Crowdsourcing the Character of a Place: Character-Level Convolutional Networks for Multilingual Geographic Text Classification

Benjamin Adams*
Department of Geography
University of Canterbury
Christchurch, New Zealand
benjamin.adams@canterbury.ac.nz

Grant McKenzie
Department of Geographical Sciences
The University of Maryland
College Park, MD, USA
gmck@umd.edu

PRE-PRINT

* Corresponding author

Abstract

This article presents a new character-level convolutional neural network model that can classify multilingual text written using any character set that can be encoded with UTF-8, a standard and widely used 8-bit character encoding. For geographic classification of text, we demonstrate that this approach is competitive with state-of-the-art word-based text classification methods. The model was tested on four crowdsourced data sets made up of Wikipedia articles, online travel blogs, Geonames toponyms, and Twitter posts. Unlike word-based methods, which require data cleaning and pre-processing, the proposed model works for any language without modification and with classification accuracy comparable to existing methods. Using a synthetic data set with introduced character-level errors, we show it is more robust to noise than word-level classification algorithms. The results indicate that UTF-8 character-level convolutional neural networks are a promising technique for georeferencing noisy text, such as found in colloquial social media posts and texts scanned with optical character recognition. However, word-based methods currently require less computation time to train, so are currently preferable for classifying well-formatted and cleaned texts in single languages.

Keywords— crowdsourcing, convolutional neural networks, text classification, geoparsing, geographic information retrieval, user-generated content

1 Introduction

A vast amount of crowdsourced, geographically-related textual content is generated online, which has created new opportunities to understand how people observe their world. Examples of crowdsourced geographic text include Wikipedia articles, travel blog entries, and Twitter observations. Crowdsourced texts about places from social sensors are often of high value because they are more up-to-date than other authoritative data sources about places (Goodchild, 2007; Sakaki, Okazaki, & Matsuo, 2010). However, a key challenge to utilizing this information is that the manner in which people communicate about places is not standardized, especially in the case of social media data, is often noisy, and can be written in different languages (Yin, Lampert, Cameron, Robinson, & Power, 2012; W. Zhang & Gelernter, 2014). Thus, it often remains difficult to correctly match crowdsourced texts to locations on the earth.

Our record of world events is increasingly granular, not only because of more sensors that measure our environment but also due to the increasing number of human observations of places and events that are recorded on the web (Silva, Martins, Chaves, Afonso, & Cardoso, 2006). This crowdsourced place-based information gives unique insight into how individuals and communities (both physical and virtual) conceptualize knowledge about the world, and allows us to capture the dynamic aspects of places to an unprecedented degree (Sui & Goodchild, 2011). Although in many cases this observational data is explicitly georeferenced, for example in the case of GPS tracked social media posts, a vast amount of this information remains exclusively in a form for human consumption and the place being observed is only implicitly referenced or ambiguous. Sometimes this is due to the fact that an ambiguous place name is referenced in the text, but often there is no place name and the only clue to the location are other features that can be learned from the data. Still, if we can positively identify the location associated with this data, then it can used for many purposes in the areas of geographic information retrieval, user geo-location, and social science and digital humanities studies with big geographic data (Jones et al., 2002; Jones & Purves, 2008; Z. Cheng, Caverlee, & Lee, 2010; Schwartz et al., 2013; Han, Cook, & Baldwin, 2014; Adams & Gahegan, 2016).

As a result, there has been interest in developing methods that can classify unstructured texts in order to determine geographic scope (Monteiro, Davis, & Fonseca, 2016). Traditionally, information retrieval indexing and text classification algorithms operate at the word level where documents are encoded as sparse vectors of unique word counts (or in some cases n-grams) (Sebastiani, 2002). However, much of the crowdsourced textual data about places currently being generated is highly unstructured, especially in the case of social media data, which poses a major challenge to word based classification algorithms (Subramaniam, Roy, Faruquie, & Negi, 2009). Word and n-gram based methods start to fail once we consider the multitude of languages, noisy data (e.g., misspellings), informal language, and other character sets being used, including emoticons, to record our observations of the world.

In this paper we explore the application of character-level convolutional neural network (CNN) models to geographically classify text from microblogs and other crowdsourced data sets like Wikipedia (X. Zhang, Zhao, & LeCun, 2015). CNNs were originally developed for image classification, and in the geosciences have recently been applied successfully to remote sensing image classification and object detection problems (G. Cheng, Zhou, & Han, 2016; Maggiori, Tarabalka, Charpiat, & Alliez, 2017; Nogueira, Penatti, & dos Santos, 2017). However, CNNs have not been used to classify geographic text previously. The benefit of using the character-level CNN method for text is that it is language-independent, and can handle noisy data (Wehrmann, Becker, Cagnini, & Barros, 2017). A key advantage of character-based classification is that very little data pre-processing or cleaning is required to build an effective classifier. Because a deep CNN learns several levels of hierarchical features, a CNN image classifier is robust to noise in the form of individual pixels being flipped. Likewise, a character-based text classifier will be robust to noise in terms of flipped characters. In contrast, in the case of non-CNN word-based methods, misspelled words are either ignored or require the application of language models that accurately correct social media-based spelling irregularities onto a common vocabulary. Despite these advantages, CNN based models are not always the best solution for image or text processing. In particular, this is due to the fact that they require very large training data sets and can be slow to train, even with modern hardware.

The key contributions of this work are as follows:

1. We extend the character-level convolutional network model to take as input any UTF-8 encoded string.
2. We demonstrate that this model is effective for classifying a variety of crowdsourced information about places, without the pre-processing required of word-based methods.
3. We show that the main added advantage for character-level CNN comes for data that is noisy (i.e., character-level errors introduced) or for data sets that have text in multiple languages.

The remainder of the paper is organized as follows. Section 2 describes related work on geocoding/geoparsing of text documents as well as background on CNNs for text classification. In Section 3 we introduce a UTF-8 encoding for a character-level CNN model. The crowdsourced data sets used in our study are listed in Section 4, and some comparative experiments between the UTF-8 CNN model and word-based classifiers are detailed. Finally, we discuss the implication of the results and conclude with future work.

2 Background

In this section we describe related work on text classification using word features, geocoding text, and background information on the use of CNNs for text classification.

2.1 Bag of words text classification

Using machine learing for text classification has long been an area of research, because it supports a variety applications in information retrieval, information filtering (e.g., to remove spam), sentiment analysis, and library science (Sebastiani, 2002). Two commonly used techniques that train on words as features are multinomial naive Bayes (McCallum & Nigam, 1998) and support vector machines (Joachims, 1998).

Naive Bayes operates under the assumption that all the words in a document are conditionally independent, and thus utilizes Bayes theorem to calculate the probability of class labels given a word. The maximum *a posteriori* method is then used to assign the class label to a new document. Despite the assumption of conditional independence, naive Bayes has proved effective for tasks such as spam filtering and is easy to train (Androutsopoulos, Koutsias, Chandrinos, & Spyropoulos, 2000).

A support vector machine (SVM) in its simplest form learns the linear threshold function that minimizes the error when splitting data into two classes (Cortes & Vapnik, 1995). A variety of kernel functions that define similarity functions for the feature space can be plugged into an SVM, which allows the SVM to learn higher-order classifiers irrespective of the dimensionality of the data. For text classification, SVMs work well to classify documents despite the data being sparse and high-dimensional (equal to the number of different words) (Joachims, 1998). Because of the kernel function, an SVM classifier does not need to make the same independence assumption that naive Bayes does.

All machine learning methods that classify text based on word features are usually paired with a prior pre-processing step. The dimensionality of the training data is often reduced by cleaning the data, removing stop words, and performing stemming (Scott & Matwin, 1999).

2.2 Geocoding text

A number of studies have explored geographic classification of textual documents. Rule-based systems for geocoding and geoparsing text have been around for a while but, due to the limitations of simple syntactic matching of place names, in recent years more sophisticated machine learning based methods have been proposed (Amitay, Har'El, Sivan, & Soffer, 2004; Clough, 2005; Melo & Martins, 2017). One ongoing challenge for geoparsing algorithms is place name disambiguation (Overell & Rüger, 2008; DeLozier, Baldridge, & London, 2015; Ju et al., 2016). The problem in that case being that the same place name can occur in many different places around the world. Data from social media and other sources have been used to aid georeferencing of crowdsourced documents, e.g., from Wikipedia (Laere, Schockaert, Tanasescu, Dhoedt, & Jones, 2014). A grid-based method that subdivides the surface of the earth for geolocation of documents was proposed in (Wing & Baldridge, 2011), although they do not use equal area sized grids.

An extensive study of text-based Twitter geo-referencing was done by Han et al. (Han et al., 2014). In that study they showed that georeferenced tweets can be used as a training sample for non-georeferenced tweets. However, it is difficult to recreate their results as the data is not available and importantly it is stated that the geographic classes being studied were not balanced in terms of the number of training examples, thus accuracy based comparisons are not meaningful. Crowdsourced texts, including Twitter and social media data, require a heavy amount of pre-processing and data cleaning when using word- or ngram-based methods (Eisenstein, O'Connor, Smith, & Xing, 2010; Han & Baldwin, 2011; Boyd & Crawford, 2012).

2.3 Character-level CNNs

As an alternative to the word-based methods for text classification, recently character-based methods have been proposed using convolutional neural networks (CNNs). The use of CNNs for supervised learning tasks in image processing and text classification has been around for many years but has seen increasing use in recent years due to fast back-propagation training algorithms implemented on systems with graphics processing units (GPUs), enabling training of much deeper models (Rumelhart, Hinton, & Williams, 1988; LeCun et al., 1989; LeCun, Bottou, Bengio, & Haffner, 1998; Krizhevsky, Sutskever, & Hinton, 2012). For many image classification and object detection tasks deep

CNN-based models have shown state-of-art performance (Girshick, Donahue, Darrell, & Malik, 2014; Szegedy et al., 2015).

Neural networks have also been used successfully to classify textual data. Liu and Inkpen (Liu & Inkpen, 2015) introduced a method for geo-locating users using stacked denoising auto-encoders. Kim (Kim, 2014) demonstrated a simple CNN sentence classifier built on word-level encoding of text. Zhang et al. (X. Zhang et al., 2015) demonstrated that CNN classifiers that are trained at the individual character-level can classify text at a level that is competitive with word-based methods. A vocabulary of 71 characters is used to one-hot encode the text input, which is then trained using 1D convolutional and max pooling layers that are similar in function to the way that 2D CNN image classifiers are designed. In (X. Zhang et al., 2015) the character-level CNN was evaluated on eight data sets ranging in size from 120,000 to 3.6 million documents. The number of classes for the data sets ranged from 2 to 14 classes. The 14 class data set comprised of Wikipedia article abstracts categorized based on DBpedia type. More complicated models have since been introduced: Kim et al. (Kim, Jernite, Sontag, & Rush, 2016) proposed a hybrid character and word classifier that uses character-based CNNs to learn a language model (i.e., the statistical distributions of sub-word information, such as morphemes, over a sequence), which feeds into a word-level recurrent neural network classifier. Character-based methods have also been proposed for other natural language processing tasks, such as named-entity recognition (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016). For a general introduction to deep neural networks and convolutional nets see (Schmidhuber, 2015).

In this work we extend the character-level CNN model from (X. Zhang et al., 2015) to support any UTF-8 encoded input, and explore its usefulness to geographically classify different kinds of crowdsourced documents.

3 Model

3.1 Character-level CNN model

The three main transformation functions in the deep character level CNN are temporal convolution, temporal max pooling, and linear transformation. The temporal convolution function (Equation 1) converts the sequence of input tensors $x_{t=1...N}$ to the outputs $y_{t=1...M}$. K is size of the kernel, δ is the incrementation size for each frame $(1 \dots N)$, allowing for sub-sampling, and b_i and $w_{i,j,k}$ are the weights in the network layer. The size of the output is $M = \frac{N-K}{\delta} + 1$.

$$y_t^i = b_i + \sum_j \sum_{k=1}^{K} w_{i,j,k} x^j_{\delta \times (t-1)+k} \tag{1}$$

The temporal max pooling function reduces the size of the network and helps to prevent overfitting. It simply sub-samples K-sized 1D blocks from the input layer and produces a new output layer based on the maximum value in each block. Equation 2 shows the maxpool function for a given block t of size K.

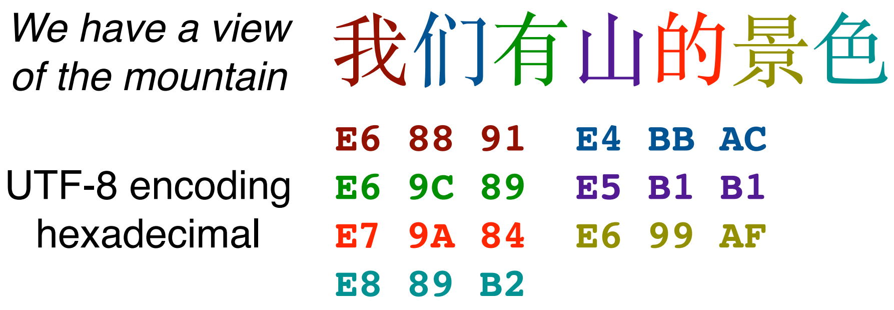


Figure 1: UTF-8 byte encoding for "We have a view of the mountain" in simplified Chinese. In UTF-8 Chinese characters are represented by three bytes of information.

$$f' = \max(f_{(t-1)\times K+1}, \dots, f_{t\times K}) \tag{2}$$

Like in a traditional neural network, the linear transformation represents a fully connected network between the input layer x and output layer y, where $y = Ax + b$ and A are the weights and b are the biases, and the final layer to the output classes is a Log Softmax function shown in Equation 3, where $a = \sum_j [e^{x_j}]$.

$$f_i(x) = log(\frac{1}{ae^{x_i}}) \tag{3}$$

3.2 UTF-8 character encoding model

In our model we extend upon the character-level encoding scheme developed in (X. Zhang et al., 2015), which is limited to the ASCII characters a-z, 0-9, and some additional punctuation characters but ignores other characters. We propose a new encoding scheme that converts any UTF-8 encoded string as an 8-bit sequence (see Figure 1), where each byte is then quantized using a one-hot encoding that is used as the input into the model (Figure 2) (Yergeau, 2003). The input feature length in our model is set to 576 bytes. Since a single UTF-8 encoded character can be one, two, or three bytes that means the model can take as input up to anywhere from 192 to 576 characters depending on the character set being used. This length was chosen to accommodate Twitter tweets of up to 140 characters using character sets in languages composed of mostly 3 byte characters as well as most Wikipedia paragraphs in English. Figure 3 shows the full network we used in training. The convolution kernel (K) and connected network (N) sizes were chosen to correspond to values used in (X. Zhang et al., 2015). This network demonstrates the generic usefulness of the method, however in an application one would want to cross-validate over a variety of alternate network configurations and learning rates in order to optimize results.

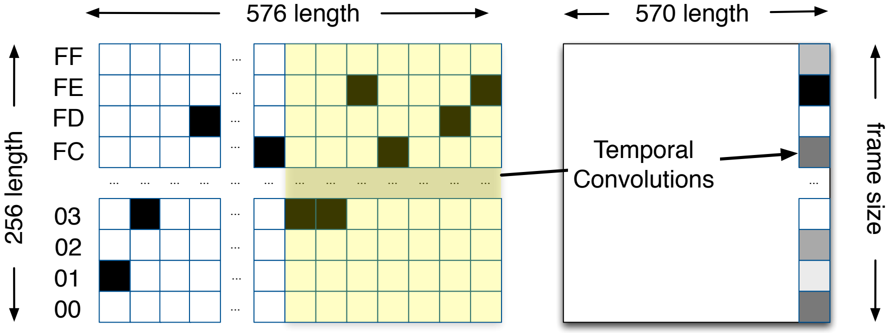


Figure 2: Illustration of one-shot encoding of the 576 byte length UTF-8 encoded input string, and the first temporal (1D) convolution. The yellow highlighted area schematically illustrates how the first convolution layer takes a sliding window of seven input features and applies a convolution function (Eq. 1) to output a weight on the next layer of the network. This is one of a series of transformation layers that are shown in Figure 3.

Word-based text classifiers are usually pre-processed into different languages using a language detection script prior to training. Using the full UTF-8 encoded character set enables our model to automatically learn patterns based on different languages as well as incorporating all possible textual characters, including emoticons (smiley faces, etc.), currency symbols, and other non-language based character sequences. For the latter, Hovy et al. (Hovy, Johannsen, & Søgaard, 2015) found that language forms including emoticon use and style and spelling are indicators of demographic characteristics, including geographic location, of the people generating content on user review sites.

To implement the model, Torch 7[1], a scientific computing framework based on the Lua language, was used (Collobert, Kavukcuoglu, & Farabet, 2011). Torch 7 has deep learning extensions that compile models to run on high-performance GPU-based systems. The code for running the Torch model using NVIDIA's Deep Learning GPU Training System (DIGITS)[2] software can be found online[3]. In the experiments, the models were trained using the AdaDelta learning rate method, because training converged in fewer iterations than stochastic gradient descent (SGD) algorithm with similar accuracy results (Zeiler, 2012).

4 Data

There are different ways in which a text can be considered geographic, so we define three categories of place-based textual data. These categories are not

[1] http://torch.ch/

[2] https://developer.nvidia.com/digits

[3] https://github.com/darwinzer0/utf8-character-cnn

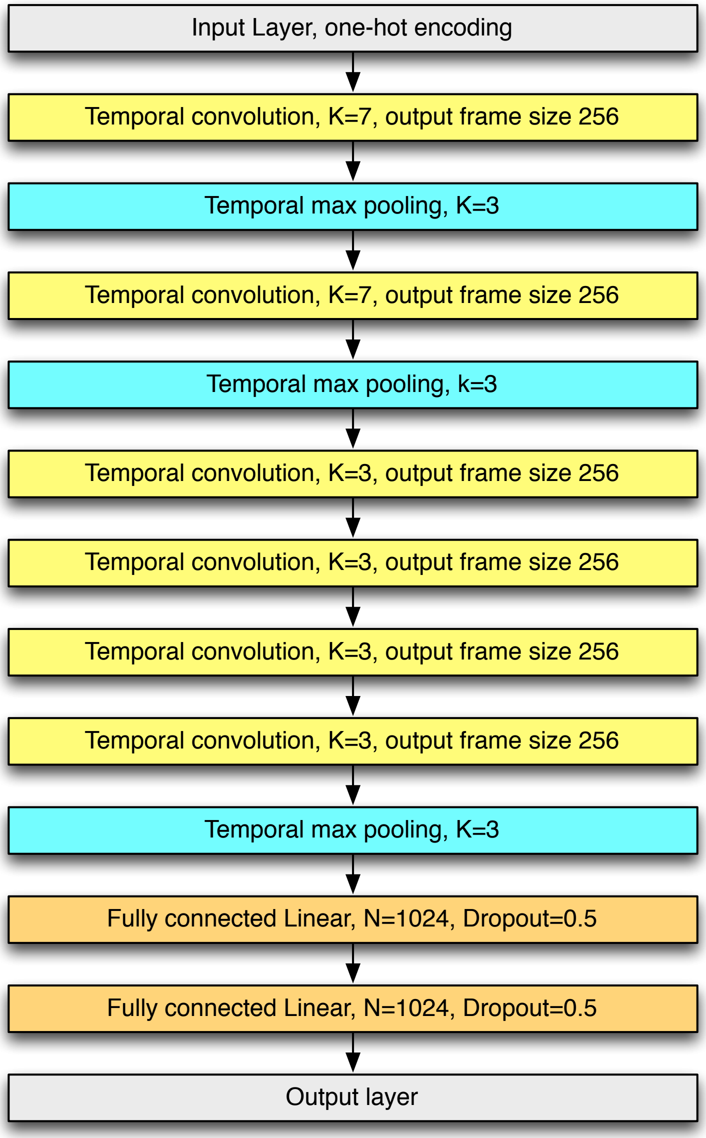

Input Layer, one-hot encoding
Temporal convolution, K=7, output frame size 256
Temporal max pooling, K=3
Temporal convolution, K=7, output frame size 256
Temporal max pooling, k=3
Temporal convolution, K=3, output frame size 256
Temporal convolution, K=3, output frame size 256
Temporal convolution, K=3, output frame size 256
Temporal convolution, K=3, output frame size 256
Temporal max pooling, K=3
Fully connected Linear, N=1024, Dropout=0.5
Fully connected Linear, N=1024, Dropout=0.5
Output layer

Figure 3: This is the basic network used in the experiments. The convolution layers are shown in yellow and correspond to the temporal convolution function described in Eq. 1. The max pooling layers are shown in blue and correspond to the function shown in Eq. 2. The fully connected layers at the bottom of the network are shown in orange. Included in the fully connected layer is a dropout of 0.5, which simply means half of the connections are dropped, and is used to prevent overfitting (Srivastava et al., 2014).

mutually exclusive but help us to define how training data is labeled in our experiments.

1. **PRIMARY** – The first kind of geographic text is based on explicit human sensor observations of geographic places. In other words, these are descriptions of specific places where the main subject of the document is the place in question. They help us to ask questions about "What do people think of the world?". Crowdsourced examples of this kind of data include Wikipedia articles about a place or its history, and travel blog entries contributed by people who have experienced a place. This data can be used to infer thematic or sentiment-based features as well as learn about human geography (Hao et al., 2010; Mitchell, Frank, Harris, Dodds, & Danforth, 2013). Because places are hierarchical (e.g., *California* is a state within the country of the *United States*), the subject need not be direct, but rather can be inferred through the hierarchy. That is, we know an article is about the *United States* if it is about *California.* In Figure 4 the Wikipedia Primary text is **PRIMARY** because it is a paragraph from an article explicitly about a place in Australia (i.e., it is the subject), though no place names are given in the text.

2. **REF** – The second kind of geographic text is text that references a named place or location on the earth within the text. Identifying references to places in text is the subject of geoparsing (Gelernter & Mushegian, 2011), and is useful for a wide variety of geographic information retrieval tools which index based on place names found in the text (Adams, McKenzie, & Gahegan, 2015). Just as with **PRIMARY**, references to places can be inferred through the place hierarchy. In Figure 4 the Wikipedia ref text is **REF** because it references 'India' in the text, although the article itself is not primarily about a place, rather it is about negative numbers.

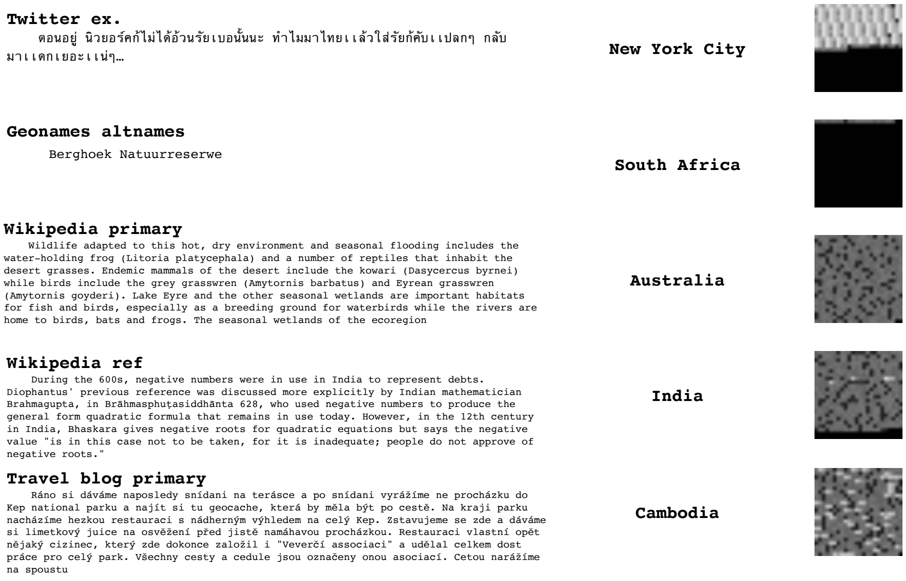
3. **USER** – The final kind of geographic text is any text that is associated with a user in a particular location, e.g., a Twitter tweet that is georeferenced. In other words, this text can include references to places (and thus also be of type **REF**) but often does not. This kind of data has the potential to aid queries that ask the question "Where is the user?". Applications that implement this kind of query include mobile and location-based personalized search tools as well as the analysis of spatial behavior for social science studies (Yi, Raghavan, & Leggetter, 2009; Lee & Sumiya, 2010).

4.1 Data sources

We collected geographic text from seven sources detailed here (see Table 1 for statistics). Figure 4 shows examples of these geographic texts encoded for the UTF-8 character-level CNN.

Table 1: Statistics on the seven data sources used in training

Source	Documents	per class
PRIMARY Wiki Country	480,000	40,000
PRIMARY Travel blog Country	1,550,000	50,000
REF Wiki country	750,000	50,000
REF Geonames Country	3,900,000	100,000
REF Geonames Admin1	3,920,000	40,000
USER Twitter Cities 20	1,000,000	50,000
USER Twitter Hexgrid	9,000,000	100,000



Twitter ex.

ตอนอยู่ นิวยอร์คก้ไม่ได้อ้วนรัยเบอนั้นนะ ทำไมมาไทยเเล้วใส่รัยก้คับเเปลกๆ กลับมาเเดกเยอะเเน่ๆ…

Geonames altnames

Berghoek Natuurreserwe

Wikipedia primary

Wildlife adapted to this hot, dry environment and seasonal flooding includes the water-holding frog (Litoria platycephala) and a number of reptiles that inhabit the desert grasses. Endemic mammals of the desert include the kowari (Dasycercus byrnei) while birds include the grey grasswren (Amytornis barbatus) and Eyrean grasswren (Amytornis goyderi). Lake Eyre and the other seasonal wetlands are important habitats for fish and birds, especially as a breeding ground for waterbirds while the rivers are home to birds, bats and frogs. The seasonal wetlands of the ecoregion

Wikipedia ref

During the 600s, negative numbers were in use in India to represent debts. Diophantus' previous reference was discussed more explicitly by Indian mathematician Brahmagupta, in Brāhmasphuṭasiddhānta 628, who used negative numbers to produce the general form quadratic formula that remains in use today. However, in the 12th century in India, Bhaskara gives negative roots for quadratic equations but says the negative value "is in this case not to be taken, for it is inadequate; people do not approve of negative roots."

Travel blog primary

Ráno si dáváme naposledy snídani na terásce a po snídani vyrážíme ne procházku do Kep national parku a najít si tu geocache, která by měla být po cestě. Na kraji parku nacházíme hezkou restauraci s nádherným výhledem na celý Kep. Zstavujeme se zde a dáváme si limetkový juice na osvěžení před jistě namáhavou procházkou. Restauraci vlastní opět nějaký cizinec, který zde dokonce založil i "Veverčí associaci" a udělal celkem dost práce pro celý park. Všechny cesty a cedule jsou označeny onou asociací. Cetou narážíme na spoustu

Figure 4: Examples of five kinds of geographic text encoded as UTF-8 bytes. The second column is the geographic class for the text. The image is a representation of the 576 byte sequence as an 8-bit greyscale image. Note that different character sets have distinctly different representations in the input data.

4.1.1 Wikipedia PRIMARY articles

The first data set we collected was a collection of the place articles constructed from the July 20, 2016 dump of the English Wikipedia. Place articles were identified by aggregating place pages from the DBpedia resource as well as crowdsourced mappings between places in the Geonames.org gazetteer and Wikipedia pages (Bizer et al., 2009). Each article was then split into paragraphs, which were in turn split to snippets of 576 bytes in length. The **PRIMARY Wiki Country** data set was built using text from articles aggregated to the country level, with 40,000 randomly selected per class. For example, an article about the Louvre in Paris will be categorized as *France*.

4.1.2 Travel blog PRIMARY articles

Travel blogs are a second PRIMARY geographic data source, for which we collected approximately 1 million entries from two popular websites: travelblog.org and travelpod.com. As user-generated content that is less curated than Wikipedia, these data sources contain more misspellings and are written in several different languages. Just as Wikipedia, the **PRIMARY Trav. blog Country** aggregates all articles about places in a country to country categories, with 50,000 entries per class.

4.1.3 Wikipedia REF snippets

These data sets were built from the graph of place references found in all English Wikipedia articles. For this task we utilized a pre-made database of place references in Wikipedia that was developed in (Adams & Gahegan, 2016). Since we were looking for differentiable text, we filtered out all paragraphs that contained more than one place reference and split into snippets of 576-bytes in the same manner as was done for the PRIMARY articles. Subsequently, the place name references were removed from the text. The purpose of this removal being that we wanted to see if there were characteristic language features other than toponyms that were indicative of references to places (Adams & Janowicz, 2012). For example, descriptions of specific geographic features such as mountains or lakes, or country specific geographic terms, such as *oblast* should help improve geo-referencing of ambiguous text.

REF Wiki country will take any text that references a place within a country to that country. For example, a reference to Sydney in an article about a person born there will be classified as *Australia*.

4.1.4 Geonames altnames REF

This source was built from the Geonames.org database of over 11 million named places. Each of those places has a set of alternate names in several different languages, and in order to build a balanced training set, we took a sample from those names (3.9 million from 25 million total) as training examples. The purpose of this data set was to see whether characteristic features of place names

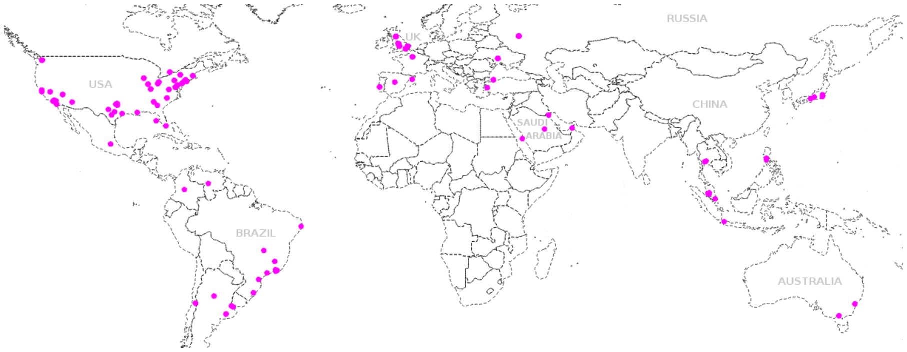


Figure 5: Ninety equal area grid cells with at least 100,000 tweets.

could be learned for geographic categories. Two training sets were built from the Geonames data. The first (**REF Geonames Country**) aggregates the training examples into Countries and the second (**REF Geonames Admin1**) aggregates the examples into Administrative level 1 regions around the world (e.g., United States states).

4.1.5 Twitter tweets, georeferenced USER

Twitter social media microblog text serves as the source for USER geographic text. A collection of 61 million tweets was gathered in mid-2016 and these tweets were used to create two training sets. For both data sets the geographic category is based on the precise location of the tweet when the user's location services are enabled. No filtering was done to remove obviously false geo-tags, though there is evidence that Twitter locations can be quite noisy in that regard (Hecht, Hong, Suh, & Chi, 2011). Tweets without a precise location are not included.

The first (**USER Tw. Cities 20**) was a set of tweets categorized into 20 major cities from around the world based on geo-location within the city limits, with 50,000 tweets randomly sampled per city. The 20 cities are Tokyo, New York, Sao Paulo, Seoul, Mexico City, Manila, Mumbai, Jakarta, Cairo, Los Angeles, Moscow, Istanbul, Paris, Melbourne, London, Toronto, Berlin, Tel Aviv, Rome, Singapore.

The second type of classification (**USER Tw. Hex**) was based on categorizing tweets into equal area grid cells on the earth. The Icosahedral Snyder Equal Area (ISEA) projection was used to generate hexagonal grid cells approximately 60 km across (3,113 km^2) (Snyder, 1992; Adams, 2017). Ninety cells (shown in Figure 5) were chosen with 100,000 tweets randomly sampled per cell.

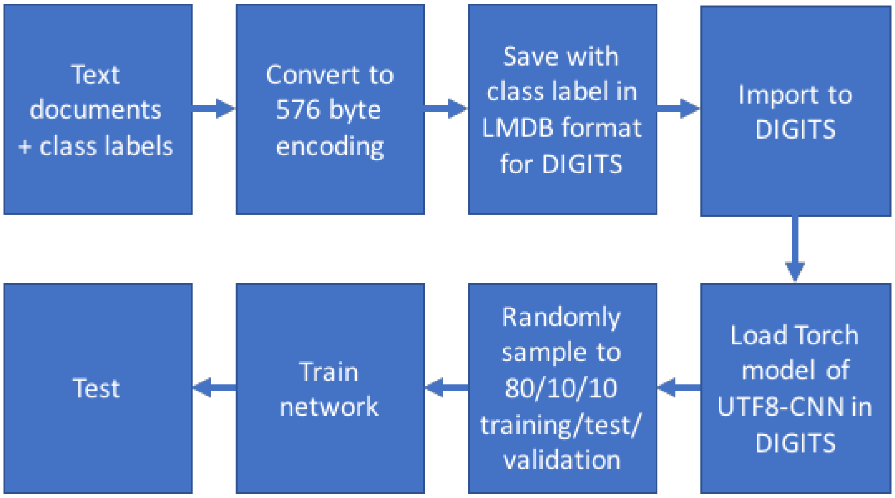

Text documents + class labels
Convert to 576 byte encoding
Save with class label in LMDB format for DIGITS
Import to DIGITS
Load Torch model of UTF8-CNN in DIGITS
Randomly sample to 80/10/10 training/test/ validation
Train network
Test

Figure 6: Workflow for creation of data sets and training.

4.2 Training workflow

Once the training data for each of these sources was collected, each training example was converted into a 576 byte data file illustrated by the greyscale images in Figure 4. The validation and test data sets were generated in the same manner, where 80% of a data set was designated training data, 10% validation data, and 10% test data. Converting the data to an image format allowed the data to be read in by NVIDIA DIGITS, which is a web-based user interface that is designed to manage image training data and train deep learning models. This conversion was simply a convenience to leverage the DIGITS exisiting user interface, which was designed to read in image files for building a training database. After training the models on the training set, the accuracy was calculated by testing the percentage of correctly classified examples from the training sets. Figure 6 shows this workflow process.

5 Experiments

Here we describe the results of our tests of the UTF-8 encoded character-level CNN model against word-based methods for geographic classification. We compared the accuracy of the different methods and, in the case of the Twitter tweets classified into hexagon cells, we calculated the average distance error for the top classification.

5.1 Alternative models

Here we describe the alternate classification models that are used to compare against the UTF-8 character-level CNN model, denoted **UTF8-CNN** in tables.

In all cases in our experiments term frequency-inverse document frequency (TF-IDF) based text classification performed better than simple term frequency (count) based methods, so we only report on TF-IDF results (Salton, Wong, & Yang, 1975). As a baseline we create a bag-of-words TF-IDF matrix, and use that as input to different classifiers. Specifically, we test **UTF8-CNN** against the following other word-based approaches:

1. **Word-TfIdf-NB** – We calculate the bag of words TF-IDF for each document and use that as input to a multinomial naive Bayes classifier (McCallum & Nigam, 1998). The top 80,000 terms are used.

2. **NGram-TfIdf-NB** – Same as the above but the TF-IDF features are based on common N-grams of size 1 to 3.

3. **Word-TfIdf-SVM** – Here we use a linear support vector machine (SVM) on the TF-IDF features (Burges, 1998; Joachims, 1998). The best results shown here are found using parameter tuning on the SVM with α ranging from .0001 to .000001 and regularization terms l2 and elastic net.

4. **NGram-TfIdf-SVM** – Same as **Word-TfIdf-SVM** above with N-grams of size 1 to 3.

5.2 Accuracy results

Table 2 shows the accuracy results for the training methods that were tested. The UTF8-CNN model has the highest accuracy for classifying the alternate names from Geonames by country and by administrative level. In addition, it scores significantly higher in the case of the tweets that were categorized into 20 cities. These 20 cities were chosen deliberately to get a selection of tweets from around the world that would represent several different languages and character sets. For the 90 equal area hexagon cells the vast majority of the hexagons fall in the US, Europe and South America, thus it is likely that the character sets being used in the tweets are predominantly using basic Latin characters with diacritics. The multinomial naive Bayes classifier performs best in this case. One additional curious result is that for the Twitter data, the bag of words results for naive Bayes and SVM were both better than the N-gram datasets. This seems to indicate that for the amount of data being investigated some key individual words were most geographically indicative. For the geographic text that is primarily in English (Wikipedia and travel blog entries) the SVM trained on Ngram-TF-IDF works the best.

There is a large difference between the accuracy results when viewed across the different datasets, which points to the difficulty of the classification task in some cases (e.g., classifying Twitter tweets to the top-1 hexagon out of 90 candidates). Nevertheless, there is value in such a seemingly inaccurate classifier when looking at top-N results instead of top-1 and when it is combined with other information for ensemble learning or human-in-the-loop augmented analysis (cf. the similar problem for image georeferencing (Hays & Efros, 2008)).

Table 2: Accuracy matrix (percent true positives) for different geographic classifiers of text. Best performing model for each data set in bold.

Model	Word-TfIdf-NB	NGram-TfIdf-NB	Word-TfIdf-SVM	NGram-TfIdf-SVM	UTF8-CNN
USER Tw. Hex	**21.4**	19.2	20.3	17.6	19.8
USER Tw. Cities 20	42.1	40.0	41.5	39.4	**50.1**
REF Geonames Country	50.7	50.2	49.4	49.4	**70.2**
REF Geonames Admin 1	25.4	24.7	21.3	21.2	**29.3**
PRIMARY Wiki Country	84.9	87.3	90.9	**91.4**	85.8
REF Wiki Country	74.7	77.5	85.3	**87.1**	85.0
PRIMARY Trav. blog	51.0	60.0	62.0	**67.6**	60.0

Table 3: Average distance error for USER Tw. Hex per classification model. Lowest average distance error in bold.

Model	Avg. Distance (km)
Word-TfIdf-NB	2493.6
NGram-TfIdf-NB	2868.7
Word-TfIdf-SVM	2617.9
NGram-TfIdf-SVM	2702.8
UTF8-CNN	**2318.1**
Random	7872.4

Overall, the two situations where a character-level approach seems to have the most advantage are when several different languages are being used and these are indicative of geographic categories, and the second being when there is a strong difference in the kinds of words being associated with instance of the categories. For example, in the case of the Geonames altnames database the word-based methods do not work nearly as well because there are not that many terms or n-grams being used across multiple training examples. Instead the regularities that can be learned are at the the sub-word level, e.g., common character combinations, prefixes, suffixes, etc. that are found in one set of place names versus another.

5.3 Geographic accuracy

For the tweet dataset organized into equal area cells (**USER Tw. Hex**) we also calculated the average distance between the ground-truth location and the predicted location. For randomly assigned classification based on the 90 classes we expect a distance error on the spheroid to be 7872.4 km. All classifiers perform significantly better than random (see Table 3), but UTF8-CNN model has the smallest error. This is an indication that the character-level classification captures some more geographic regularities that operate on the character-level rather than the word-level, which agrees with the accuracy results for the Geonames data set discussed previously.

5.4 Robustness to noise

We created synthetic versions of the **PRIMARY Wiki Country** data set with errors, in order to test the robustness of the various models to noise such as misspellings or optical character recognition (OCR) errors in the case of scanned historical documents. Four new versions were made with 2%, 5%, 10%, 15% of UTF-8 characters randomly changed, respectively. Figure 7 shows one snippet from the English Wikipedia that has had noise synthetically added for each percentage. Table 4 shows the results, which indicate that the UTF8-CNN model is more robust to noise. For example, the net negative effect of 5% noise on the UTF8-CNN model is 2.6%, whereas it is 4.8% for the NGram-TfIdf-SVM model.

6 Conclusion and Future work

In this paper we introduced a new UTF-8 encoding for character-level convolutional neural network models. We demonstrated that this approach can work for geographic classification of text and appears to capture language-based regularities (with respect to character set encodings) better than word-based methods. CNN models, in general, perform better with more data and deeper models, so different layer configurations for the CNN, including deeper models or with more parameters will likely perform even better. In addition, the character-level CNN is more resilient to synthetic noise added to the data sets.

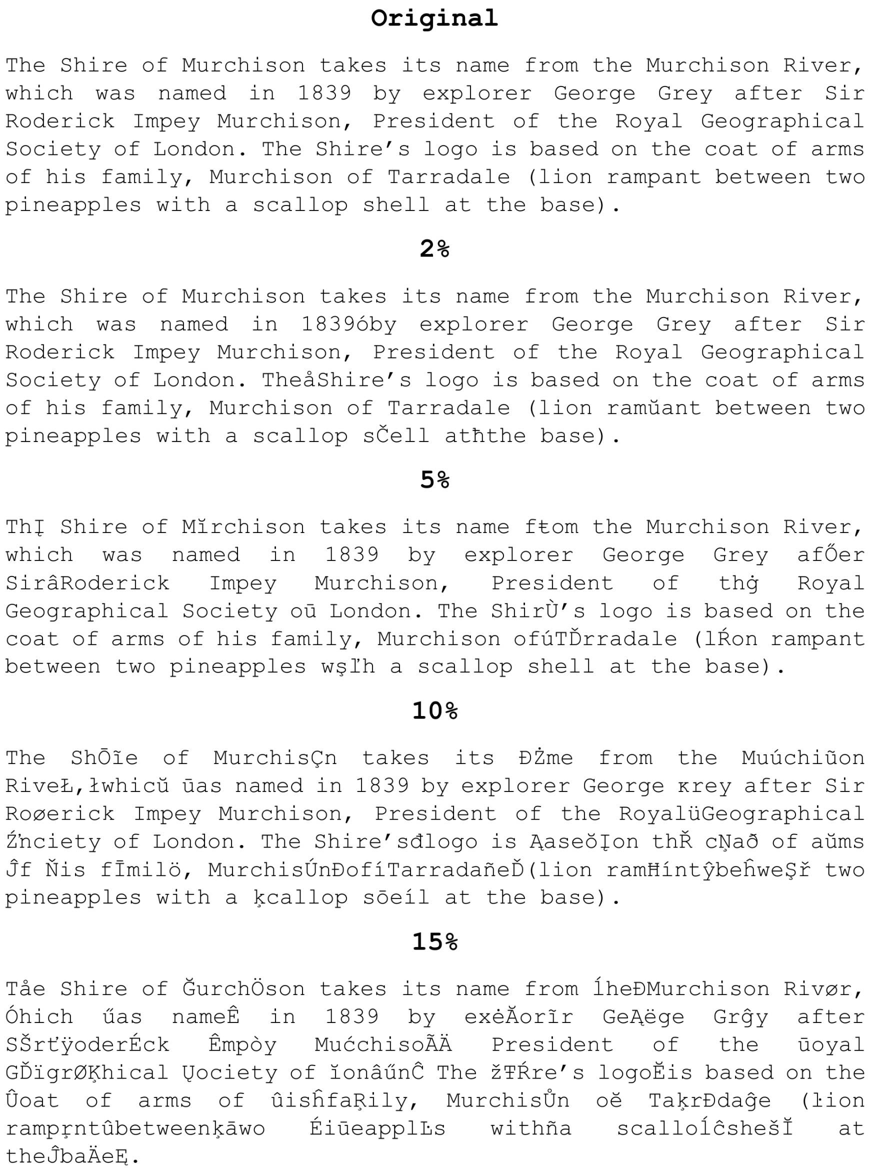
At present, it appears that word based methods are still preferable in cases where the text is primarily in one language and has few character-level errors. In addition, the training of CNNs is significantly more compute intensive than other methods such as Naive Bayes or SVM, even with optimized GPU code. A challenge going forward for all the models is how to scale the geographic classification up to thousands of classes and to train for hierarchical classes such as is common with geographic categories. Another area of future work is to investigate hybrid models that combine this method with top-down spatial semantic modeling.

Although the focus of this work was on classification of text explicitly related to geographic places, the UTF-8 level encoding of the input layer is a technique that extends the character-level convolutional neural network model to textual data of any language. Thus, it can also applied to classifying any kind of text source containing documents that come in many different languages. In addition, because it uses the same basic building blocks as image-based CNN models, there exists an opportunity to create joint text and image based geographic classifiers, e.g., on social media posted photographs of places with associated text. The results showing robustness to noise suggest that character-level CNNs might be useful for OCR error-correction algorithms (Kolak & Resnik, 2002). In the future, we intend to apply this model for improved place name disambiguation, look into how adding noise to the CNN classifier can improve classification results, such as with image classification, and explore the transfer of knowledge about places from one crowdsourced data set to another.

References

Adams, B. (2017). Wāhi, a discrete global grid gazetteer built using linked open data. *International Journal of Digital Earth*, *10*(5), 490–503.

Adams, B., & Gahegan, M. (2016). Exploratory chronotopic data analysis. In *International conference on geographic information science* (pp. 243–258).



Original

The Shire of Murchison takes its name from the Murchison River, which was named in 1839 by explorer George Grey after Sir Roderick Impey Murchison, President of the Royal Geographical Society of London. The Shire’s logo is based on the coat of arms of his family, Murchison of Tarradale (lion rampant between two pineapples with a scallop shell at the base).

2%

The Shire of Murchison takes its name from the Murchison River, which was named in 1839óby explorer George Grey after Sir Roderick Impey Murchison, President of the Royal Geographical Society of London. TheåShire’s logo is based on the coat of arms of his family, Murchison of Tarradale (lion ramŭant between two pineapples with a scallop sČell atħthe base).

5%

ThĮ Shire of Mĭrchison takes its name fŧom the Murchison River, which was named in 1839 by explorer George Grey afŐer SirâRoderick Impey Murchison, President of thġ Royal Geographical Society oū London. The ShirÙ’s logo is based on the coat of arms of his family, Murchison ofúTĎrradale (lŔon rampant between two pineapples wşľh a scallop shell at the base).

10%

The ShŌĭe of MurchisÇn takes its ĐŻme from the Muúchiũon RiveŁ,łwhicŭ ūas named in 1839 by explorer George ĸrey after Sir Roøerick Impey Murchison, President of the RoyalüGeographical Źńciety of London. The Shire’sđlogo is ĄaseŏĮon thŘ cŅað of aŭms Ĵf Ňis fĪmilö, MurchisÚnÐofíTarradañeĎ(lion ramĦíntŷbeĥweŞř two pineapples with a ķcallop sōeíl at the base).

15%

Tåe Shire of ĞurchÖson takes its name from ÍheÐMurchison Rivør, Óhich űas nameÊ in 1839 by exėĂorĩr GeĄëge Grĝy after SŠrťÿoderÉck Êmpòy MućchisoÃÄ President of the ūoyal GĎïgrØĶhical Ųociety of ĭonâűnĈ The žŦŔre’s logoĔis based on the Ûoat of arms of ûisĥfaŖily, MurchisŮn oĕ TaķrĐdaĝe (ŀion rampŗntûbetweenķāwo ÉiūeapplĿs withña scalloÍĉshešĬ at theĴbaÄeĘ.

Figure 7: Sample from the synthetic dataset demonstrating noise introduced at the levels of 2%, 5%, 10%, and 15% of the characters.

Table 4: Accuracy rates (pct) and net loss (pct difference) for PRIMARY Wiki Country with synthetic character-level errors introduced. Best performing model for each data set in bold.

Error	Word-TfIdf-NB		NGram-TfIdf-NB		Word-TfIdf-SVM		NGram-TfIdf-SVM		UTF8-CNN	
0%	84.9	-	87.3	-	90.9	-	91.4	-	85.8	-
2%	83.0	-1.9	85.3	-2.0	88.8	-2.1	89.2	-2.2	84.7	**-1.1**
5%	80.6	-4.3	83.0	-4.3	86.5	-4.4	86.8	-4.6	82.8	**-3.0**
10%	77.3	-7.6	79.9	-7.4	83.2	-7.7	83.5	-7.9	80.8	**-5.0**
15%	74.7	-10.2	77.3	-10.0	80.0	-10.9	80.8	-10.6	78.3	**-7.5**

Adams, B., & Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of the sixth international AAAI conference on weblogs and social media* (pp. 375–378).

Adams, B., McKenzie, G., & Gahegan, M. (2015). Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th international conference on world wide web* (pp. 12–22).

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 273–280).

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000). An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd annual international acm sigir conference on research and development in information retrieval* (pp. 160–167).

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, *7*(3), 154–165.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Cheng, G., Zhou, P., & Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(12), 7405–7415.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 759–768).

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on geographic information retrieval* (pp. 25–30).

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS workshop.*

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *AAAI conference on artificial intelligence* (p. 2382-2388).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277–1287).

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, *15*(6), 753–773.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

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-volume 1* (pp. 368–378).

Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, *49*, 451–500.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-M., Pang, Y., & Zhang, L. (2010). Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on world wide web* (pp. 401–410).

Hays, J., & Efros, A. A. (2008). Im2gps: estimating geographic information from a single image. In *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on* (pp. 1–8).

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 237–246).

Hovy, D., Johannsen, A., & Søgaard, A. (2015). User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on world wide web* (pp. 452–461).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137–142).

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., & Weibel, R. (2002). Spatial information retrieval and geographical ontologies an overview of the spirit project. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 387–388).

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, *22*(3), 219–228.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *Knowledge engineering and knowledge management: 20th international conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, proceedings 20* (pp. 353–367).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural lan-*

guage processing (EMNLP) (p. 1746-1751).

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *AAAI* (pp. 2741–2749).

Kolak, O., & Resnik, P. (2002). OCR error correction using a noisy channel model. In *Proceedings of the second international conference on human language technology research* (pp. 257–262).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Laere, O. V., Schockaert, S., Tanasescu, V., Dhoedt, B., & Jones, C. B. (2014). Georeferencing wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS)*, *32*(3), 12.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of naacl-hlt* (pp. 260–270).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks* (pp. 1–10).

Liu, J., & Inkpen, D. (2015). Estimating user location in social media with stacked denoising auto-encoders. In *Proceedings of NAACL-HLT* (pp. 201–210).

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), 645–657.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41–48).

Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, *21*, 3–38. doi: 10.1111/tgis.12212

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, *8*(5), e64417.

Monteiro, B. R., Davis, C. A., & Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, *96*, 23–34.

Nogueira, K., Penatti, O. A., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539–556.

Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, *22*(3), 265-287. doi: 10.1080/13658810701626236

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3), 1.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860).

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., … Ungar, L. (2013). Characterizing geographic variation in well-being using tweets. In *Proceedings of the seventh international AAAI conference on weblogs and social media* (pp. 583–591).

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of ICML-99, 16th international conference on machine learning* (pp. 379–388). Bled, SL: Morgan Kaufmann Publishers, San Francisco, US.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1–47.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., & Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, *30*(4), 378–399.

Snyder, J. P. (1992). An equal-area map projection for polyhedral globes. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *29*(1), 10–21.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Subramaniam, L. V., Roy, S., Faruquie, T. A., & Negi, S. (2009). A survey of types of text noise and techniques to handle noisy text. In *Proceedings of the third workshop on analytics for noisy unstructured text data* (pp. 115–122).

Sui, D., & Goodchild, M. (2011). The convergence of gis and social media: challenges for giscience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Wehrmann, J., Becker, W., Cagnini, H. E., & Barros, R. C. (2017). A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *Neural networks (IJCNN), 2017 international joint conference on* (pp. 2384–2391).

Wing, B. P., & Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 955–964).

Yergeau, F. (2003). *UTF-8, a transformation format of ISO 10646* (Tech. Rep.). Retrieved from `https://tools.ietf.org/html/rfc3629`

Yi, X., Raghavan, H., & Leggetter, C. (2009). Discovering users' specific geo intention in web search. In *Proceedings of the 18th international conference on world wide web* (pp. 481–490).

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, *27*(6), 52–59.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014*(9), 37–70.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).