

Smart Cities in Stars: Food Perceptions and Beyond

Maria Pontiki^(✉), Panagiota Koltsida, Dimitris Gkoumas,
Dimitris Pappas, Haris Papageorgiou, Eleni Toli,
and Yannis Ioannidis

Athena Research and Innovation Center in Information,
Communication and Knowledge Technologies,
Artemidos 6 & Epidavrou, 15125 Maroussi, Athens, Greece
{mpontiki, dgkoumas, dpappas,
xaris, yannis}@athena-innovation.gr,
{p.koltsida, elto}@di.uoa.gr

Abstract. Citizens are shaping their food preferences and expressing their food experiences on a daily basis reflecting their way of living, culture and well-being. In this paper, we focus on food perceptions and experiences in the context of smart citizen and tourist sensing. We analyze Foursquare user reviews about food-related points of interest in ten European cities, and we explore the imprint of a city as it is shaped based on the spatial distribution of food-related topics and sentiments. The topic modelling and sentiment analysis results are visualized using geo-referenced heat maps that enrich the cities maps with information that allows for a more insightful navigation across their different geographical regions providing insights not available in the original data.

Keywords: Urban sensing · Food experiences · Topic modelling · Sentiment analysis

1 Introduction

The emergence of social networks, content publishing platforms, check-in applications and smartphones/GPS devices in recent years has created huge amounts of user-related data. To exploit these rapidly growing data, recent research has focused on the use of the geographic side of information to trace events [1], analyze the sentiment of users [2], identify popular Points of Interest (POIs) [3], identify and visualize typical day-to-day traffic patterns [4], as well as to improve existing city maps. In this context, the growing need for cost-effective location-based services and effective online advertising has led leading online service providers (such as Google, Bing, Foursquare) to store and distribute information about POIs to users of these services, usually through the use of web-based programming interfaces. Users not only have access to information about nearby POIs, but they also have the ability to read and provide reviews or to inform their friends about their current location.

Foursquare data have been explored in the context of location based social networks and urban sensing in different ways to explore research challenges such as

identifying the different factors which drive check-in behavior by clustering users into meaningful groups using topic modelling techniques [5], investigating the impact of tips on user mobility using sentiment analysis techniques [6], or building location recommendation models by combining the preference extracted from check-ins and text-based tips which are processed using sentiment analysis [7].

The work presented in this paper aims at contributing to smart citizen and tourist sensing with a focus on food perceptions and further insights that can be derived from user reviews. To this end, we have been collecting and analyzing user reviews about food-related POIs in ten European cities, and we explore the imprint of each city as it is shaped based on the spatial distribution of food-related topics and sentiments discussed by food consumers. The results are visualized using geo-referenced heat maps enriching the cities maps enabling answers to questions like *“Burger food options in downtown Athens”* or *“Good Breakfast options close to Louvre Museum in Paris”*. The comparative analysis of the results for the ten cities can provide insights like *“In which European city the service is mostly evaluated as friendly?”*.

Moreover, capturing and analyzing information about people’s food choices and eating behavior can help consumers, food providers, and policy makers in making informed decisions on their activities (e.g. food recommenders, targeted nutrition education programming). The work presented in this paper constitutes a part of an ongoing work towards this direction and is developed through the CAPSELLA Social Data Platform. The CAPSELLA¹ project aims to support communities of farmers and food manufacturers in making informed decisions on their activities by offering them access to data from a variety of open data sources related to regional agro-biodiversity and food. A core activity towards this goal is the development of Social Data services based on the needs and interests of a community, depending on their respective roles (i.e. consumer, producer, etc.). A brief overview of the CAPSELLA infrastructure and Social Analytics Platform is provided in Sect. 2. Section 3 presents the topic modelling and sentiment analysis methodology. The results are presented in Sect. 4. The paper concludes with a discussion on the results and further extensions of our work.

2 The CAPSELLA Social Analytics Platform

The CAPSELLA infrastructure (Fig. 1) is the base layer of the services and applications developed in the context of the CAPSELLA project and is designed to meet various requirements coming from different communities, user profiles and to support distinct needs. It consists of several independent, but highly inter-connected systems and offers a set of functionalities covering data and metadata management and data analytics, thus supporting the complete cycle of a data infrastructure. In addition, a number of interoperable services have been developed to make the infrastructure able to interact and exchange datasets and information by exploiting, among others, well known standards. The core systems of the infrastructure are: (a) the Data Management System, offering storage, retrieval and management facilities for various data types

¹ <http://www.capsella.eu/>.

including non-relational, tabular, relational and geospatial data; (b) the Metadata Catalogue System, which offers browsing and discovery of the datasets based on a set of descriptive metadata, and (c) the Data Analytics System (Social Data Platform) that lies next to the above systems and supports data analysis and processing, as well as decision making for its client applications. Furthermore, a central horizontal service has been developed so as to manage the authentication and authorization aspects of the infrastructure concerning both the users and the datasets. Extensibility, scalability and performance are the three pillars, on which the design and the implementation of the above systems are settled. In this direction, all the systems and services can be easily extended to meet additional requirements that were not considered in the initial design.

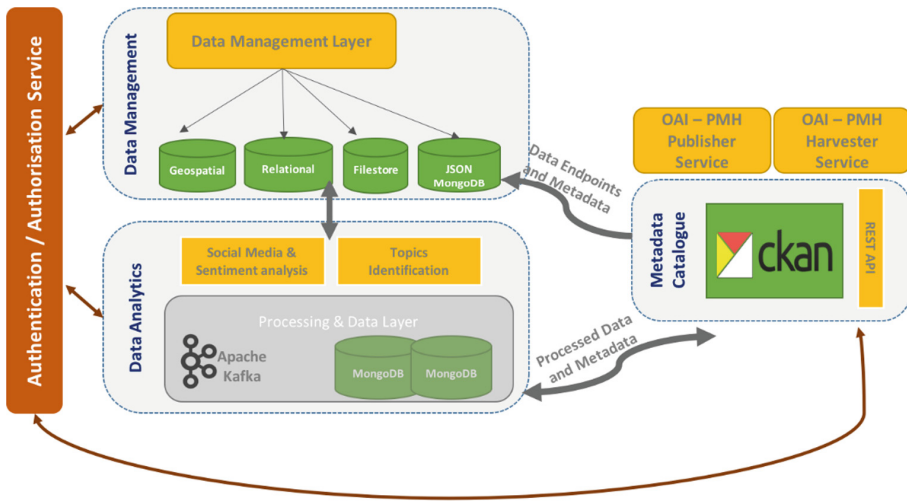


Fig. 1. The CAPSELLA infrastructure.

The CAPSELLA Social Data Platform consists of five major tiers. The Data collection tier where datasets are being captured or ingested from a variety of sources through a set of dedicated tools. Transporting data from the Collection tier to the CAPSELLA platform is facilitated by a message queuing tier (MQ tier) following the Publish/subscribe interaction pattern. In the CAPSELLA MQ tier, social analytics workflows (processing tier) play the role of the receivers enabling the efficient and effective processing of the data. In addition, data can be distributed within the platform to provide additional protections against failures, as well as significant opportunities for scaling performance. The results (e.g. annotations) are transmitted as new messages and further consumed by (a) a distributed indexing & visualization framework providing search and browsing functionalities (Data Exploration tier), (b) archiving to a storage system, and (c) a persistent storage subsystem providing REST APIs to web apps, third applications and the CAPSELLA pilot projects (access tier).

The social analytics workflows integrate several natural language processing and text analytics modules and pipelines that explore our daily food experiences, diversity in food behavior and culture, etc. and produce a wealth of annotations (e.g. sentiments, topics,

entities like nutrients, food ingredients, restaurants etc.) enriching the original data collections. The workflows are designed based on the specifications and the requirements of each pilot in order to provide the right type of information to the right community of people. An example of such a workflow is the Stevia² Application that performs sentiment analysis on Twitter data focusing on specific stevia aspects (e.g. taste, nutrition value, healthiness and price). In this paper, we present the topic modelling and sentiment analysis workflows developed in the context of smart citizen and tourist sensing by focusing on food perceptions and further insights that can be derived from user reviews.

3 Methodology

In order to explore the imprint of a city as it is shaped based on experiences expressed by food consumers, we collect and analyze user reviews about food-related POIs in ten European cities. The research hypothesis is that the spatial distribution of food-related topics and sentiments discussed by food consumers can provide to citizens and tourists meaningful information not available in the original data. In this context, we employ topic modelling, an unsupervised machine learning method, to detect topics of interest. In a second phase, we perform sentiment analysis to capture sentiments expressed towards the extracted topics for each city.

3.1 Data Collection

We collected from Foursquare³ customer reviews (written in English) about food-related POIs (i.e. POIs that belong to the food category) using the coordinates⁴ of each city of interest. Overall, we created 10 collections (one for each city) as illustrated below in Table 1.

Table 1. Amount of data collected from Foursquare for each city.

City/country	Comments	POIs
Amsterdam/Netherlands	53.767	3.851
Athens/Greece	39.359	5.432
Barcelona/Spain	44.114	6.595
Berlin/Germany	29.386	6.998
Brussels/Belgium	14.872	5.428
Lisbon/Portugal	12.295	3.519
London/United Kingdom	158.284	16.888
Paris/France	32.382	17.250
Prague/Czech Republic	49.827	4.600
Rome/Italy	14.984	5.882
Total	449.270	76.443

² <http://stevianet.gr/dashboard>.

³ <https://foursquare.com/>.

⁴ We used a specific bounding box for each city.

3.2 Topic Modelling

Topic models provide an effective way to obtain insights in large collections of unlabeled data and have been widely used for inferring low-dimensional representations that uncover the latent semantic structures of textual [8], image [9], or audio [10] data. A topic is a probability distribution over words, where distribution implies semantic coherence. The state of the art topic model method is Latent Dirichlet Allocation (LDA) [11] and its derivatives, whilst the recently proposed algorithms for Non Negative Matrix Factorization (NMF) perform also well for document clustering and topic modeling [12, 13].

We applied both LDA and NMF methods with a variable number of topics for each data collection. After many iterations and human evaluation, we decided to extract 50 topics for each collection. The results indicate that the NMF topics were of better quality in terms of topic covering and coherence, and by far faster to compute than the corresponding LDA ones. This is probably due to the tf-idf weighting scheme that the NMF is using. The input for the topic models was preprocessed data. In particular, all comments were lowercased and we removed the stopwords. Next we applied a part of speech tagger and after running some experiments we decided to keep only the nouns and the adjectives. Then, we extracted bigrams from each text targeting to obtain topics with phrase like keywords and not only single terms. For the NMF method we used tf-idf weighting scheme with l2 normalization to construct the terms – documents matrix. The output of the method is clusters of words indicating a topic (e.g. carbonara, amatriciana, ravioli, spaghetti, mimosa). No taxonomies or ontologies of topics were used to assist topic modelling. At a final step, a human evaluator inspected the output and assigned a descriptive label to the ten top topics for each city (e.g. PASTA for carbonara, amatriciana, etc. in the above example).

3.3 Sentiment Analysis

Sentiment Analysis can contribute to a better understanding of public opinions, emotions, needs and concerns. Hence, it is a key data analytics tool in the context of urban sensing and citizen behavior analysis used not only for textual data like Twitter [2, 14], but also for visual data shared on social networks (e.g. Pinterest) [15]. Sentiment Analysis solutions range from general purpose algorithms to more fine-grained approaches like aspect-based [16, 17] or topic-based [18] sentiment analysis depending on the case. State of the art techniques range from the traditional lexicon-based [19] to the current trend of deep neural network approaches [20].

We employ a neural network approach to build a model that classifies user reviews according to their positive/negative sentiment orientation. To train our model we used the Fine Foods Amazon Reviews Dataset [21] that consists of user reviews about Amazon food products. Each review is accompanied by a 1–5 score given by the users. Reviews with score 1 and 2 were labeled as Negative. For the Positive class we only use the 5-star reviews in order to create a more balanced training dataset, since the amount of the positive and negative reviews was disproportional in the original dataset. We then built a model that predicts whether a review is positive or negative. Our model uses pretrained word-embeddings; an embedding layer projects each token into an

embedding space and each token is transformed into a vector. The vectors are initialized by pretrained GloVe vectors [22], but they can also change during the training process. Then a convolutional layer consumes the vectors and learns filters that read through the bigrams and trigrams of the text. We used a LSTM layer that reads through the representations of the bigrams and trigrams produced by the convolutional layer. In order to better match the vocabulary of the pretrained embeddings, the review texts were tokenized. The input for the model was collections of topic-specific user comments derived by the topic modelling analysis for each city. Each comment is classified as positive or negative with a confidence score between 0 and 1. If a comment is associated to more than one topics, the same sentiment value is assigned to all topics.

4 Results

4.1 Food Perceptions: Food-Related Topics and Sentiments

The topic analysis output is visualized using heat maps that provide a comprehensive overview of which geographical regions of a city are associated with each topic. For example, Fig. 2 portrays the spatial distribution of the topic “BURGER FOOD” in downtown Athens according to the spatial distribution of the Foursquare user-comments about food-related POIs. More intense color indicates areas with higher density (i.e. more comments for the particular topic in a specific area). The restaurant map icon (🍔) is used only for the POIs that are indexed as burger joints in Foursquare.

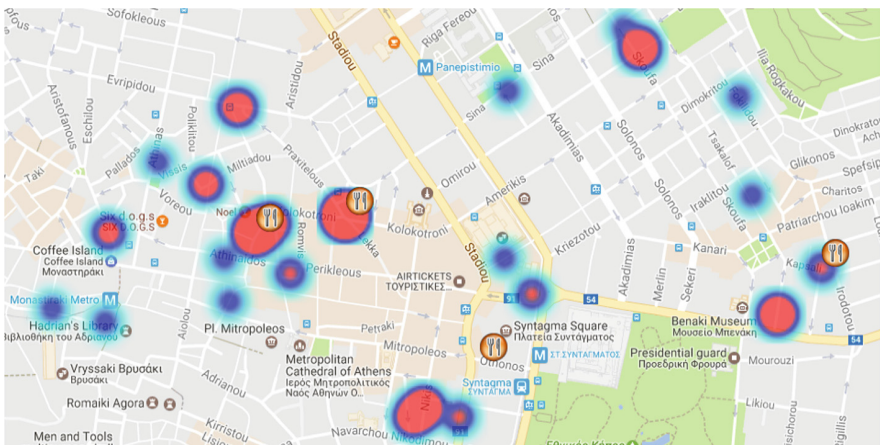


Fig. 2. Heatmap for the topic “BURGER FOOD” in Athens and POIs tagged as burger joints (Color figure online).

Given that the topics reflect the food consumers’ and not the food providers’ standpoint (i.e. owners of food-related POIs) they provide additional information about the food imprint of a city with regard to a specific topic. In other words, comments about burger food are not limited to burger joints. As it is illustrated in Fig. 2, there are

much more burger food options in downtown Athens, indicating at the same time the mostly discussed, and by extension highly visited, ones. Integrating the sentiment analysis output (see Fig. 3) enables also answering to queries like “*Where can someone find good burgers/breakfast/pizza etc. in Athens/Rome/Paris etc.*”, where “*good*” indicates positive user-comments about each topic.



Fig. 3. Heatmap of positive comments for the topic “BREAKFAST” close to Louvre Museum in Paris

The comparative analysis of the top ten topics for each city revealed that some topics are common in some cities, whilst others are unique per city indicating the local food identity/culture. For example, “ITALIAN CUISINE” is a common topic for 7 cities (Amsterdam, Athens, Berlin, Brussels, Lisbon, London, and Rome), whilst the more fine-grained topic “PASTA” is unique for Rome. Similarly, “BEER” appears in the top ten topics for the cities of Berlin, Brussels and Prague, that is cities, and subsequently countries, with brewing culture and history. As for the sentiment analysis output, “BEER” receives the most positive comments in Berlin. Further qualitative analysis of the results, could indicate also the reasons behind this.

Given that the topics and the sentiments result from Foursquare comments written in English they reflect mostly visitors’ and not residents’ perceptions and opinions. Hence, in order to be able to answer questions like “*Which city has the best beer?*” or “*Which is the best burger/breakfast in town?*” in a more accurate and “comprehensive” manner, we need to take into consideration also the residents’ perceptions and more data sources (e.g. geolocated Tweets written in other languages as well).

4.2 Further Insights

The topic modelling output includes also topics that have to do with the customers’ experience in the restaurant such as the service, the ambience, the location and the value for money. Hence, if we also take into account the dimension of time (comments’ timestamp), it is possible, for example, to monitor how the food prices are evaluated by restaurants customers in different time periods (e.g. tourist seasons) in different geographical areas of a city. The results can also be used to compare the particular

restaurants' aspects across different cities and countries e.g. which city has the worst SERVICE according to users' comments? According to the sentiment analysis output, Berlin and Brussels receive the most negative comments. Further linguistic analysis could explain the reasons behind the positive or negative evaluations. For example, Fig. 4 below presents the distribution of the words "friendly" and "slow" in all user comments for each city. The least mentions of "friendly" and the most mentions of "slow" appear in the reviews for Berlin and Brussels, so perhaps this indicates that are not very satisfied with the staff attitude and efficiency in these two cities as compared to the other eight.

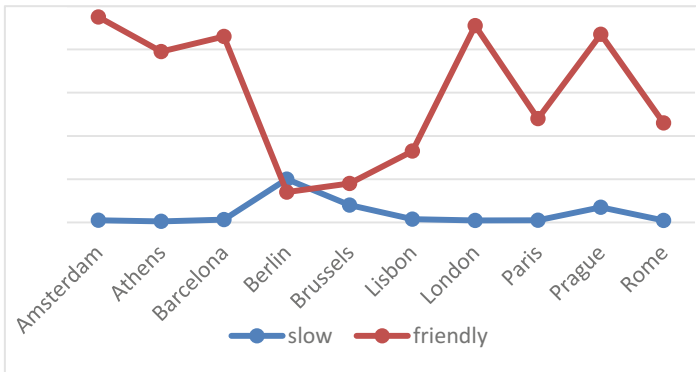


Fig. 4. Percentage of occurrences of the words "friendly" and "slow" in user reviews.

5 Conclusions and Further Work

In this paper, we presented an ongoing work exploring food experiences expressed in user reviews about food-related POIs in ten European cities with a focus on smart citizen and tourist sensing. The results indicate that the spatial distribution of food-related topics and sentiments enrich the cities maps with information that allows for a more insightful navigation across their different geographical regions, since they provide insights not available in the original data. Currently, we are working on further qualitative analysis of the results as well as on integrating results from other data sources (e.g. geolocated Tweets). Future work includes the extraction of other types of food-related insights and the combination of other types of geo-located data focusing on health issues such as obesity and diabetes, to investigate, for example, possible correlations between obesity in young ages and schools proximity to burger food.

Acknowledgements. Research presented in this paper is funded by the CAPSELLA (688813—CAPSELLA—H2020-ICT-2015) project.

References

1. Tajbakhsh, M.S., Bagherzadeh, J.: Microblogging hash tag recommendation system based on semantic TF-IDF: Twitter use case. In: 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (2016)
2. Villena-Román, J., Luna Cobos, A., González J.C.: TweetAlert: semantic analytics in social networks for citizen opinion mining in the city of the future. In: UMAP Workshops (2014)
3. Elmongui, H.G., Mansour, R., Morsy, H., Khater, S., El-Sharkasy, A., Ibrahim, R.: TRUPI: Twitter recommendation based on users' personal interests. In: Gelbukh, A. (ed.) CICALing 2015. LNCS, vol. 9042, pp. 272–284. Springer, Cham (2015). doi:[10.1007/978-3-319-18117-2_20](https://doi.org/10.1007/978-3-319-18117-2_20)
4. Karidi, D.P., Stavarakas, Y., Vassiliou, Y.: A personalized Tweet recommendation approach based on concept graphs. In: 13th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC) (2016)
5. Joseph, K., Tan, C., Carley, K.: Beyond “local,” “categories” and “friends”: clustering foursquare users with latent “topics”. In: Proceedings of ACM Ubicomp 2012, pp. 919–926 (2012)
6. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: ICWSM 2011 (2011)
7. Yang, D., Zhang, D., Yu, Z., Wang, Z.: A sentiment-enhanced personalized location recommendation system. In: 24th ACM Conference on Hypertext and Social Media (2013)
8. McAuliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in Neural Information Processing Systems, pp. 121–128 (2008)
9. Chong, W., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition (2009)
10. Hoffman, M., Cook, P., Blei, D.: Bayesian spectral matching: turning young MC into MC hammer via MCMC sampling. In: International Computer Music Conference (2009)
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
12. Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Celebi, M.Emre (ed.) Partitional Clustering Algorithms, pp. 215–243. Springer, Cham (2015). doi:[10.1007/978-3-319-09259-1_7](https://doi.org/10.1007/978-3-319-09259-1_7)
13. Greene, D., Cross, J.P.: Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Polit. Anal.* **25**(1), 77–94 (2017)
14. Estévez-Ortiz, F.-J., García-Jiménez, A., Glösekötter, P.: An application of people's sentiment from social media to smart cities. *El profesional de la información* **25**(6), 851–858 (2016)
15. Ahmed, K.B., Bouhorma, M., Ahmed, M.B.: Smart citizen sensing: a proposed computational system with visual sentiment analysis and big data architecture. *Int. J. Comput. Appl.* **152**(6), 20–27 (2016)
16. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, San Rafael (2012)
17. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: Semeval-2015 task 12: aspect based sentiment analysis. In: 9th SemEval (2015)
18. Thelwall, M., Buckley, K.: Topic-based sentiment analysis for the social web: the role of mood and issue-related words. *J. Am. Soc. Inf. Sci. Technol.* **64**(8), 1608–1617 (2013)
19. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **37**(2), 267–307 (2011)

20. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
21. McAuley, J., Leskovec, J.: From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In: WWW (2013)
22. Pennington, J., Socher, R., Manning, C.M.: GloVe: global vectors for word representation (2014)