UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE, ECONOMICHE E SOCIALI

Master's Degree Course in Data Science and Economics

HARVESTING KNOWLEDGE: ADVANCED NLP AND TEXT MINING TECHNIQUES FOR CAUSAL INFERENCE IN SUSTAINABLE AGRICULTURE RESEARCH

Supervisor: Prof Alfio Ferrara

Co-supervisors: Prof Stefano Montanelli,

Thesis by:
Aleksandra Katarzyna Łepek
Student ID: 946612

Academic Year 2023-2024

"Apteka polowa duszy?[...] - Zwycięstwo."

— *Friedrich Nietzsche*

Acknowledgments

Without your support, I wouldn't have been able to take up this challenge. Thank you, dear Mum and Dad, my husband Michele, and Federico.

Contents

Abstract

In the context of agricultural research, the vast availability of scientific data presents a unique opportunity to drive informed decision-making about future policies and promote sustainable farming practices. However, the sheer volume of literature makes it challenging to extract meaningful insights efficiently. This is where text mining and natural language processing (NLP) techniques become invaluable, providing powerful tools to identify key patterns, associations, and causal relationships within the overwhelming body of research.

The aim of this study is to analyze scientific abstracts in agronomy research to uncover associations between tillage practices, namely minimum tillage and no tillage and changes in Soil Organic Carbon (SOC) content. By applying text mining techniques and transformer-based models, this research seeks to gain a deeper understanding of causal relationships between conservation tillage and SOC dynamics. The study further evaluates the performance of different models, assessing their ability to extract and classify these relationships accurately.

Chapter 1

Introduction

Every day, the sheer volume of data created around the world presents new challenges in not only processing and storing it but also in making it accessible to those who could benefit from it. This explosion of information has sparked significant interest in leveraging this data to forecast and understand trends in different domains. In the context of agriculture research, the availability of enormous amounts of data opens up a unique opportunity to drive informed decisions and implement sustainable agricultural practices. Moreover, keeping up with the current state of scientific knowledge and emerging advancements is becoming increasingly hard. In 2022 alone, 3.3 million articles were published in the fields of Science and Engineering [1].

With such a vast volume of research, navigating through the complexities of these findings and understanding their implications often seems like a superhuman task. Here, text mining and natural language processing (NLP) techniques become invaluable, offering promising solutions for extracting key insights and causal relationships from the overwhelming body of scientific literature. By applying these technologies to analyze agricultural research, we are presented with transformative opportunities to uncover what has worked and what hasn't, ultimately contributing to more sustainable agricultural practices and a positive impact on the global community.

Agriculture employs estimated 1.23 billion people globally, which in 2024 constituted 15% of the world's population or roughly a third of the global workforce[2]. Soil is at the center of this effort, as an estimated 95% of food production relies on it directly (e.g., crops) and indirectly (e.g., livestock).[3] Soil serves as the foundation of food production. However, beyond sustaining human life by supporting the growth of edible plants, it also performs other essential ecological functions:

- regulating water (eg. water retention and aquifer replenishment, absorbing the rainfall water and preventing flooding) ,
- mitigating pollutants (eg. carbon dioxide sequestration , allowing to neutralize one of the most important green house gases as being the largest terrestrial
- cycling nutrients (e.g decomposition of animals and plants, recycling phosphorus, carbon, potassium and nitrogen)
- biodiversity host (e.g allowing organisms that constitute 25% of our planet biodiversity to thrive) [4]

As a non-renewable and irreplaceable resource (taking between 200-400 years to create 1 cm), it is critical to understand the ways in which we can protect it for us and more importantly for the

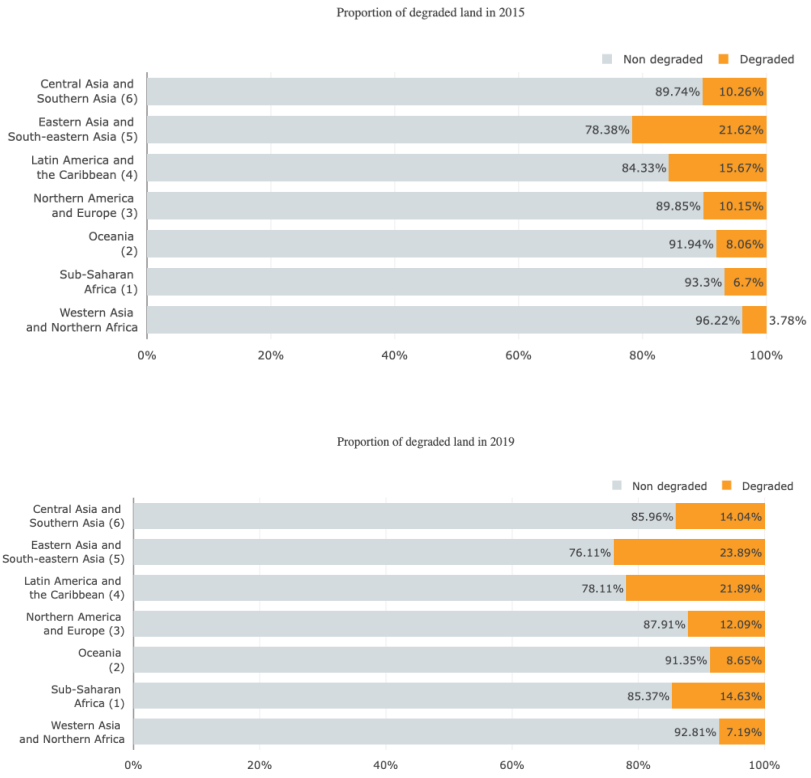

Figure 1: The change of land degradation between 2015 and 2019 across different regions. The data highlights a concerning trend particularly in Eastern Asia, Sub-Saharan Africa, and Central Asia. This visualization demonstrates the global scope of soil degradation challenges. Source: [6]

future generations. We are already facing an alarming situation, with estimated 52% of soil already being degraded and predicted 90% by 2050, if we do not take preventive actions. [5]

According to United Nations data, between 2015 and 2019, approximately 100 million hectares of healthy, productive land were lost (Figure 1), n area twice the size of Greenland, affecting 1.3 billion people. [6]. The criticality of soil as a basic resource is also reflected in the Sustainable Development Goals: 15. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, stop and reverse land degradation, and stop biodiversity loss [7].

Over the course of 4 years we have lost almost 2% of productive land in Europe and nearly 6% in Sub-Saharan Africa. The global effort to slow down land degradation can be seen through the lens of different projects. One of the most considerable and ambitious one, with a budget of 16.2 billion €, is the United Nations Convention to Combat Desertification initiative. Great Green Wall, with the objective of restoring the productivity of the 8000 km long land in the Sahel, situated on the border with the Sahara Desert. The project ambition is "to restore 100 million hectares of currently degraded land; sequester 250 million tons of carbon and create 10 million green jobs by 2030"[8] The progress that has been made can be seen on the Figure 2.

More than a third of the European Union yearly budget is spent on agriculture related activities.

Figure 2: In Bourgerba/Mauritania the soil was degraded and completely barren. Rehabilitation through dikes, half-moons and soil bunds has changed the landscape, rendering 51 ha of the land productive and fertile. Source: WFP Regional Bureau for Western Africa [9]

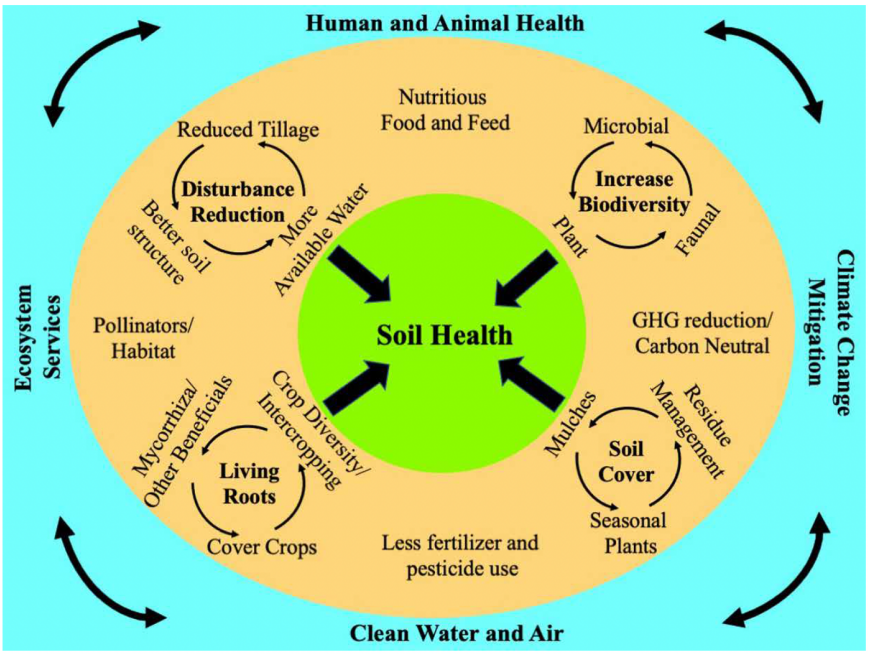

Figure 3: Diagram illustrating the interconnected aspects of soil health, showing how various biological, chemical, and physical properties contribute to overall soil functionality and agricultural productivity [12]

Common Agricultural Policy (CAP) that supports farmers is the longest-serving set of laws that aims to provide high-quality, sustainably produced, and affordable food. Each European spends 33 cents per day to support this 380 billion comprehensive policy. From 2025 around 40% of these funds will be dedicated to environmentally respectful farming. [10] Multiple efforts are on the way to ensure that citizens in EU member states have a supply of nutritious food that is produced in a sustainable way from healthy soil. Notably EU Soil Strategy for 2030 states: "As a key solution, healthy soils contribute to address our big challenges of achieving climate neutrality and becoming resilient to climate change, developing a clean and circular (bio)economy, reversing biodiversity loss, safeguarding human health, halting desertification and reversing land degradation". [11]

Soil health as a term was introduced following the green revolution in 1960s and gained traction in the 1990s, replacing terms of soil fertility and soil quality, as it encompasses a more complex function that soil plays in our lives. [12]

U.S. Department of Agriculture [13] defines main principles of soil health management as follows:

- maximizing plant roots
- minimize disturbance
- maximize soil cover
- maximize diversity

"By definition, a healthy soil will have a positive influence on human and animal health through nutritious food and feed, clean water and air through filtering and buffering of pollutants, benefiting

pollinators and habitat in general and, finally, mitigating climate change by reducing GHG emissions and through soil carbon sequestration (SOC). Whilst some of these aspects of soil health have been considered for millennia, it has only been recently understood that soil health is optimized if all these management principles are consistently practiced together. " [12]

In Europe more than 60% of soil is considered unhealthy, in the report accompanying the proposal the Directive on Soil Monitoring and Resilience lists loss of soil organic carbon as one of the main root causes of soil degradation,

Chemical, physical and biological soil health indicators are influenced by the level of soil C [12] and the most traditional way to improve soil C is the use of cover crops or the addition of animal manure, compost or other organic waste. However, presently, reducing soil disturbance is getting more traction as minimal till and no till practice allow for better soil structure, which in turn improves soil water holding capacities (see Figure. 3). Conservation tillage or zero tillage positively influences soil water availability, as well as minimize erosional losses, it also has the potential to improve the content of soil organic carbon (SOC).

The purpose of this work is to analyze the abstracts of scientific papers in the field of agronomy research papers to find an association with the tillage practices and the change in the organic carbon content of soil (SOC). Enhancing traditional text mining techniques like Entity Recognition and Pattern Matching and compare it with advanced NLP methods like Transformers based GPT-3 and GPT-4 models.

1.1 Research Question

The thesis addresses the following key research questions:

- How can existing traditional text mining techniques be optimized to better identify causal relationships between conservation agriculture practices (specifically, no-tillage and minimum tillage) and changes in soil organic carbon content?
- To what extent do modern large language models, such as GPT-3 and GPT-4, improve the extraction of research findings on sustainable agriculture, and how do they compare to established NLP techniques in terms of performance and interpretability?
- Among the diverse approaches to NLP and text mining, ranging from conventional rule-based pipelines to cutting-edge Transformer-based architectures, which yield the most accurate detection of causal relationships in agricultural studies?

1.2 Thesis structure

The thesis addresses the following key research questions:

- Chapter 2 reviews the existing literature on traditional text mining methods, advanced NLP models (including Transformers), and causal inference mining in scientific literature, with a particular focus on agriculture.
- Chapter 3 describes the methodology, detailing the data collection process, the steps taken for preprocessing and visual representation of the dataset.
- Chapter 4 shows an analysis of the dataset, highlighting key characteristics and trends in the agriculture research through visualizations to provide a comprehensive overview

- Chapter 5 presents the results obtained from both the traditional and advanced text mining techniques, highlighting their relative strengths and limitations.
- Chapter 6 concludes with a comparative discussion of these methods' performance, the implications for sustainable agriculture, and suggestions for future research directions.

Chapter 2

State of the Art

With 5.3 billion internet users - roughly 66% of the global population [14] - the volume of data that is created and shared each day is astounding. This continuous influx of new information presents substantial challenges not only for storing and organizing content, but also for making it truly usable to anyone seeking actionable insights. At the same time, it opens up exciting opportunities for leveraging these data to forecast trends and understand behaviors across diverse domains, from business and finance to health and social sciences.

In particular, *Natural Language Processing (NLP)* stands out among the key technologies that aim to transform large-scale textual data into structured formats that can be analyzed, interpreted and eventually put to work. NLP techniques - from classification and sentiment analysis to sophisticated information extraction - have seen remarkable progress, largely due to advances in machine learning and the increased capabilities of the computational infrastructure. Currently, one of the most sought-after, widely discussed, and heavily hyped technologies is Large Language Models (LLMs), such as GPT-3 and GPT-4, which have revolutionized how we approach NLP tasks. These models, trained on massive datasets and with transformer-based architectures, can generate human-like text, understand context across long passages, and even perform specific tasks such as text summarization or question answering with minimal training.

2.1 Text Mining

We live in a world that constantly produces fresh records of text. The advancement of hardware and software technology for storing data has been accompanied by the increasing popularity of social networks and easy access to the Internet. This has created an unprecedented number of various types of textual data repositories. [15] The need to uncover insights from these massive textual repositories has long been recognized, and text mining is a powerful tool that helps to achieve this objective. "Text mining is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text. This encompasses everything from information retrieval (i.e., document or web site retrieval) to text classification and clustering, to (somewhat more recently) entity, relation, and event extraction". [16]

2.1.1 Named Entity Recognition and Part-of Speech Tagging

Named Entity Recognition (NER) and Part of the Speech Tagging (POS) are classical problem in information extraction. Former aims to identify words and phrases within text that can be

categorized as entities, such as people, locations, organizations, dates, and numbers. Latter, on the other hand, adds linguistics annotations for each word.

TEXT	Part-of-speech	DEP	NER	NER Description
Apple	PROPN	nsubj	ORG	Companies, agencies, institutions.
is	AUX	aux		
looking	VERB	ROOT		
at	ADP	prep		
buying	VERB	pcomp		
U.K.	PROPN	compound	GPE	Geopolitical entity, i.e., countries, cities, states.
startup	NOUN	dobj		
for	ADP	prep		
$	SYM	quantmod	MONEY	Monetary values, including unit.
1	NUM	compound		
billion	NUM	pobj		

Table 1: Example of Part-of-speech tagging, dependency parsing, and named entity recognition based on the simple sentence. POS classification highlights syntactic categories, dependency parsing establishes grammatical relations, while NER identifies entities like organizations, geopolitical locations, and monetary values. Source:[17]

"All linguistic theories assume that words can be classified by a word class or part of speech (POS) according to their behavior within the language system".[18] Parts of Speech tagging, using English as an example, gives valuable information not only if the word is one of the common eight parts of speech like adjective, adverb, noun, verb, etc. (see Table 2), but most importantly they show how it relates to other words in particular sentence (see Figure 4). This helps extracting sentence structure and meaning. Widely used Universal Dependecies (UD) framework provides morphosyntactic annotation for more than 100 languages and establishes 17 general classes of words and other elements of text and assignes them labels. Accuracy of POS tagging task is close to 97% regardless of the algorithm used and is roughly the same to human performance. [19]

POS can accurately identify proper nouns, as illustrated in the example below: 'Apple' and 'U.K.' However, it does not indicate the types of entities they refer to—'Apple' as an organization and 'U.K.' as a geo-political entity. This task is performed by Named Entity Recognition, which introduces the problem of segmentation of the sentence and ambiguity of words. Unlike the POS which assigns each word its label, NER needs to identify entities than can span for more than one word. Additionally, depending on the context the proper noun can have different meaning eg. Washington as a person (historical Figure) and as a location (city). [17]

Entity recognition, introduced at the Sixth Message Understanding Conference (MUC-6), initially relied on rule-based approaches using manually crafted lexicons and dictionaries. While effective in specific domains, these approaches did not scale well to new contexts. Since its introduction, work on this topic has been central to language processing, as it lays the foundation for tasks like sentiment analysis,biomedical natural language processing, and machine translation, to name a few.

Over the years,Interest in improving NER has remained strong, and the field continues to evolve. Early applications of machine learning in NER used supervised approaches like Hidden Markov Models (HMM) and Conditional Random Fields (CRF). Over time, research evolved to incorporate Recurrent Neural Networks and Transformer-based architectures. [20]

Traditional POS	UPOS Category
noun	common noun
propn	proper noun
verb	main verb
aux	auxiliary verb or other tense, aspect, or mood particle
adjective	adj
det	determiner (including article)
num	numeral (cardinal)
adverb	adv
pronoun	pron
preposition	adp (adposition, preposition/postposition)
conjunction	cconj coordinating conjunctions
conj	subordinating conjunction
interjection	intj
–	part (particle, special single word markers in some languages)
–	x (other, e.g., words in foreign language expressions)
–	sym (non-punctuation symbol, e.g., a hash (#) or emoji)
–	punct (punctuation)

Table 2: Mapping of Traditional Part-of-Speech (POS) Tags to Universal POS (UPOS) Categories used in Natural Language Processing. Source: [18]

2.1.2 Rule-Based Approach for Named Entity Recognition

Rule-based methods for Named Entity Recognition (NER) rely on predefined or automatically learned sets of rules to identify entities within text. Each token in the text is represented by a set of features, such as the token itself, its part-of-speech tag, or orthographic characteristics (e.g., capitalization or numeric values). A rule consists of a pattern and an associated action.[15] Patterns are often expressed as regular expressions that define conditions over token features, while actions determine how matched token sequences should be labeled. For example, a rule could label a sequence of tokens such as "Dr. Young," where "Young" is a capitalized word, as a person entity. In this case, the rule pattern might specify that the first token must be "Dr." and the second token must have its first letter capitalized, with the action being to label the sequence as a person entity.

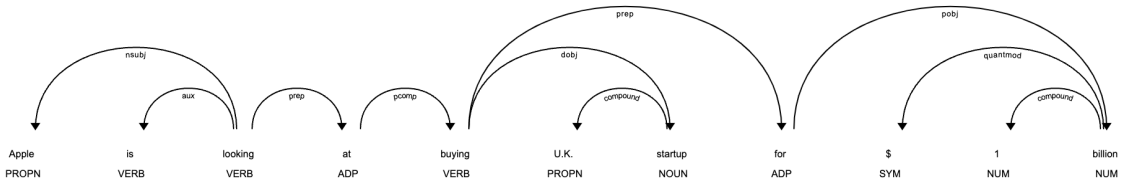

Figure 4: Detailed sentence structure breakdown demonstrating the relationship between words through dependency parsing and Part-of-Speech annotation. This example illustrates how linguistic analysis can reveal the grammatical structure of natural language. Source: [17]

[15]

One of the advantages of rule-based NER is the ability to incorporate domain-specific knowledge using ontologies or other predefined lexicons to enhance accuracy. However, conflicts can arise when multiple rules are triggered for the same sequence of tokens. To resolve such conflicts, rule-based systems often implement policies, such as ordering the rules by priority or applying them sequentially. While these approaches ensure greater control over the NER process, they also require significant manual effort to design effective rules. The creation of rules requires human expertise and is a labor-intensive process, particularly when the rules must address complex linguistic phenomena or diverse domains. [21]

2.1.3 Hidden Markow Model (HMM) for POS

HMM is one of the classic sequence labeling algorithms that generates probability distribution for each word over possible sequences of labels and assigns the top one. HMM is based on the Markow assumption, which states that the future state depends only on the current state of the system not the past ones. As well as the output observation depends only on the state that produced that observation. [22]

$$\textbf{Markov Assumption:} \quad P(q_i = a|q_1, \ldots, q_{i-1}) = P(q_i = a|q_{i-1}) \tag{1}$$

$$\textbf{Output Independence:} \quad P(o_i \mid q_1, \ldots, q_i, \ldots, q_T, o_1, \ldots, o_i, \ldots, o_T) = P(o_i \mid q_i) \tag{2}$$

- $\mathbf{Q} = q_1 q_2 \ldots q_N$: a set of N **states**, each representing a part of the speech tag (e.g., noun, verb, adjective).
- $\mathbf{A} = [a_{ij}]_{N \times N}$: atransition probability matrix where each a_{ij} represents the probability of moving from one part of speech to another. For example, a_{ij} could be the probability of a verb followed by a noun. We ensure that $\sum_{j=1}^{N} a_{ij} = 1$ for all i.
- $\mathbf{B} = b_i(o_t)$: a sequence of observation likelihoods, also called emission probabilities. These express the probability of an observation o_t (a word in the vocabulary) will be generated from a state q_i (a part of speech tag). For instance, b_i("eat") might be high if q_i is a verb.
- $\pi = [\pi_1, \pi_2, \ldots, \pi_N]$: an initial probability distribution over states. π_i is the probability that the Markov chain will start in state i. For example, π_i might be higher for nouns if sentences often start with a noun. Some states j may have $\pi_j = 0$, indicating that they cannot be initial states. We also have $\sum_{i=1}^{N} \pi_i = 1$.

HMM uses the Viterbi algorithm to compute the most likely sequence of tags for a given sentence.

2.2 Word Clouds

Word clouds are a popular visual representation of text data, where the size of each word reflects its frequency or importance within the corpus. This visualization technique provides an intuitive way to identify prominent terms and themes, making it particularly useful for exploratory text analysis. [23] Larger words typically indicate higher frequencies, while smaller ones represent less common terms. Word clouds are often used to:

- Quickly summarize the key themes in a corpus.

- Highlight differences in word usage between datasets.
- Serve as an engaging visual for presentations or reports.

However, word clouds are not without limitations. Since they rely heavily on frequency, they may not capture contextual or semantic nuances. Additionally, overly common stop words can dominate the visualization unless they are filtered appropriately. [15]

Different methods and tools can be used to create word clouds, depending on the dataset and desired outcome. There are three common approaches:

- Frequency-Based Word Clouds based on counting the frequency of each word in the corpus and assigning the sizes proportional to it. One of the variation of it is apriori based clouds that present frequently co-occurring words.
- TF-IDF Weighted Word Clouds works better whenever highly relevant but less frequent terms can be overlooked,
- Sentiment-Driven Word Clouds, focus on words with positive, negative, or neutral sentiment scores. These word clouds are particularly useful in analyzing user reviews, social media posts, or other datasets where sentiment plays a critical role.

Apriori-Based Word Clouds

While frequency-based word clouds highlight common terms, they do not reveal co-occurring words that frequently appear together. The Apriori algorithm, commonly used in association rule learning, helps identify frequent itemsets of words that appear together in a corpus. This method enhances word clouds by emphasizing associations between terms, rather than just their individual frequency.

The Apriori algorithm identifies word pairs or n-grams that meet a minimum occurrence threshold. This helps analyze word relationships that are useful for text mining within academic research.

$$\text{Support}(X) = \frac{\text{Frequency of } X}{\text{Total Words}} \tag{3}$$

Support measures how frequently a set of words appears together in the corpus and confidence evaluates the strength of the relationship between two words

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X, Y)}{\text{Support}(X)} \tag{4}$$

Apriori-based word clouds, help to visualize not only the most common words but also their frequent associations, providing a better understanding of patterns and can help tune other models.

TF-IDF Weighted Word Clouds

While frequency-based word clouds are effective for highlighting common terms, they often overlook the importance of less frequent but highly relevant words. TF-IDF (Term Frequency-Inverse Document Frequency) addresses this by assigning weights to words based on their relative importance within individual documents compared to the entire corpus. [24] This method emphasizes terms that are unique to specific subsets of the data, making it more suitable for analyzing multi-document datasets.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right)$$

Where:

- $\text{TF}(t, d)$ is the term frequency of word t in document d.
- N is the total number of documents.
- $\text{DF}(t)$ is the number of documents containing the term t.

Word clouds are a powerful starting point for exploring text data, offering an accessible and engaging overview of key terms and themes. When paired with more robust analytical methods like TF-IDF weighting or sentiment analysis, word clouds can provide deeper insights into complex datasets, including those in agricultural research.

2.3 Topic Modeling Using Latent Dirichlet Allocation

Topic modeling is a widely used text-mining technique for uncovering the underlying themes or topics present in large unstructured text corpora. Among the various methods available, Latent Dirichlet Allocation (LDA), introduced by [25], is one of the most popular probabilistic models for topic discovery. LDA represents each document as a mixture of topics, where each topic is characterized by a distribution over words. This approach has proven particularly useful for analyzing large datasets such as scientific literature. LDA is based on the assumption that documents are generated from a probabilistic model where each document is represented as a mixture of latent topics and each topic is defined by a probability distribution over the vocabulary.

Mathematically, LDA assumes the following generative process:

1. For each document, sample a distribution over topics from a Dirichlet prior.
2. For each word in the document, sample a topic from the document's topic distribution.
3. Sample a word from the chosen topic's word distribution.

The parameters of this model, including the topic distributions for each document and the word distributions for each topic, are inferred from the data using algorithms such as Variational Bayes or Gibbs Sampling.

LDA defines the probability of a corpus as:

$$P(D|\alpha, \beta) = \prod_{d=1}^{N} \int P(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{z_{dn}} P(z_{dn}|\theta_d) P(w_{dn}|\beta_{z_{dn}}) \, d\theta_d$$

Where:

- D: Collection of documents.
- w_{dn}: Word n in document d.
- z_{dn}: Topic assigned to word w_{dn}.
- θ_d: Topic distribution for document d, sampled from a Dirichlet prior with parameter α.
- β_z: Word distribution for topic z, sampled from a Dirichlet prior with parameter β.

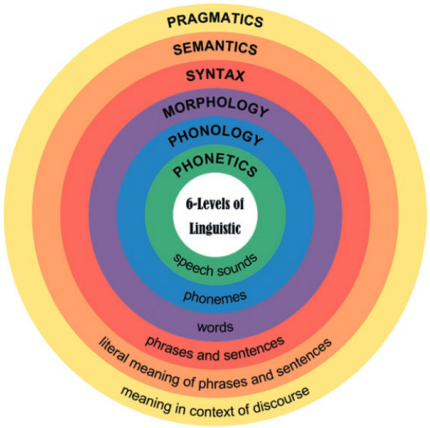

Figure 5: Hierarchical representation of linguistic levels in human languages, from phonetics to pragmatics, showing the progression from basic sound units to complex meaning interpretation in natural language processing. Source: [29]

The model learns the posterior distributions of θ and β, enabling the identification of dominant topics in the corpus.

LDA proved to be an effective tool for extracting thematic structures from a large collection of documents and in case of this work of agriculture research abstracts. By highlighting key topics, this analysis provides a foundation for further exploration of causal relationships between agricultural practices and SOC, ultimately contributing to understanding sustainable farming strategies.

2.4 Natural Language Processing

Natural Language Processing (NLP) "is the set of methods for making human language accessible to computers " [26]. and have become ubiquitous in our daily lives. Applications come in the form of search engines, translation tools, chatbots and voice assistants. The adoption of the technology is so pervasive that many people cannot imagine their lives without them. Interestingly though, methods and technologies behind above mentioned inventions arose from the famous Turing Test, which questions whether the entity we are communicating with is a human or a robot. NLP is broad filed that draws from disciplines including human linguistic, computer science , statistics, artificial intelligence, voice recognition and more Among the applications of NLP, the most prominent are Information Retirieval (IR) , Information Extraction(IE) , Machine Translations , Sentiment Analysis and Question Answering. [27]

Information Retrieval (IR) defined as: " finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [28]

To be able to succesfully identify the documents of interest there is a need to go through task of recognizing different lingusitics levels, as shown on the Figure 5. When processing speech the phonetics and phonology of the specific language will play a key role to be able to move to further analysis. When dealing with written form of language we are already able to start from morphological analysis and try to reach the level of pragmatic understanding. [27]

"Similar to an information retrieval system, an information extraction system responds to a user's

information need. Whereas an IR system identifies a subset of documents in a large text database or in a library scenario [...] an information extraction (IE) system identifies a subset of information within a document" [30]

2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of neural network specifically designed for processing sequential data. First introduced by Elman in 1990 [31], they are useful in their ability to retain memory of prior inputs through the utilization of an internal state. This internal state is propagated through the network in conjunction with each incoming input, allowing the network to effectively learn from sequences of data over time. These networks are also basis for more complex models like LTSM and even more sophisticated architecture like Transformer [22].

The core concept underlying Recurrent Neural Networks (RNNs) is the feedback loop mechanism that allows the output of a neuron to be fed back into itself. This unique structure enables RNNs to utilize both current and historical information when making decisions. However, training RNNs requires Backpropagation Through Time (BPTT), an extension of standard backpropagation that unfolds the network across time steps to compute gradients [32] As a result, RNNs are particularly well-suited for applications where contextual understanding is essential, including language modeling and text generation.

An RNN processes sequences by iterating through the sequence elements and maintaining a state that encapsulates information about what it has seen so far. The state at each step is updated by:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

where:

- h_t is the hidden state at time t,
- x_t is the input at time t,
- W_{hh} is the weight matrix for connections between hidden states,
- W_{xh} is the weight matrix for connections between the input and the hidden state,
- b_h is the bias term,
- tanh is the hyperbolic tangent activation function.

While powerful, RNNs face challenges such as the vanishing and exploding gradient problems. The vanishing gradient problem occurs when weights less than 1 lead to the exponential decay of gradients during backpropagation through time. As gradients are propagated across long sequences, repeated multiplications by small weights cause them to diminish, preventing the network from learning long-term dependencies. Conversely, the exploding gradient problem arises when weights greater than 1 cause gradients to grow exponentially over time. This results in extremely large updates to the network's weights, causing instability and preventing convergence. To address these problems, techniques like Long Short-Term Memory (LSTM) was proposed to enable effective training of RNNs for sequential data.

RNNs are particularly good for NLP tasks including: language modeling, text generation, speech recognition, or machine translation. Their main strenght being ability to maintain a memory of what has been done so far allows them to perform NLP tasks with a higher level of understanding and context awareness. When it comes to building a language model using RNN, the sentence in the document is processed one word at a time and the following word would be a prediction based on the current word and hidden state [33].

2.4.2 Transformers

Building on the success of Neural Networks in Language Processing tasks, the introduction of Transformers has significantly reshaped the field, providing a new approach to handling sequential data without the reliance on recurrence found in older models such as Recurrent Neural Networks (RNNs). Transformer-based architecture has become standard for building Large Language Models and solving a wide range of tasks, including machine translation, text classification, and named entity recognition.

Transformers rely on a key innovation: the self-attention mechanism, introduced in the seminal paper "Attention Is All You Need" [34] . Self-attention allows the model to process all tokens in a sequence simultaneously, dynamically weighing their importance in context, capturing long-range dependencies in a single step. Unlike RNNs, which read words sequentially, Transformers analyze each word in relation to all others, producing a contextual embedding. [22] This enables them to differentiate between meanings based on surrounding words—e.g., distinguishing between *"Apple"* as a company versus *"apple"* as a fruit. This is especially useful for tasks where context matters, like in above example, because the model can pay closer attention to neighboring words (tokens) that clarify meaning.

Attention Mechanisms in Transformers

An attention function maps a query and a set of key-value pairs to an output, where the output is computed as a weighted sum of the values. The weights are determined by a similarity score between the query and corresponding keys [34]. Formally, self-attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

Each token in the input sequence is transformed into three distinct vectors—known as query, key, and value. .

- **Queries** ($\mathbf{Q}$): Each input vector x_i from the sequence is transformed into a query vector $q_i = x_i\mathbf{W}^Q$. The query represents the current element being compared to others in the sequence.
- **Keys** ($\mathbf{K}$): Each input vector x_i is also transformed into a key vector $k_i = x_i\mathbf{W}^K$. The key is used to compute the match or similarity score with the queries.
- **Values** ($\mathbf{V}$): Each input vector x_i is transformed into a value vector $v_i = x_i\mathbf{W}^V$. Values are the actual data elements that are aggregated based on the attention scores.

These vectors facilitate the calculation of attention scores by enabling each token to be compared against every other token in the sequence. This comparison helps the model to focus on the most relevant parts of the text. The attention mechanism operates by first computing a score that indicates the degree of focus that each element should receive. The scores are calculated using a dot product between the queries and the keys:

$$\text{score}(x_i, x_j) = \frac{q_i \cdot k_j}{\sqrt{d_k}}$$

Where d_k is the dimension of the key vectors, and this normalization factor helps prevent extremely large values during training. These scores are then passed through a softmax function to obtain the attention weights:

$$\alpha_{ij} = \text{softmax}(\text{score}(x_i, x_j)) \quad \forall j \leq i$$

Finally, the output vector a_i for each input vector x_i is computed as a weighted sum of all value vectors, with the weights defined by the softmax output:

$$a_i = \sum_{j \leq i} \alpha_{ij} v_j$$

This formulation allows each output element to be a dynamic aggregation of inputs based on their relevance, as determined by the attention scores. The attention mechanism's ability to focus on different parts of the input sequence is what enables transformers to handle complex dependencies in data such as natural language [22].

Transformers Architecture

A key feature of Transformer architecture (Figure 6) is multi-head attention, which allows the model to capture multiple aspects of context simultaneously. Instead of relying on a single attention head, multiple heads run in parallel, each learning different relationships between words. This is especially useful for polysemous words (words with multiple meanings) and syntactic dependencies.

Unlike RNNs, which inherently capture sequence order due to their recurrent nature, Transformers process all tokens in parallel. To retain information about word position, Transformers introduce positional encodings, added to the input embeddings. These encodings use sinusoidal functions to provide unique positional information:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

where pos represents the token position and i denotes the embedding dimension. These embeddings allow the model to distinguish between words based on their placement in a sentence.

Each Transformer layer contains a position-wise feedforward network, applied to each token separately. This component enhances the expressiveness of the model, ensuring non-linearity and improved generalization. To prevent issues such as vanishing/exploding gradients, Transformers utilize **r**esidual connections and layer normalization. These mechanisms help stabilize training, allowing for deeper architectures without loss in performance.

While Transformers are not inherently sequential, they still rely on gradient-based optimization techniques used in Reccurent Neural Networks backpropagation through time (BPTT) when training for autoregressive tasks (e.g., text generation). BPTT enables the model to compute gradients across multiple steps, ensuring effective weight updates in self-attention layers[32]. However, unlike RNNs, Transformers avoid issues such as vanishing gradients due to their parallelized computations.

Transformers have revolutionized NLP by eliminating sequential dependencies and leveraging parallelization, leading to more efficient and scalable models. Through self-attention, positional encodings, and multi-head attention, these architectures provide robust language understanding capabilities.

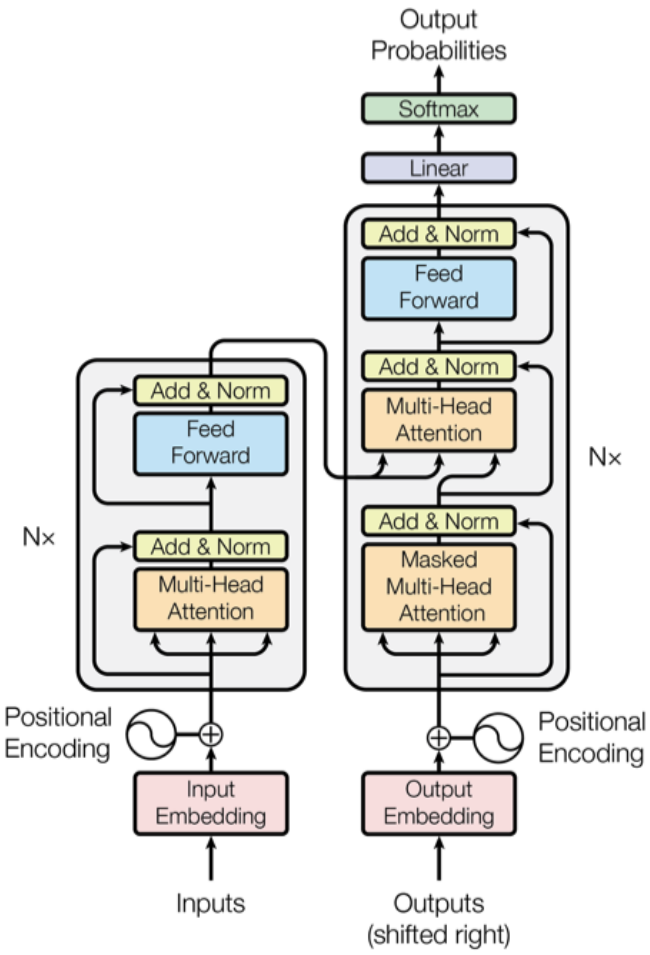

Figure 6: Transformer Architecture Overview. The model processes input sequences in parallel using self-attention and feedforward layers. Source: [34]

2.5 Large Language Models

The way we learn a language as humans is still an ongoing debate. However, more and more evidence suggests that the process of learning anything, from infancy, also the language, involves statistical learning [35][36]. Babies are learning to understand the patterns of language and apply probability to dermine which syllabels are paired together more frequently. In the spoken language there is no distinction between where one words ends and the other one starts - we learn this by experience and the more we listen and process each language the better we are at it. This is also the reason why when we hear a foreign language, for instance French, might initially sound like an indistinct string of words. Fortunately, as we hear it more often, we start to recognize the patterns and learn the probability distribution of which syllabels occure most frequently together, and from there we are able to recognize distinct words. [37] [38]

"Speech is a privileged unit for even the most basic forms of learning. From birth, when infants listen to speech, they successfully recognize individual units and their relative positions in the speech sequence . And at 1 month, infants who are conditioned to speech show a stronger response and a steeper learning curve than infants conditioned to either tones or backward speech" [39] Interestingly enough, language models work in the same way. They are trained on the large volume of text to understand the structure, rules and patterns of the language and they assign probability to each possible next word. [22].

Large Language Models (LLMs) as the name suggest are Language models that are trained on the huge amount of data for sole purpose to predict the next word based on its neighbors. There are diffrent types of the models depending on its architecture: feedforward language models intruduced by [40], recurrent language models as discussed in section 2.4.1 and transformer based language models like BERT or OpenAI's GPTs. Each of them is good when applied to certain type of NLP tasks, as shown in the Table 3

Adoption of the LLM's brought enormous advance into the speed and accuracy of various NLPs tasks, most importantly text summarization, machine translation, zero-shot and one-shot classification, question answering . For example, OpenAI's GPT is capable of accurately handling classification tasks that were traditionally done manually, thus having the potential to minimize human effort. This also holds true for translation tasks and numerous other applications. Versatility and adaptability of those models put them in the center of attention; however, accesibility of those tools is still a challenge. Due to their substantial need for computational resources, only institutions with a sufficient financial and infrastructural background can afford utilizing them.

One of the biggest challenges when it comes to LLMs, also present in all machine learning models, is their tendency to incorporate and amplify bias from training datasets. Since these models are trained on a vast amount of internet data, ensuring that they only receive accurate and factual data is an impossible task. The virality of fake news content [41], along with widespread social biases such as stereotyping, discrimination, and derogatory language, are one of the key problems facing LLMs today. [42]. Another issue that users should be aware of when using LLMs is the fact that they do not always produce factual data and have a tendency to invent, for instance, scientific papers or sources that do not exist. As always, critical evaluation of output needs to be applied to avoid propagating misinformation. [43]

Feature	BERT	GPT-3	GPT-4	RNN
Purpose	Contextual word embeddings for NLP tasks	Large-scale text generation	Improved reasoning and contextual understanding	Sequential data processing
Type	Bidirectional Transformer	Autoregressive Transformer	Advanced Autoregressive Transformer	Recurrent Neural Network
Application	Named entity recognition, sentiment analysis, text classification	Chatbots, document summarization, conversational AI	Complex reasoning, multi-modal understanding, long-context handling	Speech recognition, time-series analysis, sequential predictions
Model	Uses masked language modeling (MLM) and next sentence prediction (NSP)	Uses autoregressive modeling with left-to-right text generation	Uses improved autoregressive modeling with larger datasets	Uses sequential memory through hidden states
Example Use Case in Text	Understanding word meaning in context (e.g., "bank" as financial institution vs. riverbank)	Generating human-like responses in chatbots or creative writing	Better document comprehension and long-form content creation. Ability to add attachments	Predicting the next word in a sentence based on prior words
Training Method	Pre-trained on large text corpora using masked token prediction (cloze task)	Trained on a vast dataset using transformer-based deep learning	Fine-tuned with RLHF (Reinforcement Learning from Human Feedback) for more refined responses	Uses backpropagation through time (BPTT) to update weights sequentially
Limitations	Requires large labeled datasets for fine-tuning; limited in generative tasks	Generates text based on patterns but lacks deep reasoning	Computationally expensive; still susceptible to hallucinations and bias	Struggles with long-term dependencies due to vanishing/exploding gradient problem

Table 3: Comparison of Transformer-Based and Recurrent Neural Network Language Models

2.5.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) introduced by [44] is a different type of language model. Instead of being trained on guessing the next word, the model needs to predict the word that is masked between two other words. Models taht rely on this mechanism are called masked language models, which in contrary to left to right models allow to understand the

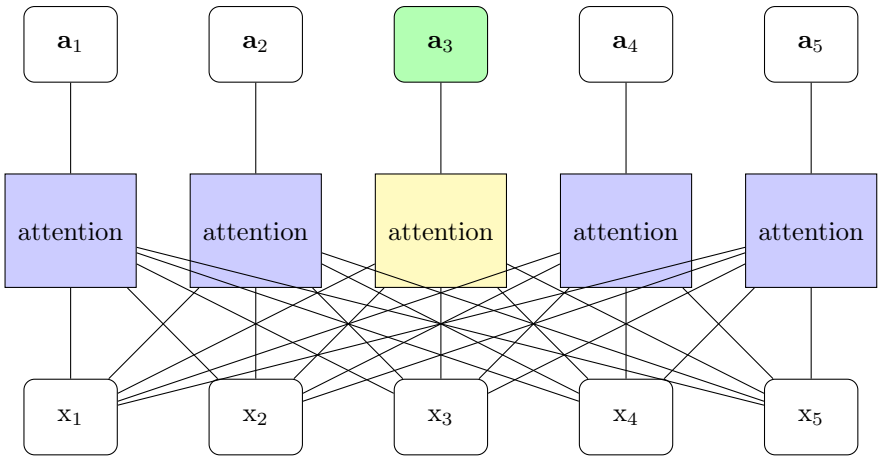

Figure 7: Representation of bidirectional self-attention layer. BERT model uses input from all the previous and following tokens to compute representation of the token, thus the name bidirectional. [22]

context on both sides of the word.. [22]

This left-to- right approach can have its limitations, it works quite well on the tasks requiring text generation, as each consecutive input relies on the previous tokens. However, bidirectional encoders are creating representation of the tokens that is based on the context, which means that all tokens in the example sentence will be impacting the tokens representation as seen on the 7 . This architecture helps whenever the task at hand requires classification or decision to be taken based on the context of the word.

"Bidirectional encoders use self-attention to map sequences of input embeddings $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ to sequences of output embeddings of the same length $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)$, where the output vectors have been contextualized using information from the entire input sequence" [22]

Biderectional encoders is trained differently to causal transformer language model. Instead of teaching the model to get better at guessing next token, we give a task of filling the blank spaces within the text with words. This is called cloze task taking the name from the psychological test that measured readability of communcation message.[45]

2.5.2 Generative Pretrained Transformer - GPT-3/GPT-4

One of the most famous LLMs that took the world by storm is OpenAI's ChatGPT. The introduction of GPT-3[46] and later GPT-4 [47] by OpenAI revolutionized the field of NLP and put Artificial Intelligence models on the radar, making them a part of public discourse. The race among the biggest tech companies, as well as economies, have begun. [48]

The GPT-3 mode, launched in 2020, is based on the transformer architecture and is designed to generate human-like responses to prompts. Like other autoregressive language models, it operates by predicting the next token in a sequence, making it capable of constructing fluent and contextually relevant response.

Mechanism behind ChatGPT is not in principle different from other language models—its goal is to generate the most probable next token given the previous ones. However, unlike bidirectional models such as BERT, which process text in both directions to understand meaning, GPT-3 and its successor, GPT-4, generate text in a strictly left-to-right manner. This design choice makes

them particularly effective at tasks like open-ended text generation, dialogue modeling, and content creation. The other thing that makes GPT 3 and GPT 4 unique is the sheer amount of data those models were trained on, making them one of the most powerful language models created so far. ChatGPT is based on the GPT 3.5 model that is smaller version of GPT 3, with 6.7 bilion parameters compared to 175. [49].

Despite these advancements, GPT-based models still face major challenges, including computational costs, accessibility, and the tendency to generate factually incorrect or biased content [50]. The ability to produce fluent text does not always mean the output is reliable, making the need for responsible AI development and fact-checking crucial.

Other phenomenon that is observed when using ChatGPT is called artificial hallucinations and is defined by itself as "Artificial hallucination refers to the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input. This can include visual, auditory, or other types of hallucinations. Artificial hallucination is not common in chatbots, as they are typically designed to respond based on pre-programmed rules and data sets rather than generating new information. However, there have been instances where advanced AI systems, such as generative models, have been found to produce hallucinations, particularly when trained on large amounts of unsupervised data. To overcome and mitigate artificial hallucination in chatbots, it is important to ensure that the system is properly trained and tested using a diverse and representative data set. Additionally, incorporating methods for monitoring and detecting hallucinations, such as human evaluation or anomaly detection, can help address this issue" [51]

Nonetheless, the rise of GPT models has propelled LLMs into mainstream discourse, sparking global debates on their implications for labor markets, education, and even regulatory policies.

GPT Mechanism

GPT is a unidirectional Transformer model that is trained in phases:

1. Pretraining: In this phase, GPT learns to predict the next word in a sentence given the previous words. The objective function is the cross-entropy loss over the predicted token probabilities:

$$L = -\sum_{t=1}^{T} \log P(w_t | w_1, w_2, \ldots, w_{t-1}; \theta)$$

   where w_t is the token at time step t, and θ represents the model parameters.

2. Fine-Tuning: After pretraining, the model is fine-tuned on domain-specific datasets or specific tasks using supervised learning. This step adjusts the pretrained weights on the smaller dataset to optimize for the task at hand, such as classification, summarization, sentiment analysis or replying in specific style. [52]

First GPT1 one was characterized by 12-layer decoder and was trained on the book corpus. Following GPT-2 was already 10 times bigger with 1.5 billion paramters and was trained on WebText. [52]

GPT-3 and GPT-4 Enhancements

Building on the original GPT architecture, GPT-3 and GPT-4 introduce significant improvements:

- GPT-3: With 175 billion parameters and trained on "Common Crawl" dataset, GPT-3 achieves good language translation, generalization and requiring minimal task-specific data to generalize effectively, however still performing poorly when it comes to explanations .
- GPT 3.5: having maximum 4096 token limit this version imrpoved significantly used experience however still performed poorly on simple math problems and contect understanding [53].
- GPT-4: This model further enhances performance and improved contextual understanding. GPT-4 is also designed to handle longer input contexts, enabling applications such as multi-document analysis and complex reasoning tasks.

2.6 Causal Inference in Natural Language Processing

While traditionally NLP models focus more on predictive tasks and detecting correlations in text, causal inference did not have the same importance in NLP community. This gap should get addressed if we want truly comprehend cause-and-effect relationships within text data and go beyond shallow understanding. The ability to move from co-occurrence and towards causality is critical in various domains, especially in scientific literature analysis .[54] In the context of agriculture and soil science, causal inference is essential for understanding the impact of conservation tillage on Soil Organic Carbon (SOC) rather than simply identifying co-existing mentions of the two concepts.

Standard text mining and machine learning models perform well on task like topic distribution, entity relations or term frequencies, however to move from this basic understanding of content to more nuanced causal relationship there are few challenges to be aware of [55]:

1. Lack of explicit causality: scientific pulbliactions rerly state cause and efect explicitly, as the language is often complex and nuanced
2. Correlation not causation: often the ussage of terms such as "associated with" or "linked to" show corralation, however not necessarly causation
3. Multiple variables: due to complexity of the domains like agriculture, there are many variables that can influance any indicator, making extraction of singular cause difficult
4. Long-span dependencies: creating a compelling scientific reasoning can span through multiple sentences which makes finding a model that can track the flow of thoughts challenging.

To address the above challenges, extraction of causal relation relied on approaches such as rule-based dependency parsing to detect key markers of causality like "caused by", etc., and pattern-based methods employing domain-specific ontologies or dictionaries to identify key terms [56]. Those methods performed well in the structured environment; however, they could run into issues when context changed.

With the advent of Transformer Based architecture new tools for Natural Language Processing became available. Starting from pre-trained contextual embeddings (BERT) that serves to better identify nuanced causal language to Chain of Thought (CoT) [57] prompting (GPT–4) that helps break down complex cause-effect reasoning. One of the most promising models for causal inference is BERT-Causal that was fine-tuned to detect cause-effect relationships within text by employing dependency parsing and attention mechanism. [58].

Chapter 3

Methodology

3.1 Information Extraction to Relation Extraction

Information Extraction turns the data in form of text, which by definition is unstructured, into structured data that can be processed further. "Relation extraction is an information extraction task that extracts entities and the relationship between them. A relation in information extraction is a descriptive relationship between entities or events" [59]

Using entity recognition can be one way of structuring data and classifying it into labels that can be processed further for relation extraction. Prior to the recognition of specific phrases or keywords, it is essential to preprocess the text. This preprocessing ensures the consistency of relevant data across various word forms and tenses. Furthermore, it is critical to eliminate any data inconsistencies and to assess data quality, as encapsulated in the well-known adage in the Data Science community: "trash in, trash out." Without good input into analysis process we are not aboe to obtain informative and valuable output.

Information extraction in this context focuses on identifying specific relationships between agricultural practices and soil organic carbon changes. The transition from general information extraction to relation extraction involves identifying not just entities (like tillage practices or SOC measurements) but also understanding how these entities interact and influence each other within the abstracts.

3.2 Data Collection and Preprocessing

3.2.1 Dataset Description

The primary data for this research consisted exclusively of scientific papers, gathering a wide range of publications within agricultural field. A total of 16,874 publications from the years 1972 to 2020 were compiled to construct the dataset for analysis. There was steady growth in amount of publications starting from late 1990's 8, which is also alligned iwith the popularity of specific terms use. These documents were sourced from scientific databases, ensuring that each paper was peer-reviewed and relevant to agronomy research.

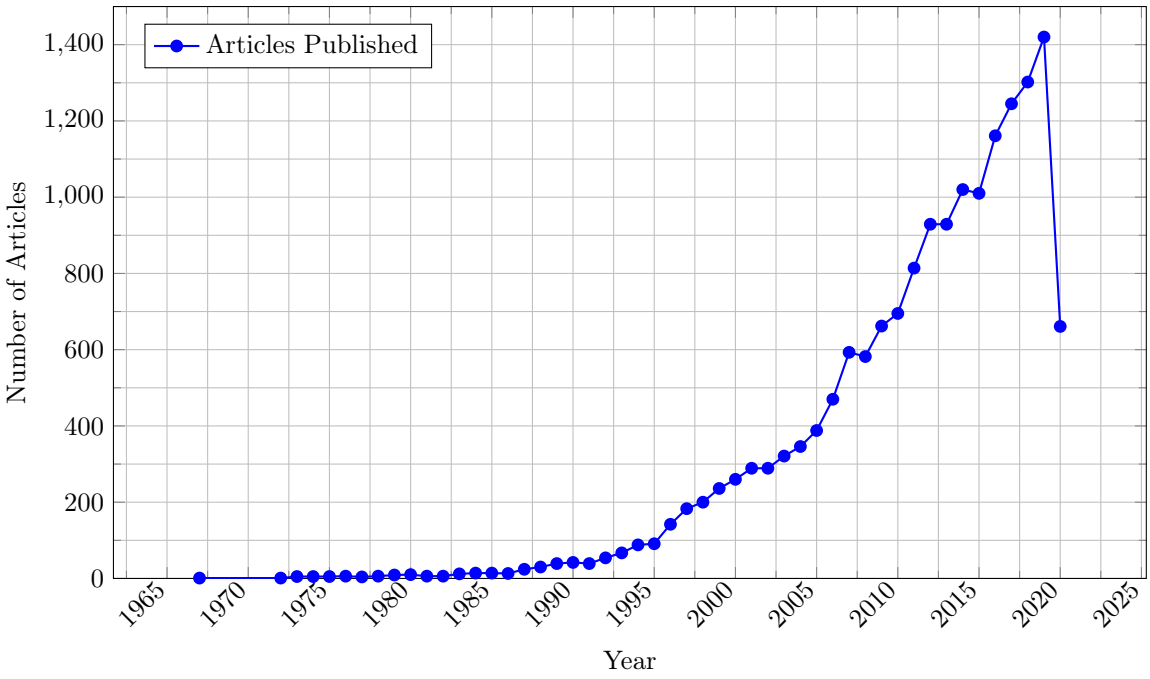

Figure 8: The timeline illustrates the progression in the number of published articles over the years. The steady increase highlights growing research activity, particularly between 2005 and 2020, where we can observe almost exponential growth with number of publications in the field.

3.2.2 Data Dictionary

An ontology, specifically provided by the Agriculture Department, served as the backbone for navigating and categorizing the collected data. The ontology included relevant terms and relationships that are key focus on this work 9 4:

- Conservation tillage practices, such as No Tillage (NT) and Minimum Tillage (MT).
- Soil organic carbon (SOC) and SOC trends (categorized under increased or decreased)

This structured approach allowed for the extraction of relevant information from complex scientific narratives. This method focuses on analyzing the effects of tillage practices on soil organic carbon (SOC). In addition to relevant labels, the data dictionary includes mappings for various agricultural practices, soil and climate types, and crops, among others. This ensures that abstracts are labeled using domain-specific knowledge. The ontology is designed to give a general overview while also providing detailed expressions to capture all information.

The list of all the phrases from ontology used to create a custom mapping is presented in Table 4 and is related to Minimal Tillage, No Tillage, SOC, SOC Increase, SOC Decrease.

The data dictionary serves as a comprehensive reference for all variables and terms used in the analysis. It includes:

- Entity categories and subcategories (e.g., tillage practices, soil properties)
- Relationship types between entities

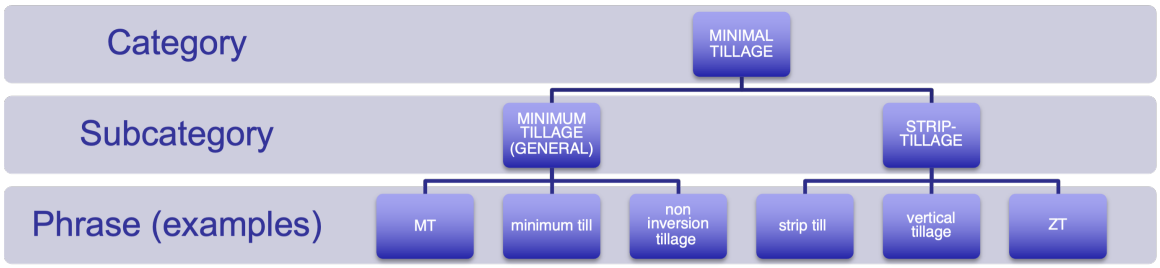

Figure 9: Data Dictionary structure illustrating the categorization of Minimum Tillage practices into Categories, Subcategories, and associated example phrases.

- Standardized terms for agricultural practices

3.2.3 Data preprocessing

All abstracts were preprocessed with the following steps using `nltk` library, which provides comprehensive open-source python Natural Language toolkit. Below steps ensured that data is ready for further modeling:

1. Gathering data: All data was combined and loaded to singular `pandas` dataframe. Each column represented different characteristics of the particular research publication, e.g. title, year, authors, etc.

2. Data cleaning and preparation: The rows that had missing values in the key columns like authors, title or abstract were removed. Conversion to lowercase, removal of irrelevant characters as well as tokenization of abstracts was performed.

3. Stop words removal: words that do not bring meaning (e.g., "the," "a") were eliminated

4. Lemmitization and stemming: reducing words to their root (e.g., "running" to "run") for easier analysis and ability to identify all key words

Dataset was preprocessed using this step with each new column gathering the specific outputs of the process like Porter-stemmed abstracts, lemmitized abstracts etc. Depending on the next steps different data columns were used. As for instance SpaCy's NER does not require preprocessing, unmodified abstract were used.

3.3 Text Mining Techniques

3.3.1 Word Cloud with Apriori Model

To visually represent the most frequent terms, a word cloud based on apriori alghoritm was generated. This model required utilizing the `wordcloud` and `efficient_apriori` library

1. Prepare transactions: Convert the preprocessed text data from previous steps into a list of transactions, where each transaction represents a document or a piece of text.

2. Apply the Apriori algorithm: Use the Apriori algorithm to identify frequent itemsets (groups of words that frequently appear together) based on a minimum support threshold (tested support of 0.5 and 0.25)

3. Prepare frequency dictionary: A dictionary called `frequency_dict` was created to store the frequency of each word or phrase obtained from the Apriori model. The keys of this dictionary were the words/phrases and the values are their frequencies.

4. Generate word cloud: Based on the frequency dictionary a `WordCloud` object was created with desired visual settings to ensure readability of the generated image. The `generate_from_frequencies` method is then used to generate the word cloud from the `frequency_dict`.

3.3.2 LDA Application

In this thesis, LDA was applied to analyze abstracts from agricultural research papers to identify among other research topics, ones that could potentially be related to tillage practices and their impact on Soil Organic Carbon (SOC). Library used in this step was `gensim` for topic modeling. Process involved:

1. Creating a dictionary and corpus**:** A dictionary is created using `corpora.Dictionary` to map words to unique IDs. Then, a corpus is created using `dictionary.doc2bow`, to represent each abstract as a bag-of-words.

2. Training the LDA model**:** An LDA model is trained using `LdaModel` on document-term matrix with the specified number of topics (20 in this case), the dictionary from the previous step and the number of passes (15).

3. Obtaining topic distribution: Get the probability distribution of topics for each document and the probability distribution of words for each topic

4. Classifying abstracts: Each abstract is classified based on the dominant topic assigned to them by the LDA model.

3.4 Custom rule-based and Pattern Matching model

3.4.1 Entity Recognition with Spacy Pipeline

The entity recognition task was designed to extract key terms and phrases relevant to the research question from the scientific abstracts. SpaCy's EntityRuler component was utilized to build a customized pipeline, enabling the identification of domain-specific entities effectively.

The recognized entities were categorized based on a predefined ontology into:

- SOC Indicators: Terms associated with SOC levels, changes, and storage (e.g., "SOC sequestration," "soil organic matter"),

- Agricultural Practices: Techniques such as no-tillage (NT) and minimum tillage (MT), conventional tillage, fertilization etc.

- Characteristics: Climate, soil, crop types, soil physical and chemical properties and geographical regions

- Increase/Decrease Trends**:**Custom labels were created to capture actions indicating improvement or reduction in SOC (e.g., "increase," "reduce") as detailed in Table 5

Label	Word
INCREASE	increase, raise, enhance, boost, higher, greater, increased, enhanced, improved, augmented, accumulated, retained, accumulation, humification, protection, retention, stabilization, stabilisation, sequestration
DECREASE	decrease, drop, reduce, lower, diminish, evolution, decomposition, degradation, emission, loss, mineralisation, mineralization, respiration

Table 5: Mapping for Words Associated with Increase and Decrease Trends in SOC. Words in the 'Increase' category reflect improvements in SOC levels (e.g., 'enhance', 'retention'), while words in the 'Decrease' category indicate reductions or losses (e.g., 'decrease', 'degradation').

The SpaCy pipeline consisted of the following key steps:

1. Pattern Creation and Categorization: Patterns were created based on a curated ontology of agricultural terms and phrases. Each pattern was mapped to its corresponding category (e.g., SOC Indicators, NT/MT practices).

2. Integration into SpaCy's EntityRuler: The EntityRuler allowed predefined patterns to be integrated directly into the SpaCy NLP pipeline. This facilitated rule-based matching of entities during text processing. Patterns were expressed using JSON objects, specifying the entity type (e.g., *SOC_INDICATOR*) and the associated terms or regular expressions. An example of this integration is shown below:

```
{
    "label": "MINIMUM TILLAGE",
    "pattern": "disking"
}
```

The outcome of this step is presented in Table 6.

3.4.2 Dependency Based Pattern Matching

SpaCy's DependencyMatcher was used to capture relationships between entities. Dependency patterns were defined to identify syntactic relationships, such as SOC being associated with trends like *increase* or *decrease*. For instance, the SOC_PLUS_INCREASE pattern captured phrases where *SOC* was syntactically related to an action like *increase*.

```
"SOC_PLUS_INCREASE": [
    {
        "RIGHT_ID": "SOC",
        "RIGHT_ATTRS": {"ENT_TYPE": "SOC"}  # Match entities tagged as SOC
    },
    {
        "LEFT_ID": "SOC",
```

```
        "REL_OP": ">",
        "RIGHT_ID": "increase",
        "RIGHT_ATTRS": {"ENT_TYPE": "INCREASE"}  # Match entities tagged as INCREASE
    }
```

Then multiple iterations of the dependency matcher and filtering took place to close the gap between all the possible combinations of labels for increasing or decreasing SOC. The dependency matcher specifically examined the abstract for sentences that contained either or both of those phrases within a single sentence.

- "SOC "+ :"Increase"/"Decrease",
- "SOC Increase"/ "SOC Decrease"
- "SOC Increase"/"SOC Decrease" + "SOC",
- "SOC Increase"/"SOC Decrease" + "No tillage"/"Minimum Tillage"

The analysis of the occurrence of each label is presented in Table 7

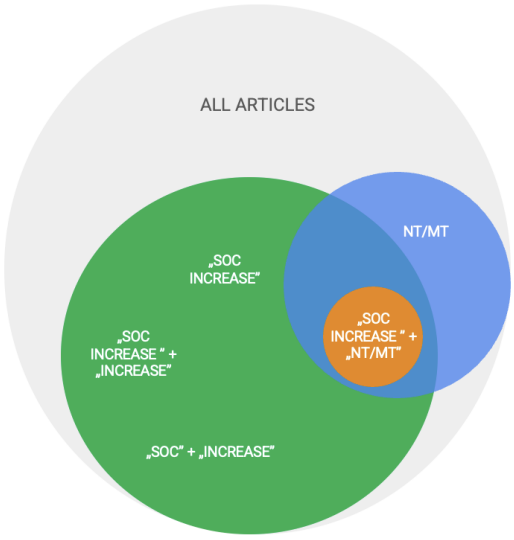

Figure 10: Example of NT/MT positive impact on SOC: Graph showing the subsets of data. Green circle represents articles that are connected to SOC Increase by any label that indicates it. Blue circle shows the publication with label: Minimum Tillage or No Tillage. The orange circle signifies subset that has SOC Increase and No tillage or Minimum Tillage within one sentence. Articles that would be relevant to identify potential casual relationship are at the cross section between NT/MT and All SOC Increase

As a next step, all the subsets that should indicate SOC Increase were included, and additional filter was applied to identify the articles that mention within abstract Conservation Agriculture practices: No Tillage and Minimum Tillage (Figure 10).

Label	Global Number of Labels	Abstracts Containing Label	% of Articles Containing Label	Phrase Occurrence
SOC	32,965	9,598	56.9%	3.43
Increase	53,149	14,579	86.4%	3.65
Decrease	39,684	12,202	72.3%	3.25
SOC Increase	914	596	3.5%	1.53
SOC Decrease	1,243	691	4.1%	1.80
No Tillage	9,452	2,433	14.4%	3.88
Minimum Tillage	4,232	1,343	8.0%	3.15
Conservation Agriculture	1,055	649	3.8%	1.63

Table 7: Presence of labels of interest in the dataset. Notably, 86.4% of the articles refer to some form of "Increase," making it the most prevalent label. Comparatively, labels related to specific changes in SOC, such as "SOC Increase" and "SOC Decrease," are less frequent but provide critical insights into the impact of agricultural practices on soil health. The column "Phrase Occurrence within Abstract" shows the average number of times each label appears within an abstract.

Based on the combination of all the possible matches we identified the subset of data that was later on used for further analysis.

3.4.3 Distance Between Entities

Basic assumption for further analysis was that if the key labels are in close proximity to each other, it would signify that there is a high probability of a causal relationship between SOC trend and tillage practice. Distance between entities of interest was calculated and different models based on that distance were used to filter the dataset. Co-ocurrence of those entities was discarded as good prediction of casual relationship. However, other distances were verified. Namely distance of maximum 3 and and maximum 5 entites between the labels of interest. Table 11 shows visually distance between labels in selected abstracts.

Histogram Analysis

The following histogram presents the distribution of distances between mentions of *SOC Increase* and labels for NT or MT practices within the corpus of agricultural research abstracts. This analysis provides foundational information for determining the cutoff point for the relevance of distance (Figure 12).

The distribution indicates that the majority of these terms appear in close textual proximity, with the highest frequency observed at distances between 0-1 and 1-2 words. The decreasing trend as distance increases suggests a strong co-occurrence pattern, reinforcing the assumption that SOC Increase is often discussed in direct connection with conservation tillage practices.

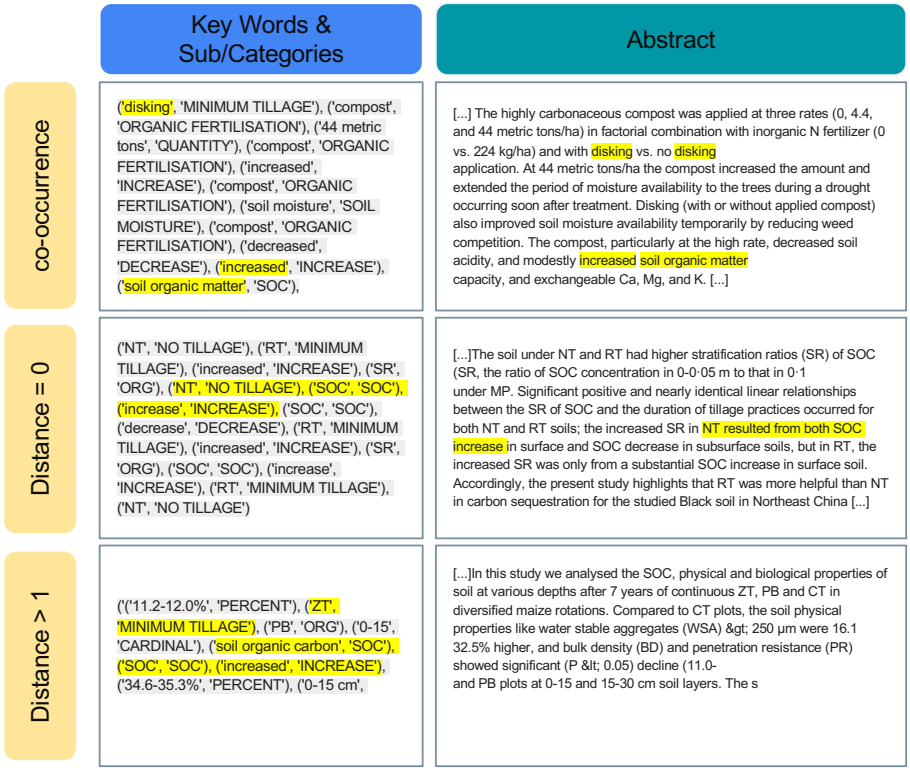

Figure 11: Visualization of the Name Entity Recognition matching outcome within abstract. The table highlights the relationships between key terms (e.g., "SOC," "Increase," "Minimum Tillage") and their context within the abstract text. The columns represent different categories: co-occurrence of terms, exact matches (distance = 0), and phrases found at greater distances (distance > 1). Examples demonstrate how different distances between terms can indicate varying strengths of causal relationships.

3.5 Transformer Based Models

3.5.1 Zero-Shot Classification with RoBERTa

Zero-shot classification enables the classification of text into predefined categories without any prior training data for those categories. This approach leverages the general knowledge captured by a pre-trained language model, such as RoBERTa, to understand the semantic level relationships between the text and the candidate labels. In this work RoBERTa-large-mnli model was employed for zero-shot classification. The model was implemented using the `transformers` library in Python in the following steps :

1. Model Loading**:** The `pipeline` class was imported from the `transformers` library. A zero-shot classification pipeline was initialized with the `RoBERTa-large-mnli` model, specifying the device to be used (GPU if available).

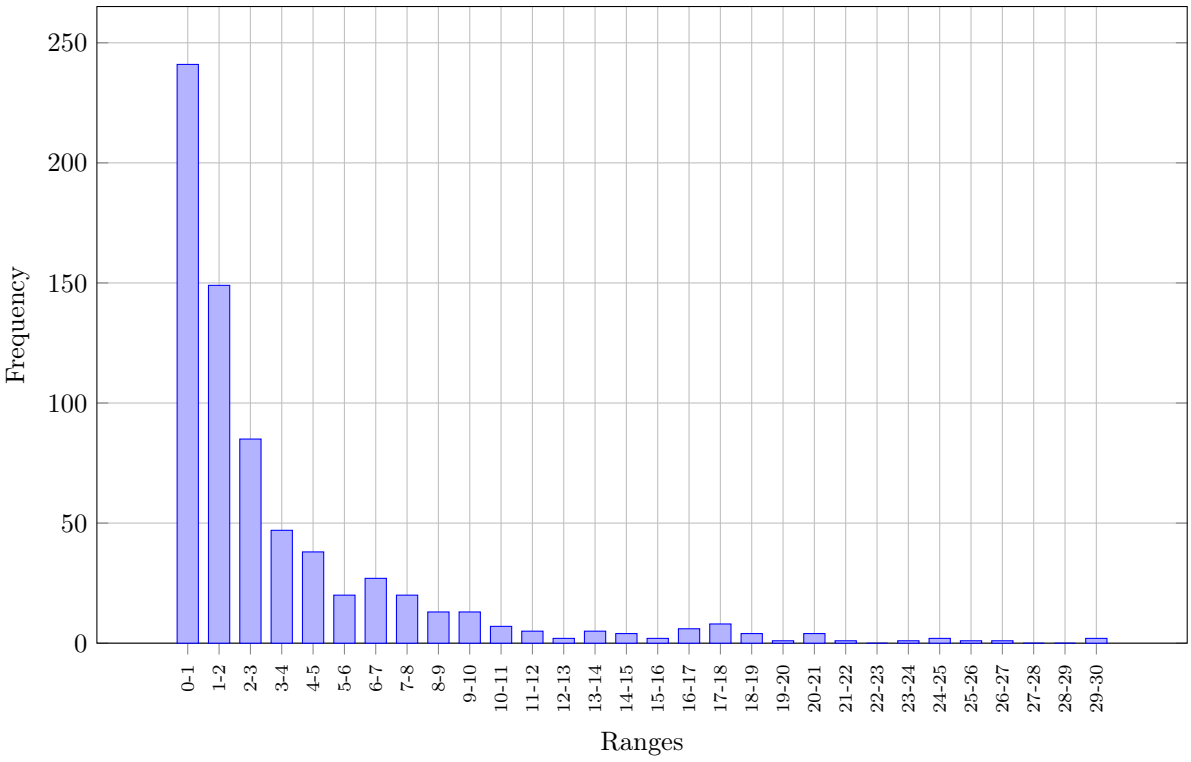

Figure 12: Histogram of the distances between SOC Increase and No Tillage or Minimum Tillage labels. Distance is measure in the entities between those 2 labels of interest

```
from transformers import pipeline

classifier = pipeline("zero-shot-classification", model="roberta-large-mnli")
```

2. Candidate Label Definition: The candidate labels for classification were defined based on the research question. Following candidate labels were used to identify the impact of conservation tillage (CT) on soil organic carbon (SOC) within each abstract:

```
candidate_labels = [
    "no tillage or minimal tillage positive impact on Soil Organic Carbon",
    "no tillage or minimal tillage negative impact on Soil Organic Carbon",
    "no impact",
    "other"
]
```

3. Classification Process: Each abstract was processed using the initialized `classifier` pipeline. The abstract and the candidate labels were input to the pipeline, which returns a dictionary containing the predicted label and its confidence score.

4. Batch Processing Implementation: For efficiency, batch processing was implemented to classify multiple abstracts concurrently. A function was defined to classify a batch of abstracts. Abstracts and their corresponding IDs are extracted in batches, processed using the `classifier`, and the results are appended to a CSV file. This approach allowed for resuming the process from an interruption point, that often occurs when using RoBERT model.

5. Output and Analysis: The output dataset contained each row representing an abstract and predicted label. The classified results were analyzed to investigate trends and correlations between the text content and the assigned categories, as discussed in the Results section.

3.5.2 Classification using GPT-3.5 and GPT-4

GPT-3 (`gpt-3.5-turbo`) and GPT-4 (`gpt-4o`) were employed to classify a representative sample of scientific abstracts. The task assigned to the models was to categorize each abstract according to the impact of tillage practice (NT/MT) on soil organic carbon (SOC) or label it as the other if it is not applicable. Prompt engineering and persona settings were used to align the model outputs with domain-specific knowledge in agriculture, ensuring the relevance of the analysis (Table 8).

To maintain consistency and prevent any contamination between abstracts, each article was processed independently using the API. This approach ensured that no prior information from other abstracts influenced the analysis. The models operated under uniform conditions, employing the same persona and prompt configuration for all inputs. The implementation was as follows:

1. Model Selection and Initialization: The OpenAI paid API was utilized to access GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4o). The API key is set, and a prompt engineering approach was adopted to guide the models towards the desired classification task. Due to the cost of the processing, the model was only applied to a sample of 200 abstarcts.

2. Prompt Engineering: A prompt was designed to provide context and instructions to the LLMs. The prompt includes information about the domain (agriculture and soil organic carbon), the target task (classification of impact), and the expected output format.

Persona definition	**Prompt**
"You are an expert in Agriculture, especially soil organic carbon and conservation tillage like for instance Minimal Tillage, No tillage etc.,"	The following text is the abstract from a scientific paper around the topic of agriculture that was published, and you want to focus on the impact of conservation tillage (CT) like Minimum Tillage (MT), No tillage (NT) on soil organic carbon (SOC). Classify it as either having a 'CT positive impact on SOC', 'CT negative impact on SOC', 'CT no impact', or 'other'. Abstract: {abstract}. Classification:

Table 8: Persona definition and prompt used for classification. The table outlines the configuration of the expert persona used in the classification task, focusing on domain-specific knowledge of soil organic carbon (SOC) and conservation tillage (CT). The accompanying prompt provides a structured instruction for categorizing abstracts based on the impact of CT practices—Minimum Tillage (MT) and No-Tillage (NT)—on SOC.

3. Classification Process: Each abstract was fed to the LLMs as input along with the crafted prompt. The models generate a response containing the predicted classification label.

```
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",  # or "gpt-4o"
    messages=messages,
    max_tokens=10,
    temperature=0.0
)
classification = response['choices'][0]['message']['content'].strip()
```

4. Batch Processing and Output: Similar to the RoBERTa implementation, batch processing was employed to handle a large number of abstracts efficiently and avoid interruption.

5. Analysis and Comparison: The classification results obtained from GPT-3.5 and GPT-4 were analyzed and compared with the results from other methods, such as custom model and RoBERTa. This analysis helps to evaluate the performance and suitability of different approaches for the given task.

3.6 Models Validation

To assess the performance of the classification models implemented for causal inference (custom model, RoBERTa, GPT-3.5, and GPT-4), a validation dataset was created and manually annotated. This dataset consisted of 200 scientific abstracts, which were deemed statistically significant out of the initial corpus of almost 17,000 abstracts. The manual classification of these abstracts served as the basis for evaluating the prediction of the model.

1. Validation Dataset Creation**:** A subset of 200 abstracts was randomly selected from the original dataset using sample method from `pandas` library. The sample size was selected based on balance between confidence level, margin of error and feasibility for manual categorization. With 200 classified abstracts, we can reach the confidence level of 95% with 6.93% margin of error. Assuming that the abstracts were relatively homogeneous, this sample should give us good understanding of models performance.

2. Manual Annotation: The 200 selected abstracts were carefully reviewed and manually classified into predefined categories related to the impact of conservation tillage on soil organic carbon (SOC). These categories were as follows: CT positive impact on SOC, CT negative impact on SOC, CT no impact, Other. Where CT refers to conservation tillage

3. Model Evaluation: Each of the implemented models (Custom model, RoBERTa, GPT-3.5, and GPT-4) was used to classify the 200 abstracts in the validation dataset. The predicted labels from each model were then compared to the ground-truth labels obtained through manual annotation.

4. Performance Metrics: Standard classification metrics were used to quantify the performance of the models. These metrics included:

- Accuracy: The percentage of correctly classified abstracts.
- Precision: The proportion of correctly classified abstracts within a specific category.

- Recall: The proportion of abstracts belonging to a specific category that were correctly classified.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of performance.
- Specificity: The ability of the model to correctly identify negatives, ensuring that abstracts without a causal link to SOC changes are not misclassified as relevant.

These metrics along the confusion matrix were calculated for each model and for "positive CT impact on SOC" label to gain a comprehensive understanding of their performance.

Category	Subcategory	Phrase
MINIMUM TILLAGE	CHISEL	chisel, chisel cultivation, chisel ploughing, chisel plowing
MINIMUM TILLAGE	HARROW	disk cultivation, disk harrowing, disk ploughing, disk plowing
MINIMUM TILLAGE	MINIMUM TILLAGE (GENERAL)	minimum till, minimum tillage, minimum-till, minimum-tillage, MT
MINIMUM TILLAGE	OTHER MINIMUM TILLAGE TECHNIQUES	hoeing, mulch till, mulch tillage, mulch-till, mulch-tillage
NO TILLAGE	NO TILLAGE	direct drill, direct drilling, direct planting, direct seeding, direct sowing, direct-seeding, direct-sowing, no till, no tillage, no tilling, no-till, no-tillage, no-tilling, NT, permanent raised bed, sod seeded, sod seeding, sod-seeded, sod-seeding
SOC	SOC QUANTITY	C stock, C storage, carbon in organic matter, carbon in the organic matter, carbon stock, carbon storage, organic C storage, organic carbon storage, SOC, soil C, soil C storage, soil carbon, soil carbon storage, soil OC, soil OM, soil organic C, soil organic carbon, soil organic matter, SOM
SOC DECREASE	SOC DECREASE (GENERAL)	soil CO2 efflux, soil CO2 evolution, SOC decomposition, SOC degradation, SOC emission, SOC loss, soil mineralisation, SOC mineralisation, soil mineralization, SOC mineralization, soil respiration, SOC respiration
SOC DECREASE	SOC INCREASE (GENERAL)	SOC accumulation
SOC INCREASE	SOC INCREASE (GENERAL)	accumulation of C, accumulation of carbon, C accumulation, carbon accumulation, SOC humification, SOC protection, SOC retention, SOC sequestration, SOC stabilisation, SOC stabilization

Table 4: Mapping of Key Phrases to Categories and Subcategories Related to Tillage Practices, SOC Quantities, and SOC Trends. This table categorizes terms associated with minimum tillage, no-tillage, and SOC increase or decrease, providing a structured framework for identifying these practices and trends in research texts

Document - Abstract	Labels
The highly carbonaceous compost was applied at three rates (0, 4.4, and 44 metric tons/ha) in factorial combination with inorganic N fertilizer (0 vs. 224 kg/ha) and with disking vs. no disking following compost application. At 44 metric tons/ha the compost increased the amount and extended the period of moisture availability to the trees during a drought occurring soon after treatment. Disking (with or without applied compost) also improved soil moisture availability temporarily by reducing weed competition. The compost, particularly at the high rate, decreased soil acidity, and modestly increased soil organic matter, cation exchange capacity, and exchangeable Ca Mg, and K. Nitrogen concentration in pine foliage was reduced following application and incorporation of the high rate of compost, but recovery to pretreatment levels was rapid.	('compost', 'ORGANIC FERTILISATION'), ('three', 'CARDINAL'), ('0', 'CARDINAL'), ('4.4', 'DATE'), ('44 metric tons', 'QUANTITY'), ('224 kg', 'QUANTITY'), ('disking', 'MINIMUM TILLAGE'), ('disking', 'MINIMUM TILLAGE'), ('compost', 'ORGANIC FERTILISATION'), ('44 metric tons', 'QUANTITY'), ('compost', 'ORGANIC FERTILISATION'), ('increased', 'INCREASE'), ('compost', 'ORGANIC FERTILISATION'), ('soil moisture', 'SOIL MOISTURE'), ('compost', 'ORGANIC FERTILISATION'), ('decreased', 'DECREASE'), ('increased', 'INCREASE'), ('soil organic matter', 'SOC'), ('cation exchange capacity', 'SOIL CHEMICAL PROPERTIES'), ('&Ca Mg&', 'WORK$_O F_A RT'$)

Table 6: Entity recognition results for domain-specific classification: This table presents the extracted entities using a rule-based approach with SpaCy's EntityRuler, customized for the agricultural domain. The categorization includes key soil organic carbon (SOC) indicators, conservation tillage practices, and their associated effects on SOC trends. The predefined ontology ensures structured extraction of meaningful entities, allowing for improved classification of abstracts based on the presence of key terms.

Chapter 4

Dataset Insights and Visualizations

To understand better the insights coming from over 40 years of research in agriculture field, the analysis of the whole dataset was conducted. To better understand the research trends and topics and identify potential gaps diffrent techniques, visual and descriptive, were applied.

Based on the ontology the abstracts coming from whole dataset were processed with Named Entity Recognition method to extract important information regarding characteristics of soil, crop and climate as well as understand the prevalence of sustainable agriculture practices in the corpus.

4.1 Research distribution by country

Geographical distribution was skewed towards China, as seen in the map (Figure 13), with 1461 articles followed by India - 501 and Brazil - 405. In Europe the most represented country was Germany with 390 articles , Spain - 364, Switzerland - 242 - and Italy - 209 were leading the way. Over representation of China needs to be also taken into consideration when interpreting the rest of the data. The number of publications from this country is growing "China's share in global agri-food publications increased from 1% in 1996 to 9% in 2012.[] However, China's global share in agri-food patents, publications and citations remains far below those in the United States and the EU28". [60] However the agricultural development was achieved at the expense of sustainable use of natural resources. [60]

4.2 Research Trends

The dominant terms considering frequency of use in the dataset, as visualized by Word Cloud (Figure 14), include "soil," "organic," "increase," "study," "result," and "carbon," which are central to discussions on soil management and conservation practices. The frequent co-occurrence of "organic soil" and "increase" suggests a strong research focus on increasing soil health. Additionally, terms such as "effect," "treatment," and "study" highlight the analytical nature of the dataset, suggesting that a significant portion of the research is dedicated to evaluating soil amendments, experimental treatments, and impacts on soil organic carbon. By applying an Apriori-based filtering, this visualization ensures that the words depicted have a strong statistical association rather than just individual frequency, providing a more meaningful interpretation of research trends in conservation tillage and soil sustainability.

To understand the trends surrounding sustainable agriculture, with the particular focus on the tillage practices, the timeline of the phrases of interest was created (Figure 15). It is evident that

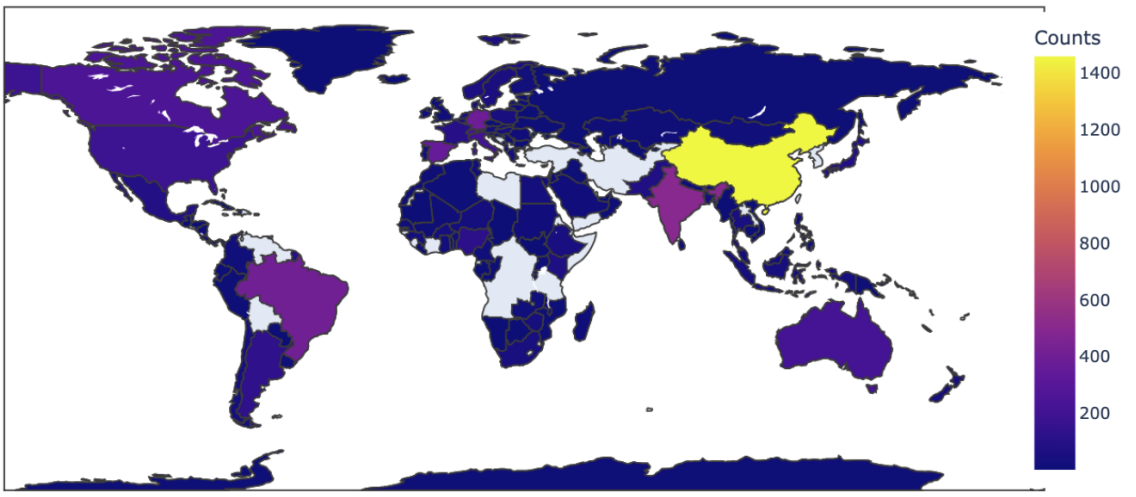

Figure 13: Heat-map of the number of articles per country. The visualization shows China's dominance in agricultural research publications, followed by India and Brazil.

conservation agriculture represented by No Tillage and Minimum Tillage has gained traction among researchers over the years, being mentioned nearly the same amount of times as conventional tillage. Considering that the number of publications has also been steadily increasing, these topics remain central to academic discourse.

Soil Organic Carbon is one of the most important, if not the most important indicator of soil health. It plays crucial role in determining soil physical and chemical properties. It is at the center of soil ability to produce as well as physical capabilities like water retention. This importance is also reflected in the number of articles that include organic carbon in their abstracts, with more than 57% of them including that reference (Figure 16).

4.2.1 Topic Modeling Using Latent Dirichlet Allocation

Applying Latent Dirichlet Allocation algorithm allowed for the automatic clustering of documents based on their word distributions, highlighting distinct thematic areas within the corpus. Classification is based on the probability of appearance of the certain key words within particular topic. Figure 9 shows the word distributions for the most significant topics identified by LDA. By analyzing these results, it becomes clear that topics revolve around soil composition, agricultural amendments and practices, and environmental impacts ("sequestration", "emission", "flux"), reinforcing the core themes of increasing soil organic matter (SOM), nutrient management, and soil health.

Topic 1 stands out as the second most dominant in the dataset and most relevant to the research question of this work. With a high count of 1572 occurrences, topic 1 indicates a strong research focus on tillage practices ("ct",'"nt"). The key terms within this topic include "soil," "tillage," "aggregate," "cm," "depth," and "system," suggesting an emphasis on how soil structure is influenced by different tillage methods and the implications for soil health. The presence of "tillage" and "aggregate" highlights the physical properties of soil, particularly how tillage practices can affect soil aggregation, porosity, and compaction.

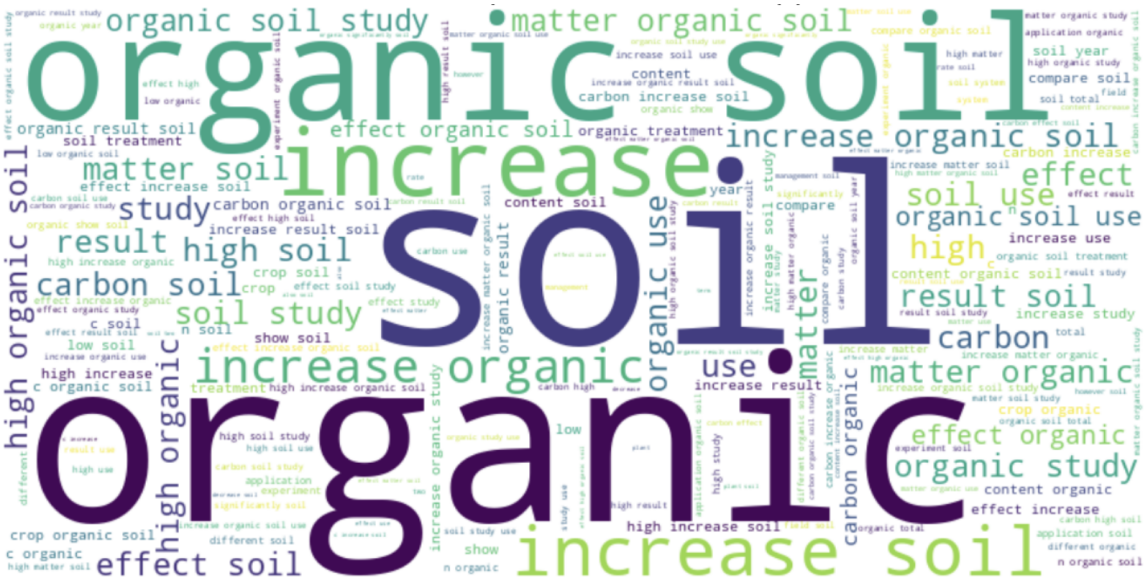

Figure 14: Apriori Model:Frequent Terms Word Cloud with support of 0.25. This word cloud visualizes the most frequently co-occurring terms in agricultural text analysis, filtered using an association rule mining approach with a support threshold of 0.25. Terms' size indicates their relative frequency and importance in the corpus.

Most represented, with 1720 occurrences, Topic 15 focuses on organic soil amendments, particularly compost and waste-derived inputs. The strong presence of "soil," "compost," "organic," "amendment," and "waste" suggests research on improving soil fertility through organic matter application. The emphasis on "increase" and "management" implies interest in nutrient release dynamics and long-term soil health benefits. This topic highlights the role of composting in agriculture.

Third in terms of articles count, with 1373 occurrences, Topic 19 focuses on crop production, soil management, and organic farming practices. The key terms "crop," "soil," "organic," "system," and "management" indicate research on how different cropping systems influence soil health and productivity. The presence of "production" and "practice" suggests studies on sustainable farming methods, including crop rotations and organic amendments.

4.3 Climate, crop type and soil type distribution across abstracts

4.3.1 Climate

Land that can be used for cultiviting crop occupies 14% of the ice-free land and pastures occupy additional 26 %. With almost half of the agricultural land being located in the dry climate (Africa and Asia). [61].

This representation of the dry climate aligns with the outcomes of analysis presented in the bar chart (Figure 17) 2,687 articles (15.92%) ("Dry": 1964 , "Arid": 723). Tropical (681 articles, 4.04%) climates follow as the next most frequently identified category. Low representation of the

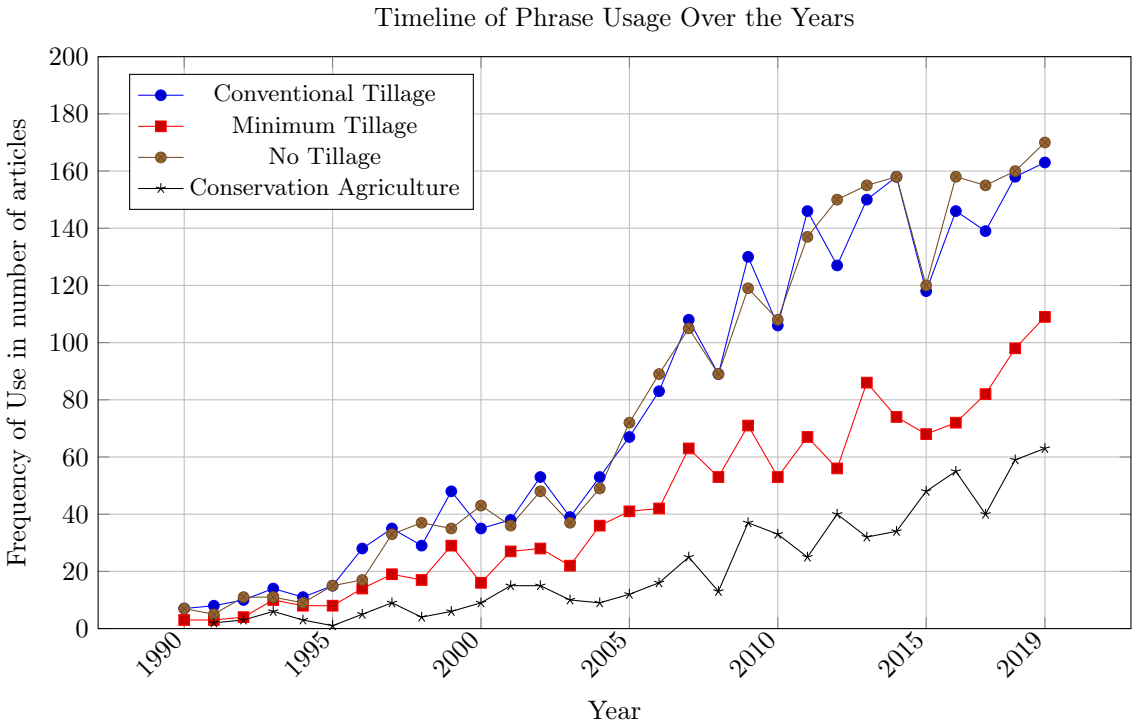

Figure 15: Timeline of Phrase Usage Over the Years. Conservation practices like No-Till (NT) and Minimum Tillage show significant increase in published articles since the early 2000s, reflecting a growing research interest in sustainable farming practices

climate type mentions within abstracts can suggests that findings related to other climates should be interpreted with caution.

4.3.2 Crop

The analysis of crop distribution across the literature reveals patterns in agricultural research focus (Figure 18). Wheat emerges as the dominant crop, with 2,568 articles (11.93% of total), followed by maize (1,873 articles) and rice (1,547 articles). This concentration on major cereal crops aligns with their global importance for food security and their substantial contribution to agricultural production [62]. The representation of these crops aligns with the geographic distribution observed in climate analysis.

The data presents a clear research focus toward crops crucial for global food security [63], with the top three cereals accounting for nearly 28% of all articles. This focus mirrors real-world agricultural priorities but also highlights potential gaps in research coverage for other important crops.

4.3.3 Soil

Sandy clay soil dominates with 784 articles, aligning with the prevalence of dry climate studies (2,687 articles, 15.92%) and wheat research (2,568 articles, 11.93%). This distribution, with China contributing 1461 articles, suggests concentrated research efforts in arid and semi-arid agricultural

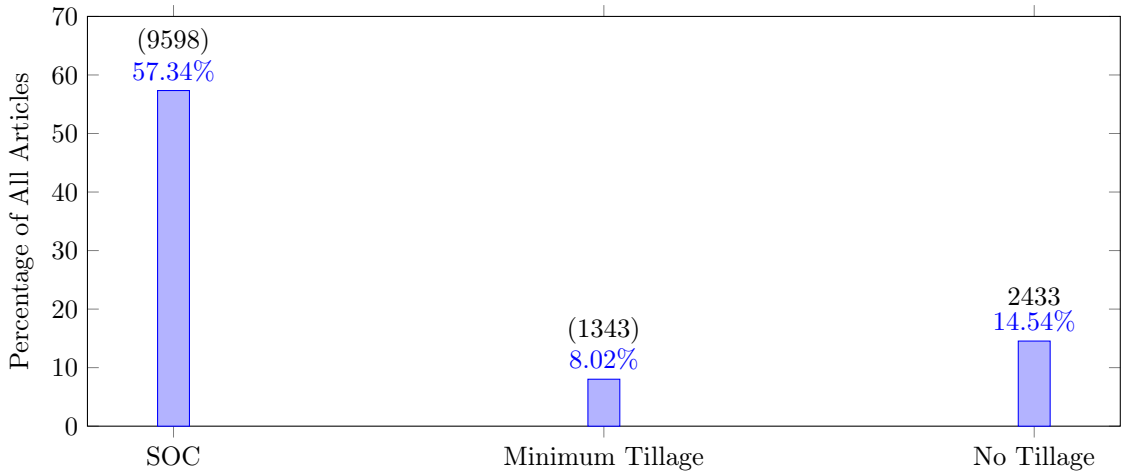

Figure 16: Articles including phrase of interest. More than half of articles was referring to SOC.

systems.

The second most represented type is clay soil (321 articles) followed by clay loam soil (249 articles). However, as in the example of the climates the other types of soil are not identified with enough frequency to be able to draw meaningful conclusions.

It should be noted that there is a mixture of taxonomies used in this graph. Loam and clay soils are classified by texture. Oxisol is USDA taxonomy, whereas Luvisol is WRB Taxonomy.

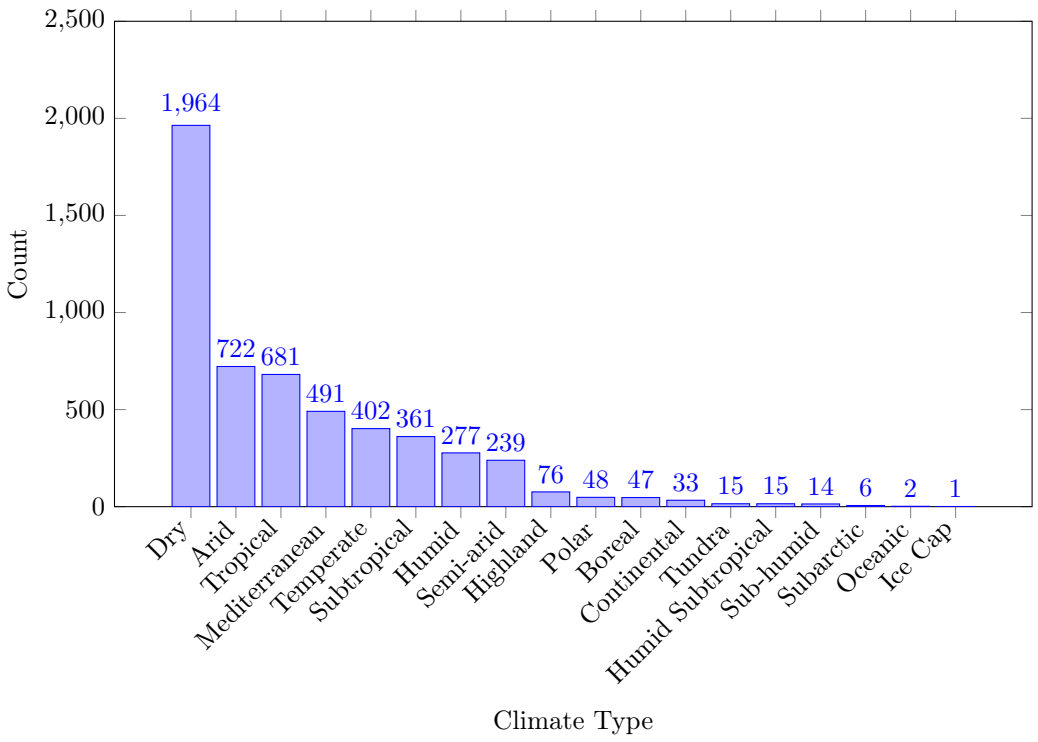

Figure 17: Distribution of Research Articles by Climate Type. The dominance of dry climates suggests significant interest in arid and semi-arid regions, while tropical and temperate zones also receive attention. Less studied climates, such as tundra, subarctic, and ice cap regions, reflect their unsuitability for agriculture due to extreme environmental conditions.

Topic	Characteristic	Count
15	15 - 0.052*"soil" + 0.044*"compost" + 0.038*"organic" + 0.021*"amendment" + 0.017*"matter" + 0.017*"increase" + 0.015*"waste" + 0.015*"application"	1720
1	0.063*"soil" + 0.062*"tillage" + 0.027*"nt" + 0.023*"aggregate" + 0.023*"cm" + 0.020*"ct" + 0.015*"depth" + 0.014*"system"	1572
19	19 - 0.024*"soil" + 0.021*"crop" + 0.020*"use" + 0.018*"system" + 0.015*"organic" + 0.015*"management" + 0.014*"production" + 0.011*"practice"	1373
3	0.043*"ha" + 0.027*"yield" + 0.025*"soil" + 0.024*"biochar" + 0.023*"application" + 0.021*"kg" + 0.018*"manure" + 0.016*"organic"	1208
17	17 - 0.113*"c" + 0.048*"soil" + 0.028*"fraction" + 0.027*"organic" + 0.020*"residue" + 0.017*"som" + 0.015*"microbial" + 0.011*"biomass"	1194
0	0.044*"soil" + 0.032*"biochar" + 0.015*"sorption" + 0.012*"organic" + 0.011*"concentration" + 0.011*"doc" + 0.010*"water" + 0.009*"carbon"	1072
12	12 - 0.046*"soil" + 0.016*"plant" + 0.012*"erosion" + 0.011*"tree" + 0.011*"specie" + 0.011*"water" + 0.010*"cover" + 0.008*"loss"	981
16	16 - 0.064*"crop" + 0.030*"wheat" + 0.027*"yield" + 0.024*"system" + 0.024*"rotation" + 0.021*"residue" + 0.019*"maize" + 0.019*"l"	916
11	11 - 0.044*"soil" + 0.042*"fertilizer" + 0.036*"treatment" + 0.034*"straw" + 0.031*"manure" + 0.030*"organic" + 0.025*"rice" + 0.023*"fertilization"	907
10	10 - 0.068*"soc" + 0.049*"c" + 0.034*"carbon" + 0.019*"change" + 0.018*"sequestration" + 0.017*"soil" + 0.017*"ha" + 0.016*"stock"	872
4	0.049*"soil" + 0.036*"community" + 0.030*"microbial" + 0.017*"bacterial" + 0.014*"diversity" + 0.013*"abundance" + 0.013*"structure" + 0.012*"composition"	828
5	0.055*"soil" + 0.020*"use" + 0.019*"model" + 0.011*"field" + 0.010*"datum" + 0.009*"study" + 0.009*"value" + 0.009*"analysis"	793
2	0.203*"n" + 0.022*"kg" + 0.021*"nitrogen" + 0.020*"soil" + 0.019*"ha" + 0.017*"rate" + 0.016*"fertilizer" + 0.012*"application"	698
8	8 - 0.052*"soil" + 0.041*"metal" + 0.035*"zn" + 0.030*"cu" + 0.029*"cd" + 0.024*"concentration" + 0.021*"pb" + 0.020*"heavy"	680
9	9 - 0.068*"soil" + 0.024*"cm" + 0.023*"land" + 0.019*"organic" + 0.017*"carbon" + 0.017*"depth" + 0.016*"site" + 0.015*"use"	622
18	18 - 0.067*"soil" + 0.050*"activity" + 0.033*"plant" + 0.031*"microbial" + 0.024*"root" + 0.023*"growth" + 0.020*"biomass" + 0.015*"enzyme"	534
14	14 - 0.087*"emission" + 0.076*"co" + 0.040*"soil" + 0.028*"flux" + 0.020*"ch" + 0.019*"gas" + 0.018*"earthworm" + 0.017*"temperature"	311
7	0.092*"soil" + 0.030*"ph" + 0.018*"clay" + 0.018*"increase" + 0.017*"property" + 0.015*"content" + 0.014*"ca" + 0.014*"om"	305
6	0.231*"p" + 0.036*"k" + 0.033*"soil" + 0.032*"lt" + 0.026*"mg" + 0.026*"r" + 0.022*"kg" + 0.020*"gt"	138
13	13 - 0.131*"sludge" + 0.082*"sewage" + 0.030*"ss" + 0.022*"disease" + 0.017*"effluent" + 0.016*"urban" + 0.014*"pathogen" + 0.013*"wastewater"	14

Table 9: Top Words for Dominant Topics Identified by LDAThe table presents the key terms associated with each topic extracted using Latent Dirichlet Allocation (LDA). The weighting of words reflects their importance within a given topic, providing insights into the thematic structure of the research corpus.

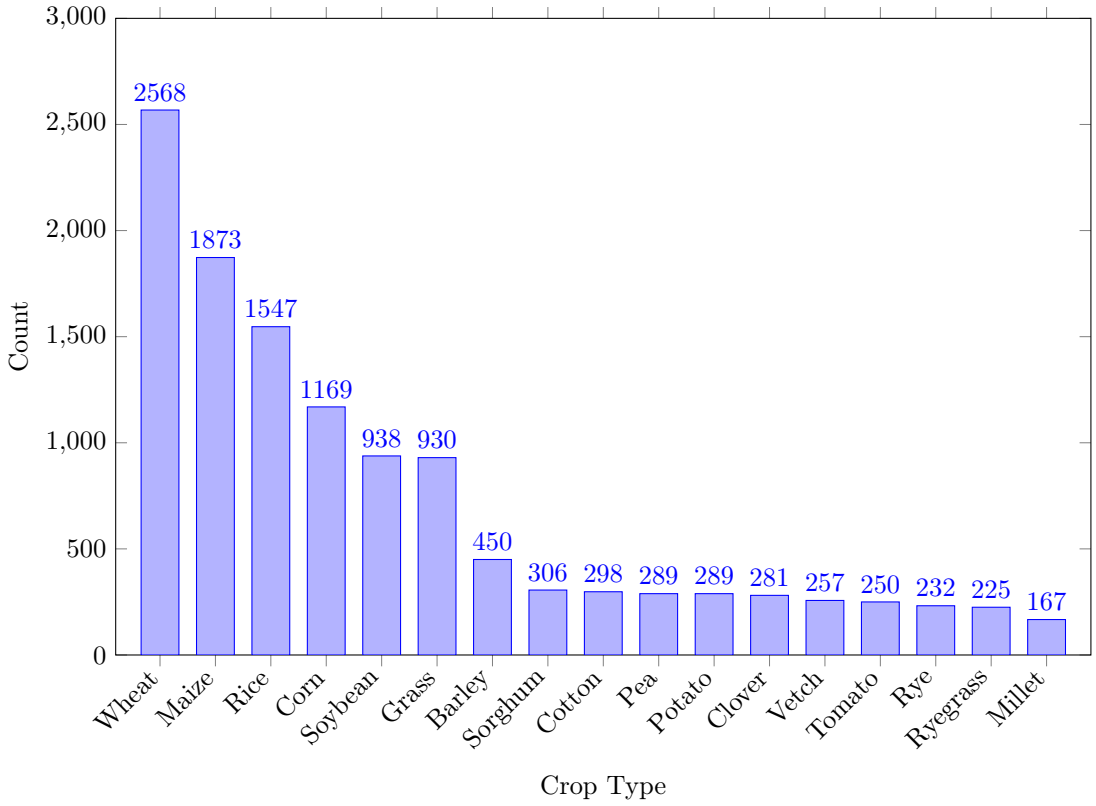

Figure 18: Distribution of Crop Types in Agricultural Publications. Wheat dominates agricultural research with 2,568 articles (11.93%), followed by maize and rice, highlighting research concentration on major cereal crops that are crucial for global food security.

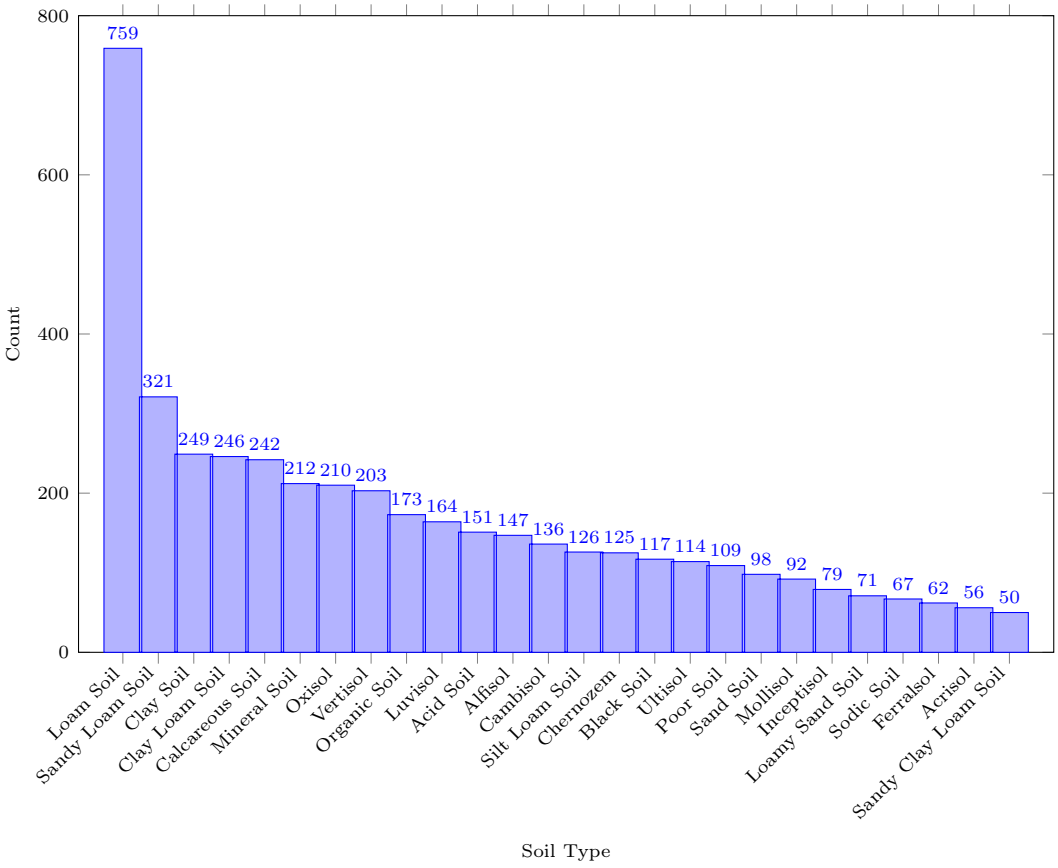

Figure 19: Distribution of Soil Types in Agricultural Research Literature. Distribution of Soil Types in Agricultural Research Literature. Sandy clay soil dominates with 784 articles, reflecting research concentration in regions with within dry climates and major cereal-producing areas.

Chapter 5

Results

This section presents the outcomes of the applied text mining and NLP techniques to answer the key question posed by this research: How do no-tillage and minimum tillage practices impact soil organic carbon (SOC)? The results are structured to first gain insights from a custom model leveraging an agriculture-specific ontology and then enhance these findings with topic modeling results. These results are subsequently compared against transformer-based models, including BERT (specifically RoBERTa) and GPT-3.5/4.

The comparative analysis is conducted using a validation set consisting of 200 randomly selected abstracts, which were manually classified to serve as ground truth for model evaluation.

Finally, the section explores the identification of causal relationships within the extracted data, assessing the extent to which different models can accurately capture the causal link between NT/MT and SOC changes.

5.1 Text Mining using Custom Model

As outlined in Chapter 3 Methodology, entity recognition was performed using SpaCy in combination with a rule-based matching model to extract domain-specific terms related to agriculture, soil health, and conservation tillage practices. This approach ensured that the model effectively captured key terms and relationships essential for addressing the research question.

The text mining pipeline consisted of the following steps:

1. Preprocessing: The dataset was prepared for further analysis by cleaning the data and identifying missing values.
2. Ontology-Based Entity Matching: A domain-specific ontology was developed to capture terms related to soil organic carbon changes, tillage types, and causal indicators such as increase, decrease, and no effect.
3. Dependency Parsing for Context Extraction: A dependency parsing model was used to detect relationships between entities in the text.
4. Co-Occurrence Analysis: The frequency and proximity of relevant terms were analyzed to identify correlations between tillage practices and SOC changes.

The key advantage of this approach is that it allows for a structured extraction of meaningful insights while maintaining interpretability, unlike black-box machine learning models.

5.1.1 Entity Recognition with Rule-Based Matching and Dependency Parsing Model

Entity recognition was applied to extract key terms associated with tillage practices and SOC changes, as well as different characteristics that should address the research question. The terms "increase", "decrease", and "no-tillage" were then aggregated and prepared for visualizations using a custom dictionary tailored to recognize SOC indicators, agricultural practices, and relevant characteristics.

Entity recognition was applied to extract key terms associated with tillage practices and soil organic carbon (SOC) changes, as well as other relevant characteristics to address the core research question. The labels "increase", "decrease", "SOC increase", "SOC decrease" were recognized as indicators of SOC impact, "no tillage" and "minimum tillage" labels were used to identify the practice. Labels of interest for the tillage SOC impact are presented in the Table 4

This approach provided a structured, rule-based framework to ensure precise term extraction while minimizing misclassification and false positives.

5.1.2 Co-Occurrence Analysis of NT/MT Practices and SOC Changes

The results of the co-occurrence analysis provide insights into how frequently no-tillage (NT) and minimum tillage (MT) practices appear in abstracts alongside positive SOC changes. Table 10 presents the outcome of the filtering and grouping process, as illustrated in Figure 10, to identify the final set of abstracts supporting the statement that NT/MT practices have a positive impact on SOC.

A co-occurrence analysis was performed to assess how frequently NT/MT practices were mentioned within the same abstract as positive SOC changes. The results indicate that 727 abstracts (4.3%) contained both "SOC Increase" and "NT/MT" within the same document.

	Co-occurence	**Distance =0**	**Distance betwen 1 and 3**	**Distance between 3 and 5**
Abstracts Count	727	192	315	68
Statement example classified correctly		"NT increased SOC compared with CT by 158% in macroaggregate fractions, but only 40% in 250-um fraction"	Aggregate dry mean weight diameter and stability in water were 1.2 and 2.2 times greater, respectively, under no-tillage than conventional tillage due to reduced mechanical disturbance and increased soil organic carbon content.	"Legumes coupled with no tillage reduced the N fertilizer requirement of corn, increased plant-available N, and augmented total soil C and N stores"

Table 10: Number of abstracts that includes terms SOC Increase + No Tillage/ Minimum Tillage (all combinations) with the distance between those entities. Distance=0 source [64]
, Distance Max 3 source [65], Distance Max 5 source [66]

When filtering for abstracts where no additional entities appeared between key terms (i.e., Distance = 0), only 192 abstracts remained. These cases represent the strongest linguistic connection between NT/MT and SOC Increase, where the impact is explicitly stated rather than inferred through broader context.

An example of such a directly causal abstract is:

> *"From a review of 20 studies in the region, SOC increased with no tillage compared with conventional tillage by 0.48 ± 0.56 Mg C ha-1 yr-1"* [67]

This example shows clearly the causal relationship between no-tillage practices and SOC accumulation, suggesting that when key terms appear in close proximity, the likelihood of a direct causal inference is higher.

5.1.3 Addressing Confounding Factors and Refining Causal Inference

A critical assumption was made that co-occurrence should not be considered sufficient evidence to imply a causal relationship. This is particularly relevant in cases where multiple agronomic factors contribute to SOC changes, potentially introducing confounding variables, as well as to the case were focus of the study is on different practice.

For example, in the following abstract, while NT/MT practices are mentioned in proximity to SOC sequestration, the presence of other contributing factors (such as plant cover, organic matter amendments, etc.) complicates the attribution of SOC changes solely to tillage practices:

> "*A promising option to sequester carbon (C) in these cropping systems is the implementation of recommended management practices (RMPs), which include plant cover in the inter-row area, minimum or no tillage and off- and on-farm organic matter amendments. However, the effects of RMPs on soil organic carbon (SOC) stocks in these cropping systems are widely overlooked, despite the critical importance of estimating their contribution on CO2 emissions for policy decisions in the agriculture sector in Mediterranean regions. We therefore conducted a meta-analysis to derive a C response ratio, soil C sequestration rate and soil C sequestration efficiency under RMPs, compared to conventional management of olive and almond orchards, and vineyards (144 data sets from 51 references). RMPs included organic amendments (OA), plant cover (CC) and a combination of the two (CMP). The highest soil C sequestration rate (5.3 t C ha−1 yr−1) was observed following the application OA in olive orchards (especially after olive mill pomace application), whereas CC management achieved the lowest C sequestration rates (1.1, 0.78 and 2.0 t C ha−1 yr−1, for olive orchards, vineyards and almond orchards, respectively). Efficiency of soil C sequestration was greater than 100% after OA and CMP managements, indicating that: i) some of the organic C inputs were unaccounted for, and ii) a positive feedback effect of the application of these amendments on SOC retention (e.g. reduction of soil erosion) and on protective mechanisms of the SOC which reduce CO2 emissions.*" [68]

In this instance, SOC increases were likely influenced by additional factors, such as organic amendments and plant cover. Consequently, it would be inappropriate to conclude that NT/MT practices alone were responsible for the observed SOC changes.

Rather than explicitly excluding other agricultural practices or addressing confounding factors directly, this was managed indirectly by restricting the analysis to abstracts where key terms related to NT/MT and SOC Increase appeared within a maximum distance of five entities. This constraint was applied to ensure a higher likelihood of contextual relevance between the two concepts while

minimizing the risk of confounding influences from unrelated agricultural practices. As a result, abstracts where these entities were separated by more than five terms were omitted from further analysis.

5.1.4 Custom Model Miss-classifications and potential for improvement

After thoroughly analyzing the validation set, several key observations emerged that could enhance the custom model's ability to detect causality more effectively.

One of the primary areas for improvement is expanding the ontology to include terms related to conservation tillage, as it serves as an umbrella term for both no-till (NT) and minimum tillage (MT) practices. The absence of this term in the ontology led to misclassifications, highlighting a key limitation of dictionary-based models. While such models can leverage domain-specific terminology effectively, their performance is inherently constrained by the completeness of the ontology. Ontology-based models will always be as effective as the knowledge embedded in the ontology itself.

A notable example of this limitation is the following sentence, which was not classified as a true positive:

> *Amount of soil organic C (SOC) present will increase with specific land-cover/use changes (e.g., conservation tillage, mulching, agroforestry), and will be reduced or even eliminated by others.*[69]

Here, "conservation tillage" is directly associated with SOC increase, yet the model failed to recognize this causal link due to the term's absence from the predefined ontology. Interestingly, 5 out of 10 false positives referenced conservation tillage as having a positive impact on SOC, further supporting the need for its inclusion in the model's knowledge base.

Another key limitation in the custom model's design is its inability to detect reverse causality—cases where conventional tillage (CT) is explicitly shown to reduce SOC, which indirectly implies that NT/MT either maintains or increases SOC levels. This oversight stems from the fact that the model primarily focuses on direct mentions of NT/MT rather than inverse relationships.

An example of a misclassified sentence demonstrating reverse causality is:

> *The different tillage systems showed a significant effect with respect to the amount of organic matter and aggregate stability in the soil. Organic matter values were lower (49-60%) under CT practices, and residue burning accelerated the loss of organic carbon content. The highest aggregate stability values were found for NT (over 38%)."*[70]

In this case, the model did not classify it as a positive NT/MT impact, despite the evidence that conventional tillage reduced SOC, which inherently suggests that NT/MT had a relatively beneficial effect. Addressing this issue would require modifying the model to recognize indirect causal patterns, particularly when conventional tillage is shown to deplete SOC.

By refining the ontology and enhancing the model's ability to capture both direct and inverse causal relationships, future iterations of the custom model could achieve higher classification accuracy and a more comprehensive understanding of tillage-related SOC changes

5.1.5 Tillage impact on SOC and association with dominant topic

To further explore the textual patterns associated with the impact of no-tillage (NT) and minimum tillage (MT) on soil organic carbon (SOC), the outcomes of topic modeling analysis using Latent Dirichlet Allocation (LDA) were compared against results of custom model.

Table 11 presents the most prevalent topics identified through Latent Dirichlet Allocation (LDA) within the subset of abstracts where the custom model identified an association between tillage and SOC impact. The topic modeling results reveal that a significant majority of abstracts reporting a positive impact of No-Tillage (NT) on Soil Organic Carbon (SOC) are concentrated within a small subset of topics. Specifically, the top 6 out of 19 topics account for 80.3% of all NT-positive impact abstracts (653 out of 727). This indicates that discussions on NT and SOC enhancement tend to cluster around a few dominant themes in the literature.

Additionally, in some topic clusters we can see that the fraction of negative impact of SOC is increasing (topic 17), which could show the specific characteristics in which determining impact can be harder, potentially due to confounding factor of microbial activity etc.

Topic	**Dominant Topic Top Words**	**Dataset occurrence Count**	**NT on SOC Positive Impact Count**
1	0.063*"soil" + 0.062*"tillage" + 0.027*"nt" + 0.023*"aggregate" + 0.023*"cm" + 0.020*"ct" + 0.015*"depth" + 0.014*"system"	1720 (9.4%)	295 (40.1%)
10	0.068*"soc" + 0.049*"c" + 0.034*"carbon" + 0.019*"change" + 0.018*"sequestration" + 0.017*"soil" + 0.017*"ha" + 0.016*"stock"	872 (4.8%)	155 (21.3%)
16	0.064*"crop" + 0.030*"wheat" + 0.027*"yield" + 0.024*"system" + 0.024*"rotation" + 0.021*"residue" + 0.019*"maize" + 0.019*"l"	916 (5.0%)	64 (8.8%)
19	0.024*"soil" + 0.021*"crop" + 0.020*"use" + 0.018*"system" + 0.015*"organic" + 0.015*"management" + 0.014*"production" + 0.011*"practice"	1373 (7.5%)	55 (7.6%)
17	0.113*"c" + 0.048*"soil" + 0.028*"fraction" + 0.027*"organic" + 0.020*"residue" + 0.017*"som" + 0.015*"microbial" + 0.011*"biomass"	1194 (6.5%)	48 (6.6%)
9	0.068*"soil" + 0.024*"cm" + 0.023*"land" + 0.019*"organic" + 0.017*"carbon" + 0.017*"depth" + 0.016*"site" + 0.015*"use"	622 (3.4%)	36 (5%)

Table 11: Tillage impact on SOC and association with dominant topics.

5.2 Model Evaluation and Performance Analysis

This section presents a comparative evaluation of different models used to classify the relationship between No-Tillage (NT), Minimum Tillage (MT), and Soil Organic Carbon (SOC) changes. The performance of these models is assessed through confusion matrix metrics and classification scores, focusing on precision, recall, F1-score, and specificity.

Confusion Matrix

The confusion matrix results (Table 12) shows key differences in model performance, particularly in terms of true positives (TP) and false positives (FP). The custom distance-based models, which apply different maximum entity distances (Dist max 3 and Dist max 5), shows an improvement when increasing the distance in detecting true positives, while maintaining the same number of false positives as the more restrictive Dist max 3 model. A followup analysis could be conducted to understand which distance would bring the best result between accuracy of the model and false positive rate.

Among the transformer-based models, GPT-4 stands out as the most effective, identifying the highest number of true positives while keeping the false positive rate low. GPT-3, in contrast, struggles with specificity, producing the highest false positive count, which suggests that it tends to over lassify abstracts as relevant even when the relationship between NT/MT and SOC impact is weaker. This also might mean that it is analysing the abtracts with the confirmation bias for the initial hypothesis presented in the prompt.. BERT Zero-Shot, although not fine-tuned for this specific task, maintains a balanced performance, achieving moderate levels of precision and recall. Considering the task at hand, the ideal approach would be to minimize false negative rates as much as possible while maintaining a balanced rate of false positives. This would allow for the identification of nearly all cases of causal relationships while keeping the error rate of falsely identifying causality—where none exists—low. Taking this goal into account, the custom model still has the potential for improvement by identifying the optimal cutoff point for the distance between entities of interest.

Model	**TP**	**TN**	**FP**	**FN**
Actual	25	175	-	-
Custom Model Dist max 3	11	170	5	14
Custom Model Dist max 5	15	170	5	10
GPT-4	22	172	3	3
GPT-3	20	101	74	5
BERT Zero Shot	14	166	9	11

Table 12: Table presents the confusion matrix metrics (True Positives, True Negatives, False Positives, and False Negatives) for different models applied in the study. The models include baseline performance, custom distance-based models, GPT-3, GPT-4, and a zero-shot BERT classifier.

5.3 Classification Metrics: Accuracy, Precision, Recall, and Specificity

A deeper analysis of accuracy, precision, recall, and F1-score (Table 13) provides further insights into each model's ability to correctly classify relationships in the dataset. The custom model with Dist max 5 improves recall (0.600) and achieves a classification accuracy of 93.5%, demonstrating that increasing the allowed distance between NT/MT and SOC-related terms improves the model's capacity to detect relevant abstracts.

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Baseline	100.0%	-	-	-	-
Custom Model Dist max 3	91.2%	0.688	0.440	0.537	0.971
Custom Model Dist max 5	93.5%	0.750	0.600	0.667	0.971
GPT-4	97.6%	0.880	0.880	0.880	0.983
GPT-3	70.8%	0.213	0.800	0.333	0.577
BERT Zero Shot	86.5%	0.609	0.560	0.583	0.949

Table 13: Table presents a comparative analysis of accuracy, precision, recall, F1 score, and specificity across different models. The metrics provide insights into each model's ability to balance classification performance.

GPT-4 consistently outperforms all other models, achieving an accuracy of 97.6% and a precision, recall, and F1-score of 0.880, making it the most reliable classifier for this task. The model's high recall and precision indicate that it successfully identifies relevant abstracts while minimizing misclassifications. In contrast, GPT-3 exhibits low precision (0.213) but high recall (0.800), meaning that it classifies many abstracts as relevant but does so at the expense of specificity. This suggests that while GPT-3 is effective at capturing possible SOC impact mentions, it frequently includes irrelevant abstracts, reducing its reliability.

BERT Zero-Shot, despite not being fine-tuned for the dataset, performs relatively well, with an accuracy of 86.5% and a balanced precision (0.609) and recall (0.560). This result highlights the model's adaptability and its ability to generalize across different scientific texts without requiring domain-specific training.

5.4 Model Limitations and Implications for Agricultural Research

Although this thesis provides information on the relationships between tillage practices (No Tillage/Minimum Tillage) and soil organic carbon (SOC), certain limitations must be taken into account. One of the main constraints lies in the reliance of transformer-based models, such as GPT-3.5 and GPT-4, on their training data and the biases present within it. These models do not possess true domain expertise but rather generate outputs based on probability distributions, which in some cases may lead to the misclassification of scientific findings or the overgeneralization of relationships between agricultural practices and SOC changes. The challenge of ensuring factual accuracy remains and continues to require human verification.

Another limitation comes from the domain-specific nature of the dataset. The abstracts analyzed in this study are structured differently from full-length scientific papers, policy reports, or other relevant sources. Although abstracts provide a summary of research findings, they may lack critical methodological details or nuanced discussions that are essential to establish causality.

The approach of the study to causal inference is also limited by the methods employed. Named Entity Recognition (NER) and dependency-based pattern matching, although effective in identifying associations, do not explicitly establish causation [56]. The presence of key terms within proximity does not necessarily indicate a direct cause-and-effect relationship, especially in cases where multiple factors influence SOC changes. The complexity of agricultural systems—where climatic conditions, soil properties, and crop management practices interact in complex ways—makes extracting causal conclusions from text mining particularly challenging. While these techniques help uncover patterns

in the literature, they do not replace controlled experimental studies or meta-analyses. Rather, they serve as a useful tool for selecting relevant papers for further investigation.

From a computational perspective, accessibility remains an important consideration. Transformer-based models require significant computational resources, making them less practical for researchers or institutions with limited access to high-performance computing. While rule-based approaches and traditional NLP methods offer more transparent and computationally efficient alternatives, they often lack the flexibility and adaptability of deep learning models. s demonstrated in our case, the simple omission of 'Conservation Tillage' from the ontology resulted in limiting the number of correctly identified abstracts. The trade-off between interpretability and predictive power remains an ongoing challenge when applying NLP techniques in specialized research domains like agriculture.

Despite these limitations, the study presents significant implications for both agricultural research and policymaking. The ability to extract structured insights from vast amounts of scientific literature enhances the potential for evidence-based decision-making in soil management and sustainable agriculture. By automating the identification of research trends and associations, text mining techniques can accelerate knowledge synthesis, reducing the time required to aggregate findings from multiple studies. This is particularly valuable in the context of sustainability research, where timely and informed decisions are crucial for addressing challenges related to soil degradation and climate change.

Beyond research, there is also potential for these methodologies to inform agricultural policy. Understanding the broad landscape of scientific findings can help policy makers design more targeted and region-specific soil conservation strategies. Different tillage practices can produce varying results depending on soil conditions, climate, and crops. Automating the extraction of research insights can facilitate the adaptation of policies to the unique challenges faced by different agricultural regions, ensuring that recommendations are grounded in empirical evidence rather than generalized assumptions.

There is also an opportunity to bridge the gap between academic research and practical agricultural applications. Farmers and agronomists often struggle to navigate the vast and growing body of scientific literature relevant to sustainable soil management. If refined and integrated into user-friendly platforms, language processing tools could play a role in making research findings more accessible to practitioners. This could support knowledge transfer efforts, enabling farmers to adopt best practices based on the latest scientific evidence without requiring direct engagement with complex academic papers.

Chapter 6

Conclusions and future work

The results demonstrate that increasing the allowed distance between entities improves classification by enhancing recall while maintaining acceptable precision. The superior performance of GPT-4 confirms the advantage of transformer-based models in complex text classification tasks, particularly in scientific literature analysis. However, the trade-offs observed in GPT-3's high recall but low precision indicate a tendency to overclassify, which could lead to the inclusion of irrelevant abstracts.

Custom rule-based models remain a viable approach, particularly for cases requiring structured extraction, but they lack the flexibility to handle nuanced contextual relationships as effectively as transformers. The BERT Zero-Shot model provides an alternative that balances generalization and specificity, demonstrating its potential for classification tasks where fine-tuning is not feasible.

Beyond performance, cost considerations play a crucial role in model selection. BERT and the custom rule-based model are entirely free, making them practical choices for large-scale document classification, especially when computational resources are limited. In contrast, using GPT-4 and GPT-3 for analyzing 200 abstracts incurred a processing cost of approximately 2$, which could be a limiting factor when scaling to larger datasets. While GPT-4 delivers the highest accuracy and robustness, its usage comes at a financial cost, making it less feasible for routine large-scale applications unless a cost-benefit trade-off is carefully considered.

These findings highlight that while transformers like GPT-4 are becoming increasingly good in terms of accuracy and contextual understanding, however cost constraints may favor alternative models such as BERT or rule-based approaches, depending on the scale and objectives of the analysis. Future research should explore hybrid models that combine the affordability of rule-based methods with the flexibility of transformer architectures, optimizing both efficiency and cost-effectiveness.

6.1 Implications for Agricultural Research and Policy Makers

Agronomists typically rely on manual literature reviews to stay informed about best practices in soil preservation and restoration. However, with the exponential growth of scientific publications, manually identifying relevant research has become an increasingly time-consuming and labor-intensive process. As the volume of agronomic research continues to expand, it becomes more challenging to synthesize findings efficiently and translate them into actionable recommendations. The integration of advanced transformer-based models and text mining techniques presents a unique opportunity to automate the extraction of key insights, significantly reducing the effort required to stay up to date with the latest advancements in conservation tillage and soil health. The application of data-driven methodologies in agricultural research has the potential to bridge the gap between science

and practice, ensuring that evidence-based policies are developed using the most comprehensive and up-to-date knowledge available.

The Text Mining and NLP Transformer based models applied in this study offer an efficient and scalable solution capable of processing vast amounts of scientific abstracts to extract meaningful insights on NT/MT practices and their impact on SOC. By automating literature analysis, these techniques can help detect emerging trends, identify research gaps, and support systematic reviews in soil conservation research. [71]The same approach could be extended to analyze the impact of other agricultural practices, such as cover cropping, crop rotation, and organic amendments, allowing researchers to compare multiple strategies for enhancing soil carbon sequestration and sustainability.

Researchers are increasingly using machine learning and text mining to understand better different soil health indicators, assess soil conservation practices and other important issues that farming faces nowadays. [72] [73]. Given the pressing challenges surrounding soil degradation and carbon sequestration, this is an area where further exploration is not only valuable but necessary.

Appendix

6.2 Crop Type Distribution in Research vs. Global Production

Understanding the representation of different crop types within the analyzed dataset is essential to put in the perspective the findings on no-tillage (NT) and minimum tillage (MT) practices and their impact on soil organic carbon (SOC). The distribution of crop types in the dataset was compared against global crop production data obtained from the Food and Agriculture Organization (FAO) [62].

Crop category	**Dataset occurrence Count**	**NT on SOC Positive Impact Count**	**World data by FAO in %**	**World data by FAO in million ton**
Cereals	6395	425	32.5%	3008
Leguminous	1545	182	Missing data	91
Grasses and fodder	1754	104	Missing data	310
Fiber	366	42	Missing data	31.8
Vegetable	1040	19	12%	1144
Oilseed	311	19	12%	417
Sugar	160	8	23%	2132
Root and tuber	354	7	9%	877
Other	45	3	Missing data	Missing data
Beverage	170	1	Missing data	Missing data
Unknown	8297	236		

Table 14: Comparison of crop category distributions among different subsets of articles and the word production data coming from FAO report [62]

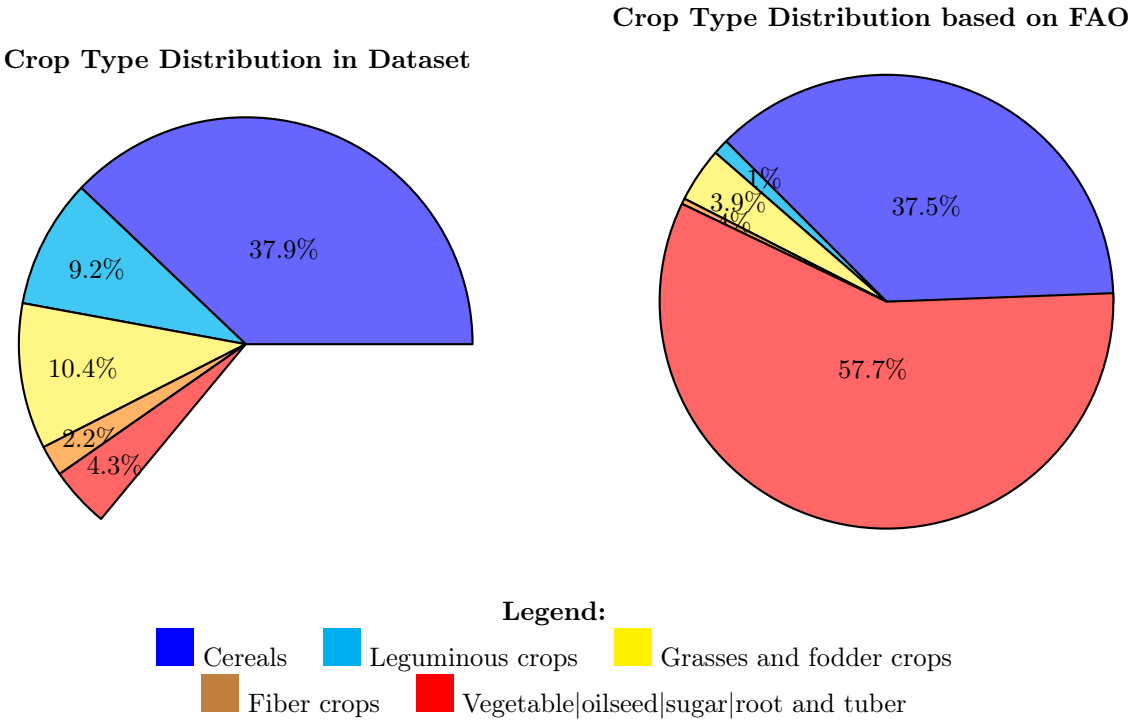

Figure 20: Comparison of crop category distributions based on "New Data Count" and "Million Tones".

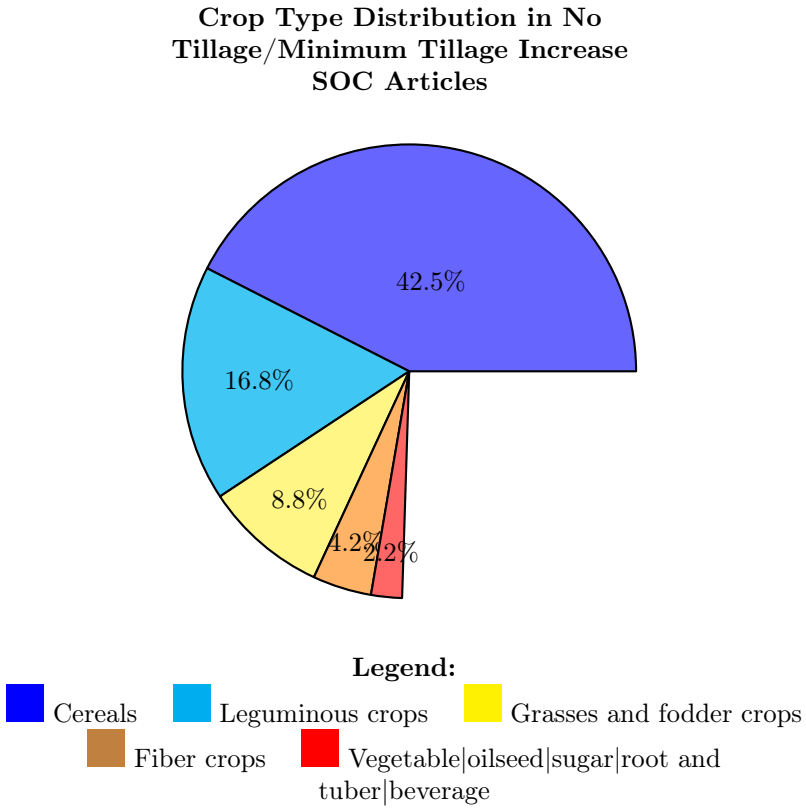

Figure 21: Distribution of Crop category within articles for NT/MT practices that Increase SOC (excluding "not present").Important to note: article can have more than one crop category mentioned in abstract, however "not present" means that there was no crop specified whatever.

Key Observations

These findings emphasize that while research on NT/MT practices covers a broad spectrum of crops, there is a potential bias toward specific categories that could be particularly relevant to soil health and carbon sequestration research. This insight is crucial for understanding how we can generalize the findings across different agricultural systems.

- Cereal crops: The prevalence of cereals in both datasets confirms their central role in global agriculture and their significant representation in conservation tillage studies.
- Leguminous crops: Their higher occurrence in the research dataset may indicate a focus on nitrogen-fixing species, which are commonly associated with sustainable farming systems.
- Grasses and fodder crops: While frequently appearing in research, their global production share is relatively small, highlighting a potential research bias towards soil-related studies rather than food production.
- Underrepresented categories: Root and tuber crops, sugar crops, and oilseeds are relatively less frequent in conservation tillage research compared to their global production significance.

6.3 Code Repository

The implementation of the text mining and NLP techniques used in this study has been documented in publicly accessible code repository in Github. Repository contain the scripts for data preprocessing, entity recognition, topic modeling, causal inference analysis, and model evaluation, ensuring full reproducibility of the research.

The following repository is available for reference: https://www.github.com/olalepek/Text–Mining——Agriculture/

Bibliography

[1] Benjamin Schneider, Jeffrey Alexander, and Patrick Thomas. *Publications Output: U.S. Trends and International Comparisons | NSF - National Science Foundation*. Tech. rep. National Center for Science and Engineering Statistics, Dec. 2023. URL: https://ncses.nsf.gov/pubs/nsb202333/publication-output-by-region-country-or-economy-and-by-scientific-field#utm_source=chatgpt.com.

[2] FAO. *The State of Food and Agriculture 2024 – Value-driven transformation of agrifood systems*. Food & Agriculture Org, 2024.

[3] *Healthy Soils for a Healthy People and Planet: FAO Calls for Reversal of Soil Degradation*. Publication Title: Newsroom. URL: https://www.fao.org/newsroom/detail/agriculture-soils-degradation-FAO-GFFA-2022/en.

[4] F.A.O.U. Nations et al. *State of knowledge of soil biodiversity - Status, challenges and potentialities: Report 2020*. FAO, 2020. ISBN: 978-92-5-133582-6. URL: https://books.google.it/books?id=1zoQEAAAQBAJ.

[5] Stacey Noel et al. *Report for policy and decision makers: Reaping economic and environmental benefits from sustainable land management*. Tech. rep. Sept. 2015.

[6] *UNCCD Data Dashboard Sustainable Development Goal (SDG) Indicator 15.3.1*. URL: https://data.unccd.int/land-degradation.

[7] *Goal 15 Department of Economic and Social Affairs*. URL: https://sdgs.un.org/goals/goal15.

[8] *Great Green Wall Initiative*. Publication Title: UNCCD. URL: https://www.unccd.int/our-work/ggwi.

[9] *Greening the Sahel*. Oct. 2022. URL: https://www.wfp.org/publications/greening-sahel.

[10] *CAP at a glance - European Commission*. en. Jan. 2025. URL: https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance_en.

[11] Directorate-General for Environment European Commission. *Soil strategy - European Commission*. en. 2024. URL: https://environment.ec.europa.eu/topics/soil-and-land/soil-strategy_en.

[12] William Horwath. *Improving soil health*. eng. Burleigh Dodds series in agricultural science ; Number 109. Publication Title: Improving soil health. Cambridge, England ; Burleigh Dodds Science Publishing, 2023. ISBN: 1-78676-672-8.

[13] *Soil Health Natural Resources Conservation Service*. Nov. 2024. URL: https://www.nrcs.usda.gov/conservation-basics/natural-resource-concerns/soils/soil-health.

[14] *Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper*. en. URL: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[15] Charu C. Aggarwal and ChengXiang Zhai. *Mining Text Data*. en. Springer Science & Business Media, Feb. 2012. ISBN: 978-1-4614-3223-4.

[16] Anne Kao and Stephen R. Poteet, eds. *Natural Language Processing and Text Mining*. en. London: Springer, 2007. ISBN: 978-1-84628-175-4. DOI: 10.1007/978-1-84628-754-1. URL: http://link.springer.com/10.1007/978-1-84628-754-1.

[17] *Linguistic Features · spaCy Usage Documentation*. en. URL: https://spacy.io/usage/linguistic-features.

[18] Marie-Catherine Marneffe et al. "Universal Dependencies". In: *Computational Linguistics* 47 (July 2021), pp. 255–308. DOI: 10.1162/coli_a_00402.

[19] Christopher D. Manning. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" en. In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander F. Gelbukh. Berlin, Heidelberg: Springer, 2011, pp. 171–189. ISBN: 978-3-642-19400-9. DOI: 10.1007/978-3-642-19400-9_14.

[20] James H. Martin and Daniel Jurafsky. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition / Daniel Jurafsky and James H. Martin ; contributing writers: Andrew Kehler, Keith Vander Linden, and Nigel Ward*. eng. Upper Saddle River (NJ): Prentice Hall, 2000. ISBN: 978-0-13-095069-7.

[21] Stephen Soderland. "Learning Information Extraction Rules for Semi-Structured and Free Text". en. In: *Machine Learning* 34.1 (Feb. 1999), pp. 233–272. ISSN: 1573-0565. DOI: 10.1023/A:1007562322031.

[22] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. en. 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.

[23] Florian Heimerl et al. "Word Cloud Explorer: Text Analytics Based on Word Clouds". In: *2014 47th Hawaii International Conference on System Sciences*. ISSN: 1530-1605. Jan. 2014, pp. 1833–1842. DOI: 10.1109/HICSS.2014.231. URL: https://ieeexplore.ieee.org/document/6758829/?arnumber=6758829.

[24] Karen Spärck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 60.5 (Jan. 2004). Publisher: Emerald Group Publishing Limited, pp. 493–502. ISSN: 0022-0418. DOI: 10.1108/00220410410560573.

[25] David Blei, Andrew Ng, and Michael Jordan. *Latent Dirichlet Allocation*. Vol. 3. Journal Abbreviation: The Journal of Machine Learning Research Pages: 608 Publication Title: The Journal of Machine Learning Research. Jan. 2001.

[26] Jacob Eisenstein. *Introduction to Natural Language Processing*. en. Google-Books-ID: 72yuDwAAQBAJ. MIT Press, Oct. 2019. ISBN: 978-0-262-04284-0.

[27] Raymond Lee. *Natural Language Processing : A Textbook with Python Implementation*. Nov. 2023. ISBN: 978-981-9919-99-4. DOI: 10.1007/978-981-99-1999-4.

[28] Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. "Introduction to Information Retrieval". en. In: (2009).

[29] James J. Thomas and Kristin A. Cook. "Illuminating the Path: The Research and Development Agenda for Visual Analytics". In: 2005. URL: https://www.semanticscholar.org/paper/Illuminating-the-Path%3A-The-Research-and-Development-Thomas-Cook/e0d3ea6f33a2712b9f33707fd33443c1df354ecc.

[30] Mary Ellen Okurowski. "Information extraction overview". In: *Proceedings of a workshop on held at Fredericksburg, Virginia: September 19-23, 1993*. TIPSTER '93. USA: Association for Computational Linguistics, Sept. 1993, pp. 117–121. DOI: 10.3115/1119149.1119164. URL: https://dl.acm.org/doi/10.3115/1119149.1119164.

[31] Jeffrey L. Elman. "Finding Structure in Time". en. In: *Cognitive Science* 14.2 (1990). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1, pp. 179–211. ISSN: 1551-6709. DOI: 10.1207/s15516709cog1402_1.

[32] P.J. Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10 (Oct. 1990). Conference Name: Proceedings of the IEEE, pp. 1550–1560. ISSN: 1558-2256. DOI: 10.1109/5.58337.

[33] Tomáš Mikolov et al. "Recurrent neural network based language model". en. In: *Interspeech 2010*. ISCA, Sept. 2010, pp. 1045–1048. DOI: 10.21437/Interspeech.2010-343. URL: https://www.isca-archive.org/interspeech_2010/mikolov10_interspeech.html.

[34] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[35] Patricia Kuhl and Maritza Rivera-Gaxiola. "Neural Substrates of Language Acquisition". en. In: *Annual Review of Neuroscience* 31.Volume 31, 2008 (July 2008). Publisher: Annual Reviews, pp. 511–534. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev.neuro.30.051606.094321.

[36] Henry M. Wellman et al. "Infants Use Statistical Sampling to Understand the Psychological World". en. In: *Infancy* 21.5 (2016). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/infa.12131, pp. 668–676. ISSN: 1532-7078. DOI: 10.1111/infa.12131.

[37] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. How emotions are made: The secret life of the brain. Pages: xv, 425. Boston, MA: Houghton Mifflin Harcourt, 2017. ISBN: 978-0-544-13331-0 978-0-544-12996-2.

[38] Erno Téglás et al. "Pure reasoning in 12-month-old infants as probabilistic inference". eng. In: *Science (New York, N.Y.)* 332.6033 (May 2011), pp. 1054–1059. ISSN: 1095-9203. DOI: 10.1126/science.1196404.

[39] Athena Vouloumanos and Sandra R. Waxman. "Listen up! Speech is for thinking during infancy". In: *Trends in Cognitive Sciences* 18.12 (Dec. 2014), pp. 642–646. ISSN: 1364-6613. DOI: 10.1016/j.tics.2014.10.001.

[40] Yoshua Bengio et al. "A Neural Probabilistic Language Model". en. In: *Journal of Machine Learning Research* (2003).

[41] Max Fisher. *The Chaos Machine: The Inside Story of How Social Media...* en. Little, Brown and Company, 2022. URL: https://www.goodreads.com/book/show/58950736-the-chaos-machine.

[42] Isabel O. Gallegos et al. "Bias and Fairness in Large Language Models: A Survey". In: *Computational Linguistics* 50.3 (Sept. 2024), pp. 1097–1179. ISSN: 0891-2017. DOI: 10.1162/coli_a_00524.

[43] Dipto Barman, Ziyi Guo, and Owen Conlan. "The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination". In: *Machine Learning with Applications* 16 (June 2024), p. 100545. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2024.100545.

[44] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[45] Wilson L. Taylor. ""Cloze Procedure": A New Tool for Measuring Readability". en. In: *Journalism Quarterly* 30.4 (Sept. 1953), pp. 415–433. ISSN: 0022-5533. DOI: 10.1177/107769905303000401.

[46] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[47] OpenAI et al. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2024. DOI: 10.48550/arXiv.2303.08774. URL: http://arxiv.org/abs/2303.08774.

[48] Alessandro Annoni et al. *Artificial Intelligence: A European Perspective*. en. ISBN: 9789279972171 9789279972195 9789279982132 ISSN: 1831-9424, 1018-5593, 1831-9424. 2018. DOI: 10.2760/11251. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC113826.

[49] Partha Pratim Ray. "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope". In: *Internet of Things and Cyber-Physical Systems* 3 (Jan. 2023), pp. 121–154. ISSN: 2667-3452. DOI: 10.1016/j.iotcps.2023.04.003.

[50] Ali Borji. *A Categorical Archive of ChatGPT Failures*. arXiv:2302.03494 [cs]. Apr. 2023. DOI: 10.48550/arXiv.2302.03494. URL: http://arxiv.org/abs/2302.03494.

[51] Hussam Alkaissi and Samy I McFarlane. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing". In: *Cureus* 15.2 (), e35179. ISSN: 2168-8184. DOI: 10.7759/cureus.35179.

[52] Gokul Yenduri et al. *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. arXiv:2305.10435 [cs]. May 2023. DOI: 10.48550/arXiv.2305.10435. URL: http://arxiv.org/abs/2305.10435.

[53] Luciano Floridi and Massimo Chiriatti. "GPT-3: Its Nature, Scope, Limits, and Consequences". en. In: *Minds and Machines* 30.4 (Dec. 2020), pp. 681–694. ISSN: 1572-8641. DOI: 10.1007/s11023-020-09548-1.

[54] Amir Feder et al. *Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond*. arXiv:2109.00725 [cs]. July 2022. DOI: 10.48550/arXiv.2109.00725. URL: http://arxiv.org/abs/2109.00725.

[55] Xiao Liu et al. "Eliciting and Improving the Causal Reasoning Abilities of Large Language Models with Conditional Statements". In: *Computational Linguistics* (Jan. 2025), pp. 1–38. ISSN: 0891-2017. DOI: 10.1162/coli_a_00548.

[56] Jie Yang, Soyeon Caren Han, and Josiah Poon. "A survey on extraction of causal relations from natural language text". en. In: *Knowledge and Information Systems* 64.5 (May 2022), pp. 1161–1186. ISSN: 0219-3116. DOI: 10.1007/s10115-022-01665-w.

[57] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2201.11903. URL: http://arxiv.org/abs/2201.11903.

[58] Vivek Khetan et al. *Causal BERT : Language models for causality detection between events expressed in text*. arXiv:2012.05453 [cs]. Jan. 2021. DOI: 10.48550/arXiv.2012.05453. URL: http://arxiv.org/abs/2012.05453.

[59] Brett Drury and Mathieu Roche. "A survey of the applications of text mining for agriculture". In: *Computers and Electronics in Agriculture* 163 (Aug. 2019), p. 104864. ISSN: 0168-1699. DOI: 10.1016/j.compag.2019.104864.

[60] *Innovation, Agricultural Productivity and Sustainability in China*. en. Oct. 2018. URL: https://www.oecd.org/en/publications/innovation-agricultural-productivity-and-sustainability-in-china_9789264085299-en.html.

[61] Michael Cherlet et al. *World Atlas of Desertification*. en. ISBN: 9789279753503 9789279753497. 2018. DOI: 10.2760/06292. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC111155.

[62] *FAOSTAT*. URL: https://www.fao.org/faostat/en/#home.

[63] Olivier (ESS) LavagnedOrtigue. "Agricultural production statistics 2000–2020". en. In: . *Production* (2005).

[64] Zhongkui Luo, Enli Wang, and Osbert J. Sun. "Can no-tillage stimulate carbon sequestration in agricultural soils? A meta-analysis of paired experiments". In: *Agriculture, Ecosystems & Environment* 139.1 (Oct. 2010), pp. 224–231. ISSN: 0167-8809. DOI: 10.1016/j.agee.2010.08.006.

[65] O. Fernández-Ugalde et al. "No-tillage improvement of soil physical quality in calcareous, degradation-prone, semiarid soils". In: *Soil and Tillage Research* 106.1 (Dec. 2009), pp. 29–35. ISSN: 0167-1987. DOI: 10.1016/j.still.2009.09.012.

[66] Yuhua Bai et al. "Soil Structure and Crop Performance After 10 Years of Controlled Traffic and Traditional Tillage Cropping in the Dryland Loess Plateau in China". en-US. In: *Soil Science* 174.2 (Feb. 2009), p. 113. ISSN: 0038-075X. DOI: 10.1097/SS.0b013e3181981ddc.

[67] H. J. Causarano et al. "Soil organic carbon sequestration in cotton production systems of the southeastern United States: a review". eng. In: *Journal of Environmental Quality* 35.4 (2006), pp. 1374–1383. ISSN: 0047-2425. DOI: 10.2134/jeq2005.0150.

[68] José Luis Vicente-Vicente et al. "Soil carbon sequestration rates under Mediterranean woody crops using recommended management practices: A meta-analysis". In: *Agriculture, Ecosystems & Environment* 235 (Nov. 2016), pp. 204–214. ISSN: 0167-8809. DOI: 10.1016/j.agee.2016.10.024.

[69] Karl-Heinz Feger and Daniel Hawtree. "Soil Carbon and Water Security". en. In: *Ecosystem Services and Carbon Sequestration in the Biosphere*. Ed. by Rattan Lal et al. Dordrecht: Springer Netherlands, 2013, pp. 79–99. ISBN: 978-94-007-6455-2. DOI: 10.1007/978-94-007-6455-2_5. URL: https://doi.org/10.1007/978-94-007-6455-2_5.

[70] Z. Barut and I. Celik. "Tillage Effects on Some Soil Physical Properties in a Semi- Arid Mediterranean Region of Turkey". en. In: *Chemical Engineering Transactions* 58 (June 2017), pp. 217–222. ISSN: 2283-9216. DOI: 10.3303/CET1758037.

[71] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. "Information extraction from scientific articles: a survey". en. In: *Scientometrics* 117.3 (Dec. 2018), pp. 1931–1990. ISSN: 1588-2861. DOI: 10.1007/s11192-018-2921-5.

[72] Guillaume Blanchy et al. "Potential of natural language processing for metadata extraction from environmental scientific publications". English. In: *SOIL* 9.1 (Mar. 2023). Publisher: Copernicus GmbH, pp. 155–168. ISSN: 2199-3971. DOI: 10.5194/soil-9-155-2023.

[73] Saed Rezayi et al. *Exploring New Frontiers in Agricultural NLP: Investigating the Potential of Large Language Models for Food Applications*. arXiv:2306.11892 [cs]. June 2023. DOI: 10.48550/arXiv.2306.11892. URL: http://arxiv.org/abs/2306.11892.