ACCEPTED AUTHOR MANUSCRIPT
Pengyuan Liu, Stefano De Sabbata,
A graph-based semi-supervised approach to classification learning in digital geographies,
Computers, Environment and Urban Systems, Volume 86, 2021, 101583, ISSN 0198-9715,
https://doi.org/10.1016/j.compenvurbsys.2020.101583.

# A Graph-Based Semi-supervised Approach to Classification Learning in Digital Geographies

Pengyuan Liu and Stefano De Sabbata

School of Geography, Geology and Environment, University of Leicester

*November 6, 2020*

**Abstract:** As the distinction between online and physical spaces rapidly degrades, social media have now become an integral component of how many people's everyday experiences are mediated. As such, increasing interest has emerged in exploring how the content shared through those online platforms comes to contribute to the collaborative creation of places in physical space at the urban scale. Exploring digital geographies of social media data using methods such as qualitative coding (i.e., content labelling) is a flexible but complex task, commonly limited to small samples due to its impracticality over large datasets. In this paper, we propose a new tool for studies in digital geographies, bridging qualitative and quantitative approaches, able to learn a set of arbitrary labels (qualitative codes) on a small, manually-created sample and apply the same labels on a larger set. We introduce a semi-supervised, deep neural network approach to classify geo-located social media posts based on their textual and image content, as well as geographical and temporal aspects. Our innovative approach is rooted in our understanding of social media posts as augmentations of the time-space configurations that places are, and it comprises a stacked multi-modal autoencoder neural network to create joint representations of text and images, and a spatio-temporal graph convolution neural network for semi-supervised classification. The results presented in this paper show that our approach performs the classification of social media content with higher accuracy than traditional machine learning models as well as two state-of-art deep learning frameworks.

**Keywords:** social media, multi-modal autoencoder, neural network, graph convolutional network, digital geographies.

# 1 Introduction

Social media have become major platforms for people to communicate and exchange information regarding a wide range of topics. The information created and distributed through such platforms is now a significant source for scholars to understand the reproduction of urban spaces (Shaw and Graham, 2017). The intersections between the "code" (Dodge and Kitchin, 2004) of social media platforms and space capture the "localities" of users' everyday activities, augment spatial experiences (Elwood and Leszczynski, 2013), and shape the representations of places emerging from those platforms. The representation and interpretation of data retrieved from geo-referenced social media provide a means by which to assess different urban dynamics (e.g., mobility, land use and urban activities, event detection, etc.) (Martí et al., 2019), and further contribute to a digitally layered urban environment (Zook and Graham, 2007; Shaw and Graham, 2017). Despite the unequal geographies of social media platforms (Ballatore and De Sabbata, 2018), there is a growing interest in analysing such information from a geographic perspective within the field of digital geographies (Ash et al., 2018). However, traditional qualitative analysis often struggles with tackling large datasets, and the volume of data produced daily on social media is enormous. Thus quantitative analysis and summarisation are frequently necessary steps in digital geographies. That creates a strong association with GIScience, where data mining approaches have been applied to identify users' opinions and online trends, to study the emergence of place from space through content production (Graham et al., 2015), or to monitor events from football to earthquakes (Frias-Martinez and Frias-Martinez, 2014; Ifrim et al., 2014; Sechelea et al., 2016; Zahra et al., 2017) and to understand the digital representations of a place (Ballatore and De Sabbata, 2020).

In computer science and related disciplines, sentence-level topic extraction from social media posts has attracted wide attention, where research has mainly focused on supervised learning approaches with well labelled and balanced data (Medhat et al., 2014). However, labelling large volumes of social media posts can be a lengthy and costly procedure as it requires a significant amount of human intervention. Such approaches are only viable when a pre-defined set of topics or labels has been agreed upon by a large number of stakeholders, for instance, for monitoring scheduled events or natural disasters. Such approaches are more difficult to employ effectively for exploratory analysis or monitoring of unexpected events. Other studies have adopted unsupervised approaches in the context of summarising social media posts, such as methods based on n-grams (e.g., Poorthuis and Zook, 2017; Hamid et al., 2005). So far, however, limited attention has been given to the study of exploratory analysis, where only vague or no labels at all have been pre-defined. Conversely, semi-supervised learning approaches, which do not require complete labelled training, have achieved competitive results in learning accuracy, without the time and costs needed for the training data preparation step of supervised learning (Zhu and Goldberg, 2009). However, there are severe concerns about the uncertainties of social media analysis, as raw data collected from social media platforms tend to be noisy and fuzzy, rendering any approach problematic when applied to "live" data (Sommer, 2016). Thus, creating a robust framework, able to produce consistent results with imbalanced datasets is a crucial task in digital geographies.

In this paper, we present and test an approach for the exploratory analysis of social media content, capable of automatically classifying a large volume of posts for a user-defined set of labels. That is, the labels are not an integral part of the framework here presented,

but they can instead be customised for any specific research project a user might be developing. The classification process takes into account not only on the textual component but also taking into account their visual content. The latter is a crucial contribution of our approach, as only a handful of papers account for images when conducting quantitative analyses of social media content (Gao et al., 2015; Xu et al., 2017; Huang et al., 2018b). Furthermore, conceptualising posts as "augmentations"(Graham et al., 2015) of places, understood as "time-space configurations"(Agnew, 2011), we go beyond the geo-tag (Crampton et al., 2013a) by developing graph convolutional networks that account for the relationships between each post and its spatio-temporal neighbours. To the best of our knowledge, the proposed model is the first to account for all four aspects (text and media, as well as geographical and temporal information) using a deep learning approach.

Our approach thus comprises two main components. The first component is a stacked multi-modal autoencoder we developed (Liu and De Sabbata, 2019b) to create dense representations of text and image content from social media posts. The second component extends earlier work on a spatial graph convolutional network (Liu and De Sabbata, 2019a), originally developed to explore how the spatial component of social media posts benefit the labelling (i.e., semantic categorisation) of their contents, into a spatio-temporal graph convolutional network, which encodes the geographical and temporal proximity relationship between social media posts.

In this paper:

1. we explore the effects of accounting for the spatio-temporal aspects of social media posts in using a graph convolutional network to classify them;
2. we provide robust and detailed evaluation of the presented framework through a series of comparisons between different set-ups and other baseline machine learning approaches;
3. we explore the effects of imbalanced datasets on the labelling accuracy and evaluate how data uncertainty (e.g., variations in the number of cases in each label category) affects the labelling results.

## 2 Related Work

Due to the potential of social media platforms for exploring human activities in space and the narrative of places (Abernathy, 2016), social media platforms in general, and Twitter in particular, have been at the centre of data-driven analysis in GIScience and quantitative geography for about a decade (Miller and Goodchild, 2015) from location mining to residence location prediction (Resch et al., 2015; Abrol et al., 2012; Dan et al., 2014). Lee et al. (2011b) proposed a geo-social event detection system to detect unusual regional social activities based on geotagged Twitter messages and identified a strong point of connection between local events and crowd behaviours detected from posts shared on the platform. Tsou et al. (2013) showed the effectiveness of using combined content from tweets and multiple web sources to identify relevant spatial and temporal patterns regarding specific events. Studies by Longley and Adnan (2016) and Wakamiya et al. (2011) also illustrate how the spatial analysis of social media progresses our understandings of socio-spatial patterns and crowd behaviour within cities.

Despite the growing popularity of visual content in social media, limited work has been done so far on such content within the field of GIScience. That is a severe limitation, as

image content is a key component of social media posts – especially considering the rise of image-focused platforms such as Instagram or Flickr. According to a recent survey[1], images or photos constitute around 36% of posts on Twitter, which renders the analysis of visual data, an interesting area to explore. Visual content can also provide rich information regarding places, the use of space, and people's experiences of landscape. Earlier work on the visual content of geo-located media mostly focused on tags or meta-data. For instance, Hollenstein and Purves (2010) investigated the use of geo-located photos from Flickr that users have tagged with keywords such as "downtown" or "citycentre" to explore the user-defined centre of a city. Hu et al. (2015) introduced a coherent three-layer (data layer, spatio-temporal layer and semantic layer) framework for studying urban areas of interest (AOI) spatially as well as temporally using geotagged photos extracted from social media. The semantic layer of this framework serves the purpose of discovering knowledge from the extracted AOI combining the use of photos (based on a defined similarity matrix) and their tags. Their proposed method quantitatively studies AOI from geotagged photos and contributes a better understanding of how areas of interest form over time. Panteras et al. (2015) developed a social multimedia triangulation process to identify natural disasters using Twitter text content and Flicker image meta-data. With the help from the rising methodologies of traditional machine learning, Gao et al. (2015) proposed a method for geo-located event detection from micro-blogs, they generated an intermediate semantic entity, named micro-blog clique (MC) based on text similarities, image similarities, location similarities and temporal similarities to explore the highly correlated information among the noisy and short micro-blogs. Xu et al. (2017) introduced their framework to detect urban emergency events based on the correlations among images, texts and locations. Previous research mainly focused on using traditional machine training approaches based on low- or mid-level attributes from texts and images – e.g., td-idf for text representation, spatial pyramid image features for image representation.

Current developments in deep learning technologies open new opportunities for bridging the gap, and moving beyond the use of tags chosen by the users and low- or mid- level attributes, and combine high-level representations extracted directly from the visual content (e.g., images). Within the discipline of computer science, some work (Xu et al., 2014; You et al., 2015; Gajarla and Gupta, 2015) has been devoted to using convolutional neural networks to analyse users' sentiment directly from the images posted on social media posts. Attracted by the growing popularity of multimedia content on social media posts, research has been widely conducted from uni-modal analysis to multimedia content study. For example, Cai and Xia (2015) proposed a framework using two separate convolutional neural networks (CNN) to extract representations from text and images from tweets, and feed them into another CNN to predict sentiment on Twitter multimedia content. Chen et al. (2017) trained separated classifiers for images and text simultaneously and proposed a weighted co-training approach for the joint visual-textual sentiment analysis. Mouzannar et al. (2018) combined multiple pre-trained convolutional neural networks that extract representations from raw text and images independently and classify them to identify damage related information. Huang et al. (2018a) proposed a framework based on Deep Canonical Correlation Analysis (Andrew et al., 2013) to jointly learning representations from text and image and further identify event-related tweets.

To our best knowledge, within the discipline of GIScience, with the exception of the work we presented (Liu and De Sabbata, 2019b), Huang et al. (2018b) are the only other

[1]https://www.adweek.com/digital/is-the-status-update-dead-36-of-tweets-are-photos-infographic/

authors to apply this technology on text and images to analyse geo-located social media posts. Huang et al. (2018b) proposed an end-to-end, fully supervised framework to report geo-located flood events using Twitter posts. They adopted two convolutional neural network architectures to extract representations from texts and images, and combine both representations for filtering out flood-related tweets from a massive tweets pool. However, their approach focuses on a binary classification and does not account for the geo-location of social media post in the context of the classification.

# 3 Case study

Our case study includes a geographical analysis of geo-located, multimedia posts on a social media platform, regarding a certain set of topics (e.g., posts about personal life, trending news, or entertainment), in London. The focus is on exploring how users' activities are reflected through their related geo-located social media posts, and how those posts contribute to the digital representation (Graham et al., 2013) of the city. We have collected all geo-located tweets posted within Greater London (UK) between January 1st, 2018 and December 31st, 2018, through the Twitter API[2]. The geo-located tweets we retrieved from Twitter contain both image and text, and precise geographic coordinates. In order to filter bots, we limited the overall amount of tweets per account to 365. The final sample is composed of 16,950 tweets, which clearly would still not be easy to analyse manually, in a tweet-by-tweet manner.

In digital (geographies) studies, it is common to explore a sample of a few hundred tweets and conduct a qualitative content and visual analysis by labelling the sampled tweets through a relatively small number of "codes" (see, e.g., Felt, 2016; Awcock, 2018). The use of larger samples is rare, due to the resources that would be required to engage with tens of thousands (or millions) of posts.

In this scenario, the proposed framework (described in Section 4) would be able to learn the set of labels (i.e., "codes") defined on a small sample by a researcher in digital (geographies) studies and deploy the same labels to a larger dataset. The labelled dataset could then be used for further exploration of the specific topics. For the scope of this paper, we randomly sampled 701 tweets from the dataset, that have been manually labelled as discussed below.

## 3.1 Labelling

We manually labelled (i.e., "coded" or classified) the set of randomly sampled 701 tweets using 11 different labels: Animals, Entertainment, Food, Nature, News, Personal, Places and attractions, Social, Sports, Work and Not informative. These labels are for testing purposes only, and they are not an integral part of the framework, which can instead be used with any set of labels (i.e., "codes") that a researcher might consider relevant for a dataset they aim to analyse.

The Not informative label includes advertisement and other content that was difficult to interpret. The label Personal includes content related to personal daily activities such as shopping or selfies, whereas tweets in the label Social are related to social activities (e.g., parities). The label Work mostly contains tweets related to offices environment or the

[2]https://developer.twitter.com/en/docs.html

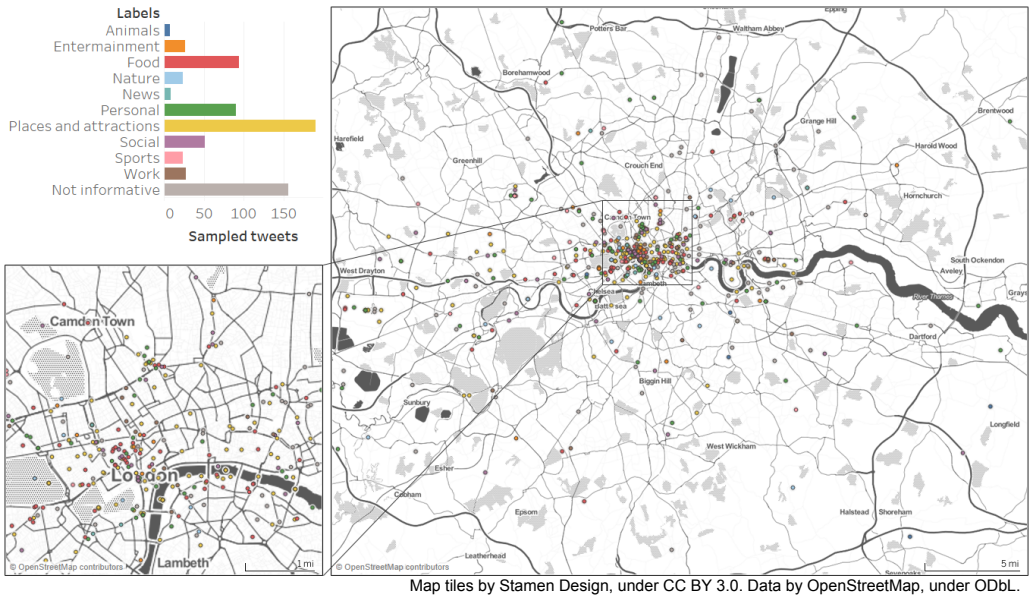



Figure 1: Distribution of labelled tweets used for training and testing.

description of users' work. As illustrated in Figure 1, the sampled dataset is unbalanced as certain labels are more represented than other, for instance, there are 166 tweets regarding Places and attractions while only 8 tweets in the label Animals.

It is important to emphasis that we labelled our social media only based on their text and image content rather than labelling them based on their location or geographic content explicitly. That is, labels used in the case study are not geographical per se. Pre-defined labels that we created is relatively generic, but still very subjective and expressing the interests and understandings of the authors on the classified content. Other authors might prefer to incorporate the label Work into Personal, or clearly differentiate a diverse set of Sports (e.g., football or tennis). However, this is not an issue in the scope of this experiment. The objective of our proposed approach is to provide a framework to classify large volumes of social media posts that is unrealistic to process manually, based on a set of labels tailored to a specific project or task.

Traditional classification tasks in computer science tend to use datasets created using a top-down approach with a set of well-balanced categories as benchmarks, which are optimised to test the effectiveness of new algorithms. Given the aim of our approach, we decided to use a "real-life" dataset, retrieved from Twitter directly, which is much noisier and could be difficult to label even for human assessors. Even for tweets within the same label category, information tends to be much fuzzier compared with datasets used in traditional classification tasks.

As shown in Figure 1, the distribution of the tweets in our dataset is heavily concentrated in the central area of the inner boroughs of London, while only a few tweets are located in the suburban areas of the external boroughs. The impact of this skewed geographic distribution on the creation of the graphs was one of the reasons that led to testing the diverse set of approaches which will be discussed in Section 4.3. Most labels seem to

follow this general pattern, and while some expected cluster can be identified (e.g., Food in Soho, or Nature in Hyde Park), there seems to be no clear-cut geographic clustering of the labels among the 701 sampled tweets.

# 4 Methodology

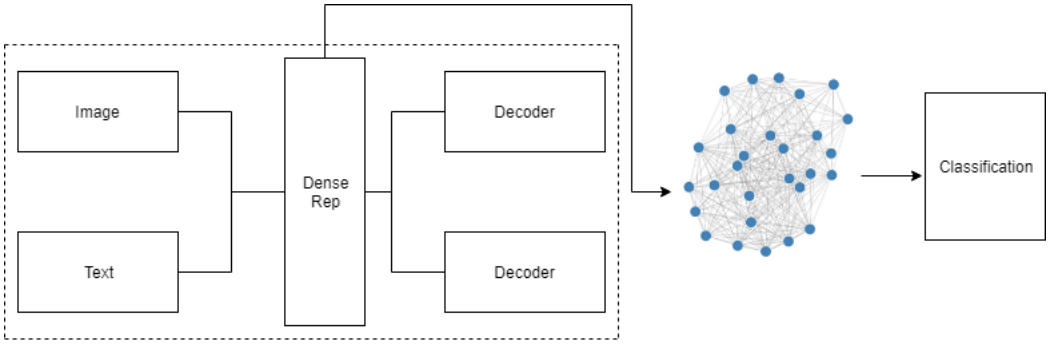



Figure 2: Methodology flowchart. The multi-modal autoencoder extracts dense representations from the images and text; such representations together with the spatial or spatio-temporal components of the posts are fed into a graph convolutional network as its input for the classification task.

In this paper, we propose an approach consisting of two components, illustrated in Figure 2. First, a stacked multi-modal autoencoder model is used to extract dense representations from both texts and images of tweets. Second, graph convolutional network (GCN) is applied based on the graph constructed with geo-coordinates from the social media posts (see details in Section 4.3) to do the semi-supervised classification. That is, the relationship between features and labels is learnt throughout the process and updated based on the information that neighbours exchange with each other. As such, each node of the neural network learns locally, focusing on one social media post. The neural network node works towards understanding the relationship between content and assigned labels in a locally defined subset, taking into account that particular post and all its spatial neighbours. The knowledge acquired locally for each social media post at one layer is added to the information available for that post at the following layer.

We then postulate that the local learning process described above should allow the neural network to take better advantage of spatial clusters of information. In turn, that approach should deliver better performance in understanding labels that are spatially clustered, as it is commonly the case in geo-located social media, which focus on local content.

The model here presented builds upon our previous work (Liu and De Sabbata, 2019b), and it was inspired by approaches such as the Deep Embedding Clustering (DEC) (Xie et al., 2016) and the Correlational Neural Network (Cornet) (Chandar et al., 2016). Our model includes two main components.

## 4.1 Multi-Modal Autoencoder

First, we replace the dense layers in Corrnet with stacked Resnet-style convolution layers for extracting image representations (Mao et al., 2016) and Long Short-Term Memory neural network (LSTM) layers for extracting textual representations. The objective was not only to minimise the self-construction error, but also the cross-reconstruction error from image and texts, and maximise the correlation between the hidden representations of both parts. We achieved this by minimising the objective function introduced in the original Corrnet paper:

$$\mathcal{J}_{\mathcal{Z}} = \sum_{i=1}^{N}(L(z_i, g(h(z_i))) + L(z_i, g(h(x_i))) + L(z_i, g(h(y_i)))) - \lambda corr(h(X), h(Y)) \quad (1)$$

$$corr(h(X), h(Y)) = \frac{\sum_{i=1}^{N}(h(x_i - \overline{h(X)})(h(y_i - \overline{h(Y)}))}{\sqrt{(\sum_{i=1}^{N}(h(x_i - \overline{h(X)})^2(\sum_{i=1}^{N}(h(y_i - \overline{h(Y)})^2}} \quad (2)$$

considering a dataset $\mathcal{Z} = \{z_i\}_{i=1}^{N}$ where all data have inputs from two channels of media text and images $X$ and $Y$. Each data $z_i$ can be represented as $z_i = (x_i, y_i)$, where $x_i \in X$ and $y_i \in Y$. $L$ is the squared error reconstruction error, and $\lambda$ is the scaling parameter. $\overline{h(X)}$ is the mean vector for the hidden representation $h(x_i)$ of the text part input, and $\overline{h(Y)}$ is the mean vector for the hidden representation $h(y_i)$ of the image part input. $h(z) = f(Wx + Vy + b)$, where $W$ and $V$ are two $k \times d_i$ project matrix, and $b$ is a $k \times 1$ bias vector. Equation 1 is the objective function of the proposed multi-modal autoencoder. The first term is the objective function that allows learning meaningful hidden representations. The second term ensures that both images and text output from the decoder can be reconstructed using only text representations. Similarly, the third term ensures that both images and text output from the decoder can be reconstructed using only image representations. The fourth term ensures that the combined representations are highly correlated, and it is defined in Equation 2.

## 4.2 Graph Convolutional Network

The second component of our model encodes the geographies of the content as a graph network, thus allowing us to take advantage of recent advances in graph convolutional neural networks. Indeed, the use of graph networks to represent proximity has long been used in geographic information science (O'Sullivan and Unwin, 2014). In our approach, each social media post is rendered as a node in a graph network, and their proximity is represented by the presence and weight of the edges between nodes, depending on the definition, as detailed in next section.

We then frame our classification problem as a graph-based semi-supervised learning task, and a graph convolution network (Kipf and Welling, 2016) is adopted for efficient information propagation through the graph.

The task is specified as *f(X,A)*, where *X* is the extracted information from the multi-modal autoencoder for each post, and *A* is the adjacency matrix for the graph. We expect the model to produce a node-level output *Z* as:

$$Z = f(X, A) = \textit{softmax}(H^{(L)}) \quad (3)$$

which satisfies the layer-wise propagation rule for GCN:

$$H^{(L+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(L)}W^{(L)}) \tag{4}$$

with $\hat{A} = A + I_N$. $I_N$ is the identity matrix of $A$ and $W^{(L)}$ denotes the trainable weight matrix of the $L$th layer of the neural network. $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$, and $\sigma(\cdot)$ represents a non-linear activation function which in our case we use $ReLu(\cdot) = max(0, \cdot)$. $H^{(L)}$ is the activation matrix for the $L$th layer; for example, $H^{(0)} = X$ and $H^{(L)} = \hat{A}ReLu(H^{(L-1)})W^{(L)}$. The softmax activation in formula (1) is used for classifying nodes with their corresponding labels. We calculate the cross-entropy error over all labelled nodes in the graph:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} \mathcal{Y}_{lf} \ln Z_{lf} \tag{5}$$

where $\mathcal{Y}_L$ is the set of nodes that have labels.

## 4.3 Graph Construction

We tested a variety of graphs that were constructed using the tweets presented in the case study. We classified the graphs based on whether they account for the absolute positions of tweets and distances between the tweet pairs into three different categories: *a-spatial graphs*, *semi-spatial graphs* and *spatial graphs*.

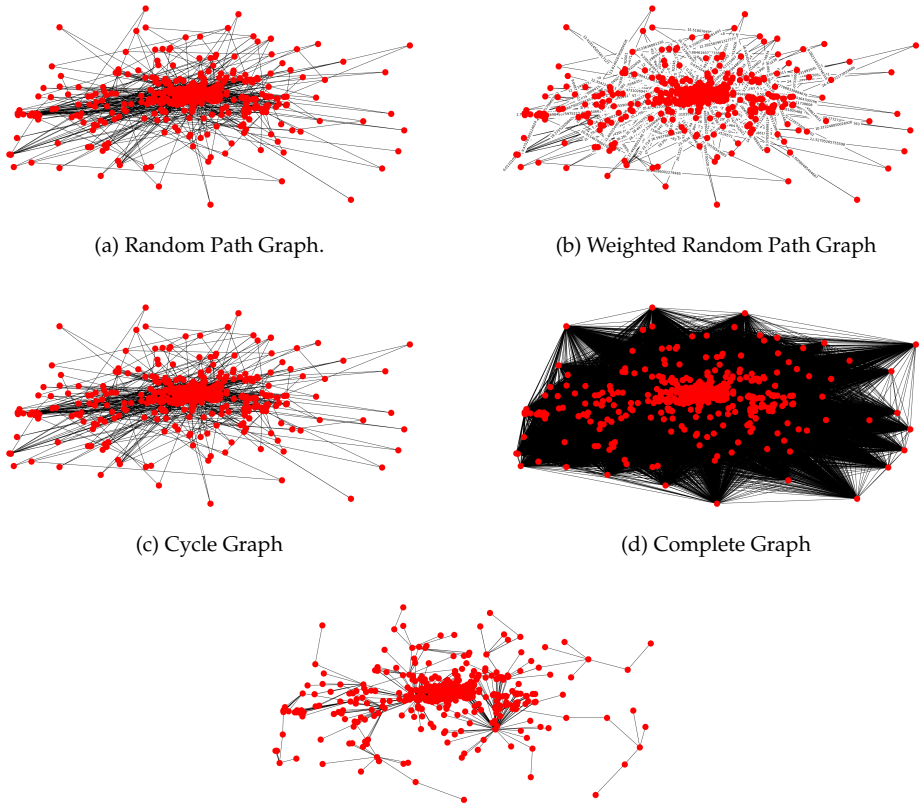
In constructing the graphs, we employ Euclidean distance as a reasonable estimation of distance in urban spaces, as illustrated by Boscoe et al. (2012).

### 4.3.1 A-spatial Graphs

*A-spatial graphs* do not take into account the absolute positions of the tweets and distances between the pairs of nodes in the graph. We tested three different *a-spatial graphs* in the experiments:

- **Random Path Graph:** A path graph is a graph that can be drawn so that all of its vertices and edges lie on a single straight line (Gross and Yellen, 1999). We randomly assign tweets in a line so that they are linked to each other one by one, as shown in Figure 3(a). If two nodes are connected to each other $A_{ij} = 1$ in its adjacency matrix, otherwise $A_{ij} = 0$.
- **Random Cycle Graph:** A cycle graph is a graph containing a single cycle through all nodes shown in Figure 3(d). It is randomly generated in the same way as the Random Path Graph, plus adding a link between the beginning and the end nodes.
- **Complete Graph:** A complete graph is a graph in which each pair of graph vertices is connected by an edge shown in Figure 3(c).

Note that although *Random Path Graph*, *Cycle Graph* and *Complete Graph* are connected in the form that each tweet is connected to another, such graphs do not take into account the absolute positions of the tweets and distances between the pairs of nodes in the graph. For example, *Random Path Graph* is constructed by connecting all the nodes in the graph with a straight line. It can start from any arbitrary node as long as all the nodes can lie on the same line by the end of the graph construction. Thus, the absolute positions of tweets are



(a) Random Path Graph.

(b) Weighted Random Path Graph

(c) Cycle Graph

(d) Complete Graph

(e) Minimum Spanning Tree

Figure 3: Different graph structures

not useful in such a case. Therefore, such graphs can be seen as the nodes are connected without the spatial component, whereby the spatial locations of tweets have no impacts in those graph construction processes.

#### 4.3.2 Semi-spatial Graphs

*Semi-spatial graphs* do not take into account the absolute positions of the tweets but are constructed with the information of the distances between the pairs of nodes in the graph. Following the experiment in section 4.3.1, we tested three different *semi-spatial graphs* in the experiments:

- **Weighted Random Path Graph:** Same structure as *Path Graph* shown in Figure 3(b); however, the weights for edges are defined by spatial interaction as:

$$A_{ij} = 1/(1 + distance) \tag{6}$$

- **Weighted Random Circle Graph:** Same structure as Cycle Graph; however, the weights for edges are defined by spatial interaction.

- **Weighted Complete Graph:** Same structure as Complete Graph, but the weights for edges are defined by spatial interaction.

#### 4.3.3 Spatial Graphs

*Spatial graphs* take into account the absolute positions of the tweets as well as the information of the distances between the pairs of nodes in the graph. We tested two *spatial graphs* as follow:

- **Minimum Spanning Tree (MST):** We first generate a series of graphs based on spatial adjacency using distances ranging from 2 kilometres to 15 kilometres. We then calculate the minimum spanning tree for each one of those graphs to further minimise the number of connections. Figure 3 (e) is an example of a Minimum Spanning Tree calculated starting from the 9 kilometres spatial adjacency. If two nodes are connected to each other $A_{ij} = 1$ in its adjacency matrix, otherwise $A_{ij} = 0$.
- **Weighted Minimum Spanning Tree (Weighted MST):** Same structure as Minimum Spanning Tree, but the weights for edges are defined by the same spatial interaction defined in Section 4.3.2 (Equation (6)).

### 4.4 Spatio-Temporal graph

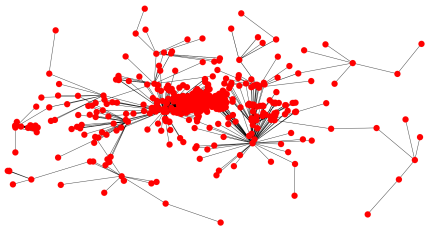
The temporal component of social media post (Yang and Leskovec, 2011) is a key aspect to take into account to move beyond the simple geotag (Crampton et al., 2013a). The temporal evolution of the social media trend has clear links to emerging events in the physical world (Wang et al., 2016a), which leave "data shadows" behind them (Shelton et al., 2014). Spatio-temporal analysis has been widely adopted in the study of digital geographies (Cheng and Wicks, 2014; Gomide et al., 2011; Lee et al., 2011a), to identify sociospatial patterns of online events (Crampton et al., 2013b; Luo et al., 2016), or to monitor and surveillance nature disasters (Wang et al., 2016b; Martín et al., 2017). To explore the usefulness of the temporal component of social media posts in better understanding its relationship with the assigned labels, we tested two graphs based on two different spatio-temporal distances and a weighted graph with distance information on the edges.

- **Spatio-temporal neighbourhood (StN), Euclidean Distance:** To be consistent with spatial distance, we transform the time series information equivalent to the spatial distance, and we define such a process as *temporal-spatial distance transformation*. That is, the temporal differences between social media posts are measured in a defined spatial distance (see details in Section 5.2). We define the first spatio-temporal distance as:

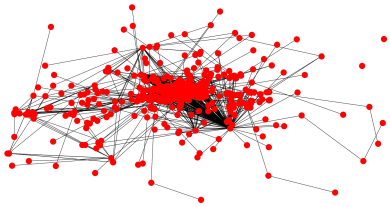
$$STDist = \sqrt{easting_{dist}^2 + northing_{dist}^2 + time\_series_{dist}^2} \tag{7}$$

  where distance is calculated after projecting the longitude and latitude of each post to the respective values as $easting$ and $northing$ in the British National Grid[3]; $time\_series_{dist}$ denotes to the defined *temporal-spatial distance transformation*, for example, $time\_series_{dist}$ = 1 $metre$ if the temporal difference between two posts is 12 hours. An example of the constructed graph using Euclidean distance is shown in

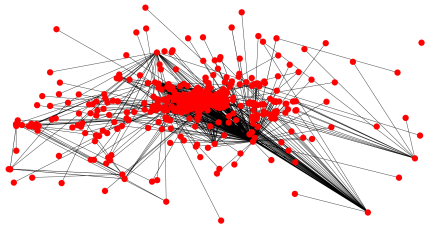
[3]https://epsg.io/27700



(a) Spatio-temporal neighbourhood graph, Euclidean Distance (9 km).



(b) Spatio-temporal neighbourhood graph, distance defined in Chang et al. (2007) (9km)



(c) Spatio-temporal neighbourhood graph, temporally-weighted Euclidean Distance (9 km).

Figure 4: Different spatio-temporal graph structures

Figure 4(a). We ran a series of experiments using different distances to equate time and space, and the results are presented in Section 5.2.

- **Spatio-temporal neighbourhood (StN), temporally-weighted Euclidean Distance:** Chang et al. (2007) define a spatio-temporal similarity measure to compute spatio-temporal relevance between two trajectories of moving objects on road networks, which is known as spatio-temporal distance:

$$STDist = (SD + \delta * TD)/2 \tag{8}$$

where $\delta$ is the spatio-temporal weight; *SD* and *TD* denote for spatial distance and temporal distance respectively. An example of using such a distance can be seen in Figure 4(b). As each entity in our dataset represents a point in the space-time continuum, rather than a trajectory, we propose the following definition of the distance between two points into:

$$STDist = \sqrt{SD^2/2 + (\delta * TD)^2/2} \tag{9}$$

where *SD* is defined as $\sqrt{easting_{dist}^2 + northing_{dist}^2}$. It is a variation on the Euclidean distance, but taking into account an additional spatial weight $\delta$ defined in formula (8) to define the impact of the temporal distance. In this paper, we keep $\delta$ as 20 same in Chang et al. (2007), and the results are presented in Section 5.2. Thus, we define such approach as *Temporal weighted Euclidean Distance*. An example of the constructed graph using temporal weighted Euclidean distance is shown in Figure 4(c).

- **Spatio-temporal neighbourhood (StN), distance and temporally weighted Graph:** Given the best results reported in Section 5.1 is achieved by using graph defined

by Equation 9, we define this graph as same structure as the temporally weighted Euclidean Distance model, but the weights for edges are defined by spatial-temporal interaction as:

$$A_{ij} = 1/(1 + distance_{ST}) \tag{10}$$

where $distance_{ST}$ is the distance calculated by Equation (9).

## 4.5 Baseline methods

In order to test the capability of our proposed semi-supervised multimedia classification framework, We compare it with eight baselines developed from various methods focusing on text content and image content, as well as the spatial component of the tweets:

### 4.5.1 A-spatial Baselines with Text and Images

We set up baselines which include a traditional machine learning algorithm and two neural network-based on deep learning approaches to compare their performance with our proposed graph-based semi-supervised classification framework.

- **SVM**: We adopt a traditional machine learning approach Support Vector Machine (SVM) (Cortes and Vapnik, 1995) on the extracted representations from multi-modal autoencoder to classify tweets. Traditional machine learning methods such as SVM has a long-standing history being adopted for social media classification and spatial analysis within the field of geography (e.g., —-see, Guo and Chen (2014), Qi et al. (2019)). Although recent research shows that deep learning methods outperform that traditional machine learning method in various disciplines, SVM is still worth to be set up as a basic baseline in comparisons with the proposed GCN framework due to its popularity within academic studies.
- **Dense Neural Network (DNN)**: We adopt a 3-layer dense neural network (DNN) on the extracted representations from multi-modal autoencoder to classify tweets. Due to its strong ability of generalisation, DNN as one type of deep learning techniques has been widely adopted in various social media analytic studies (—-see, a survey was done by Ghani et al. (2019)). Thus, the 3-layer DNN is chosen as another baseline.
- **Visual-textual Fused CNN (VTCNN)**: Inspired by Huang et al. (2018b), we design an end-to-end deep learning framework using two stacked CNN to extract representations from images and text simultaneously and concatenate them in the middle layer of the framework for twitter classification. Huang et al. (2018b) can be seen as a direct comparison to our proposed framework, although such a method is primarily a supervised training framework which usually requires training data in a large size and to be well-labelled.

Note that the baseline VTCNN is the only end-to-end training framework among all the baselines. For other baselines introduced in this section, representation extraction (from images, text or both) and classification are two separated steps. It is also important to highlight that these three baselines do not take into account the locational information of the tweets, and they perform the labelling process purely based on the multimedia content (images and text) from tweets. Thus, they are set up as comparisons for the version of our framework set up on *a-spatial graphs* introduced in Section 4.3.1.

#### 4.5.2 A-spatial Baseline with Text Only

**Doc2Vec + Label Propagation**: We use Doc2Vec (Le and Mikolov, 2014) to extract text representation and a traditional semi-supervised machine learning approach Label Propagation (LP) (Zhu and Ghahramani, 2002) to classify tweets. Such a method has been widely adopted on online content analysis (e.g., sentiment analysis (Mishra et al., 2019; Wadawadagi and Pagi, 2020)). As a semi-supervised machine learning approach, LP is used to assess the performance of the GCN framework, which is also a semi-supervised learning framework. Such a framework aims to classify posts based on their text content; thus, it can be used for demonstrating whether the multimedia content analysis is superior to content analysis which only using text.

#### 4.5.3 A-spatial Baseline with Images Only

**CNN autoencoder + Label Propagation**: We use a CNN autoencoder (Mao et al., 2016) which is the same structure as we adopted in the multi-modal autoencoder to extract image representation, and use LP approach to classify tweets.

#### 4.5.4 Spatial Baselines with Text Only

**Doc2Vec + GCN (MST)**: We use Doc2Vec to extract text representation and GCN on a spatially constructed graph (MST) to classify tweets. This baseline is set up as a direct comparison to the previous baseline (**Doc2Vec + Label Propagation**).
**LSTM autoencoder + GCN (MST)**: We use an LSTM autoencoder to extract text representation and GCN on a spatially constructed graph to classify tweets. Although the two previous baselines have been designed to illustrate classification results based only on text, Doc2Vec is not considered as a deep learning approach to extract text representation. As mentioned in Section 4.1, our proposed multi-modal autoencoder contains an LSTM encoder extracting text representations from UGC content. Thus, we further design such a baseline as one of the comparisons to assess whether the multimedia content analysis is superior to content analysis which only using text.

#### 4.5.5 Spatial Baselines with Images Only

**CNN autoencoder + GCN (MST)**: We use the CNN autoencoder to extract image representation, and GCN on a spatially constructed graph (MST) to classify tweets. This baseline and the previous baseline (**CNN autoencoder + Label Propagation**) are designed to assess whether the multimedia content analysis is superior to content analysis which only using images.

### 4.6 Model training

The image and text content of each tweet was pre-processed. Tokenisation, stop words removal, and case folding was applied to the text, then vectorised using Word2Vec-Twitter[4]. Images were converted to greyscale and re-sized to a $158 \times 158$ uniform size. A sample of cases was selected (as discussed below) to train the GCN, ensuring that at least four tweets from each label were selected.

[4]https://github.com/loretoparisi/word2vec-twitter

We implemented the model in Python, using Keras[5] with Tensorflow[6], as a two-layer GCN model with 0.5 dropout rate for both layers, $L2$ regularisation factor for the first GCN layer and 8 hidden units. We trained the model using a Nivida GPU Geforce GTX 1080[7] for a maximum of 3000 epochs (training iterations) using Adam (Kingma and Ba, 2014) with a learning rate of 0.01, and early stopping with a window size of 300. Trainable weights initialisation and feature vectors normalisation remain the same, as in Kipf and Welling (2016). We evaluated it using both classification accuracy and F1 score.

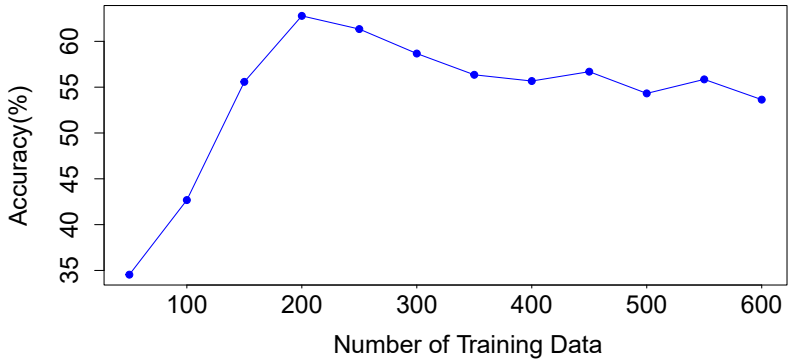



Figure 5: Variation of accuracy based on the number of training data.

We explored the impact of different training sample sizes on the performance of the model by testing randomly selected training sets with increasing sizes from 50 to 600, incrementing the sample size by 50 at each step and using the Random Path Graph structure. As illustrated in Figure 5, model performance peaks at 200, with a slight decrease afterwards, while accuracy tends to be stable. Our interpretation is that, while fewer than 200 cases are insufficient to train the model adequately, larger samples tend to become too unbalance, which affects the performance.

# 5 Results

## 5.1 Spatial graph

The experiments started by focusing on GCN with a-spatial graph structures with no defined spatial interaction (see Section 4.3.1, Equation (6)), and the assessments are based on the classification accuracy and $Micro - F1$ score. The $Micro - F1$ score is a type of $F1$ score suitable for multi-label classification tasks aiming to provide robust evaluations for machine learning or deep learning models compared to accuracy-based assessments.

As summarised in Table 1, the results reveal that our best GCN approach successfully categorises each tweet with its corresponding label based on partially labelled data with an accuracy of $72.57\%$. Figure 6 shows how the manually assigned labels compare to the model output, and it illustrates how most of the errors are due to some *Food* and most of

[5]https://keras.io/
[6]https://www.tensorflow.org/
[7]https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1080/

| Model input | Representation Extractor | Model | Accuracy | Micro-F1 Score |
|---|---|---|---|---|
| A-spatial with Images and Text | Multi-modal Autoencoder | SVM (no graph structure) | 15.87% | 9.13% |
| | Multi-modal Autoencoder | DNN (no graph structure) | 11.20% | 4.35% |
| | (VTCNN itself) | VTCNN (no graph structure) | 16.00% | 8.37% |
| | Multi-modal Autoencoder | GCN (Random Path Graph) | 62.78% | 56.87% |
| | Multi-modal Autoencoder | GCN (Cycle Graph) | 68.63% | 65.94% |
| | Multi-modal Autoencoder | GCN (Complete Graph) | 23.75% | 15.65% |
| A-spatial with Text | Doc2Vec | Label Propagation | 18.31% | 3.40% |
| A-spatial with Images | CNN autoencoder | Label Propagation | 26.76% | 4.20% |
| Spatial with Text | Doc2Vec | GCN (MST (3 km)) | 26.43% | 24.32% |
| | LSTM autoencoder | GCN (MST (3 km)) | 36.66% | 35.95% |
| Spatial with Images | CNN autoencoder | GCN (MST (3 km)) | 71.07% | 70.51% |
| Semi-spatial with Images and text | Multi-modal Autoencoder | GCN (Weighted Random Path Graph ) | 65.34% | 63.15% |
| | Multi-modal Autoencoder | GCN (Weighted Cycle Graph) | 68.83% | 67.67% |
| | Multi-modal Autoencoder | GCN (Weight Complete Graph) | 23.66% | 18.15% |
| Spatial with Images and text | Multi-modal Autoencoder | GCN (MST (2 km)) | 56.73% | 51.89% |
| | Multi-modal Autoencoder | **GCN (MST (3 km))** | **72.57%** | **69.10%** |
| | Multi-modal Autoencoder | GCN (MST (5 km)) | 61.60% | 57.83% |
| | Multi-modal Autoencoder | GCN (MST (8 km)) | 55.55% | 52.24% |
| | Multi-modal Autoencoder | GCN (MST (10 km)) | 54.67% | 48.67% |
| | Multi-modal Autoencoder | GCN (MST (15 km)) | 51.64% | 47.25% |
| | Multi-modal Autoencoder | **GCN (Weighted MST (3 km))** | **73.57%** | **72.89%** |
| Spatio-temporal with Images and Text | Multi-modal Autoencoder | GCN (StN (Euclidean, 2 km)) | 67.33% | 64.53% |
| | Multi-modal Autoencoder | GCN (StN (Euclidean, 3 km)) | 70.28% | 68.45% |
| | Multi-modal Autoencoder | GCN (StN (Euclidean, 4 km)) | 69.58% | 67.25% |
| | Multi-modal Autoencoder | GCN (StN (Euclidean, 5 km)) | 69.15% | 66.73% |
| | Multi-modal Autoencoder | GCN (StN (as defined in Chang et al. (2007), 2 km)) | 63.24% | 60.24% |
| | Multi-modal Autoencoder | GCN (StN (as defined in Chang et al. (2007), 3 km)) | 66.57% | 63.83% |
| | Multi-modal Autoencoder | GCN (StN (as defined in Chang et al. (2007), 4 km)) | 69.89% | 65.27% |
| | Multi-modal Autoencoder | GCN (StN (as defined in Chang et al. (2007), 5 km)) | 69.24% | 67.51% |
| | Multi-modal Autoencoder | GCN ((StN (temporally-weighted, 2 km)) | 69.58% | 65.32% |
| | Multi-modal Autoencoder | GCN (StN (temporally-weighted, 3 km)) | 72.32% | 69.68% |
| | Multi-modal Autoencoder | **GCN (StN (temporally-weighted, 4 km))** | **78.98%** | **76.72%** |
| | Multi-modal Autoencoder | GCN (StN (temporally-weighted, 5 km)) | 74.56% | 71.41% |
| | Multi-modal Autoencoder | **GCN (StN (distance-temp.-weighted, 4 km))** | **80.08%** | **78.65%** |

Table 1: Comparisons of different graph structures. (Best results achieved.)

*Nature* tweets being labelled as *Personal* by the model, many *Sports* tweets being labelled as *Places and attractions*, and most *Work* tweets being labelled as *Not informative*. Errors do not seem to display any specific spatial pattern.

Those results are achieved on a training sample of 200 randomly selected tweets and despite a fairly imbalanced and noisy dataset. The GCN seems to perform better on a sparse – non necessarily simple, but less dense graph structure, as the best results are obtained with a graph structure constructed by creating Weighted Minimum Spanning Tree using a 3 kilometres range, whereas the classification accuracy and F1 score on the two complete graphs are much lower compared to the other spatial graph structures. Clearly, choosing a suitable distance range for creating graph structure is essential within the framework. The results show that identifying a geographic graph with an appropriate density of connections within a reasonable distance range can significantly improve the performance of our graph-based semi-supervised framework.

As shown in the table, GCN on the spatial graph constructed using Minimum Spanning Tree with 3 kilometres radius achieves the best results among other structures; we choose the same distance radius for constructing the Weighted Minimum Spanning Tree. The results show an even better accuracy of $73.57\%$, and it illustrates that knowing local context (i.e., tweets posted nearby) can help our framework to understand content better.

Based on whether the models account for the absolute positions of tweets and spatial distances between the pairs of nodes in the graph or not, we classified the graphs into three

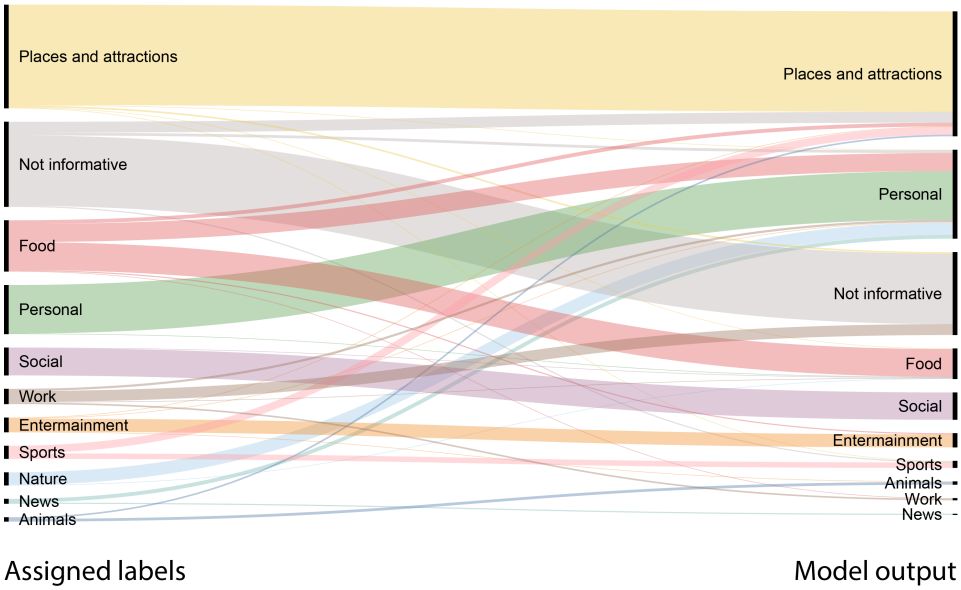



Figure 6: Comparing manually assigned labels and the output from the model based on a Minimum Spanning Tree (3 km).

major categories as introduced in Section 4.3 and shown in Table 1: *a-spatial graphs*, *semi-spatial graphs* and *spatial graphs*. It is evident that the more abundant spatial information the model has, the higher performance the GCN can achieve. GCN on Weighted Minimum Spanning Tree is clearly higher than the results achieved by GCN on the semi-spatial graphs (weighted Random Path Graph, weighted Cycle Graph and weighted Complete Graph), and the results achieved by GCN on the a-spatial graphs (Random Path Graph, Cycle Graph and Complete Graph).

It is also interesting to see that the results achieved by GCN are significantly higher than the traditional supervised learning method SVM. The latter used features exacted from the stacked multi-modal autoencoder and their corresponding labels, but it didn't account for the geographies of the posts. As already mentioned, labels have not been assigned based on geography or location, but solely based on content, which is the information provided to the SVM. Similarly, our proposed framework outperforms the two deep learning methods DNN and VTCNN. It is, however, important to highlight that the latter two were originally designed for supervised learning tasks with large and well-defined training data. In the context of our task, those two frameworks tend to overfit when trained on a relatively small and noisy sample. That is instead not a problem for the GCN framework.

These findings are particularly interesting, as they attest to the relevance of the geographies of information in analysing social media content. By propagating information through a geographically-constructed network, our approach seems to be able to exploit the information implicitly encoded by the location of the posts and the resulting proximity and clustering of content. As such, the GCN approaches on spatial graphs are clearly superior to the a-spatial and semi-spatial models, including models that use random or complete graphs.

Additionally, following our experiment design in Section 4.5, Table 1 also summarises the results of the baseline methods which only adopt text content and image respectively rather than using combined representations for the classification. As GCN with spatial graph constructed using Minimum Spanning Tree with 3 kilometres as radius achieves reasonably good labelling, we implement the same settings for GCN models in the baseline experiments. The result shows that the GCN model outperforms the traditional machine learning semi-supervised approach Label Propagation. Also, the labelling solely relying on text content proves to be unreliable with comparably low accuracy. Furthermore, although the labelling based on image content achieves worse results compared with multimedia content, it produces a competitive labelling output with relatively high accuracy and F1 score.

Those results particularly interesting from a social science perspective, as it proves the evidence that visual content offers richer complementary information than what the accompanying text reveals (Borth et al., 2013), and the image content of tweets dominates human judgment at the labelling stage.

## 5.2 Spatio-Temporal graph

As mentioned in Section 4.4, to be consistent with spatial distance, we transform the time series information equivalent to the spatial distance. In Table 1, we summarise the results of experiments on spatio-temporal graphs using 10 $meters = 12$ $hours$ (see the paragraph after next for further discussion). The topological structure using Minimum Spanning Tree based on temporal weighted Euclidean distance with a 4 kilometres radius achieves the best results for both accuracy (78.98%) and F1 score (76.72%), which are significantly higher than the results achieved by GCN on the graphs merely with spatial information. Further performance improvement is obtained when applying the Weighted Minimum Spanning Tree using the same distance radius ($80.08\%$ accuracy and $78.65\%$ F1 score). The performance is superior compared with the results achieved by spatial graphs discussed above.

The findings also illustrate that despite the variation of the graphs constructed using different types of spatio-temporal distance and distance radius, the results achieved prove to be relatively stable with higher accuracy and F1 score comparing with spatial graphs. These findings are interesting from spatio-temporal analysis perspective, as they illustrate that adding the temporal component of tweets can help the GCN model to produce better labelling (i.e., semantic categorisation) on their multimedia contents.

We also designed further experiments using different temporal-spatial distance transformations on the graphs, and explore their impacts on classification accuracy. As shown in Table 1, the best results for graphs constructed using Spatio-Temporal Euclidean Distance and Temporal Weighted Euclidean Distance are achieved with a radius equal to 3 and 4 kilometres. As such, we use 3 and 4 kilometres as default radius to construct graphs respectively, for these two approaches. Table 2 shows the results obtained on different temporal-spatial distance transformations including 1 $metre = 12$ $hours$, 10 $metres = 6$ $hours$, 10 $metres = 8$ $hours$, 10 $metres = 12$ $hours$ and 10 $metres = 24$ $hours$. These test allowed us to test different temporal "localities" and how they compare against spatial "localities" in capturing events and spatio-temporal patterns. The results indicate that 10 $metres = 12$ $hours$ perform best in the context of our dataset.

As shown in Table 1 and Table 2, GCN performs best when used in combination with our proposed spatio-temporal weighted distance. As we discussed in Section 4.4, the dis-

| Transformations | Spatio-temporal Euclidean Distance | Temporal Weighted Euclidean Distance |
|---|---|---|
| 1m = 12hr | 60.53% | 65.56% |
| 10m = 6 hr | 65.47% | 72.08% |
| 10m = 8 hr | 68.23% | 75.82% |
| **10m = 12 hr** | **70.28%** | **80.08%** |
| 10m = 24 hr | 68.85% | 77.07% |

Table 2: Comparisons between different temporal-spatial distance transformations. (Best results achieved.)

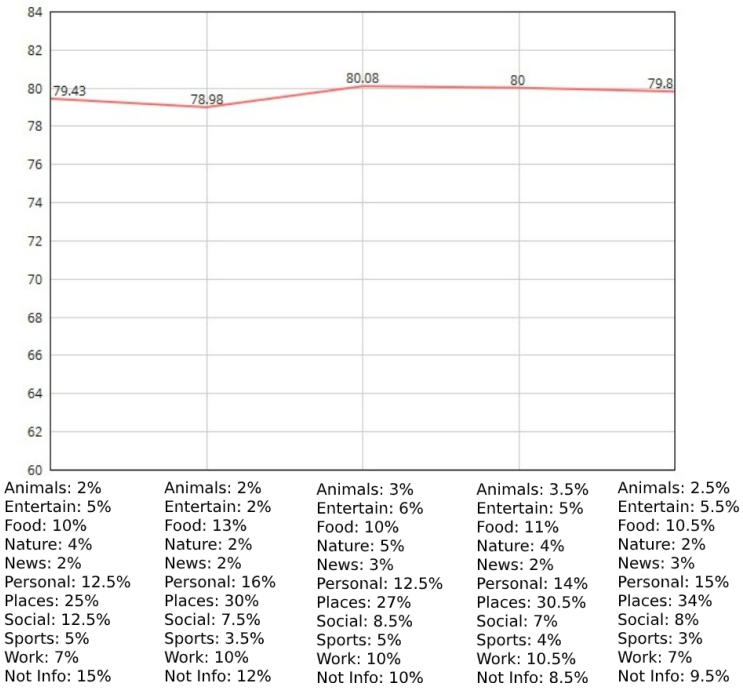



Figure 7: Performance comparisons using different training data.

tance proposed by Chang et al. (2007) was originally devised for analysing trajectory data rather than social media posts. Further extensive research into such spatio-temporal modelling issues is clearly needed. Despite the fact that there is a wide literature focusing on spatio-temporal analysis of social media data, we argue our paper is the first to embed spatio-temporal distance in a deep learning approach to achieve a semantic understanding of content. How to best model spatio-temporal distance in this context is an interesting research area that we hope to explore further in our future work.

## 5.3 Framework Robustness

As mentioned in Section 3, the dataset used for the experiments presented in this paper is noisier and more imbalanced than classic benchmarks used in traditional classification tasks. It is therefore important that we explore the effects of such imbalances on the classi-


a) **Not informative**

The tweet clearly states the name "Hyde Park" but in the attached photo a bicycle is only partially visible

b) **Food**

The tweet is very short but includes the word "veggy" and the name "Greater London" the photo shows a sandwich, a cup and a sweet

c) **Personal**

An institutional accunt wishing happy Easter and including a cartoon of an egg as image

d) **Places and attractions**

The tweet clearly states that the user is at the ZSL London Zoo and the photo shows a board with the map of the zoo

Figure 8: Results of the prediction test on a further unlabelled sample.

fication task, and evaluate the robustness of our framework against variations in training data.

Therefore, we design an additional experiment using five different samples from our datasets. Each sample has at least four tweets for each label, but the proportion of tweets assigned with the different labels is slightly adjusted. The experiment is conducted using the best performing approach in Table 1, that is a weighted graph constructed using the temporal weighted Euclidean distance with Weighted Minimum Spanning Tree (4 km).

The results are shown in Figure 7. Although model performance is slightly affected by the variation in the sample, the classification results are fairly consistent and stable. The results illustrate the robustness of our proposed framework on heavily imbalanced datasets such as "live" social media streams, and thus its relevance for applications in digital geographies.

## 5.4 Showcases

In order to showcase the capabilities of our framework, we briefly include two showcases. First, Figure 8 illustrates how the model trained for the case study above can be used to

| Model | Accuracy | Micro-F1 Score |
|---|---|---|
| GCN (StN(distance-temp.-weighted, 4km) | 78.50% | 76% |

Table 3: Showcase of the framework on a new dataset

classify a further sample of unlabelled data (1200 tweets) with a spatially constructed graph using Minimum Spanning Tree (3 km). As discussed above, the classification is noisy and, for instance, the tweet in Figure 8a is labelled as Not informative, as the model struggles to reconcile the location in a park and a text that indicates being at an attraction with the image of a bicycle. At the same time, the remaining three tweets indicated in Figure 8 seem to have been assigned a fairly accurate label among those we defined for our case study and considering that the aim of the tool is to allow users to define their own labels.

Second, we extracted a new sample of 200 tweets and labelled them following the same labelling procedure introduced in Section 3.1. The new dataset has 30 tweets that are labelled as *Not Informative*, 2 tweets as *Animals*, 52 tweets as *Places and Attractions*, 43 tweets as *Personal*, 50 tweets as *Food*,1 tweet as *News*, 3 tweets as *Entertainment*, 3 tweets as *Nature*, 8 tweets as *Social*, 4 tweets as *Sports* and 4 tweets as *Work*. We then used the model described above using a spatio-temporally constructed graph using Minimum Spanning Tree (4 km) to predict the labels of the new sample, obtaining the results presented in Table 3.

# 6 Discussion

In the sections above, we introduced a semi-supervised learning framework based on geographic adjacency networks to label social media posts based on their textual and visual content, as well as spatial and temporal aspects. The results demonstrate that taking into account the geography of each post is crucial to achieving a semantic understanding of content and enable labelling. In particular, while the labels used in the experiment were not assigned based on the location of social media posts, spatially-enabled classifiers performed better than a-spatial ones. The temporal component was also established as a key aspect in encapsulating the concept of place, and taking into account spatio-temporal relationships between social media post led to better labelling. The results show that our framework can produce good labelling results with partially labelled data, even on noisy and imbalanced data such as the one used for the case study presented above. Although we used Twitter as our case study, our framework has the flexibility to be extended to any other social media platform providing location-based services. As such, our approach has the potential to be developed into a flexible tool for the study of digital geographies.

The majority of quantitative research on social media analysis in geography focuses on the text, whereas qualitative research maintains the importance of visual content (Ash et al., 2018). As such, we based our work on the assumption that including the visual component of a post provides key information in understanding its content. To test that assumption we designed a set of experiments to compare the labelling resulting from including both text and images, only text and only images, using our GCN model, as well as the semi-supervised approach Label Propagation (Zhu and Ghahramani, 2002) as our baseline. The outcomes show that GCN provides the best results on Weighted Minimum Spanning Tree (3 km), which takes into account the geographies of social media content (more on this below), as well as text and image. That indicates that including both the textual and media

component improves the labelling results compared to traditional text-based social media analysis, confirming our assumption above. These results are particularly important in a time where visual content such as images have become an integral and growing part of social media communication, as users shift from text-based posts to multimedia content (Weller et al., 2014). By taking advantage of recent developments in deep learning technologies, our paper is a first step towards bridging the gap between text-based quantitative analysis and visual methodologies in digital geographies.

To explore how to best encode the spatial information of social media posts in our model, we tested the effect of different graph structures on the performance of the GCN. We implemented our proposed GCN model on different structures, from semi-spatial graphs (e.g., Weighted Path Graph and Weighted Complete Graph) to complex structures taking different approaches to encoding the geographies of posts as network links and distances. The results show that constructing a geographic graph taking into account distances between posts and with an appropriate density of connections (e.g., Minimum Spanning Tree) can significantly improve the performance of our graph-based semi-supervised framework compared to random or complete graph structures. The performance of our model is clearly superior to the traditional machine learning approach SVM (Cortes and Vapnik, 1995), which does not take into account spatial graph structures, and classifies tweets based solely on the extracted feature representations. The comparison with the results obtained by GCN on three a-spatial graphs (i.e., Random Path Graph, Cycle Graph and Complete Graph) demonstrate that a graph-based deep neural network which takes into account the geographies of social media post provides not only better labelling results compared to traditional machine learning methods, but also better results compared to itself on the graphs with no geographies encoded in. Furthermore, the outcomes obtained by using different spatial graphs demonstrate that selecting an appropriate spatial (topological) structure can significantly improve the labelling results.

The results ultimately highlight the importance of understanding social media content geographically. The geotag specifying the geographic location of a post is not merely a point, but it is an integral part of the augmentations that bring the place into being(Graham et al., 2015). As such, taking into account the spatial relations between posts via the convolution of content through the spatial graphs allows us to go beyond the geo-tag(Crampton et al., 2013a), and provides the GCN with key contextual information, that is crucial in the semantic understanding of social media content and thus the digital representations of the city (Ballatore and De Sabbata, 2020)

However, places do not merely exist in space, but they are "specific time-space configurations made up of the intersection of many encounters between ‘actants’ (people and things)" (Agnew, 2011). In fact, our experiments indicate that the labelling (i.e., semantic categorisation) of social media posts benefits significantly from including not only the spatial but also the temporal aspects of social media content. We experimented with graphs based on spatio-temporal distance which take the temporal element of tweets into account during the construction of the graph. We proposed two distance calculation approaches, one based on a spatio-temporal Euclidean distance and one based on a temporal weighted Euclidean distance. The former simply considers the temporal element as a third, separate dimension, whereas the latter uses a mathematical weight to equate space and time, to control the impact of time on distance. These versions of the GCN thus take into account not merely the spatial neighbours of a tweet to understand the local context, but its spatio-temporal neighbours. The results show that taking into account the temporal component

improves the quality of the labelling and the stability of the model. The GCN model on the graph constructed using temporal weighted Euclidean distance also achieves the overall best results, which does not only illustrate the effectiveness of our distance calculation approach but also indicates that a social media analysis requires sophisticated modelling of the temporal element. The GCN seems to successfully capture the in-depth connections between similar events that might be spatially distant from each other but temporally close, and vice versa.

As such, a GCN on a well-defined spatio-temporal graph achieves better results through a deeper understanding of places as "time-space configurations" (Agnew, 2011) and social media posts as "intersection of many encounters between 'actants'" (Agnew, 2011), thus contextualising each post within its spatio-temporal neighbours. To the best of our knowledge, this is the first paper to embed a spatio-temporal distance into a deep learning approach to achieve semantic understandings of social media content. While our approach in this paper has achieved reasonable performance, we suggest that further research is necessary with regard to this aspect.

Finally, we explored the robustness of our framework and evaluate whether data variability (e.g., variations in the proportion of data for each label in training data) might affect the labelling results. The experiments demonstrate that our framework is robust and produces stable, consistent labellings. As such, we argue that our proposed framework has the potential to be developed into a powerful tool for analysis of noisy and imbalanced social media datasets in digital geographies.

## 7 Conclusions and Future Work

In this paper, we presented a novel approach to the exploratory analysis of geo-located social media content capable of classifying posts based not only on their textual content but also taking into account their visual content, and embedding the concept of place through spatio-temporal graph convolutional networks, thus breaking new ground in the use of deep learning in GIScience. Furthermore, our experiments show that our framework can also benefit research in digital geographies (Ash et al., 2018) in the analysis of large volumes of data, where a mixed-method approach combining quantitative and qualitative analysis might be necessary.

We outlined a stacked multi-modal autoencoder able to extract combined representations from multimedia content of social media posts, and a graph convolutional network with encoded geographical information developed to label social media posts based on their content and the place where they are posted. The outcomes indicate that our framework can produce good labelling results with partially labelled data, even if the dataset is heavily noisy and imbalanced. Our experiments also demonstrate that spatio-temporal graph convolutional networks are an effective way to encapsulate and understand social media posts as "augmentations"(Graham et al., 2015) of places as "time-space configurations"(Agnew, 2011) and thus enable a better semantic characterisation of content. As such, our approach is a first step towards bridging the gap between quantitative textual processing and visual content analysis in digital geographies, illustrating how visual content is an indispensable part of social media analysis, and it has the potential to be developed into a flexible tool for the study of digital geographies.

This paper is but one step towards a better understanding of our digitally-mediated urban spaces. The approach here presented is by no means meant to replace in-depth qualitative analyses of social media (or indeed urban spaces). Rather, it aims to complement and take those findings further, expanding in-depth analyses of small volumes of social media posts to large volumes of posts, which would not be unrealistic to process manually. Our framework aims to bridge the qualitative-quantitative divide and provide a useful tool for researchers in digital geographies, where the latter is understood in a broader sense, including both quantitative and qualitative (and indeed mixed) methods.

In our future work, we aim to test the scalability of our framework further. As mentioned above, any labelling process, whether manual or automated, brings significant levels of uncertainty and subjectiveness, thus rendering the type of classification task tackled in the paper even more challenging, as well as difficult to assess. We are currently experimenting a fuzzy logic classification, where multiple labels (or "codes") can be attached to each tweet. We believe that such an approach might be suited to case studies in digital geographies, where tagging a single piece of content with multiple labels can be extremely valuable in dealing with uncertainty and subjectiveness. Moreover, the current approach deals with feature extraction and semi-supervised training as two separated, subsequent stages, and we are working towards combining them into an end-to-end training framework.

# References

Abernathy, David. 2016. *Using Geodata and Geolocation in the Social Sciences: Mapping Our Connected World*. Sage.

Abrol, Satyen, Latifur Khan, and Bhavani Thuraisingham. 2012. "Tweecalization: Efficient and Intelligent Location Mining in Twitter Using Semi-Supervised Learning". In: *International Conference on Collaborative Computing: Networking*.

Agnew, John. 2011. "Space and place". In: *The SAGE handbook of geographical knowledge*. Ed. by John Agnew and David N. Livingstone. Sage London. Chap. 23, pp. 316–330.

Andrew, Galen et al. 2013. "Deep canonical correlation analysis". In: *International conference on machine learning*, pp. 1247–1255.

Ash, James, Rob Kitchin, and Agnieszka Leszczynski. 2018. "Digital turn, digital geographies?" In: *Progress in Human Geography* 42.1, pp. 25–43.

Awcock, Hannah. 2018. "Contesting the Capital: Space, Place, and Protest in London, 1780-2010". English. PhD thesis. Royal Holloway, University of London.

Ballatore, Andrea and Stefano De Sabbata. 2018. "Charting the Geographies of Crowdsourced Information in Greater London". In: *Geospatial Technologies for All*. Ed. by Ali Mansourian et al. Cham: Springer International Publishing, pp. 149–168. ISBN: 978-3-319-78208-9.

— 2020. "Los Angeles as a digital place: The geographies of user-generated content". In: *Transactions in GIS* 24.4, pp. 880–902. DOI: 10.1111/tgis.12600. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12600. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12600.

Borth, Damian et al. 2013. "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content". In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM, pp. 459–460.

Boscoe, Francis P., Kevin A. Henry, and Michael S. Zdeb. 2012. "A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals". In: *The Professional Geographer* 64.2, pp. 188–196. DOI: 10.1080/00330124.2011.583586. URL: https://doi.org/10.1080/00330124.2011.583586.

Cai, Guoyong and Binbin Xia. 2015. "Convolutional neural networks for multimedia sentiment analysis". In: *Natural Language Processing and Chinese Computing*. Springer, pp. 159–167.

Chandar, S et al. 2016. "Correlational Neural Networks." In: *Neural Computation* 28.2, p. 257.

Chang, Jae-Woo et al. 2007. "Spatio-temporal similarity measure algorithm for moving objects on spatial networks". In: *International Conference on Computational Science and Its Applications*. Springer, pp. 1165–1178.

Chen, Meng et al. 2017. "Weighted co-training for cross-domain image sentiment classification". In: *Journal of Computer Science and Technology* 32.4, pp. 714–725.

Cheng, Tao and Thomas Wicks. 2014. "Event detection using Twitter: a spatio-temporal approach". In: *PloS one* 9.6, e97807.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Crampton, Jeremy W. et al. 2013a. "Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb". In: *Cartography and Geographic Information Science* 40.2, pp. 130–139. DOI: 10.1080/15230406.2013.777137. eprint: https://doi.org/10.1080/15230406.2013.777137. URL: https://doi.org/10.1080/15230406.2013.777137.

Crampton, Jeremy W et al. 2013b. "Beyond the geotag: situating 'big data'and leveraging the potential of the geoweb". In: *Cartography and geographic information science* 40.2, pp. 130–139.

Dan, Xu et al. 2014. "Find you from your friends: Graph-based residence location prediction for users in social media". In: *IEEE International Conference on Multimedia Expo*.

Dodge, Martin and Rob Kitchin. 2004. "Flying through code/space: the real virtuality of air travel". In: *Environment and planning A* 36.2, pp. 195–211.

Elwood, Sarah and Agnieszka Leszczynski. 2013. "New spatial media, new knowledge politics". In: *Transactions of the Institute of British Geographers* 38.4, pp. 544–559.

Felt, Mylynn. 2016. "Social media and the social sciences: How researchers employ Big Data analytics". In: *Big Data & Society* 3.1, p. 2053951716645828. DOI: 10.1177/2053951716645828. eprint: https://doi.org/10.1177/2053951716645828. URL: https://doi.org/10.1177/2053951716645828.

Frias-Martinez, Vanessa and Enrique Frias-Martinez. 2014. "Spectral clustering for sensing urban land use using Twitter activity". In: *Engineering Applications of Artificial Intelligence* 35, pp. 237–245.

Gajarla, Vasavi and Aditi Gupta. 2015. "Emotion Detection and Sentiment Analysis of Images". In: *Georgia Institute of Technology*.

Gao, Yue et al. 2015. "Multimedia social event detection in microblog". In: *International Conference on Multimedia Modeling*. Springer, pp. 269–281.

Ghani, Norjihan Abdul et al. 2019. "Social media big data analytics: A survey". In: *Computers in Human Behavior* 101, pp. 417–428.

Gomide, Janaína et al. 2011. "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter". In: *Proceedings of the 3rd international web science conference*. ACM, p. 3.

Graham, Mark, Stefano De Sabbata, and Matthew A. Zook. 2015. "Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information". In: *Geo: Geography and Environment* 2.1, pp. 88–105. DOI: 10.1002/geo2.8. eprint: https://rgs-ibg.onlinelibrary.wiley.com/doi/pdf/10.1002/geo2.8. URL: https://rgs-ibg.onlinelibrary.wiley.com/doi/abs/10.1002/geo2.8.

Graham, Mark, Matthew Zook, and Andrew Boulton. 2013. "Augmented reality in urban places: contested content and the duplicity of code". In: *Transactions of the Institute of British Geographers* 38.3, pp. 464–479. DOI: 10.1111/j.1475-5661.2012.00539.x. eprint: https://rgs-ibg.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-5661.2012.00539.x. URL: https://rgs-ibg.onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-5661.2012.00539.x.

Gross, J. L. and J. Yellen. 1999. *Graph theory and its applications /*.

Guo, Diansheng and Chao Chen. 2014. "Detecting non-personal and spam users on geotagged Twitter network". In: *Transactions in GIS* 18.3, pp. 370–384.

Hamid, Raffay et al. 2005. "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE, pp. 1031–1038.

Hollenstein, Livia and Ross Purves. 2010. "Exploring place through user-generated content: Using Flickr tags to describe city cores". In: *Journal of Spatial Information Science* 2010.1, pp. 21–48.

Hu, Yingjie et al. 2015. "Extracting and understanding urban areas of interest using geotagged photos". In: *Computers, Environment and Urban Systems* 54, pp. 240–254.

Huang, Po-Yao et al. 2018a. "Multimodal Filtering of Social Media for Temporal Monitoring and Event Analysis". In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, pp. 450–457.

Huang, Xiao et al. 2018b. "A visual–textual fused approach to automated tagging of flood-related tweets during a flood event". In: *International Journal of Digital Earth*, pp. 1–17.

Ifrim, Georgiana, Bichen Shi, and Igor Brigadir. 2014. "Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering." In: *SNOW-DC@ WWW*, pp. 33–40.

Kingma, Diederik P and Jimmy Ba. 2014. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kipf, Thomas N. and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks". In:

Le, Quoc and Tomas Mikolov. 2014. "Distributed representations of sentences and documents". In: *International conference on machine learning*, pp. 1188–1196.

Lee, Chung-Hong et al. 2011a. "A novel approach for event detection by mining spatio-temporal information on microblogs". In: *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 254–259.

Lee, Ryong, Shoko Wakamiya, and Kazutoshi Sumiya. 2011b. "Discovery of unusual regional social activities using geo-tagged microblogs". In: *World Wide Web* 14.4, pp. 321–349.

Liu, Pengyuan and Stefano De Sabbata. 2019a. "Learning Digital Geographies through a Graph-Based Semi-supervised Approach". In: *the 15th International Conference on GeoComputation*. Queenstown, New Zealanda.

— 2019b. "Learning Digital Geographies through a Multi-modal Autoencoder". In: *GISRUK 2019, the 27th annual GIScience Research UK conference*. Newcastle, UK.

Longley, Paul A and Muhammad Adnan. 2016. "Geo-temporal Twitter demographics". In: *International Journal of Geographical Information Science* 30.2, pp. 369–389.

Luo, Feixiong et al. 2016. "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago". In: *Applied Geography* 70, pp. 11–25.

Mao, Xiao Jiao, Chunhua Shen, and Yu Bin Yang. 2016. "Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections". In:

Martí, Pablo, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. "Social media data: Challenges, opportunities and limitations in urban studies". In: *Computers, Environment and Urban Systems* 74, pp. 161–174.

Martín, Yago, Zhenlong Li, and Susan L Cutter. 2017. "Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew". In: *PLoS one* 12.7, e0181701.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4, pp. 1093–1113.

Miller, Harvey J. and Michael F. Goodchild. 2015. "Data-driven geography". In: *GeoJournal* 80.4, pp. 449–461. ISSN: 1572-9893. DOI: 10.1007/s10708-014-9602-6. URL: https://doi.org/10.1007/s10708-014-9602-6.

Mishra, Shaunak, Aasish Pappu, and Narayan Bhamidipati. 2019. "Inferring advertiser sentiment in online articles using wikipedia footnotes". In: *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1224–1231.

Mouzannar, Hussein, Yara Rizk, and Mariette Awad. 2018. *Damage Identification in Social Media Posts using Multimodal Deep Learning*.

O'Sullivan, David and David Unwin. 2014. *Geographic information analysis*. John Wiley & Sons.

Panteras, George et al. 2015. "Triangulating social multimedia content for event localization using Flickr and Twitter". In: *Transactions in GIS* 19.5, pp. 694–715.

Poorthuis, Ate and Matthew Zook. 2017. "Making big data small: strategies to expand urban and geographical research using social media". In: *Journal of Urban Technology* 24.4, pp. 115–135.

Qi, Weijie et al. 2019. "Mapping consumer sentiment toward wireless services using geospatial twitter data". In: *IEEE Access* 7, pp. 113726–113739.

Resch, Bernd et al. 2015. *Urban Emotions—Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data*.

Sechelea, Andrei et al. 2016. "Twitter data clustering and visualization". In: *2016 23rd International Conference on Telecommunications (ICT)*. IEEE, pp. 1–5.

Shaw, Joe and Mark Graham. 2017. "An informational right to the city? Code, content, control, and the urbanization of information". In: *Antipode* 49.4, pp. 907–927.

Shelton, Taylor et al. 2014. "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'". In: *Geoforum* 52, pp. 167 –179. ISSN: 0016-7185. DOI: https://doi.org/10.1016/j.geoforum.2014.01.006. URL: http://www.sciencedirect.com/science/article/pii/S0016718514000207.

Sommer, Alfred. 2016. "The utility of "big data" and social media for anticipating, preventing, and treating disease". In: *JAMA ophthalmology* 134.9, pp. 1030–1031.

Tsou, Ming-Hsiang et al. 2013. "Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election". In: *Cartography and Geographic Information Science* 40.4, pp. 337–348.

Wadawadagi, Ramesh S and Veerappa B Pagi. 2020. "Sentiment Analysis on Social Media: Recent Trends in Machine Learning". In: *Handbook of Research on Emerging Trends and Applications of Machine Learning*. IGI Global, pp. 508–527.

Wakamiya, Shoko, Ryong Lee, and Kazutoshi Sumiya. 2011. "Urban area characterization based on semantics of crowd activities in twitter". In: *International Conference on GeoSpatial Sematics*. Springer, pp. 108–123.

Wang, Zheye, Xinyue Ye, and Ming-Hsiang Tsou. 2016a. "Spatial, temporal, and content analysis of Twitter for wildfire hazards". In: *Natural Hazards* 83.1, pp. 523–540.

— 2016b. "Spatial, temporal, and content analysis of Twitter for wildfire hazards". In: *Natural Hazards* 83.1, pp. 523–540. ISSN: 1573-0840. DOI: 10.1007/s11069-016-2329-6. URL: https://doi.org/10.1007/s11069-016-2329-6.

Weller, Katrin et al. 2014. *Twitter and society*. Vol. 89. Peter Lang.

Xie, Junyuan, Ross Girshick, and Ali Farhadi. 2016. "Unsupervised deep embedding for clustering analysis". In: *International conference on machine learning*, pp. 478–487.

Xu, Can et al. 2014. "Visual sentiment prediction with deep convolutional neural networks". In: *arXiv preprint arXiv:1411.5731*.

Xu, Zheng et al. 2017. "Building the multi-modal storytelling of urban emergency events based on crowdsensing of social media analytics". In: *Mobile Networks and Applications* 22.2, pp. 218–227.

Yang, Jaewon and Jure Leskovec. 2011. "Patterns of temporal variation in online media". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 177–186.

You, Quanzeng et al. 2015. "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks." In: *AAAI*, pp. 381–388.

Zahra, Kiran, Frank O Ostermann, and Ross S Purves. 2017. "Geographic variability of Twitter usage characteristics during disaster events". In: *Geo-spatial information science* 20.3, pp. 231–240.

Zhu, Xiaojin and Zoubin Ghahramani. 2002. *Learning from labeled and unlabeled data with label propagation*. Tech. rep. Citeseer.

Zhu, Xiaojin and Andrew B Goldberg. 2009. "Introduction to semi-supervised learning". In: *Synthesis lectures on artificial intelligence and machine learning* 3.1, pp. 1–130.

Zook, Matthew A and Mark Graham. 2007. "Mapping DigiPlace: geocoded Internet data and the representation of place". In: *Environment and Planning B: Planning and Design* 34.3, pp. 466–482.