Journal of Digital Terminology and Lexicography - ISSN:3103-3601
Vol. 1, No. 2 (2025) DOI: 10.25430/pupj.jdtl.1762210900

Framing Food Online Discourse: Employing Generative AI and Semantic Analysis for Digital Lexicography and Terminology Extraction

Malamatenia Panagiotou[1,2*†], Konstantinos Gkatzionis[1†] and Efstathios Kaloudis[2†]

[1*]Laboratory of Consumer and Sensory Perception of Food and Drinks, Department of Food Science and Nutrition, University of the Aegean, Metropolite Ioakeim 2, Myrina, 81400, Lemnos, Greece.

[2]Computer Simulation, Genomics and Data Analysis Laboratory, Department of Food Science and Nutrition, University of the Aegean, Leoforos Dimokratias 66, Myrina, 81400, Lemnos, Greece.

*Corresponding author(s). E-mail(s): teniapanag@aegean.gr
Contributing authors: kgkatzionis@aegean.gr; stathiskaloudis@aegean.gr
†These authors contributed equally to this work.

Abstract

Social media platforms provide vast amounts of authentic, user-generated linguistic data that can reveal conceptual structures, terminology, and cultural discourse. This study analyzes Instagram posts related to traditional and local foods in Greece, focusing on semantic frames, key concepts, and lexically salient items that contribute to food consumer identity. Using a bottom-up approach based on hashtag co-occurrence, we identified frames of locality, tradition, and food consumption as culturally embedded structures. Two generative AI tools, ChatGPT (English-based) and Llama-Krikri (Greek-based), were employed for food-relatedness and sentiment classification; ChatGPT achieved very high agreement with human judgments (F1 > 90%), while Llama-Krikri showed more modest performance. Case studies on Mediterranean snails and North Aegean cheeses revealed prototypical associations, with graviera and feta emerging as central references. Importantly, food names and related concepts are treated as terminological units, and selected cases were transformed into structured entries using TBX and OntoLex-Lemon, demonstrating applications to digital terminology and lexicography. The methodology highlights both the opportunities (frame-based vocabulary extraction, resource building) and challenges (Greeklish, spelling variation) of applying large language models to under-resourced languages. This work shows how social-media-based analyses can support the creation of interoperable, frame-based terminological and lexicographical resources.

Keywords: social media, large language models, specialized lexicography, digital terminology, semantic frames

1 Introduction

Product failure in the marketplace happens even if products show high liking scores by consumers in lab and in-house tests. Thus, alternative methods are being used in sensory and consumer studies to obtain data from consumers in environments outside the lab, one being social media-based methods. Online social media networking sites, content communities, online reviews, forums, and blogs provide a rich and

expansive source of qualitative data that can be analyzed in a quantitative manner (Hutchings et al., 2023). Examples of social media platforms currently used for language and food related research are Facebook, Instagram, Twitter, and Reddit.

1.1 Related Work and State of the Art

Manual handling is not feasible when managing substantial amounts of data nor can it be a measure of accuracy, but it can provide insight into the accuracy of current Natural Language Processing (NLP) tools (Hutchings et al., 2023). Data collection can be done either by web scraping tools that collect publicly available data from social media websites quickly and automatically, and extract it into a well-structured format, readable by humans and machines, easy to access, and lightweight for storage (Nigam & Biswas, 2021). A wide range of programming languages support reading and processing of collected data, such as Python.

Sentiment analysis is an analysis method also applicable in food research on social media. Sentiment analysis is the computational study of people's opinions, emotions, and attitudes towards entities, and topics (Aggarwal & Zhai, 2012). Sentiment classification of posts is either formulated as a two-class (positive, negative) or three-class (positive, neutral, negative) supervised learning problem. Machine Learning algorithms, which learn how to identify the valence of each word, i.e., the dimensional aspect of emotional experience varying from pleasant to unpleasant (Barrett, 2006) within a specific context, are commonly used in sentiment analysis tasks. When every word of the post has been assigned a score, the sum of scores is computed, thus determining whether the post is positive, negative, or neutral (and how much so) (Tao, Yang, & Feng, 2020). This type of information is valuable to food and marketing companies who want to know how consumers feel about their products and brand. It is also of interest to linguists and anyone studying culture by providing insight into words in context.

1.2 Theoretical Background for Data Analysis

In linguistic and cognitive theory, a semantic frame is a cognitive structure that represents a particular type of event, scenario, or object and the participants, props, and roles associated with it. The concept was developed within Frame Semantics, a theory proposed by Charles J. Fillmore in the 1970s (Fillmore, 1977). According to this theory, understanding the meaning of a word requires knowledge of the broader conceptual structure – or frame – it evokes. For example, the word "buy" evokes a commercial transaction frame, which includes roles such as buyer, seller, goods, and money. Other words like "sell", "pay", and "cost" also evoke this frame, highlighting the interconnectedness of vocabulary through shared conceptual structures. As regards lexicography, frame-based lexicons organize words according to the frames they evoke, helping lexicographers group related lexical items and explain their meaning through contextual roles and semantic relations (Ostermann, 2014). As regards terminology, frames provide a structured, concept-oriented approach to capturing domain knowledge, and support ontological modeling (Faber, 2015).

1.3 Present Work and Contribution

The aim of the present study is to understand how geographical factors, namely the country or place of origin of a food product or a recipe, affect consumers' response and emotions to foods, using a novel methodology that incorporates NLP models. The main objectives are a) to identify the way consumers feel and talk about traditional and local foods as opposed to relevant novel ones on social media, b) to investigate how sensory attributes, geographical characteristics, nutritiousness, and environmental concerns impact consumers' food choice, and c) to identify relevant concepts. The methodology developed was evaluated in the case studies of a) snails, a traditional food for the Mediterranean market and a sustainable alternative to meat, and b) local cheeses of the North-Aegean Sea islands. The present study is original as regards the methodology and tools applied. More specifically, a new methodology has been developed for the mining and handling of data on Instagram for specific language related purposes, combining existing programming and NLP tools in an original way to decrease the need for manual data handling. In addition, ChatGPT and Llama KriKri are used for automated application of criteria as a substitute for human to save time and ensure repeatability of results, and, to compensate for the lack of geotagging information provided for Instagram posts, a geographical tag (using the prefectures of Greece) based on relevant

preliminary hashtag identification per case study			
identify spelling variations		identify respective Greeklish forms	
↓			
collect posts			
merge posts per case study	remove duplicate posts	remove posts in languages other than Greek, Greeklish, English	
↓			
clean data			
nonsensical words	pronouns, articles, prepositions	names of businesses	hashtags used for data collection
↓			
remove non food-related posts			
using an AI application		manually	
↓			
perform sentiment analysis			
using an AI application		manually	
↓			
perform post-geotagging			
↓			
semantic grouping of hashtags			
↓			
hashtag measurements			
frequency of appearance		engagement*	
↓			
analyse food-related concepts per case study			
key drivers of consumption	conceptual profile	sensory profile	emotional profile

Table 1 Diagram of workflow for data collection and analysis (*engagement measurement is not presented in this paper as it is outside the scope of the present publication).

hashtags has been added to the respective posts. The present paper builds upon previously published work (Panagiotou, Gkatzionis, & Kaloudis, 2025a) and focuses on key semantic frames and concepts/words that populate them were identified – and analyzed – relevant to locality, tradition, and food consumption.

2 Methodology

Instagram was the social media platform chosen for this study because Instagram users interact with companies more often than on other platforms in Greece, and cooking comes second (together with health/fitness) among the most common interests of Greek Instagram users (Gewiese & Rau, 2023). The workflow for data collection and analysis is presented in Table 1. The complete hashtag datasets and categorization files are available in the project's online repository[1], in addition to the Supplementary Materials.

2.1 Data Collection, Cleaning of Duplicates and Irrelevant Posts, and Preparation for Analysis

A preliminary search was conducted using Apify (Apify, 2024), a web scraping tool, in order to identify Greeklish (i.e., non-standardized idiom of Greek in which words are transliterated using the Latin alphabet based on how they are written or pronounced)[2] (see Supplementary Table 1) and misspelled forms of hashtags (i.e., words or phrases preceded by the symbol # used to classify the accompanying text) of interest, as such hashtags are a common phenomenon on social media. The hashtags used for post collection were specific to each case study (see Supplementary Table 2). The rendered posts included the same hashtags in various Greeklish forms, various misspelled forms, and relevant hashtags. The top 100 hashtags (appearing the most frequently) in Greek, English, and Greeklish were then used to collect the posts for the main study. Social media users tend to use English hashtags extensively, even when English is not their native language, to increase audience reach. The coexistence of English and Greek hashtags elicits

[1] https://zenodo.org/records/17478417

[2] The initial list of hashtags included Greeklish hashtags following a rule-based transliteration. For example, the Greek word for snails can be transliterated as "saligkaria", "saliggaria", "saligaria", or in more regional forms as "xoxlioi", "chochlioi", or "hohlioi".

the assumption that the account is of Greek origin and, thus, these posts were included in the study. The dataset consists of hashtags and terms referring to cheeses, their origins, qualities, and contexts of use. In this study, we treat all names of foods and the related concepts that describe their production, consumption, and evaluation as food terminology. This framing positions the lexical material not simply as casual labels but as structured terminological units that can be analyzed, compared, and integrated into formal resources.

Posts from April 2012 (Instagram platform release) to March 2023 were collected. Posts were merged into a single file and duplicates were removed. Data obtained for each post contained: post id, type of post [image, sidecar (i.e., group of images), video], shortCode (shortened URL of the post), caption, hashtags, number of comments, number of likes, timestamp, and whether it appeared on a professional account or not. The data were further cleaned by removing hashtags that belong to languages other than Greek, Greeklish, or English, nonsensical words, parts of speech that were not meaningful for the present study (pronouns, articles, and prepositions), names of businesses (producers, sellers, restaurants), and the original hashtags that had been used for data collection. In addition to Apify, a set of Python scripts was developed specifically for this study to automate the cleaning, merging, and filtering of Instagram posts, as well as to preprocess hashtag metadata and detect Greeklish variations. These scripts also allowed seamless interaction with large language models (LLMs), enabling automated application of classification criteria (e.g., food-relatedness and sentiment) and structured storage of model responses. This toolset supported reproducible workflows and minimized manual effort across multiple stages of data handling and analysis.

2.2 Applying the Food-Relatedness Criterion

ChatGPT, a novel Artificial Intelligence (AI) application (edition 3.5; OpenAI, San Francisco, CA, USA), and Llama-Krikri-8B-Instruct (Roussis et al., 2025), a locally deployed instruction-tuned model based on Meta's Llama 3 architecture, were used for the automated application of the food-relatedness criterion, so that only food-related posts would be left for further analysis[3]. This was performed on snail posts only, as posts on cheese refer to food *de facto*. ChatGPT was chosen because: a) it can process data in multiple languages, b) it can automate the processing of data in Python, and c) it is low cost. The instruction was: "Process only food related posts". Llama KriKri was chosen as it is the first AI application trained on Greek language data. The linguist of the research team subsequently performed a manual check regarding food-relatedness. This was done to evaluate the degree of agreement between human and machine responses and explore the potential of the machine in substituting manual data management to save time and effort.

2.3 Sentiment Analysis of Posts

ChatGPT and Llama KriKri checked the posts and provided a response (i.e., positive, neutral, negative) for each post regarding sentiment, instructed to take into consideration the following components of each post: caption, emojis/emoticons, hashtag(s)[4]. The linguist of the research team subsequently performed a manual check regarding sentiment. As previously described, this was done for comparison purposes. The posts were grouped per sentiment category (positive, neutral, negative) and per hashtag.

2.4 Tagging for Regionality of Posts

Instagram does not offer post-geotagging information due to personal information protection reasons. Since locality was one of the focus points of the study, a region tag was added to those hashtags that were names of a city, island, or area of Greece, using the thirteen prefectures of Greece as reference.

2.5 Identifying Key Concepts and Concept Frames

The data collected for the present study were analyzed within the theoretical framework of semantic frames, employing a bottom-up, data-driven approach. The analysis began with the actual linguistic material – specifically, the hashtags used in the Instagram posts – and identified semantic associations by measuring their co-occurrence within posts. This method allowed for the emergence of key concepts,

[3]Prompt: Classify the following Instagram post as related or non-related to food: {post}. Respond with one word: true or false.
[4]Prompt: Classify the sentiment of this Instagram post: {post}. Respond with one word: Positive, Negative, Neutral.

the mapping of relationships among them, and the subsequent construction of semantic frames that capture the conceptual structure underlying the discourse. Given the study's focus on local and traditional foods, particular attention was paid to aspects of food consumption that shape consumer perception, cultural identity, and emotional engagement. The frames that emerged reflect how food is not only a nutritional or economic commodity but also a socially and symbolically charged activity, embedded in broader experiential and cultural contexts.

2.6 From Frames to Terminological Entries

The transition from semantic frames to terminological entries was operationalized in four stages. First, frames relevant to food discourse (e.g. *Consumption*) were identified from the co-occurrence structure of hashtags. Second, frame elements (roles such as *CheeseVariety, Origin, Occasion, FoodPairing, Assessment, Emotion*) were derived from semantic categories. Third, lexical units (hashtags and terms) were assigned to frame elements, creating inventories of language that populate the frames. Finally, these units were transformed into structured terminological entries that record the term, definition, usage context, semantic roles, and relations to other concepts. The study did not aim to systematically transform all hashtags or terms into structured terminological or lexicographical entries; rather, selected cases were developed as illustrative examples to demonstrate the potential extension of the methodology. This approach makes it explicit how linguistic patterns in social media can be transferred into structured terminology resources.

3 Results and Discussion

3.1 Human and Machine Agreement on Food Relatedness Criterion and Sentiment Analysis

Both ChatGPT and human identified 44% of the posts as food-related and 47% of them as not food-related, thus reaching an overall agreement rate of 91%. Beyond-chance agreement was confirmed using kappa test. Additionally, sentiment analysis revealed high agreement between ChatGPT and human evaluations, with 61% of the posts identified as positive and 37% as neutral, resulting in an overall agreement rate of 98%, beyond chance. Quantitative evaluation metrics also confirmed the effectiveness of ChatGPT in both classification tasks. For the food-relatedness classification, ChatGPT achieved a precision of 92.4%, a recall of 88.9%, and an F1-score of 90.6%. In the sentiment classification task, class-wise performance was also high: for the "positive" sentiment class, precision was 99.8%, recall 96.6%, and F1-score 98.2%; for the "neutral" class, precision was 94.6%, recall 99.7%, and F1-score 97.1%. No post was identified as negative by either human or ChatGPT.

On the other hand, the Krikri model achieved an overall agreement rate of 63.3% for the food-relatedness classification, with beyond-chance agreement confirmed by Cohen's kappa (κ=0.27). Sentiment analysis showed a higher level of agreement, with an overall accuracy of 72.3% and a moderate beyond-chance agreement (κ=0.42). Quantitative evaluation metrics indicated that for the food-relatedness classification, the Krikri model achieved a precision of 61.5%, a recall of 68.9%, and an F1-score of 65.0% for the "true" class. For the sentiment classification, the "positive" class achieved a precision of 79.3%, recall of 79.6%, and F1-score of 79.4%; the "neutral" class showed a precision of 64.5%, recall of 59.9%, and F1-score of 62.1%. The "negative" sentiment class appeared in only two cases, limiting the robustness of its evaluation.

However, ChatGPT's and Llama Krikri's performance in identifying "Greeklish" was unsatisfactory and inconsistent. More specifically, when instructed to categorize hashtags according to language (Greek, English, Greeklish), in order to subsequently transcribe Greeklish into Greek for further analysis, the applications consistently identified Greeklish as English. In an attempt to rectify that, they were provided with a list of hashtags in Greeklish matched to their Greek equivalents and instructed to categorize the original list again based on this input. However, the new output was again erroneous, identifying Greeklish as either Greeklish or English. Every time they were asked to perform the same task the categorization of the same hashtags was different. As a result, the task was performed manually.

Regarding the performance of the two GenAI applications in the tasks assigned to them, when compared with ChatGPT, which previously demonstrated exceptionally high agreement rates with human annotators (91% for food-relatedness and 98% for sentiment classification) the Krikri model shows more modest alignment with human evaluations. ChatGPT's advantage is partly attributable to its substantially

larger model size and far broader multilingual training, which allow it to capture nuanced semantic and emotional cues across a wide range of contexts and languages. In contrast, Llama-Krikri-8B, although a cutting-edge large language model built on Meta's Llama 3.1-8B and extensively fine-tuned for Greek, operates with a smaller parameter count relative to ChatGPT and has been primarily trained on Greek-language data. This strong Greek linguistic focus, including Modern, Ancient, and polytonic Greek, equips Krikri with exceptional capabilities for deep Greek text understanding but may limit its adaptability to broader sentiment and thematic classification tasks, especially when domain-specific or cross-linguistic nuances are involved. Furthermore, given that the dataset used in this study is predominantly in Greek, Krikri's moderate performance relative to ChatGPT suggests that, while language specialization provides advantages in text comprehension, ChatGPT's scale, training diversity, and advanced contextual modeling still give it an edge in accurately distinguishing sentiment and identifying food-related content. This is particularly evident in the handling of ambiguous sentiment cases, where Krikri tends to conflate neutral and positive categories, contributing to lower agreement rates.

3.2 Food-Related Concepts Identified

3.2.1 Sensory, Nutritional, and Environmental Aspects and Drivers of Consumption

Certain key concepts related to the foods under study were identified in the collected posts. Firstly, concepts related specifically to the sensory aspect of eating were taste (the sense), tasting (the act), tastiness (the food quality), and gourmet eating. They were found as hashtags, such as #taste, #winetasting, #eat, #slurp, #wow, #delicious, #tasty, #tastyrecipes, appearing with high frequency. The co-occurrence of hashtags suggests that traditional and local cheeses are frequently described with sensory-evaluative language (e.g., tasty, authentic, delicious). While this points to an association between local food and sensory satisfaction, we acknowledge that such patterns may be amplified by the positive and promotional character of Instagram food discourse. Secondly, concepts referring to nutritional value (e.g., #protein, #omega3, #organic, #vitamins) and dieting styles (e.g., #keto, #vegan, #glutenfree, #eatclean, #healthylifestyle, #vegetarianfood) were identified but did not appear with high frequency. This may mean that nutritional content is not the focus or goal when consuming or posting about traditional and local foods. Thirdly, hashtags pertaining to the environmental aspect of food production, such as #cleanandgreen, #ethical, sustainable, #bio, #greenfood, were not frequently mentioned. This implies either that the Greek consumer has not cultivated environmental concerns to a significant degree yet, or that this aspect of food production is not relevant to traditional and local food consumption. The fact that the two concepts coappear in posts, however infrequent, suggests that the two are connected, and it was identified that it is important for the food to be locally and organically produced.

The 1836 word-hashtags collected were categorized into semantic categories, using a Greek thesaurus and the researchers' opinion for cases that could not fit like multi-word hashtags or new social-media-native words. Hashtags that belonged to more than one category were duplicated (e.g., #tastyrecipes was categorized as post content and assessment). The twenty categories identified in the snail study were: activity, assessment, dance, diet, drink, emotion, food, food preparation technique, music, nature, nutritional content, occasion, origin, people, personal touch, place of consumption, place of purchase, post content, region, time. The fifteen categories identified in the cheese study were: cheese variety, food pairing, origin, occasion, emotion, cooking, social media, identity, assessment, nutrition, aesthetics, commerce, sustainability, utensils, time (see Supplementary Table 3). These semantic categories led to drawing conclusions in a more meaningful and organized way. Hashtag grouping into semantic categories was done manually in the snail case study, but in the cheese case study Generative Artificial Intelligence (GenAI) (ChatGPT and Copilot) was assessed - as in other food related case studies - and was able to group hashtags in a meaningful way. For cheeses, categories and hashtag categorization by GenAI were checked by humans and the ones by ChatGPT were preferred and kept for further analysis. Automated categorization is necessary to minimize time and effort.

We were able to identify key drivers of consumption, and draw the conceptual, sensory, and emotional profiles of the foods under study. For example, as regards cheeses, intended use, familiarity, and price are the main drivers of consumption (in the order mentioned). Local cheeses correlate with concepts such as tradition, family, granny, summer, bread, salad, olive oil, and honey, while non-Greek cheeses with concepts such as wine, Italy, pasta, fast food, pizza, and mushrooms. Sensorially, local cheeses are thought

to be hard and salty, while non-Greek cheeses are considered mainly gummy, creamy, fatty, and soft. Local cheeses elicit emotions by bringing forth childhood memories, as opposed to non-Greek ones that elicit emotions of sensuality (Panagiotou et al., 2024).

As regards the snail-related study, when consuming (or posting about) traditional foods, nutrition or dieting styles seem to be irrelevant, and the emotional, cultural, and regional aspects of the foods are highlighted more often than taste (in any sense of the word). Traditional food consumption relates to authenticity, health, simplicity, freshness of ingredients, homemade cooking, hospitality, respect for the cook, passion (a.k.a. Greek meraki [5]), and emotions of happiness, love, care, nostalgia, liveliness, bliss, fun, and comfort, thus depicting a link between traditional food consumption and community in interaction. Community has always been an integral part of the Greek culture (Katsanevaki, 2010; SBS, International Education Services Multicultural NSW, 2016). Food, from production to consumption, is a social activity, especially for the Greeks (Delormier, Frohlich, & Potvin, 2009; Dunbar, 2017; Hanna, Cross, Nicholls, & Gallegos, 2023). On the other hand, non-traditional foods containing snails (e.g., in powder or dried fillet forms) highlighted the nutritional content and the fact that snail meat is protein without being what one would consider meat. It is considered a fasting dish, and it appears with hashtags referring to keto, paleo, and gluten-free eating, all generally promoted as healthy dieting styles. Dishes containing alternative snail meat forms are also characterized as creative, original, and gourmet, thus depicting a wellness-oriented lifestyle, a person-centered approach to eating, with a focus on health, beauty, and nutrition, which stand in contrast to the tradition- and community-oriented aspects of snail eating.

3.2.2 The Semantic Frames of "Locality", "Tradition", and "Food Consumption", and Prototypical Foods

Semantic frames were identified using GenAI applications and manually checked and evaluated. The analysis revealed that the semantic frames identified were closely related to the overarching concepts of locality, tradition, and food consumption as a culturally embedded activity. These frames provided valuable insight into the phenomena of food consumer ethnocentrism and regiocentrism, reflecting preferences for national and local foods, respectively. Using the methodological approach presented in this study, the frames were automatically populated with relevant words and concepts (Fig. 1), illustrating the potential of bottom-up data analysis in uncovering the structure and meaning of food-related discourse. The frame of food consumption, in particular, was enriched with vocabulary pertaining to meat and cheese consumption, reflecting the study's thematic focus. Notably, references to meatless eating appeared in connection with religious or, to a lesser extent, health-related motivations among Greek consumers, while cheese consumption was often described in relation to its intended culinary use.

All three frames elicited emotionally charged associations, highlighting the symbolic and experiential dimensions of food. The locality frame included references to the geographical origin of ingredients or recipes (e.g., place names), sites of purchase and consumption (e.g., local businesses, taverns, or family homes), and regional or national identities (e.g., Greek, Cretan). These linguistic patterns underscored the strong connection between place-based identity and traditional food production. The tradition frame was populated with words linked to culinary practices (e.g., boubouristoi, tsigariasto) and consumption occasions tied to social or religious rituals (e.g., Easter, Mother's Day). Finally, the frame of food consumption as an activity was conceptually interwoven with other frames of human experience, such as dance, music, drinking, and nature, as well as with social actors like family members, community figures, producers, and cooks. These findings demonstrate how food-related discourse is deeply embedded in broader cultural narratives and practices, reinforcing the interconnectedness of language, identity, and experience in the context of food.

We adopt a frame-semantic approach in which a semantic frame is a schematic representation of a recurrent situation evoked by lexical units and structured by core and non-core frame elements. In our domain, we instantiate a Consumption frame with core elements CheeseVariety (the ingested item), FoodPairing, and Utensil/Instrument, and non-core elements Origin, Occasion, Assessment, Emotion, and Time. To avoid treating broad semantic fields as frames, we treat Locality and Tradition as non-core facets realized within Consumption rather than as standalone frames, that in turn are realized by other elements like

[5] "Meraki" is a Greek word that describes putting your heart, soul, and creativity into something (e.g., work or an artistic endeavor). It conveys a deep sense of love, care, and personal investment in what one does.

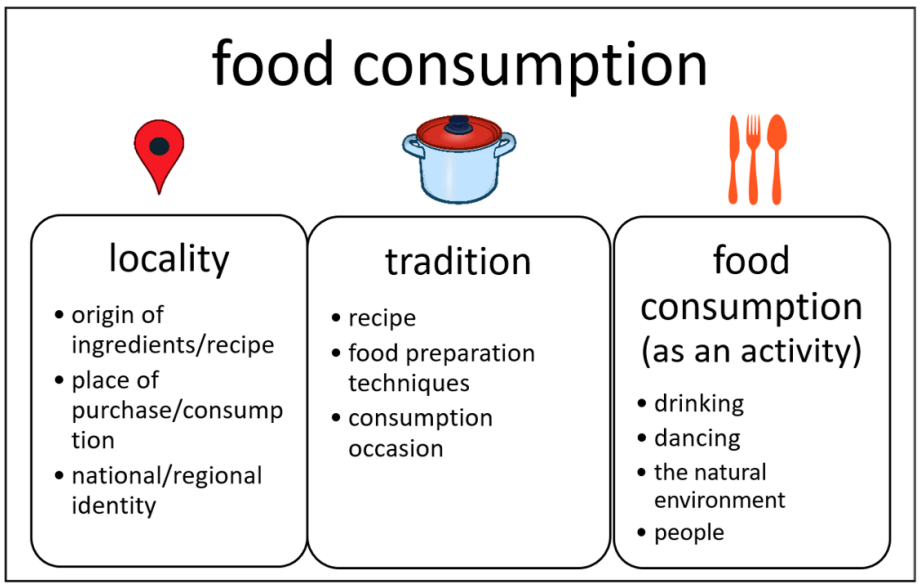

Fig. 1 Semantic frames identified in the present study focusing on consumer ethnocentrism/regiocentrism and consumption of traditional and local foods (meat alternatives and cheeses specifically)

origin, place of consumption, place of purchase, recipe, occasion, people etc. (Fig. 1). Lexical units (hashtags/terms) populate the frame by aligning to the most specific applicable element; multi-facet items are linked to multiple elements. This structure supports both analysis and downstream resource building (terminological and lexicographical entries). For example, #graviera populates CheeseVariety (core), co-occurs with Origin = Crete/Naxos, FoodPairing = wine/pasta/salad, Occasion = meze/dinner, and contributes Assessment (authentic/traditional) and Emotion (nostalgia/pride). This frame structure feeds a domain-specific definition template ("X is a [cheese variety] produced in [Origin] ...") and yields entries that are expressible both as terminological records (e.g., TBX) (see Supplementary Table 4) and as lexicographical entries (definition, usage, collocations).

Furthermore, the present methodology was able to provide prototypical cheeses per use, such as imported yellow soft and gummy cheeses for everyday snacks (e.g., gouda or cheddar for toasties and pizza). The first cheese that comes to mind when cheese is mentioned (also confirmed from studies with focus groups and questionnaires) is graviera, emerging as the prototypical cheese. On the other hand, when Greek cheeses are discussed, the prototypical one is feta.

3.3 Example of Frame-Based Terminological Entry

The hashtag #graviera exemplifies how frames give rise to terminological entries. Within the Consumption frame, it was linked to multiple elements: CheeseVariety (graviera), Origin (Crete, Naxos), Occasion (meze, dinner, holiday meal), FoodPairing (wine, pasta, salad), Assessment (authentic, traditional, tasty, PDO), and Emotion (nostalgia, pride, enjoyment). These roles supply the internal structure of the entry, specifying how the term relates to other concepts and contexts of use. A TBX version (International Organization for Standardization, 2019) of this entry is provided in the Supplementary, alongside an OntoLex-Lemon representation (McCrae, Bosque-Gil, Gracia, Buitelaar, & Cimiano, 2017) (see Supplementary Table 4), demonstrating how the same data can be encoded in two complementary models. The explicit encoding of frames and frame elements within terminological entries enables integration into existing infrastructures. In TBX, frame and role information can be represented in descrip elements within term entries, facilitating export to multilingual termbases such as IATE[6] or national terminology banks. In OntoLex-Lemon, lexical entries evoke frames through the ontolex:evokes relation, while frame elements are modeled as properties linking to related entities. This alignment ensures interoperability with the Linguistic Linked Open Data cloud and other linked lexica (Declerck, Lendvai, Mörth, Budin, & Váradi, 2012). The approach thus bridges social-media-based analyses and standard terminological practice, expanding the coverage

[6] https://iate.europa.eu/

of under-represented domains such as food and gastronomy. In the same way that our data can be transformed into structured terminological entries, the same frame-based inventories of terms, definitions, usage examples, and semantic relations also constitute the basic microstructure of a lexicographical entry. Thus, the results are not only suitable for termbases and standard models such as TBX or OntoLex-Lemon but can also feed directly into dictionary-making processes, ensuring consistency between terminology management and lexicography.

3.4 Limitations and Future Perspectives

The present study was culture oriented and thus worked on identifying and solving problems specific to posts containing Greek, Greeklish, and English hashtags of interest. Post collection was also restricted to Instagram platform. However, the methodology and tools applied can be extended to other languages and platforms. Challenges in data collection from other platforms, such as Reddit, remain to be identified. Certain tools, like ChatGPT and Llama Krikri, could be further tested using different settings to deliver more accurate and repeatable results. What is more, translation of hashtags into English could be a route to be tested to ensure uniformity in the language of hashtags before further analysis. In that case, a wider variety of NLP tools would be available.

Future research will build upon the current study by extending data collection beyond social media platforms to a broader range of online sources, including consumer reviews, food blogs, digital magazines, and e-commerce platforms. Because Instagram posts typically highlight appetitive, positive qualities of food, the frequent pairing of traditional/local terms with sensory descriptors may reflect platform conventions and trending practices as much as genuine consumer evaluation. This ongoing extension outside beyond social media aims to further investigate the digital identity of agrifood products, with particular focus on the representation of sensory attributes, salient conceptual associations, and consumer attitudes and expectations. Expanding the corpus in this way will allow for the identification of cross-platform patterns and a more comprehensive understanding of how traditional and local foods are conceptualized and communicated in digital environments. In addition, the databases generated through this methodology can be further enriched with hashtags and key terms related to other culturally significant food products, including those from different linguistic backgrounds. Addressing the limitations encountered – such as the inconsistent handling of Greeklish and orthographic variation by current NLP systems – will be central to improving the accuracy and adaptability of language models, particularly for under-resourced languages like Greek. Advances in this area will enhance the potential of large language models and semantic frame analysis to support applications in digital lexicography, food communication research, terminology management, and the monitoring of evolving consumer discourse in the agrifood sector.

3.5 Interested Parties

The findings and methodology presented in this study are relevant to a diverse range of stakeholders across multiple disciplines. Linguists and applied linguists focusing on semantic frames, corpus analysis, and digital language data will find value in the bottom-up approach to conceptual structure identification using social media data. Researchers in food studies and cultural anthropology can leverage these insights to better understand the construction of food consumer identity and the cultural embedding of traditional and local foods in Greece. The use of GenAI tools for sentiment analysis and classification also appeals to computational linguists and digital humanities scholars interested in advancing automated methods for language and cultural analysis. Marketing and consumer behavior analysts focusing on food consumption trends and identity formation via social media platforms are another key audience. Moreover, the inclusion of case studies on sustainable alternatives like Mediterranean snails and regional cheeses highlights the relevance for sustainability experts and food innovation stakeholders. Finally, Greek cultural institutions, tourism bodies, policy makers, and agricultural product promoters may utilize these findings to enhance the digital representation and promotion of Greek culinary heritage and local food products.

Beyond consumer and industry stakeholders, the results can be directly applied to the development of termbanks and lexicons in the agrifood domain. By treating food names and associated concepts as terminological units, our frame-based inventories support the population of structured resources that are compatible with international standards (e.g., TBX, OntoLex-Lemon). This ensures that insights derived

from social media discourse can be reused in terminology management, dictionary-making, and multilingual food vocabularies, expanding the reach of the study to both professional practice and linguistic resource building. Additionally, terminologists and lexicographers will benefit from the semi-automated extraction of neologisms and domain-specific terminology, which supports digital lexicography and terminology development (Panagiotou, Gkatzionis, & Kaloudis, 2025b).

4 Conclusions

The present study explored the use of existing and the development of new LLMs to collect, manage, and analyse linguistic data for specific purposes from a social media platform. ChatGPT was validated as a tool for automated application of criteria as a substitute for human handling to save time and ensure repeatability of results. The use of Python scripts was effective and successful in analyzing data collected from Instagram. Misspelled hashtags in Greek and English, and the use of Greeklish created problems during post collection, hashtag identification, and hashtag analysis. ChatGPT and Llama KriKri were not able to identify Greeklish in a correct and consistent manner, even following attempts to train them by providing lists of Greek words with their Greeklish counterparts as feedback. Areas for improvement in NLP systems still exist, especially regarding the Greek language. The databases created using this methodology can be further populated, by collecting hashtags relevant to other traditional and local foods and in other languages.

The findings of this study highlight how semantic frames related to locality, tradition, and food consumption serve as powerful lenses through which food-related discourse can be understood. These frames offered valuable insight into the cultural preferences of consumers, particularly in relation to food ethnocentrism and regiocentrism. The bottom-up, data-driven methodology employed enabled the automatic population of frames with relevant concepts and vocabulary, demonstrating its effectiveness in capturing the structure and symbolic dimensions of food discourse. Importantly, the study also revealed prototypical associations within the domain of cheese consumption. These findings reinforce the potential of semantic frame analysis as a tool for exploring culturally embedded food meanings and consumer perceptions. The study also demonstrates that frame-based analysis of social media discourse can yield structured outputs that function both as terminological entries, interoperable with standards such as TBX and OntoLex-Lemon, and as lexicographical entries, providing definitions, usage examples, and contextual relations suitable for dictionary compilation.

By situating social media–derived data within structured frames and converting them into entries, the study extends beyond discourse analysis to contribute directly to the fields of terminology and digital lexicography. The approach shows how emerging data sources can be mobilized for the systematic creation of food vocabularies, which are exportable to termbanks, interoperable with existing lexical standards, and reusable in multilingual lexica. In this way, the methodology not only enriches our understanding of food discourse but also strengthens the infrastructure of linguistic resources available to both scholars and practitioners in the agrifood domain.

Acknowledgements

The present study has been funded by the Greek National Development Program 2021-2025 through the General Secretariat for Research and Innovation under the call "Flagship action in sustainable agri-food systems - applied research, development of infrastructure and services for the sustainability of the sector - (Sust.Agri.Food)" [MIS code: 5201774].

Data Availability and Supplementary Files

All data supporting this study are provided as Supplementary Material. This includes the categorized hashtag datasets for snails and cheeses, the comparative cheese file, the top-frequency tables, and the Supplementary file containing Supplementary Tables 1-4. For transparency and reuse, the files are deposited in the project's public repository at https://zenodo.org/records/17478417.

References

Aggarwal, C.C., & Zhai, C. (2012). A survey of text classification algorithms. In C.C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163–222). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6

Apify (2024). *Apify – full-stack web scraping and data extraction platform.* https://apify.com.

Barrett, L.F. (2006, February). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10*(1), 20–46, https://doi.org/10.1207/s15327957pspr1001_2

Declerck, T., Lendvai, P., Mörth, K., Budin, G., Váradi, T. (2012). Towards linked language data for digital humanities. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked data in linguistics: Representing and connecting language data and language metadata* (pp. 109–116). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28249-2_11

Delormier, T., Frohlich, K.L., Potvin, L. (2009). Food and eating as social practice – understanding eating patterns as social phenomena and implications for public health. *Sociology of Health and Illness*, *31*(2), 215–228, https://doi.org/10.1111/j.1467-9566.2008.01128.x

Dunbar, R.I.M. (2017, March). Breaking bread: the functions of social eating. *Adaptive Human Behavior and Physiology*, *3*(3), 198–211, https://doi.org/10.1007/s40750-017-0061-4

Faber, P. (2015). Frames as a framework for terminology. In *Handbook of terminology* (p. 14–33). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.02fra1

Fillmore, C.J. (1977, December). The case for case reopened. In *Grammatical relations* (p. 59–81). BRILL.

Gewiese, & Rau. (2023). *Instagram users in Greece - March 2023 — napoleoncat.com.* https://napoleoncat.com/stats/instagram-users-in-greece/2023/03/. (Accessed 01-07-2025)

Hanna, K., Cross, J., Nicholls, A., Gallegos, D. (2023). The association between loneliness or social isolation and food and eating behaviours: A scoping review. *Appetite*, *191*, 107051, https://doi.org/10.1016/j.appet.2023.107051

Hutchings, S.C., Dixit, Y., Al-Sarayreh, M., Torrico, D.D., Realini, C.E., Jaeger, S.R., Reis, M.M. (2023). A comprehensive guide to digital terminology. *Food Research International*, *165*(2), 112494, https://doi.org/10.1016/j.foodres.2023.112494

International Organization for Standardization (2019). *ISO 30042:2019: Terminology work and terminology management — TermBase eXchange (TBX).* https://www.iso.org/standard/62510.html. (Standard published by the International Organization for Standardization, Geneva, Switzerland)

Katsanevaki, A. (2010). The importance of the community: Its dynamics and its impact on contemporary research (greece as a case-study). *Journal of Ethnography and Folklore/ New Series*, *1-2*, 72-102, Retrieved from http://ikee.lib.auth.gr/record/311873

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017, September). The ontolex-lemon model: Development and applications. I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubicek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. proceedings of elex 2017* (p. 587-597). Lexical Computing CZ s.r.o., Brno, Czech Republic.

Nigam, H., & Biswas, P. (2021). Web scraping: From tools to related legislation and implementation using python. In *Innovative data communication technologies and application* (p. 149–164). Springer Singapore. https://doi.org/10.1007/978-981-15-9651-3_13

Ostermann, C. (2014, jul). Frame semantics and learner's dictionaries: Frame example sections as a new dictionary feature. A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the 16th euralex international congress* (p. 1153-1162). Bolzano, Italy: EURAC research.

Panagiotou, M., Gkatzionis, K., Kaloudis, E. (2025a, June). Compiling linguistic resources for specific purposes: Using LLMs to collect data from social media. F. Vezzani, G.M. Di Nunzio, E. Loupaki, G. Meditskos, & M. Papoutsoglou (Eds.), *4th international conference on multilingual digital terminology today. design, representation formats and management systems* (Vol. 3990). CEUR Workshop Proceedings. Retrieved from https://ceur-ws.org/Vol-3990/short1.pdf

Panagiotou, M., Gkatzionis, K., Kaloudis, E. (2025b, June). Food neologisms and word formation trends identified on social media posts using LLMs for hashtag collection. F. Vezzani, A.O. Anic, A. Salgado, & G. Tallarico (Eds.), *Proceedings of the 1st international workshop on terminological neologism management* (Vol. 3972). CEUR Workshop Proceedings. Retrieved from https://ceur-ws.org/Vol-3972/paper1.pdf

Panagiotou, M., Kaloudis, E., Koukoumaki, D.I., Bountziouka, V., Giannakou, E., Pandi, M., Gkatzionis, K. (2024). Key drivers of consumption, conceptual, sensory, and emotional profiling of cheeses based on origin and consumer familiarity: A case study of local and imported cheeses in Greece. *Gastronomy*, *2*(4), 141–154, https://doi.org/10.3390/gastronomy2040011

Roussis, D., Voukoutis, L., Paraskevopoulos, G., Sofianopoulos, S., Prokopidis, P., Papavasileiou, V., … Katsouros, V. (2025). *Krikri: Advancing open large language models for Greek.* Retrieved from https://arxiv.org/abs/2505.13772

SBS, International Education Services Multicultural NSW (2016). *Greek - Core Concepts — culturalatlas.sbs.com.au.* https://culturalatlas.sbs.com.au/greek-culture/greek-culture-core-concepts. (Accessed 01-07-2025)

Tao, D., Yang, P., Feng, H. (2020, February). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *19*(2), 875–894, https://doi.org/10.1111/1541-4337.12540