Compiling Linguistic Resources for Specific Purposes: Using LLMs to Collect Data From Social Media[1]

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

[1]*Laboratory of Consumer and Sensory Perception of Food & 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, Metropolite Ioakeim 2, Myrina, 81400, Lemnos, Greece*

Abstract

Social media generate vast amounts of authentic linguistic data daily, offering insights into users' conceptualizations, language use, and cultural phenomena. However, managing this data manually is challenging due to its volume and evolving algorithms. This study identifies key concepts and specific words that shape Greek consumer identity by analyzing Instagram posts on traditional and local food consumption. Sentiment analysis was conducted using generative AI, and new computational tools were developed for automated data collection and analysis. AI-generated responses were evaluated against human assessments, and they proved to be successful in identifying and categorizing posts and hashtags according to criteria set for the task. Case studies included Mediterranean snails as a sustainable meat alternative and local cheeses from the North-Aegean islands. The methodology can be applied to specialized lexicography, linguistic and cultural studies, and terminology research, such as identifying neologisms and tracking term usage over time.

Keywords

social media, large language models, generative artificial intelligence, specialized lexicography

1. Introduction

1.1. General background

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 [1]. Examples of social media platforms currently used for language and food related research are Facebook, Instagram, Twitter, and Reddit.

1.2. 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 [1]. 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 [2].

[1] *4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.*

*Corresponding author.

[†] These authors contributed equally.

teniapanag@aegean.gr (M.Panagiotou); kgkatzionis@aegean.gr (K.Gkatzionis); stathiskaloudis@aegean.gr (E.Kaloudis)

0000-0001-8179-1163 (M.Panagiotou); 0000-0002-8904-6448 (K.Gkatzionis); 0000-0001-7602-3282 (E.Kaloudis)

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 [3]. 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 [4] 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) [5]. 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.

1.3. 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 is 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 hashtags has been added to the respective posts.

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 [6]. The workflow for data collection and analysis is presented in Figure 1.

2.1. Data collection, cleaning of duplicates and irrelevant posts, and preparation for analysis

For data collection, Apify [7], a web scraping tool, was used for automatic multi-word search. Posts about snails and the cheeses under study were searched for using hashtags (i.e., words or phrases preceded by the symbol # used to classify the accompanying text) in Greek, English, and 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). 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 the assumption that the account is of Greek origin and, thus, these posts were included in the study. A preliminary search was previously conducted using software specifically developed for this task to identify Greeklish and misspelled forms of the hashtags of interest, as misspelled hashtags are a common phenomenon on social media.

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.

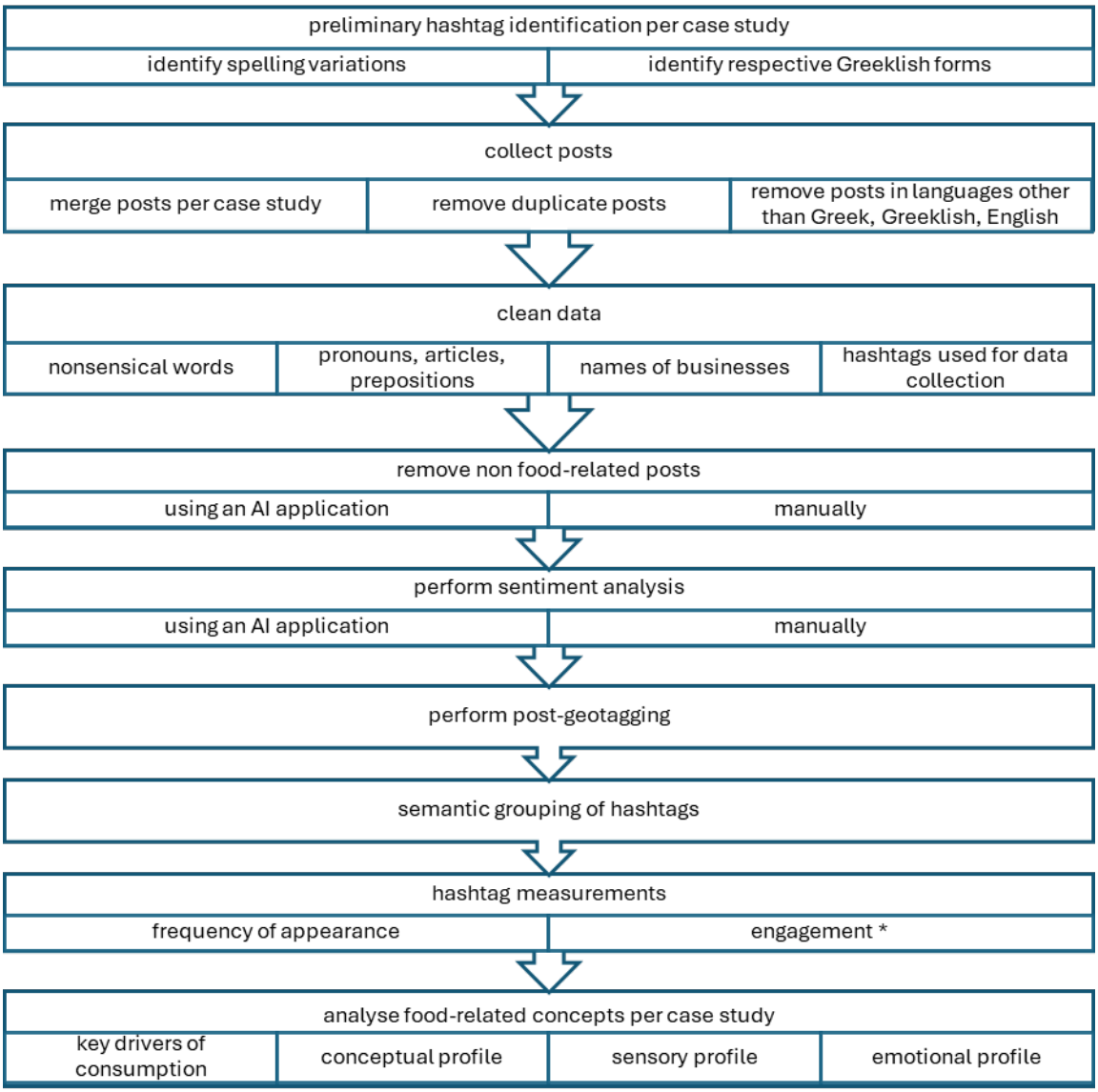

Figure 1: Diagram of workflow for data collection and analysis (*engagement measurement is not presented in this paper as it is outside the scope of the present publication).

2.2. Applying the food-relatedness criterion

ChatGPT, a novel Artificial Intelligence application (edition 3.5), was used for the automated application of the food-relatedness criterion, so that only food-related posts would be left for further analysis. 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". 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 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). 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.

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 AI. However, ChatGPT'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 application consistently identified Greeklish as English. In an attempt to rectify that, it was 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 it was asked to perform the same task the categorization of the same hashtags was different. As a result, the task was performed manually.

3.2. Food-related concepts identified

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 which probably means that traditional and local food consumption closely correlates with sensory satisfaction. 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 twenty semantic categories, using a Greek thesaurus. Hashtags that belonged to more than one category were duplicated (e.g., #tastyrecipes was categorized as post content and assessment). These semantic categories led to drawing conclusions in a more meaningful and organized way. Hashtag grouping into semantic categories was done manually in this case, but GenAI (ChatGPT and Copilot) has been subsequently assessed in food related case studies and was able to group hashtags in a meaningful way. 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 [8].

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[2]), 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 [9,10]. Food, from production to consumption, is a social activity, especially for the Greeks [11–13]. 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.

3.3. 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

[2] **"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.

other languages and platforms. Challenges in data collection from other platforms, such as Reddit, remain to be identified. Certain tools, like ChatGPT, could be further tested using different settings to deliver more accurate and repeatable results, and new similar applications, like the Greek KriKri, could be assessed for the same tasks. 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. An extension of this study, already in progress is the collection and analysis of data from online sources in general (reviews, blogs, magazines, etc.) to investigate the digital identity of agrifood products (sensory attributes, key relevant concepts, consumer attitudes and expectations, etc.).

3.4. Interested parties

On the one hand, food producing and marketing companies could use hashtag analysis, in particular the frequency of appearance and co-appearances in posts, to gain insight into how consumers view their products and brand. For example, the link identified between meat alternatives and fasting or gourmet dining could be used to promote relevant products. On the other hand, linguists, and anyone studying linguistic and cultural phenomena (food-related or not) could also use hashtag frequency of appearance and co-appearances in posts. These hashtags, grouped in semantic categories, form a network of related concept frames that depict cultural aspects of food consumption. These semantic groups can also be used to compile domain-specific glossaries (e.g., for traditional food consumption to be used in travel guides, menus, and consumer studies), or to identify concepts that correlate positively or negatively with a phenomenon under study in (cross-) linguistic and (cross-) cultural studies (e.g., to study consumer ethnocentrism). Theoretical and applied linguists can additionally extract lists of consistently misspelled words to gain some understanding of specific spelling difficulties concerning users of the languages under study. Finally, neologisms can be identified and studied regarding frequency and context of appearance, as well as use over time.

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 was not able to identify Greeklish in a correct and consistent manner, even following attempts to train it 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.

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].

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling checks. The authors have subsequently reviewed and edited the content and take full responsibility for the publication’s final version.

References

[1] S. C. Hutchings, Y. Dixit, M. Al-Sarayreh, D. D. Torrico, C. E. Realini, S. R. Jaeger, M. M. Reis, A critical review of social media research in sensory-consumer science, Food Res. Int. (2023) 112494. doi:10.1016/j.foodres.2023.112494.

[2] H. Nigam, P. Biswas, Web Scraping: From Tools to Related Legislation and Implementation Using Python, in: Innovative Data Communication Technologies and Application, Springer Singapore, Singapore (2021)149–164. doi:10.1007/978-981-15-9651-3_13.

[3] C. C. Aggarwal, C. Zhai, A Survey of Text Classification Algorithms, in: Mining Text Data, Springer US, Boston, MA (2012) 163–222. doi:10.1007/978-1-4614-3223-4_6.

[4] L. F. Barrett, Solving the Emotion Paradox: Categorization and the Experience of Emotion, Personal. Soc. Psychol. Rev. 10.1 (2006) 20–46. doi:10.1207/s15327957pspr1001_2.

[5] D. Tao, P. Yang, H. Feng, Utilization of text mining as a big data analysis tool for food science and nutrition, Compr. Rev. Food Sci. Food Saf. 19.2 (2020) 875–894. doi:10.1111/1541-4337.12540.

[6] J. Gewiese, S. Rau, Instagram Users in Greece, Statista, 2023. URL: https://www.statista.com/study/141743/instagram-users-in-greece/.

[7] Apify. Full-stack web scraping and data extraction platform. URL: https://apify.com/store

[8] M. Panagiotou, E. Kaloudis, D. I. Koukoumaki, V. Bountziouka, E. Giannakou, M. Pandi, K. Gkatzionis, 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, vol. 2 (2024) 141–154 doi:10.3390/gastronomy2040011.

[9] A. Katsanevaki, The Importance of the Community: Its Dynamics and Its Impact on Contemporary Research (Greece as a Case-Study), J. Ethnogr. Folk. New Ser. (2010) 72–102.

[10] SBS, International Education Services, Multicultural NSW, Cultural Atlas - Greek Culture - Core Concepts, 2016. URL: https://culturalatlas.sbs.com.au/greek-culture/greek-culture-core-concepts

[11] T. Delormier, K. L. Frohlich, L. Potvin, Food and Eating as Social Practice – Understanding Eating Patterns as Social Phenomena and Implications for Public Health, Sociology of Health & Illness, 31(2) (2009) 215–228. doi:10.1111/j.1467-9566.2008.01128.x.

[12] R. I. M. Dunbar, Breaking Bread: The Functions of Social Eating, Adaptive Human Behavior and Physiology, 3(3) (2017) 198–211. doi:10.1007/s40750-017-0061-4.

[13] K. Hanna, J. Cross, A. Nicholls, D. Gallegos, The Association Between Loneliness or Social Isolation and Food and Eating Behaviours: A Scoping Review, Appetite, 191 (2023) 107051. doi:10.1016/j.appet.2023.107051.