*algorithms*

MDPI

*Article*

# A Largely Unsupervised Domain-Independent Qualitative Data Extraction Approach for Empirical Agent-Based Model Development

**Rajiv Paudel [1] and Arika Ligmann-Zielinska [2,*]**

[1] Operation Research and Analysis, Idaho National Laboratory, 1955 Fremont Ave., Idaho Falls, ID 83415, USA; rajiv.paudel@inl.gov
[2] Department of Geography, Environment, and Spatial Sciences, Michigan State University, Geography Building, 673 Auditorium Rd, Room 121, East Lansing, MI 48824, USA
* Correspondence: arika@msu.edu

**Abstract:** Agent-based model (ABM) development needs information on system components and interactions. Qualitative narratives contain contextually rich system information beneficial for ABM conceptualization. Traditional qualitative data extraction is manual, complex, and time- and resource-consuming. Moreover, manual data extraction is often biased and may produce questionable and unreliable models. A possible alternative is to employ automated approaches borrowed from Artificial Intelligence. This study presents a largely unsupervised qualitative data extraction framework for ABM development. Using semantic and syntactic Natural Language Processing tools, our methodology extracts information on system agents, their attributes, and actions and interactions. In addition to expediting information extraction for ABM, the largely unsupervised approach also minimizes biases arising from modelers' preconceptions about target systems. We also introduce automatic and manual noise-reduction stages to make the framework usable on large semi-structured datasets. We demonstrate the approach by developing a conceptual ABM of household food security in rural Mali. The data for the model contain a large set of semi-structured qualitative field interviews. The data extraction is swift, predominantly automatic, and devoid of human manipulation. We contextualize the model manually using the extracted information. We also put the conceptual model to stakeholder evaluation for added credibility and validity.

**Keywords:** agent-based modeling; natural language processing; unsupervised data extraction; model contextualization



**Citation:** Paudel, R.; Ligmann-Zielinska, A. A Largely Unsupervised Domain-Independent Qualitative Data Extraction Approach for Empirical Agent-Based Model Development. *Algorithms* **2023**, *16*, 338. https://doi.org/10.3390/a16070338

Academic Editors: Nuno Fachada and Nuno David

Received: 30 May 2023
Revised: 27 June 2023
Accepted: 29 June 2023
Published: 14 July 2023



## 1. Introduction

Qualitative data provide thick contextual information [1–4] that can support reliable complex system model development. Qualitative data analysis explores systems components, their complex relationships, and behavior [3–5]) and provides a structured framework that can guide the formulation of quantitative models [6–10]. However, qualitative research is complex, and time- and resource-consuming [1,4]. Data analysis usually involves keyword-based data extraction and evaluation that requires multiple coders to reduce biases. Moreover, model development using qualitative data requires multiple, lengthy, and expensive stakeholder interactions [11,12], which adds to its inconvenience. Consequently, quantitative modelers often avoid using qualitative data for their model development. Modelers often skip qualitative data analysis or use unorthodox approaches for framework development, which may lead to failed capturing of target systems' complex dynamics and produce inaccurate and unreliable outputs [13].

The development in the information technology sector has substantially increased access to qualitative data over the past few decades. Harvesting extensive credible data is crucial for reliable model development. Increased access to voluminous data presents a

challenge and an opportunity for model developers [14]. However, qualitative data analysis has always been a hard nut to crack for complex modelers. Most existing qualitative data analyses are highly supervised (i.e., performed mainly by humans) and hence, bias-prone and inefficient for large datasets.

This study proposes a methodology that uses an efficient, largely unsupervised qualitative data extraction for credible Agent-Based Model (ABM) development using Natural Language Processing (NLP) toolkits. ABM requires information on agents (emulating the target system's decision makers), their attributes, actions, and interactions for its development. The development of a model greatly depends on its intended purpose. Abstract theoretical models concentrate on establishing new relationships and theories with less emphasis on data requirements and structure. In contrast, application-driven models aim to explain specific target systems and tend to be data-intensive. They require a higher degree of adherence to data requirements, validity, feasibility, and transferability [15–17]. Our methodology is particularly applicable to application-driven models rich in empirical data.

ABMs help understand phenomena that emerge from nonlinear interactions of autonomous and heterogeneous constituents of complex systems [18–20]. ABM is a bottom-up approach; interactions at the micro-level produce complex and emergent phenomena at a macro (higher) level. As micro-scale data become more accessible to the research community, modelers increasingly use empirical data for more realistic system representation and simulation [11,21–24].

Quantitative data are primarily useful as inputs for parameterizing and running simulations. Additionally, quantitative model outputs are also used for model verification and validation. Qualitative data, on the other hand, find uses at various stages of the model cycle [25]. Apart from the routine tasks of identifying systems constituents and behaviors for model development, qualitative data support the model structure and output representations [26,27]. Qualitative model representations facilitate communication for learning, model evaluation, and replication.

Various approaches have been proposed to conceptualize computational models. First of all, selected quantitative models have predefined structures for model representation. System dynamics, for instance, uses Causal Loop Diagrams as qualitative tools [28]. Causal Loop Diagrams elucidate systems components, their interrelationships, and feedback that can be used for learning and developing quantitative system dynamics models. ABM, however, does not have a predefined structure for model representation; models are primarily based on either highly theoretical or best-guess ad-hoc structures, which are problematic for model structural validation [16,29].

As a consequence, social and cognitive theories [30–34] often form the basis for translating qualitative data to empirical ABM [35]. Since social behavior is complex and challenging to comprehend, using social and cognition theories helps determine the system's expected behavior. Moreover, using theories streamlines data management and analysis for model development.

Another school of thought bases model development on stakeholder cognition. Rather than relying mainly on social theories, this approach focuses on extracting empirical information about system components and behaviors. Participatory or companion modeling [36], as well as role-playing games [11], are some of the conventional approaches to eliciting stakeholder knowledge for model development [23,24]. Stakeholders usually develop model structures in real time, while some modelers prefer to process stakeholders' information after the discussions. For instance, [37] employs computer technologies to post-process stakeholder responses to develop a rule-induction algorithm for her ABM.

Stakeholders are assumed to be the experts of their systems, and using their knowledge in model building makes the model valid and reliable. However, stakeholder involvement is not always feasible; for instance, when modeling remote places or historical events. In such cases, modelers resort to information elicitation tools for information extraction. In the context of ABM, translating empirical textual data into agent architecture is complex and requires concrete algorithms and structures [25,38]. Therefore, modelers first explore the

context of the narratives and then identify potential context-specific scopes. Determining narrative elements becomes straightforward once context and scopes are identified [38].

Many ABM modelers have formulated structures for organizing qualitative data for model development. For instance, [39] used Institutional Analysis and Development framework for managing qualitative data in their Modeling Agent system based on Institutions Analysis (MAIA). MAIA comprises five structures: collective, constitutional, physical, operational, and evaluative. Information on agents is populated in a collective structure, while behavior rules and environment go in constitutional and physical structures. Similar frameworks were introduced by [40,41], to name but a few. However, all these structures use manual, slow, and bias-prone data processing and extraction. A potential solution presented in this paper is to employ AI tools, such as NLP, for unsupervised information extraction for model development [30].

The remainder of the paper is structured as follows: After the introduction, we briefly characterize NPL, focusing on its utility for ABM conceptualization. Next, we describe the proposed methodology. Finally, we demonstrate the framework in a case study of processing narratives from in-depth household interviews on individual food security in Mali, noting the framework's advantages and limitations.

## 2. Background, Materials, and Methods

Software engineers have been exploring various supervised and unsupervised approaches for information extraction. In supervised approaches, syntactical patterns are defined [42], and text is manually scanned for such patterns. In unsupervised information extraction, the machine does the pattern matching.

Supervised approaches are reliable but slow. Contrarily, the faster, unsupervised approaches are difficult and prone to errors, mainly due to word sense ambiguation [43]. A purely syntactical analysis cannot capture the nuanced meaning of texts, which is often the culprit of the problem. Recently, pattern matching also involves semantic analysis. External databases of hierarchically structured words such as WordNet or VerbNet [44] and machine learning tools are increasingly used for understanding semantics for reduced word sense ambiguity [45,46].

NLP toolkits are increasingly used for unsupervised pattern matching and information extraction [47–52]. Tools such as lemmatizing, tokenizing, stemming, and part-of-speech tagging [53] are helpful for syntactic information extraction. These tools can normalize texts and identify subjects and main verbs from their sentences.

The ability to convert highly unstructured texts to structured information through predominantly unsupervised approaches is one of the main advantages of NLP in qualitative data analysis. NLP efficiently analyzes intertextual relationships using syntactic and semantics algorithms. Approaches such as word co-occurrence statistics and sentiment analysis [54] are beneficial for domain modeling [42] and for exploring contextual and behavioral information from textual data. Similarly, its efficiency in pattern matching for information extraction is essential for model development.

Although present for decades in object-oriented programming and database development [55,56], NLP has a minimal footprint in ABM development. The introduction of NLP in ABM development is very recent. Refs. [14,57] used NLP to model human cognition through word embedding, which is a contextually analyzed vector representation of a text. The procedure places closely related texts next to each other. Specifically, placing agents with similar worldviews together to support theorizing agent decision-making. Although their approach helps develop agent decision-making, it is not well-equipped for developing a comprehensive agent architecture.

Another example is the study by [58], who applied NLP in conjunction with machine learning to create an ABM structure from unstructured textual data. In their framework, texts are translated to the agent–attribute–rule framework. They define agents as nouns (e.g., person and place) that perform some actions and attributes as words that represent some variables. Similarly, sentences containing agents or attributes and action verbs are

considered rules. The primary goal of their approach is to create an ABM structure mainly for communicating the model to non-modelers.

As with machine learning approaches in general, Padilla et al.'s approach required a large amount of training data. They used ten highly concise, formally written ABM descriptions from published journals as training datasets. Another limitation, according to the researchers, was the lack of precise distinctions between agents and attributes, attributes could also be nouns, which might confuse the machine. Repeated training with extensive data effectively increased the accuracy of the agent–attribute rule detection. However, the model frequently resulted in underpredictions and overpredictions.

Our work is of significant importance in the context of ABM development, particularly in relation to the utilization of machine learning and artificial intelligence. The recent research papers, refs. [30,59,60], also emphasized the role of these technologies in this domain. Ref. [60] go as far as to argue that natural language processing (NLP) can potentially replace the conventional method of developing ABMs, which heavily relies on field interviews. This perspective highlights the relevance and timeliness of our work, as we effectively incorporated machine learning and artificial intelligence techniques into our ABM development process.

Additionally, we discussed the relevance of prior works such as [14,57,58], and that exhibit similarities to our approach. However, these studies lack certain aspects of model development that our proposed methodology aims to address. For instance, ref. [58] relied on extensive training datasets and struggled to differentiate between agents and attributes effectively, whereas our methodology overcomes these limitations. Furthermore, unlike Runck's approach, which primarily focuses on developing agents' decision-making abilities, our approach strives to create a comprehensive agent architecture.

## 3. The Proposed Framework

In response to these limitations, our study proposes and tests a largely unsupervised domain-independent approach for developing ABM structures from informal semi-structured interviews using Python-based semantic and syntactic NLP tools (Figure 1). The method primarily uses syntactic NLP approaches for information extraction directly to the object-oriented programming (OOP) framework (i.e., agents, attributes, and actions/interactions) using widely accepted approaches in database design and OOP [61]. Database designers and OOP programmers generally exploit the syntactic structure of sentences for information extraction. Syntactic analysis usually treats the subject of a sentence as a class (an entity for a database) and the main verb as a method (a relationship for a database). Since the approach is not based on machine learning, it does not require large training data. The semantic analysis is limited to external static datasets such as WordNet (https://wordnet.princeton.edu/) (accessed on 8 July 2020) and VerbNet (https://verbs.colorado.edu/verbnet/) (accessed on 21 July 2020).

In the proposed approach, information extraction includes systems agents, their actions, and interactions from qualitative data for model development using syntactic and semantic NLP tools. As our information extraction approach is primarily unsupervised and does not require manual interventions, we argue that, in addition to being efficient, it reduces the potential for subjectivity and biases arising from modelers' preconceptions about target systems.

The extracted information is then represented using Unified Modeling Language (UML) for an object-oriented model development platform. UML is a standardized graphical representation of software development [62]. It has a set of well-defined class and activity diagrams that effectively represent the inner workings of ABMs [63]. UML diagrams represent systems classes, their attributes, and actions. Identified candidate agents, attributes, and actions were manually arranged in the UML structure for supporting model software development. Although there are other forms of graphical ABM representations such as Petri Nets [64], Conceptual Model for Simulation [29], and sequence and activity

diagrams [65], UML is natural in representing ABM, named by [66] the default lingua franca of ABM.

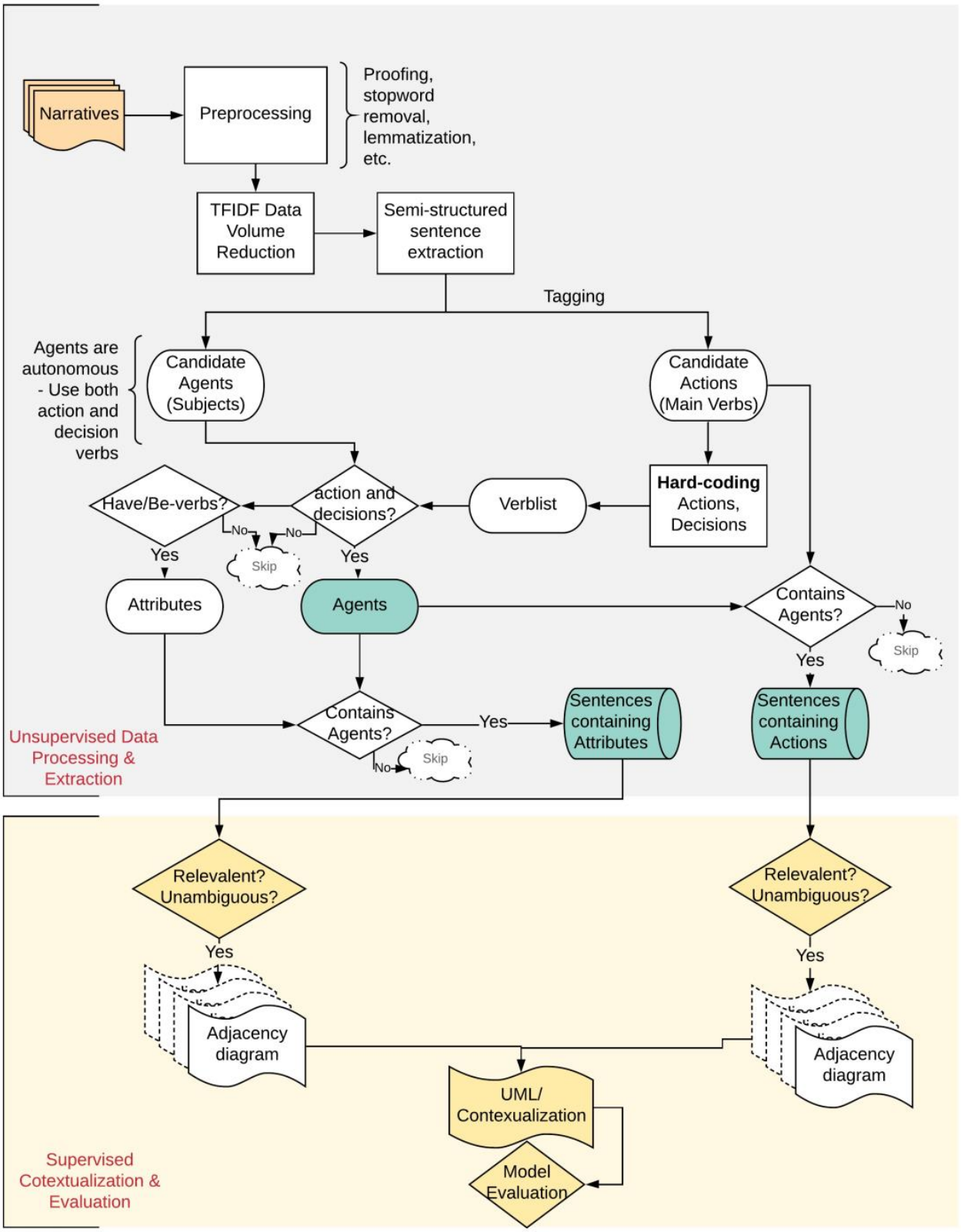


**Figure 1.** Largely unsupervised information extraction for ABM development. TFIDF: term frequency inverse document frequency.

In our approach, model development is mainly unsupervised and involves the following steps (Figure 1):

1. Unsupervised data processing and extraction;
2. Data preprocessing (cleaning and normalization);
3. Data volume reduction;
4. Tagging and information extraction;
5. Supervised contextualization and evaluation;
6. UML/Model conceptualization;
7. Model evaluation.

Steps one and two are required since semi-structured interviews often contain redundant or inflected texts that can bog down NLP analysis. Hence, removing non-informative contents from large textual data is highly recommended at the start of the analysis. NLP is well-equipped with stop words removal tools that can effectively remove redundant texts. Similarly, tools such as stemming and lemmatizing help normalize texts to their base forms [67].

Step three is data volume reduction, which can tremendously speed up NLP analyses. Traditional volume reduction approaches usually contain highly supervised keyword-based methods. Data analysts use predefined keywords to select and extract sentences perceived to be relevant [68]. Keyword identification generally requires a priori knowledge of the system and is often bias-prone. Consequently, we recommend a domain-independent unsupervised Term Frequency Inverse Document Frequency (TFIDF) approach [69] that eliminates manual keyword identification requirements. The approach provides weightage to individual words based on their uniqueness and machine-perceived importance. The TFIDF differentiates between important and common words by comparing their frequency in individual documents and across entire texts. Sentences that have high cumulative TFIDF scores are perceived to have higher importance. Given a document collection $D$, a word $w$, and an individual document $d \, \varepsilon \, D$, TFIIDF can be defined as follows:

$$fw, d * log(|D|/fw, D) \quad (1)$$

where *fw,d* equals the number of times $w$ appears in $d$, $|D|$ is the size of the corpus, and *fw,D* equals the number of documents in which $w$ appears in $D$ [69].

Step four involves tagging and information extraction. Once the preprocessed data are reduced, we move to tagging agents, attributes, and actions/interactions that can occur. We propose the following approaches for tagging agent architecture:

*Candidate agents:* Following the conventional approaches in database design and OOP [61], we propose identifying the subjects of sentences as candidate agents. For instance, the *farmer* in 'the farmer grows cotton' can be a candidate agent. NLP has well-developed tools such as part-of-speech tagger and named-entity tagger that can be used to detect subjects of sentences.

*Candidate actions:* The main verbs of sentences can become candidate actions. The main verbs need candidate agents as the subject of the sentences. For example, in the sentence 'the farmer grows cotton,' the *farmer* is a candidate agent, and the subject of the sentence; *grows* is the main verb and, hence, a candidate action.

*Candidate attributes*: Attributes are properties inherent to the agents. Sentences containing candidate agents as subjects and *be* or *have* as their primary (non-auxiliary) verbs provide attribute information, e.g., 'the farmer *is* a member of a *cooperative*,' and 'the farmer *has 10 ha* of land.' Additionally, the use of possessive words also indicates attributes, e.g., *the cow* in the sentence 'my cow is very small' is an attribute.

*Candidate interactions*: Main verbs indicating relationships between two candidate agents are identified as interactions. Hence the sentences containing **two or more candidate agents** provide information on *interactions*, e.g., '**The government** *trains* **the farmers**.'

Since the data tagging is strictly unsupervised, false positives are likely to occur. The algorithm can over-predict agents, as the subjects of all the sentences are treated as

candidate agents. In ABM, however, agents are defined as autonomous actors, they act and make decisions. Hence, we propose to use a hard-coded list of action verbs (e.g., eat, grow, and walk) and decision verbs (e.g., choose, decide, and think) to filter agents from the list of candidate agents. Only the candidate agents that use both types of verbs qualify as agents. Candidate agents not using both verbs are categorized as *entities* that may be subjected to manual evaluation. Similarly, people use different terminologies that are semantically similar. We recommend using external databases such as WordNet to group semantically similar terminologies.

Step five involves supervised contextualization and evaluation. While the unsupervised analysis reduces data volume and translates semi-structured interviews to the agent–action–attribute structure, noise can percolate to the outputs since the process is unsupervised. Additionally, the outputs need to be contextualized. Consequently, we suggest performing a series of supervised output filtration followed by manual contextualization and validation. The domain-independent unsupervised analysis extracts individual sentences that can sometimes be ambiguous or domain irrelevant. Hence the output should be filtered based on ambiguity and domain relevancy. Once output filtration is performed, contextual structures can be developed and validated with domain experts and stakeholders.

The last two steps (UML/model conceptualization and model evaluation) are described in the following sections.

For this study, we used Python 3.7 programming language (https://www.python.org/) (accessed on 10 May 2020) along with a plethora of NLP libraries (e.g., scikit-learn, NLTK, spaCy, and textacy) to perform data reduction, tagging, extraction, and structuration. Scikit-learn provides a wide range of machine learning algorithms for classification, regression, clustering, and dimensionality reduction tasks. Similarly, NLTK, spaCy, and textacy are useful for analyzing natural language data. We primarily used scikit-learn for dimensionality reduction and NLTK, spaCy, and textacy for tokenization and part of speech tagging.

## 4. Results and Discussion

We tested the above approach by developing a structural ABM of household food security using semi-structured field interviews, for example, the excerpt in Figure 2. Our qualitative data contain 42 semi-structured interviews from different members (young and old, male and female) of farming households in Koutiala, Southern Mali. The interviews were initially conducted to develop mental models of household food security in the region [70]. Verbal consent was obtained from the participants prior to the interviews. The interviews were originally conducted in the local Bambara dialect and then translated into English for model development. The mental model development followed the lengthy conventional qualitative data analysis approach that used multiple coders and keyword-based sentence extraction. That inspired the research team to develop a more efficient alternative data processing and extraction approach for ABM development, presented here.

First, we grouped the interviews by the member types (i.e., elder male, younger male, elder female, and younger female) and analyzed the grouped narratives collectively. After preprocessing the interviews using NLTK tools, we used the scikit-learn Tfidf Vectorizer to reduce the volume of qualitative data. Textacy was primarily used for identifying candidate agents, actions, and attributes. Additionally, textacy extract (textacy.extract. semi-structured statements) was used in converting sentences to structured outputs. Finally, we manually filtered the unsupervised outputs based on their domain relevancy and ambiguity. The final outputs were then visualized and conceptualized using Gephi (https://gephi.org/) (accessed on 18 May 2020) and Lucid Chart (https://www.lucidchart.com/) (accessed on 21 May 2020) platforms (Figure 3).

**Enumerator:** When the production reaches home, how do you consume it?

**Surveyed:** First of all, we are 48 people in my household. After harvesting, we beat and weigh the crops. Then, we pick the quantity of food that is needed to feed the household. That quantity is given to women for cooking. We continue with that quantity until the new crop is harvested. When food is cooked and ready for eating, it is distributed by a plate to five and six persons who eat together. As the daily food consumption is known, we know what quantity we can sell to address household needs. You know, when crops are harvested, we tie them. Then we pick some pieces for beating. After beating, we weigh and stock. Then, we know the existing quantity in the store. If it is maize, for example, we have a cart and beating machine. We take one load of maize in the cart. Then we beat and weigh it and finally stock in the store. Farmers start beating crops and stock them. Because, a researcher like you, **IER** of **Sikasso** and **CMDT** trained us on post-harvest management. By following lessons learned from training, our food won't be over. Even if it happens that our food finishes, we would know what food to buy without any difficulties. With lessons acquired, we can produce and increase our production that will feed all our household over the year. Unless our food is stolen, it will cover for the year.

**Enumerator:** What are the main crops you are producing?

**Surveyed:** We produce maize, sorghum, millet, and cowpea in one hand and use the same crops to feed ourselves on the other hand. Some farmers prefer feeding themselves with millet all over the year. Even if they produce maize, they sell it. Although other farmers prefer maize, we consume all the crops we produce. We may consume one crop in one month and shift to another for another month. This is how we shift food consumption.

**Enumerator:** Who decides to shift from one food to another?

**Surveyed:** It depends. Some crops are easier to process than others. So, when there is the farm, we decide to consume crop, which is easy and quick to process. But, only one person is responsible for picking the food for daily consumption.

**Figure 2.** Excerpt from sample interview.

As expected, the unsupervised tagging overpredicted the agents. Subjects that do not make decisions were also identified as candidate agents. To overcome the issue, we created an external database of action/decision verbs (Table 1) that somewhat addressed the problem. Using the external database resulted in more than 60% reduction in the number of agents (e.g., Figure 4). The filtration process discarded the initially identified candidates, such as porridge, food, cereal, or farm. We also obtained multiple similar actions (synonyms). We used an external WordNet database to group semantically similar actions. The process resulted in a highly manageable and structured output for model conceptualization.

Next, we used the extracted information to develop UML class diagrams (Figure 5) and contextual diagrams (Figure 6). The diagrams revealed that different members of households support household food security differently. Male members of the households are generally involved in farming. They grow cereal crops and vegetables and are also into cash cropping, i.e., growing plants for selling on the market rather than subsistence farming to feed their families. Women principally look after household work and assist men in the fields. Households consume the food they produce. During food shortages, households seek help from their fellow villagers or buy food from the market. They use money obtained from cash crops to buy food.

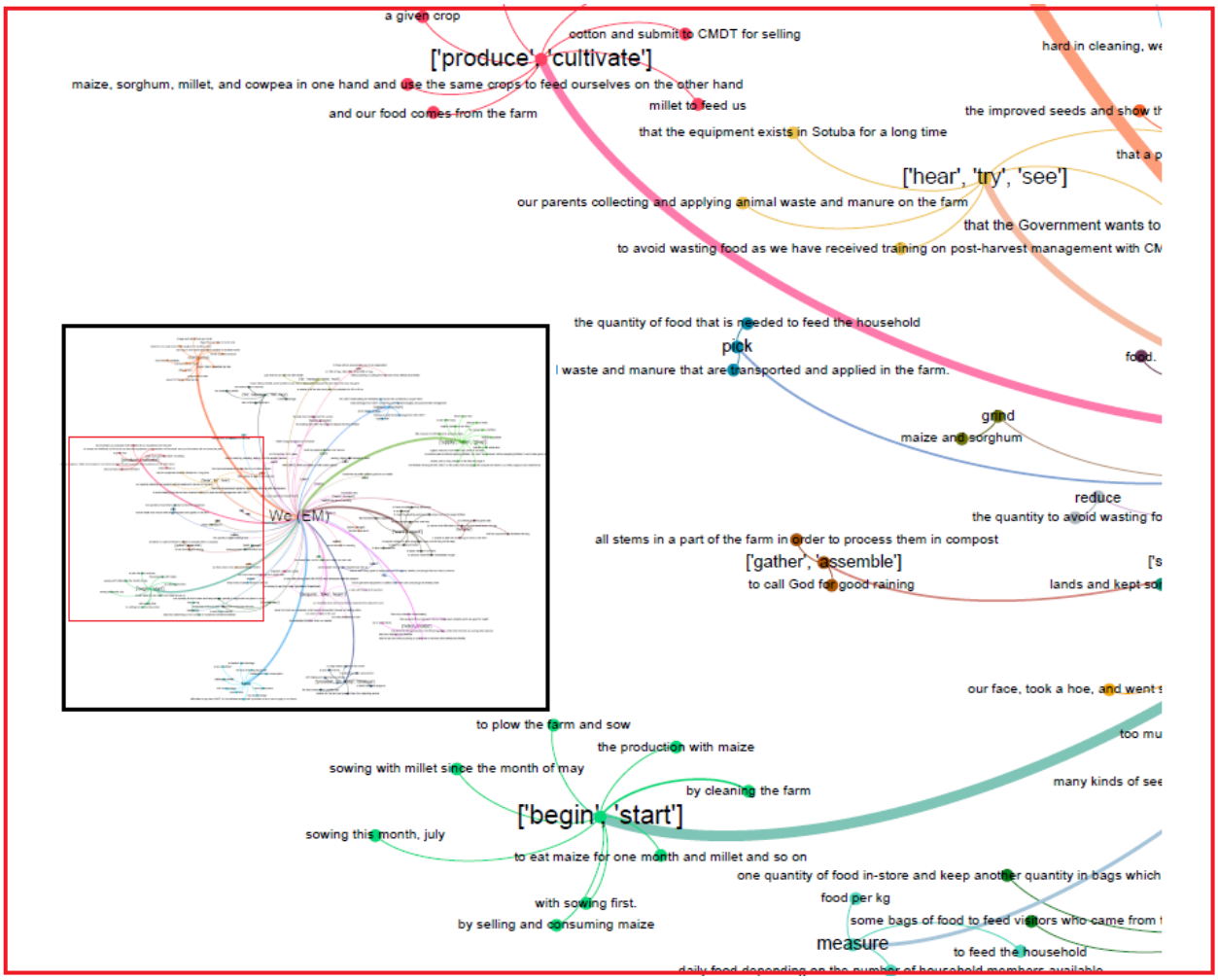


**Figure 3.** Visual representation of data processing and extraction (fragment).

**Table 1.** An excerpt of action and decision verbs.

| Decision Verbs | Action Verbs |
|---|---|
| adhere | abandon |
| advise | accelerate |
| approve | accept |
| assess | access |
| choose | accompany |
| comply | accord |
| consult | achieve |
| decide | acquaint |
| determine | acquire |
| discourage | add |
| educate | adjust |
| encourage | adopt |
| expect | advertise |
| favor | affect |
| guide | afford |
| instruct | aim |
| learn | allow |
| obey | analyze |
| oblige | apply |
| plan | argue |

```
candidate agents: {'money', 'woman', 'water', 'porridge', 'household', 'father', 'i', 'food, 'it', 'we', 'they', 'she', 'he', 'cereal', 'farm'}

agents: ['i', 'it', 'we', 'they', 'she']
```

**Figure 4.** Candidate agents before and after using the external database.

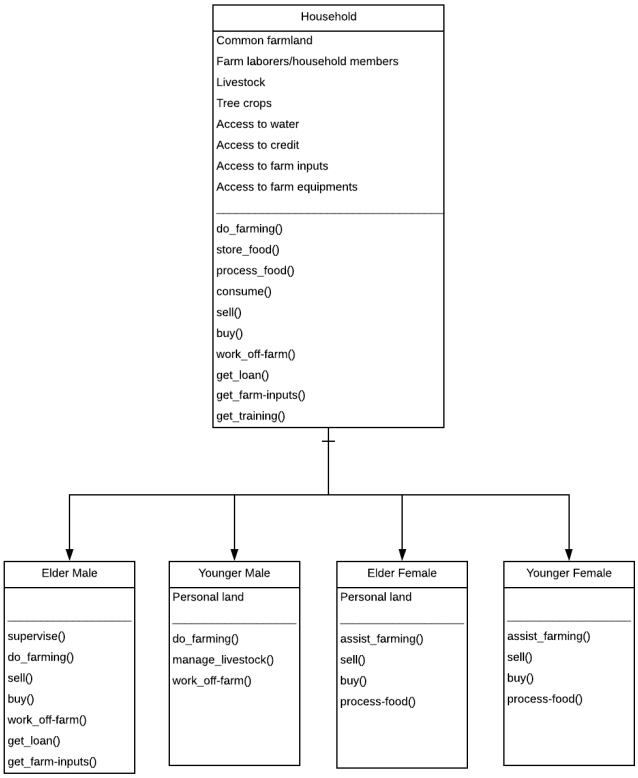


**Figure 5.** UML class diagram of agents of household food security.

Additionally, household women are involved in small businesses that can support food purchases. For example, some households might need to rely on off-farm jobs or sell their livestock to buy food. Other organizations and credit agencies provide households with credits and support.

Following our framework, the conceptual model required evaluation. We applied model-to-model (M2M) comparison [71,72] and stakeholder validation. M2M involved comparing model output with the mental model of household food security developed using the same dataset, reported by [70]. We found that our approach captured all the essential components of household food security that were identified in the mental model.

Initially, we aimed to develop an efficient, bias-free, completely unsupervised information extraction for conceptualizing an ABM. However, after preliminary algorithm development, we realized that entirely unsupervised data processing and conceptualization is unrealistic with the current NLP capabilities. Therefore, we decided to use manual filtration and contextualization that potentially introduced subjectivity and biases in model development. We performed a stakeholder validation to address this deficiency to check for subjectivity and biases. Consequently, we converted the contextual model to a pictorial representation (Figure 7) and brought it to the stakeholders (interviewees) for validation.

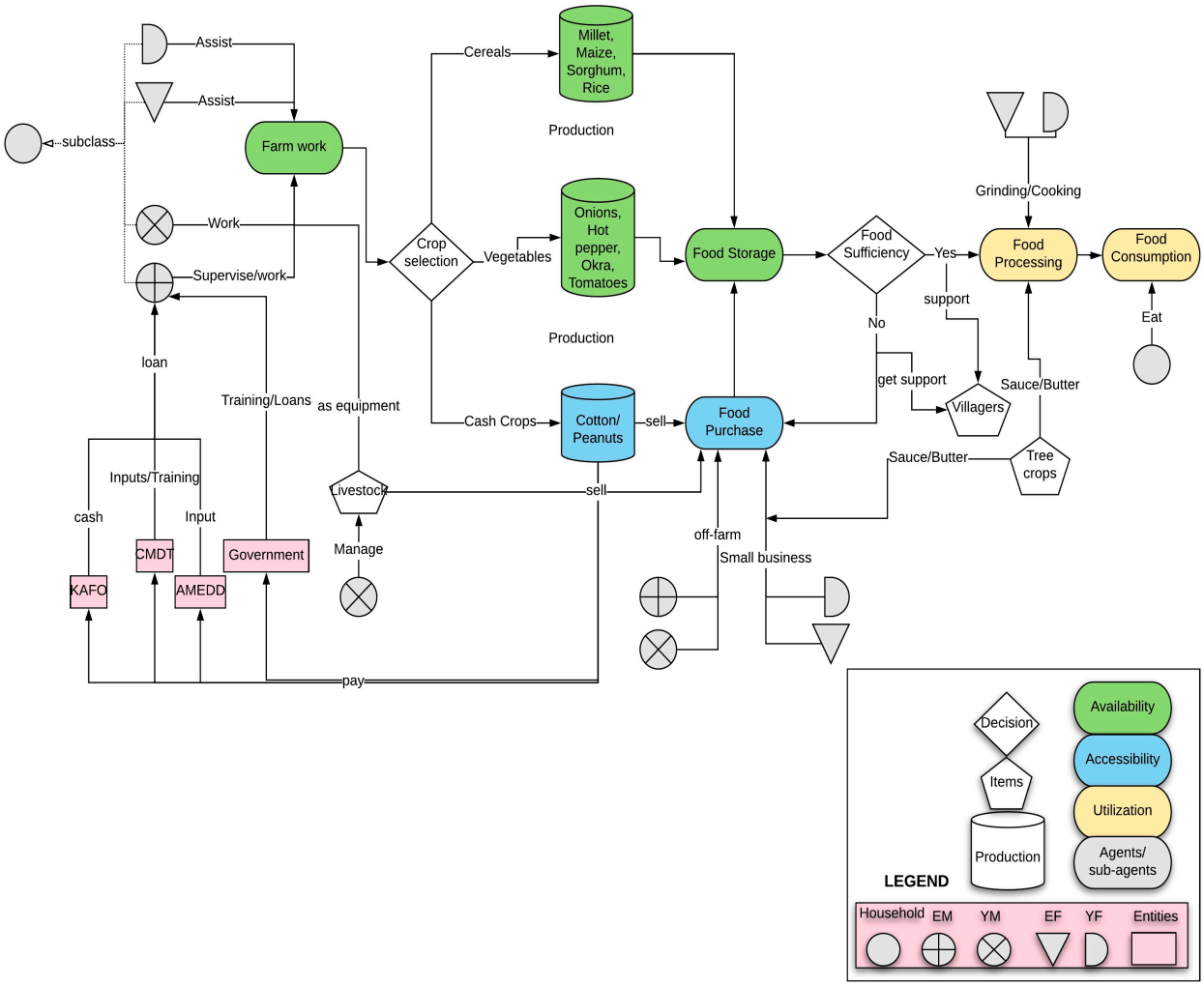


**Figure 6.** Conceptual model of household food security in Koutiala, Mali (EM: Elderly male; YM:. Younger male; EF: Elderly female; YF: Younger female).

The stakeholders positively evaluated the model and acknowledged that it included all the principal dynamics of the household food system. They, however, pointed out that the contextualized structure did not provide the dynamics of the government and non-government actors. Since the input data only contained interviews from farm households, we failed to capture the dynamics occurring outside of the households. Consequently, the model revealed data gaps where more information needs to be gathered on household food security's government and non-government actors.

The proposed unsupervised information extraction picked individual sentences based on their cumulative TFIDF weights. However, some of the individually extracted sentences lacked contextuality and were ambiguous. To add context and reduce this ambiguity, we used neighboring sentences during the unsupervised data extraction and processing phase (Figure 1). We hypothesized that extracting a tuple of preceding and trailing sentences along with the identified sentence can provide vital contextual information; for example, some of the extracted sentences contained pronouns. These pronouns were impossible to resolve without the information in the preceding sentences. Therefore, extracting the preceding sentence should help in resolving their references.

The NLP also has a coreference resolution tool that automatically replaces pronouns with their referenced nouns. However, the tool is in development. We found that it generated too many errors that would require manual checks. Hence, we proceeded without using the tool, and the pronouns identified as agents were ignored.

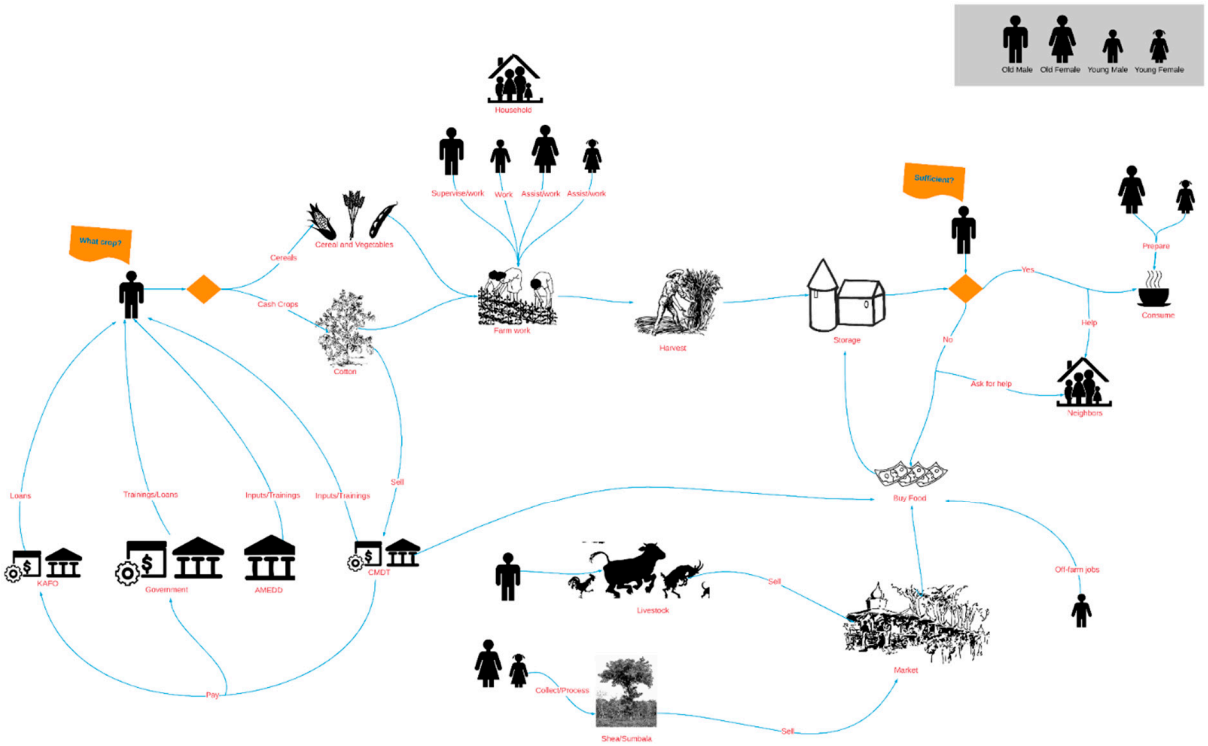


**Figure 7.** Pictorial representation of the conceptual model presented in Figure 6.

Using our framework, we only collected information on agents, attributes, and actions/interactions. However, ABM also requires information on agent decision-making. Although using social and behavioral theories in defining agent decision-making is predominant, empirically derived decision-making frameworks are context-specific and, therefore, more desirable when ABMs are applied in real-world situations [15,24]. We realize that some sentences are particularly useful in deriving agent decision-making. Specifically, conditional sentences such as 'if it rains, we plant maize' and compound sentences such as 'when production is low, we buy food from the market' can reveal decision-making. Harvesting these sentences with semantics and machine learning approaches can open new avenues for formulating empirically based decision-making rules for ABM.

It is important to note that the derived information is limited by the information contained in the input. For example, we noticed that agent tagging underpredicted agents after using the action and decision verbs. Entities such as 'father' and 'the government' should also be identified as agents of this particular system. However, some information was missed since subjects did not use both types of verbs (action and decision) in the provided interviews. Additionally, stakeholders pointed out that our model structure did not include the dynamics of the governmental and non-governmental actors. It prompts a need for a careful analysis of entities that failed to qualify as agents for data gaps. Furthermore, the interviews went through different translation stages (from local dialects to French and English) that could have corrupted some of their original meanings.

## 5. Conclusions

Complexities, ambiguities, and difficulties in data processing often discourage ABM developers from using qualitative data for model development, preventing modelers from using rich contextual information about their target systems. ABMs are often developed using ad-hoc approaches, potentially producing models that lack credibility and reliability. We introduced a systematic approach for ABM development from semi-structured qualitative interviews using NLP to address these gaps. The proposed methodology contained a largely unsupervised, domain-independent, efficient, and bias-controlled data processing and extraction approach aimed at ABM conceptualization. We demonstrated its effective-

ness by developing an ABM of household food security from large open-ended qualitative field interviews.

Additionally, we outlined some of the significant limitations of the approach and recommended improvements for future development. Our framework is only relevant to data-driven models that focus on applications and address specific geographic regions and localities. It is not aimed at theory-driven modeling, which requires generalizable observations, where other methods, such as metamodeling, are more appropriate. It is also important to note that the proposed framework was developed only to handle information derived from text. Future improvements should focus on algorithms and tools combining text-derived and quantitative information using data analytics tools. Moreover, our framework requires further testing and experimentation, for example, contrasting it with alternative approaches, which is one of the objectives of our future research. Hopefully, since the NLP development community is highly active, these limitations will soon be resolved, making semantic and syntactic NLP more effective for unsupervised information extraction and model conceptualization.

Although we could not fully develop a completely unsupervised approach, we successfully managed to reduce subjectivity and biases by limiting data extraction manipulation. Data processing and extraction were fully unsupervised, and manual inputs were only required towards the end of model conceptualization, limiting the opportunities for introducing human bias in model development. Furthermore, the unsupervised approach was much faster compared with manual coding.

**Author Contributions:** Conceptualization: R.P. and A.L.-Z.; Methodology: R.P.; Code: R.P.; Formal Analysis: R.P.; Investigation: A.L.-Z.; Writing—Original Draft Preparation: R.P.; Writing—Review and Editing: A.L.-Z.; Manuscript Visualization: R.P.; Graphical Abstract: A.L.-Z.; Supervision, Project Administration and Funding Acquisition: A.L.-Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project was supported by National Science Foundation Grant SMA 1416730, titled IBSS: Participatory-Ensemble Modeling to Study the Multiscale Social and Behavioral Dynamics of Food Security, Ligmann-Zielinska A (PI) et al. The sponsors or funders played no role in the study design, data collection, analysis, decision to publish, or manuscript preparation.

**Data Availability Statement:** The results presented in this manuscript were extracted from unprocessed (i.e., original) data collected through open-ended interviews from individual households. The interviews contain confidential information such as geographic location, employment, ethnicity, household structure, relationships, and interactions among household members, or number of dependents. The Institutional Review Board (Michigan State University) approval restricted the data access only to the research team. All subjects who provided the interviews provided their informed consent for inclusion before participating in the study. Consequently, the data are protected under the rights of privacy.

**Acknowledgments:** We want to thank Laura Schmitt-Olabisi from Michigan State University, USA, and Amadou Sidibe' from 4IPR/IFRA de Katibougouand, Koulikoro, Mali, for assistance in data collection and result validation.

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

## References

1. Miles, M.B. Qualitative data as an attractive nuisance: The problem of analysis. *Adm. Sci. Q.* **1979**, *24*, 590–601. [CrossRef]
2. Mortelmans, D. Analyzing qualitative data using NVivo. In *The Palgrave Handbook of Methods for Media Policy Research*; Palgrave Macmillan: London, UK, 2019; pp. 435–450.
3. Rich, M.; Ginsburg, K.R. The reason and rhyme of qualitative research: Why, when, and how to use qualitative methods in the study of adolescent health. *J. Adolesc. Health* **1999**, *25*, 371–378. [CrossRef] [PubMed]
4. Watkins, D.C. Qualitative research: The importance of conducting research that doesn't "count". *Health Promot. Pract.* **2012**, *13*, 153–158. [CrossRef]
5. Kemp-Benedict, E. From Narrative to Number: A Role for Quantitative Models in Scenario analysis. In Proceedings of the International Congress on Environmental Modelling and Software, Osnabrück, Germany, 1 July 2004.

6. Ackermann, F.; Eden, C.; Williams, T. Modeling for litigation: Mixing qualitative and quantitative approaches. *Interfaces* **1997**, *27*, 48–65. [CrossRef]
7. Coyle, G. Qualitative and quantitative modelling in system dynamics: Some research questions. *Syst. Dyn. Rev. J. Syst. Dyn. Soc.* **2000**, *16*, 225–244. [CrossRef]
8. Forbus, K.D.; Falkenhainer, B. Self-Explanatory Simulations: An Integration of Qualitative and Quantitative Knowledge. In Proceedings of the AAAI, Boston, MA, USA, 29 July–3 August 1990; pp. 380–387.
9. Jo, H.I.; Jeon, J.Y. Compatibility of quantitative and qualitative data-collection protocols for urban soundscape evaluation. *Sustain. Cities Soc.* **2021**, *74*, 103259. [CrossRef]
10. Wolstenholme, E.F. Qualitative vs quantitative modelling: The evolving balance. *J. Oper. Res. Soc.* **1999**, *50*, 422–428. [CrossRef]
11. Djenontin, I.N.S.; Zulu, L.C.; Ligmann-Zielinska, A. Improving representation of decision rules in LUCC-ABM: An example with an elicitation of farmers' decision making for landscape restoration in central Malawi. *Sustainability* **2020**, *12*, 5380. [CrossRef]
12. Polhill, J.G.; Sutherland, L.-A.; Gotts, N.M. Using qualitative evidence to enhance an agent-based modelling system for studying land use change. *J. Artif. Soc. Soc. Simul.* **2010**, *13*, 10. [CrossRef]
13. Landrum, B.; Garza, G. Mending fences: Defining the domains and approaches of quantitative and qualitative research. *Qual. Psychol.* **2015**, *2*, 199. [CrossRef]
14. Runck, B. GeoComputational Approaches to Evaluate the Impacts of Communication on Decision-Making in Agriculture. Ph.D. Thesis, University of Minnesota, Minneapolis, MN, USA, 2018.
15. Du, J.; Ligmann-Zielinska, A. The Volatility of Data Space: Topology Oriented Sensitivity Analysis. *PLoS ONE* **2015**, *10*, e0137591. [CrossRef] [PubMed]
16. Grimm, V.; Augusiak, J.; Focks, A.; Frank, B.M.; Gabsi, F.; Johnston, A.S.; Liu, C.; Martin, B.T.; Meli, M.; Radchuk, V. Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecol. Model.* **2014**, *280*, 129–139. [CrossRef]
17. Ligmann-Zielinska, A.; Siebers, P.-O.; Magliocca, N.; Parker, D.C.; Grimm, V.; Du, J.; Cenek, M.; Radchuk, V.; Arbab, N.N.; Li, S. 'One size does not fit all': A roadmap of purpose-driven mixed-method pathways for sensitivity analysis of agent-based models. *J. Artif. Soc. Soc. Simul.* **2020**, *23*. [CrossRef]
18. An, L.; Linderman, M.; Qi, J.; Shortridge, A.; Liu, J. Exploring Complexity in a Human–Environment System: An Agent-Based Spatial Model for Multidisciplinary and Multiscale Integration. *Ann. Assoc. Am. Geogr.* **2005**, *95*, 54–79. [CrossRef]
19. Railsback, S.F.; Grimm, V. *Agent-Based and Individual-Based Modeling: A Practical Introduction*; Princeton University Press: Princeton, NJ, USA, 2019.
20. Wilensky, U.; Rand, W. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*; Mit Press: Cambridge, MA, USA, 2015.
21. Janssen, M.; Ostrom, E. Empirically based, agent-based models. *Ecol. Soc.* **2006**, *11*. [CrossRef]
22. O'Sullivan, D.; Evans, T.; Manson, S.; Metcalf, S.; Ligmann-Zielinska, A.; Bone, C. Strategic directions for agent-based modeling: Avoiding the YAAWN syndrome. *J. Land. Use Sci.* **2016**, *11*, 177–187. [CrossRef] [PubMed]
23. Robinson, D.T.; Brown, D.G.; Parker, D.C.; Schreinemachers, P.; Janssen, M.A.; Huigen, M.; Wittmer, H.; Gotts, N.; Promburom, P.; Irwin, E.; et al. Comparison of empirical methods for building agent-based models in land use science. *J. Land. Use Sci.* **2007**, *2*, 31–55. [CrossRef]
24. Smajgl, A.; Barreteau, O. *Empirical Agent-Based Modelling-Challenges and Solutions*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 1.
25. Seidl, R. Social scientists, qualitative data, and agent-based modeling. In Proceedings of the Social Simulation Conference, Barcelona, Spain, 1–5 September 2014.
26. Grimm, V.; Berger, U.; DeAngelis, D.L.; Polhill, J.G.; Giske, J.; Railsback, S.F. The ODD protocol: A review and first update. *Ecol. Model.* **2010**, *221*, 2760–2768. [CrossRef]
27. Müller, B.; Balbi, S.; Buchmann, C.M.; De Sousa, L.; Dressler, G.; Groeneveld, J.; Klassert, C.J.; Le, Q.B.; Millington, J.D.A.; Nolzen, H. Standardised and transparent model descriptions for agent-based models: Current status and prospects. *Environ. Model. Softw.* **2014**, *55*, 156–163. [CrossRef]
28. Ford, A.; Ford, F.A. *Modeling the Environment: An Introduction to System Dynamics Models of Environmental Systems*; Island press: Washington, DC, USA, 1999.
29. Heath, B.L.; Ciarallo, F.W.; Hill, R.R. Validation in the agent-based modelling paradigm: Problems and a solution. *Int. J. Simul. Process Model.* **2012**, *7*, 229–239. [CrossRef]
30. An, L.; Grimm, V.; Bai, Y.; Sullivan, A.; Turner II, B.; Malleson, N.; Heppenstall, A.; Vincenot, C.; Robinson, D.; Ye, X. Modeling agent decision and behavior in the light of data science and artificial intelligence. *Environ. Model. Softw.* **2023**, *166*, 105713. [CrossRef]
31. Balke, T.; Gilbert, N. How Do Agents Make Decisions? A Survey. *J. Artif. Soc. Soc. Simul.* **2014**, *17*, 13. [CrossRef]
32. Doscher, C.; Moore, K.; Smallman, C.; Wilson, J.; Simmons, D. An Agent-Based Model of Tourist Movements in New Zealand. In *Empirical Agent-Based Modelling-Challenges and Solutions: Volume 1, The Characterisation and Parameterisation of Empirical Agent-Based Models*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 39–51.
33. Edwards-Jones, G. Modelling farmer decision-making: Concepts, progress and challenges. *Anim. Sci.* **2006**, *82*, 783–790. [CrossRef]

34. Janssen, M.; Jager, W. An integrated approach to simulating behavioural processes: A case study of the lock-in of consumption patterns. *J. Artif. Soc. Soc. Simul.* **1999**, *2*, 21–35.
35. Becu, N.; Barreteau, O.; Perez, P.; Saising, J.; Sungted, S. A methodology for identifying and for-malizing farmers' representations of watershed management: A case study from northern Thailand. In *Companion Modeling and Multi-Agent Systems for Integrated Natural Resource Management in Asia*; International Rice Research Institute: Manila, Philippines, 2005; p. 41.
36. Voinov, A.; Bousquet, F. Modelling with stakeholders. *Environ. Model. Softw.* **2010**, *25*, 1268–1281. [CrossRef]
37. Bharwani, S. Understanding complex behavior and decision making using ethnographic knowledge elicitation tools (KnETs). *Soc. Sci. Comput. Rev.* **2006**, *24*, 78–105. [CrossRef]
38. Edmonds, B. A context-and scope-sensitive analysis of narrative data to aid the specification of agent behaviour. *J. Artif. Soc. Soc. Simul.* **2015**, *18*, 17. [CrossRef]
39. Ghorbani, A.; Schrauwen, N.; Dijkema, G.P.J. Using Ethnographic Information to Conceptualize Agent-based Models. In Proceedings of the European Social Simulation Association Conference, Warsaw, Poland, 16–20 September 2013.
40. Gilbert, N.; Terna, P. How to build and use agent-based models in social science. *Mind Soc.* **2000**, *1*, 57–72. [CrossRef]
41. Huigen, M.G. First principles of the MameLuke multi-actor modelling framework for land use change, illustrated with a Philippine case study. *J. Environ. Manag.* **2004**, *72*, 5–21. [CrossRef]
42. Clark, M.; Kim, Y.; Kruschwitz, U.; Song, D.; Albakour, D.; Dignum, S.; Beresi, U.C.; Fasli, M.; De Roeck, A. Automatically structuring domain knowledge from text: An overview of current research. *Inf. Process. Manag.* **2012**, *48*, 552–568. [CrossRef]
43. Al-Safadi, L.A.E. Natural Language Processing for Conceptual Modeling. *JDCTA* **2009**, *3*, 47–59.
44. Navigli, R. Word sense disambiguation: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–69. [CrossRef]
45. Husain, M.S.; Khanum, M.A. Word Sense Disambiguation in Software Requirement Specifications Using WordNet and Association Mining Rule. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, Udaipur, India, 4–5 March 2016; pp. 1–4.
46. Orkphol, K.; Yang, W. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet* **2019**, *11*, 114. [CrossRef]
47. Fraga, A.; Moreno, V.; Parra, E.; Garcia, J. Extraction of Patterns Using NLP: Genetic Deafness. In Proceedings of the SEKE, Pittsburgh, PA, USA, 5–7 July 2017; pp. 428–431.
48. Liddy, E.D. Natural Language Processing. 2001. Available online: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub (accessed on 28 June 2023).
49. Loper, E.; Bird, S. NLTK: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, PA, USA, 7 July 2002; pp. 63–70.
50. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd annual meeting of the association for computational linguistics: System demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
51. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. Using text mining techniques for extracting information from research articles. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 373–397.
52. Sun, S.; Luo, C.; Chen, J. A review of natural language processing techniques for opinion mining systems. *Inf. Fusion.* **2017**, *36*, 10–25. [CrossRef]
53. Bird, S.; Klein, E.; Loper, E. *Natural language processing with Python: Analyzing text with the natural language toolkit*; O'Reilly Media Inc.: Sebastopol, CA, USA, 2009.
54. Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, FL, USA, 23–25 October 2003; pp. 70–77.
55. Harris, L.R. The ROBOT System: Natural language processing applied to data base query. In Proceedings of the 1978 Annual Conference, Washington, DC, USA, 4–6 December 1978; pp. 165–172.
56. Lees, B. Artificial Intelligence Education for Software Engineers. In *WIT Transactions on Information and Communication Technologies*; 1970; Volume 12. Available online: https://www.witpress.com/elibrary/wit-transactions-on-information-and-communication-technologies/12/10537 (accessed on 28 June 2023).
57. Runck, B.C.; Manson, S.; Shook, E.; Gini, M.; Jordan, N. Using word embeddings to generate data-driven human agent decision-making from natural language. *GeoInformatica* **2019**, *23*, 221–242. [CrossRef]
58. Padilla, J.J.; Shuttleworth, D.; O'Brien, K. Agent-Based Model Characterization Using Natural Language Processing. In Proceedings of the 2019 Winter Simulation Conference (WSC), National Harbor, MD, USA, 8–11 December 2019; pp. 560–571.
59. Heppenstall, A.; Crooks, A.; Malleson, N.; Manley, E.; Ge, J.; Batty, M. Future Developments in Geographical Agent-Based Models: Challenges and Opportunities. *Geogr. Anal.* **2021**, *53*, 76–91. [CrossRef] [PubMed]
60. Liang, X.; Luo, L.; Hu, S.; Li, Y. Mapping the knowledge frontiers and evolution of decision making based on agent-based modeling. *Knowl.-Based Syst.* **2022**, *250*, 108982. [CrossRef]
61. Harmain, H.M.; Gaizauskas, R. CM-Builder: An automated NL-based CASE tool. In Proceedings of the ASE 2000 Fifteenth IEEE International Conference on Automated Software Engineering, Grenoble, France, 11–15 September 2000; pp. 45–53.
62. Bersini, H. UML for ABM. *J. Artif. Soc. Soc. Simul.* **2012**, *15*, 9. [CrossRef]

63. Collins, A.; Petty, M.; Vernon-Bido, D.; Sherfey, S. A Call to Arms: Standards for Agent-Based Modeling and Simulation. *J. Artif. Soc. Soc. Simul.* **2015**, *18*, 12. [CrossRef]
64. Bakam, I.; Kordon, F.; Le Page, C.; Bousquet, F. Formalization of a spatialized multiagent model using coloured petri nets for the study of an hunting management system. In Proceedings of the International Workshop on Formal Approaches to Agent-Based Systems, Greenbelt, MD, USA, 5–7 April 2000; pp. 123–132.
65. Gilbert, N. Agent-based social simulation: Dealing with complexity. *Complex. Syst. Netw. Excell.* **2004**, *9*, 1–14.
66. Miller, J.H.; Page, S.E. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*; Princeton University Press: Princeton, NJ, USA, 2009.
67. Manning, C.; Raghavan, P.; Schütze, H. Introduction to information retrieval. *Nat. Lang. Eng.* **2010**, *16*, 100–103.
68. Namey, E.; Guest, G.; Thairu, L.; Johnson, L. Data reduction techniques for large qualitative data sets. *Handb. Team-Based Qual. Res.* **2008**, *2*, 137–161.
69. Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, Piscataway, NJ, USA, 3–8 December 2003; pp. 133–142.
70. Rivers III, L.; Sanga, U.; Sidibe, A.; Wood, A.; Paudel, R.; Marquart-Pyatt, S.T.; Ligmann-Zielinska, A.; Olabisi, L.S.; Du, E.J.; Liverpool-Tasie, S. Mental models of food security in rural Mali. *Environ. Syst. Decis.* **2017**, *38*, 33–51. [CrossRef]
71. Ligmann-Zielinska, A.; Sun, L. Applying time-dependent variance-based global sensitivity analysis to represent the dynamics of an agent-based model of land use change. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1829–1850. [CrossRef]
72. Xiang, X.; Kennedy, R.; Madey, G.; Cabaniss, S. Verification and validation of agent-based scientific simulation models. In Proceedings of the Agent-Directed Simulation Conference, San Diego, CA, USA, 3 April 2005; p. 55.