GI_Forum 2021, Issue 1
Page: 68 - 75
Research Paper
Corresponding Author:
martin.werner@tum.de
DOI: 10.1553/giscience2021_01_s68

Spatially Supervised Text Mining for Social Media Cleaning and Preprocessing

Martin Werner[1]
[1]Technical University of Munich, Germany

Abstract

In this paper, we show a framework for partial bot rejection based on spatially supervised text mining from social media messages. We show qualitative results towards the reduction of known bots and give hints on how this cleaning technique can help us in filling gaps of current signals related to human life on Earth based on social media. The bot rejection framework is based on using a spatial signal for supervising a machine learning model with extreme label noise still being able to reject some of the unwanted components of the social media stream. Furthermore, we comment that such models show significant biases and can, therefore, not be used responsibly without bias analysis and mitigation per application.

Keywords: social media analysis, text mining, data cleaning

1 Introduction

Urbanization is one of the most pressing and challenging megatrends for human life on Earth. As depicted in Figure 1, the rural population has constantly been increasing up to today, but with a slowing effect, it is expected to start decreasing by the mid of the current century. In contrast, the urban population is expected to have at least linear growth in the time such that by 2050 urban areas give a home to more than double as many people as the rural areas (United Nations Department of Economic Affairs, 2018). Moreover, the local dynamics of this development are surprising, if not daunting. For example, it is expected that Delhi, India, will become the largest city by 2030, overtaking Tokyo. In 2018, however, the United Nations report 37.4 million inhabitants for Tokyo and only 28.5 million for Delhi. The expectation formulated for 2030 is that Tokyo will shrink to 36.5 million inhabitants while Delhi will grow to nearly 39 million inhabitants. This is a growth of 11 million inhabitants in 12 years or about one million inhabitants per year. This extreme local variability of the dynamics implies heavy challenges, for example, for the transport system (food, mobility, waste disposal etc.), for the infrastructure (electricity, water, healthcare, police, etc.), and for the environment (e.g., air and water pollution).

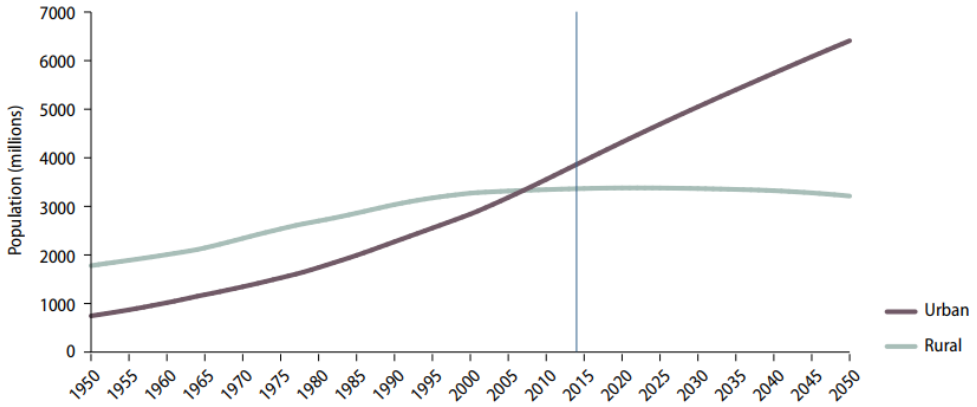
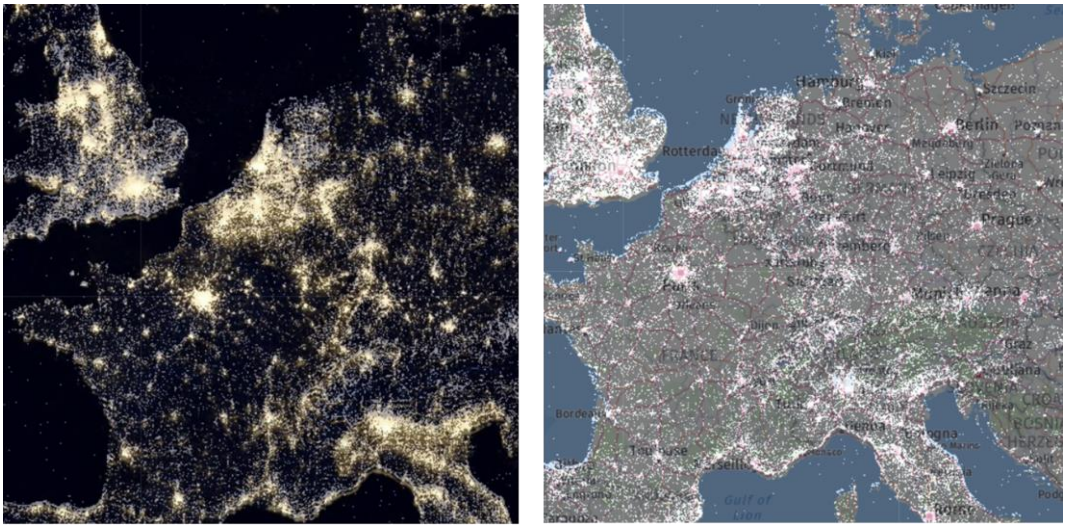


Figure 1: Global Urban Population Compared to Rural Population - 1950 - 2050 as Expected by the United Nations.

The United Nations have established 17 Sustainable Development Goals (SDGs), many of which have strong interaction with the process of urbanization (United Nations, 2019). For example, urbanization is related to zero hunger and no poverty, as the hope for jobs and fleeing from rural poverty is one reason people move into the city. Good health and wellbeing, as well as quality education, are challenged as well because these rely on infrastructures that might be difficult to grow at the needed pace and at the same time motivate people to relocate to the urban areas. Furthermore, clean water and affordable and clean energy is similarly challenging as the energy density needed in megacities is difficult to provide with renewable energies today The consequences of quick urbanization processes directly challenge sustainable cities and communities, climate action, life on land, and life below water in terms of pollution.

In order to cope with this situation on a global scale, innovative methods of data acquisition and data analysis are needed, which go beyond the current observational capabilities mainly based on remote sensing from space. Because these overhead observation systems do not observe the process of urbanization, but rather the impact of urbanization on morphological structures, while it is comparably easy to see cities grow from a spaceborne platform, it might be difficult to get a reliable signal on the expected minor shrinkage of Tokyo. It is unlikely that this will result in a major change in the morphology. Therefore, we propose and follow a different path of using additional signals with strong anthropogenic components to better understand these dynamics.

One such signal is represented by night light observations as, for example, provided by NASA and NOAA. These images represent the amount of light emitted at night, which correlates with human settlements quite strongly. In addition, the amount of light has been used to estimate census parameters in the United States.The more light is being observed, the higher the population density and the average income (Chen & Nordhaus, 2019). Figure 2a depicts an example of such night light observations. The limitation of these observations is twofold: long integration times are used in order to come up with clear signals, and the resolution remains limited. That is, light gives us kind-of an upper bound to the urban extent as light is among the first persistent signals in settled areas.



a) NASA Night Light Map Over Europe

b) Twitter Occurence over Europe

Figure 2: NASA Night Light Imagery and Twitter Occurrence over Europe.

Another promising signal can be extracted from social media depicted in Figure 2b. Social media message frequency also correlates to population density in areas of social network adoption (Li et al., 2013). However, social media is full of special noise patterns induced by a high number of bots sending messages and frequent trends that have a varying spatial resonance ranging from global (#metoo) to very local resonance (e.g., hashtags related to local events). Therefore, long integration times are needed as well, such that the social media data represents a reasonable average behaviour. However, when looking closely at densities, it seems that social media is more focused on city centres and, therefore, a more selective signal compared to night light emission None of these signals can truthfully represent the sociodemographic indicators of interest, including population density, wealth, and income, but all of them show a slightly different pattern of correlation with these signals of interest. Therefore, we expect a joint observation of all of these signals towards unexpected diverging patterns is a suitable monitoring aid for systematic urbanization analytics.

This paper shows how a spatial knowledge injection method applied to text mining can be used to reduce some unwanted signals from social media, making social media a more reliable signal. In order to clean up the Twitter signal, the aim is to remove components that are just due to bots or automated messaging. In order to detect a component of such bot messages, we apply text mining to the social media messages in order to detect a very spurious pattern of bots, namely, that many bots are not using sensible location information. We learn a bot rejection model based on training it with all precisely geolocated tweets based on whether the location is over land or ocean. While some of these messages over the ocean might originate from shipping, many of these messages are expected to be blurring the patterns of urbanization we want to observe.

2 Methodology

2.1 Datasets

In this study, we use three datasets. The first dataset is the NASA/NOAA Night Light Imagery for 2018. It represents the average light emission in 2018 across the globe in a medium resolution of about 500m per pixel (of course varying across the globe due to the WGS84 projection). This data has been acquired by the Suomi NPP satellite and processed by NASA to account for the moon phase dynamics trying to normalize towards a moon-phase independent representation of the light emission. This dataset comprises 3.73 billion pixels. The second dataset is a sample of all observed social media messages throughout 2018 acquired from the public Twitter stream, representing about one per cent of the total social media messages on this platform. We sampled a set of 220 million precisely geolocated tweets (note that these include bots and retweets due to the specification of the stream API endpoint) and process both the geospatial location and the raw text, including hashtags and punctuation in all observed languages. The third dataset is the dataset representing country boundaries across the world. For this purpose, we take the LSIB 2017 Large Scale International Boundary Polygons Dataset as published by the United States Department of State at the Office of Geographer. It presents 284 countries in 312 features modelled with 2,342,905 points.

2.2 Labeling

In a first step, we label Twitter data from the first three months based on the country dataset in two categories: land and water . As we already expect very high label noise in this dataset as some tweets might be from very good bots or human beings around the ocean, we do not create geospatial buffers around the countries to take care of coastal areas into account. Instead, we rely on the fact that most tweets in the ocean are observed far enough from the nearest country In order to do this efficiently, we need to rely on a dedicated implementation based on well-performing bulk loaded in-memory R*-trees to speed up point in polygon queries. We rely on HDF5 and boost::geometry for the core operations and modern C++, including OpenMP for parallel processing. We follow a strict property map interface, that is, records that are implicitly linked by their primary key, which is just the row number in the memory block allowing for constant-time access to individual records. With an average Gaming PC (Intel i7, 32 GB RAM), we process the point in polygon join in this way in 8 hours without simplifying geometry. The resulting dataset is heavily imbalanced, with only 5.7 million tweets observed over water. Hence, we then create a class-balanced dataset by sampling alternating between land and water classes such that we gain a temporally ordered sub-dataset with the same numbers of water and land classes and a total of 11 million[1].

[1] Source codes and details of this project and are available at https://www.bgd.lrg.tum.de/code/2021-landwatersplit.

2.3 Text Mining

The data mining problem induced by the labelling process is to develop a text mining model that can be applied across many languages, including non-human languages like hex-codes observed for some bots. As explained, we have now a labelled dataset of tweets based on whether it was observed over the ocean or a country. In a second step, we train a skip-gram model with subword information on the tweet text towards detecting the class "water" or "land" (Bojanowski et al., 2016). This model is based on cutting text into small pieces of n consecutive characters, so-called n-grams, and assigning a randomly initialized vector of chosen dimension with each n-gram. Then, we minimize an objective function using a variant of gradient descent which balances two aspects: one loss term pulls vectors associated with textually nearby n-grams (those that appear not farther away than a chosen parameter "context window" in the text) towards each other minimizing their Euclidean distance in the embedding space while a second loss term compares with random non-neighbouring word vectors and pushes the representing vectors away from each other. Word embeddings obtained in this unsupervised way are then used to numerically represent words or sentences (by taking the mean of the words or n-gram tokens). We apply a deep neural network with one softmax layer to directly transform these learnt word embeddings into a classification result for tweets. As expected, the model's performance is not excellent, as calling for a land/water split from textual data is not plausible. Nevertheless, it gives us an interesting signal regarding the trustworthiness of tweet messages, as we explain in the sequel. More concretely, we train a model with an embedding dimension of 10 and tune parameters for an optimal overall F1-score. Therefore, we train on the first million entries in the balanced sample, use the second million entries to validate hyperparameters, and evaluate over time in slices of one million tweets. Results are depicted in Figure 3. The model reaches a performance of about 0.8 F1-score, keeping in a window of less than 5 per cent around. It is interesting to see that numbers degrade only a neglectable amount over time and stabilize around 0.80 overall F1 quickly. This is a hint that only a small fraction of the model does not generalize over time.

Furthermore, it is nice to see that the precision of the water class is higher instead of the land class. The surprising characteristics of this model are visualized as well in Figure 3 as a ROC curve which shows the behaviour of the false-positive rate as opposed to the true positive rate when changing the threshold parameter τ at which the devision between land and water is made. Depending on the actual application and its demands, a suitable τ can be chosen to trade-off precision and recall.

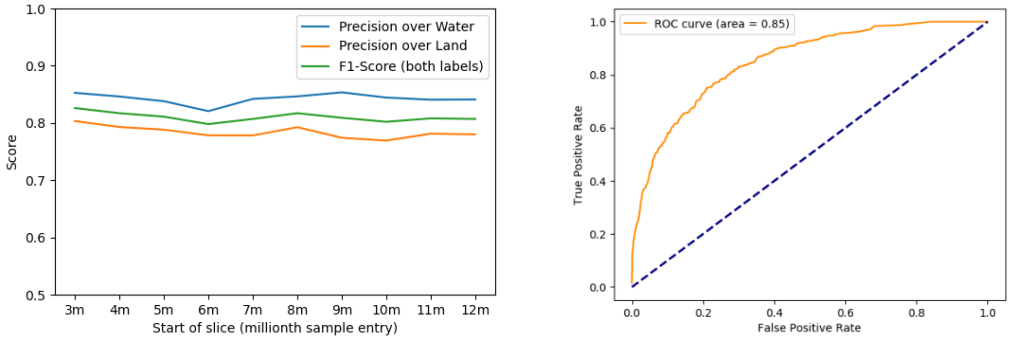


Figure 3: Performance of the classifier over time and in relation to choosing a classification threshold

3 Results

We apply this model over land and reject tweets that are similar to those observed throughout the oceans. Figure 4 depicts an application of this framework to a one-month data sample taken from the Twitter social network. That is, we trained in the past and take fresh data and classify it into the two classes "ocean" and "land". This figure is representative of all the one-million slices.

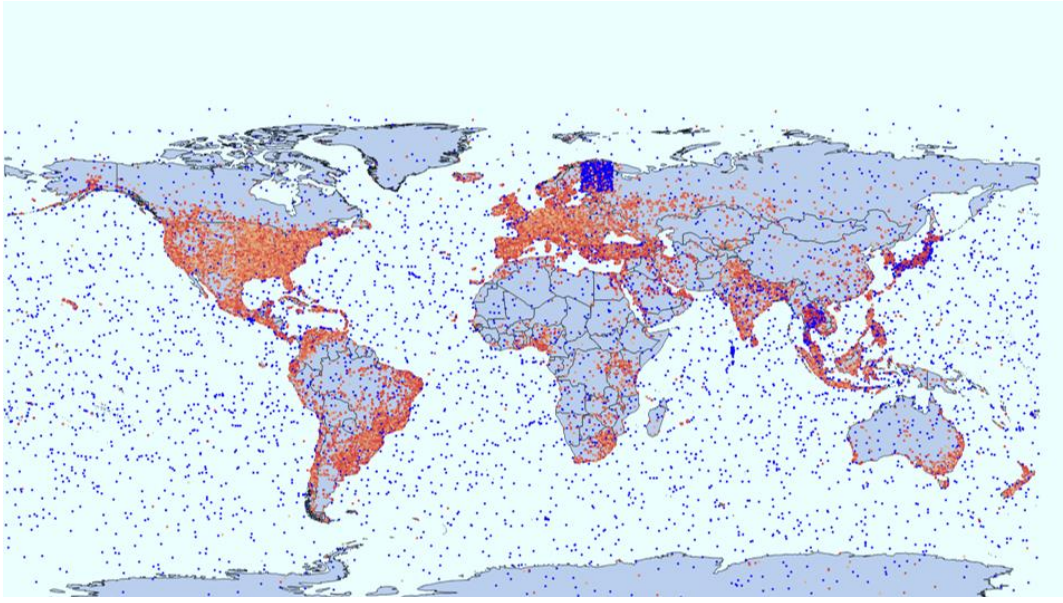


Figure 4: Illustration of Bot Rejection Result on a One-Month Data Sample.

Without knowing exactly, what the model rejects over land, Figure 4 shows the behaviour of the trained model. As one can see, the model predicts low values over the ocean and higher values over land while it predicts surprisingly low values, for example, for the rectangle over Finland, which represents a known bot using fake locations from this rectangle. This illustrates that a spatially semisupervised bot rejection scheme is able to correctly reject some of the fake messages that we observe in social media datasets. At the same time, however, it is easy to see some unwanted results. For example, in Japan and more generally around Asia, we are rejecting many more tweets as in countries with western languages. This is a severe bias, which is easy to explain. Most of the Twitter social network data is communicated in the English language, and non-western languages take only a small fraction of the data. Therefore, the model is overfitted to English (or more generally Western) languages and has problems learning Asian languages from the given sample or because of the pictographic script. Still, with a thorough case by case evaluation, it seems to be viable to apply this model at least in Europe and the United States and it can, for example, enable the detection and analysis of urban structures below the very noisy Finland bot which is difficult without such a scheme.

Further research is needed to assess for each possible social media mining application independently whether such a bot rejection scheme is helpful (increasing correlation) or not (e.g., ethically unsound due to biases) and where to put the threshold on the bot scores. This is a difficult question that needs to be answered in the light of individual applications as it depends on the spatial integration area (how much data is left for further analysis in each analysis unit), the spatial focus (are we interested in the city centres, where social media presents a strong signal or more in the extended urban space and the borders of cities, where social media messages become rare). Furthermore, the rejection scheme puts a tradeoff between preprocessing and data mining in the sense that even if the model was correctly able to reject tweets originating from bots, it would as well reject some messages (false positives) that weaken the spatial signal. Therefore, a selective threshold leads to less data in the following data-mining stag;, a weak threshold reduces the impact of the current approach. Finally, one might want to probabilistically calibrate the classifier and use the calibrated scores for upstream processing instead of simple thresholding. This might mitigate some difficulties of setting a threshold but implies a more complex input of weighted messages to the upstream data mining stages.

4 Conclusion

This paper explored how the injection of spatial knowledge into a text mining problem through labelling can help filter streams of location-based social network messages sensibly. We were able to reject the most obvious bot over Finland. We were able to reject the most obvious bot over Finland. This qualitative result is not enough to understand the behaviour of this model. We will emphasise possible applications in future work, especially towards propaganda awareness, social media trend analysis, outlier and event detection, and land cover classification. This is, to the best of our knowledge, the first time that a spatially semisupervised bot detection and rejection model was designed and showed to perform well with an area under curve measure (ROC_AUC) of 0.85. For clarity, we do not claim that this model rejects bots. Any claim towards this direction would ignore that language models like GPT-3 (Brown

et al., 2020) and BERT (Devlin et al., 2018) can generate text in a quality that is nearly indistinguishable from human text and that human beings are often steering bot networks to, for example, disseminate fake news or bots just pick up valid messages for retweets. We claim to be helpful to filter a very specific component of communication samples that overlaps with bots. We envision using this framework of spatial supervision as well beyond social media classification.

We expect that models that allow us to observe and compare anthropogenic signals from a multitude of decoupled sensing systems (social media, light, activity, prosperity, …) help to put in place global indicators for many of the United Nations Sustainable Development Goals, most importantly, "sustained communities" and "life on land". However, more research in bias estimation, de-biasing, and more generally in the ethical implications of using social media signals is needed before a wide adoption is encouraged.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, X., & Nordhaus, W. D. (2019). VIIRS nighttime lights in the estimation of cross-sectional and time-series GDP. *Remote Sensing*, *11*(9), 1057.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

United Nations. (2019). *Sustainable Development Goals*. https://sustainabledevelopment.un.org/

United Nations Department of Economic Affairs. (2018). *2018 Revision of World Urbanization Prospects*. https://population.un.org/wup/