ARTIFICIAL INTELLIGENCE AND NLP ON REDDIT: UNSUPERVISED DETECTION OF FOOD TRENDS AND HEALTHY EATING PATTERNS

Rocío del Campo-Pedrosa [1] , Diego del Campo-Pedrosa [1] , Bettina Merlin [2] and Ana González-Marcos [1]

[1] Department of Mechanical Engineering, Universidad de La Rioja, Logroño, La Rioja, Spain
[2] Fakultät International Business, Hochschule Heilbronn, Heilbronn, Germany

ABSTRACT

Traditional sensory analysis in food innovation provides limited insight into consumer behavior, whereas social platforms such as Reddit offer large-scale, real-time textual data on food-related practices and perceptions. This study evaluates Reddit as a scalable source for detecting food trends and healthy eating patterns in Spanish-language discussions using artificial intelligence (AI) and natural language processing (NLP). An end-to-end pipeline was implemented, including targeted data scraping across seven food-related domains, Spanish-language filtering (≥70% confidence), customized preprocessing, and unsupervised topic discovery via k-means clustering. The system processed 17,774 Spanish-language posts from an initial corpus of 92,949 entries. Despite linguistic challenges such as polysemy and lemmatization errors, the method produced coherent and representative themes, including barriers to home cooking, weight management concerns, economic factors, food categories, and nutrition-related consultations. These results demonstrate the effectiveness of unsupervised NLP techniques for large-scale monitoring of food-related discourse on social media.

KEYWORDS

Natural Language Processing, Unsupervised Learning, Social Media Mining, Artificial Intelligence

1. INTRODUCTION

The development of new food products plays a critical role in the modern food industry. Nevertheless, a considerable proportion of these products fail once they reach supermarket shelves, leading to significant financial losses and missed business opportunities. Among the main causes of this high failure rate are the limited investment in research and development and the insufficient or inadequate integration of consumer perspectives during the product development stages [1], [2]. In general, the reason for the high failure rate is that the experts decided on the innovative aspects of these products, instead of the consumers themselves [3].

For the last few decades there has been a change in the trend, that is, away from an evaluation approach based on trained sensory panels (or experts) to greater incorporation of consumers' experience. This experience is based on sensory analysis and the collection of opinions. Nonetheless, earlier studies examining how people perceive food quality and make product choices indicate that these are highly complex phenomena [4], [5]. Nevertheless, one must consider that the current information gathering practices require consumers to reflect on their behavior [6]. This can compromise the validity and reliability of the data that is obtained [7],

Bibhu Dash et al: EDUPAN, NLPAI, IOTCB, DBML, ICEEE - 2026
pp. 43-57, 2026. IJCI – 2026 DOI:10.5121/ijci.2026.150204

especially, because most of the decisions that are made daily are made without much effort or deliberation. Instead, they are determined by experiential, affective and intuitive processes [8]. In other words, the request for information by questioning consumers or users can cause them to offer, sometimes unconsciously, socially desirable and excessively rationalized responses [9]. Another factor to consider is that these common practices require that an effective balance be found between the cost and time required to collect information and the accuracy of the estimates that is made, which reduces and limits the sample size.

Therefore, while sensory analysis remains a valuable tool in food innovation for assessing the organoleptic characteristics of products and gauging consumer experience and acceptance, it is essential to integrate additional factors to achieve a more comprehensive and realistic understanding of consumer perceptions, expectations, and attitudes. For this reason, this research work proposes the use of other sources of information. These include what can be obtained from comments on social networks and, more specifically, Reddit [10], since they enable a large amount of data to be obtained for free or at very low cost. Furthermore, the data gathered is not guided by any question, but is provided on the users' own initiative.

The analysis of online comments that have been provided by users and consumers has been explored in different sectors. One of the most noteworthy examples is the tourism industry, where travelers share their opinions on dedicated platforms such as Booking [11], TripAdvisor [12], and Airbnb [13], [14]. These websites are analyzed in order to understand the factors that are most relevant to the creation of satisfactory experiences for users in different types of tourist accommodation.

Within the food sector, several studies have analyzed user comments posted on specialized forums or websites to explore aspects such as customer satisfaction with restaurant experiences [15] or the perceptions of German- and English-speaking consumers regarding organic products [16]. Regarding content from social media, one example includes an analysis of posts and comments from the Twitter and Facebook accounts of three major U.S. pizza chains over the course of a month [17]. This research aimed to assess the influence of social media on customer service and to evaluate how monitoring competitors' online activity can support business decision-making. Another example involved the analysis of Twitter messages by users of a U.S. municipality for a period of five weeks to investigate the relationships between the food choices made by users and their local environment (availability of supermarkets, fresh product stores and fast-food restaurants, etc.) [18].

Specifically, this work seeks to study whether Reddit [10] can be an effective source of information in the food sector. Reddit has emerged as a leading social platform online, boasting 52 million users engaging daily in 2020 and more than 138,000 active thematic communities referred to as "subreddits" [19]. This social network has been used for studying messages and tendencies across different fields. For instance, Reddit has been examined for its role in health information engagement, where users seek and possibly enact health-related information found on the platform [20]. To add, the preferences and outlook of tourists, both pre-pandemic and during the COVID-19 outbreak, were examined through an analysis of over one million Reddit posts from travel-oriented subreddits [21].

With regard to food habits, there is a scarcity of studies conducted on a pure dataset based only on Reddit. For instance, sentiment and emotion analysis of nutrition, food and cooking related content on mixed platforms and social media such as Twitter, YouTube, Instagram, Reddit, Pinterest and Sina Weibo [22]. There are few food-oriented topics analyzed just through a Reddit database in food themes. Among them, the interest in veganism during the COVID-19 pandemic [23] or food safety information-seeking behaviors of young adults on Reddit [24] were studied.

However, none of these investigations examine consumer behavior across the broader food market or explore its evolution over multiple years.

In general, although there are studies that have conducted text analysis in different fields gathering Reddit information, to the best of our knowledge, there is not any research addressing tendencies in the food sector with information gathering solely on Reddit. Consequently, the objective of this study is to evaluate Reddit's potential as a data source for identifying patterns in the food industry and in population nutritional habits.

For this purpose, this work focuses on the analysis of Reddit along different topics related to nutrition and healthy diet, and analyzes them by text mining and natural language processing (NLP) techniques. In addition, it examines discussions spanning nearly eight years to assess the platform's capability for long-term trend detection. Finally, the study outlines the principal advantages and limitations of utilizing Reddit as an information source within this domain.

2. MATERIALS AND METHODS

Text mining can be defined as the process of extracting implicit knowledge from unstructured textual data [25]. This implicit knowledge is not explicitly stored in databases or text repositories; rather, it emerges from analytical processes applied to the data. Therefore, text mining extends beyond basic information retrieval, seeking to identify hidden patterns, structures, and relationships within large bodies of textual information [26].

In this study, a standard text mining pipeline was implemented, comprising the following phases: data acquisition, preprocessing, transformation, analysis, and evaluation (Figure 1). All text processing and analytical procedures were conducted using Python.

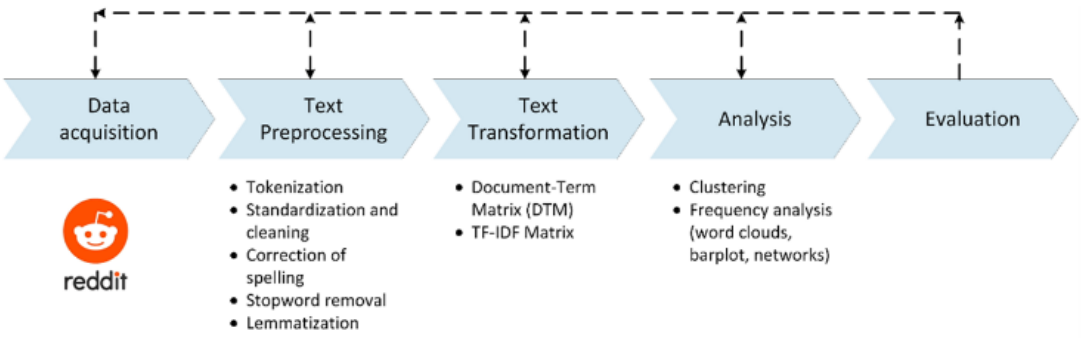

Figure 1. Implemented text mining and NLP pipeline

2.1. Data Acquisition

The effectiveness of text mining outcomes largely depends on the quality and characteristics of the input data. Due to the unstructured and heterogeneous nature of textual information, data acquisition is often more challenging than for structured data sources [27].

The input dataset consisted of Spanish-language texts related to food and eating practices, extracted from the social media platform Reddit. Data were collected using automated web-scraping techniques through Reddit's official API, accessed via the Python Reddit API Wrapper (*PRAW*) library [28]. This approach ensured compliance with Reddit's usage policies, including request rate limitations.

Reddit content can be retrieved either through predefined subreddits (topic-specific communities) or through platform-wide keyword searches. While Spanish-language subreddits such as *r/askspain* or *r/GoingToSpain* offer structured discussions, this study prioritized open keyword-based searches to capture a broader range of users, contexts, and cultural perspectives. This strategy was chosen to maximize data diversity and thematic coverage related to food trends.

Various criteria can be established for keyword searching on Reddit, including relevance, popularity ("hotness"), top-ranking based on positive votes within specified time frames, newest posts, or those generating the highest number of comments. Filtering by "relevance" is particularly useful for finding directly related posts, while sorting by "hotness" aids in locating popular and active discussions. The "top-ranking" option orders results by total positive votes, offering high-quality or historically popular content, and the "new" option displays the most recent posts chronologically. In this study, the "relevance" criterion was selected to ensure the retrieval of the most pertinent and widely discussed information on food-related topics. Additionally, a consensus was established to include only posts containing at least 30 replies, as an initial evaluation of forum activity indicated that such discussions represent a meaningful threshold for engagement and content relevance. Moreover, posts containing images or graphics were excluded, since the accompanying text could depend on visual elements, potentially leading to a loss of contextual meaning in the textual analysis.

Thus, the methodology of searching was settled to a general search across the platform using specific terms chosen to maximize the amount of information accessible from any user or community. This approach led to the establishment of 7 search focuses, identified as key interest areas within food trends: Fast Food, Precooked Food, Healthy Food, Diet, Nutrition/Nutri-Score, Takeaway Food and Supermarkets.

A key challenge identified after data collection was the presence of multilingual content, due to Reddit's global and linguistically diverse user base and the lack of native language filtering in keyword searches. To address this issue, a secondary automated filtering step was applied using Python-based language identification. Only messages classified as Spanish with a minimum confidence of 70% were retained. As a result, from an initial corpus of 92,949 retrieved messages, 17,774 posts (19.1%) were selected for subsequent analysis (Table 1).

Table 1. Volume of messages collected by topic.

Topic	**Number at scrapping**	**Number after Spanish selection**
Fast Food	5,702	980
Precooked Food	599	116
Healthy Food	14,528	1,347
Diet	14,361	372
Nutrition/Nutri-Score	34,359	644
Takeaway Food	22,987	11,823
Supermarkets	7,763	2,492
Total	92,949	17,774

Regarding dataset representativeness, it is acknowledged that social media sampling may introduce voluntary participation biases. To address this, we implemented a targeted relevance-based strategy (rather than pure random sampling) combined with engagement thresholds (≥30 replies), ensuring thematic coherence and interaction quality. Language filtering (≥70% Spanish certainty) further minimized multilingual noise. While demographic comprehensiveness is

limited, the resulting 17,774 posts across 7 diverse topics provide robust coverage for exploratory trend analysis.

2.2. Preprocessing

Preprocessing constitutes one of the most critical stages of the text mining pipeline, as the quality of downstream analyses depends not only on the original data but also on the effectiveness of the transformations applied to reduce noise and standardize textual content. Due to the informal and highly variable nature of social media language, a multi-phase preprocessing procedure was applied after the data collection stage.

The preprocessing steps were as follows:

1. **Tokenization**. In this first step, texts were divided into the selected unit of analysis. In general, word is used as the basic unit.
2. **Standardization and cleaning**. After the texts were segmented, it was essential to standardize and clean them to prevent the same terms (words) from being considered to be different due to their appearance in uppercase or lowercase or with a comma or other punctuation mark, etc. This activity was carried out with tools that were specifically designed in Python for this research and based on natural language processing (NLP) libraries (the developed functions are available at https://github.com/RocioDelCampo/Spanish-Standardization.-NLP-aplication-in-text-mining-preprocessing). In summary, it included several steps:

- Treatment of regular expressions to unify the text. This is accomplished by ensuring that all text is written in lowercase. Also, numbers, emoticons and characters that are not used in Spanish are eliminated. In addition, hashtags and web addresses that appear in the messages are removed. Finally, given the nature of these social networks, abbreviations, which frequently appear, will be changed to their orthographically correct equivalent. For example, *k*, *q*, *ke* or *qe* should be replaced by *que*.
- Removal of infinitive verb endings: numerous verbs in their infinitive form occur alongside direct or indirect objects and occasionally in reflexive constructions. This structure complicates subsequent processes such as spelling correction and stemming. Therefore, all *-lo*, *-la*, *-le*, *-los*, *-las*, *-les*, *-me*, *-te*, *-se*, *-selo*, *-sela*, *-sele* endings are dropped.
- Elimination of punctuation marks, spaces, page breaks and tabs.

3. **Correction of spelling**. Because texts that are extracted from open social networks often contain spelling and grammatical errors, a correction strategy must be designed. This step is the most critical since it guarantees that the value of the meaning of each word will be retained. Therefore, this is the step that requires further elaboration. Specifically, a method that is based on two correctors that work in a coordinated manner has been employed: the Python Hunspell Library [29] and a corrector that is based on Levenshtein's distance through a bottom-up type algorithm with three types of operation: edition, insertion and substitution [30]. The reason for using two correctors is their ability to complement each other. The former perfectly corrects any term that is usually used in common language, whereas the latter has a repertoire of typical and specific vocabulary in the field that this work studies, namely food.
4. **Elimination of empty words or stop words**. In this step, all common and frequent words that have little or no semantic meaning are eliminated. These words (determiners, conjunctions and prepositions) are dropped as they do not add value in subsequent analyses.
5. **Lemmatization**. Generally, preprocessing ends with the reduction of each word to its root (a method that is known as stemming) or with the lemmatization [31]. The latter consists of finding the lemma, a form that is accepted by agreement as a representative of all the inflected forms of the same word (plural, conjugated, feminine, etc.). In this research,

lemmatization is used. Unlike the stemming method, lemmatization assumes that a word can have many meanings depending on the context. This improves the process of grouping words [32]. This process is executed by the use of a Python library called *spaCy* [33].

2.3. Transformation

After preprocessing, the textual dataset was converted into numerical formats appropriate for computational analysis. This conversion allows the implementation of machine learning and statistical techniques by representing text as structured feature vectors [32].

Two conventional representations were employed: the document–term matrix (DTM) and the term frequency–inverse document frequency (TF–IDF). Both representations were generated using the *scikit-learn* Python library [34], which provides dedicated functions for text vectorization. Since the analytical results obtained from both representations were largely comparable, subsequent analyses were based on the TF–IDF weighting scheme.

2.4. Analysis

To address the objective of identifying food-related trends, both thematic and temporal analyses were conducted on the collected messages. For result representation in both analyses, word clouds were used as an exploratory visualization tool. Word clouds display the most relevant terms within a corpus, with word size proportional to frequency. These visualizations were generated using Python libraries including *NLTK* [35], *wordcloud*, and *matplotlib* [36].

The thematic analysis was carried out using clustering methods applied independently to each predefined topic. Clustering represents an unsupervised machine learning approach well-suited for exploratory analysis scenarios in which no prior labeling or predefined group structure is available. These methods allow the identification of text groups that share lexical and semantic similarities, revealing dominant themes and latent structures within the corpus.

Specifically, this study employed the *k-means* clustering algorithm, implemented in the *scikit-learn* Python library [34]. The algorithm iteratively partitions the data into a predefined number of k clusters by minimizing intra-cluster variance. Word clouds were subsequently used to visualize the most representative terms within each cluster, providing an interpretable summary of the resulting thematic groupings.

The choice of *k-means* was motivated by its efficiency, scalability, and interpretability when handling high-dimensional sparse representations such as TF–IDF vectors. In contrast to probabilistic topic models like Latent Dirichlet Allocation (LDA) or Correlated Topic Model (CTM)—which require assumptions about topic distributions and are computationally more demanding—*k-means* offers a straightforward, parameter-light alternative that is particularly effective for large datasets where preliminary thematic groupings are sought. Moreover, our objective was to conduct an exploratory extraction of lexical patterns rather than infer probabilistic topic hierarchies. Therefore, the use of *k-means* allowed rapid and transparent inspection of clusters, providing an empirical foundation for future research where more complex topic modeling (LDA or CTM) can be implemented for validation and comparison purposes.

For the temporal analysis, messages were grouped by year, covering the period from 2016 to 2024 (with data for 2024 limited to January). As a preliminary exploration, yearly word clouds were created to visualize the most recurrent terms within each time frame. To further explore temporal dynamics, TF–IDF matrices were used as the basis for frequency and co-occurrence analyses. From these representations, word relationship graphs were constructed to visualize both

term prominence and associations between terms over time. In these networks, nodes represent terms, with node size proportional to term frequency. Edges represent relationships between terms, with directionality and edge weight indicating the frequency and strength of co-occurrence. Edge thickness and color intensity increase with higher relationship frequency. These network visualizations were generated using the *NetworkX* Python library [37], enabling a structured and interpretable representation of evolving discourse patterns across years.

3. RESULTS

Once the texts were collected, they underwent preprocessing following the procedure described in the corresponding section. Although this stage is particularly time-consuming, each step plays a crucial role in converting the raw data into a clean and standardized format suitable for subsequent analysis. Figure 2 illustrates this process by displaying a word cloud generated from the unprocessed dataset. As shown, the visualization fails to convey meaningful information due to the high level of noise introduced by internet references (e.g., "*http*" or "*www*"), common stop words (e.g., "*que*", "*en*", "*la*", "*de*", "*para*"), and stop words from other languages (e.g., "*ñao*", "*da*", "*mai*", "*in*").

Figure 2. Word cloud of all raw texts

The methodology that was used to process collected messages was advantageous as it automatically permits a large number of texts to be adapted for later analysis. However, the stemming step is not perfect. The libraries that were used consider that the lemma of some nouns and adjectives correspond to an infinitive form, as their roots coincide with that verb form. In fact, the library sometimes has difficulty discerning when to use the most appropriate grammatical category. For example, in Figure 4, we observe the words "*paso*" and "*pasar*", as "*paso*" in Spanish can mean either the noun "step" or the first-person singular present simple form of the verb "*pasar*". Similarly, in Figure 5, we notice that mostly the word "*grasa*" has been lemmatized as "*grasa*", but a few instances appear in their infinitive form "*grasar*." Another clear example is seen in terms of health and healthy, for instance, in Figures 3 and 5 it appears as "*saludable*" what is the pure adjective "*heatlhy*", although in Figure 6 it is written as an infinitive "*saludar*" with literally means "to greet".

3.1. Thematic analysis

The analysis was first conducted separately for each topic—Fast Food, Precooked Food, Healthy Food, Diet, Nutrition/Nutri-Score, Takeaway Food, and Supermarkets—using *k-means* clustering to identify the dominant themes and concerns within each category.

For the "Fast Food" topic, clustering revealed three main areas of discussion:

- Health-related concerns.
- Frequent references to major fast-food chains, particularly McDonald's, Burger King, and pizza.
- Low wages and economic constraints.

Within "Precooked Food", the identified clusters focused on:

- Types of food, including rice, eggs, meat, sandwiches, ham, cheese, and chorizo.
- Time constraints associated with cooking and the convenience of precooked food, especially in work-related contexts.
- Cross-country comparisons, mainly involving Spain, Italy, and England.

The "Healthy Food" topic exhibited a broader thematic diversity, with clusters centered on:

- Weight loss and nutritional concerns related to sugar, calorie, and fat intake, as well as strategies such as consulting nutritionists or practicing intermittent fasting.
- The relationship between fast or low-cost eating and weight gain or calorie consumption.
- Government taxation policies affecting food products.
- Discussions linking salary, working conditions, and quality of life, particularly in Chile, Colombia, and Mexico

For the "Diet topic", the clustering highlighted:

- References to politicians' salaries and allowances (diets).
- Dietary strategies aimed at reducing fat and calories.
- Muscle gain.
- The role of nutritionists and physical exercise.
- Weight loss approaches involving diet and tea consumption.

In the "Nutrition/Nutri-Score" topic, the main clusters reflected:

- Concerns about weight, calorie intake, and overall health.
- Associations between sugar, fat, and health outcomes.
- References to nutritionists in the context of both weight loss and weight gain.
- Eating in moderation and the use of quantitative or mathematical approaches to portion control.
- Concerns about excessive intake of specific nutrients, including carbohydrates, fat, sugar, sodium, and protein.

Analysis of “Takeaway Food” revealed several distinct themes:

- Economic concerns related to spending money and tipping practices.
- Poverty-related issues, including asking for money and difficulties in paying.
- Cultural food references, particularly Argentine cuisine and Chinese food.
- Comparisons with homemade food, emphasizing time constraints, perceived difficulty, and trade-offs between cost and health.
- Mentions of food types such as meat, rice, milk, eggs, chicken, and vegetables.
- Protein scarcity in connection with barbecuing practices.

Finally, messages related to “Supermarkets” clustered around:

- Food categories, including meat, fish, protein sources, vegetables, and legumes.
- Trade-offs between quality, price, and variety.
- Comparisons between supermarket chains such as Alcampo, Mercadona, Carrefour, Aldi, and Lidl.
- References to new products and cashback offers.

Figure-based examples illustrate these thematic structures. In the “Healthy Food” topic, weight loss emerges as a dominant theme, characterized by terms such as eat, calorie, lose, weight, intermittent, fasting, healthy, nutritionist, and exercise (*comer*, *caloría*, *bajar*, *peso*, *ayuno*, *intermitente*, *saludable*, *nutricionista*, *ejercicio*) (Figure 3). In “Takeaway Food”, comparisons with homemade food are reflected through terms such as time, buy, cook, food, money, think, difficult, cheap, and healthy (*tiempo*, *comprar*, *cocinar*, *comida*, *dinero*, *pensar*, *difícil*, *barato*, *sano*) (Figure 4). Finally, in “Nutrition/Nutri-Score”, concerns about excessive nutrient consumption are highlighted by frequent mentions of food, excess, carbohydrate, fat, sugar, sodium, and protein (*alimento*, *exceso*, *carbohidrato*, *grasa*, *azúcar*, *sodio*, *proteína*) (Figure 5).

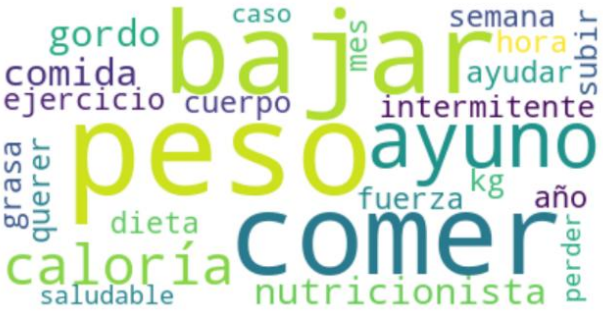

Figure 3. Concern about losing weight in “Healthy Food” topic

Figure 4. Concern about homemade cooking difficulty in “Takeaway Food” topic

Figure 5. Concern about excess consumption of certain nutrients in "Nutrition/Nutri-Score" topic

3.2. Temporal analysis

After preprocessing, all messages were grouped by year to examine temporal trends and variations in food-related discourse. Table 2 summarizes the annual volume of collected messages between 2016 and January 2024, showing a substantial increase in data availability from 2021 onwards.

As an initial exploratory step, yearly word clouds were generated to identify the most frequently mentioned terms. To improve interpretability, highly generic words unrelated to specific content (e.g., cook, eat, want, people) were excluded. This analysis revealed distinct thematic emphases across years. In 2016, discussions were dominated by references to civil insecurity and socioeconomic issues, with frequent terms such as economy, rights, criminal, violence, and political (*economía*, *derecho*, *criminal*, *violence*, *politico*). In 2017, references were mainly limited to specific food items, particularly pizza and brownies. In 2018 and 2019, discourse shifted toward diet- and weight-related topics, with frequent mentions of calories, carbohydrates, sugar, fat, and diet (*calorías*, *carbohidratos*, *azúcar*, *grasa*, *dieta*). In 2020, Argentina and Urugay were the most commented words. In 2020, country-related terms such as Argentina and Uruguay became prominent. From 2021 to 2024, no single topic clearly dominated according to word frequency alone.

Table 2. Volume of messages collected by year.

Year	Number messages
2016	71
2017	38
2018	565
2019	568
2020	777
2021	1,615
2022	4,528
2023	7,528

To extract more informative patterns, we complemented frequency-based analysis with network analysis, modeling co-occurrence relationships between frequently mentioned terms. This approach captures not only term prevalence but also the strength of associations between concepts, enabling a more nuanced view of how discourse evolves over time. The most relevant word relationships identified for each year were:

- 2016: Free-market (*Librar-mercar*), earn-money (*ganar dinero*) and violence-arthritis (*violencia-artritismo*).
- 2017: no stable or interpretable relationships were identified.
- 2018: low-wage (*bajar-soldar*, *soldar* as infinitive for *sueldo*), want-stop-eating (*querer-dejar-comer*), eat-home (*comer-casa*), eat-hamburger (*comer-hamburguesa*), diet-week (*dieta-semana*) and raise-wage (*subir-soldar*, *soldar* as infinitive for *sueldo*).
- 2019: eat-health (*comer-saludar*), carbohydrate-fat (*carbohidrato-graso*), carbohydrate-sugar (*carbohidrato-azúcar*), obesity-people-problem (*problema-gente-gorda*) and high-consumption (*alto-consumo*).
- 2020: lose-weight (*bajar-kilogramo*) and eat-quantity (*comer-cantidad*).
- 2021: person-health (*persona-salud*), leave-tip (*dejar-propina*), deny-pay (*privar-pagar*) and pay-pass (*pasar-pagar*).
- 2022: eat-fast (*comer-rápido*), eat-health (*comer-saludar*) and eat-restaurant (*comer-restaurante*) (see Figure 6).
- 2023: ask-money (*pedir-dinero*), buy-eat (*comprar-comer*), want-money-eat (*querer-dinero-comer*) and leave-tip (*dejar-propina*) (see Figure 7).
- 2024: workhard-twelve (*pringar-doce*) and mercadona-lidl-aldi (*mercar-lidl-aldi*).

Overall, the network-based analysis provides a richer temporal characterization of food-related discussions than word frequency alone, highlighting shifts from nutrition- and diet-centered discourse toward economic and consumption-related concerns in later years.

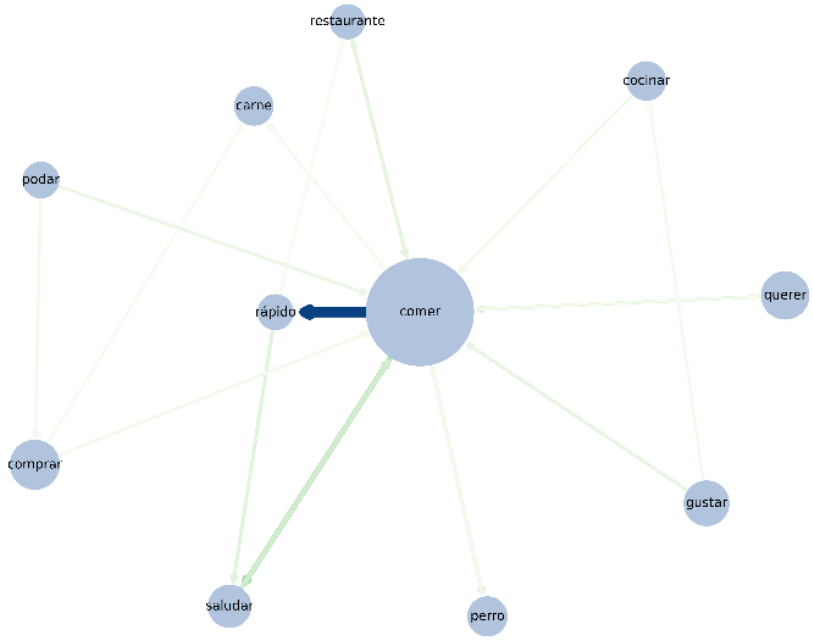

Figure 6. Network diagram for 2022

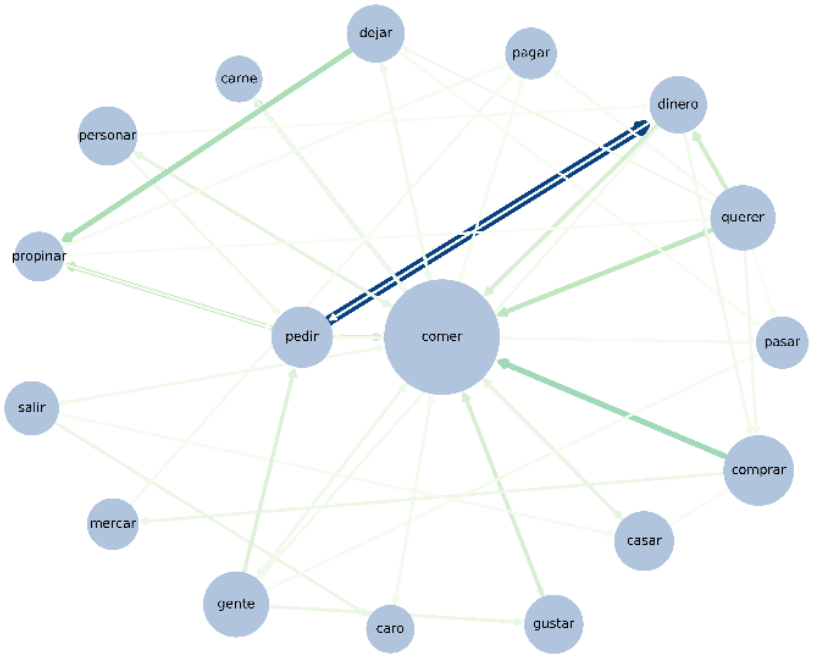

Figure 7. Network diagram for 2023

CONCLUSIONS

This study demonstrates the feasibility and practical value of applying a standard text mining and NLP pipeline, based on unsupervised artificial intelligence (AI) methods, to the analysis of food-related discourse on Reddit. Despite the heterogeneous, informal, and noisy nature of social media data, the results show that Reddit can constitute a meaningful source for exploring eating-related concerns and trends, provided that data collection strategies and language filtering are carefully designed to ensure relevance and interpretability.

A principal contribution of this research lies in its ability to autonomously collect and process large volumes of user-generated text, enabling a scalable and time-efficient alternative to manual qualitative analysis. The combination of automated extraction, statistical preprocessing, numerical representation (TF–IDF), and unsupervised clustering (*k-means*) transforms unstructured data into coherent and interpretable thematic structures. The resulting clusters captured recurrent themes such as the challenges of home cooking, weight management and dietary concerns, economic aspects of food consumption, discussion of various food categories, and the growing importance of professional nutritional advice. These findings highlight the utility of unsupervised AI techniques in domains where labeled data are limited and the primary goal is exploratory pattern discovery rather than prediction.

Beyond methodological contributions, the results offer actionable insights for the food industry and public policy. Increasing discussions about weight management, calorie reduction, and nutritionist guidance reflect heightened consumer interest in healthier and transparently labeled products. Food manufacturers and retailers can leverage this information to reformulate products, improve labeling practices, and enhance communication with health-conscious consumers. Likewise, recurring themes around affordability, wages, and tipping practices underscore economic sensitivity in food choices, suggesting that companies may benefit from adopting value-oriented marketing and adaptive pricing strategies. Persistent contrasts between homemade

and fast or takeaway food also point to shifting work–life dynamics and open opportunities for innovation in convenient yet nutritious meal solutions.

At a policy level, the prevalence of conversations about obesity, sugar and fat intake, and nutritional inequalities emphasizes the potential of social media data to complement traditional surveys. Integrating insights from platforms like Reddit into public health monitoring systems could enable more adaptive, real-time strategies for nutrition policy development.

Several limitations must be acknowledged. Although engagement-based filtering helped reduce sampling biases, Reddit's voluntary and platform-specific participation may still underrepresent certain demographic groups, limiting generalizability. Future studies could address this through stratified sampling or demographic weighting when auxiliary data are accessible. Additionally, the platform's multilingual environment, lexical overlap among languages, and polysemy introduced residual noise that may affect downstream NLP performance. These challenges reveal the limitations of word-level text representations and highlight the potential of context-aware approaches—such as multilingual or transformer-based embeddings (e.g., BERT, mBERT, or XLM-R)—to capture semantic context more effectively. Incorporating such models, together with alternative unsupervised topic modeling frameworks (e.g., LDA or CTM), could substantially enhance thematic coherence, cross-lingual consistency, and semantic precision in future analyses.

Overall, this work provides a strong foundation for subsequent AI-driven research into food discourse on digital platforms. Improving multilingual preprocessing pipelines, integrating context-sensitive NLP models, and expanding the analysis across cultural and linguistic settings represent promising paths for refining both the analytical robustness and the real-world applicability of AI techniques in food trend detection.

ACKNOWLEDGEMENTS

All authors appreciate the support from Elizabeth Marie Lavadia for her external grammatical review of the manuscript.

REFERENCES

[1] K. G. Grunert et al., "Consumer-oriented innovation in the food and personal care products sectors: Understanding consumers and using their insights in the innovation process," in Consumer-Driven Innovation in Food and Personal Care Products, 2010. doi: 10.1533/9781845699970.1.3.

[2] S. E. Kemp, "Consumers as part of food and beverage industry innovation," in Open Innovation in the Food and Beverage Industry, 2013. doi: 10.1533/9780857097248.2.109.

[3] N. V. Olsen, "Design Thinking and food innovation," Trends in Food Science and Technology, vol. 41, no. 2. 2015. doi: 10.1016/j.tifs.2014.10.001.

[4] E. P. Köster, "Diversity in the determinants of food choice: A psychological perspective," Food Qual Prefer, vol. 20, no. 2, 2009, doi: 10.1016/j.foodqual.2007.11.002.

[5] D. Asioli, P. Varela, M. Hersleth, V. L. Almli, N. V. Olsen, and T. Næs, "A discussion of recent methodologies for combining sensory and extrinsic product properties in consumer studies," Food Qual Prefer, vol. 56, 2017, doi: 10.1016/j.foodqual.2016.03.015.

[6] R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing, vol. 27, no. 4, 2010, doi: 10.1016/j.ijresmar.2010.09.001.

[7] E. P. Köster, "The psychology of food choice: Some often encountered fallacies," Food Qual Prefer, vol. 14, no. 5–6, 2003, doi: 10.1016/S0950-3293(03)00017-X.

[8] D. Kahneman, "Maps of bounded rationality: Psychology for behavioral economics," American Economic Review, vol. 93, no. 5. 2003. doi: 10.1257/000282803322655392.

[9] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," Journal of Applied Psychology, vol. 88, no. 5, pp. 879–903, 2003, doi: 10.1037/0021-9010.88.5.879.
[10] "Reddit." Accessed: Feb. 25, 2024. [Online]. Available: Specifically, this work seeks to study whether Reddit can be an effective source of information in the food sector.
[11] Xu and Y. Li, "The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach," Int J Hosp Manag, vol. 55, 2016, doi: 10.1016/j.ijhm.2016.03.003.
[12] F. Hu and R. H. Trivedi, "Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content," Int J Hosp Manag, vol. 84, 2020, doi: 10.1016/j.ijhm.2019.102317.
[13] H. Villeneuve and W. O'Brien, "Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality," Build Environ, vol. 169, p. 106555, Feb. 2020, doi: 10.1016/J.BUILDENV.2019.106555.
[14] K. Kiatkawsin, I. Sutherland, and J. Y. Kim, "A comparative automated text analysis of airbnb reviews in Hong Kong and Singapore using latent dirichlet allocation," Sustainability (Switzerland), vol. 12, no. 16, 2020, doi: 10.3390/su12166673.
[15] S. (Sixue) Jia, "Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews," Tour Manag, vol. 78, Jun. 2020, doi: 10.1016/j.tourman.2019.104071.
[16] H. Danner and L. Menapace, "Using online comments to explore consumer beliefs regarding organic food in German-speaking countries and the United States," Food Qual Prefer, vol. 83, Jul. 2020, doi: 10.1016/j.foodqual.2020.103912.
[17] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," Int J Inf Manage, vol. 33, no. 3, 2013, doi: 10.1016/j.ijinfomgt.2013.01.001.
[18] X. Chen and X. Yang, "Does food environment influence food choices? A geographical analysis through 'tweets,'" Applied Geography, vol. 51, 2014, doi: 10.1016/j.apgeog.2014.04.003.
[19] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics," Social Media and Society, vol. 7, no. 2, 2021, doi: 10.1177/20563051211019004.
[20] R. A. Record, W. R. Silberman, J. E. Santiago, and T. Ham, "I Sought It, I Reddit: Examining Health Information Engagement Behaviors among Reddit Users," J Health Commun, vol. 23, no. 5, 2018, doi: 10.1080/10810730.2018.1465493.
[21] D. Hardt and F. K. Glückstad, "A social media analysis of travel preferences and attitudes, before and during Covid-19," Tour Manag, vol. 100, 2024, doi: 10.1016/j.tourman.2023.104821.
[22] A. Molenaar, E. L. Jenkins, L. Brennan, D. Lukose, and T. A. McCaffrey, "The use of sentiment and emotion analysis and data science to assess the language of nutrition, food and cooking related content on social media: A systematic scoping review," Nutrition Research Reviews. 2023. doi: 10.1017/S0954422423000069.
[23] E. Park and S. B. Kim, "Veganism during the COVID-19 pandemic: Vegans' and nonvegans' perspectives," Appetite, vol. 175, 2022, doi: 10.1016/j.appet.2022.106082.
[24] M. Espedido and I. Young, "I read it on reddit: Food safety information-seeking preferences and practices of young adults online," Food Prot Trends, vol. 41, no. 2, 2021, doi: 10.4315/1541-9576-41.2.204.
[25] R. Feldman and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, vol. 44, no. 10. New York: Cambridge, 2007. doi: 10.5860/choice.44-5684.
[26] T. Jo, Text Mining Concepts, Implementation, and Big Data Challenge. Switzerland: Springer, 2019. doi: 10.1007/978-3-319-91815-0.
[27] C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. San Rafael: Morgan & Claypool, 2016. doi: 10.1145/2915031.
[28] "PRAW: The Python Reddit API Wrapper." Accessed: Feb. 25, 2024. [Online]. Available: https://praw.readthedocs.io/en/stable/index.html
[29] M. Wu, "Spacy_hunspell." [Online]. Available: https://github.com/tokestermw/spacy_hunspell
[30] P. Norvig, "How to Write a Spelling Corrector." [Online]. Available: http://www.norvig.com/spell-correct.html
[31] S. Bird, E. Klein, and E. Loper, "Natural language processing with Python: [analyzing text with the natural language toolkit]." [Online]. Available: http://www.nltk.org/book/

[32] M. Anandarajan, C. Hill, and T. Nolan, Practical text analytics. Maximizing the value of text data. Switzerland: Springer, 2019.
[33] Montani et al., “spaCy version 3.7.2.” Accessed: Feb. 25, 2024. [Online]. Available: https://github.com/explosion/spaCy
[34] Grisel et al., “scikit-learn,” 2024, Accessed: Feb. 26, 2024. [Online]. Available: https://github.com/scikit-learn/scikit-learn
[35] Bird et al., “Natural Language Toolkit (NLTK) - version 3.8.1.,” 2024, Accessed: Feb. 26, 2024. [Online]. Available: https://github.com/nltk/nltk
[36] The Matplotlib development team, “Matplotlib v3.8,” 2024, Accessed: Feb. 26, 2024. [Online]. Available: https://matplotlib.org/
[37] Hagberg et al., “Networkx.” Accessed: Feb. 26, 2024. [Online]. Available: https://github.com/networkx/networkx