2021**, *3*, lqab090. [CrossRef]
36. Chu, C.Y.; Park, K.; Kremer, G.E. A Global Supply Chain Risk Management Framework: An Application of Text-Mining to Identify Region-Specific Supply Chain Risks. *Adv. Eng. Inform.* **2020**, *45*, 101053. [CrossRef]
37. Nota, G.; Postiglione, A.; Carvello, R. Text Mining Techniques for the Management of Predictive Maintenance. In Proceedings of the 3rd International Conference on Industry 4.0 and Smart Manufacturing, ISM 2021, Linz, Austria, 17–19 November 2021; Longo, F., Affenzeller, M.P.A., Eds.; Elsevier: Amsterdam, The Netherlands, 2022; Volume 200, pp. 778–792. [CrossRef]
38. Kumar, B.S.; Ravi, V. A Survey of the Applications of Text Mining in Financial Domain. *Knowl.-Based Syst.* **2016**, *114*, 128–147. [CrossRef]
39. Gupta, A.; Dengre, V.; Kheruwala, H.A.; Shah, M. Comprehensive review of text-mining applications in finance. *Financ. Innov.* **2020**, *6*, 39. [CrossRef]
40. Kumar, S.; Kar, A.K.; Ilavarasan, P.V. Applications of text mining in services management: A systematic literature review. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100008. [CrossRef]
41. Ngai, E.; Lee, P. A review of the literature on applications of text mining in policy making. In Proceedings of the Pacific Asia Conference on Information Systems, PACIS 2016, Chiayi, Taiwan, 27 June–1 July 2016.
42. Fenza, G.; Orciuoli, F.; Peduto, A.; Postiglione, A. Healthcare Conversational Agents: Chatbot for Improving Patient-Reported Outcomes. *Lect. Notes Netw. Syst.* **2023**, *661*, 137–148. [CrossRef]
43. Cheerkoot-Jalim, S.; Khedo, K.K. A systematic review of text mining approaches applied to various application areas in the biomedical domain. *J. Knowl. Manag.* **2020**, *25*, 642–668. [CrossRef]
44. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Shirvanizadeh, N.; Ortiz, A.; Pardo-Quiles, D.J. Applications of artificial intelligence, machine learning, big data and the internet of things to the COVID-19 pandemic: A scientometric review using text mining. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8578. [CrossRef]
45. Abbe, A.; Grouin, C.; Zweigenbaum, P.; Falissard, B. Text mining applications in psychiatry: A systematic literature review. *Int. J. Methods Psychiatr. Res.* **2016**, *25*, 86–100. [CrossRef]
46. Drury, B.; Roche, M. A Survey of the Applications of Text Mining for Agriculture. *Comput. Electron. Agric.* **2019**, *163*, 104864. [CrossRef]
47. Irfan, R.; King, C.K.; Grages, D.; Ewen, S.; Khan, S.U.; Madani, S.A.; Kolodziej, J.; Wang, L.; Chen, D.; Rayes, A.; et al. A Survey on Text Mining in Social Networks. *Knowl. Eng. Rev.* **2015**, *30*, 157–170. [CrossRef]
48. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. *Adv. Sci. Technol. Eng. Syst.* **2017**, *2*, 127–133. [CrossRef]
49. Ferreira-Mello, R.; André, M.; Pinheiro, A.; Costa, E.; Romero, C. Text Mining in Education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1332. [CrossRef]
50. Postiglione, A. *Text Mining with Finite State Automata via Compound Words Ontologies*; Lecture Notes on Data Engineering and Communications Technologies; Springer Nature: Berlin, Germany, 2024; Volume 193, pp. 1–12.
51. Gross, M. Lexicon-Grammar and the Syntactic Analysis of French. In Proceedings of the 10th International Conference on Computational Linguistics, COLING 1984 and 22nd Annual Meeting of the Association for Computational Linguistics, ACL 1984, Stanford, CA, USA, 2–6 July 1984; pp. 275–282.
52. Gross, M. The construction of electronic dictionaries; [La construction de dictionnaires électroniques]. *Ann. Télécommun.* **1989**, *44*, 4–19. [CrossRef]
53. Gross, M. The Use of Finite Automata in the Lexical Representation of Natural Language. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **1989**, *377*, 34–50. [CrossRef]
54. Monteleone, M. NooJ for Artificial Intelligence: An Anthropic Approach. *Commun. Comput. Inf. Sci.* **2021**, *1389*, 173–184. [CrossRef]
55. Aho, A.V.; Corasick, M.J. Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM* **1975**, *18*, 333–340. [CrossRef]

56. Boyer, R.S.; Moore, J.S. A Fast String Searching Algorithm. *Commun. ACM* **1977**, *20*, 762–772. [CrossRef]
57. Crochemore, M.; Hancart, C.; Lecroq, T. *Algorithms Strings*; Cambridge University Press: Cambridge, UK, 2007; Volume 9780521848992, pp. 1–383. [CrossRef]
58. Hakak, S.I.; Kamsin, A.; Shivakumara, P.; Gilkar, G.A.; Khan, W.Z.; Imran, M. Exact String Matching Algorithms: Survey, Issues, and Future Research Directions. *IEEE Access Pract. Innov. Open Solut.* **2019**, *7*, 69614–69637. [CrossRef]
59. Postiglione, A.; Monteleone, M. Towards Automatic Filing of Corpora. In Proceedings of the 18ème COLLOQUE INTERNATIONAL "Lexique et Grammaires Comparçs", Parco Scientifico e Tecnologico Di Salerno e Delle Aree Interne Della Campania, Salerno, Italy, 6–9 October 1999; pp. 1–9.
60. Elia, A.; Monteleone, M.; Postiglione, A. Cataloga: A Software for Semantic-Based Terminological Data Mining. In Proceedings of the 1st International Conference on Data Compression, Communication and Processing, Palinuro, Italy, 21–24 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 153–156. [CrossRef]
61. Elia, A.; Postiglione, A.; Monteleone, M.; Monti, J.; Guglielmo, D. CATALOGA: A Software for Semantic and Terminological Information Retrieval. In Proceedings of the ACM International Conference Proceeding Series, Wuhan, China, 13–14 August 2011; pp. 1–9. [CrossRef]
62. Postiglione, A.; Monteleone, M. Semantic-Based Bilingual Text-Mining. In Proceedings of the Second International Conference on Data Compression, Communication, Processing and Security (CCPS 2016), Cetara, Italy, 22–23 September 2016; pp. 1–4.
63. Hadi, M.U.; Tashi, Q.A.; Qureshi, R.; Shah, A.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. *Authorea Prepr.* **2023**. [CrossRef]
64. Min, B.; Ross, H.; Sulem, E.; Veyseh, A.P.B.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* **2023**, *56*, 1–40. [CrossRef]
65. Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Wong, D.F.; Chao, L.S. A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. *arXiv* **2023**, arXiv:cs.CL/2310.14724.
66. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 2017, pp. 5999–6009.
67. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
68. Silva Barbon, R.; Akabane, A.T. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors* **2022**, *22*, 8184. [CrossRef]
69. Onita, D. Active Learning Based on Transfer Learning Techniques for Text Classification. *IEEE Access* **2023**, *11*, 28751–28761. [CrossRef]
70. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186.
71. Rogers, A.; Kovaleva, O.; Rumshisky, A. A primer in bertology: What we know about how bert works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [CrossRef]
72. Kaliyar, R.K. A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of BERT. In Proceedings of the Confluence 2020—10th International Conference on Cloud Computing, Data Science and Engineering, Noida, India, 29–31 January 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; pp. 336–340. [CrossRef]
73. Xia, P.; Wu, S.; van Durme, B. Which *BERT? A survey organizing contextualized encoders. In Proceedings of the EMNLP 2020—2020 Conference on Empirical Methods in Natural Language Processing, Virtual, 16–20 November 2020; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2020; pp. 7516–7533.
74. Mohammed, A.H.; Ali, A.H. Survey of BERT (Bidirectional Encoder Representation Transformer) types. *J. Phys. Conf. Ser.* **2021**, *1963*, 012173. [CrossRef]
75. Aftan, S.; Shah, H. A Survey on BERT and Its Applications. In Proceedings of the 20th International Learning and Technology Conference, Jeddah, Saudi Arabia, 26 January 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 161–166. [CrossRef]
76. Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L. A Comprehensive Survey on Pretrained Foundation Models: A History from Bert to Chatgpt. *arXiv* **2023**, arXiv:2302.09419.
77. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *OpenAI Blog* **2018**, 1–12. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 3 February 2024).
78. Zhang, C.; Zhang, C.; Zheng, S.; Qiao, Y.; Li, C.; Zhang, M.; Dam, S.K.; Thwal, C.M.; Tun, Y.L.; Huy, L.L.; et al. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *arXiv* **2023**, arXiv:cs.AI/2303.11717.
79. Kalyan, K.S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2024**, *6*, 100048. [CrossRef]

80. Calzolari, N.; Fillmore, C.J.; Grishman, R.; Ide, N.; Lenci, A.; MacLeod, C.; Zampolli, A. Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas, Spain, 27 May–2 June 2002; pp. 1934–1940.
81. Sag, I.A.; Baldwin, T.; Bond, F.; Copestake, A.; Flickinger, D. Multiword Expressions: A Pain in the Neck for NLP. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **2002**, *2276*, 1–15. [CrossRef]
82. Constant, M.; Eryiğit, G.; Monti, J.; Van Der Plas, L.; Ramisch, C.; Rosner, M.; Todirascu, A. Multiword Expression Processing: A Survey. *Comput. Linguist.* **2017**, *43*, 837–892. [CrossRef]
83. Zhang, M.; Li, J. A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* **2021**, *1*, 831–833. [CrossRef]
84. Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C.A.; Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **2022**, *4*, 258–268. [CrossRef]
85. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 2020-December.
86. Huang, J.; Shao, H.; Chang, K.C.C. Are Large Pre-Trained Language Models Leaking Your Personal Information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2022; pp. 2038–2047.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.