AGILE: GIScience Series, 2, 1, 2021. https://doi.org/10.5194/agile-giss-2-1-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

Information-optimal Abstaining for Reliable Classification of Building Functions

Gabriel Dax and Martin Werner

gabriel.dax@tum.de, martin.werner@tum.de

Technical University of Munich, Department of Aerospace and Geodesy, Big Geospatial Data Management, Munich, Germany

Abstract. In the past decade, major breakthroughs in sensor technology and algorithms have enabled the functional analysis of urban regions based on Earth observation data. It has, for example, become possible to assign functions to areas in cities on a regional scale. With this paper, we develop a novel method for extracting building functions from social media text alone. Therefore, a technique of abstaining is applied in order to overcome the fact that most tweets will not contain information related to a building function albeit they have been sent from a specific building as well as the problem that classification schemes for building functions are overlapping.

Keywords. Probabilistic Classification, Social Media Text Mining, Land Use, Urban Analysis, Building Functions

1 Introduction

The fusion of social media data and remote sensing data for urban studies is an emerging research area (Salcedo-Sanz et al., 2020). While remote sensing sensors reach resolutions in which the analysis of urban regions with respect to building-level scale becomes possible, the limitations of the birds-eye perspective of satellite-acquired data is getting more and more problematic.

In general, there is a rising interest in urban regions due to the fact that it is expected that the majority of people will live in urban regions in the next years (Taubenböck and Wurm, 2015). At the same time, many global challenges arise inside cities or due to urbanization (Cohen, 2006).

In this context, data fusion has become a major research trend in the data science and earth observation field: how can we augment the highly accurate, morphological data acquired from space with local, ground-level information in order to resolve ambiguities inherent to satellite imagery? In general, the extraction of spatial data from social media has been widely discussed, for example for geo-tagged photos (Paldino et al., 2015; Zhu and Newsam, 2016), social media text (Crooks et al., 2013), and mobility data (Veloso et al., 2011; Yuan et al., 2012).

One particularly interesting area of research is the question of building functions. When we can assign functions to buildings and at the same time have access to the morphological parameters of buildings (footprint, height, etc.), applications including expected population density are within reach. Though many highly complicated and culturally variable concepts of building functions exist, we concentrate on the simplest yet most important distinction: *residential* and *commercial*. These two classes clearly combine to the majority of buildings in cities, though other classes like religious places, amenities, or industrial might be interesting as well.

Of course, the classification into *residential* and *commercial* is not well-defined. First of all, some buildings are neither of those classes like public buildings and religious places, and some are both like buildings where the upper floors are *residential* while the first floor is used for shops. Any promising classification system must therefore be able to deal with class overlap as well as with outliers.

With this paper, we address the challenging question whether social media text collected from the Twitter social network can be used to assign buildings into the two classes *residential* and *commercial*. While text

mining is a well-established research area with many techniques, the performance for social media is still limited due to the very short texts, the use of slang, as well as the general low quality of language. In addition, text classification is often done with respect to classification schemes that are actually related to the text: traditional examples include the classification of movie reviews into a scale from positive to negative (Maas et al., 2011) or newsgroup postings into topics (Mitchell; Nigam et al., 2000). For social media, the sentiment has been a traditional research area (Go et al., 2009) and this is clearly related to the text itself.

The problem discussed in this paper is different, though: we expect that most of the tweets are not related to a building function at all. In other words, we are trying to develop a classification system that is able to filter irrelevant data items automatically and –at the same time– work with very small support.

The third challenge for the task described in this paper is class imbalance: while it is clear that most of the buildings are *residential*, the fusion dataset is just opposite: when we assign tweets to buildings, we have so many more tweets per *commercial* building that the number of tweets in *commercial* buildings is significantly larger than the number of tweets in *residential* buildings. As described before, however, we might be more interested in the *residential* class for the envisioned applications related to population in cities. That is, the minority class is given by social media near *residential* buildings and is the main class of interest. Hence, our classification system should be able to deal well with the minority class. There are many systems that are actually designed to detect a minority in the area of anomaly and outlier detection (Kiermeier et al., 2017). Still, due to the incompleteness and overlap of our set of classes, this extreme case doesn't fit. Instead, we need a full classification that is putting some efforts into understanding the minority class even if this is not usually supported by typical classification metrics such as cross-entropy loss or F1 scores, and consequently not reachable by optimization-based learning like deep learning alone.

We apply a technique known as abstaining (Chow, 1970) in order to solve the three outlined challenges of our problem setting: (1) class overlap and class ambiguity, (2) irrelevant data, and (3) class skew. In abstaining, a classifier is given the option to classify a data instance into an additional class which basically means that there is no evidence of putting it into one of the real classes. The central challenge in this area is how the cost of abstaining from classification relates to the cost of a wrong classification. In cost-sensitive classification (Elkan, 2001), this trade-off is the central research topic explicit in assigning cost to errors and it is surprisingly hard to come up with a non-subjective optimal cost setting. Fortunately, in the case of probabilistic classifiers it is possible to develop such systems based on information theory avoiding the subjective and complicated choice of parameters related to misclassification cost.

Another twist on the problem is given by the spatial nature of language. Geo-located tweets often contain spatial references in the text. This includes the name of places or restaurants. As we are interested in spatial generalization, we have to make sure that we are training our classifiers in a different area than where we apply them. Classical random train-test splits that do not account to spatial autocorrelation will provide very optimistic estimations of performance.

The main contribution of this paper is the development and analysis of a methodology to ensemble many sparse text mining models in order to extract a highly reliable label for at least a few buildings giving geo-referenced tweet text alone. Note that one should not expect to get very high accuracies with this approach, but the purpose of this paper is not to show the best way of assigning building functions. Instead, it shows that the social media text contains an independent and important contribution to building function classification and builds an informed basis for fusion with remote sensing imagery as well as with social media imagery and other data sources.

The remainder of this paper is structured as follows: in the next Section 2, we introduce fundamental principles on abstaining in the context of probabilistic classifiers as well as model blending techniques. Section 3 describes the construction of the dataset. Then, Section 4 introduces a case study in the Los Angeles area. Section 5 discusses the results of this case study and, finally, concludes the paper.

2 Fundamental Principles

In this section, we introduce some background on relevant topics to the special classification problem. First, we give an introduction to the classical technique of abstaining and the recent developments related to using the modified normalized mutual information to guide the parameter choice for abstaining costs. Furthermore, we introduce some basic ensembling techniques in which a set of different classifiers can be combined to an overall classification result.

2.1 Text Classification

In the Internet area, huge collections of text are accessible to the public including all web pages, curated collections like Wikipedia or news pages, email and social media messages. The methods of text classification aim at classifying documents into classes. In order to do that, features can be extracted from higher order

language patterns such as grammar or just by word occurrences.

As this paper is concerned with rather informal and very short documents (e.g., tweets), we decide to use only low-level structures including words, characters and character n-grams. Character n-grams are subsequences of n-characters and thereby capture the concept of syllables to a certain extent.

Given a set D of documents (e.g., tweets), a basic approach to text mining is based on first splitting the documents into words (tokenization) and using the occurrence statistics of words in documents for information representation. That is, we fix a set of words called *vocabulary* and create a vector for each document containing the number of times each word of the vocabulary occurred in the document. This results in a sparse integer vector for each documents and, thus, the corpus D of documents can naturally be represented as a sparse integer matrix S.

However, the raw frequencies are not very useful as many frequent words are uninformative in general ("we", "he", "are", etc.) and should be removed. At the same time, rare words cannot be used in machine learning setting as it is difficult to infer the meaning of a word from the statistics of word occurences if the number of occurences is small. Therefore, it is customary to remove a certain, language-specific set of words called stop-words, a certain fraction of the frequent and rare words and to build the vocabulary somewhere in the middle of the trade-off between highly frequent words and rare words. The technique of term-frequency-inverse-document frequency (TF-IDF) has further been proposed to normalize raw frequencies of words in single documents by expected frequencies of these words occurring in documents.

For tweets, these document-word matrices are very sparse as tweets contain only a handful of words. Therefore, we face a high risk of overfitting and apply simple classification schemes such as logistic regression and multinomial Naïve Bayes. In addition, it makes very clear that we should not expect that each and every tweet contributes to our problem of assigning building functions: only some of the words of every tweet are part of the vocabulary and only some of these words actually are non-neutral with respect to the given classification task.

Two traditional approaches to text mining address the problem that the overlap between two documents in terms of vocabulary might be small. One is topic mining in which a set of words from the vocabulary is grouped together into a topic. The other approach is text embedding in which words are assigned to positions in a chosen low-dimensional space such that the Euclidean distance captures aspects of meaning. However, these two techniques need huge amounts of training data and / or a clear topic structure of all documents.

2.2 Learning under Class Imbalance

In the past, there have been many efforts to deal with class imbalance. In machine learning, it is quite common that the interesting class only has a few examples while the majority class is defined as the less important default behavior. A broad range of specialized methods have been proposed, we want to give an overview of the most important directions of dealing with imbalance:

Collect more data: This one-fits-all rule of machine learning is, of course, also valid for imbalanced datasets. If it is possible to extract more examples from both classes this can be very helpful.

Change your metric: If you know the imbalance of the dataset and you also have a good argumentation to fix the misclassification cost of both classes, you can try to reflect this in the metric used for optimization-based machine learning including, but not limited to, deep learning.

Resample the dataset: Of course, one simple way of getting rid of class imbalance is to randomly sample the same amount of data from all classes. Two major approaches can be distinguished: undersampling the minority class and oversampling the majority class. While undersampling has the advantage of being conceptually simple, it reduces the amount of data that can be effectively used. Various methods of sampling have been studied in the literature (Tomek, 1976; Chawla et al., 2002; Japkowicz, 2000).

For undersampling, it has been discussed, from which region of the feature space of the classifier it is best to draw the examples. For oversampling, it has been studied whether the data should just be repeated or how synthetic examples can be generated. An advanced method of this type is SMOTE in which a combination of over- and undersampling is applied in order to maximize performance. The oversampling is done by generating new examples in feature space by choosing a random example, computing its k nearest neighbors in feature space and creating new instances of the given class by interpolating along the line connecting the example with its k nearest neighbors. This approach leads to better decision boundaries in classification such that the classifier is not picking up details of the shape induced by the real examples, but rather something related to a locally convex closure of this shape. Adaptive variants of SMOTE including Borderline-SMOTE and Adaptive Synthetic Sampling (ADA-SYN) have been proposed that account for the fact that SMOTE might increase the overlap of classes near the boundary (Han et al., 2005; He et al., 2008).

Select Classifier: Some classifiers are known to work better than others with imbalanced classes. For example, trees and random forests are a good family of algorithms for imbalanced classification due to the splitting rules employed. Some algorithms have actually been modified to account for class imbalance in model building. The interested reader is referred to a survey of He and Garcia (He and Garcia, 2009).

Problem Reformulation: If the imbalance is rather extreme, it might as well be advisable to change the perspective to anomaly detection. In this perspective, a model is learned that basically describes the majority class only. For a given instance, we then test whether it is inside the expectation of the model or an anomaly.

2.3 Abstaining

While the methods from the previous section are helpful in order to deal with class imbalance, they are not designed to work with blurry classification schemes in which not every instance can be safely assigned to a class. For example, it is –in general– not possible to assign a class like *commercial* or *residential* to each and every building. Some are different (e.g., industrial) and some are mixtures (e.g., a shop and some apartments). The situation gets even worse when the relation between the observation and the problem is not clear: while some tweets will contain information about building functions, there are also tweets that do not contain such information at all. Therefore, we should not expect that the classification can be performed for each instance. This ability of abstaining from classification has been well studied in decision theory (Chow, 1970) and has been successfully applied in diverse domains (Pietraszek, 2007). Given a probabilistic classifier ϕ that assigns a class probability vector ϕ_i to an instance x_l, we can first inject a vector of decision thresholds $0 \leq \tau \leq 1$ as it is described in Eq. (1).

$$y_l = \arg\max\left(\frac{\phi_i(x_l)}{\tau_i}\right), 0 < \tau_i \leq 1 \tag{1}$$

This vector τ can be used to vary the weight of probabilities per class. We will use an optimization based on mutual information to find good values of τ while it is possible to manually adjust this vector.

In abstaining classification, this rule is being extended to include the case of an additional class $m+1$. This represented in Eq. (2).

$$y_l = \begin{cases} \arg\max\left(\frac{\phi_i(x_l)}{\tau_i}\right) & \text{if } \max\left(\frac{\phi_i(x_l)}{\tau_i}\right) \geq 1 \\ m+1 & \text{else} \end{cases} \tag{2}$$

In general, the vector τ can be selected in various ways based on domain knowledge or by optimizing case by case. In contrast to such subjective choices, Zhang and Hu proposed a strategy based on information theory alone and showed that it is comparable to the best known techniques including SMOTE (Chawla et al., 2002), Chow's rejection rule (Chow, 1970), as well as to rejection based on the geometric mean over a large range of datasets covering single-class and multiclass problems as well as abstaining and non-abstaining situations. We adopt this mechanism, because it is completely parameter-free and clearly rooted in theory.

Normalized mutual information is a traditional measure for the degree of dependence between two random variables T and Y. It is defined to be

$$\mathrm{NI}(T,Y) = \frac{I(T,Y)}{H(T)},$$

where H is the Shannon entropy, descried in Eq. (3), and I is the mutual information, which is described in Eq. (4).

$$H(T) = -\sum_{i=1}^{m} P(T=i)\log_2 P(T=i) \tag{3}$$

$$I(T,Y) = \sum_{i=1}^{m}\sum_{j=1}^{m+1} P(T=i,Y=j)\cdot \cdot\log_2\frac{P(T=i,Y=j)}{P(T=i)P(Y=j)} \tag{4}$$

In general, it is difficult to calculate the involved probabilities. Still, Hu et al. proposed an empirical estimation based on the entries of the confusion matrix.

T	**Y**				
	1	2	…	m	$m+1$
1	c_{11}	c_{12}	$\cdots$	c_{1m}	$c_{1(m+1)}$
2	c_{21}	c_{22}	$\cdots$	c_{2m}	$c_{2(m+1)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
m	c_{m1}	c_{m2}	$\cdots$	c_{mm}	$c_{m(m+1)}$

Table 1. Representation of an confusion matrix.

Given a confusion matrix, such as in Table 1 of T and Y with an additional column $m+1$ covering the possible cases of abstaining, mutual information can be approximated by Eq. (5) according to (Bao-Gang et al., 2012).

$$I(T,Y) \approx I(C) = -\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\log_2\left(\frac{c_{ij}}{C_i\sum_{i=1}^{m}\frac{c_{ij}}{n}}\right)}{\sum_{i=1}^{m} C_i\log_2\frac{C_i}{n}}, \tag{5}$$

where C_i are the sum of the i-th row and $n = \sum_i\sum_j c_{ij}$ is the total number of samples. Note that

the second sum goes to m instead of $m+1$, which is not rigorously correct, but overcomes the limitation of NI not changing value if rejections are made within a single class, compare (HU and WANG, 2008).

Using this measure as a measure for the quality of an abstained classifier, we can optimize the value of τ.

Optimizing Abstaining Classifiers: A central challenge in cost-sensitive and abstaining classification is to assign the weightings or costs in an optimal manner. We apply a simple grid search and Powell's algorithm (Powell, 1964) in order to optimize for the best classifier, that is, the classifier using a threshold vector τ such that its decisions maximize normalized mutual information with the ground truth, see Eq. (6).

$$\tau^* = \arg\max \mathrm{NI}(t, y = \phi^\tau(x)) \tag{6}$$

The ϕ^τ is given by abstaining from classification for a probabilistic classifier ϕ from Eq. (7).

It is worth noting that normalized information is biased towards the minority class. That is, abstaining will improve the error behavior of the minority class more than of the majority class in unbalanced situations as ours.

2.4 Ensembling Models

Ensembling many weak classifiers in order to obtain a better overall classification has long been discussed. For example, in 1984 Granger already writes: "The common practice, however, is to obtain a weighted average of forecasts [...]". That is, already in 1984 it was widely accepted that averaging machine learning models increases the performance.

While ensembling can be formulated quite generic by saying that ensembling covers the case of building a novel classification problem by applying several classification models and building the model from their output or intermediate information, we concentrate on several basic approaches in order to remedy the impact of singular choices.

The simplest way of combining probabilistic models is through averaging. Given n probabilistic classifiers $\phi_1 \dots \phi_n$, the classifier

$$\phi_*(x) = \frac{1}{n} \sum_i \phi_i(x) \tag{7}$$

is a probabilistic classifier which is surprisingly strong, especially when the individual classifiers ϕ_i show a certain diversity (good performance, but low pairwise correlation). This approach is also known as *model blending*. A more involved approach is to use the classifiers ϕ_i to generate a novel machine learning problem, namely, predict y from the vector $\phi_i(x)$. A typical choice is to use logistic regression for this step. In contrast to the model blending approach, this way of model ensembling is more stable with respect to correlated classifiers and can perform more complex model combinations. It is also known as *model stacking*.

While there are many other methods of model ensembling, the given methods are chosen for their unbeaten performance given their simplicity and the fact that they do not need too much additional data for training and verification.

3 Dataset Description

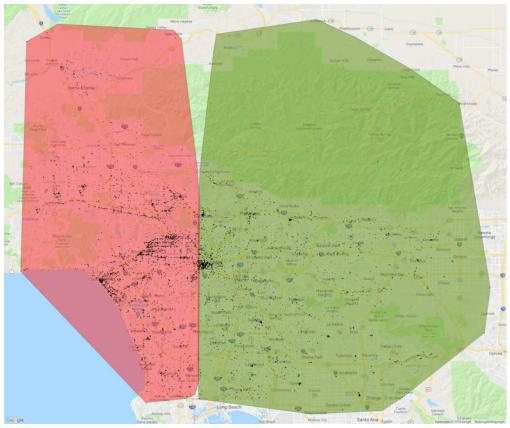
During this experiment a dataset has been created where a million of tweets has been assigned to two classes in Los Angeles. Therefore, we first collect geolocated tweets in this area, relate them to the nearest building in OpenStreetMap, and prepare a text mining problem by assigning the functional class of a building as derived from OSM to the nearby tweets.

Twitter Data Preparation: The public Twitter API provides a function for streaming up to one percent of all tweets published on the Twitter platform. We collected a dataset of nearly 4TB of tweets using this API. From this dataset, we filtered only those tweets that are published with a precise geo-location. We expect that most of these tweets are associated to the location, which has been assign by the user or the application itself. It needs to be mentioned that this is not true for all tweets. For example, Twitter bots can create arbitrary spatial patterns by publishing tweets in fake locations. Still, we assume that a significant amount of geo-located tweets originates from this location and the fact that we can reach high classification precision from Twitter text alone confirms this assumption.

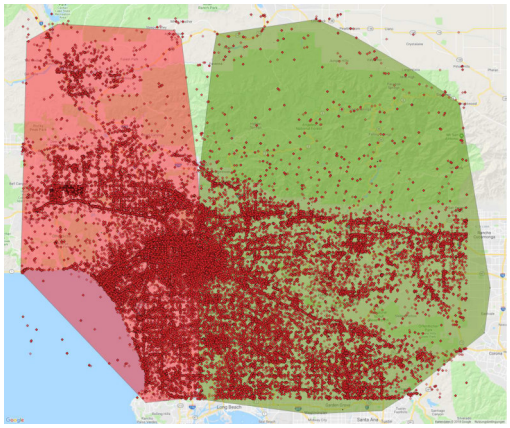
OSM Buildings: In Los Angeles, our study region, the OpenStreetMap contains $24,898$ building polygons that are clearly specified as *residential* or *commercial*. With the study (see Section 4), we concentrate on those buildings and assume that the labels assigned by the OSM community are largely correct.

Spatial Join of Tweets and OSM Buildings: In the spatial nearest neighbour join phase, we assign each tweet to the nearest building of the OSM building polygons dataset. After joining, we remove assignments that appear to be too far away by introducing a pseudo-distance threshold of 0.001, which corresponds to roughly 100 meters. This distance is measured as the Euclidean distance in the WGS84 coordinate space and, therefore, has a varying interpretation across Earth.

Dataset Split: To evaluate the system, we train the ensemble using half the available data in a given region and test on the other half. While this type of a spatial train-test split is needed to get reliable estimations of performance, we should avoid having a different distri-



(a) Well-classified OSM buildings and a spatial train-test split.



(b) Tweets collected for the area of Los Angeles.

Figure 1. The dataset of the Los Angeles area including spatial split information, ground truth, and tweet locations (Map Data ©2018 Google).

bution of building or district functions. Figure 1 depicts the chosen split.

Los Angeles Tweets - Dataset: In Los Angeles, we extracted $1,223,037$ precisely geo-located tweets in half a year between November 2017 and May 2018. Furthermore, the dataset has been split vertically in equal partitions and has been balanced such that both classes contain roughly the same number of samples. Each of the splits, resulting from splitting west and east as well as *commercial* and *residential*, contains $16,133$ examples.

In this way, we obtain a dataset in which most buildings are set into relation with many tweets. Figure 2 depicts the distribution of tweets per building for this study. One can see that tweets concentrate on a minority number of buildings. While the average number of tweets per building is 47.23 in this dataset, only 30% of the buildings have more than ten tweets, and 19% have more than 20 tweets and 5.6% have more than 100 tweets.

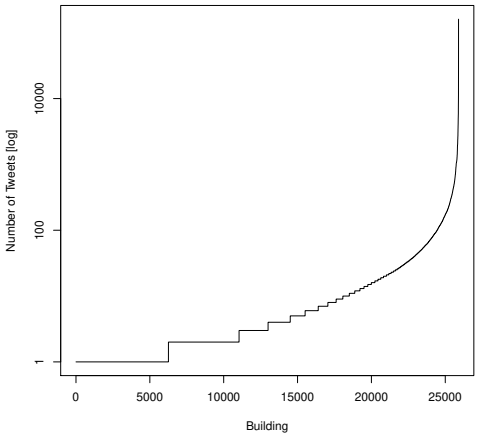


Figure 2. Number of tweets per building for the dataset

4 Case Study in Los Angeles

We decided to answer the question of whether function-related information can be unlocked from tweet text alone in the area of Los Angeles. First, there is a significant amount of social media available, second, the English language is dominant in this area for which the largest collections of text mining training data are available (English Wikipedia and News), and, third, the OSM contains very many buildings explicitly labelled *commercial* or *residential*.

We proceed as follows: first, we fix baselines by applying a wide range of sparse text mining models to the problem. Then, we introduce information-optimal abstaining. Finally, we analyze the behavior of simple ensembles.

4.1 Sparse Text Mining Models

We start by analyzing a single classifier in the area of Los Angeles. We perform analysis on a tweet by tweet basis and do not aggregate tweets per building as we expect that certain tweets lead to abstaining as they are unrelated to the building function while other tweets clearly contain a hint on the building function. First, we extract a sparse matrix based on counting all word occurrences in the tweets and normalizing those using the term frequency inverse document frequency strategy (TF-IDF). This leads to a total of $71,994$ columns and given that tweets are very short texts to a very sparse representation of each and every tweet.

Classifier	Training		Test	
	Commercial	Residential	Commercial	Residential
Ridge	0.97	0.97	0.50	0.50
Perceptron	1.00	1.00	0.50	0.48
kNN	0.72	0.79	0.40	0.57
RF	1.00	1.00	0.49	0.53
X-Tree	1.00	1.00	0.50	0.52
SVC-L2	0.99	0.99	0.50	0.50
SVC-L1	0.91	0.91	0.50	0.51
ElasticNet	0.77	0.70	0.54	0.42
MN-NB	0.99	0.99	0.52	0.52
SVC-L1/2	0.94	0.94	0.50	0.50

Table 2. Precision of selected single classifiers for sparse text representation using described dataset and framework.

We conducted experiments with various classifiers including Ridge regression, a Perceptron trained for 50 steps, a Passive-Aggressive Classifier (Crammer et al., 2006), kNN classification, Random forests with 100 trees, and several support vector machine and neural network classifiers, sometimes regularized with l_1 or l_2 penalties, as well as Naïve Bayes algorithms using Multinomial distributions. In addition, we performed feature selection using an l_1-penalized support vector machine classifier and trained an l_2 penalized model only on the selected features.

While all these classifiers are heavily overfitting the training set and showing poor generalization, we expect that some of the classifications in the test set are not made by chance alone and are going to try to find them via abstaining in the next section.

Table 2 shows results of the classifiers. In general, the picture is clear: The algorithms are highly overfitting on the training split and do not perform significantly better than random on the test set. However, there seems to be some information extracted at least for a few tweets and we want to extract exactly this knowledge using abstaining. It is clear that most of the tweets do not contain information about the building function at all. On the other hand, the classifiers might have collected information in their probabilities that allow us to select those instances where there is information and use those for classification.

We apply information-optimal abstaining, as explained in Section 3,to find a threshold vector τ such that the joint information is optimized. However, we can only apply abstaining to probabilistic classifiers directly. In order to get a clear picture, we restrict attention to those classifiers, where a probability is naturally available. Note however, that many classifiers can be calibrated to give probabilities. Still, this would need another dataset split in order to use different sets for training and calibration. Given the spatial nature of our problem, however, this would greatly reduce the amount of available information for training and at the same time it is unclear whether a calibration on a spatial disjoint set is actually working well. Therefore, we omit this option for urban-scale studies as it is too likely that a spatially disjoint split of the training set does cover different functional regions of a city.

As abstained classification basically tries to increase precision by reducing recall, we shift attention from the F1 score to the per-class precision and recall. We only present numbers for the test set in Table 3. The multinomial Naïve Bayes has been trained with different Laplace smoothing parameters (MN-NB1 with 0.001, MN-NB2 with 0.01, MN-NB3 with 0.1). This parameter steers how the probabilities are adjusted for words in the test set that have not been in the train set. Of course, this has an interesting effect on abstaining as it directly modifies the probabilities.

As you can see from Table 3, a larger smoothing parameter for Multinomial Naïve Bayes increases the information-optimal abstain rate. In essence, the precision of the minority class increases while the recall is decreasing. One also observers that the classifiers based on regularized stochastic gradient decent show good values for the minority class.

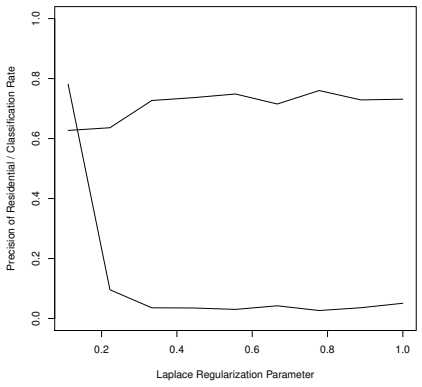


Figure 3. Laplace Smoothing of Naïve Bayes and Abstain Rate

Classifier	Abstain-Rate	Commercial		Residential	
		Precision	Recall	Precision	Recall
MN-NB1	63%	0.54	0.19	0.57	0.21
MN-NB2	72%	0.53	0.14	0.58	0.17
MN-NB3	89%	0.55	0.04	0.62	0.09
SGD-L2	99%	0.17	0.00	0.83	0.01
SGD-L1	96%	0.56	0.01	0.76	0.04

Table 3. Precision and recall for the abstained classifiers.

Figure 3 depicts the influence of the smoothing parameter on model behaviour. Larger regularization parameters lead to higher performance for the minority class yet at the same time to a lower fraction of classified elements. That is, choosing a larger number is more conservative.

In summary, this section showed that it is very difficult to learn the association of tweets and building functions. In fact, classifiers quickly overfit the training set and do not generalize. However, the discussion of abstaining classifiers revealed that some knowledge is embedded and that rejecting more examples consistently increases the precision of the minority class to more than 80% for only one percent of the test samples. Still, this means that we are able to assign a label to 168 buildings. Given the fact, that we have many unclassified buildings in OSM, assigning a class with 80% for one percent of those buildings is still a very valuable result and encourages our vision that social media is an interesting data space augmenting Earth observation in urban areas. Combining this with a human operator could speed up building mapping, for example.

4.2 Ensembling Abstaining Models

The previous sections have shown that a multitude of models and approaches is able to unlock a little bit of information about the building function from using Tweet text alone. In this section, we are going to ensemble the various models, because we expect that the information, they learned is different for each model and that they can be combined to a stronger model through ensembling. For the final ensemble in this paper, we applied sparse text mining models on the given dataset of one million tweets near Los Angeles. In a first step, we rejected words from the vocabulary that occur in less than 0.1% of the document and those that occur in more than 20% of the documents. The first rejection threshold is related to rare words that won't generalize and the second threshold has been chosen quite low and is related to corpus-specific stop-words. This includes many hashtags, smileys, and emoticons. This results in a vocabulary of $1,032$ words. With this data, we trained nine different classifiers from the Naïve Bayes family as well as support vector machines

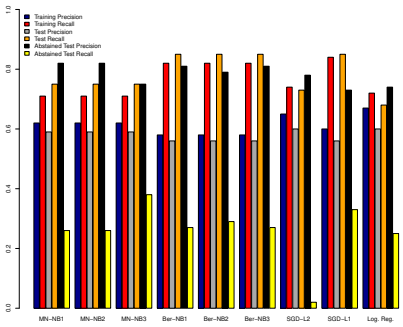


Figure 4. Performance of the Ensembles with Abstaining of different classifiers. Where blue is the training precision, red is the training recall, gray is the test precision, orange is the test recall, black is the abstained test precision and yellow is the abstained test recall.

and logistic regressions. Given the large corpus, these all reached good performances depicted in Figure 4.

This Figure depicts the performance of the minority class (*residential* buildings). We give precision and recall for the training, test, and abstained validation. You can see that with this large corpus of more than one million tweets and by rejecting rare and frequent words, the bias is significantly reduced. Precision and recall of train and test set are in the same range of about 60% to 80%. Information-optimal abstaining was applied to each and every classifier individually leading to high precision values in the range of 70% to 80% with reduced recall. Finally, we build an ensemble of all of these models by building a weighted average of these models. The weights are taken from the expected precision of the *residential* class as we are not interested in high recall, but in very high reliability for a few buildings. This gives us a new probabilistic classifier P. This results in a classifier with 85% accuracy and a recall of 2% translating to 1937 classifications that have been done with this performance. In all other cases, the classifier decided that there is not enough certainty to assign a class and abstained from classification. Still, we can also apply the information-optimal abstaining machinery to this ensemble classifier. As we already put a lot on emphasis on the *residential* class, it is not surprising that this one is worse on the minority class as opposed to the previous one, which largely

Classifier	Abstain-Rate	Training		Test	
		Commercial	Residential	Commercial	Residential
BIRP	54 %	0.60	0.78	0.82	0.26
HRF1	58 %	0.70	0.23	0.75	0.38
AVE	-	0.59	0.85	0.73	0.41
AVE-A	16 %	0.61	0.72	0.75	0.37
AVE-F1	-	0.59	0.86	0.74	0.52
AVE-F1-A	16 %	0.61	0.72	0.75	0.37

Table 4. Final Classification Ensembles.

ignored negative effects on the *commercial* class by averaging using *residential* layer performance. Still, it is a very good, general purpose classifier leading to 61% precision for the *commercial* class with a recall of 72% and 75% precision for the *residential* class with a high recall of 37%. In fact, this enables us to classify $32,342$ buildings in Los Angeles with an expected precision of 75%. By the way, this is a very nice example of the balancing effect of information-optimality. As long as one class is dominant (in the beginning, this was the *commercial* class) it helps to focus on the less dominant class. But, when you cross a point where you lose too much performance for the *commercial* class, it starts doing the reverse. Table 4 lists the final performance measures for the ensembles studied in this section in comparison to selected members including the best individual *residential* precision (BIRP) given by Multinomial Naïve Bayes with smallest smoothing parameter, the highest *residential* F1 (HRF1) given also by Multinomial Naïve Bayes, but with the largest smoothing parameter. In addition, we give ensembles with basic model averaging (AVE), with averaging based on the expected F1 score (AVE-F1). Each of those is also given with an information-optimal abstaining variant (*-A).

This family of final classifiers provides different tradeoffs for the given problem. While the best individual *residential* precision classifier (BIRP) is a very good general classifier, it suffers from a high abstaining rate of 54%. Similarly, the best individual classifier measured using the F1 score on the test set (HRF1) is abstaining in 58% of the cases. Building ensembles through averaging, however, increases the support of the classifier significantly. In summary, all of those classifiers perform very good irrespective of the weighting scheme. The precision of the *residential* class is in the range of 73% -75% while the recall ranges from 37% to 52%. It turns out that an averaged ensemble with weights given by the F1 score presents a good general candidate. It is worth noting that though the table does not show abstaining rates for the averaged ensembles that do not abstain in the ensembling step, the individual models of the ensemble are all abstaining. It is an interesting coincidence that applying information-optimal abstaining also to the ensemble is bringing the ensembles into the same working region reducing the impact of the actual weighting.

5 Conclusion and Future Work

With the work in this paper, we have shown that social media text actually contains information about building functions. However, it is very difficult to spot and advanced techniques like abstaining are needed in order to remove the many text messages that are simply unrelated to the nearest building of their occurrence. Without such techniques, all models trained to high performance, but did not generalize at all. However, this lack in generalization is not so much related to the models not learning something, but rather to the fact that forcing a classification of unrelated tweets results in random choices and significantly influences performance.

We were able to construct a diverse set of models through abstaining and model blending. Some of them have high precision, but classify only few instances at all, others have lower precision, yet better recall.

Given the absolute numbers, however, we can also conclude that social media text alone is not sufficient to understand building functions. Though we never expected this, it is an important fact to know when considering data fusion. In the future, we plan to study social media text in the context of data fusion with satellite data, mainly Landsat and Sentinel. An interesting area of research is about techniques that combat spatial overfitting. While it was possible for Los Angeles to find a spatial split such that both parts of the city have similar settlement patterns, this is not true for many other cities. Therefore, novel techniques for sampling need to be developed and applied in order to prevent spatial overfitting and to predict performance of classifiers on unseen data. Due to some local aspects of language, however, we might actually want to allow some spatial overfitting. This leads to questions of how global ensembles can be constructed and evaluated that can deal with all languages of the world and reliably aid in the building classification task on a global scale.

In addition, advanced ensembling methods like boosting or stacking can be taken into account, though av-

eraging is known to work surprisingly well in many small-data situations. In fact, the boosting or stacking steps might increase the risk of overfitting as they are implicitly consuming additional data. We think that this additional held-out data is possibly better invested in finding the abstaining parameters or training the underlying models.

6 Data and Software Availability

The source code is accessible via: https://github.com/mwernerds/agile21_abstaining under MIT license, respectively some files are made public under BSD license.

Social media data cannot be made accessible due to GDPR and Twitter license regulations. Open Street Map data exported under ODbL 1.0 License from.

References

Bao-Gang, H., Ran, H., and Xiao-Tong, Y.: Information-theoretic measures for objective evaluation of classifications, Acta Automatica Sinica, 38, 1169–1182, 2012.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321–357, 2002.

Chow, C.: On optimum recognition error and reject tradeoff, IEEE Transactions on information theory, 16, 41–46, 1970.

Cohen, B.: Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability, Technology in society, 28, 63–80, 2006.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y.: Online passive-aggressive algorithms, Journal of Machine Learning Research, 7, 551–585, 2006.

Crooks, A., Croitoru, A., Stefanidis, A., and Radzikowski, J.: # Earthquake: Twitter as a distributed sensor system, Transactions in GIS, 17, 124–147, 2013.

Elkan, C.: The foundations of cost-sensitive learning, in: International joint conference on artificial intelligence, vol. 17, pp. 973–978, Lawrence Erlbaum Associates Ltd, 2001.

Go, A., Bhayani, R., and Huang, L.: Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, 1, 2009.

Han, H., Wang, W.-Y., and Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: International Conference on Intelligent Computing, pp. 878–887, Springer, 2005.

He, H. and Garcia, E. A.: Learning from imbalanced data, IEEE Transactions on knowledge and data engineering, 21, 1263–1284, 2009.

He, H., Bai, Y., Garcia, E. A., and Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 1322–1328, IEEE, 2008.

HU, B.-G. and WANG, Y.: Evaluation criteria based on mutual information for classifications including rejected class, Acta Automatica Sinica, 34, 1396–1403, 2008.

Japkowicz, N.: The class imbalance problem: Significance and strategies, in: Proc. of the Int'l Conf. on Artificial Intelligence, 2000.

Kiermeier, M., Werner, M., Linnhoff-Popien, C., Sauer, H., and Wieghardt, J.: Anomaly detection in self-organizing industrial systems using pathlets, in: Proceedings of the 18th Annual International Conference on Industrial Technology (ICIT), pp. 1226–1231, IEEE, 2017.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C.: Learning Word Vectors for Sentiment Analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150, Association for Computational Linguistics, Portland, Oregon, USA, http://www.aclweb.org/anthology/P11-1015, 2011.

Mitchell, T.: Twenty Newsgroups Data Set, available: https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T.: Text classification from labeled and unlabeled documents using EM, Machine learning, 39, 103–134, 2000.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., and González, M. C.: Urban magnetism through the lens of geo-tagged photography, EPJ Data Science, 4, 5, 2015.

Pietraszek, T.: Classification of intrusion detection alerts using abstaining classifiers, Intelligent Data Analysis, 11, 293–316, 2007.

Powell, M. J.: An efficient method for finding the minimum of a function of several variables without calculating derivatives, The computer journal, 7, 155–162, 1964.

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., and Camps-Valls, G.: Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources, Information Fusion, 63, 256–272, 2020.

Taubenböck, H. and Wurm, M.: Globale Urbanisierung–Markenzeichen des 21. Jahrhunderts, in: Globale Urbanisierung, pp. 5–10, Springer, 2015.

Tomek, I.: Two modifications of CNN, IEEE Trans. Systems, Man and Cybernetics, 6, 769–772, 1976.

Veloso, M., Phithakkitnukoon, S., and Bento, C.: Urban mobility study using taxi traces, in: Proceedings of the 2011 international workshop on Trajectory data mining and analysis, pp. 23–30, ACM, 2011.

Yuan, J., Zheng, Y., and Xie, X.: Discovering regions of different functions in a city using human mobility and POIs, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 186–194, ACM, 2012.

Zhu, Y. and Newsam, S.: Spatio-temporal sentiment hotspot detection using geotagged photos, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 76, ACM, 2016.