ISPRS Journal of Photogrammetry and Remote Sensing 188 (2022) 255–268

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Can linguistic features extracted from geo-referenced tweets help building function classification in remote sensing?

Matthias Häberle [a,b], Eike Jens Hoffmann [a], Xiao Xiang Zhu [a,b,*]

[a] Technical University of Munich (TUM), Data Science in Earth Observation (SiPEO), Arcisstraße 21, 80333 Munich, Germany
[b] German Aerospace Center (DLR), EO Data Science, Münchener Straße 20, 82234 Weßling, Germany

ARTICLE INFO

Keywords:
Remote sensing
Decision fusion
Building function classification
Deep learning
Natural language processing
Word embedding

ABSTRACT

The fusion of two or more different data sources is a widely accepted technique in remote sensing while becoming increasingly important due to the availability of big Earth Observation satellite data. As a complementary source of geo-information to satellite data, massive text messages from social media form a temporally quasi-seamless, spatially multi-perspective stream, but with unknown and diverse quality. Despite the uncontrolled quality: can linguistic features extracted from geo-referenced tweets support remote sensing tasks? This work presents a straightforward decision fusion framework for very high-resolution remote sensing images and Twitter text messages. We apply our proposed fusion framework to a land-use classification task – the building function classification task – in which we classify building functions like commercial or residential based on linguistic features derived from tweets and remote sensing images. Using building tags from OpenStreetMap (OSM), we labeled tweets and very high-resolution (VHR) images from Google Maps. We collected English tweets from San Francisco, New York City, Los Angeles, and Washington D.C. and trained a stacked bi-directional LSTM neural network with these tweets. For the aerial images, we predicted building functions with state-of-the-art Convolutional Neural Network (CNN) architectures fine-tuned from ImageNet on the given task. After predicting each modality separately, we combined the prediction probabilities of both models building-wise at a decision level. We show that the proposed fusion framework can improve the classification results of the building type classification task. To the best of our knowledge, we are the first to use semantic contents of Twitter messages and fusing them with remote sensing images to classify building functions at a single building level.

1. Introduction

Today, migration into cities shapes fast-growing and dynamic urban structures. According to the United Nations (2018), in 2050, about 68% of the world population will live in cities. Therefore, information about urban structures, their properties, and their dynamics is important. This includes, for example, fine-grained land-use classification such as buildings functions. Henüce, building functions such as *commercial* or *residential* contain valuable knowledge about a settlement and its composition.

The standard approach to acquiring urban structures such as building functions is analyzing remote sensing data combined with deep learning methods or simply querying municipal land registers. However, classifying building functions with optical sensors could bring challenges like a coarse spatial resolution of the imagery, uniform rooftops, or invariable building shapes. Subtle changes on the ground might not be detected or, in the case of informal settlements, not even documented in official databases (Baud et al., 2010).

A way to tackle the mentioned issues is combining deep learning methods, remote sensing images, and additional *in-situ* sensors and apply decision fusion (Ghamisi et al., 2019; Salcedo-Sanz et al., 2020). Such an in-situ sensor could be a very present and digitally global phenomenon: social media. It is pretty common these days to post information on social media platforms like Facebook, TikTok, Instagram, or Twitter. Twitter, for example, is connecting hundred of millions of active users around the globe. Users report their activities, show pictures from a place they visited, or share ideas and observations of their environment in text form, so-called *tweets*. Furthermore, users can tag

* Corresponding author at: Technical University of Munich (TUM), Data Science in Earth Observation (SiPEO), Arcisstraße 21, 80333 Munich, Germany. (Xiao Xiang Zhu)
E-mail addresses: Matthias.Haeberle@tum.de (M. Häberle), Eike.Jens.Hoffmann@tum.de (E.J. Hoffmann), Xiaoxiang.Zhu@dlr.de (X.X. Zhu).
URL: https://www.asg.ed.tum.de/sipeo (M. Häberle), https://www.asg.ed.tum.de/sipeo (E.J. Hoffmann), https://www.asg.ed.tum.de/sipeo (X.X. Zhu).

https://doi.org/10.1016/j.isprsjprs.2022.04.006
Received 27 September 2021; Received in revised form 21 February 2022; Accepted 14 April 2022
Available online 28 April 2022
0924-2716/© 2022 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

locations like bars, restaurants, museums, businesses, landmarks, or streets to share exciting places with their followers (cf. paragraph 3.2). The tagged locations are scattered throughout the city and surrounding areas and thus, can be seen as an in-situ sensor. They can contain information about the aforementioned urban characteristics and can be utilized to classify building functions. This approach is well known as citizen sensors (Goodchild, 2007).

Deep learning has been established as an omnipresent tool in remote sensing (Zhu et al., 2017; Ma et al., 2019). A significant research community focuses on land-use and building function classification tasks by exploiting deep learning methods and remote sensing images in various characteristics (Albert et al., 2017a; Srivastava et al., 2018; Kang et al., 2018; Hoffmann et al., 2019).

The latter studies showed that additional ground information could contribute to object-level building function classification tasks or urban mapping. Thus, additional data sources can be added to keep up with a modern city's dynamics and rapid changes. Chen et al. (2020) for example, studied the relationship between the urban environment and the distribution of geo-referenced tweets collected in Chicago. Therefore, inhabitants could act like a detector of urban attributes in the city they live (Ertiö, 2015; Ertiö and Bhagwatwar, 2017; Jones et al., 2015).

In contrast, research which combines remote sensing methods and natural language processing is rare. Lobry et al. (2019) merged natural language processing methods and remote sensing images. They developed a visual answering system based on a recurrent neural network and a ResNet50 fine-tuned with Sentinel-2 images. The features from both modalities have been fused point-wise, and the resulting model was able to answer questions like "Is there rural area?".

As of today, most research on urban land-use classification utilizes each of the mentioned methods alone or, regarding the text part, often employ methods like term frequency-inverse document frequency (TF-IDF) at a block-level which can determine specific topics of texts. However, the actual meaning is not considered (e.g., Zhang et al., 2017). Therefore, we aim to close the gap using pre-trained word embeddings to preserve the meaning of the social media texts for the classification of buildings.

Thus, we present in this work the decision fusion of VHR remote sensing images and linguistic features from Twitter text messages based on a pre-trained English word embedding to classify with deep learning-based methods building functions in Los Angeles, New York City, San Francisco, and Washington D.C. To the best of our knowledge, there is no such work combining previously mentioned methods classifying building functions at an individual building level.

The first part of this work includes performing a land-use classification task–the building function classification task[1]–at an individual building level. The considered functions are *commercial*, *residential*, and *other* and have been obtained from OpenStreetMap. This task has the goal to estimate the type of a building next to a geo-referenced tweet based on its linguistic features derived from a pre-trained English word embedding trained by the *fastText* algorithm (Bojanowski et al., 2017). For text classification, we made use of a bi-directional long short-term memory network (LSTM). We trained the LSTM with two different data splits: the first split is on the city level, which means that the training and inferring is performed on one city. We call this split *intra city random split*. In addition to that, we included a second split–the *inter city cross-validation split*. This split comprises the training data of three cities for training and the test split of the fourth city for predicting. We utilized the latter split to analyze the impact i) of spatial variability (spatial over-fitting) and ii) dataset size on text classification. For the remote sensing part, we apply image classification models with VGG16 (Simonyan and Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), and ResNet50 (He et al., 2016) as architectures. All models are pre-trained on ImageNet (Russakovsky et al., 2015) and fine-tuned with U.S.-wide building aerial images obtained from Bing Maps (Hoffmann et al., 2019). For testing, we obtained Google Maps images at zoom level 18, which corresponds to a spatial resolution of approx. 0.5m in our areas of interest.

The second part comprises investigating the impact of the fusion of the text classification results with classification results from remote sensing images on building type classification. For this, we utilize decision level fusion by averaging the prediction probabilities of both classifiers for each building in the test set.

The results show that the fusion of linguistic features and remote sensing imagery is rewarding on building function classification.

1.1. Contributions

In this study, we show how linguistic features from Twitter tweets can be used to predict urban land use on building level instance. Our proposed method is computationally efficient and improves the results of established remote sensing models significantly. In our evaluation we demonstrate the effectiveness using spatial leave-one-out validation in four U.S. cities.

2. Related work

Because of Twitter's worldwide active user base, information extraction from Twitter text messages offers many applications in geospatial research. For example Hamstead et al. (2018) used geo-referenced Twitter and Flickr data to evaluate how public parks are used and visited in New York City. Chen et al. (2018) employed tweets and TF-IDF to annotate OpenStreetMap objects in Great Britain. Furthermore, Twitter text messages can provide insights into the demographic characteristics of a country by analyzing language patterns (Bokányi et al., 2016). Terroso-Saenz and Muñoz (2020) used tweets and Flickr images for a fine-grained land-use classification in New York and San Francisco. They applied Latent Dirichlet allocation (LDA) (Blei et al., 2003) to extract relevant topics which relate to Foursquare venues. In this work, no remote sensing imagery was used. For the building function classification, previous work (Huang et al., 2018b and Häberle et al., 2019b; Häberle et al., 2019a) showed applicability of Twitter data and natural language processing.

Word embeddings established a widely accepted technique to represent text in machine learning tasks. Word embeddings provide a vector space representation of words such that vector similarity resembles semantic and syntactic features of a given text corpus (Bengio et al., 2003; Bojanowski et al., 2017; Collobert et al., 2011; Mikolov et al., 2013a; Pennington et al., 2014). Word embeddings showed good performance in text classification tasks such as an election classification task with tweets (Yang et al., 2018), sentiment analysis in transportation (Ali et al., 2019), or the training from scratch for domain-specific applications in geo-science (Padarian and Fuentes, 2019).

Classifying urban land-cover and land-use has been a well-studied task in the remote sensing community. Early works used decision trees based on handcrafted features (Hu and Wang, 2013), but with the rise of deep learning methods, they were quickly applied to this task as well (Marmanis et al., 2015). The strength of such deep architectures is the ability to discover latent features in large-scale datasets (Cheng et al., 2017). For remote sensing, CNN-based architectures achieve high classification scores in scene classification tasks (Cheng et al., 2017) even with randomized and frozen weights (Risojevic, 2016). Especially urban land-cover predictions gained a considerable benefit from deep models if the classification schema is at a very fine-grained level (Albert et al., 2017b). Splitting urban land cover into detailed morphological classes yields the local climate zone classification schema, which can be predicted using multi-temporal remote sensing data (Qiu et al., 2019; Qiu et al., 2020). Since urban land-use is more difficult than urban land-cover from an aerial view, several studies investigated the feasibility

[1] In this work, the terms *building function* and *building type* are used as synonyms.

of deep architectures for predicting the functions of regions. Multi-spectral remote sensing data was shown to be a valuable data source used in combination with a two-stream network and a skeleton-based decomposition network (Huang et al., 2018a). Due to the close relation of both tasks, they can benefit from each other's predictions: in built-up areas, there are fewer options for possible functions and vice versa. If an industrial function is obvious, there are just a few land cover options available. Learning these relationships can be either done in a joint top-down and bottom-up approach (Zhang et al., 2018) or using an iterative method by alternating training of land-cover and land-use networks while using the other one's prediction as a prior of the trained one (Zhang et al., 2019). However, a major challenge in remote sensing image classification (in urban) areas is the diversity and similarity between classes. Therefore, (Cheng et al., 2018) proposed a method to enhance the scene classification accuracy for similar object shapes such as churches or palaces by applying a metric learning regularization term during the training of CNN architectures. Although deep architectures achieve good results in remote sensing image classification the issue of diversity within a class still persists. Endless color variations and shapes could be present within a class like *residential* or *industrial* (Cheng et al., 2020).

Therefore, additional data from the ground, e.g., citizen sensors (Goodchild, 2007), could be used to improve the remote sensing classification at an individual object level (Cheng et al., 2017). The fusion (Schmitt and Zhu, 2016) of two or more data sources can improve classification results in urban land-use. Zhang et al. (2017), for example, examined urban land-use in Haidian District, Beijing, China using Weibo and Gaofen-2 imagery. They divided the district into fields via OpenStreetMap road data. For the land-use classification, they used textural and spectral features from the imagery and the density and temporal patterns from geo-referenced Weibo posts. The classification was performed by a Random Forest classifier, and they achieved an accuracy of 77.83%. In addition to temporal and remote sensing features, Fu et al. (2019) integrated linguistic features obtained from Twitter messages to determine land-use and land change. Including the Twitter-derived features produced a land-use classification accuracy of 81% vis-á-vis 72% without the Twitter features. Hoffmann et al. (2019) explored fusion methods of nadir satellite/aerial images with street view images within the framework of a building function classification task. The results show that decision-level fusion *(model blending)* achieves the best classification performance. Srivastava et al. (2019) showed the fusion of several data sources for urban land-use mapping. Google StreetView images for the ground perspective and Google Maps aerial images for the remote scene are used in that work. The proposed model can outperform models using only one data source. Also, the fusion of social media data and remote sensing images was proposed to generate flood maps (Wang et al., 2018) and damage estimation (Cervone et al., 2016). However, none of the studies previously mentioned explicitly exploit linguistic features derived from social media at a building level.

3. Dataset

For our fusion experiment, we have a text dataset and an image dataset. First, we describe the Twitter data collection for Washington D.C., Los Angeles, New York City, and San Francisco. Also, we discuss the Twitter geo-reference accuracy and introduce our labeling methodology and explain our developed train-test split approach.

3.1. Twitter streaming

We used the free Twitter Application Programming Interface (API), which allows us to collect 1% of the daily amount of tweets (Twitter, 2021) in the area of interest, which is, in our case, the whole world in the period from January 2018 to December 2019. For this study, we only used tweets were the `.coordinates.coordinates` field of the tweet JSON is not *null*, i.e., coming with point coordinates. From this data, we derived sub-samples for Los Angeles, New York City, San Francisco, and Washington D.C. In this work, we are only using tweets written in English.

3.2. Twitter geo-reference accuracy

In June 2019, Twitter announced,[2,3,4] to cease the precise geo-referencing functionality because only a small number of users actually used this feature. On the one hand, this step increases the geo-privacy of Twitter users, but on the other hand, this action could have an impact on geo-spatial research using Twitter data (Ballatore and Sabbata, 2020). Even though 88% of the data is posted via a third-party app like Instagram or Foursquare (Hu and Wang, 2020; Kruspe et al., 2021), precise point coordinates of places of interest, like museums or restaurants, are still provided. Hu and Wang (2020) point out that

> […] about 72% to 88% of precisely geotagged tweets were from third-party apps, such as Instagram, whereas only about 8% to 25% were directly from Twitter. Although the three datasets cannot exhaust all possible datasets that can be retrieved through Twitter API, these results suggest that Twitter's decision may not have an earth-shaking impact on research relying on geotagged tweets. (p. 1220)

Therefore, it can be argued that it is still possible to utilize geo-referenced Twitter data as a data source for geo-spatial research to a certain extent. For this study, however, we excluded tweets that were posted after the announcement. Additionally, it is possible to tag tweets with cities or neighborhoods and concentrate at a single coordinate. However, the exact location of the Twitter user who posted a tweet remains uncertain. Therefore, to avoid over-weighing tweets tagged with a city or a neighborhood, we limit the tweets per building to the average number of tweets per building of the explored city. With this measure, we treat every building equally to prevent biasing the classification towards a building that happens to be next to city or neighborhood-level tweets. At tops, such tweets are treated as noise.

Furthermore, we stress that we are not claiming that a tweet was exactly posted within a building or that the Twitter user stood right next to the building. It has been proven on a block-level that topics of tweets can contribute to building function classification (cf. Section 1 and 2). Therefore, we hypothesize that the linguistic features of nearby tweets are meaningful enough to estimate the function of the building next to them.

3.3. Labeling tweets

Before we split the text data into a train and test set, we must label the data. Several options, e.g., an open cadaster database, are possible. However, we decided to use OpenStreetMap (OSM) for our approach. First, cadaster data is not (publicly) available for every city. Second, there is no standard labeling schema for cadastral data. Third, we want to facilitate replication of our experiment. Within this process, we tag each tweet based on the next building to the tweet's geo-location. If the distance between the point location of the tweet and the building polygon is less than 50 meters and the building has a valid function tag, the tweets get labeled based on this functional tag.

OSM provides labeling schemes for building functions,[5] amenities, and shop types, and encourages its contributors to tag building polygons based on these schemes. We evaluated all of these tags for each building

[2] https://twitter.com/twittersupport/status/1141039841993355264.
[3] https://twitter.com/twittersupport/status/1142130343715078144.
[4] https://www.theverge.com/2019/6/19/18691174/twitter-location-tagging-geotagging-discontinued-removal.
[5] https://wiki.openstreetmap.org/w/index.php?title=Key:building&oldid=1576985.

in the four study areas and labeled them with one of *commercial*, *residential*, and *other* if the tags were unanimously and present. Otherwise, a building did not obtain a label. Using this method, we aim at overcoming the sparsity of semantic tags in OSM (Fan et al., 2014).

Equipped with the generated labeling scheme, we can now assign tweets to close buildings. Our rule to assign a tweet to a building is performed by OpenStreetMap building polygons stored in the geo-spatial database created by PostgreSQL[6] with the PostGIS[7] extension (Owusu et al., 2021). The distance of a tweet to the buildings of a city is measured using a geo-spatial distance function provided by the database environment. For the present study, we only used tweets closer to its next building as 50 m since a taller building could have a plaza surrounding it and further cover tweets in residential areas with a more scattered urban configuration. This yields a $1:n$ relationship between buildings and tweets so that one building can be assigned to several tweets, whereas a single tweet can be solely related to exactly one building.

The following paragraph will explain the train-test-split process, which reduces the numbers mentioned above to a certain extent. To avoid too much noise in the data, we manually excluded buildings associated with many very similar tweets like weather forecasts or traffic information. To remove such tweets, we filtered the user ids associated the most with such tweets.

3.4. Train-test-split

After we labeled the tweets, we split the data into train and test sets for each city (intra city random split). Typically, one would separate the data by a certain ratio and select the data points by chance. However, since we generated a $1:n$ relationship of buildings and tweets, it is likely that after a random train-test split, tweets assigned to the same building are at the same time in the train and test set. That means that tweets of a specific building can appear in the train and test set. To prevent such a data leakage (Kaufman et al., 2012), we developed a train-test split method that generates a list with all unique OSM building IDs assigned to a tweet and random-splits that list by a specified ratio (cf. Fig. 1). In our case, we set the split ratio to 75% to 25% and utilizing the random seed 1337. Additionally, we balanced the number of buildings per function. For this, we down-sampled the buildings to the minority class. The tweets remain unbalanced.

The next step includes de-duplication of the tweet text and text pre-processing. Text pre-processing is necessary to reduce artifacts and generate a more clean structure and representation of the irregular tweet texts (Atefeh and Khreich, 2015; Han and Baldwin, 2011; Hong et al., 2011). First, we strip all numbers and almost all punctuation from the text except apostrophes and dashes to preserve words like “wasn’t” or “part-time”. We also maintained the upper or lower casing of a specific word to cover semantic differences in words like “apple” and “Apple” (fruit, tech company). After that, we further delete URLs and emojis (Yao and Wang, 2020). The last step of the pre-processing tries to normalize spelling like “greaaaaat” into “greaat”[8] to minimize out of vocabulary words. We would like to point out that we keep so-called stop-words[9] because they could contain for the geo-spatial NLP task investigated in this paper helpful phrases like “I am at …” (Samad et al., 2020).

Now, the tweets are divided by the assigned OSM building ID either into the train or test set. As noted in paragraph 3.2, we limit the total amount of tweets per building to the mean number of tweets per building of the studied city. We limit Los Angeles to 37 tweets per building, in New York to 42, in San Francisco to 87, and in Washington 40. The tweets for each building are randomly drawn. After these steps, we have 55,910 labeled tweets for Los Angeles, 111,353 for New York, 36,313 for San Francisco, and 40,077 for Washington. See Table 1 for a detailed overview.

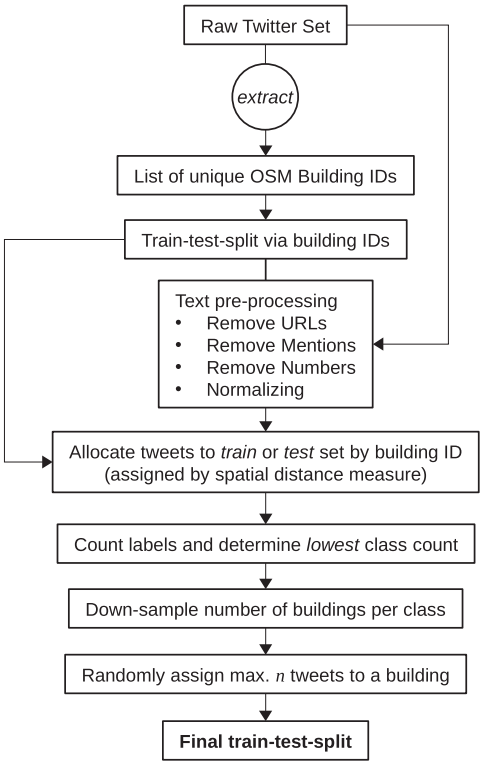


Fig. 1. Train-test-split and down-sampling process.

The inter city cross-validation split of a city consists of the training data of three other cities generated via the above method. For example, if the task is to predict building functions of New York City, we train the text classifier with the *training* data of Washington D.C., Los Angeles, and San Francisco. The model is then evaluated with the *test* data of New York City.

3.5. Remote sensing images

Based on the OSM building IDs obtained during tweet labeling we downloaded their corresponding aerial images patches focused on the building polygon centroid. Our patches were created from the Google Maps Satellite layer (Ghaffarian and Ghaffarian, 2014; Zhang et al., 2016; Li et al., 2020) at zoom level 18 yielding a spatial resolution of approximately $0.48m$ in our study areas.

Google Maps uses the WGS84 standard and provides tiles with a resolution of 256×256 pixels on up to 22 zoom levels.[10] Therefore, the ground sample distance gsd on a given zoom level z and a latitude lat is

$$gsd(z, lat) = \frac{2\pi r_E \cos(lat)}{2^{(z+8)}} \tag{1}$$

with r_E as the equatorial radius of $6,378,137m$.[11] Our test area in New York is at latitude 40 and the one in Los Angeles at latitude 33. Hence, the ground sample distance of our image patches is from $0.46m$ to

[6] https://www.postgresql.org/.
[7] https://postgis.net/.
[8] The double-a is not a mistake. The normalization keeps at least two letters of the same letter because some words have correct double characters (e.g., *letter*).
[9] Stop-words are words like *and*, *at*, or *the*, which carry no valuable information for some NLP tasks.
[10] https://developers.google.com/maps/documentation/javascript/coordinates.
[11] https://wiki.openstreetmap.org/wiki/Zoom_levels.

Table 1
Class distribution and train-test split numbers by city. The columns *tweets train* and *tweets test* show the amount of tweets after the building down-sampling as well as after the reduction to the mean tweets per city.

		raw tweet count	tweets train	tweets test	buildings train/test	aerial images
LA						
	commercial	800,137	17,552	5,720	2,004/681	681
	residential	338,722	5,369	1,619	2,004/681	681
	other	204,206	18,935	6,715	2,004/681	681
	total tweets	1,343,065	41,856	14,054		
	mean tweets per building		6.96	6.88		
	mean tweets per coordinate		3.81	3.86		
	individual buildings	67,855	6,012	2,043		2,043
NYC						
	commercial	262,343	30,694	10,159	2,509/814	814
	residential	94,571	24,808	7,970	2,509/814	814
	other	364,803	28,398	9,324	2,509/814	814
	total tweets	721,717	83,900	27,453		
	mean tweets per building		11.15	11.24		
	mean tweets per coordinate		4.35	4.55		
	individual buildings	15,449	7,527	2,442		2,442
SF						
	commercial	107,337	10,614	3,582	678/217	217
	residential	24,880	6,584	1,842	678/217	217
	other	154,552	10,431	3,260	678/217	217
	total tweets	286,769	27,629	8,684		
	mean tweets per building		13.58	13.34		
	mean tweets per coordinate		3.45	3.47		
	individual buildings	4,410	2,034	651		651
WDC						
	commercial	92,586	12,130	3,941	985/331	331
	residential	54,618	7,554	2,347	985/331	331
	other	40,746	10,748	3,357	985/331	331
	total tweets	187,950	30,432	9,645		
	mean tweets per building		10.3	9.71		
	mean tweets per coordinate		3.74	3.9		
	individual buildings	4,619	2,955	993		993

$0.50m$.

Based on multiple map tiles stitched together, we cropped out a patch of 512×512 pixels centered on the building centroid. Hence, each aerial image covers an area of approximately $65,000$ square meters. We will refer to this dataset as zoom level 18 large.

To investigate the effect of the window size on the classification results, we added a second dataset for which we cropped patches from zoom level 18 to 256×256. These patches have the same center, i.e., the building centroid, but cover only a quarter of the aforementioned original patches and are referred to as zoom level 18 small.

4. Fusion framework

This section introduces our fusion framework and explains the individual classification methodologies for the text and image datasets. Fig. 2 depicts the proposed fusion framework. It consists of a text classification stream that describes the used methodologies transferring the text into a machine-readable representation and the classification architecture. The second stream shows the image classification part. Both streams flow together at the fusion block, where the prediction probabilities are averaged building-wise and outputted. A more detailed explanation is given in the upcoming sections.

4.1. Text classification

Furthermore, the training procedure is shown, and some hyperparameters are noted. In the upcoming paragraphs we will use *word* and *token* as synonyms. Note that a word or token does not necessarily have to be a word but can also be a number or entity of a given text sequence.

4.1.1. Word embeddings

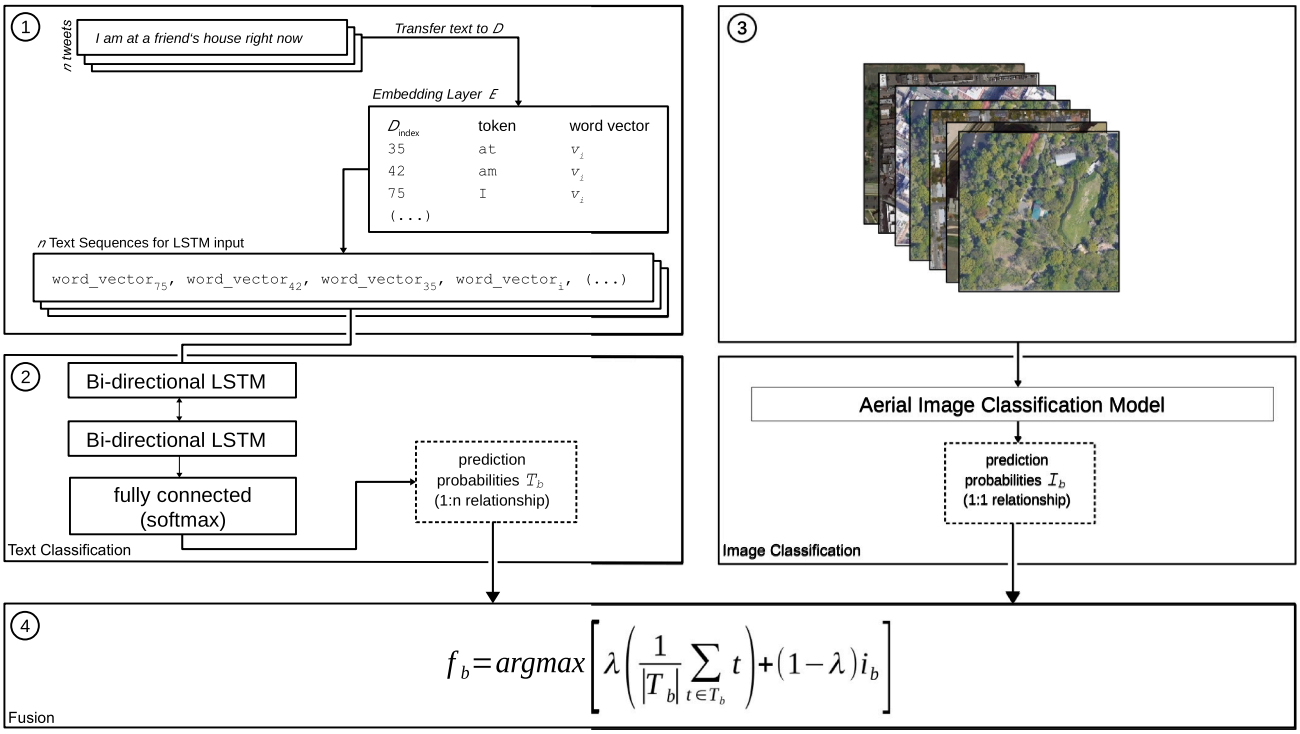
Before we can feed the text into a neural network, we have to transfer the text into a machine-readable format. Methods like TF-IDF (Spärck Jones, 1972) taking word frequencies in a document into account and producing a score for each word. However, the context of a word within a text is ignored, and semantic and syntactic features are lost.

Word embedding methods (Bengio et al., 2003; Collobert et al., 2011) such as word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), or fastText (Bojanowski et al., 2017), for example, are able to preserve semantic and syntactic features of a word up to some degree. This is achieved by taking the word's context within a sequence, i.e., neighboring words, into account (Firth, 1957; Schütze, 1992). This information is embedded into a real-valued n-dimensional feature vector for each word–a word vector. The final embedding can be queried for word similarities or word analogies for a word with algebraic operations. For example the operation *paris-france* + *italy* evaluates approximately to the vector which represents the word *rome* (Mikolov et al., 2013a; Mikolov et al., 2013b).

In this work, we apply the word embedding algorithm fastText, which could be seen as a further development of word2vec. A significant difference is that fastText considers subword information represented by character n-grams additionally. In the end, a word vector of a word is computed by the sum of its character n-grams. For this reason, fastText can improve the representations of words of morphologically rich languages, for example, German, Hebrew, or Arabic (Tsarfaty et al., 2010). Furthermore, out of vocabulary words and word compositions can be easier approximated by its n-gram structure (Bojanowski et al., 2017).

A further reason why we are using fastText is that the research team provides pre-trained word embeddings in 157 languages (Grave et al., 2018).[12] Even for non-regular languages like Bavarian or Volapük are word embeddings available. For the training procedure, they used Wikipedia dumps because of high textual quality. To add more

[12] https://fasttext.cc/docs/en/crawl-vectors.html.



$$f_b = argmax\left[\lambda\left(\frac{1}{|T_b|}\sum_{t\in T_b} t\right)+(1-\lambda)i_b\right]$$

Fig. 2. Fusion Framework. *1)* Mapping the text to machine-readable representation. *2)* Text classification with stacked bi-directional LSTMs. *3)* Remote Sensing Image Classification with DenseNet121,InceptionV3, ResNet50, VGG16, and Xception. *4)* Decision-level fusion by weight-averaging the prediction probabilities of text and image classification with Eq. (2). Where $b \in B$, and B denotes a non-empty set of building IDs. Background images ©TerraMetrics 2021, Google.

heterogeneous language, they additionally utilized CommonCrawl[13] to train the word embedding using fastText, and words were left in upper and lower case to add some semantics and context to words (cf. Section 3.4). For our work, we downloaded the pre-trained English word embedding with 2 million tokens. Each of the 2 million tokens in the English embedding has a 300-dimensional word vector assigned. In the end, we can now produce sequences of word vectors as input for a nüral network.

4.1.2. Bi-directional LSTM

The standard way a neural network reads a text input is from "left-to-right". A bi-directional LSTM, however, reads the input not just from "left-to-right" but also backwards–from "right-to-left". This behavior adds some further context to the sequence. The first bi-directional recurrent neural network was used by Schuster and Paliwal (1997) and subsequent studies added this technique to LSTM networks, for example for phoneme classification (Graves et al., 2005).

To generate the sequential input for the neural network, every word of the pre-processed tweets was transferred from human-readable text into machine-readable sequences. To perform the transfer, for each city, a word dictionary D_{city} was created by mapping every single token found in the tweets of a city to a positive integer. Next, all tokens in a tweet are replaced by the corresponding integer representing the word in the dictionary. Each dictionary was limited to the top $45,000$ words for each city (cf. Fig. 2, ①).

The final word vector sequences are produced before the actual LSTM layers by an embedding layer E. This layer was initialized with the downloaded fastText word embedding as weights and serves as look-up table for the integer index sequences (cf. Fig. 2, ①). The word index of D_{city} is also the index of the tokens and corresponding word vectors in E. Not the whole embedding with all 2 M tokens[14] is initialized in E but only the words w which are $w \in D_{city}$. In addition to that, a token $t \in D_{city}$, which is not present in the fastText embedding, is omitted. We limit the input sequence length to 30 tokens $t \in D_{city}$ for each tweet, and layer E is not trainable. Table 2 shows the total number of found words. It can be noticed, not all words within a city's vocabulary can be found in the embedding. That could have three reasons. First, a word was not in the training vocabulary of the embedding. Second, the word cannot be found due to a spelling error or irregular spelling. Third, due to the limitation to the top city 45,000 words and maximum sequence length of 30, the word was not considered for D, and therefore, it is not available.

For the text classification (cf. Fig. 2, ②), we implemented a stacked bi-directional long short-term memory (bi-LSTM) model making use of TensorFlow (Abadi et al., 2015) and Keras (Chollet, 2015). The LSTM directly after the embedding layer has 256 units, and the top LSTM layer has 128 units. For the output, we added a softmax fully-connected layer, and as an optimizer, Adam (Kingma and Ba, 2017) was used, and cross-entropy calculated the loss. Each individual-city model and cross-

Table 2
Total count of unique words after the pre-processing and building down-sampling steps in the train-test-split and words found in the fastText word embedding after the limitation to the top 45,000.

City	Unique Words	Words in Embedding
Los Angeles	95,811	28,159
New York	134,223	30,731
San Francisco	61,582	29,645
Washington	66,361	31,092

[13] https://commoncrawl.org/.

[14] A token could be a word, punctuation mark, or an emoji–any entity of a text sequence which bears meaning.

validation model was trained for 5 epochs with a batch size of 128, a learning rate of 1e-3, and 40% dropout was applied to each layer to prevent the network from overfitting to the relatively short sequences. In addition to that, the embedding layer was non-trainable.

4.2. Image classification

To predict the building function based on the aerial image, we used five fine-tuned, state-of-the-art architectures: VGG16 (Simonyan and Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), Xception (Chollet, 2017), and DenseNet121 (Huang et al., 2017). All of them are based on ImageNet (Russakovsky et al., 2015) weights and fine-tuned on a dataset of aerial images covering buildings across the U.S. Hoffmann et al. (2019) using Adam Kingma and Ba, 2017 and categorical cross-entropy (cf. Fig. 2, ③). Each architecture was adapted to the building image dataset in two steps with 16 epochs each. First, only a new, final dense layer for classification was trained, with the remaining network being frozen. In the subsequent step, all layers of the networks were unfrozen starting from the end and trained further on the dataset while decreasing the learning rate (Hoffmann et al., 2019). The exact protocols of training can be found in Table 3.

Using a different dataset from another data provider (here, Microsoft Bing Maps) for training made sure that it has never seen our test images before and obtains real-world prediction scores. For each architecture and patch size, there is an individual model specialized in the task.

4.3. Fusion method

The fusion method denoted in (2) we are using is straightforward. In short, we are weight-averaging the predicted softmax probabilities of the test sets for each building and classifier (cf. Fig. 2, ④). This method is based at the decision level fusion or, in other words: Decision-In-Decision-Out (DEI-DEO) approach (Dasarathy, 1997; Salcedo-Sanz et al., 2020) and can also be referred to as model blending (e.g., Hoffmann et al., 2019). A significant pro argument using this procedure is that if one classifier cannot perform well on a specific data point, the other classifier could compensate. In that case, the "Twitter sensor" can counteract this issue by providing a classification result. Furthermore, this approach is computationally reasonable. However, as a counterpoint, there is a possibility that information, i.e., features, are lost during the decision process (Dasarathy, 1997).

Besides the here applied fusion method, several techniques are available. For example, the data could be fused using a feature-based method. The features of both modalities are fused before a decision was made. That could be realized by fusing the features after applying intra-modality feature extraction or, on the other hand, in an inter-modality approach. Another variant of the decision level fusion is to integrate the features after a decision was made. In contrast to the decision method we apply, the blending takes place before softmax is applied–so to speak, the decision features are fused instead of the decision probabilities. For further details please consider (Dasarathy, 1997; Schmitt and Zhu, 2016; Ghamisi et al., 2019; Salcedo-Sanz et al., 2020). In addition, a recent study found that decision fusion at a feature level cannot improve the classification results of a land-use classification task if both modalities are fundamentally different (Hoffmann et al., 2019). First, we gather all predictions from the text and image classifiers and fully separate the training and prediction phase from the fusion procedure. That course of action allows us to analyze the predictions detached from the fusion process. That can be useful, for example, if one would like to study the performance or impact of specific images or tweets on the classification at an individual data point level.

Since the $1:n$ relationship between buildings and tweets in the text set, we first average all predicted probabilities of a building ID produced by the text classification part if one building possesses more than one prediction. In this way, we generate a $1:1$ relationship between text classification and image classification predictions before the actual fusion process, i.e., a mini-fusion before the fusion. Now we can fuse the building-wise averaged text probabilities and the probabilities predicted by the image classifiers building-wise. The fusion process per building $b \in B$, where B denotes a non-empty set of building IDs. Finally, the latter steps can be summarized and can be formally defined by:

$$f_b = \operatorname{argmax}\left[\lambda\left(\frac{1}{|T_b|}\sum_{\mathbf{t}\in T_b}\mathbf{t}\right) + (1-\lambda)\mathbf{i_b}\right] \tag{2}$$

The predictions of both models are represented as probability vectors of the size $n_{classes}$. Thus, the text model predictions are represented as a set of probability vectors T, and therefore, a prediction from a tweet for a building is noted as $\mathbf{t} \in T$. Analog to the text predictions, predictions for a building from an image are determined as a set of probability vectors I where $\mathbf{i} \in I$. $\lambda \in (0, 1]$ defines the weight given to the predictions from the linguistic features. An optimal λ value yields the best fusion results.

Table 3
Fine-tuning protocol on aerial imagery applied to selected architectures.

Architecture	Step	Learning Rate	# Trained Layers
DenseNet121	1	1e−4	1
	2	1e−5	427
InceptionV3	1	1e−4	1
	2	1e−5	311
ResNet50	1	1e−4	1
	2	1e−5	175
VGG16	1	1e−4	1
	2	1e−5	21
Xception	1	1e−4	1
	2	1e−5	132

5. Results and discussion

The following section shows the classification results of the text, image, and data fusion parts. First, we discuss the text classification results, followed by the aerial image classification, and finally, we present the fusion results of the two modalities.

5.1. Text classification results

As pointed out in paragraph 3.4, we used two datasets for the classification task. The *intra city random split* comprises data for one city. We want to examine, in general, how the LSTM network performs in a single city. By performing an *inter city cross-validation*, we want to investigate two goals. First, we check the impact of spatial variability, i.e., spatial over-fitting, on text classification. In other words: how is the locally specific vocabulary of a city influencing the text classification? Moreover, is there a measurable generalization such that the classification performance does not drop? Even if the model is trained with a dataset that is from an unseen and distant area. Second, we want to investigate the effect of the dataset size. In the last paragraph, we stated that the classification results for cities does not depend on dataset size. For the cross-validation, we trained four additional text models with the same configuration as noted in Section 4.1.2. However, we extended the training data of a city with the training data of two other cities. We validated the model by predicting the test set on the fourth (cf. Table 4, column *inter city cross-validation*).

Interestingly, the models deliver different results for every city. Also, for cities with less data such as San Francisco and Washington D.C., the overall performance appear not dependent on the total amount of tweets (cf. Table 4, column *intra city random split, bi-LSTM (text)*). Despite we had less Twitter data and available buildings for Washington (cf. Table 1), the overall accuracy and Kappa score is higher as for New York. Additionally, the New York data coming with the highest number of individual buildings. Also, we cannot measure a positive impact on the classification results. If we compare the classification results of single

Table 4
Classification and fusion results in more detail. All numbers reflect the F1 score of the classification results except Kappa (κ) and overall accuracy (OA). The λ value in brackets denotes the one determined for the inter city cross-validation.

			intra city random split		inter city cross-validation	
		VGG16 18 small	bi-LSTM (text)	fusion	bi-LSTM (text)	fusion
Los Angeles $\lambda = 0.49(0.55)$	*commercial*	0.73	0.69	**0.77**	0.65	**0.77**
	residential	0.82	0.70	**0.85**	0.52	0.82
	other	0.64	0.68	**0.72**	0.61	0.72
	OA	0.73	0.69	**0.78**	0.60	0.77
	κ	0.60	0.54	**0.67**	0.39	0.65
New York City $\lambda = 0.45(0.45)$	*commercial*	0.68	0.64	**0.73**	0.61	0.72
	residential	**0.63**	0.50	0.62	0.31	0.60
	other	0.59	0.64	**0.66**	0.63	0.65
	OA	0.64	0.60	**0.67**	0.55	0.67
	κ	0.45	0.40	**0.51**	0.32	0.50
San Francisco $\lambda = 0.72(0.65)$	*commercial*	0.59	0.64	**0.68**	0.63	**0.66**
	residential	**0.63**	0.52	0.61	0.52	0.62
	other	0.53	0.59	0.61	0.59	**0.62**
	OA	0.58	0.59	**0.64**	0.59	0.63
	κ	0.37	0.38	**0.45**	0.38	**0.45**
Washington D.C. $\lambda = 0.46(0.50)$	*commercial*	0.71	0.69	0.76	0.67	**0.76**
	residential	**0.72**	0.56	0.72	0.45	0.70
	other	0.68	0.67	0.73	0.67	**0.73**
	OA	0.70	0.65	**0.74**	0.62	0.73
	κ	0.55	0.47	0.60	0.42	**0.60**

classes of San Francisco and New York, we learn that the results of San Francisco are almost the same (with a small bias towards N.Y.). The results of Washington outperform New York's clearly and slightly lower than the Los Angeles results, although L.A. has more data and far more buildings. By comparing Los Angeles and New York, we can find the same performance pattern. L.A. has less data than N.Y. but performs much better. This finding is the first takeaway message: for individual building function classification, the claim that more data improves text classification performance seems not to hold. We further analyze the impact of the dataset size. Additionally, we investigate a possible impact of spatial variability by performing an inter city cross-validation.

Table 4 depicts all the (class-wise) results in more detail. The Los Angeles text classification results of the intra city random split show an overall accuracy of 0.69 and a Kappa score of 0.54. Further, a F1 score of 0.70 at the *residential* class can be observed. The *commercial* class shows a F1 score of 0.69 and the *other* class 0.68. For Los Angeles, the inter city cross-validation brought no additional performance. However, it is interesting that the classification results are almost as good as the intra city random split. This finding could indicate that the classification of tweet text can work outside of the pre-defined area. That spatial variability has an impact but is not as explicit as might be expected.

New York City shows for the text classification an overall accuracy of 0.60 and a Kappa score of 0.40. The *commercial* and *other* class performs best with a F1 score of 0.64 whereas the *residential* class shows with 0.50 the weakest result. A comparison of the intra city random split and inter city cross-validation reveals that the cross-validation could not boost the results. On the contrary: the performance of the *residential* class dropped clearly to 0.31 as well as the Kappa score to 0.32. Nonetheless, we want to point out that (except for the *residential* class) the results of the individual classes are almost the same. Which supports the claim that spatial variability has a limited impact on classification.

San Francisco shows at the intra city random split an overall accuracy of 0.59 and a Kappa score of 0.38. The *residential* class has the lowest F1 score with 0.52 and the *other* class 0.59. However, the *commercial* class shows with 0.64 the best result. The numbers of the inter city cross-validation split do not draw a different picture: the performance is equal. The claim that more training data can positively influence the classification performance for cities with small text training datasets does not hold for San Francisco. On the other hand, since the performance did not decrease, the effect of spatial variability also here a minor one.

The results for Washington D.C. show an overall accuracy of 0.65 and a Kappa score of 0.47. The F1 for *commercial* is 0.69, for *residential* 0.56, and for *other* 0.67. Regarding the inter city cross-validation, we could observe a similar output even though it is not so clear as for San Francisco. For the *commercial* class, the F1 slightly decreased from 0.69 to 0.67 but *residential* from 0.56 to 0.45. The overall accuracy declined from 0.65 to 0.62 as well as the Kappa score from 0.47 to 0.42. As for the other cities, the classification performance did not become dramatically less. With one exception: the *residential* class.

We can observe in all cities the drop of performance of the *residential* class. Of course, such results can be attributed to the smaller number of tweets available for the *residential* class (cf. Table 1). However, why do we have less data for this class in the first place? A possible answer to this question could be the accuracy of the Twitter data. According to the comments below the announcement of the deactivation of the precise geo-referencing of a tweet, users welcomed this step since more and more people having privacy concerns. As mentioned before, the amount of precise geo-referenced tweets are decreased over time. Most likely, the residential areas and private buildings are not as precisely covered as *commercial* or *other* buildings where users can tag points of interest in their tweets. The source of *residential* tweets could stem from a provided pre-defined area provided by the used app like a street or a certain neighborhood. Therefore, the superior performance of *commercial* and *other* buildings could arise from the possibility to tag a pre-defined point of interest line a landmark, shopping mall, museum, university, car repair shop, et cetera.

5.2. Text classification examples

Furthermore, we assume that some of the text models have difficulties covering the linguistic diversity of certain areas. In more detail, that, for example, tweets about work or carrier sent nearby a *residential* building are confused with an actual commercial tweet (cf. Fig. 3). However, since people can discuss work or career questions from home, the classifier could be biased towards one specific word, somehow prototypical for such a class. For example, the wrongly classified *residential* buildings of Los Angeles and San Francisco might show exactly that (cf. Fig. 3 D, P). The tweets containing *breakfast*, *toast*, *cook*, or *restaurant* that are clearly foot-related and therefore might be associated with restaurant tweets. In addition to that, the falsely classified tweet of the Los Angeles *other* building, also shows *Grill* which is also a foot-related term (cf. Fig. 3 F). An interesting example is the wrongly classified commercial building of Los Angeles (cf. Fig. 3 B). Here, the tweet is

Fig. 3. Remote Sensing images and one selected example tweet of the corresponding OSM building. The abbreviations in the brackets denote the falsely classified class (*com* = commercial, *res* = residential, *oth* = other). Real names have been overwritten with *xxx* or *yyy* to preserve the privacy of the user. Background images ©TerraMetrics 2021, Google.

about the Los Angeles Housing Agency, which is a governmental institution. Those government-related buildings are summarized in the other class. By scrutinizing the text, the classifier correctly identified the tweet content as government-related and, consequently, classified it to *other*. Nevertheless, what went wrong? By a close look, we found that the assigned building is incorrect. The tweet is labeled with the building function across the street because it is closer to this building. However, the behavior of the classifier, in this case, gives some evidence that the

classification of buildings with pure text data is possible.

Nonetheless, one can also "work" in your garden or on your painting skills, but one is still at home and tweet about non-commercial topics. That opens further interesting research questions we would like to address in the future. On the other hand, the correctly classified *commercial* buildings of Washington, New York, San Francisco, and Los Angeles suggest that food-related words are associated with *commercial* places. *Residential* tweets include words like *cat* or *Super Bowl*, perhaps this is why the *commercial* tweet of New York are classified as *residential* even though *commercial* words are occurring. The words used in *other*-tweets seem more diverse like *University* or *Elementary* (Fig. 3 , Q, W).

5.3. Aerial image classification results

The results for all vision models are shown in the Appendix, Table 5. For the sake of brevity, we will focus on the best model, VGG16, on zoom level 18 small in this section. This model is at least as good as all other models but outperforms them in most cases. A larger spatial context including more neighborhoods does not improve the classification results. Instead, focusing on the building instances themselves shows a higher classification performance.

Moreover, although InceptionV3 and ResNet50 were published after VGG16, their results are below the VGG16 models. The similar behavior of Inception and ResNet models is most likely related to their high structural similarity McNeely-White et al., 2020. The better performance of VGG16 indicates that its features generalize better to other domains than the other two models.

Table 4, column *VGG16 18 small* shows the results of the remote sensing image classification class-wise for each city. The VGG models outperform every text model with respect to overall accuracy and Kappa score except for San Francisco.

For Los Angeles, the predictions indicate an overall accuracy of 0.73 and a Kappa score of 0.60. The *residential* class shows the best F1 score of 0.82 which could be explained by different building shapes by comparison with commercial buildings. This results for the *residential* class is the best amongst all other city VGG models. *Commercial* exhibits the second-best F1 score with 0.73 and is followed by the *other* class with 0.64. Here, the text model of the intra city random split can outperform the image classification. The overall accuracy is higher (0.73 vs. 0.69).

The VGG predictions for New York show an overall accuracy of 0.64 and a Kappa score of 0.45. The *commercial* class is the top class with a F1 score of 0.68 followed by the *residential* class (0.63). *Other* shows a F1 of 0.59. The text model, however, can outperform the score of the *other* class (0.59 vs. 0.64). Such findings could demonstrate the usefulness of geo-referenced on-site text data because they could deliver additional (latent) information about the area or building.

In San Francisco, the model outperforms the VGG16 network (except the *residential* class). The VGG shows an overall accuracy of 0.58 (text 0.59) and a Kappa score of 0.37 (text 0.38). *Residential* is the best performing class with a F1 of 0.63 followed by the *commercial* class with 0.59. The *other* class showing the poorest classification results with 0.53. The difference between the image and text model here is even more pronounced than before. The text model can achieve higher F1 scores for *commercial* and the *other* class. Which further substantiate that geo-referenced text can contribute to building function classification.

For Washington, the VGG predictions outperform the text model. The overall accuracy is 0.70 and Kappa 0.55 better than the text model. Only Los Angeles presents with 0.60, a higher Kappa value. The *commercial* class classification results show with a F1 score of 0.71 a good result. For the *residential* class, the VGG model can achieve a strong F1 of 0.72 which is followed by the *other* class with 0.68.

The most exciting finding of the image classification is that the performance of the *residential* class shows opposite results by comparison with the text classification outcome. The second compelling discovery is (as it has already been indicated above in the paragraph reviewing the text classification results) that the highest amount of buildings do not necessarily result in better performance. New York has the most significant number of individual buildings, but the accuracy score is in third place behind Los Angeles and Washington D.C. The following paragraph discusses the remaining open question if data fusion can improve the results of individual classes.

5.4. Fusion results

Fig. 4 shows the overall results of the weighted fusion process. For Los Angeles (0.48 and 0.49), New York City (0.45), and Washington D.C (0.46). In contrast, the San Francisco fusion results demonstrate higher scores when using higher λ weights (0.71 and 0.72). This means the linguistic features require a higher weight to achieve optimal results. The numbers in Table 4 of the decision fusion of text and remote sensing image classification results draw a clear picture: for all cities, the applied method can outperform all of the image and text classification results.

The overall accuracy of Los Angeles improved to 0.78 and the Kappa score to 0.67. All classed benefit from the data fusion. Even though the

Table 5
Additional Remote Sensing image classification results with respect to vision models in more detail. All numbers denote the F1 score except Kappa κ and overall accuracy. L = Large, S = Small.

		Inception		ResNet50		VGG16		DenseNet		Xception	
		18L	18S	18L	18S	18L	18S	18L	18S	18L	18S
LA	*commercial*	0.68	0.70	0.65	0.69	0.71	0.73	0.69	0.70	0.68	0.71
	other	0.62	0.62	0.57	0.61	0.80	0.64	0.62	0.63	0.58	0.60
	residential	0.78	0.79	0.75	0.78	0.63	0.82	0.76	0.78	0.74	0.77
	Overall Accuracy	0.69	0.70	0.66	0.69	0.71	0.73	0.69	0.70	0.67	0.70
	κ	0.54	0.55	0.49	0.54	0.57	0.60	0.54	0.55	0.50	0.54
NY	*commercial*	0.62	0.63	0.60	0.62	0.65	0.68	0.62	0.66	0.63	0.64
	other	0.51	0.53	0.46	0.54	0.57	0.59	0.50	0.56	0.48	0.52
	residential	0.60	0.56	0.58	0.57	0.54	0.63	0.59	0.59	0.61	0.56
	Overall Accuracy	0.58	0.58	0.55	0.58	0.59	0.64	0.57	0.60	0.58	0.58
	κ	0.37	0.37	0.33	0.37	0.39	0.45	0.36	0.40	0.37	0.37
SF	*commercial*	0.57	0.54	0.53	0.52	0.54	0.59	0.52	0.54	0.55	0.53
	other	0.50	0.53	0.49	0.51	0.49	0.53	0.51	0.52	0.51	0.50
	residential	0.59	0.59	0.54	0.59	0.41	0.63	0.41	0.50	0.48	0.50
	Overall Accuracy	0.56	0.55	0.52	0.53	0.49	0.58	0.49	0.52	0.52	0.52
	κ	0.34	0.33	0.28	0.30	0.24	0.37	0.24	0.28	0.28	0.27
WDC	*commercial*	0.66	0.68	0.61	0.65	0.69	0.71	0.64	0.66	0.65	0.67
	other	0.62	0.63	0.55	0.59	0.63	0.68	0.60	0.60	0.62	0.62
	residential	0.64	0.67	0.60	0.65	0.65	0.72	0.61	0.63	0.67	0.63
	Overall Accuracy	0.64	0.66	0.59	0.63	0.66	0.70	0.62	0.63	0.65	0.64
	κ	0.46	0.49	0.38	0.44	0.49	0.55	0.43	0.45	0.47	0.46

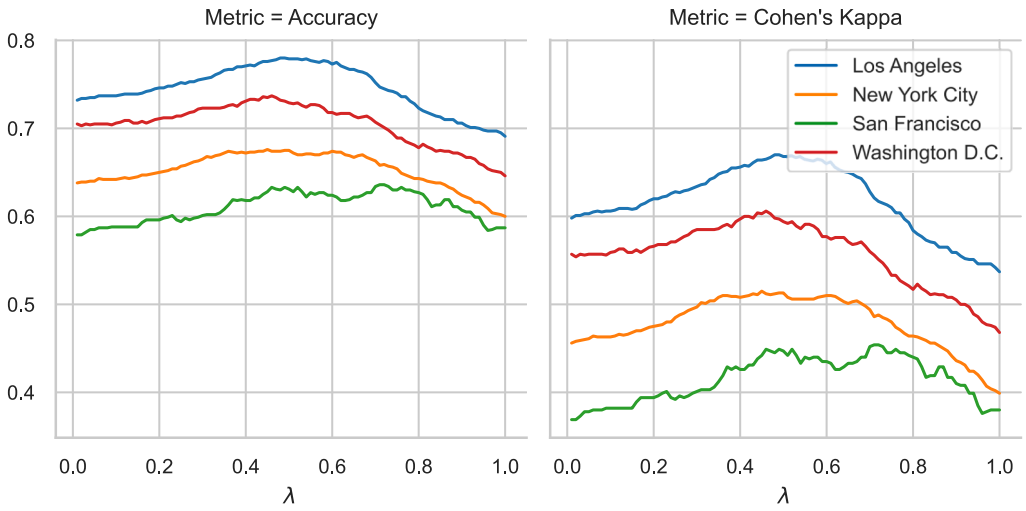


Fig. 4. Overall results of weighted mean fusion with λ as the weight for the linguistic features. Most of the cities show the best results using $\lambda \in \{0.45, 0.46, 0.48, 0.49\}$. San Francisco is an outlier with an optimal λ of 0.71 and 0.72.

other class is not the weakest, it can benefit most from the fusion process. It improved from 0.64 achieved by the vision model to 0.72 F1. However, the fusion of the text predictions of the inter city cross-validation split cannot reach the fusion results of the intra city random split results but can improve the results almost to the same level as of the intra city random split. All classes can improve by the data fusion process. The other model compensates the weaknesses of one model (cf. paragraph 4.3). A good example is the performance of the *other* class: the VGG F1 is 0.64 and the text's 0.68. The mean fusion of the classification results of the individual buildings lifted the results to 0.72.

Fusing the results of image and text results for New York City also revealed similar findings. The results are improved after the decision fusion with an overall accuracy of 0.67. As for Los Angeles, the weak *other* class improved most. Moreover, the intra city random split fusion outperforms the inter city cross-validation split fusion. As for Los Angeles, the fusion can increase the numbers of the cross-validation results almost to the same level as of the intra city random split.

For San Francisco, however, we could also observe an improvement after combining the two modalities. In contrast to the findings of Los Angeles and New York, the fusion of the inter city cross-validation text models and the image results can outperform the random split for the *other* and *residential* class. However, the *residential* class' fusion results are outperformed by the VGG results. That could further reflect evidence for an inter-area generalization of the inter city cross-validation text models.

Washington's fusion results, like all other results, demonstrate the same pattern: the fusion of the two modalities increases the classification results from 0.70 of the VGG to 0.74 overall accuracy of the intra city random split. The results of the inter city cross-validation fusion improved for the *other* class and the overall accuracy. Additionally, as mentioned above, the text models of the intra city random split and the inter city cross-validation performing almost on the same level. This substantiates the findings of San Francisco and backs the claim that the models could generalize beyond a region.

The results suggest that the advantage of the weighted Decision-In-Decision-Out fusion method mentioned in Section 4.3, namely the mutual support of the classifiers, can be seen in our classification results. Furthermore, the discoveries of San Francisco and Washington D.C. point to the generalization capability of the text models across different areas to a certain amount.

6. Conclusion

In this work, we classified buildings labeled by OpenStreetMap building tags. The building tags have been summarized to *commercial*, *residential*, and *other*. For the text classification part, we used a bi-directional LSTM architecture and a pre-trained English fastText word embedding. To classify the images, we used DenseNet,InceptionV3, ResNet50, VGG16, and Xception as vision architectures pre-trained with ImageNet and fine-tuned on building patches from remote sensing data. The prediction probabilities of the test set have been fused building-wise at a decision level. We can show that the fusion of linguistic features extracted from social media text messages (as in-situ sensors) and remote sensing images can improve the building classification task results. Additionally, the results indicate that the amount of data is not the holy grail for classification performance. This finding opens research opportunities that investigate which data contributes most to geo-spatial research tasks like building function classification. Finally, the impact of spatial variability remained small in our study. We could not observe a major decline in overall classification performance except for single classes. Especially the *residential* class showed clear performance drops after the inter city cross-validation. A possible explanation for this might be that *commercial* and *other* class comprise a more general vocabulary like business terms, whereas the *residential* class has more tweets with more regional phrases or words.

However, building type classification at an individual building level remains a challenging task. Labeled OpenStreetMap buildings are sparse, and therefore, the number of different buildings and the tweets assigned to them might not be optimal. In addition to that, Twitter deactivated in June 2019 the precise geo-referencing of tweets. Nonetheless, we could show that even with sparse data, careful parameter tuning, and a straightforward decision level fusion, the classification performance is quite good. The results encourage further research regarding fusing remote sensing images and linguistic features extracted from geo-referenced texts like social media text messages.

Utilizing Twitter data is just a snapshot of current geo-referenced text sources. In the future, a new text source with a precise geo-reference might be established, which could be used to explore urban characteristics like building functions. As a thought experiment, in a visionary Smart City, a new text feedback system could be installed by the municipal government to collect information about the communities of the city reported by their citizens (Jones et al., 2015). Therefore, the approach discussed in this paper is not limited to Twitter text messages and can also be applied to other geo-referenced text resources that might come up in the future.

7. Future work

A certainly interesting research step is the accuracy of geo-referenced tweets. Even though Twitter deactivated this feature, first studies Hu and Wang (2020) and Kruspe et al. (2021) took a closer look at the *sources*, i.e., from what platform a tweet was released into the internet. In addition to that, how the place object in the metadata of a tweet can be exploited. Since it contains further geo-spatial information to a certain granularity. First findings indicate that geo-spatial natural language processing with Twitter data is still possible but remains a valuable source for geo-spatial research. A further research point might be the assigned of tweets to buildings. We saw in paragraph 5.2 that a tweet was labeled with the wrong building. The labeling process could be further improved by studying the relationship between distance and classification accuracy.

In the current study, we only used English tweets sampled from U.S. cities. Even though English seems to be the *lingua franca* on Twitter (Kim et al., 2014), future studies could encompass European cities and the exploration of the impact of a multilanguage dataset. Adequate approaches like MUSE (Yang et al., 2019) or a multilanguage variant of BERT (Devlin et al., 2018) can be applied on building function classification. Furthermore, the challenge of a possible spatial overfitting phenomenon could be further investigated. For example, which impact has a certain (prototype) word on a building function prediction? Are there even such prototype words? Investigating this, we could add some explanation why a tweet is classified as a tweet from a *commercial* or *residential* building and add more meaning into the text and remote sensing imagery fusion. Finally, additional data fusion concepts like adequate fusion methods, for example, at the feature level and uncertainty measures, could be further investigated.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xiaoxiang Zhu reports financial support was provided by European Research Council. Xiaoxiang Zhu reports financial support was provided by the Helmholtz Association of German Research Centers eV. Xiaoxiang Zhu reports financial support was provided by German Federal Ministry of Education and Research.

Acknowledgment

The work is mainly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*). The work of X. Zhu is also supported by the Helmholtz Association through the Framework of Helmholtz AI [Grant No.: ZT-I-PF-5-01] - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research" (W2-W3-100) and by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future A.I. lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant No.: 01DD20001).

Appendix A. Detailed remote sensing image classification results

Table 5 shows an overview of the classification results of the different deep learning models on both image datasets. Notably, the VGG16-18-Small fusion can outperform almost every other model. Even the VGG-18-large is clearly exceeded by the VGG16-18-small model. The Inception and ResNet models can not compete with both of the VGG models. Also, newer models, such as Xception (Chollet, 2017) or DenseNet (Huang et al., 2017) cannot achieve the VGG's performance.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/. software available from tensorflow.org.

Albert, A., Kaur, J., Gonzalez, M.C., 2017a. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 1357–1366. URL: http://doi.acm.org/10.1145/3097983.3098070, doi:10.1145/3097983.3098070. event-place: Halifax, NS, Canada.

Albert, A., Kaur, J., Gonzalez, M.C., 2017b. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1357–1366.

Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H., Kwak, K.S., 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowl.-Based Syst. 174, 27–42. https://doi.org/10.1016/j.knosys.2019.02.033.

Atefeh, F., Khreich, W., 2015. A survey of techniques for event detection in twitter. Comput. Intell. 31, 132–164. https://doi.org/10.1111/coin.12017.

Ballatore, A., Sabbata, S.D., 2020. Los Angeles as a digital place: The geographies of user-generated content. Trans. GIS 24, 880–902. https://doi.org/10.1111/tgis.12600.

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R., Karuppannan, S., 2010. Understanding heterogeneity in metropolitan india: The added value of remote sensing data for analyzing sub-standard residential areas. Int. J. Appl. Earth Obs. Geoinf. 12, 359–374. https://doi.org/10.1016/j.jag.2010.04.008.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A Neural Probabilistic Language Model. J. Mach. Learn. Res. 3, 1137–1155.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022. URL: http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

Bojanowski, P., Grave, E., Mikolov, T., 2017. Enriching Word Vectors with Subword Information. Trans. Assoc. Comput. Linguist. 5, 135–146. URL: http://www.aclweb.org/anthology/Q17-1010.

Bokányi, E., Kondor, D., Dobos, L., Sebok, T., Stéger, J., Csabai, I., Vattay, G., 2016. Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States. Palgrave Commun. 2, 16010. https://doi.org/10.1057/palcomms.2016.10. URL: https://www.nature.com/articles/palcomms201610.

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., Waters, N., 2016. Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. Int. J. Remote Sens. 37, 100–124. https://doi.org/10.1080/01431161.2015.1117684.

Chen, X., Vo, H., Yu, W., Wang, F., 2018. A framework for annotating OpenStreetMap objects using geo-tagged tweets. Geoinformatica 589–613.

Chen, Y., Song, Y., Li, C., 2020. Where do people tweet? the relationship of the built environment to tweeting in chicago. Sustain. Cities Soc. 52, 101817. https://doi.org/10.1016/j.scs.2019.101817.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE 105, 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 13, 3735–3756. https://doi.org/10.1109/JSTARS.2020.3005403.

Cheng, G., Yang, C., Yao, X., Guo, L., Han, J., 2018. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. IEEE Trans. Geosci. Remote Sens. 56, 2811–2821. https://doi.org/10.1109/TGRS.2017.2783902.

Chollet, F., 2015. Keras. URL: https://keras.io.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcouglu, K., Kuksa, P., 2011. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res. 12, 2493–2537. URL: http://jmlr.org/papers/volume12/collobert11a/collobert11a.pdf.

Dasarathy, B., 1997. Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proc. IEEE 85, 24–38. https://doi.org/10.1109/5.554206.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] URL: http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Ertiö, T.P., 2015. Participatory Apps for Urban Planning-Space for Improvement. Plan. Pract. Res. 30, 303–321. https://doi.org/10.1080/02697459.2015.1052942. URL: https://doi.org/10.1080/02697459.2015.1052942.

Ertiö, T.P., Bhagwatwar, A., 2017. Citizens as planners: Harnessing information and values from the bottom-up. Int. J. Inform. Manage. 37, 111–113. URL: http://www.sciencedirect.com/science/article/pii/S026840121630473X, doi: https://doi.org/10.1016/j.ijinfomgt.2017.01.001.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on openstreetmap. Int. J. Geogr. Inform. Sci. 28, 700–719. https://doi.org/10.1080/13658816.2013.867495.

Firth, J.R., 1957. A synopsis of linguistic theory 1930–1955. Stud. Linguist. Anal. 1–32.
Fu, C., Song, X.P., Stewart, K., 2019. Integrating Activity-Based Geographic Information and Long-Term Remote Sensing to Characterize Urban Land Use Change. Remote Sens. 11, 2965. https://doi.org/10.3390/rs11242965.
Ghaffarian, S., Ghaffarian, S., 2014. Automatic building detection based on Purposive FastICA (PFICA) algorithm using monocular high resolution Google Earth images. ISPRS J. Photogram. Remote Sens. 97, 152–159. https://doi.org/10.1016/j.isprsjprs.2014.08.017.
Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A., 2019. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. IEEE Geosci. Remote Sens. Mag. 7, 6–39. https://doi.org/10.1109/MGRS.2018.2890023.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. https://doi.org/10.1007/s10708-007-9111-y.
Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T., 2018. Learning Word Vectors for 157 Languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. pp. 3483–3487. URL: http://www.aclweb.org/anthology/L18-1550.
Graves, A., Fernández, S., Schmidhuber, J., 2005. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S. (Eds.), Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005. Springer, Berlin, Heidelberg, pp. 799–804. https://doi.org/10.1007/11550907_126.
Hamstead, Z.A., Fisher, D., Ilieva, R.T., Wood, S.A., McPhearson, T., Kremer, P., 2018. Geolocated social media as a rapid indicator of park visitation and equitable park access. Comput. Environ. Urban Syst. 72, 38–50. URL: http://www.sciencedirect.com/science/article/pii/S0198971517303538, doi: https://doi.org/10.1016/j.compenvurbsys.2018.01.007.
Han, B., Baldwin, T., 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - vol. 1, pp. 368–378. URL: http://dl.acm.org/citation.cfm?id=2002472.2002520.
Häberle, M., Werner, M., Zhu, X.X., 2019a. Building Type Classification from Social Media Texts via Geo-Spatial Textmining. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 10047–10050. doi: 10.1109/IGARSS.2019.8898836.
Häberle, M., Werner, M., Zhu, X.X., 2019b. Geo-spatial text-mining from Twitter - a feature space analysis with a view toward building classification in urban regions. Eur. J. Remote Sens. 52, 2–11. https://doi.org/10.1080/22797254.2019.1586451. URL: https://www.tandfonline.com/doi/full/10.1080/22797254.2019.1586451.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: European conference on computer vision. Springer, pp. 630–645.
Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model fusion for building type classification from aerial and street view images. Remote Sens. 11 https://doi.org/10.3390/rs11111259. URL: https://www.mdpi.com/2072-4292/11/11/1259.
Hong, L., Convertino, G., Chi, E.H., 2011. Language matters in twitter: A large scale study. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 518–521. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2856.
Hu, S., Wang, L., 2013. Automated urban land-use classification with remote sensing. Int. J. Remote Sens. 34, 790–803. https://doi.org/10.1080/01431161.2012.714510.
Hu, Y., Wang, R.Q., 2020. Understanding the removal of precise geotagging in tweets. Nat. Hum. Behav. 4, 1219–1221. https://doi.org/10.1038/s41562-020-00949-x.
Huang, B., Zhao, B., Song, Y., 2018a. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sensing of Environment 214, 73–86. URL: http://www.sciencedirect.com/science/article/pii/S0034425718302074, doi: 10.1016/j.rse.2018.04.050.
Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708.
Huang, R., Taubenböck, H., Mou, L., Zhu, X.X., 2018b. Classification of Settlement Types from Tweets Using LDA and LSTM. In: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 6408–6411. doi:10.1109/IGARSS.2018.8519240.
Jones, P., Layard, A., Speed, C., Lorne, C., 2015. MapLocal: Use of Smartphones for Crowdsourced Planning. Plan. Pract. Res. 30, 322–336. https://doi.org/10.1080/02697459.2015.1052940.
Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogram. Remote Sens. 145, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.
Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data 6, 15:1–15:21. doi:10.1145/2382577.2382579.
Kim, S., Weber, I., Wei, L., Oh, A., 2014. Sociolinguistic analysis of twitter in multilingual societies. In: Proceedings of the 25th ACM conference on Hypertext and social media, Association for Computing Machinery. pp. 243–248. doi:10.1145/2631775.2631824.URL: https://doi.org/10.1145/2631775.2631824.
Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv: 1412.6980.
Kruspe, A., Häberle, M., Hoffmann, E.J., Rode-Hasinger, S., Abdulahhad, K., Zhu, X.X., 2021. Changes in Twitter geolocations: Insights and suggestions for future usage. arXiv:2108.12251 [cs] URL: http://arxiv.org/abs/2108.12251. arXiv: 2108.12251.
Li, W., Dong, R., Fu, H., Wang, J., Yu, L., Gong, P., 2020. Integrating Google Earth imagery with Landsat data to improve 30-m resolution land cover mapping. Remote Sens. Environ. 237, 111563. https://doi.org/10.1016/j.rse.2019.111563.
Lobry, S., Murray, J., Marcos, D., Tuia, D., 2019. Visual question answering from remote sensing images. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 4951–4954. doi:10.1109/IGARSS.2019.8898891.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogram. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.
Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2015. Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13, 105–109.
McNeely-White, D., Beveridge, J.R., Draper, B.A., 2020. Inception and resnet features are (almost) equivalent. Cogn. Syst. Res. 59, 312–318.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3 [cs] URL: https://arxiv.org/abs/1301.3781v3.
Mikolov, T., Le, Q.V., Sutskever, I., 2013b. Exploiting Similarities among Languages for Machine Translation. arXiv:1309.4168 [cs] URL: http://arxiv.org/abs/1309.4168. arXiv: 1309.4168.
Owusu, M., Kuffer, M., Belgiu, M., Grippa, T., Lennert, M., Georganos, S., Vanhuysse, S., 2021. Towards user-driven earth observation-based slum mapping. Comput. Environ. Urban Syst. 89, 101681. https://doi.org/10.1016/j.compenvurbsys.2021.101681.
Padarian, J., Fuentes, I., 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. SOIL 5, 177–187. https://doi.org/10.5194/soil-5-177-2019.
Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. pp. 1532–1543. URL: http://www.anthology.aclweb.org/D/D14/D14-1162.pdf.
Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network. ISPRS J. Photogram. Remote Sens. 154, 151–162.
Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2020. Fusing Multiseasonal Sentinel-2 Imagery for Urban Land Cover Classification With Multibranch Residual Convolutional Neural Networks. IEEE Geosci. Remote Sens. Lett.
Risojevic, V., 2016. Analysis of learned features for remote sensing image classification. In: 2016 13th Symposium on Neural Networks and Applications (NEUREL), pp. 1–6. doi:10.1109/NEUREL.2016.7800145.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision (IJCV) 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., Camps-Valls, G., 2020. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. Inform. Fusion 63, 256–272. https://doi.org/10.1016/j.inffus.2020.07.004.
Samad, M.D., Khounviengxay, N.D., Witherow, M.A., 2020. Effect of Text Processing Steps on Twitter Sentiment Classification using Word Embedding. arXiv:2007.13027 [cs] URL: http://arxiv.org/abs/2007.13027. arXiv: 2007.13027.
Schmitt, M., Zhu, X.X., 2016. Data Fusion and Remote Sensing: An ever-growing relationship. IEEE Geosci. Remote Sens. Mag. 4, 6–23. https://doi.org/10.1109/MGRS.2016.2561021.
Schütze, H., 1992. Dimensions of meaning. In: Supercomputing '92:Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, pp. 787–796. doi:10.1109/SUPERC.1992.236684.
Schuster, M., Paliwal, K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45, 2673–2681. https://doi.org/10.1109/78.650093 conference Name: IEEE Transactions on Signal Processing.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Spärck Jones, K., 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. J. Document. 28, 11–21.
Srivastava, S., Vargas-Muñoz, J.E., Swinkels, D., Tuia, D., 2018. Multilabel building functions classification from ground pictures using convolutional neural networks. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Association for Computing Machinery. pp. 43–46. doi:10.1145/3281548.3281559.URL: doi: 10.1145/3281548.3281559.
Srivastava, S., Vargas-Muñoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. Remote Sens. Environ. 228, 129–143. https://doi.org/10.1016/j.rse.2019.04.014.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.
Terroso-Saenz, F., Muñoz, A., 2020. Land use discovery based on Volunteer Geographic Information classification. Expert Syst. Appl. 140, 112892. https://doi.org/10.1016/j.eswa.2019.112892.
Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., Tounsi, L., 2010. Statistical Parsing of Morphologically Rich Languages (SPMRL). What, How and Wither. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, Los Angeles, CA, USA. pp. 1–12. URL: http://dl.acm.org/citation.cfm?id=1868771.1868772.
Twitter, 2021. Sampled Stream. URL: https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction.

United Nations, 2018. World urbanization prospects 2018 (keyfacts). URL: https://esa.un.org/unpd/wup/Publications/Files/WUP2018-KeyFacts.pdf.

Wang, H., Skau, E., Krim, H., Cervone, G., 2018. Fusing Heterogeneous Data: A Case for Remote Sensing and Social Media. IEEE Trans. Geosci. Remote Sens. 56, 6956–6968. https://doi.org/10.1109/TGRS.2018.2846199.

Yang, X., Macdonald, C., Ounis, I., 2018. Using word embeddings in Twitter election classification. Inform. Retriev. J. 21, 183–207. URL: https://link.springer.com/article/10.1007/s10791-017-9319-5.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R., 2019. Multilingual universal sentence encoder for semantic retrieval. arXiv:1907.04307.

Yao, F., Wang, Y., 2020. Tracking urban geo-topics based on dynamic topic model. Comput. Environ. Urban Syst. 79, 101419. https://doi.org/10.1016/j.compenvurbsys.2019.101419.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019. Joint Deep Learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187.

Zhang, Q., Wang, Y., Liu, Q., Liu, X., Wang, W., 2016. CNN based suburban building detection using monocular high resolution Google Earth images. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 661–664. doi:10.1109/IGARSS.2016.7729166.

Zhang, X., Du, S., Wang, Q., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sens. Environ. 212, 231–248.

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., Wang, H., 2017. The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China. Remote Sens. 9, 865. https://doi.org/10.3390/rs9090865.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307.