Fuzzy Systems and Data Mining X
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241455

Some Brief Considerations on Computational Statistics Effectiveness and Appropriateness in Natural Language Processing Applications

Mario MONTELEONE [a] and Alberto POSTIGLIONE [b,1]

[a] *Department of Political Sciences and Communication (DiSPC), University of Salerno, ITALY*

[b] *Department of Business Sciences - Management & Innovation Systems (DISA-MIS), University of Salerno, ITALY*

ORCiD ID: Mario Monteleone https://orcid.org/0000-0002-9829-4275, Alberto Postiglione https://orcid.org/0000-0001-6411-6529

Abstract. This paper addresses the challenges of managing and processing unstructured or semi-structured text, particularly in the context of increasing data volumes that traditional linguistic databases and algorithms struggle to handle in real-time scenarios. While humans can easily navigate linguistic complexities, computational systems face significant difficulties due to algorithmic limitations and the shortcomings of Large Language Models (LLMs). These challenges often result in issues such as a lack of standardized formats, malformed expressions, semantic and lexical ambiguities, hallucinations, and failures to produce outputs aligned with the intricate meaning layers present in human language.

As for the automatic analysis of linguistic data, is well known that Natural Language Processing (NLP) uses two different approaches, coming from diverse cultural and experiential backgrounds. The first approach is based on probabilistic computational statistics (PCS), which underpins most Machine Learning (ML), LLMs, and Artificial Intelligence (AI) techniques. The second approach is based, for each specific language, on the formalization of morpho-syntactic features and constraints used by humans in ordinary communication activities. At first glance, the second approach appears more effective in addressing linguistic phenomena such as polysemy and the formation of meaningful distributional sequences or, more precisely, acceptable and grammatical morpho-syntactic contexts.

In this paper, we initiate a scientific discussion on the differences between these two approaches, aiming to shed light on their respective advantages and limitations.

Keywords. Natural Natural Language Processing, Rule-Based Natural Language Processing, Probabilistic Computational Statistics, Large Language Models, Statistical Natural Language Processing, Machine Learning, Formal Semantics.

[1] Corresponding Author: Alberto Postiglione, DISA-MIS, University of Salerno, ITALY, ap@unisa.it.

1. Foreword

This paper addresses the multifaceted challenges of processing and managing unstructured or semi-structured text, especially given the rapid increase in data volumes that traditional databases and algorithms struggle to handle effectively in real-time scenarios. Unlike structured data, which is systematically organized in databases or spreadsheets, unstructured and semi-structured data lack a consistent organizational framework.

Unstructured text data, which constitutes the vast majority of digital information, includes a wide range of documents such as books, scientific publications, news articles, web content, Word documents, and PDFs. This type of data lacks a predefined schema, making it challenging to process with conventional data analytics methods. On the other hand, semi-structured text contains elements of both structured and unstructured data. It often features irregular formats like XML or JSON, which complicates analysis due to variations in schema and content representation.

According to various estimates, over 80% of global data is unstructured or semi-structured [1,2]. The rapid advancements in Big Data technologies [3,4,5,6,7,8,9], alongside the growing adoption of the Internet of Things (IoT) [10,11,12,13,14,15,16] and Industrial IoT (IIoT) [17,18,19,20,21,22], have contributed to a massive explosion of textual data. The proliferation of interactive technologies like chatbots has further amplified this data surge [23,24,25,26]. In general, various types of data found on the web, such as emails, HTML and CSS files, social media posts, news, and blogs, are practically indestructible, rarely disappearing from the web. This data often presents significant redundancy and grows in a non-linear manner [27,28,29,30,31]. The volume of unstructured medical data is also increasing at an astonishing rate [32,33,34,35,36]. The research community contributes significantly to this growth, with a continuous rise in the number of scientific papers and the proliferation of accredited journals and conferences [37,38,39,40]. Furthermore, global governmental initiatives aimed at reducing paper-based documentation have bolstered this trend. Additionally, the Deep Web, which contains trillions of unindexed documents, significantly adds to the volume of text-based data.

The current landscape of unstructured and semi-structured data aligns with the defining characteristics of Big Data, notably in terms of volume, variety, lack of standardization, and continuous generation. Traditional databases and conventional algorithmic approaches are often inadequate to handle this vast and complex data, particularly in real-time processing scenarios. This challenge highlights the pressing need for advanced data management and analysis techniques, such as Natural Language Processing (NLP), to effectively store, manage, and extract value from this expanding data landscape. The key issues involve:

Lack of Standardized Formats The diversity of formats in unstructured and semi-structured text presents a significant challenge. Data sources range widely from user-generated content on social media and emails to technical documents and logs, each with its own unique format, syntax, and structure. This heterogeneity complicates the preprocessing and parsing stages, making it difficult to extract relevant information uniformly. Techniques such as data normalization, schema matching, and document classification are essential to address this variability, enabling a more consistent and standardized representation of the text.

Malformed Expressions and Noisy Data Text data frequently contains errors such as typos, grammatical mistakes, and incomplete sentences, especially in user-generated content. For instance, social media posts and product reviews often include informal language, slang, and emojis, which introduce noise into the data. This noise can significantly degrade the performance of NLP models, particularly those relying on rule-based methods or shallow machine learning techniques. Robust text-cleaning methods, such as spell-checking, grammatical corrections, and noise filtering, are necessary to enhance the quality of the input data for subsequent analysis.

Semantic and Lexical Ambiguities Semantic ambiguity occurs when a sentence or phrase can be interpreted in multiple ways. For example, the sentence "The chicken is ready to eat" is difficult to interpret (the chicken could be a meal to be eaten, or a live animal ready to eat), due to its ambiguous syntactic structure. Similarly, lexical ambiguity arises when a word has multiple meanings (polysemy) or when different words have similar pronunciations or spellings (homonymy). Addressing these ambiguities is critical for tasks such as word sense disambiguation (WSD), which aims to assign the correct meaning to a word based on its context, and named entity recognition (NER), which seeks to identify and classify proper nouns into predefined categories such as people, organizations, or locations.

Hallucinations An artificial hallucination occurs when an LLM-type AI, during the generation or summarization of a text, in relation to specific knowledge, randomly produces illogical, imprecise, contradictory or false information. AI hallucinations also occur during text generation, specifically when in terms information completeness and organicity, the texts produced do not respect the seven constitutive traits defined by De Beaugrande and Dressler [41] and fundamental for the recognition of a text as such. These traits are coherence, cohesion, intentionality, acceptability, informativeness, situationality, and intertextuality.

Complex Meaning Levels in Human Language Human language is inherently complex, encompassing multiple layers of meaning, including literal, figurative, and contextual interpretations. To understand this complexity, computational systems must capture not only surface-level syntax and semantics but also deeper, pragmatic meanings. Figures of speech such as metaphors, idioms, and sarcasm add an additional layer of difficulty, as their deeper meanings often diverge from their literal interpretations. Advanced NLP techniques, such as transformer-based models (e.g., BERT [42,43], GPT [44,45]), have shown promise in capturing contextual nuances by learning rich representations of words in different contexts, but they are far from having satisfactorily solved this challenge. Accurately interpreting figurative language and resolving deep semantic ambiguities thus remain open challenges in the field of NLP research.

Real-Time Processing of High-Volume Text Data With the exponential growth of data generation from sources like social media, sensor logs, and customer interactions, real-time processing of text data has become increasingly challenging. Traditional NLP models and database systems often fall short due to their limited scalability and inability to handle large datasets with low latency. Modern approaches leverage streaming data frameworks (e.g., Apache Kafka [46,47], Apache Flink [48]) along with real-time NLP pipelines, employing techniques like incremental parsing and online learning to efficiently process continuous data flows. Additionally,

cloud-based solutions and distributed architectures are utilized to manage the computational load, enabling scalable and responsive text processing systems.

Adopting a Semantic-Centered Approach Given these challenges, adopting a semantic-centered approach is essential for improving text processing and understanding. This approach emphasizes leveraging linguistic principles, such as identifying multi-word units (MWUs) and understanding their compositional meanings, to extract more accurate and semantically rich information from text. Finite automata ([49]) can be employed effectively to recognize specific patterns and MWUs within the text ([50,51,52]. Furthermore, integrating ontologies provides a structured framework for assigning knowledge domains and resolving ambiguities by linking textual elements to predefined concepts and relationships. This combination enhances the system's ability to capture the underlying meaning and context of the text, facilitating more effective natural language understanding and information retrieval.

While humans can effortlessly interpret and navigate these complexities using cognitive skills and contextual understanding, these tasks present significant challenges for computational systems.

As is known, there are currently two different approaches to Natural Language Processing (NLP), one based on the use of Computational Statistics (CS), on which Machine Learning (ML) is largely based; the other on the formalization of the morphosyntax of each language. In this study, we intend to analyse as objectively as possible the differences between these two approaches, as well as the strengths and weaknesses of both, as regards the structuring of the analysis and the application of tools and routines. Already in this preliminary phase of the study, it is evident that both the approaches we are about to investigate are based on the analysis of word combination, interconnection, governance and co-occurrence. This means, or at least should mean, that to perform NLP analyses, both approaches may make use of sets of formal rules and descriptions built on exhaustive descriptive linguistic resources. As we shall instead see, only the formalizing morphosyntax approach (which is the base for Formal Semantics (FS)) manifests and exploits this attention to the specific characteristics of natural language. On the contrary, CS mainly looks at words as simple sequences of signs delimited by two blank spaces and very often is unable to provide any precise and beyond question linguistic classification of them. That is, in its applications, CS mainly uses lists of words elaborated and tagged automatically, which often turn out to contain many inaccuracies. Therefore, we take the liberty of anticipating here that we are more inclined to place greater value on those linguistic resources made manually by qualified linguists, rather than on those elaborated by means of statistical or (supposed) algorithmic methods. To endorse such choice of ours, in the following page we will give specific examples, directly linked to the linguistic quality of any of the analytic procedures we will evaluate.

2. Machine-Learning vs. Grammar Engineering

As for the already mentioned two NLP methodologies, in Fig. 1 we give the most important pros and cons[2].

[2] *https://www.linkedin.com/pulse/pros-cons-two-approaches-machine-learning-grammar-engineering-wei-li/*, last accessed 2023/03/22.

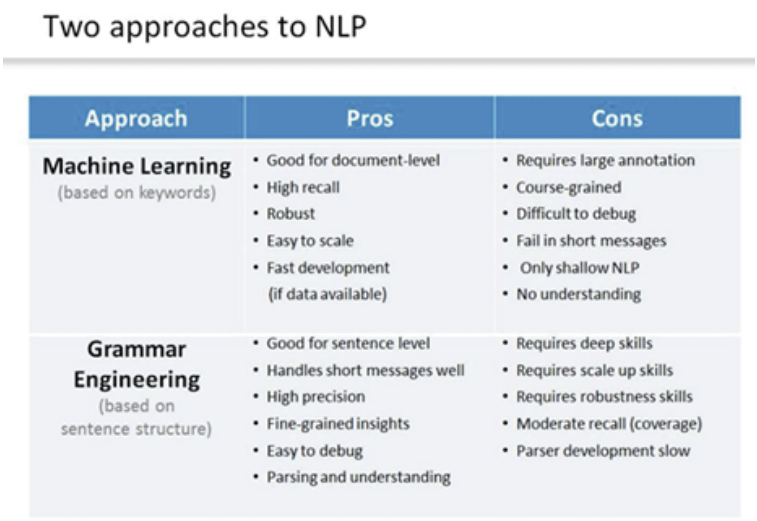

Figure 1. Two Approaches To NLP

These pros and cons help us deducing which is the most important methodological and functional difference between the two methods. Actually, ML is essentially statistical: it explores and processes linguistic data as Markov Chains [53]. This means that the concatenation of words is inferred on a probabilistic basis, i.e., without resorting to the application of specific formalized morphosyntactic rules. For this reason, ML is referred to as ruleless, or Statistical/Stochastic Natural Language Processing (SNLP). On the contrary, Grammar Engineering (GE) results to be [54]: "the creation of linguistically motivated electronic grammars is a key aspect of natural language processing (NLP). These grammars, developed within theoretical frameworks like the Lexicon NLP method by Maurice Gross [55,56,57,58], provide detailed descriptions of natural language. Initially focused on syntax, how sentence components relate, these grammars now also encompass functional structure (FS) information. This precise grammar engineering supports natural language understanding and generation, proving essential for applications such as textual entailment, dialogue systems, and machine translation".

For these reasons, GE is referred to as rule-based NLP. In the following pages, we will see how and how much the difference here outlined between SNLP and rule-based NLP, together with the adoption of one of the two methods, are decisive for the successful processing of linguistic data. Above all, we seldom may experience how the statistical/stochastic approach, in its non-specificity, often produces imprecise and non-reusable results, failing to reach deep and well-structured levels of linguistic analysis.

2.1. Statistical Natural Language Processing

In broad terms, Statistics (STAT) [59] can be defined as a set of scientific methods aimed at the quantitative and qualitative knowledge of collective phenomena through the collection, sorting, synthesis and data analysis. In other words, STAT is supposed to be a tool that translates information into knowledge. It studies collective phenomena (observation of a set of individual manifestations), in order to obtain information, describe a

phenomenon, and identify relationships. Hence, in a positive perspective, we could say that STAT has the advantage of being applicable to all domains from which it is possible to collect and store data. At the same time, but in a negative perspective, we could say that STAT is a non-specific discipline: although equipped with refined calculation tools, in fact STAT is applicable in the same way to very different sectors and domains, such as for example weather forecasts or the average kilos of apples consumed annually by the inhabitants of a given city. Among the various sets to which STAT can be applied, there are therefore also those that have already well identified and established their structural, productive, and iteratively-applicable rules, such as for instance morphology, syntax and sentence semantics.

As for SNLP [59], it aims to perform Statistical Inference (SI) for the field of natural language. SI in general consists of taking some data (generated in accordance with probability distribution) and then making some inferences about this distribution. In this sense, the current trend is to base SNLP on the use of Large Linguistic Models (LLMs)[60]. The results obtainable by means of this tool remain controversial: in fact, although they exponentially increase the quantity of "examples" on which to statistically base the analysis, at the same time they increase also all potential analysis errors. It also to consider the possibility of "manipulating" the responses obtainable from (alleged) Generative Artificial Intelligence systems based on LLMs. For example, we might look at lots of instances of prepositional phrase attachments in an English corpus and use them to try to predict, in general, prepositional phrase attachments for English. In this sense, the task of language modeling[3] is fundamental to speech or optical character recognition, and is used also for spelling correction, handwriting recognition, and statistical machine translation. At first glance, SI definition seems to encompass some controversial aspects, which may lead to possible methodological and applicative limits, especially when natural language is the object of the analysis. We will deal with these aspects in detail in the following pages. For now, we essentially want to highlight how natural language data are to be analysed necessarily distinguishing at least three descriptive levels phonetic-phonological, morphosyntactic, semantic of the sentence each of them having their specific usage rules and also having, among them, points of extreme contiguity, together with others of strong differentiation. Actually, it is not possible to define natural language as a completely heterogeneous or homogeneous set of items, nor can it be studied based on such an assumption. On the contrary, it is possible to state that in its entirety, the study of natural language seems to need the adoption of "quantum" inferential functions, which are extremely different from and more fine-grained than those offered by SI. As well, the previous definition of SNLP also calls into question, albeit indirectly, the already mentioned Markov Chains, in the point in which it refers to the classic task of language modeling, where the problem is to predict the next word given the previous words. However, what does not seem to be coped with in this definition is the fact that natural language can be studied as an autonomous element, i.e., by means of its specific idiosyncrasies. As is known, in each given language, correct word concatenations are not subject to probabilistic laws: there are specific linguistic use rules to validate and govern them. Such rules exist outside and before any text or linguistic dataset, in which these same rules must always find correct application[4]. Without these rules, no idiom could be

[3] Here, the problem is to predict the next word occurrence, given those of previous words.

[4] Yet, sometimes, linguistic datasets, including LLMs, offer imprecise readings and applications of these linguistic usage norms.

written, spoken, or understood. Therefore, the previous definition of SNLP leaves room to several doubts, which are:

(a) Is it correct that SNLP tries to model natural language without comprehending its internal dynamics in detail? I.e., is SNLP suitable concretely for language modelling? Can SNLP really produce inferences about natural language?
(b) Is SNLP not too dependent on the data it observes, while it ignores the dynamics of those data that escape its observation, or which it does not observe with the correct approach?
(c) Finally, natural language is by its nature ambiguous, that is, its concatenations of words often lend themselves to having more than one acceptable meaning. Is SNLP suitable to deal with natural language ambiguity? Is it sufficient for SNLP to analyse a vast amount of linguistic data[5] in order to produce a scientifically precise prospect of how natural language works? Should it not analyse all possible word combinations?

2.2. *Contextual Inference vs. Statistical Inference*

In the previous paragraph, regarding the definition of SNLP and, essentially, with reference to ML, we encountered the term "statistical inference", which lexically, semantically, as well as in relation to the study of logic, is opposed to that of "human inference". As is known, human inference is an associative procedure through which, starting from a certain premise or from the observation of a fact, our mind draws one or more consequences, or elaborates one or more judgments. However, the question about how our minds concretely make inferences remains yet unresolved. Along with the problem of accurately establishing the difference between simple perception and inference or cognition, it is not yet clear whether:

- our mind produces inferences via the "specialized modules" postulated by Fodor[6]; or
- inference is:
  - a behaviourist process, as argued by John Watson[7];
  - a reductionist procedure connected to the theory of identity, as maintained by Bronisaw Malinowski[8];
  - a connectionist and computational operation, as indicated by Hilary Putnam[9].

What is certain is that humans use inference to solve problems, as well as to make choices, and that often inferences may even occur within not completely "controlled levels" of the human brain (to simplify, we could say within unconscious or subconscious levels). Therefore, human inference appears to be a procedure classifiable as a crucial part of the problem-solving domain. Besides, what seems more plausible is that the inferential procedures of the human mind tend to be non-deterministic, i.e., not always following

[5] As it happens for instance today with Large Linguistic Models.
[6] $https://plato.stanford.edu/entries/modularity-mind/$.
[7] $https://www.ufrgs.br/psicoeduc/chasqueweb/edu01011/behaviorist-watson.pdf$.
[8] $https://anthropology.ua.edu/theory/functionalism/$.
[9] $https://plato.stanford.edu/entries/computational-mind/$.

the binary logic "0 vs. 1", nor that of the consequential variables of the "if ... then" type. Rather, it seems that our inferential procedures contemplate, at least in an initial state, multiple resolutions to a given problem, and that it then arrives at the solution, or choice, based on a computational process, which includes certain possibilities, and excludes others, until it finds the (hypothetically) right one. It would therefore look more like a quantum decision-making process, initially structured as a finite-state automaton, which tends to drop its paths when they are no longer useful. Always intuitively, we can say that inferential processes, including especially those related to natural language, are strictly dependent on the functioning of our cerebral or biological networks. These networks are made up of closely interconnected neurons, and among other, they allow us to reason, make calculations in parallel, recognize sounds, images and faces, or learn how to take specific actions. In practice, they actively allow our mind to "express its intelligence". Thanks to complex organizations of nerve cells that have different "tasks", such as perception of the environment, recognition of stimuli, and so on, a cerebral neural network works through the reception of data and signals internal and external as well as through sensory perceptions. These "become" information and knowledge through an impressive number of biological units, our neurons, which represent the computing capacity of our brain. Neurons interconnect with each other in a non-linear structure that responds to external data and stimuli. A cerebral neural network, therefore, presents itself as an "adaptive" system capable of modifying its structure, in response to both internal and external data, perceptions, experiences, and information[10]. Moreover,

"The brain consists of seven main networks at a high level: the sensorimotor system, visual system, limbic system, central executive network (CEN), default mode network (DMN), salience network, and dorsal attention network (DAN). While some of these networks can function independently, many complex cognitive functions arise from interactions among them. A more detailed examination reveals specific subnetworks dedicated to particular tasks, often involving components from multiple main networks. Notably, the Language network has significantly evolved since the Broca-Wernicke model, now incorporating a newly identified area, 55B, which is linked to muscle control during speech production[11]."

Hence, it seems almost certain that all natural language produced by human beings is the result of a series of inferences, supported by a specific cerebral neural subnetwork, which, in turn, interacts with other neural networks and subnetworks. It also takes into account and complies with the specific idiosyncrasies of the language in which a human being wants to express himself, i.e., its phonetics and phonology, as well as its morphosyntax and semantics. This procedure, that we may call Human Linguistic Inference (HLI), is therefore contextual: in each language, it produces acceptable utterances thanks to its interaction with all certified governance and co-occurrence relationships between words inside word groups, propositions, and sentences. Recalling the definition of SI given in the previous paragraph and comparing it to the one about human inference and contextual HLI, we can single out straightforwardly important differences, both methodological and structural.

[10] *https://www.biorxiv.org/content/10.1101/2021.07.15.452445v1.full*.
[11] *https://www.o8t.com/blog/brain-networks*.

Besides, it is worth stressing that among the tools and routines used in SI and SNLP, we find also Artificial Neural Transition Networks Transition (ANTN), more commonly called Artificial Neural Networks (ANNs). From the Web[12], we read that

"Artificial Neural Networks (ANNs) are computational models inspired by the human brain, using interconnected artificial neurons to simulate information processing. These neurons are connected through adjustable weights, allowing the network to learn by detecting patterns in data. ANNs can perform tasks such as classification, regression, and forecasting, and deep neural networks (DNNs), with multiple hidden layers, allow for more complex data processing. While neural networks have been around for years, advancements in hardware like GPUs have made them faster and more practical for large datasets. ANNs are now applied in a wide range of areas, including image processing, NLP, healthcare, and recommendation systems, demonstrating impressive performance in detecting patterns and making predictions".

From this (perhaps excessively) enthusiastic definition, it is possible to extract some points of discussion which, at first sight, arouse perplexity. The first point is the one associating the functioning of ANNs to that of cerebral or biological neural networks. At first glance, the association seems not only risky, but also devoid of effective evidence. In fact, we still know little about the functioning of the human brain, which we cannot yet fully study during its ordinary functioning. This is because there are still no adequate observation tools that can taxonomically describe its activities, also in terms of element and part interconnections. Besides, among humans ones, the brain is most delicate organ, i.e. it is not possible to carry out "open-brain" studies, as for example it is possible with the human heart. In http://www.thehighestofthemountains.com/images/thehighestofthemountains_brain_map_brainwithquotes28-125px.jpg you find a simplified description of human brain structure.

As for the previous definition on ANNs, a second perplexity arises from those passages in which sketches are given about how a biological neural network would work. These passages seem inspired by [61], in which we read that "Neurons are the basic building blocks of the nervous system, responsible for information processing and communication in animals. They consist of a centralized cell body and two types of processes axons and dendrites that connect to one another." Actually, a more convincing explanation on how biological neural networks work, and how they may help in predicting mammals behaviour, including humans, is available on Web[13]. It is supported by a 3D animation, and specifically testifies how "a group of Harvard University neuroscientists and Google engineers released the first wiring diagram of a piece of the human brain. The tissue, about the size of a pinhead, had been preserved, stained with heavy metals, cut into 5,000 slices and imaged under an electron microscope. This cubic millimetre of tissue accounts for only one-millionth of the entire human brain. Yet the vast trove of data depicting it com prises 1.4 petabytes worth of brightly coloured microscopy images of nerve cells, blood vessels and more". Also, it states that "this approach is leading to impressive progress in understanding these animals". No mention is made regarding the possibility of replicating, even summarily, the functions of this wiring diagram. In practice, no comparison is made with the functionality of ANNs, either directly or indirectly.

Considering all this, one therefore wonders what concrete relationship there may be, in terms of similarities, between biological networks and ANNs. A specific answer in

[12] *https://gemmo.ai/what-is-an-artificial-neural-network/*

[13] *https://www.quantamagazine.org/new-brain-maps-can-predict-behaviors-*20211206/

this regard comes from Web[14], in which it is stated that ANN computational power is still much lower than that of the human brain, as well as more consuming in terms of power, as "the number of neurons in our brain is about 86 billion." Currently, the largest artificial neural networks, built on supercomputers, have the size of a frog brain (about 16 million neurons). Besides, the actual largest artificial neural network is supposed to include 160B parameters, where a parameter roughly corresponds to a synapse in the human brain. Given the estimation that the human brain has about 100T synapses, this largest artificial neural network could be said to be about 0.16% of the human brain. As for sustainability, "research also suggests that the power consumption by biological neural networks is around 20W whereas by artificial neural networks is around 300W." On these considerations, and with such a reduced potential, it is difficult to imagine how an ANN can correctly mimic, among other, the linguistic capabilities of the human brain. Therefore, we take here the liberty of stating that if ANNs are among the basis of ML and AI, then these disciplines are still concretely groping in the dark and will continue to do so for a long time. All this clearly arises from the lack of attention that STAT dedicates to the management of the idiosyncrasies of natural language, starting from the (erroneous) assumption that it is possible to replace the precise dynamics of natural language with the inaccurate ones of SI.

3. Conclusions

This paper delves into the inherent challenges of processing natural language text, which is often found in unstructured or semi-structured formats. Traditional database systems and algorithms struggle to handle such data effectively, particularly in scenarios requiring rapid, real-time analysis. The variability in language usage, influenced by diverse authorship, cultural nuances, and different writing styles, introduces significant ambiguity. Lexical and semantic variations complicate the interpretation process, making it difficult for computational systems to consistently extract meaningful insights. Unlike humans, who can intuitively understand the context and subtleties of language, machines lack the external or sensory cues needed to accurately interpret semantic content.

Much more could and should be said about the topics we have tried to address in this paper. Above all, we would have liked to demonstrate how to overcome all the STAT-approach inaccuracies highlighted, and how to solve the problems they create using NLP rule-based methods and software of grammar engineering linguistic analysis. In this specific analysis, for the sake of brevity, this was not possible, but it will certainly be the subject of future publications, which we will take care of producing with all the details necessary to support our theses.

We advocate for the integration of computational methodologies with linguistic approaches [50,62,63,64], utilizing tools like CATALOGA [65,66], AUTOMETA [22,52,67], and Nooj [68,69,70]. Linguistic-based computational techniques (refer to examples such as [71,72,73,74,75,76]) are designed to extract semantically relevant information from text. They focus on processing, transforming, and analyzing textual data to uncover patterns, entities, relationships, and insights, effectively converting unstructured text into structured, analyzable formats. This strategy leverages Natural Language Pro-

[14] $https://medium.com/@eraiitk/brain-and-artificial-neural-networks-differences-and-similarities-1d337fe50168$

cessing (NLP) techniques rooted in linguistic theory, employing methods like rule-based linguistic processing, text annotation, and semantic analysis. Key methods in this approach include discourse analysis, syntactic parsing, part-of-speech tagging, sentiment analysis, named entity recognition, and knowledge extraction. These techniques are essential for deriving structured insights from raw text data, supporting enhanced analysis and informed decision-making.

References

[1] Chengqing Zong, Rui Xia, and Jiajun Zhang. *Text Data Mining*. Springer Singapore, 2021. Cited by: 23.

[2] Sayali Sunil Tandel, Abhishek Jamadar, and Siddharth Dudugu. A survey on text mining techniques. In *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, pages 1022–1026. Institute of Electrical and Electronics Engineers Inc., 2019. Cited by: 31.

[3] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014. Cited by: 2252.

[4] C.L. Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275:314–347, 2014. Cited by: 2208.

[5] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos. Big data analytics: A survey. *Journal of Big Data*, 2(1), 2015. Cited by: 546.

[6] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4):431–448, 2018. Cited by: 581.

[7] Amina Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1), 2021. Cited by: 80.

[8] Liang Zhao. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys*, 54(5), 2022. Cited by: 63.

[9] Haolan Zhang, Sanghyuk Lee, Yifan Lu, Xin Yu, and Huanda Lu. A survey on big data technologies and their applications to the metaverse: Past, current and future. *Mathematics*, 11(1), 2023.

[10] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787–2805, 2010. Cited by: 10936.

[11] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang. Data mining for internet of things: A survey. *IEEE Communications Surveys and Tutorials*, 16(1):77–97, 2014. Cited by: 496.

[12] Ala Al-Fuqaha and al. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys and Tutorials*, 17(4):2347–2376, 2015. Cited by: 5626.

[13] Li Da Xu, Wu He, and Shancang Li. Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243, 2014. Cited by: 3715.

[14] Lalit Chettri and Rabindranath Bera. A comprehensive survey on internet of things (IoT) toward 5G wireless systems. *IEEE Internet of Things Journal*, 7(1):16–32, 2020. Cited by: 1174.

[15] Dinh C. Nguyen and al. 6G internet of things: A comprehensive survey. *IEEE Internet of Things Journal*, 9(1):359–383, 2022. Cited by: 587.

[16] Ibrahim Kahraman, Alper Kose, Mutlu Koca, and Emin Anarim. Age of information in internet of things: A survey. *IEEE Internet of Things Journal*, 11(6):9896–9914, 2024.

[17] Emiliano Sisinni, Abusayeed Saifullah, Song Han, Ulf Jennehag, and Mikael Gidlund. Industrial internet of things: Challenges, opportunities, and directions. *IEEE Transactions on Industrial Informatics*, 14(11):4724–4734, 2018. Cited by: 1317.

[18] Hugh Boyes, Bil Hallaq, Joe Cunningham, and Tim Watson. The industrial internet of things (IIoT): An analysis framework. *Computers in Industry*, 101:1–12, 2018. Cited by: 824.

[19] Diego G.S. Pivoto and al. Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review. *Journal of Manufacturing Systems*, 58:176–192, 2021. Cited by: 285.

[20] Yujiao Hu and al. Industrial internet of things intelligence empowering smart manufacturing: A literature review. *IEEE Internet of Things Journal*, 11(11):19143–19167, 2024.

[21] Xiaoshao Mu and Maxwell Fordjour Antwi-Afari. The applications of Internet of Things (IoT) in industrial management: A science mapping review. *International Journal of Production Research*, 62(5):1928–1952, 2024. Cited by: 20.

[22] Alberto Postiglione and Mario Monteleone. Predictive maintenance with linguistic text mining. *Mathematics*, 12(7):1–18, paper num. 1089, 2024.

[23] Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? A survey on social characteristics in Human–Chatbot interaction design. *International Journal of Human-Computer Interaction*, 37(8):729–758, 2021. Cited by: 156.

[24] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human Computer Studies*, 151, 2021. Cited by: 152.

[25] Bei Luo, Raymond Y. K. Lau, Chunping Li, and Yain-Whar Si. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1), 2022. Cited by: 61.

[26] Giuseppe Fenza and al. Healthcare conversational agents: Chatbot forimproving patient-reported outcomes. *Lecture Notes in Networks and Systems*, 661 LNNS:137 – 148, 2023.

[27] Veena Jose, V.P. Jagathy Raj, and Shine K. George. Ontology-based information extraction framework for academic knowledge repository. *Advances in Intelligent Systems and Computing*, 1184:73–80, 2021.

[28] Momin Saniya Parvez and al. Analysis of different web data extraction techniques. In *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018*. Institute of Electrical and Electronics Engineers Inc., 2018. Cited by: 21.

[29] Alberto Postiglione and Giustino De Bueriis. On Web's contact structure. *Journal of Ambient Intelligence and Humanized Computing*, 10(7):2829–2841, 2019.

[30] Mosa Salah, Basem Al Okush, and Mustafa Al Rifaee. A comparison of web data extraction techniques. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*, pages 785–789. Institute of Electrical and Electronics Engineers Inc., 2019.

[31] Vlad Krotov and Leigh Johnson. Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons*, 66(4):481–491, 2023. Cited by: 12.

[32] Jose A. Reyes-Ortiz, Beatriz A. Gonzalez-Beltran, and Lizbeth Gallardo-Lopez. Clinical decision support systems: A survey of NLP-based approaches from unstructured data. In Spies M., Wagner R.R., and Tjoa A.M., editors, *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, volume 2016-February, pages 163–167. Institute of Electrical and Electronics Engineers Inc., 2016. Cited by: 21.

[33] Sumati Boyapati, Srinivasa Rao Swarna, Vishal Dutt, and Narayan Vyas. Big data approach for medical data classification: A review study. In *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, pages 762–766. Institute of Electrical and Electronics Engineers Inc., 2020. Cited by: 14.

[34] Irene Li and al. Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46, 2022. Cited by: 70.

[35] V. Giannella and al. Neural networks for fatigue crack propagation predictions in real-time under uncertainty. *Computers and Structures*, 288:1–12, paper num. 107157, 2023.

[36] Yixuan Qiu, Feng Lin, Weitong Chen, and Miao Xu. Pre-training in medical data: A survey. *Machine Intelligence Research*, 20(2):147–179, 2023. Cited by: 10.

[37] Kavitha Jayaram and K. Sangeeta. A review: Information extraction techniques from research papers. In *IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017 - Proceedings*, pages 56–59. Institute of Electrical and Electronics Engineers Inc., 2017. Cited by: 20.

[38] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. Using text mining techniques for extracting information from research articles. *Studies in Computational Intelligence*, 740:373–397, 2018. Cited by: 130.

[39] Gohar Zaman, Hairulnizam Mahdin, Khalid Hussain, and Atta-Ur-Rahman. Information extraction from semi and unstructured data sources: A systematic literature review. *ICIC Express Letters*, 14(6):593–603, 2020. Cited by: 21.

[40] Murtaza Ashiq, Muhammad Haroon Usmani, and Muhammad Naeem. A systematic literature review on research data management practices and services. *Global Knowledge, Memory and Communication*, 71(8-9):649–671, 2022. Cited by: 33.

[41] Robert-Alain De Beaugrande and Wolfgang U Dressler. *Introduction to Text Linguistics*, volume 1. longman London, 1981.

[42] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021. Cited by: 288.

[43] Sulaiman Aftan and Habib Shah. A survey on BERT and its applications. In Sarirete A., Balfagih Z., Brahimi T., El-Amin Mousa M.F., and Elkafrawy P.M., editors, *20th International Learning and Technology Conference, L and T 2023*, pages 161–166. Institute of Electrical and Electronics Engineers Inc., 2023. Cited by: 14.

[44] Gokul Yenduri and al. GPT (generative pre-trained transformer) - a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE access : practical innovations, open solutions*, 12:54608–54649, 2024. Cited by: 27.

[45] Jack Gallifant and al. Peer review of GPT-4 technical report and systems card. *PLOS Digital Health*, 3(1 January), 2024. Cited by: 14.

[46] Guozhang Wang and al. Building a replicated logging system with apache kafka. *Proceedings of the VLDB Endowment*, 8(12 12):1654–1655, 2015. Cited by: 121.

[47] Theofanis P. Raptis and Andrea Passarella. A survey on networked data streaming with apache kafka. *IEEE access : practical innovations, open solutions*, 11:85333–85350, 2023. Cited by: 18.

[48] Paris Carboney, Stephan Ewenz, Gyula Fóra, Seif Haridi, Stefan Richter, and Kostas Tzoumas. State management in Apache Flink: ® consistent stateful distributed stream processing. *Proceedings of the VLDB Endowment*, 10(12):1718–1729, 2017. Cited by: 177.

[49] Jacques Sakarovitch and Reuben Thomas. *Elements of Automata Theory*. Cambridge University Press, 2011. Cited by: 432.

[50] Maurice Gross. The use of finite automata in the lexical representation of natural language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 377 LNCS:34–50, 1989. Cited by: 25.

[51] Sudersan Behera. Implementation of a finite state automaton to recognize and remove stop words in english text on its retrieval. In *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, pages 476–480. Institute of Electrical and Electronics Engineers Inc., 2018. Cited by: 10.

[52] Alberto Postiglione. Text mining with finite state automata via compound words ontologies. *Lecture Notes on Data Engineering and Communications Technologies*, 193:194–205, 2024.

[53] Paul A Gagniuc. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons, 2017.

[54] D. Duchier and Y. Parmentier. High-level methodologies for grammar engineering, introduction to the special issue. *Journal of Language Modelling*, 3(1):5–19, 2015.

[55] Zellig Harris. Grammar on mathematical principles. *Journal of Linguistics*, 14(1):1–20, 1978.

[56] Maurice Gross. Observations on semantic theories. *Theoretical Linguistics*, 5(1-3):1–18, 1978.

[57] Maurice Gross. Lexicon-grammar and the syntactic analysis of French. In *10th International Conference on Computational Linguistics, COLING 1984 and 22nd Annual Meeting of the Association for Computational Linguistics, ACL 1984*, pages 275–282. Association for Computational Linguistics (ACL), 1984. Cited by: 26.

[58] Mario Monteleone. Zellig S. Harris' transfer grammar and its application with NooJ. *Communications in Computer and Information Science*, 1758 CCIS:65–75, 2022.

[59] Christopher D Manning. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[60] Ashish Vaswani and al. Attention is all you need. In Guyon I., Fergus R., and et al. H., editors, *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5999–6009. Neural information processing systems foundation, 2017. Cited by: 46600.

[61] Paheli Desai-Chowdhry, Alexander Brummer, and Van Savage. How axon and dendrite branching are governed by time, energy, and spatial constraints. *bioRxiv : the preprint server for biology*, 2021.

[62] Mario Monteleone. NooJ local grammars for endophora resolution. *Communications in Computer and Information Science*, 667:182–195, 2016.

[63] Maurice Gross. A few analogies with computing. *Behavioral and Brain Sciences*, 6(3):407–408, 1983.

[64] Maurice Gross. The construction of electronic dictionaries; [La construction de dictionnaires électroniques]. *Annales Des Télécommunications*, 44(1-2):4–19, 1989. Cited by: 21.

[65] Annibale Elia, Mario Monteleone, and Alberto Postiglione. Cataloga: A software for semantic-based

terminological data mining. In *1st International Conference on Data Compression, Communication and Processing,IEEE,Palinuro(SA),June 21-24*, pages 153–156. IEEE Computer Society, 2011.

[66] Annibale Elia, Alberto Postiglione, Mario Monteleone, Johanna Monti, and Daniela Guglielmo. CATALOGA©: A software for semantic and terminological information retrieval. In *ACM International Conference Proceeding Series*, pages 1–9, 2011.

[67] Alberto Postiglione. Finite state automata on multi-word units for efficient text-mining . *Mathematics*, 12(4):1–20, paper num. 506, 2024.

[68] Mario Monteleone. NooJ grammars for morphophonemic continuity and semantic discontinuity. *Communications in Computer and Information Science*, 1816 CCIS:139–149, 2024.

[69] Mario Monteleone. NooJ grammars and ethical algorithms: Tackling on-line hate speech. *Communications in Computer and Information Science*, 987:180–191, 2019.

[70] Mario Monteleone. NooJ for artificial intelligence: An anthropic approach. *Communications in Computer and Information Science*, 1389:173–184, 2021.

[71] Jan Dědek, Peter Vojtáš, and Marta Vomlelová. Fuzzy ILP Classification of web reports after linguistic text mining. *Information Processing and Management*, 48(3):438–450, 2012.

[72] Yang-Cheng Lu, Chung-Hua Shen, and Yu-Chen Wei. Revisiting early warning signals of corporate credit default using linguistic analysis. *Pacific Basin Finance Journal*, 24:1–21, 2013. Cited by: 28.

[73] Frederik Cailliau and Ariane Cavet. Mining automatic speech transcripts for the retrieval of problematic calls. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7817 LNCS(PART 2):83–95, 2013.

[74] Francisco Villarroel Ordenes, Babis Theodoulidis, Jamie Burton, Thorsten Gruber, and Mohamed Zaki. Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research*, 17(3):278–295, 2014. Cited by: 139.

[75] R. Jindal and S. Taneja. A novel weighted classification approach using linguistic text mining. *International journal of computer applications*, 180(2):9–15, 2017.

[76] Tobias Nießner, Daniel H. Gross, and Matthias Schumann. Evidential strategies in financial statement analysis: A corpus linguistic text mining approach to bankruptcy prediction. *Journal of Risk and Financial Management*, 15(10), 2022.