IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 13, 2021, accepted March 26, 2021, date of publication April 7, 2021, date of current version April 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071620

Geo-Spatial Market Segmentation & Characterization Exploiting User Generated Text Through Transformers & Density-Based Clustering

LUIS E. FERRO-DÍEZ[1,2], **NORHA M. VILLEGAS**[1,3], **(Senior Member, IEEE), JAVIER DÍAZ-CELY**[1,4], **AND SEBASTIÁN G. ACOSTA**[1]

[1]Department of Information and Communication Technologies, Universidad Icesi, Cali 760031, Colombia
[2]Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria
[3]Department of Computer Science, University of Victoria, Victoria, BC V8P 5C2, Canada
[4]MO Tecnologías, 110221 Bogotá, Colombia

Corresponding author: Norha M. Villegas (nvillega@icesi.edu.co)

This work was fully supported by Universidad Icesi, Colombia.

ABSTRACT In data analysis, context information plays a significant role in enhancing the quality of the insight obtained. Furthermore, spatial analysis helps understand spatial relationships among entities. Nevertheless, findings of a comprehensive literature review show that the characterization of geographic areas based on user generated content, such as text messages, has not been sufficiently explored. This paper focuses on investigating how to combine and exploit geographic information with user generated text content to detect geographic clusters of textual events, and infer relationships between each cluster and a fixed set of retail product categories, which we consider as an insightful way to perform spatial market segmentation. We propose a workflow composed of several machine learning models incorporating Transformers as an attention mechanism and BERT-based data augmentation capable of predicting product classes from Amazon product reviews and Twitter message corpora, and then characterizing the obtained geographic clusters based on their aggregated scores. The output of our system is an effective visualization of the geographic areas with their corresponding relevance score against a fixed set of categories. We trained a product document classifier achieving an F1-Score of 86% in the test set for product reviews, and of 76% in the test set for tweets; and validated our approach by manually annotating a subset of Twitter data with respect to ten product categories. Our approach provides practitioners with a mechanism to combine location context, a Transformer encoder, and transfer learning to derive insights from geo-spatial and text data; and researchers with opportunities to continue advancing the field.

INDEX TERMS Advertising, context awareness, machine learning, natural language processing, clustering algorithms, transformers.

I. INTRODUCTION

During the last decade, the computer-science research community has been showing interest in exploiting context information to improve interactions, decisions, and analysis in systems. There are five categories of context information: individual, time, activity, relation, and location [1]. Location context refers to the information about the place in which an event occurs. This information can be in the form of geographical coordinates, or a discrete named location, like a Point of Interest (POI). There are many uses for location context in information systems. For example, modeling venue characteristics for user geo location [2], or improving recommender systems predictions [3].

The proliferation and subsequent availability of data from Location-Based Social Networks (LBSN) [4] have made an impact in the research community by enabling researchers to study and elaborate use cases that were difficult before,

The associate editor coordinating the review of this manuscript and approving it for publication was Fatos Xhafa.

e.g., monitoring and visualizing collective behavior [5]. Furthermore, LBSN provide detailed information about the location in which events and interactions take place. From this information, several studies, for example in recommendation systems, have been elaborated [3], [6], [7].

One of the use cases of location context is the characterization of geographic areas. This can be useful to understand the properties and traits that distinguish certain locations, which is important for example in Location-Based Advertising (LBA), where the advertising elements are tailored to particular locations. Nevertheless, according to our comprehensive literature review [8], few studies have focused on categorizing geographic areas based on user generated content, particularly targeting market segmentation objectives. We found approaches exploiting social context [9], POI descriptions with user traces [10], and semantic relationships of supply-domain text corpora [11]. However, none has exploited user generated geo tagged text to model or characterize locations.

This work focuses on exploiting both location context and user generated content in the form of geo tagged text messages, to detect and characterize geographic areas. The idea of our research is inspired by the LBA concept aiming for a spatial market segmentation strategy. In this sense, we wanted to exploit geo tagged messages generated directly by users to associate geographic regions to products and services, by extracting semantic relationships between the text messages and a selected group of ten product categories.

One of the major challenges we faced was the data. To the best of our knowledge, there are no publicly available datasets good enough to satisfy our objectives, i.e., geo tagged and product-annotated, user-generated text datasets from which to obtain all the information we needed. This forced us to use sources for which not all the needed information was available (location and product categories), and manually annotate a subset of one of the sources. To validate our approach, we employed an Amazon retail product review dataset (written in English) [12], [13], consisting of 8 million records, which did not contain geo-spatial coordinates, to train an attention-based neural network classifier from which we could extract semantic features to associate a vocabulary of one million words with ten product categories. Furthermore, we employed three months (from July to September 2013) of Twitter data [14] from the city of Los Angeles - USA, which was not labeled for product categories. In order to validate the results of our work, we manually annotated the Twitter dataset with respect to our category set of interest. The Twitter data consisted of a total of 36,634 geo tagged and product-annotated records. Despite the Twitter dataset being the ground of truth for our validation, the Amazon review dataset was crucial to our success since, compared to the obtained Twitter data, it was large enough to train a classifier from which we could later do transfer learning to solve the final challenge.

We implemented a series of machine learning models to classify user generated text into a fixed set of categories using state-of-the-art attention-based architectures in deep neural networks such as Transformers; and then, to detect density-based clusters in selected cities in order to obtain aggregated category scores calculated by the text classifier. We built a product category classifier with high accuracy from both the Amazon review and Twitter datasets. During this process, the major challenges were the noisiness of the Twitter data and the lack of usable messages. We addressed these problems by using BERT-based data augmentation and by doing transfer learning from a more robust review classifier we also implemented. First, we trained a deep neural network from the review data, by applying Transformers in combination with convolutional layers, achieving an F1-Score of 86% in the test set. Second, the attention layers from the review classifier model were transferred and fine-tuned to implement the final product category classifier model from the Twitter data to finally achieve an F1-Score of 76% in the test set. Furthermore, we used a density-based clustering technique to discover spatial clusters from the geo tagged messages. With the derived clusters, we employed the product category classifier to characterize the geographic areas, aggregating the individual scores of each message to get the final characterization of each cluster.

While conducting the experiments, we observed several interesting aspects. From the product classifier perspective, as expected, attention based architectures were superior to other model types such as those based on document embeddings. Moreover, we observed that applying convolutional layers on top of the Transformer encoder, followed by dense/fully connected layers, increased the model performance and yielded more accurate results compared to only using fully connected layers on top of the Transformer encoder. Additionally, we tried several data augmentation techniques in the Twitter dataset, finding that the ones that led to better results were those based on BERT, which suggests that this is indeed an excellent way to deal with poor or limited text datasets.

We built a system that integrates all the mentioned pieces together and, as a result, delivers easy-to-read interactive HTML maps, where users can explore the characterization of particular cities in the form of probability distributions over product categories. Our approach is easy to understand and it can be applied to other business processes such as customer service or operation performance, and to different contexts such as psychology, politics, and other human traits, subject to be re-trained and adapted with the corresponding datasets. This way, we aim at helping practitioners and researchers to further advance in knowledge retrieval and understanding from user generated content by exploiting location context to attain enhanced insights.

This paper is organized as follows: Section II describes our research objective and contribution; Section III discusses related work; Section IV presents the methodological aspects; Section V elaborates on the foundational aspects and techniques that support our contribution; Section VI presents the datasets and their preparation process; Section VII explains the details about the first component of our solution, the product document classifier; Section VIII illustrates the details related to the second component of our solution, the cluster calculation and the characterization of geographical areas; Section IX reports on the details and results of our experiment; Section X discusses our results. Finally, Section XI concludes this paper and outlines future work.

II. RESEARCH OBJECTIVE & CONTRIBUTION OVERVIEW

Artificial Intelligence (AI) has become an indispensable asset for business analysis, with plenty of applications in several areas, particularly in marketing [15]. For example, AI applied to location-based advertising (LBA), which is the practice of controlling the marketer information tailored for the place where the user may interact with the advertising medium [16]. The concept of LBA can be interpreted as organizing and controlling advertising information so that the location is a common ground between the ads and the user.

Inspired by the idea of LBA, our research objective is to systematically identify locations that can be associated with certain products by exploiting user generated content in form of user generated text messages. We considered this to be of interest because marketing campaigns can be costly and may suffer from low user response compared to the original investment. Moreover, there might be other reasons to estimate the optimal location for a particular activity [17], such as opening a new store [6], [7], or estimating the interest for something in particular locations to validate assumptions [18]. Our work focuses on how to characterize geographic areas in relation to a selected set of product categories. Our approach differentiates from other studies in the fact that we exploit user generated geo tagged messages as a criteria to characterize segmented geographic areas.

We employed a series of machine learning models to implement a system with the ability to train a text classifier of product categories with high accuracy using a user generated corpus from a retail product review dataset. Also, the system is capable of detecting density based clusters of geo tagged messages, aggregate the text from the clusters, and characterize each cluster with a probability distribution over ten product categories.

As a summary, our main contributions are as follows:

1) We demonstrate the usefulness of Transformers and transfer learning from a pre-trained knowledge corpus of topics, based on user generated text messages that users post on the internet. Furthermore, we demonstrate the worth of location context to characterize dense and partitioned geographic areas, and we present the results of the analysis in an effective visualization that helps researchers and practitioners gain insight from the characterization.
2) We propose a new approach to perform spatial marketing segmentation exploiting user generated content in the form of text messages, which we validate using datasets of considerable size. Furthermore, we argue that our approach can be adapted and applied to other scenarios such as customer engagement, or even geographical analysis of human thinking.
3) We created and made publicly available both the product review and Twitter datasets, which consist of eight million and 36,634 documents respectively. Furthermore, we made available the review classifier base model for further benchmarking and research.[1]

III. RELATED WORK

We conducted a comprehensive SLR (based on 168 papers) that allowed us to understand the state of the art of the usage of location context in business value chain processes [8]. Based on this characterization, we concluded that there are no contributions that approach the problem the way we do it. Our main objective was to propose strategies to perform spatial marketing segmentation, i.e., characterize geographical areas, by exploiting opinion-related user generated geo tagged data. Despite the demonstrated interest from the research community in studying location context as an asset for better knowledge-based systems, we were not able to find studies exploiting natural language features, specifically user generated text content, to characterize geographic areas.

Numerous studies have been conducted with the purpose of exploiting location context to improve data analysis, particularly in business-oriented settings. For example, East *et al.* [19] combined location context with survey data to better understand visitor behavior in a zoo; Gao *et al.* [20] implemented a system to build gazetteers[2] from volunteered geo datasets; Yu *et al.* [6] and Mao *et al.* [7] leveraged social media and other data sources along with location-based services to recommend shop types given a particular location; and Chang and Li [21] proposed a framework to predict business performance in relation to location context elements such as intrinsic attributes and competitors.

Lloyd and Cheshire [22] implemented spatial analysis and clustering approaches to derive retail center locations by exploiting geo tagged Twitter data. Another set of approaches focused on inferring characteristics associated with particular locations, such as Liu *et al.* [23], who proposed a workflow to exploit taxi trajectory data to derive optimal locations for billboards. Korakakis *et al.* [3] proposed a system to improve POI recommendations and tourism routes by exploiting social media information. Similarly, [24] worked on a location-aware personal assistant for retrieving POI and services. And Fernańdez-Gavilanes *et al.* [25] implemented a methodology to differentiate users by language and location.

Table 1 presents a comparison between the related work and our approach using the following criteria: *Exploits location data*, denoting that the authors exploit the spatial component of the dataset; *Exploits text data*, indicating that the

[1] https://ohtar10.github.io/wtsp/

[2] https://en.wikipedia.org/wiki/Gazetteer

TABLE 1. Related work - Comparing our research with other studies with applications related to location context, market segmentation and natural language approaches for business cases. ELD: Exploits location data, ETD: Exploits text data, CGA: Characterizes geographic areas, CUNL: Characterizes using natural language, SMS: Spatial market segmentation.

Author/Year	ESD	ETD	CGA	CUNL	SMS
[6], [7]	✓		✓		✓
[10]	✓		✓	✓	
[22]	✓	✓			
[19], [20], [23]	✓		✓		
[3]	✓	✓	✓		
[9]	✓	✓	✓		✓
[21]	✓	✓		✓	
[24], [25]	✓	✓			
[11]		✓		✓	
Our approach	✓	✓	✓	✓	✓

authors exploit textual properties to derive insight; *Char. geographic areas*, denoting that the presented results can be used to characterize geographical areas according to some criteria; *Char. using natural text*, indicating that natural language processing techniques were employed to perform characterization aspects; and *Spatial market segmentation*, denoting that the approach is focused on, or can be used to perform some level of spatial market segmentation.

We found three studies that result more relevant to our approach. The main difference among our approach and these three studies is the exploitation of user generated textual features in combination with location context to characterize geographic areas. The first relevant approach was the work of Anagnostopoulos *et al.* [9], which focused on exploiting social context and curated topic lists associated with certain users from which interests could be derived, then the zones are equally divided and the location traces from users are analyzed to characterize areas. The second relevant approach is the research by Dashdorj and Sobolevsky [10], which focused on analyzing GPS mobile traces of users and comparing them with POI descriptions to characterize behavioral patterns. None of these approaches did fully exploit semantic aspects of language. Finally, He *et al.* [11] exploited more in depth natural language techniques such as semantic relationships between text corpora to find similarities for supply-demand texts. However, their approach did not explored location context which is a major element in our research. Although our approach is oblivious to social context or POI, we focused on deriving characteristics from raw textual and geo tagged opinions from users compared with a knowledge corpus of a specific domain, in our case, retail products. Furthermore, we include the location context associated with the text data to segment and characterize geographic areas. We believe that our approach is complementary to the aforementioned studies, and that it adds value to the characterization of geographic areas and spatial market segmentation based on user generated content and traces.

IV. PROBLEM DEFINITION & METHODOLOGY

In this section, we present a formal definition of the problem that drove our research, as well as the methodology we followed to implement our solution.

A. PROBLEM DEFINITION

This study focuses on investigating how to characterize geographical areas by exploiting spatial relationships and semantic properties in text corpora. For example, if we observe a major region, e.g., a city, a system should be able to calculate sub-regions based on the spatial densities of geo tagged text messages, to then classify the sub-regions according to semantic relationships against a fixed set of categories.

From the problem statement (cf. Figure 1), we can derive that we have as input a set of geo tagged messages M, from which we can extract subsets corresponding to messages for a particular city $C \subset M$, and each message $m \in C$ contains two basic properties: a message m_t, corresponding to free text; and coordinates m_c, corresponding to the latitude and longitude associated with the message. We want to partition the messages from the area spanned by C into a set of smaller areas A, such that we can submit each corpus $a_{i;1 \leq i \leq N} \in A$ to classification using a fixed set of categories K, where $k_j \in K$ represents a single category of products. Given that a_i can group multiple messages with diverse topics, we want the output of the categories associated with a_i to be $k_j \in a_iK$, where $a_iK \subset K$, and every $k_j \in a_iK$ corresponds to a probability of a_i being of category k_j. Finally, the members of a_iK depend on some threshold t such that $\forall k_j \in a_iK, k_j \geq t$.

B. METHODOLOGY

Our methodology consisted of six stages: *first*, conduct an SLR; *second* collect the required data; *third*, analyze the data and prepare it for modeling and training; *fourth*, train the product categories classifier; *fifth*, implement a tool to automate the clustering aggregation and classification; and *sixth*, obtain and document the final results.

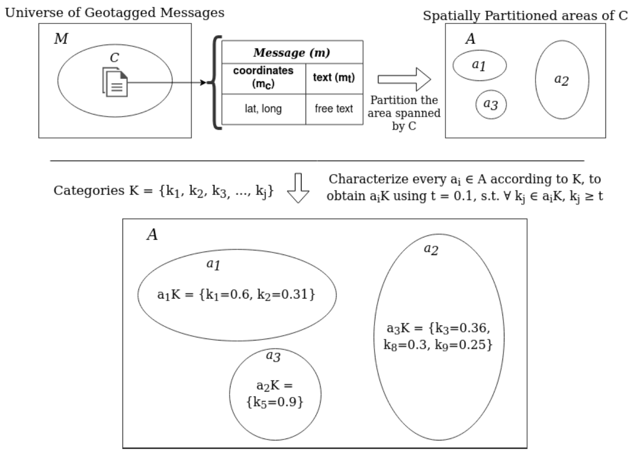

Universe of Geotagged Messages
M
C
Message (m)
coordinates (m_c)
text (m_t)
lat, long
free text
Partition the area spanned by C
Spatially Partitioned areas of C
A
a_1
a_2
a_3
Categories K = {k_1, k_2, k_3, ..., k_j}
Characterize every a_i ∈ A according to K, to obtain a_iK using t = 0.1, s.t. ∀ k_j ∈ a_iK, k_j ≥ t
A
a_1
a_1K = {k_1=0.6, k_2=0.31}
a_2
a_3K = {k_3=0.36, k_8=0.3, k_9=0.25}
a_3
a_2K = {k_5=0.9}

FIGURE 1. **Graphical representation of the problem definition. From the universe of geo tagged messages M we take subsets per city C. Every message contains coordinates m_c and text m_t. The messages are partitioned exploiting the coordinate values into partition set A. Finally, every partition $a_i \in A$ is characterized according to a set of categories K using an example threshold $t = 0.1$, such that $\forall k_j \in a_i K, k_j \geq t$.**

The findings of our SLR [8], confirm our research problem and objectives. To develop our contribution, we divided our work into two major components: *The product document classifier*, which consists of all the processes and necessary experiments related to the product document classifier responsible for the characterization of the clusters; and *the geo tagged text cluster aggregation & characterization*, which comprises everything needed to calculate and aggregate the spatial clusters and subsequent classification. Figure 2 summarizes our approach and illustrates the result delivered to users.

To avoid potential conflicts of interest, we used publicly available datasets. As for product related data, we used the Amazon product review dataset collected by [12] and [13]. We considered this to be a useful dataset for building a solid semantic knowledge base and vocabulary for product related text data, considering that the reviews are written opinions about categorizable items. However, the lack of location context in the reviews hampers our final objective. Thus, we used Twitter data [14] since it is known that Twitter allows to share the geographical locations along with the messages published, at the price of lacking product-annotated text messages and having noisy data. A new challenge arised and it was to find ways to take advantage of the available data to solve and validate our objective.

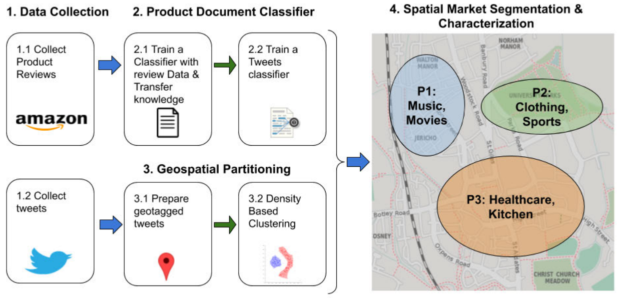

1. Data Collection
2. Product Document Classifier
1.1 Collect Product Reviews
amazon
2.1 Train a Classifier with review Data & Transfer knowledge
2.2 Train a Tweets classifier
3. Geospatial Partitioning
1.2 Collect tweets
3.1 Prepare geotagged tweets
3.2 Density Based Clustering
4. Spatial Market Segmentation & Characterization
P1: Music, Movies
P2: Clothing, Sports
P3: Healthcare, Kitchen

FIGURE 2. **Proposed approach.**

V. BACKGROUND

In this section we briefly discuss the techniques we employed in the development of our solution, as well as some alternatives to the selected techniques.

A. ATTENTION MECHANISMS & TRANSFORMERS

Humans that can read have the ability to transform user generated text into meaning. However, this is a challenging task for computers, particularly because for a machine, the information must be encoded as numbers. It is possible to represent, or *encode* words in a numerical form so the computer can process them. Nevertheless, another challenge appears when we want to analyze semantic relationships between words.

Deep learning has become the most promising field in several machine learning tasks. Despite being complex, deep learning has been showing prowess in urban geography [26], and text related tasks such as sentiment analysis [27].

In the context of deep learning for text analysis, there are certain types of neural network architectures that can be employed. Long-Short Term Memory (LSTM) [28], [29] is a common architecture for analyzing text as sequences of words. Convolutional Neural Networks (CNN) [30] is another architecture that is widely used for image related workflows, but it also has applications with text [31].

In recent years, new architectures have emerged with remarkable results in Natural Language Understanding (NLU), such as Transformers [32]. The intuition behind Transformers is that sequences can be observed in parallel and independently by multiple attention heads that compute scores for all the sequence elements and positions. The attention score is computed through linear transformations and softmax activations on each element in the sequence, and it tells us the relevance of each element towards a particular goal, e.g., classification. The *scaled dot-product attention equation* 1 summarizes the calculation of the attention score. We must provide matrices Q (queries), V (values), and K (keys), which are abstractions of word embeddings [33] for each word in a sentence, and d_k is the number of dimensions of the K embedding. When the term $\frac{QK^T}{\sqrt{d_k}}$ is passed through a softmax activation to get weight probabilities, and by multiplying the final vector V, we obtain the attention score of each vector in the sequence. Furthermore, when matrices Q, K, and V are the same, the mechanism is called "self-attention".

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

This kind of Attention Mechanism has been successfully used in Transformer architectures for NLU, such as the case of BERT [34], achieving better results compared to traditional language representation models. One particular example in which Transformers show superiority is when compared against document embeddings (Doc2Vec) [35] for contextualized representation of complete sequential data. One common use case of Transformers and Attention Mechanisms is Neural Machine Translation (NMT), which uses an encoder-decoder architecture. The encoder abstracts the input language and the decoder transforms the abstraction into the target language. We focused on classification, hence, we only needed to use the encoder element to extract semantic abstractions of documents. The abstractions obtained by the Transformer encoder can be used as predictive features for classification tasks.

1) DENSITY BASED CLUSTERING

Clustering is the task of grouping objects based on some criteria such that similar elements can be identified as a group. There are several clustering criteria, deriving into different clustering techniques, such as centroid-based clustering [36], distribution-based clustering [37], self-organized maps [38], and density-based clustering [39].

The idea of density-based clustering is that the grouping of objects is determined by how dense the members of a cluster are, i.e., the clusters are defined by higher density areas, while disperse objects are considered noise, and not associated with any cluster. Algorithms such as *Density-Based Spatial Clustering for Applications* (DBSCAN) [40] exploit spatial relationships between objects, grouping together closely packed points. Additionally, it automatically detects noisy elements that can be easily discarded. Another algorithm is the *Ordering Points to Identify The Clustering Structure* (OPTICS) [41], [42] which is a generalization of DBSCAN, and it also overcomes the limitation of having data with varying density. In both cases, there are two required parameters: ε, which represents the maximum distance between objects to be considered neighbors; and *MinPts*, which is the minimum number of neighbors before a group can be considered a cluster.

DBSCAN and some of its variants have been used in location context problems before [3], [17]. Location information encoded as geographical coordinates in a surface fits well in location context use cases. Additionally, in our use case, we argue that the density of the clusters is meaningful for the interpretation of the output, hence we considered OPTICS to be an appropriate technique in this case.

VI. DATASETS

In this section we briefly describe the datasets used in our approach and the pre-processing we implemented, as well as show some examples of the data we encountered as part of the task.

A. TWITTER DATA

We employed the Twitter Stream Grab dataset [14] for the period of June to September of 2013, which consists of 607.7 million tweets. We selected this period because the product reviews from Amazon available were from about the same time frame. The variables used were the tweet text, its geographical coordinates, place name, and country name. We were not interested in the users as individuals for this experiment, hence, we did not include any of the social relationships nor interactions with other entities or users. Also, we filtered out the tweets that did not have coordinates informed. After the pre-processing, we obtained a total of 9.1 million tweets, which is $\sim 1.48\%$ of the original dataset. Furthermore, this dataset did not have any ground truth with respect to product categories. Due to limited resources and time constraints, we were able to manually annotate a subset of 36,634 tweets corresponding to the city of Los Angeles in United States. We defined a set of ten labels. Then we semantically associated the content of each message to one or more of these labels:

- **Movies & TV**: Mentions about movie or TV show names, actors, or directors, regardless of the genre.
- **Clothing, Shoes & Jewelry**: Mentions about clothing, fashion items, or attires, including shoes or accessories.
- **Music**: Similar to movies, the mention of names, artists, performers, shows, concerts, producers, among others, regardless of the genre.
- **Technology, Electronics & Accessories**: Mentions about computing hardware, or software, as well as gadgets, mobile phones, or other technology items.
- **Books**: Similar to movies and music, mentions about literature pieces, authors, commentators, among others, regardless of the topic.
- **Toys & Games**: As with technology, the mention of games, video games, toys, or similar.
- **Home & Kitchen**: Tweets about being at home, doing kitchen activities, or mentioning particular items associated with this topic.
- **Office & School Supplies**: Similar to home & kitchen, messages about being at the office, or doing school work.
- **Health & Personal Care**: Like the topic itself suggests, messages that can be related to health or personal care items or activities.
- **Sports & Outdoors**: Messages about sport or outdoor activities, or items associated with the topic.

Additional to the aforementioned categories, we included three special categories to denote what we considered cases outside our primary objective, such as unrelated messages or noise, for a total of thirteen labels as follows:

- **Other Products**: This category was dedicated to messages that can be interpreted as talking about other kind of products outside of the first ten in this list. For example, food.
- **Other Topics**: Similarly, this category was dedicated to messages that carry a semantic meaning about something that is not, or can not be associated with any particular product or service. For example, politics.
- **Not Applicable**: Finally, this category is meant to be used as noise annotation, i.e., Twitter is notorious for carrying noisy data, such as, people just laughing at

something. These type of messages do not carry meaningful information for our subject of interest. Hence, for simplicity, we filtered them out from the classification.

It is relevant to clarify that we annotated the messages regardless of dealing with specific products or services. We considered that if a particular message, from a human standpoint, is considered to carry enough meaning such that it can make us think there can be a relationship with one of the mentioned categories, it was valuable to be taken into account. We also considered the possibility of a message carrying more than one meaning, hence, we performed a multi-label annotation for those messages we deemed relatable to two or more categories.

After annotation, only the messages labeled with at least one of the first ten categories listed above were considered. Hence, our final dataset was of 8,691 geo tagged and product annotated text messages. This imposed a significant challenge in our objective as we knew this would not be enough to derive meaningful and general results, and was the main reason for us to complement this dataset with additional sources, as explained in the next subsection. Finally, considering we could only use geo tagged messages, it was infeasible to validate whether this subset was a truly representative sample of all possible tweet messages. Table 2 shows a few examples of the annotation results. The most frequent category in this final subset was "Music" which represents 25% of the dataset, thus, the classification baseline for this task.

TABLE 2. Product documents classifier results sample. Predicted categories score in parentheses. Actual user names and links were altered.

Tweet	True Categories
En route 2 Jakarta for a show w music @usermention drums @usermention	Music
There should be school tomorrow.	Office & School Supplies
@usermention I can honestly say I have no idea when Christians birthday party was and if I was there	Other Topic
@usermention @usermention lmfaooo	Not Applicable
#Dodgers #ThinkBlue Dodgers lose, learn they're Atlanta-bound - ESPN (blog) http://somelink #SportsRoadhouse	Sports & Outdoors
Guatemalan coffee is good but this stuff is pretty awesome too...good to be back in the #USA.... http://somelink	Other Products

B. PRODUCT REVIEW DATASET

To cope with the scarce data product of the Twitter annotation task, we used an Amazon product review [12], [13] (version 2014) dataset, which includes 83 million reviews, and 9.4 million product metadata records. From the product review dataset, we selected the product identifier, summary, and review text. For the product metadata, we selected the categories, title and description. For every product in the metadata, there could be one or more product reviews, and the relationship was determined by the product identifier. Also, product categories can be hierarchical and one product can be associated with more than one category. However, for simplicity, we worked with only one category level.

Because of the large amount of data and the hierarchical nature of the categories, we pre-processed the data to obtain a sample of 59.6 million of "product documents", consisting of the review text or product description. Furthermore, similar categories were merged and used as the labels for each product document. For example, T-shirts, shirts, and dresses are related to clothing. We manually selected the categories that could be combined, and created a mapping for them according to the ones mentioned in Section VI-A. We worked with the same ten category set as mentioned in Section VI-A. Product documents that did not belong to the selected categories were left out.

After the pre-processing stage, we took a stratified sample of eight million product documents for training, we considered this to be a good enough amount to perform our experiments. Table 3 shows the count per document category in the final set. Products with more than one category were counted multiple times. The predominant category is books, representing 29.4% of the dataset, thus, we considered it as our classification baseline for this task. Table 4 shows a few examples of the final product documents for predictions.

TABLE 3. Sample document count per category.

Category	Count
Books	2,340,373
Technology, Electronics & Accessories	1,249,200
Home & Kitchen	1,152,838
Clothing, Shoes & Jewelry	731,258
Health & Personal Care	623,679
Toys & Games	601,138
Sports & Outdoors	415,380
Music	415,013
Movies & TV	404,341
Office & School Supplies	145,078

TABLE 4. Product documents classifier results sample. Predicted categories score in parentheses.

Document	True Categories
Disappointed. Plastic is too flimsy. Its not rubber outer core, its slick. Thought it wouldnt slide off tables, counters, dash. It does!	Technology, Electronics & Accessories
Jim from Ocala, Fl. great replacement product. Has been in use for several months with no problem. This is a very good replacement battery. would order again inthe future.	Home & Kitchen; Technology, Electronics & Accessories

VII. PRODUCT DOCUMENT CLASSIFIER

Our intention is to exploit semantic properties to predict characteristics in user generated opinions in the form of text messages. Hence, the first component of our system is a product document classifier.

The major challenge with our classification task was the nature of the data we were able to obtain. The Twitter dataset was not good enough and the product review dataset did

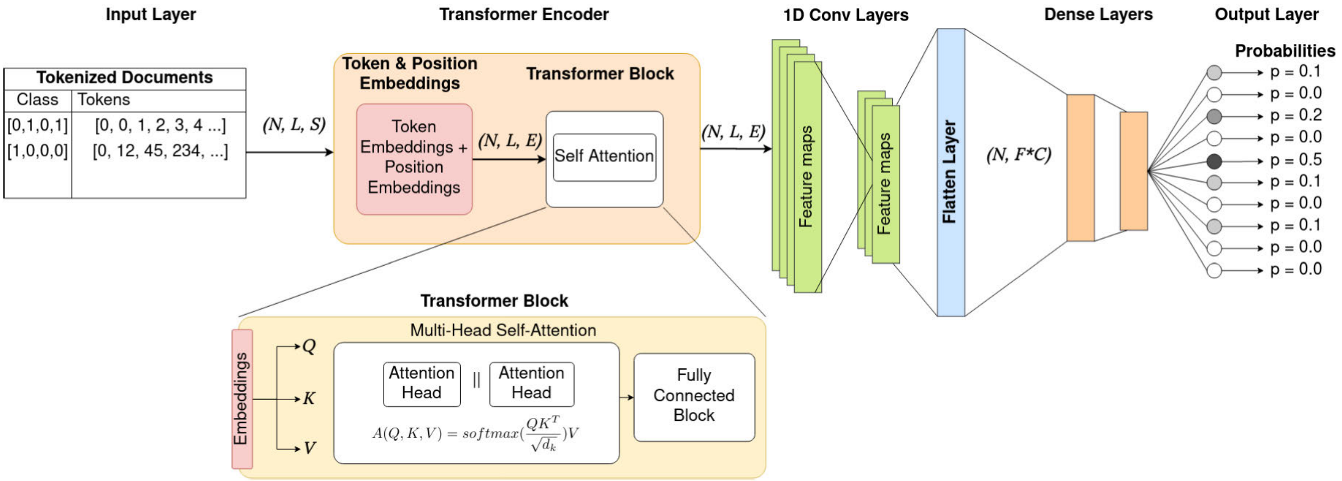


FIGURE 3. General network architecture for document classification. Where, N is the batch size, L is the input sequence length, S is the input vocabulary size, E is the embedding size, F is the output filter size, C is the number of convolutional filters on the last CNN layer, and Q, K, and V are the abstractions of the resulting embeddings to compute the attention for each word.

not have location context. The tweet classifier was the most important piece of the solution and despite using modern deep learning architectures, the model would barely learn to generalize well with little data. At the same time, if we trained a robust model with only the review dataset, there was no guarantee that it would generalize well enough by predicting data outside the training distribution. Moreover, since the review dataset did not include location context, we could not achieve the final objective of performing spatial market segmentation.

We proposed a transfer learning based solution, in which we first would train a robust model, capable of extracting meaningful semantic relationships between a large corpus, such as the review dataset, and the selected categories using Transformers. Then, we would transfer this knowledge, in the form of the learned Transformer encoder, into a final tweet classifier. This way, we were able to give a considerable and much needed boost to our classifier, and still exploit the location context.

As per our experiments (c.f., Section IX), we compared different architecture configurations and confirmed the superiority of the attention-based model compared to Doc2Vec and convolutional-based approaches. It is common to directly connect the output of the Transformer encoder directly to fully connected layers prior classification. However, we also observed that for our particular task, the combination of convolutional layers after Transformer blocks gives a considerable boost to the classification task. CNN layers calculate feature maps from input data. Feature maps are in essence activations of different parts of the input, meaning that a strong activation signals the existence of a certain feature in the input. In our use case, the CNN layers are responsible for detecting dimension features in the Transformer encoder output that is later fed into dense layers to make the final predictions. The improvement we observed by using this approach is the reason why we decided our network architecture to have both layer types.

Figure 3 shows the general idea of the network architecture we employed to classify document embeddings using Transformers and CNN. The input layer corresponds to the pre-tokenized product documents. Next, we have a positional encoding, an embedding layer and a transformer block, and a set of 1D convolutional layers to extract feature maps from the Transformer encoder activations. In a subsequent step, we flatten the output from the convolutional layers to feed them to fully connected (dense) layers. Finally, the output layer employs a softmax function to emit probabilities for each class, which we can interpret as the probability of a sample document of being in every category. One important aspect to note here is that the three matrices used as the input to the Transformer block correspond to three different linear transformations of the same word embeddings, thus our model is a self-attention mechanism.

The same architecture concept was used for training a document classifier from the reviews and another from the tweets, with the difference that the tweet classifier received the Transformer encoder pre-trained from the reviews as a starting point, and then fine tuned it, this was crucial for the success of our work.

VIII. GEO TAGGED TEXT CLUSTER AGGREGATION & CHARACTERIZATION

The second phase of our approach focuses on the clustering and score aggregation tasks. This phase allows for the partitioning of geographic areas and subsequent characterization using the product document classifier, and for the exploitation of the location context provided by the geo tagged messages. This section describes these mechanisms.

A. GEO SPATIAL CLUSTERING

We wanted to analyze geographic areas in general, without being limited by predefined regions, such as cities, neighborhood polygons or bounding boxes. Our intention was to let the points of the geo tagged messages to naturally form clusters. Moreover, because depending on the problem, the size of the clusters could represent relevant context information such as *audience size*. Under these circumstances, it would be wiser to select a region with bigger size.

Density-based clustering algorithms discover groupings based on how packed the instances are, i.e., their density. Density-based techniques have been used for geo-spatial information in the past [3], [17], particularly DBSCAN. In our approach we employed OPTICS instead, because it deals better with uneven cluster densities, as we expect from Twitter data.

B. CLUSTER CHARACTERIZATION

The final step in our approach is the characterization of geo-spatial clusters. After calculating the clusters, we submit to classification every tweet belonging to the cluster, then we average the individual scores to obtain a final score per cluster. For classification, we are deliberately omitting tweets labeled as 'Not Applicable' since these do not contribute to our goal. Furthermore, from the cluster results, we filtered out clusters that are below a predefined score threshold.

IX. VALIDATION

This section describes the experimental setting and results.

A. EXPERIMENTS & RESULTS

1) PRODUCT DOCUMENT CLASSIFIER TRAINING

We trained a model directly from Twitter data with an architecture similar to the one described in Section VII. However, despite the prowess of Transformers, the data we had available for training was not sufficient to learn meaningful features, and the model was not able to surpass the twitter classification baseline (25%). We overcame this challenge by first learning the semantic features we needed from another similar classification task. We took the subset of eight million annotated Amazon product documents, and trained a robust classifier achieving an F1-Score of 86% in the test set, surpassing the review baseline for this dataset (29.4%). We also compared the performance between a Doc2Vec based model consisting of 1D convolutional layers on top of pre-computed document embeddings and an Attention based model as previously described, demonstrating the superiority of the Attention Mechanism (cf. Table 5). We pre-tokenized the review documents and built a vocabulary out of the first million from the most common words, which we also reused later for the tweet classification task. Only tokens within the vocabulary, including trend terms, were used to form the encoded review/tweet sequences. For both review and tweet classifiers, we defined an input sequence length of 300.

TABLE 5. Product review classifiers performance comparison in the test set.

Metric	Doc2Vec-Based	Attention-Based
Accuracy	73%	81%
F1-Score	77%	86%

Table 6 shows some real examples of product review documents with the original categories, and the predicted ones with their corresponding probability scores. We used a threshold of $t = 0.1$, hence for the selected records we only show the scores for categories greater or equals than 0.1, i.e., $a_{ki} \geq 0.1$.

With a robust review classifier trained, we proceeded with a transfer learning and fine tuning approach by borrowing the Transformer encoder of this model and building a new one specifically for the tweet classification task. The model performance doubled, we were able to surpass the tweet baseline and achieved a 43% F1-Score in the test set. However, despite being a better result, we did not considered it great, once more, the size of the dataset made it difficult to improve the performance. Hence, we resorted to perform some level of data augmentation. We employed nlpaug library [43], and tried different data augmentation approaches, e.g., keyboard noise, synonyms, word embeddings for insertion and substitution, as well as context based augmentation with a pre-trained BERT model in which we could replace existing words, insert new ones, or augment the sentence based on the original message. From the mentioned augmentation approaches, the BERT-based ones yielded the best results. Hence, we only used BERT-based augmentation, making the tweet dataset ten times larger to finally work with 85,497 documents. This data augmentation mechanism helped us to improve the tweet classification F1-Score to 58% in the test set.

Finally, the final tool that helped us improve the model's performance was the inclusion of convolutional layers on top of the Transformer encoder followed by the fully connected layers. Our intuition was that the Transformer encoder still outputs multidimensional vectors per document, hence, the feature maps produced by convolutional blocks would help extract additional information. Since each word embedding is a vector, convolutional layers can extract 1D features. By including 1D convolutions, we achieved 76% F1-Score in the test set. Table 7 shows the comparison of the most important milestones in the tweet classification model.

We used Keras [44] and Tensorflow [45] to implement both classifiers using a machine with one GPU. We defined a network architecture based on what we described in Section VII. The final model for both classification tasks consisted of one transformer block with four self-attention heads, two 1D convolutional layers of 128 and 64 filters, and kernel sizes of five and three respectively. The fully connected layers consisted of three layers of 1024, 512, and 256 units with dropout regularization between layers. We used softmax as the activation function in the last layer, and Kullback-Leibler

TABLE 6. Product document classifier results sample. Predicted categories score is presented in parentheses.

Document	True Categories	Predicted Categories
Disappointed. Plastic is too flimsy. Its not rubber outer core, its slick. Thought it wouldnt slide off tables, counters, dash. It does!	Technology, Electronics & Accessories	Home & Kitchen (**0.37**); Technology, Electronics & Accessories (**0.37**); Toys & Games (**0.18**)
Jim from Ocala, Fl. great replacement product. Has been in use for several months with no problem. This is a very good replacement battery. would order again inthe future.	Home & Kitchen; Technology, Electronics & Accessories	Health & Personal Care (**0.10**); Home & Kitchen (**0.24**); Technology, Electronics & Accessories: (**0.64**)

TABLE 7. Tweet classifiers performance comparison in the test set.

Metric	Att-FC (no aug)	Att-FC (with aug)	Att-Conv1D-FC (with aug)
Accuracy	41%	54%	72%
F1-Score	43%	58%	77%

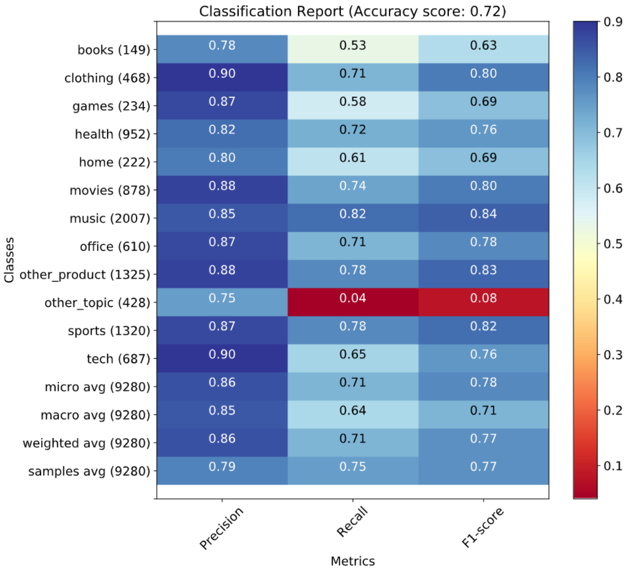


FIGURE 4. Att-Conv1D-FC model classification report.

Divergence (KLD) as the loss function with an Adam optimizer.

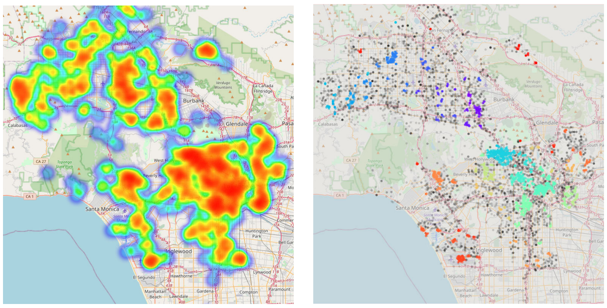
Figure 4 presents the classification report for the test set of the final tweet classification model. We can observe consistent results in all categories except for ''Other Topics'', which considering that our training sample only consists of messages with at least one of the main ten category group, we attribute this low score to the lack of individual examples for this category. However, for simplicity in our experiments and final objective, we chose to ignore this issue for now.

Finally, Table 8 shows some tweet classification examples from the test set. We observe that some examples seem to be ambiguous, yet the model activates at least one of the true categories (c.f., last example). Furthermore, for metric calculation, we used a decision threshold of 0.5 for each class, meaning that if a predicted category logit was greater than this threshold, we rounded it up to 1.0 in order to compare it with the true one-hot encoded category. We observed that several of the classification errors were because the score obtained for the categories did not surpassed the decision threshold. However, since we were interested in the logits per se, we deliberately omitted further inspection.

2) GEO SPATIAL CLUSTERING CALCULATION

We employed the Python module Scikit-learn [46], and its implementation of the OPTICS algorithm. As mentioned in Section V-A1, there are two important parameters we should provide, one of them is epsilon ϵ, which represents the maximum distance to consider two instances belonging to the same cluster. We employed the technique described by Rahmah and Sitanggang [47], in which we can find a suitable ϵ by calculating the distance to the nearest n points for each point, then sorting, plotting the results and finding the point of inflection on the curve to select the optimal ϵ. The second parameter is the minimum number of points. We chose a value of $MinPts = 10$, since we wanted aggregated corpora of at least 10 messages per cluster.

The mentioned clustering configuration allows us to discover clusters over any subset C of cities. In Figure 5, the subplot (a) shows a heatmap of geo tagged messages in Los Angeles - USA from a subset of 1,000 tweets. The subplot(b) shows the calculated clusters from the geo tagged messages. The colored dots represent clusters, and the gray dots represent noise data points that were not considered as part of any cluster.



(a) Los Angeles Heatmap (b) Los Angeles Clusters

FIGURE 5. Los Angeles geo tagged messages and clusters.

TABLE 8. Product document classifier results sample. Predicted categories score in parentheses.

Tweet	True	Predicted
#WeLoveLA Pat Haden meets with NCAA to seek easing of penalties against USC http://someurl #SportsRoadhouse	Sports, Fitness & Dieting	Sports, Fitness & Dieting (0.99)
@usermention: I don't understand how people beat GTAV in 4 days or less from when it launched, no life. Longest game ever" lmao :/	Toys & Games	Toys & Games (0.99)
Hi @usermention @usermention @usermnetion @usermention. im looking for a #coconut brand that wants to be included in my YouTube videos.	Health & Personal Care	Health & Personal Care (0.98)
people at work thinking of " have a good life " like these three play like quiet streets and shieet.	Office & School Supplies, Sports & Outdoors	Sports & Outdoors (0.75)
" @usermention : they caníd t possibly wait to go play these la town shows!!!! sickkkkkkk " its so not fair im currently stuck at another home while you fucking guys are done here -, - fml	Music	Music (0.46), Sports & Outdoors (0.47)

B. CLUSTER AGGREGATION & CHARACTERIZATION

Once the clusters were calculated, we classified all individual messages per cluster, and averaged the scores to get the final characterization scores per cluster. To visualize the results, we also needed to aggregate the cluster points, since the idea was to calculate polygons associated with each cluster. Visualizing cluster polygons in a map allows us to clearly see the shape, boundaries, a subjective sense of coverage among each cluster, and their characterization.

In the context of spatial data, there are several methods to partition a surface. We employed convex hulls [20], [48], through the Quickhull algorithm [49] because we were interested in the polygons enclosing the cluster data points, instead of a span region based on the distribution of points.

Figure 6 shows an example of the final output generated by our solution. This figure represents a portion of the city of Los Angeles, showing their predicted product categories from July to September 2013. We can observe several polygons of different shapes, which represent the enclosing boundaries of each cluster. For some of them, we show their characterization resulting from our tweet classifier. The output of our system is an HTML interactive map.[3] We employed the Folium Python module on top of OpenStreetMap [50] to render and save the HTML file.

It is possible to observe how the clusters tend to vary in shape and size (according to their density), as well as their predicted categories. For example, cluster #88 shows user generated content related to Music (26.32%), Sports (19.09%), Health (18.55%) and Technology (11.10%). We employed a threshold $t = 0.1$ to filter out scores below this limit. Furthermore, if we inspect the word cloud based on word frequencies of cluster #88 (see Figure 7), we can observe words like "football" and "season", which we can semantically relate to Sports. Or similarly, "fat", "food", or "sick", which we can associate with Health. However, it is interesting to see that "Office & School Supplies" did not appear considering the presence of words such as "school" and "university", this definitely makes up for future work and experimentation.

X. DISCUSSION & LIMITATIONS

A. RESEARCH DISCUSSION

We argue that our contributions are twofold. First, we highlight the importance of context information in decision support systems, in our particular case, location context. Second, Attention Mechanisms not only help achieving better results in multi-label classification tasks, they also help augment the datasets when examples are scarce, and the knowledge learned from similar tasks is transferred with good results in presence of larger amounts of meaningful data. Moreover, the nature of the output tensors of the Transformer encoder can be suitable to convolutional layers as well, and allows gaining a boost in performance compared to simple fully connected layers. Our scenario and results demonstrate that, whenever possible, we should analyze and explore location information around the events of systems, even if the relationship could appear futile at the beginning. Additionally, the combination of the different context information types could further enhance the quality of the results, or reveal crucial and interesting insights that we were not considering before.

In terms of theoretical implications, the results presented in Section IX-A demonstrate that we were able to successfully partition and subsequently characterize geographic areas by exploiting text data and location context. In this sense, we extend the theoretical applications of [9] and [10]. Our implementation is capable of calculating geographical clusters at city level in any country, and it is applicable to different languages, if trained with the corresponding language corpus, which expands the scope presented by [11]. Moreover, the results are easy to read and interpret, since the spatial scope and area densities are easy to understand. Furthermore, our product document classifier is simple yet effective. Our validation suggests that it is capable of predicting product categories with high accuracy. This was a key aspect to the success of this project. Considering we

[3] https://ohtar10.github.io/wtsp/

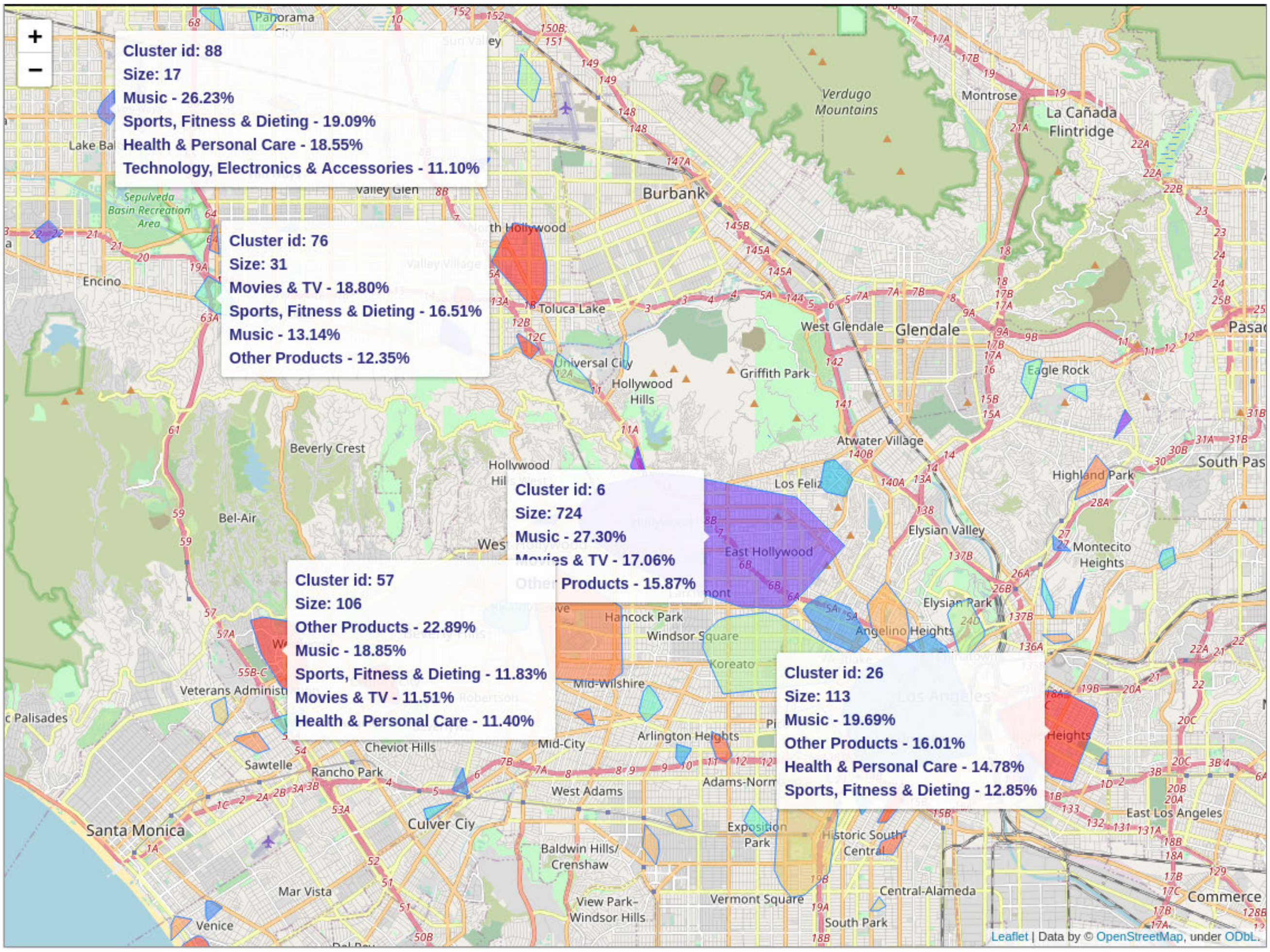


FIGURE 6. Geo spatial market segmentation of Los Angeles - July-September 2013.

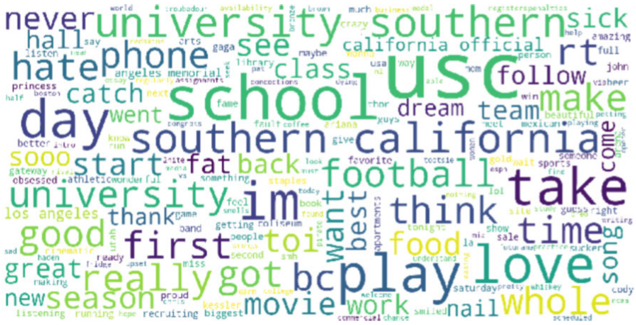


FIGURE 7. Word cloud of cluster id 88 from Figure 6.

had to use different data sources to exploit the geo-spatial context, we needed to have a robust document classifier to get meaningful results. However, we deliberately excluded noisy and non product-related tweets from our experiments to facilitate progress. In more realistic scenarios, practitioners should consider building additional models or data pipelines to deal with the noisy data, or expand the classification task even further.

Our approach can leverage on the exploitation of other context types such as time [51]. Given that geo localized messages are dynamic in time, by calculating the spatial segments in different time spans, our solution can make explicit the dynamic nature of the variables observed in the geographic region over time. This is a powerful feature for marketing business applications. Another context type that can be leveraged with our solution is social context. For example, by considering social context, we can build geographically associated networks or create stronger relationships with products and services, which has been demonstrated by [9]. Individual context can contribute with additional features to refine geographical partitions; and activity context can contribute to better describe the events that are occurring at a particular location and time, such as the case of [10].

In the context of LBA and similar, our solution has the potential of becoming an insightful source for marketing campaign planning. For example, being able to look ahead for the locations at which there are people with demonstrated interest in what a business wants to communicate can alleviate the resources needed to roll out marketing campaigns, and subsequently increase the chances of success. Hence, our solution has the potential of complementing shop-type recommendation approaches [6], [7]. The temporal characteristics that can be exploited as well, demonstrate that time context is also important to elaborate communication strategies such as seasonal marketing campaigns. Being able to observe these variations geographically can definitely add significant value to the decisions made by businesses.

Our application can also be applied to service-related processes. For example, practitioners could spatially analyze user-generated customer service support messages. This could lead to geographically visualize and design better service strategies tailored to specific areas and use cases, e.g., geo localizing problems in the service chain that need to be addressed. Another business process that could benefit from our approach is operations. For instance, in geographically distributed outlets or service stations that continuously generate reports in natural language to assess or guarantee operational continuity, our approach can help understand and categorize operational events. Finally, regarding logistics, we argue that for transportation networks or services that are frequently moving and generating reports, by incorporating additional elements, such as speech recognition, our approach can contribute to a dynamic geo-spatial event mapping for better routing or distribution planning.

Outside the business context, our approach has tremendous potential in other applications. For instance, characterizing other human traits through spatially generated content can help model locations and understand properties that could only be obtained through other methods, e.g., surveys. Moreover, surveys can be costly, designed with specific objectives, and biased. By extracting knowledge directly from user words, we can obtain additional insights to enrich surveyed data. A specific use case could be politics. Our approach can facilitate the geographical modeling of cities towards political affiliation based on what the users are saying online, in their own words. Additionally, because we are using user generated content, we could add other NLU practices, such as sentiment analysis and topic modeling to further characterize the interest on different product categories.

In general, our approach contributes to the analysis and geographic visualization of textual data. This eventually fosters more options to perform collective and population analysis. Our solution is versatile, and it can be incorporated with other approaches to expand its scope. For example, voice recognition techniques, such as speech-to-text, can also be employed to extract verbal input prior executing our solution, this would enable practitioners not to limit their analysis to textual datasets. Moreover, our workflow can conceptually be adapted beyond natural language analysis. We observe great potential in the characterization of geographical areas by exploiting other types of unstructured data, such as photos and videos.

As for practical implications, we recommend performing the spatial characterization as per city basis, this is mostly due to how the density based clusters vary according to the observed region and the computational power available. Moreover, in similar cases to Twitter, where none or only a small portion of messages carry the location context, incorporating other solutions to predict location context from messages lacking this information will significantly increase the observation surface [18]. Furthermore, practitioners might want to tune or extend upon our classification results, particularly for the classes with lower F1-scores. With respect to the semantic aspects of text, Transformer encoders deliver superior results compared to plain word embeddings, and combined with convolutional layers on top of Transformer encoders, there is plenty of space to extract representations of the semantic space despite the corpus size. Another technical aspect we consider relevant to explore is to use the recently proposed Modern Hopfield Networks [52] in place of the Transformer encoder or Attention Mechanism, considering their cited benefits such as their large memory capacity in pattern storage and retrieval as well as the few training steps they require compared to traditional architectures.

Finally, we created and made publicly available the knowledge base corpus of products, the classified tweets, the product reviews classifier, the twitter classifier, and a spatial segmentation tool for further research and experiments, allowing researchers and practitioners to implement and compare other approaches against ours.

B. LIMITATIONS

We encountered some limitations regarding the data used. In particular, the lack of dynamic and geo tagged opinion data. However, in the case of Twitter, other researches have proposed approaches to infer the locations of messages when lacking this information [18], [53]. Furthermore, to the best of our knowledge, there were few public geo tagged free text datasets that we could exploit. Regarding the product reviews, we were not able to find a geo tagged version, or a similar dataset that presented free text with associated categories along with geographical coordinates per text fragment. The limited resources we had at hand, prevented us from acquiring larger datasets, which could hamper the generalization potential of our solution.

When this study was being conducted, the Amazon product review dataset was available for reviews up to 2014. In order to minimize a possible temporal drift, product of semantic, topic, or entity differences in datasets from two different periods, we decided to work with Twitter data from a similar time frame at the cost of working with outdated data. However, from the experimental and proof of concept standpoint, we did not considered this to be a blocking limitation. Finally, we had no instruments for validating our characterization with a real business situation.

XI. CONCLUSION & FUTURE WORK

From the obtained results, we conclude that the characterization of geographic areas exploiting user generated content in the form of text messages is possible. Furthermore, we achieved a simple yet effective category classification model good enough for our LBA scenario which is applicable to other domains. Our approach is one of a set of different solutions and variants that can be applied to geographic characterization. In particular, one aspect that we ignored was sentiment analysis, which can become a useful trait to observe in the characterization. We argue that our approach contributes to the exploitation of location context and textual data to derive insight from our environment, and that it can become a valuable asset in different business endeavors and other study fields.

One aspect we highlight is the fact that the datasets we selected are not strictly related to each other, i.e., Twitter is not necessarily focused on product reviews, and Amazon product reviews are not necessarily for opinions about any topic. Despite this, the learned knowledge from the product reviews successfully adapted to the tweet classification task and delivered great results. Additionally, we could not have achieved our results without the Transformers applied to both the model training and the data augmentation tasks.

We identify several paths for future work. For example, we see opportunities in obtaining other geo tagged textual datasets, or explore options to estimate the location of existing non-tagged datasets. Also, we did not fully explored time context in this study, and given the ever-evolving nature of opinions, we believe this will definitely add value to the interpretation of the results. Other aspects that we consider interesting to explore refer to language dialects and even other languages, since naturally these vary depending on the geographic region.

From the technical standpoint, despite narrowing down the amount of data points to those of selected cities, the amount of data points can be large. We observed that clustering calculation was a bottleneck to obtain faster results. Thus, investigating techniques for online and batch training, or even considering other spatial partitioning approaches, might help speed up the clustering detection process while preserving the density-based requirement. Another technical aspect we will explore in the future is trying with other network architectures and layers such as Hopfield.

Incorporating other context information can boost up the quality of the insights obtained even further. For example, social context as the relationships among users or other content creators can strength the model performance. Moreover, we consider our workflow can be applied in other scenarios such as psychological traits, geographic sentiment analysis, tourism & leisure, and cultural aspects, among others. For instance, our approach can be applied to characterize geographical areas in terms of politic affiliation. Another example could be the detection of harmful speech and situations that could signal a threat to safety of living beings. However, depending on the final objective, there might be other challenges, such as privacy & security enforcement, or authorization of data providers as well as end users.

Finally, although our application and validation were focused on a specific business process, our approach is applicable to other processes and scenarios as well. Our problem definition can be adjusted and implemented outside the business realm. We designed our application with the necessary flexibility to be applied in other contexts. This not only facilitates studying other scenarios with geo tagged textual data, but also establishes a basis to add and implement other relevant techniques and practices, such as speech recognition, semantic analysis, and topic modeling, or even combine them with other practices outside NLU. Furthermore, our work can motivate researchers advance in similar study fields. For example, in the characterization of geographic areas by exploiting other input sources such as audio visual data with computer vision techniques to analyze video footage of urban, rural, and natural areas.

CONFLICT OF INTEREST

None.

ACKNOWLEDGMENT

Base map and data are from OpenStreetMap and OpenStreetMap Foundation.

REFERENCES

[1] N. M. Villegas and H. A. Müller, “Managing dynamic context to optimize smart interactions and services,” in *The Smart Internet* (Lecture Notes in Computer Science), vol. 6400. Berlin, Germany: Springer, 2010, pp. 289–318.

[2] W.-H. Chong and E.-P. Lim, “Exploiting user and venue characteristics for fine-grained tweet geolocation,” *ACM Trans. Inf. Syst.*, vol. 36, no. 3, pp. 1–34, Apr. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3156667

[3] M. Korakakis, E. Spyrou, P. Mylonas, and S. J. Perantonis, “Exploiting social media information toward a context-aware recommendation system,” *Social Netw. Anal. Mining*, vol. 7, no. 1, p. 42, Dec. 2017. [Online]. Available: http://link.springer.com/10.1007/s13278-017-0459-9

[4] Y. Zheng, “Location-based social networks: Users,” in *Computing With Spatial Trajectories*. New York, NY, USA: Springer, 2010, pp. 243–276.

[5] D. Yang, D. Zhang, L. Chen, and B. Qu, “NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs,” *J. Netw. Comput. Appl.*, vol. 55, pp. 170–180, Sep. 2015.

[6] Z. Yu, M. Tian, Z. Wang, B. Guo, and T. Mei, “Shop-type recommendation leveraging the data from social media and location-based services,” *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 1, pp. 1–21, Aug. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2974720.2930671

[7] X. Mao, X. Zhao, J. Lin, and E. Herrera-Viedma, “Utilizing multi-source data in popularity prediction for shop-type recommendation,” *Knowl.-Based Syst.*, vol. 165, pp. 253–267, Feb. 2019.

[8] L. E. Ferro-Diez, N. M. Villegas, and J. Diaz-Cely, “Location data analytics in the business value chain: A systematic literature review,” *IEEE Access*, vol. 8, pp. 204639–204659, Nov. 2020.

[9] A. Anagnostopoulos, F. Petroni, and M. Sorella, “Targeted interest-driven advertising in cities using Twitter,” *Data Mining Knowl. Discovery*, vol. 32, no. 3, pp. 737–763, May 2018. [Online]. Available: http://link.springer.com/10.1007/s10618-017-0529-7

[10] Z. Dashdorj and S. Sobolevsky, “Characterization of behavioral patterns exploiting description of geographical areas,” in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVII* (Lecture Notes in Computer Science), vol. 9860. Springer, 2016, pp. 159–176.

[11] X. He, X. Meng, Y. Wu, C. S. Chan, and T. Pang, “Semantic matching efficiency of supply and demand texts on online technology trading platforms: Taking the electronic information of three platforms as an example,” *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102258.

[12] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, “Image-based recommendations on styles and substitutes,” in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 43–52. [Online]. Available: http://jinni.com

[13] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 507–517, doi: 10.1145/2872427.2883037.

[14] J. Scott. (2012). *Archive Team: The Twitter Stream Grab*. [Online]. Available: https://archive.org/details/twitterstream?tab=about

[15] C. Campbell, S. Sands, C. Ferraro, H.-Y.-J. Tsao, and A. Mavrommatis, "From data to action: How marketers can leverage AI," *Bus. Horizons*, vol. 63, no. 2, pp. 227–243, Mar. 2020.

[16] G. C. Bruner and A. Kumar, "Attitude toward location-based advertising," *J. Interact. Advertising*, vol. 7, no. 2, pp. 3–15, Mar. 2007.

[17] F. Chen, J. Qi, H. Lin, Y. Gao, and D. Lu, "GOAL: A clustering-based method for the group optimal location problem," *Knowl. Inf. Syst.*, vol. 61, no. 2, pp. 873–903, Nov. 2019. [Online]. Available: http://link.springer.com/10.1007/s10115-018-1307-6

[18] P. Zola, C. Ragno, and P. Cortez, "A Google trends spatial clustering approach for a worldwide Twitter user geolocation," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102312.

[19] D. East, P. Osborne, S. Kemp, and T. Woodfine, "Combining GPS & survey data improves understanding of visitor behaviour," *Tourism Manage.*, vol. 61, pp. 307–320, Aug. 2017.

[20] S. Gao, L. Li, W. Li, K. Janowicz, and Y. Zhang, "Constructing gazetteers from volunteered big geo-data based on Hadoop," *Comput., Environ. Urban Syst.*, vol. 61, pp. 172–186, Jan. 2017.

[21] X. Chang and J. Li, "Business performance prediction in location-based social commerce," *Expert Syst. Appl.*, vol. 126, pp. 112–123, Jul. 2019.

[22] A. Lloyd and J. Cheshire, "Deriving retail centre locations and catchments from geo-tagged Twitter data," *Comput., Environ. Urban Syst.*, vol. 61, pp. 108–118, Jan. 2017.

[23] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu, "SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 1–10, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7534856/

[24] L. Massai, P. Nesi, and G. Pantaleo, "PAVAL: A location-aware virtual personal assistant for retrieving geolocated points of interest and location-based services," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 70–85, Jan. 2019.

[25] M. Fernández-Gavilanes, J. Juncal-Martínez, S. García-Méndez, E. Costa-Montenegro, and F. Javier González-Castaño, "Differentiating users by language and location estimation in sentiment analisys of informal text during major public events," *Expert Syst. Appl.*, vol. 117, pp. 15–28, Mar. 2019.

[26] G. Grekousis, "Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis," *Comput., Environ. Urban Syst.*, vol. 74, pp. 244–256, Mar. 2019.

[27] H. H. Do, P. W. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Aug. 2020, Art. no. 132306. [Online]. Available: http://arxiv.org/abs/1808.03314, doi: 10.1016/j.physd.2019.132306.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998.

[31] S. V. Georgakopoulos, A. G. Vrahatis, S. K. Tasoulis, and V. P. Plagianakos, "Convolutional neural networks for toxic comment classification," in *Proc. ACM Int. Conf. Ser.* New York: Association Computing Machinery, Jul. 2018, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3200947.3208069

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6009. [Online]. Available: https://arxiv.org/abs/1706.03762v5

[33] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Oct. 2013, pp. II-1188–II-1196. [Online]. Available: http://arxiv.org/abs/1310.4546

[34] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL HLT-Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Stroudsburg, PA, USA: Association Computational Linguistics, Oct. 2019, pp. 4171–4186. [Online]. Available: http://arxiv.org/abs/1810.04805

[35] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Intl. Conf. Mach. Learn. (ICML)*, vol. 4, May 2014, pp. 2931–2939. [Online]. Available: http://arxiv.org/abs/1405.4053

[36] S. K. Uppada, "Centroid based clustering algorithms—A clarion study," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, pp. 7309–7313, 2014.

[37] Z. Yu, X. Zhu, H.-S. Wong, J. You, J. Zhang, and G. Han, "Distribution-based cluster structure selection," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3554–3567, Nov. 2017.

[38] H. Yin, "The self-organizing maps: Background, theories, extensions and applications," in *Computational Intelligence: A Compendium* (Studies in Computational Intelligence), vol. 115. Berlin, Germany: Springer, 2008, pp. 715–762. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-78293-3_17

[39] G. I. Webb, J. Fürnkranz, J. Fürnkranz, J. Fürnkranz, G. Hinton, C. Sammut, J. Sander, M. Vlachos, Y. W. Teh, Y. Yang, D. Mladeni, J. Brank, M. Grobelnik, Y. Zhao, G. Karypis, S. Craw, M. L. Puterman, and J. Patrick, "Density-based clustering," in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2011, pp. 270–273.

[40] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Aug. 1996, pp. 226–231.

[41] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *SIGMOD Rec. (ACM Special Interest Group Manage. Data)*, vol. 28, no. 2, pp. 49–60, Jun. 1999. [Online]. Available: http://portal.acm.org/citation.cfm?doid=304181.304187

[42] E. Schubert and M. Gertz, "Improving the cluster structure extracted from OPTICS plots," in *Proc. CEUR Workshop*, vol. 2191, 2018, pp. 318–329.

[43] E. Ma. (Aug. 2019). *Nlp Augmentation*. [Online]. Available: https://github.com/makcedward/nlp

[44] F. E. A. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io and https://github.com/fchollet/keras

[45] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[47] N. Rahmah and I. S. Sitanggang, "Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra," in *Proc. IOP Conf. Ser., Earth Environ. Sci.*, 2016, vol. 31, no. 1, Art. no. 012012.

[48] S. Kisilevich, M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko, "Event-based analysis of People's activities and behavior using flickr and panoramio geotagged photo collections," in *Proc. 14th Int. Conf. Inf. Visualisation*, Jul. 2010, pp. 289–296. [Online]. Available: http://ieeexplore.ieee.org/document/5571255/

[49] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, Dec. 1996. [Online]. Available: http://dl.acm.org/doi/10.1145/235815.235821

[50] C. OpenStreetMap. (2017). *Planet Dump*. [Online]. Available: https://planet.osm.org and https://www.openstreetmap.org

[51] N. M. Villegas, C. Sánchez, J. Díaz-Cely, and G. Tamura, "Characterizing context-aware recommender systems: A systematic literature review," *Knowl.-Based Syst.*, vol. 140, pp. 173–200, Jan. 2018.

[52] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlovic, G. Kjetil Sandve, V. Greiff, D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "Hopfield networks is all you need," 2020, *arXiv:2008.02217*. [Online]. Available: http://arxiv.org/abs/2008.02217

[53] P. Thomas and L. Hennig, "Twitter geolocation prediction using neural networks," in *Language Technologies for the Challenges of the Digital Age* (Lecture Notes in Computer Science), vol. 10713. New York, NY, USA: Springer-Verlag, 2018, pp. 248–255.

LUIS E. FERRO-DÍEZ received the B.Sc. degree in software systems engineering from Institución Universitaria Antonio José Camacho, Cali, Colombia, and the Master of Science degree in informatics and telecommunications from Universidad Icesi, Cali, with a focus on machine learning engineering. He has been a team lead for different top-tier location and technology projects, helping with design, implementation, validation, and delivery endeavors. He has worked with important clients in the technology industry, including HERE Technologies and IPG Mediabrands (Kinesso) under PSL-Perficient, Colombia, designing and implementing data-oriented solutions to support critical business processes. He is currently a Software and Machine Learning Research Engineer with over 13 years of professional experience in several technology sectors. He is also working as a Researcher with the Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria. His research interests include software engineering and development, system and architecture design, big data platforms, cloud environments, and machine learning.

NORHA M. VILLEGAS (Senior Member, IEEE) received the B.Sc. degree in software systems engineering and the Specialist Diploma degree in management of information systems from Universidad Icesi, Cali, Colombia, and the Ph.D. degree in computer science from the University of Victoria, Canada, with a focus on software engineering. She is currently an Associate Professor with the Department of Information and Communication Technologies, Universidad Icesi; the Director of the Software Systems Engineering Bachelor Program; and an Adjunct Assistant Professor with the University of Victoria. She has 20 years of professional experience in academia and industry. Her research has been recognized internationally by her peers for its quality and relevance. She has published an important number of book chapters and refereed journal and conference proceedings papers that have been coauthored with top researchers worldwide. She investigates the application of software engineering models, techniques, and architectures to the development of smart software systems, that is, systems that are context-aware and self-adaptive. In particular, she is interested in the application of feedback loops, dynamic context management, and autonomic computing mechanisms to interdisciplinary areas that are crucial for the advancement of society, such as smart cyber-physical systems, cognitive computing, and digital twins. She is also highly engaged in engineering education research. Over the last nine years, she has co-chaired several international workshops and served as a program committee member for many IEEE international conferences and symposia in the field of software engineering.

JAVIER DÍAZ-CELY received the B.Sc. degree in computer engineering from Pontificia Universidad Javeriana, Cali, Colombia, the master's degree in artificial intelligence, pattern recognition and applications and the Ph.D. degree in informatics from Sorbonne Université, and the master's degree in corporate finance from the Conservatoire National des Arts et Métiers, Paris, France. He is currently the Chief Analytics Officer of MO Technologies, a Fintech that aims at revolutionizing the credit scoring industry and promoting financial inclusion, by giving everyone, independently from their credit history, the possibility to access nano and micro loans in a simple and effective way, leveraging innovative credit scoring through the use of alternative data sources. He believes that business analytics offers an incredible set of tools that enable organizations to achieve strategical goals by optimizing processes, solving problems, and discovering new and disruptive opportunities of creating value from data. With more than 15 years of experience in the field of artificial intelligence in France and Colombia, he has gathered considerable experience applying machine learning to real-world problems in several sectors, such as banking and finance, telecommunications, retail, health, and education, either working directly for a company (Orange France Télécom, Sociéte Générale, Banco Falabella) or as a Consultant (Altran, Icesi University). He has also had the chance to work as a University Professor and Applied Analytics Researcher, first as a part of the LIP6 Laboratory, Paris. He was the Director of the Master's in Data Science Program, Universidad Icesi, where he is also a Faculty Member of the School of Engineering.

SEBASTIÁN G. ACOSTA is currently pursuing the B.Sc. degree in software systems engineering with Universidad Icesi, Cali, Colombia. He has two years of experience as a Research Assistant and Analytics Solutions Developer with the ICESI's Finance and Economics Research Centre (CIENFI, originally in Spanish), working on projects that involve hybrid recommender systems, advanced clustering algorithms, and data engineering tasks. He is also an Artificial Intelligence Enthusiast and has been studying machine learning independently for two years, his subfields of expertise are natural language processing (NLP) and computer vision. He also has had the chance to work as a researcher on multilingual text processing projects, applying cutting-edge language modeling techniques on business cases.